Human diseases and exposure to pesticides in developing countries: temporal spatial analysis, risk mitigation strategies and spatial decision support system

by

Jörg Matthias Rapp

from San Miguel de Tucuman /Argentina

Accepted Dissertation thesis for the partial fulfilment of the requirements

for a

Doctor of Natural Science

Fachbereich Natur- und Umweltwissenschaften

Rheinland-Pfälzische Technische Universität

Kaiserslautern-Landau

Thesis examiners:

Prof. Dr. Engelbert Niehaus

Prof. Dr. Ralf Schulz

Date of the oral examination: 05.09.2025

Frontmatter

Kurzfassung/ Abstract

Abstract

In developing countries, an increasing number of cases of chronic kidney disease (CKD) is observed, in which traditional risk factors for CKD can be ruled out. In such cases, the term chronic kidney disease of unknown etiology (CKDu) is used. A causal relationship between potential risk factors and the onset of the disease has not been clearly established, yet. However, agrochemicals are suspected to be associated with the disease. This dissertation aims to contribute to the reduction of exposure to the potential risk factor agrochemicals.

The first part of the work describes the characteristics of CKDu, the situation in developing countries, and their agricultural practices. Developing countries and their agricultural sectors are often marked by a lack of financial and infrastructural resources as well as comparatively low levels of education. The occurrence of CKDu has been documented in various regions around the world.

Based on this, the second part of the work adapts a research and development environment to local conditions, using the concept Living Lab. The Living Lab serves to develop, validate, and improve tailored risk mitigation strategies concerning agrochemical use. These strategies are introduced by various stakeholders involved in the Living Lab as potential solutions. Both, the Living Lab itself and the resulting risk mitigation measures are designed as low-cost approaches, for instance through the use of open-source software.

The third part of the dissertation describes and develops exemplary low-cost risk mitigation strategies from the fields of ecotoxicology and mathematical modeling. These serve as a base for further development of risk mitigation strategies within the research and development cycle of the Living Lab.

In the final part of the work, various mathematical methods are presented for the development of an adaptive decision support system. This system is capable of processing spatial, temporal, and fuzzy data to generate appropriate decision support related to the potential

risk factor agrochemicals.

Kurzfassung

In Entwicklungsländern ist eine zunehmende Anzahl von Fällen chronischer Niereninsuffizienz (CKD) zu beobachten, wobei traditionelle Risikofaktoren für CKD ausgeschlossen werden können. In solchen Fällen spricht man von chronischer Niereninsuffizienz unbekannter Herkunft (CKDu). Eine ursächliche Verbindung zwischen potenziellen Risikofaktoren und dem Auftreten der Krankheit konnte bislang nicht eindeutig nachgewiesen werden. Es besteht jedoch der Verdacht, dass Agrochemikalien mit der Krankheit in Zusammenhang stehen könnten. Die vorliegende Dissertation soll einen Beitrag zur Reduzierung der Exposition gegenüber dem potenziellen Risikofaktor Agrochemikalien leisten.

Im ersten Teil der Arbeit werden zunächst die Charakteristika von CKDu, die Situation in Entwicklungsländern sowie die dortige Landwirtschaft beschrieben. Entwicklungsländer und deren Agrarsektor sind häufig durch einen Mangel an finanziellen und infrastrukturellen Ressourcen sowie ein vergleichsweise niedriges Bildungsniveau gekennzeichnet. Das Auftreten von CKDu ist in verschiedenen Regionen weltweit dokumentiert.

Auf dieser Grundlage wird im zweiten Teil der Arbeit eine Forschungs- und Entwicklungsumgebung basierend auf dem Konzept des Living Lab an die beschriebenen Gegebenheiten angepasst. Das Living Lab dient der Entwicklung, Validierung und Verbesserung angepasster Risikominimierungsstrategien im Umgang mit Agrochemikalien. Diese Strategien werden von den verschiedenen beteiligten Akteuren im Rahmen des Living Lab als Lösungsvorschläge eingebracht. Sowohl das Living Lab als auch die entwickelten Maßnahmen zur Risikominimierung werden als kostengünstige Methoden konzipiert, beispielsweise durch den Einsatz von Open-Source-Software.

Im dritten Teil der Arbeit werden exemplarisch Low-Cost-Risikominimierungsstrategien aus den Bereichen Ökotoxikologie und mathematische Modellbildung beschrieben und entwickelt. Diese dienen als Grundlage für weiterentwickelbare Strategien innerhalb des Forschungs- und Entwicklungszyklus des Living Lab.

Im abschließenden Teil der Arbeit werden verschiedene mathematische Methoden vorgestellt, mit deren Hilfe ein adaptives Entscheidungsunterstützungssystem entwickelt werden kann. Dieses System ist in der Lage, räumliche, zeitliche und unscharfe Informationen zu verarbeiten und daraus fundierte Entscheidungshilfen im Bereich der Risikominimierung abzuleiten.

Contents

Frontmatter **ii**

Kurzfassung/ Abstract **iv**

Contents **viii**

Abbrevations **xvi**

I Introduction: CKDu, agriculture, pesticides, risk and developing countries **1**

1 Introduction **3**

1.1 Problem formulation 3

1.2 Aim of the thesis and research hypotheses 7

1.3 Structure of the thesis 9

1.4 Related projects 9

2 CKDu, pesticides, risk, and the situation in less-developed countries **11**

2.1 CKDu and less-developed countries 11

2.1.1 CKDu, a disease with unknown etiology 11

2.1.2 Introduction to the term less-developed countries 17

2.1.3 El Salvador and the pilot region Bajo Lempa 19

2.1.4 Health situation in less-developed countries and factors influencing access to the healthcare system 22

2.2 Pesticides and CKDu 24

2.2.1 Pests, pesticides, and alternative pest management strategies 24

2.2.2 Pesticides in the environment, fate, and their behavior 28

2.2.3 Human exposure on different levels and uptake routes 33
2.2.4 Agriculture in less-developed countries, risk groups and factors influencing the risk 37
2.2.5 Models to estimate the fate of chemicals and human exposure 41
2.2.6 Effects for public health and ecosystems caused by pesticides 44
2.2.6.1 One Health approach 44
2.2.6.2 Hazardous effect of pesticides: toxicity 45
2.2.6.3 Effects of pesticides on human health 47
2.2.6.4 Ecological problems caused by pesticides 48
2.2.6.5 Pesticides and their benefits 50
2.3 Chapter conclusion 51

3 Risk and risk assessment 52
3.1 The term risk as it relates to pesticides 52
3.1.1 Introduction to the term risk and some definitions 52
3.1.2 Discussion of the relation between the risk definitions 56
3.2 Risk assessment and risk management 57
3.2.1 Human risk assessment 58
3.2.2 Ecotoxicological risk assessment 61
3.2.3 Mixture toxicity 62
3.2.4 Pesticide legislation in developing countries 63
3.3 Chapter conclusion 63

4 Mathematical modeling and CKD 65
4.1 Mathematical modeling and system analysis 65
4.2 Chapter conclusion 69

II Research and development environment and data sampling 71

5 A research and development environment in the described framework 73
5.1 Introduction 73
5.2 Requirements for a research and development environment 75
5.3 A Living Lab as research and development environment in the proposed framework 78

5.3.1 The theory of a Living Lab 78
5.3.2 A Living Lab in the proposed framework and adaptations 82
5.3.2.1 Open community approach and open source 83
5.3.2.2 Selection of stakeholders 87
5.3.2.3 Pilot region 88
5.3.2.4 Motivation for community members to participate 88
5.4 Establishing a Living Lab 89
5.5 Chapter conclusion 90

6 Data sampling in a Living Lab in less-developed countries 92
6.1 Crowdsourcing 92
6.2 Citizen science 93
6.3 Information and Communication Technology (ICT) in crowdsourcing and citizen-science projects 96
6.4 Data sampling and delivering in less-developed countries 98
6.5 Requirements for an app for data sampling and delivering 99
6.6 Example for an open-source app for citizen-science projects in less-developed countries 100
6.7 SDAPS as a sensor for complex and dynamic systems 102
6.8 Citizen science in the described approach 103
6.9 Chapter conclusion 104

7 Dealing with incomplete data and quality of data 106
7.1 Introduction 106
7.2 Interpolation and extrapolation of temporal and spatial data 107
7.3 Data density, interpolation quality and data quality 118
7.4 Methods for dealing with missing data 125
7.5 Chapter conclusion 127

8 LLinES application and a Living Lab in El Salvador 129

III Initial knowledge base for risk mitigation strategies 133

9 Risk mitigation strategies in an agrochemical related Living Lab (LL) 135

9.1 Requirements for risk mitigation strategies in less-developed countries in an LL approach . . . 135
9.2 Possible risk mitigation strategies . . . 137

10 The adaptation of precision farming and Variable-rate technology (VRT) for less-developed countries 139
10.1 Introduction: precision farming and VRT . . . 139
10.2 Vegetation indices derived by remote sensing . . . 141
10.3 Remote sensing-based VRT as a risk mitigation strategy in less-developed countries . . . 143
10.4 Remote-sensed data derived from satellite images . . . 147
10.5 Processing of satellite images to derive vegetation indices . . . 149
10.6 From vegetation indices to spatial decision support . . . 153
10.7 Interim conclusion . . . 160

11 Creation of pesticide application maps 162
11.1 Methods . . . 162
11.1.1 Software used . . . 162
11.1.2 Used external maps and their modification . . . 163
11.1.3 Data used to calculate yearly application rates per crop and country . 168
11.1.4 Data used to validate the model . . . 173
11.1.5 Creation of application maps in Geographic Information System (GIS) and validation analysis . . . 174
11.2 Results . . . 175
11.3 Discussion . . . 181
11.4 Use of the described model and application maps for an Spatial Decision Support System (SDSS) in an LL . . . 188

12 Generating pesticide application maps with crowdsourcing 190
12.1 Overview over the framework . . . 190
12.2 Transmitted data items by citizen scientists . . . 190
12.3 From static to dynamic risk maps: fate models . . . 193
12.4 Results and discussion: generating risk maps with crowdsourcing . . . 194

13 Calculating the individual impact induced by a walk through a contaminated area **195**

13.1 Needed data to estimate the individual impact 195

13.2 Graph theory 197

13.3 Concentration and uptake reduction function 202

13.4 Integral over a path 203

13.5 Finding a way with a minimum of exposure 205

13.6 Discussion and conclusion 206

14 Part conclusion **208**

IV SDSS **209**

15 SDSS and requirements in the described LL approach **211**

15.1 Problem, aim and background of an SDSS in the described framework 211

15.2 Requirements for an SDSS in the described framework 213

16 Fuzzy sets, fuzzy logic, and Artificial Neural Network (ANN) and their use in an SDSS **216**

16.1 Introduction: the use of fuzzy sets and fuzzy logic in the described LL framework 216

16.2 Fuzzy set theory 217

16.2.1 Introduction to fuzzy set theory 217

16.2.2 Characteristics of fuzzy sets 224

16.2.3 Relations of fuzzy sets 227

16.2.4 Fuzzy Logic Controller theory 236

16.2.5 The application of fuzzy logic for generating SDSS in an LL approach 242

16.2.6 Limitations in the application of a standalone Fuzzy Logic Controller (FLC) 251

16.3 Artificial neural networks 252

16.3.1 Artificial neurons and biological background 253

16.3.2 The perceptron 256

16.3.3 Introduction to ANNs 259

16.4 The application of ANN in combination with FLC in an LL framework 264

16.4.1 A concurrent neuro-fuzzy system to give SDSS with incomplete or noisy data . . . 268
16.4.1.1 Data completion and noise reduction with autoassociative networks . . . 269
16.4.1.2 Autoassociative ANN with supervised learning: multilayer perceptron and the backpropagation algorithm . . . 270
16.4.1.3 Interim conclusion: a concurrent neuro-fuzzy to pre-process incomplete or noisy data . . . 277
16.4.2 A neuro-fuzzy system to determine parameters for a FLC . . . 278
16.4.3 A neuro-fuzzy system to determine rules for a FLC . . . 283
16.4.3.1 Generating fuzzy rules with the fuzzy C-means (FCM) algorithm 284
16.4.3.2 Generating fuzzy rules with the Wang and Mendel algorithm 287
16.4.4 An algorithm for generating fuzzy rules and membership function optimization . . . 289
16.4.5 Interim conclusion . . . 294
16.5 Chapter conclusion: fuzzy logic and ANN . . . 295

17 Data delivering and needed ICT infrastructure 296
17.1 Media to deliver spatial support . . . 296
17.2 Online and offline mode . . . 297
17.3 Hardware . . . 298
17.4 Possible software solutions for the described LL approach . . . 300
17.5 Chapter conclusion . . . 302

18 Development of an SDSS in the described LL framework 303
18.1 Mathematical description of the adaptive mapping machine in the SDSS . . . 303
18.2 Domain and co-domain of the proposed SDSS . . . 304
18.3 Developed SDSS . . . 305
18.4 Chapter discussion . . . 308

19 Discussion and conclusion 310

Appendices 315

A Equations for the calculation of application rates 316

B Maps visualizing the crop responsible for the highest applied substance amount **318**

C Table tasks in the LL and possible open-source software solutions **321**

Bibliography **325**

List of Figures **405**

List of Tables **413**

Index **414**

Erklärung **415**

Danksagung **417**

Lebenslauf **419**

Abbreviations

6S *Second Simulation of the Satellite Signal in the Solar Spectrum*
ACR Albumin Creatinine Ratio
ADI acceptable daily intake
a.i. active ingredient
AI artificial intelligence
ANFIS *Adaptive Neuro-Fuzzy Inference System*
ANN Artificial Neural Network
AR augmented reality
ASTER *Advanced Spaceborne Thermal Emission and Reflection Radiometer*
AT6 FUI *Action Team 6 Follow up Initiative*
AVHRR *Advanced Very High Resolution Radiometer*
CAPE Continuous Ambulatory Peritoneal Dialysis
CKD chronic kidney disease
CKDu chronic kidney disease of unknown etiology
COA *Center-of-Area Method*
COM *Center-of-Maximum Method*
CPU central processing unit
CSIR *Council for Scientific and Industrial Research*
CSV *character-separated values*
DDT Dichloro-Diphenyl-Trichloroethane
DEM digital elevation model
DENFIS *Dynamic evolving neural-fuzzy inference system*
DNA Deoxyribonucleic Acid
DOS *dark-object subtraction*
DSS Decision Support System
ECM evolving clustering method
EFSA *European Food Safety Authority*
EU *European Union*
FALCON *Fuzzy Adaptive Learning Control Network*
FAO *Food and Agriculture Organization of the United Nations*
FCM fuzzy C-means
FLC Fuzzy Logic Controller
FOCUS *FOrum for the Co-ordination of pesticide fate models and their USe*
GDEM *Global Digital Elevation Model*
GDP Gross Domestic Product
GFR Glomerual Filtration Rate
GIS Geographic Information System
GNI Gross National Income
GUI graphical user interface
GNU GPL *GNU General Public License*
GPRS General Packet Radio Service
GPS Global Positioning System
HDI Human Development Index
HyFIS *hybrid neural fuzzy inference system*
ICT Information and Communication Technology
IPM Integrated Pest Management
IT Information Technology
IRSS *Influential Rule Search Scheme*

k-NN *K-Nearest-Neighbor*
LC Lethal Concentration
LD Lethal Dose
LL Living Lab
LLinES Living Lab in El Salvador
LLM *large language model*
ML machine learning
MOM *Mean-of-Maxima Method*
MSDS material safety data sheet
NAIP *National Agriculture Imagery Program*
NASA *National Aeronautics and Space Administration*
NDVI *Normalized Difference Vegetation Index*
NEFPROX *NEuro Fuzzy function apPROXimation*
NOAA *National Oceanic and Atmospheric Administration*
NOAEL No Observed Adverse Effect Level
NOEC No Observed Effect Concentration
NOEL No Observed Effect Level
NPUD 2002 *National Pesticide Use Database 2002*
NSAID non-steroidal anti-inflammatory drugs
ODK *Open Data Kit*
OECD *Organisation for Economic Co-operation and Development*
OSM *OpenStreetMap*
PEC Predicted Environmental Concentration
POES *Polar Orbiting Environmental Satellites*
POP Persistent Organic Pollutants
QA quality assessment
Q-Q-plot Quantile-Quantile-plot
QUEFTS *Quantitative Evaluation of the Fertility of Tropical Soils*
RAM random-access memory
ReGLaN *Research-Group Learning and Neurosciences*
RfC reference concentration
RfD reference dose
SAVI *Soil-adjusted Vegetation Index*
SDAPS Scripts for data acquisition with paper-based surveys
SDSS Spatial Decision Support System
SIM subscriber identification module
SMS Short Message Service
SONFIN *Self-cOnstructing Neural Fuzzy Inference Network*
SRTM *Shuttle Radar Topography Mission*
TER Toxicity Exposure Ratio
TV television
UN *United Nations*
UNEP *United Nations Environment Programme*
URL uniform resource locator
US EPA *United States Environmental Protection Agency*
USA United States of America
USB Universal Serial Bus
USGS *United States Geological Survey*
VRT Variable-rate technology
WHO *World Health Organization*
WiFi Wireless Fidelity

I | Introduction: CKDu, agriculture, pesticides, risk and developing countries

1 Introduction

1.1 Problem formulation

Globally, around 750 million people are affected by a disease called chronic kidney disease (CKD) [BPR+18]. Traditional risk factors for CKD are, inter alia, hypertension, diabetes mellitus, and obesity [ET11].

However, in less-developed countries a steadily increasing number of people suffering from CKD can also be observed for about three decades, and traditional risk factors for CKD can be excluded as cause for these disease cases [(PA17]. In these cases, when traditional risk factors can be excluded, the disease is called chronic kidney disease of unknown etiology (CKDu). To date, a causality between potential risk factors besides the mentioned traditional risk factors could not be found, and only different hypotheses about the origin of CKDu and related risk factors are mentioned in scientific literature. Possible nontraditional risk factors are exposure to agrochemicals [SLW+10, OHA+11] , dehydration and hard working conditions [OTW+10], simultaneous exposure to pesticides and hard water [JGS14], or intake of alcohol, nicotine, and non-steroidal anti-inflammatory drugs (NSAID) [RJLJ+15]. A characteristic group of people affected by CKDu cannot be determined; it is observed in males and females, children and adults, people working in agriculture and people not involved in the agricultural process. CKDu can also be observed in different regions of the world, e.g., Central America, Asia, Northern Africa, or in the Balkan states [(PA17]. Instances of high case numbers of CKDu share commonalities, i.e., they are mostly located in less-developed countries with the accompanying characteristics, e.g., low economic performance, an underdeveloped healthcare system, a high number of people living in rural areas, and an insufficient infrastructure [fECG19].

In some of the countries and areas with high CKDu case numbers, CKDu leads to a high degree of suffering in the population, which is caused by high mortality, the related negative effects on the populace, and a high level of effort in the treatment of people affected by CKDu.

This leads to a level of high pressure on the healthcare system in affected regions [OHA+11]. This thesis was inspired, inter alia, by the case of CKDu in El Salvador and is located within a scientific project called Living Lab in El Salvador (LLinES), funded by the German Federal Ministry of Education and Research. A detailed description of the project can be found in section 1.4. The aim of the project was to initiate a research and development environment called LL (section 5.3.1) in a CKDu affected rural community in El Salvador in a multidisciplinary open community approach. In the LL, possible risk mitigation strategies should be developed and evaluated. Additionally, the LL acts as an sensor for data sampling. The approach described in this thesis should be a generic approach not only usable for the case of CKDu in El Salvador but also for other agrochemical related diseases with a unknown etiology and temporal and spatial dimensions.

Finding and establishing countermeasures against CKDu has two main obstacles: First, as the etiology of CKDu is unknown at the moment, countermeasures against CKDu cannot get to the root of the problem; they can only treat symptoms of CKD and try to reduce the exposure to suspected risk factors. Second, as long as there is no causality found between a risk factor and the outbreak of CKDu, mechanisms for reducing the case numbers and severity of the disease could focus inter alia in three fields: treating symptoms of CKD, helping to find the structure behind the epidemic of CKDu in less-developed countries, and, until the structure and mechanisms are discovered, mitigating suspected risk factors, like avoiding unnecessary exposure to pesticides, alcohol, nicotine, dehydration etc.

The results of this thesis contribute to improving the situation related to CKD in less-developed countries. Therefore, methods from three main disciplines, namely mathematical modeling, (eco-)toxicology, and research design (LL approach) are combined with the overall aim of mitigating the risk for CKDu in the affected population. In figure 1.1, this issue is visualized; the context of this thesis lies in the intersection between these three disciplines.

In this thesis methods from the discipline mathematical modeling are used to generate temporal and spatial risk maps for CKD, develop an algorithm helping to detect the structure behind the disease, and give spatial decision support in terms of mitigating the risk factors for CKD. Methods from (eco-)toxicology are needed in this framework because substances that can act as toxins, e.g., heavy metals and pesticides, are suspected risk factors for CKDu, and knowledge about the fate and behavior of these substances might help to reduce exposure to the substances. Methods from research design are required because existing risk mitigation

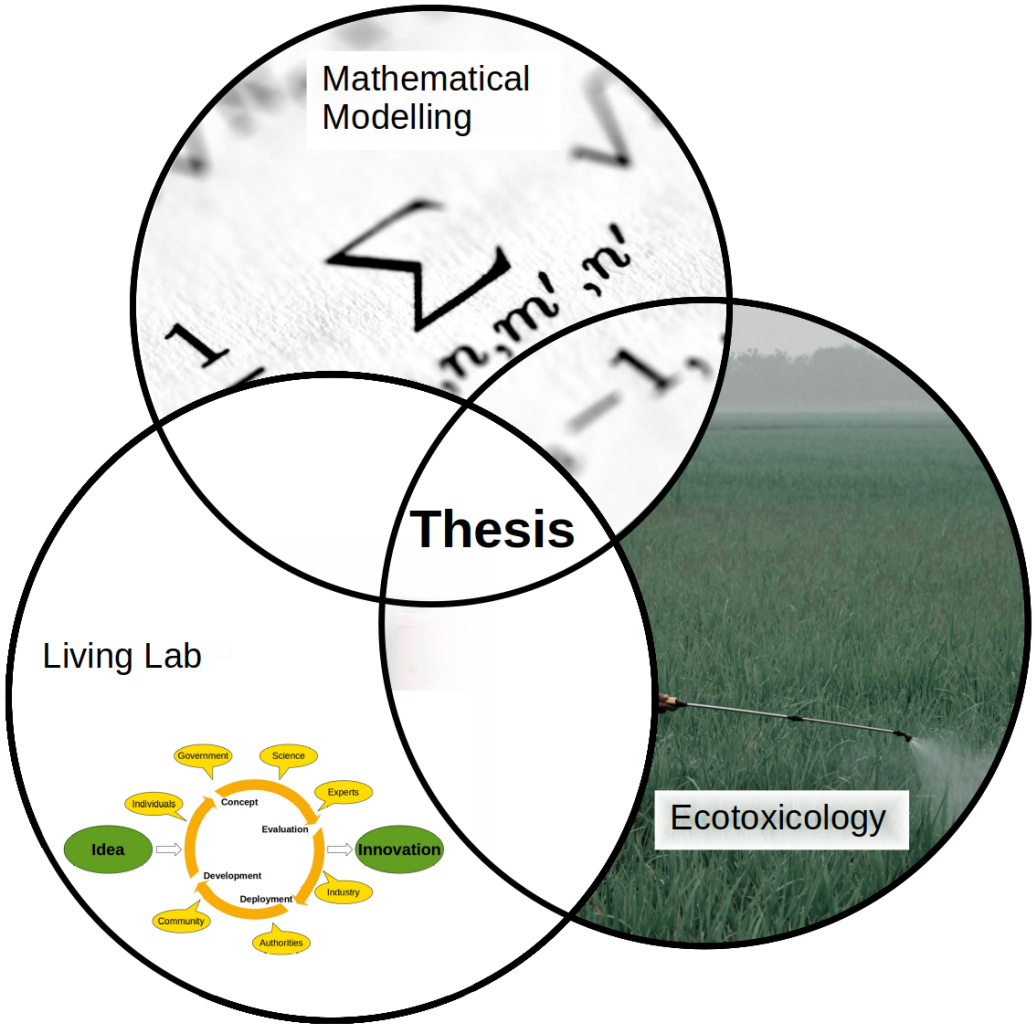

Figure 1.1: Overview about the scientific main disciplines involved in this thesis: the content of the thesis lies in the intersection between ecotoxicology, mathematical modeling, and the LL approach (figure generated with *LibreOffice Draw*, image source: Bozhin Karaivanov and Ilham Wicaksono via Unsplash `https://unsplash.com/`, Herselman unpublished).

strategies used in highly developed countries are not applicable in most of the affected areas because the areas primarily affected by CKDu are poor rural communities in less-developed countries [DE11]. Therefore, existing risk mitigation strategies must be adapted and tailored to the regional, social and cultural constrains of people affected by CKDu in less-developed countries.

Because it pertains to a research and development environment, a user-centered innovation process called LL is used. In this thesis, the idea of a LL goes beyond the usage as a research and development environment [BKS09]. Risk mitigation strategies are developed further, and their effectivity and usability are tested and validated in a user-centered research and develop-

ment process. Additionally, the LL can be regarded as a sensor, whereby relevant parameters related to CKD can be sampled from the population involved in the LL.

In this thesis, the focus is on one suspected risk factor, pesticides, for CKDu [JPA+15, GSS+17, OHA+11]. Besides a possible connection with CKDu, pesticides can have negative effects on human health [HCB12] and environmental health [MIS+16]. Reducing the use of pesticides to only the necessary amount might have therefore positive effects on human and ecosystem health and a positive effect on the financial situation of farmers using pesticides, i.e., reducing the amount of pesticide used also reduces the amount of money spent for agricultural inputs. Effects on human and ecosystem health are inter alia long-term effects and the results of a lower exposure to pesticides might lie in positive effects that could be effectuated in the remote future that are not so obvious in the near future. However, the positive effect of a reduced pesticide amount on the financial situation of farmers is immediately economically visible. This short-term benefit might help to motivate people to contribute in the LL approach mentioned. According to [RPH+12], personal benefits are one of the short-term motivation factors to contribute in citizen-science projects.

Regarding the substance group of pesticides and negative health effects, pesticide-related health problems are more usual in less-developed countries. In less-developed countries, limited economic, medical, and educational resources lead to a high number of residents suffering from pesticide-related health problems that are not seen in industrialized agriculture [STM+12]. Because of the situation in less-developed countries, e.g., financial limitations, missing resources [Noh98], highly toxic pesticides, [Eco01] and high temperatures [Sam24], countermeasures as known as in developed countries are not applicable and must be adapted to the needs of the populations in less-developed countries [SGKJ21] or new measures must be invented.

Adoption of pesticide-related risk mitigation strategies to the needs of a rural community in a less-developed country is a multidisciplinary approach with elements from ecotoxicology, agriculture, resource management, and medical and material science. Therefore, different stakeholders from these disciplines must work together in the LL. As most stakeholders do not come from such rural communities and since community members have the best knowledge about their own situation and cultural and religious behavior, they must be included in the innovation process [VDWBZVV09]. As cultural behavior and the availability of necessary resources have a temporal and spatial dimensions and since the fitness of risk mitigation strategies is inter alia dependent on personal parameters like grade of literacy and education,

skills, and personal history, not every risk mitigation strategy fits for every single person or group.

In rural communities in less-developed countries, there is also the problem that people have only limited access to information [LC17], e.g., regarding the availability of resources or risk mitigation strategies in their area or the location of medical resources. If it is not available in local institutions, such information can be generated or sampled, for example, in a collaborative mapping approach [Bla15]. To gain the maximum benefit from a system that mitigates the risk caused by agrochemicals, people can be assisted in their decisions, e.g., through a SDSS [Kee03], considering the mentioned factors as well as personal data. To increase the benefits of and adapt it to similar problems, such a system must be able to learn from the feedback of the users or related sensors, e.g., if the risk mitigation strategy was successful. Adaptivity is necessary because not every risk mitigation strategy fits to every location, as there might be different characteristics determining the fitness of risk mitigation strategies.

1.2 Aim of the thesis and research hypotheses

The overall aim of this thesis is to develop, describe, and evaluate a structure with which it is possible to give personalized adapted decision support to a person in the field of agrochemical-related risk mitigation in a less-developed country. The system should be adaptive with the aim that it can be used in different areas of the world with their own characteristics.

As the topic deals with less-developed countries and pesticide-related risks, the characteristics of less-developed countries and agriculture in CKDu affected countries are elaborated.

SDSS and input data have spatial and temporal dimensions, for example, exposure maps, and quality of risk mitigation strategies or the generated spatial decision support. As the quality of the generated spatial decision support is dependent on the quality of the input parameters, one of the aims of this thesis is to develop methods to rate the spatial and temporal quality of data.

The decision support should be generated through different environmental and personal input parameters and parameters sampled by citizen observers in an LL approach. These parameters can be processed in a software system with a mathematical structure to gain output parameters with which it is possible to give decision support to people living in a CKDu affected community. These mathematical methods must be adapted to the given structure.

In a mathematical sense, a system is required that transforms a vector with n input parame-

ters into a decision support vector with m output parameters from the input space $K \subseteq \mathbb{R}^n$ to the output space $L \subseteq \mathbb{R}^m$:

$$K \to L \tag{1.1}$$

In this thesis, the mathematical methods necessary for an SDSS in an open community approach should be evaluated according to their use for the described approach.
A concrete application for risk mitigation in the field of CKD using the LL approach will be described and developed in this thesis. The LL serves as an environment for the data sampling, development, and local evaluation of the described system and to find and evaluate new risk mitigation strategies, which are accepted and used by users in the long term, in a user-centric innovation process.
The temporal and spatial distribution of risk can be visualized in risk maps. In this thesis, methods will be developed with which it is possible to generate spatial and temporal data related to risk and risk maps.
Additionally, an example for a base repository of low-cost risk mitigation strategies from the disciplines mathematical modeling and ecotoxicology will be developed.
Summarizing, the overall aim of this thesis is to combine ecotoxicological risks and risk mitigation strategies with mathematical modeling and an LL approach to create and analyze a system giving spatial decision support that is adapted to local economic, social and cultural requirements and constraints in a less-developed country in order to improve the pesticide-related health situation.
Within this thesis, the following research hypotheses are examined:

H_1: A user-centered research and development environment can be adapted with the aim of developing risk mitigation strategies related to agrochemicals or CKDu in less-developed countries.

H_2: It is possible to generate an initial base repository of risk mitigation strategies for a user-centered research and development environment.

H_3: Mathematical methods can be used to develop an SDSS to find the best-fitting risk mitigation strategies tailored to personal characteristics and available resources with fuzzy data.

H_4: Related Information Technology (IT) systems and risk mitigation strategies can be developed with open-source and low-cost methods.

1.3 Structure of the thesis

The thesis is structured in the following parts: chapter 2 consists of an introduction to CKDu, pesticide use and exposure, as well as agriculture and healthcare system in less-developed countries. In the following chapter 3, the term risk is discussed. The theoretical and mathematical background of model building and mappings are highlighted in chapter 4. The following two chapters deal with the applied research and development environment LL. First, in chapter 5, a requirement and constraints analysis for the research and development environment is performed. Derived from this requirement and the constraints analysis, the concept LL is theoretically introduced and adapted to the needs of a research and development environment in the described context of CKDu. Chapter 6 deals with data sampling in a citizen-science approach, as used in this context, and how an LL can be used as a sensor for the data acquisition of CKDu relevant parameters.
The methods for the data processing and quality estimators of temporal and spatial data are highlighted in chapter 7. In part III, first, the characteristics of risk mitigation strategies usable in the described LL approach in less-developed countries are worked out. Second, an initial base repository consisting of possible risk mitigation strategies usable in the research and development cycle in the LL approach is developed. Requirements for an SDSS that can be used in the described approach are defined in chapter 15.In chapter 16, the mathematical tools necessary for the modeling and SDSS approach, such as fuzzy sets, fuzzy logic, ANN and optimization techniques, are introduced and described, including how these mathematical tools can be used for generating SDSS. In chapter 18, mathematical approaches are described how an SDSS can be developed by combining fuzzy logic and ANN and how a system must be constructed with mathematical tools in order to find a possible structure behind a disease with unknown etiology. Chapter 19 contains a summary about the results of this thesis and a discussion about the limitations of the described approach as well as an outlook on possible related research.

1.4 Related projects

This thesis was part of the scientific project LLinES.

The project LLinES, funded by the *German Federal Ministry of Education and Research*, started in 2014 as a collaboration of different scientific research institutes and other authorities with the aim of building an infrastructure for a CKD-related LL in El Salvador. Originally, the project was initiated by stakeholders of different disciplines and from different agencies and authorities like members of the *University of Koblenz-Landau*, the *University of El Salvador*, *Research-Group Learning and Neurosciences* (ReGLaN) and the *Council for Scientific and Industrial Research* (CSIR) *Meraka Institute* but also by stakeholders involved in politics, such as the embassy of El Salvador in Germany, and from authorities like the *National Health Institute of El Salvador* and other institutions, such as the *Action Team 6 Follow up Initiative* (AT6 FUI). The project ended in 2018 because gang violence was so high in El Salvador that project related work could not be performed in the pilot region Bajo Lempa [iES20]. Within the project, some field trips were made to the pilot region in El Salvador; as a result, observations made during such field trips found their way into this thesis.
On the other side, it was intended that results from this thesis find their way into the planned LL in El Salvador. Due to the unplanned end of the project LLinES and the delayed completion of this thesis, this was not possible.

2 CKDu, pesticides, risk, and the situation in less-developed countries

This thesis is inspired by a concrete case of a disease with unknown etiology in a less-developed country, called CKDu. In the thesis, an approach is developed whereby mathematical modeling is used in an user-centered innovation process called LL with the aim of developing methods that help discover the etiology of the disease and to give decision support in terms of mitigating the risk caused by the disease. However, the approach should be not only for the concrete case of CKDu but should also work for other diseases with a structural similarity.
In the following chapter, first the concrete case of CKDu is described. As CKDu is related with less-developed countries, agriculture, and healthcare, and to understand the general structure behind the disease and to bring it to a broader base, the theoretical background in terms of the general situation in less-developed countries, agriculture, and healthcare are analyzed.
The focus of this thesis lies on one of the suspected drivers for CKDu: pesticides. Therefore, an introduction to the term pesticides, the fate of pesticides, and the effects of pesticides on humans and ecosystems are also part of this chapter.

2.1 CKDu and less-developed countries

2.1.1 CKDu, a disease with unknown etiology

In industrialized nations, a steady increase in the number of people suffering from CKD can be observed, with an annual growth of 5-8% [ENB05]. It is estimated that by 2030 about 70% of patients with CKD are located in less-developed countries [AF04, Bar06].
In general, a patient has CKD if they have "[k]idney damage for three or more months, as defined by structural or functional abnormalities of the kidney, with or without decreased GFR, manifested by pathological abnormalities or markers of kidney damage, including abnormalities in the composition of the blood or urine or abnormalities in imaging tests" [JLC+04, p.

871] .

CKD is not a disease itself; it consists of symptoms caused by different ailments [Fou02]. There are traditional risk factors associated with the occurrence of CKD, like hypertension, diabetes mellitus, unhealthy lifestyle, [Bar02, LC12], old age, cardiovascular disease, and obesity [LC12], etc. Also, genetic factors that make a person susceptible to kidney disease are recognized as potential risk factors [LC12]. In areas like El Salvador, Nicaragua, Mexico, Costa Rica, Belize, Sri Lanka, India, Pakistan, Tunisia and Egypt – countries with a high prevalence of CKD – these traditional risk factors can be excluded as a responsible factor for the high case numbers [OHA$^+$11, (PA17]. In these regions, CKD is called agricultural nephropathy, CKD of nontraditional causes, or CKD of uncertain etiology (CKDu).

The case numbers of CKDu in Central America and Asia have been rapidly increasing for about three decades; researchers call it an epidemic. Those affected by this epidemic are mostly poor, rural communities in less-developed countries [Jay14]. In El Salvador, one of the countries with the highest case numbers of CKDu in relation to the population, CKD is a huge health problem. Nationally, it is the leading cause of death in the adult population treated in hospitals, the second leading cause of death in the male population, and the fifth leading cause of death among all adults [dNeHAdES10]. CKDu is mostly observed in middle-aged farm workers; however, children and women not working in the agricultural sector are also affected [ONHVAL$^+$17].

Nowadays, as CKDu is regarded as a real problem in the mentioned areas, several institutions and scientific working groups are working on the field of CKDu [(PA13, ONHVAL$^+$16, yAS18, oSLOw18].

Suspected related factors that favor the occurrence of CKDu in these areas include exposure to heavy metals, pesticides, high temperature, hard working conditions, and nephrotoxic substances [SLW$^+$10]. In a review study, [VLdSW17] found no strong epidemiological evidence that pesticides in general are the single cause of CKDu. However, pesticides consist of hundreds of toxins, some of which are nephrotoxic. Other studies show relations between pesticide use and drinking water from a well with water contaminated with the pesticide glyphosate and CKDu in Sri Lanka [JPA$^+$15]. [GSS$^+$17] showed a relation between organochlorine pesticide residues in human samples and CKDu. However, a causality could not be found.

Investigations in the El Salvadorian Bajo Lempa region show significant correlations between the presence of CKD and work in the agricultural sector as well as exposure to pesticides [OHA$^+$11]. [VLOR14] have shown that high temperature, a factor associated with agriculture

in the mentioned countries affected by CKDu, cannot be linked exclusively for the occurrence of CKDu. They found indications that people suffering from CKDu in El Salvador live in agricultural areas with high pesticide use, high agricultural working intensity, and high temperatures, e.g., areas with a high cultivation intensity of sugarcane. As pesticides and heavy metals are some of the suspected risk factors and are often mentioned in scientific literature about CKD, research in El Salvador is focused mostly on these substances.

In general, in El Salvador pesticides are in use that are listed in the Rotterdam Convention and not allowed to be used in the *European Union* (EU), for example [MQL$^+$14]. A study conducted in a village with a high CKD prevalence analyzed causes for pesticide exposure. People involved in the farming process have a high exposure to pesticides through the misuse and mishandling of pesticides caused by a lack of knowledge about correct usage. Causes for the missing knowledge and misuse are not mentioned in the study. However, not only farm workers are exposed to pesticides. In this region pesticides are applied in high quantities by airplane; therefore, the pesticides also reach non-agricultural areas. Another source of pesticide exposure for non-farm workers is through stockpiles of old pesticides. Drinking water in the investigated community is also contaminated with heavy metals. Farm workers have a particularly high exposure to pesticides because of the lack of protective clothing use when applying pesticides [QRM$^+$17].

Several studies have shown that in general, in El Salvador drinking and surface water have high concentrations of heavy metals, like aluminum and nickel, for which the source was both natural and anthropogenic caused, inter alia, through metal mining, discharges from industry, or illegal waste dumps [ACR$^+$17, GZ14, RACR19]. Similar relationships have been observed in other countries such as Nicaragua and Sri Lanka [SCA$^+$10]. In all these countries, the cultivation and export of agricultural goods is one of the main economic factors. However, this intensive agriculture requires large amounts of pesticides and fertilizers [Eco01]. In some of the mentioned countries, pesticides with a high toxicity for humans that are restricted in some way in industrialized countries are also used because they are still allowed or their prohibition is not adequately controlled in some of these countries [WCB05, KFP$^+$09, HM11]. Table 2.1 gives a summary of the mentioned potential risk factors.

Nowadays, multideterminant models for the etiology of CKDu are developed; however, a concrete causality has not been observed until now. [(PA17] developed a multideterminant model with potential causes for CKDu. In general, besides traditional factors, they attribute the occurrence of CKDu to hard agricultural working conditions, dehydration, exposure to pesti-

cides and heavy metals, and behavioral factors like smoking, alcohol, or ingestion of NSAID. They determine it to a lack of hygiene practices and healthcare, unsustainable agricultural working practice and a low socioeconomic status of the affected people. Within this multideterminant model, it is assumed that CKDu occurs if a person has two or more of the mentioned risk factors. However, concrete evidence does not exist [(PA17]. [JGS14] propose another multifactorial model. CKDu occurs if three risk factors come together at the same time: hard water, exposure to the herbicide glyphosate and nephrotoxic metals.

Table 2.1: Potential risk factors for CKD and uptake routes

Potential risk factor	**Source**
Agrochemicals [SLW+10]; [OHA+11]	Uptake during preparation of application process [FW05]
	Uptake during application process itself [FW05]
	Uptake after the application process [FW05]
	Uptake because of re-entry of sprayed fields [FW05]
	Drift from contaminated clothes [FW05]
	Uptake with drinking water [OHL+15] - wrong disposal of empty containers [MQL+14] - waste deposits of old agrochemicals [OHL+15] - wrong application [MPW+99] Inhalation due to: - spray drift [FW05] - storage of equipment and pesticides in the house [MQL+14] - use of (nearly) empty containers in the house [GLR99]
Dehydration [RJLJ+15]	Low water uptake [CRWJ14] High temperatures [CRWJ14] Duration of work [CRWJ14]
Alcohol [OTW+10]	Personal uptake
Smoking [OTW+10]	Personal uptake
NSAID [OTW+10]	Personal uptake: Pain and inflammation
Heavy Metals [SLW+10]	Uptake with drinking water [ACR+17]
Hard water [JGS14]	Uptake with drinking water

Other risk factors for CKD not listed in table 2.1 are missing or limited access to healthcare and missing treatment of CKD [(PA17].

The progress of CKD can be categorized into 5 stages, according to the damage and the function of the kidney, measured with the Glomerual Filtration Rate (GFR), a parameter describing the level of kidney function. With a higher stage of CKD, the kidney function decreases, which can lead to death by renal failure. Renal failure is observed as the highest stage of CKD. Table 2.2 gives an overview of the different stages of CKD. Patients in the early stages of CKD often do not have symptoms, and the disease is often only recognized in routine examinations. The early stages of CKD are reversible; that means that the disease can disappear. Higher stages of CKD can lead to death within a few months. Severe symptoms can only be treated with kidney transplantation or with dialysis. CKD in a stage in which transplantation or dialysis is needed is also called end-stage renal disease [LC12].

There are biomarkers and routine laboratory tests available to detect CKD [OMAB$^+$14]. Abnormalities in kidney function can be detected by a decreased GFR or by other biomarkers, like the presence of proteins in the urine [Gro09]. In practice, the urine of a person is screened to the protein albumin, the presence of which in the urine is a marker for an abnormal kidney function. The screening for albumin can be easily conducted with test strips that are available in pharmacies [OHA$^+$11]. The GFR of a person is not measured directly; it can be calculated [Gro09]. Different formulas are available to calculate the GFR; the most used is the formula developed by the Modification of Diet in Renal Disease Study Group [LBL$^+$99], called MDRD formula, and the CKD epidemiology collaboration creatinine equation [LS10], further called the CKD-EPI equation. Their commonalities are that personal attributes, such as age, sex, skin color, weight, and height, are implemented into the formulas in addition to laboratory values [LBL$^+$99, LS10]. To use the MDRD formula, values for the creatinine, albumin, protein, and urea content in the blood serum as well as the content of urea in the urine are needed [LBL$^+$99]; for the CKD-EPI equation, only the creatinine content in the blood serum is required [LS10]. The creatinine content in the blood serum can be measured with laboratory equipment, such as a photometer [Gro09]. In [OHA$^+$11], a photometer was used to analyze the creatinine content in the blood serum, with the laboratory equipment located in a rural healthcare facility and operated by laboratory staff.

The results of tests for the stage of CKD can be used to confirm the results of the later developed theoretical approach for modeling risk and the success of the proposed risk mitigation strategies. In practice, CKD is diagnosed if in two tests albumin is present in the urine,

expressed as a Albumin Creatinine Ratio (ACR) $\geq 30\ \frac{mg}{g}$, or if GFR $< 60\ \frac{mL}{min}$ per 1.73 m^2 [(PA17].

In relation to the CKD stage, therapy includes the administration of medications, dialysis,

Table 2.2: Stages of CKD, classification and GFR according to [Amm20]

Stage	**GFR in $\frac{mL}{min}$ per 1.73 m^2**	**Classification**
1	≥ 90	Normal or high
2	60 - 89	Slightly decreased
3 A	45 - 59	Mild to moderately decreased
3 B	30 - 44	Moderately to severely decreased
4	15 - 29	Severely decreased
5	< 15	Kidney failure

or kidney transplantation [CRMT11]. As kidney damage is in most cases irreversible, therapy has the aim of mitigating kidney damage [LRB+13]. However, due to the lack of financial resources, therapy is often not affordable or specialized healthcare facilities not available for people living in less-developed countries, leading to a higher mortality of CKD patients in less-developed countries [CRMT11].

Up to now, hemodialysis is the most appropriate therapy for higher stage CKD patients. However, specialized physicians (nephrologists) are needed, and hemodialysis is conducted in specialized healthcare facilities that are often only located in metropolitan areas. A dialysis method, for which a specialized healthcare facility is not necessary, and which can be conducted as an outpatient procedure, is called Continuous Ambulatory Peritoneal Dialysis (CAPE). It can be conducted by a trained, but not specialized, physician; the patients do not must travel to a specialized healthcare facility, and it can also be conducted in a local healthcare facility [NWG+17]. However, medical equipment, like catheters and cleaning fluid are necessary to conduct the CAPE but sometimes not affordable by affected people [LC01]. As the etiology of CKDu is not clarified at the moment, surveillance of CKDu cases, inter alia, is regarded as an important instrument [KS17, (PA17]. [(PA17] have developed a framework for passive, active, and population-based surveillance. In the proposed framework of [(PA17], passive surveillance means that national databases of reported deaths caused by CKDu or registries of transplantation or dialysis are used and statistically analyzed with other variables,

like location of residence, age, sex and socioeconomic conditions to find patterns. Active surveillance means that surveillance is undertaken in well-defined communities, e.g., through a questionnaire, taking into account health parameters, like biomarkers for CKD, as well as parameters describing occupational and socioeconomic factors. In population-based surveys, the whole population or occupational groups are surveyed. In the surveys, questionnaires must be answered with personal information and information about the presence of potential risk factors; additionally, physical examinations can be conducted. The aim is to find risk factors for the disease [(PA17]. There are several approaches available for population-based surveillance in order to identify risk factors, e.g., the STEPwise approach to surveillance (STEPS) of the *World Health Organization* (WHO) [Org18] or the Chronic Kidney Disease Surveillance Project of the National Health and Nutrition Examination Survey of the United States [fDCP18].

In this section, the case of CKDu was described. Derived from this case, this work intends to develop a general framework for the development and application of risk mitigation strategies in this thesis. Therefore, it is necessary to describe the case of CKDu on a more abstract level.

CKDu can be described as a noncommunicable disease with unknown etiology that mostly occurs in poor rural areas with high agricultural activity and limited access to healthcare facilities. One of the suspected main factors is agrochemicals. CKDu results in a high number of death and affects many people. Geographically, it is not restricted to a specific area or continent. The etiology of the disease is unknown. Potential risk factors were identified; however, their interrelations and mechanisms and the assessment of which of them is responsible for the occurrence of CKD has not been clarified until now.

The progress of the disease can be measured with routine laboratory tests. Research about the disease is mostly done in basic research. The disease has a temporal and spatial dimension. Different types of people are affected, not only farm workers, but also the general population. A heterogeneous group of affected people distributed all over the world can be recognized.

2.1.2 Introduction to the term less-developed countries

In the last section, the case of CKD was described. As most cases of CKDu are located in less-developed countries, and as CKDu in El Salvador is used a case study, some basic background information to these terms are given in the following chapter.

This thesis deals with the case of CKD in El Salvador and related risk mitigation strategies. The approach used in this thesis should not only be for the described case but also for a more general framework. As CKDu is a problem of less-developed countries, the characteristics of these countries are described in the following chapter.

Overall, there is no unique definition of what a less-developed country is. Institutions like the *World Bank*, *Organisation for Economic Co-operation and Development* (OECD) or the *United Nations* (UN) use indicators to categorize countries into less-developed countries and developed or industrialized countries with different subgroups. The categorization into developed and less-developed countries is used to decide which countries can receive development assistance by international organizations such as the OECD. Recipient countries of official development assistance are all low- and middle-income countries as categorized by the *World Bank* and all least developed countries as listed by the UN [fDP18]. The *World Bank* uses as an indicator only the Gross National Income (GNI) per capita, with an upper threshold value for the GNI per capita of $12235 in 2016; the UN also incorporates indicators for human assets and economic vulnerability [fECoO19].

Less-developed countries are in general characterized by economic underdevelopment in contrast to industrially developed countries. The economies of most less-developed countries are characterized by a structure in which traditional modes of production, predominantly in agriculture, are confronted with a modern, dynamic economic sector – mainly in the industrial sector. The economy is characterized by a lack of capital and external difficulties due to the fact that many less-developed countries are heavily in debt [fECG19].

Economically, they can be characterized by a low per capita income, low savings rate and investment activity, low productivity of labor, low technical education, dominance of the primary economic sector, and a lack of material infrastructure. This results in a low level of healthcare with low life expectancy, high child mortality, poor nutrition, and food insecurity, and a low educational level with high illiteracy rates [Noh98]. Additionally, high unemployment, an overall low standard of living, and often extremely uneven distribution of existing goods characterize these countries [fECG19].

In this section the common characteristics of less-developed countries were analyzed. This knowledge is necessary to find appropriate risk mitigation strategies, applicable also in less-developed countries. From these findings, requirements for risk mitigation strategies can be derived in general for less-developed countries. However, the mentioned characteristics can differ between different less-developed countries. To improve the effectiveness of the risk mit-

igation strategies, they must be adapted to the countries' characteristics.

This thesis was inspired by the case of CKD in El Salvador and is embedded in a project in this country. Therefore, in the following section, the situation related to CKDu in El Salvador is highlighted.

2.1.3 El Salvador and the pilot region Bajo Lempa

El Salvador is a middle-income country located in Central America between 13 and 15 north latitude and 87 and 91 west longitude and is bordered by Honduras and Guatemala. It has a coastline with the Pacific. The capital of El Salvador is San Salvador [Geo02]. It has an area of about 21,000 km^2 [Ban19g], a population of about 6.421 million inhabitants [Nat19b], and a population density of about 309.9 people per km^2 [Nat19a] (the last two values for the year 2018). It is the smallest country with the highest population density in Central America.

The landscape of El Salvador is characterized by four landscape types: There are the central highlands, consisting of about 20 volcanoes, some of which are active. Between the volcanoes, there are basins with fertile soil with volcanic origin used for agriculture. North of the central highlands borders an interior plain, and at the Pacific coast there is a coastal plain. In the north of the interior plain along the border with Honduras are the highlands. Only about 20% of the soil is usable for agriculture. This soil is mostly in regions with volcanic activity or in areas where volcanic ash and sediment are transported by rivers. Regions with fertile soil are therefore in the interior and the coastal plain, and fertile soil in regions with high elevation is in danger of erosion [Bri19].

The longest river of El Salvador is the Lempa River, originating in Guatemala and flowing through the central highlands and plains and flowing into the Pacific Ocean in the Bajo Lempa region after having crossed the coastal plain. A second main river is the Rio Grande, draining the eastern part of El Salvador [Bri19].

The environment in El Salvador is mainly affected by deforestation caused by agricultural activities and a high population density. However, industrial, agricultural, and urban pollution of the environment are also severe problems [QRM+17]. The literacy rate in the adult population in El Salvador is about 88%, which is around the world average (86%) and the average rate for Central American countries (86%) but higher than the average of low-income countries (61%) [Ban19d, Ban19f]. The school system is divided into preschool, primary, and secondary education, and primary school is compulsory and free [Bri19]. In the rural areas, people have received on average about 2.2 years of school education, and a large part have no

school education [FotUN19]. The climate of El Salvador can be characterized as tropical with a dry season in the summer and a wet season with heavy rainfalls in the winter. However, due to the elevation differences within the country, differences in terms of temperature and precipitation can be observed. Precipitation in the coastal plain is lower than in the mountainous regions, and temperatures in the coastal plain are higher than in the mountainous region [Bri19]. The annual mean temperature is about $24.8°C$ with an annual mean precipitation of about 1824 mm [Pro19a].

About 60% of the population lives in urban areas. Spanish is the official language in El Salvador [Bri19].

The country has only a low growth Gross Domestic Product (GDP), with an estimated GDP growth of about 2.6% in 2019, and about 20% of the GDP comes from emigrated workers. The poverty rate is about 31% [Ban19a]. Poverty in the rural areas is characterized by a lack of basic services like electricity, water access, and sanitation [FotUN19].

The main agricultural crops are cotton, sugarcane, maize, and coffee. Industry is mainly based on service [Ban19a]. Of farmers, 82% are small-scale farmers with less than 3 ha land. These small-scale farmers mostly produce basic grains [Ban19a]. Of all employees, 19% are employed in agriculture [Ban16]. In 2019, El Salvador had a Human Development Index (HDI) value of 0.674 and is categorized as a country with medium human development [Pro19b]. The GDP per capita was about US$4058.20 in the year 2018, lower than the world average (US$11298.30) and lower than the average for Central America (US$5449.20) GDP per capita [Ban19b]. The main problems in the country are natural disasters, like volcanoes and tropical storms, and crime conducted by gangs, also called maras. El Salvador has one of the highest homicide rates in the world [Bri19].

Child mortality in El Salvador is about 14 per 1000 birth, lower than the world average (39 deaths per 1000 birth) but much higher than the OECD member state average (6) [Ban19e]. Life expectancy at birth is about 72 years (world average 72 years, OECD members 80) [Ban19c].

The project LLinES was planned in the pilot region of Bajo Lempa. The Bajo Lempa region is a rural region characterized mainly by agriculture, about 150 km southeast from the capital San Salvador on the Pacific coast at lower reaches of the Rio Lempa. The location of the Bajo Lempa region within El Salvador is visualized in figure 2.1.

The region is frequently affected by flooding and has a high CKDu prevalence. The main agricultural products are sugar cane and maize. Bajo Lempa communities were resettled

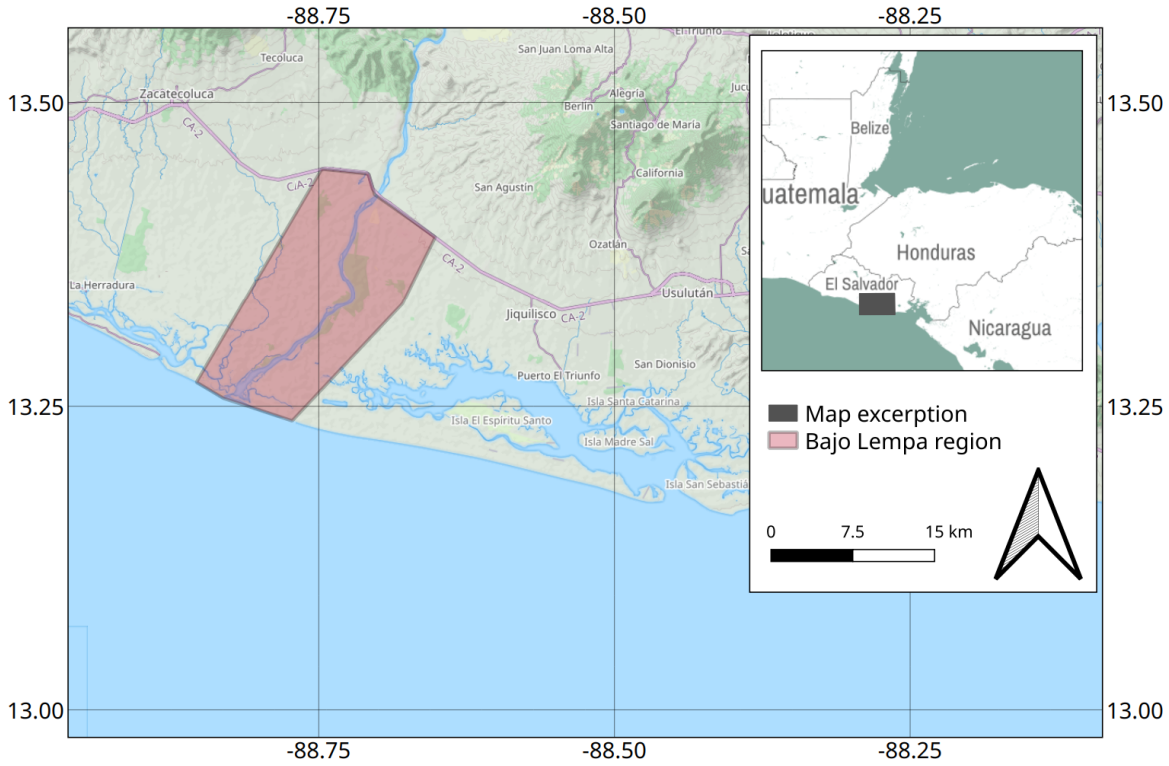

Figure 2.1: Map of the pilot region Bajo Lempa in El Salvador (figure generated with *QGIS*).

after the end of the civil war in 1992 – mainly by refugees from other parts of the country. Society in the Bajo Lempa region is characterized by a high degree of organization with agricultural cooperatives, like *ACUDESBAL (Asociacion Intercomunal de Comunidades Unidas para el desarollo economico y social del Bajo Lempa)*, or grassroots communities with the aim of developing social, environmental and participatory progress, like the *Asociacion Mangle* [Man24]. However, the region is also characterized by high poverty and gang violence. Basic healthcare services like a local medical station are offered in the communities (personal communication). Overall, three field visits were carried out during the LLinES project, and some findings are presented in chapter 2.2.4.

The Bajo Lempa region was chosen as the pilot region because the chance of suffering from CKD is quite high and relations with the Salvadorian project partners already existed; furthermore, the local population and cooperatives wanted interventions against CKD and were willing to cooperate with foreign project partners.

In this section, the situation related to CKDu in El Salvador was discussed. From the results of this chapter, options for risk mitigation strategies can be derived with a higher

granularity and tailored to the situation in selected region than by using the general characterization of less-developed countries given in section 2.1.2. However, if no information about the situation in specific region or country are available, risk mitigation strategies can also be developed for the general characteristics of less-developed countries; however, the effectiveness might be lower than if more specific information about a specific area is used in the risk mitigation strategy developing process.

2.1.4 Health situation in less-developed countries and factors influencing access to the healthcare system

An introduction to the healthcare situation in less-developed countries is presented in the following chapter. For example, one possible risk mitigation strategy can be the correct and appropriate treatment of agrochemical-related diseases.

In general, the healthcare situation in less-developed countries is not sufficient enough that people's health can benefit highly from it. In less-developed countries, wealthier people benefit more from the existing healthcare system than poorer people [O'D07]. In less-developed countries, the number of hospital beds, doctors, and nurses per inhabitant is more than five times less than in developed countries [PGB+08].

In the less-developed world, the main problem in the field of healthcare is the accessibility of medical prevention and treatment methods. [Wag04] describes that three diseases, namely malaria, diarrhea, and pneumonia, for which prevention and treatment methods readily exist in more developed countries, are still responsible for over 50% of child deaths worldwide. These deaths could be prevented if access to the mentioned basic prevention and treatment methods were available. Therefore, there is a gap between potential healthcare and healthcare used, and this gap has different reasons.

[O'D07] mentions different reasons for this gap. On the demand or patient side, a lack of education and typical cultural behavior lead to the fact that symptoms are not recognized as being a reason for visiting a healthcare facility, e.g., the disease is not recognized as severe or is not recognized at all. On the other side, the effect and the benefits of medical treatment are not known. The high number of child deaths in Latin American countries are, according to [HKE04], attributed to this fact.

[Sac01] analyzed the difference between the money necessary for countries to offer a base healthcare level and the amount spent in reality. They found a large gap between these two variables and state that this gap shows that required healthcare elements are not offered in

a range necessary for basic healthcare. This indicates that the supply side is not developed enough for a working healthcare system.

[O'D07] reports that existing healthcare facilities in less-developed countries often are low quality. This low quality is also a reason for a lack of healthcare. According to [O'D07], the lack of effective healthcare in less-developed countries is related to a meager supply of effective high-quality healthcare and that people will not use healthcare services from which they do not get a benefit. On the supply side, in less-developed countries the offered healthcare is often of low quality. This is connected to the demand observations that patients will not use a healthcare facility that they think is of a bad quality and little or no benefit.

In less-developed countries, there is also a gap between rural and urban areas in the field of healthcare. [Str03] attributes this to higher poverty and fewer healthcare resources in rural areas. Additional factors influencing this trend can be seen in the behavior of rural communities. They have their own traditions and believe in a traditional way of treating diseases. On the other side, in many traditional communities there is a way of doing things and finishing them, even, if the people have diseases. This often leads to a worsening disease with no or insufficient treatment.

E-health is a concept that is increasingly used in less-developed countries, especially in rural areas, to decrease the costs and counter geographical barriers [LSLS12]. The WHO defines E-health as "the transfer of health resources and healthcare by electronic means" [Org15b]. According to the WHO, E-Health implements, inter alia:

- delivering health information via telecommunication and the internet for both patients and health professionals and
- the use of IT for the improvement of public healthcare.

Healthcare professionals and patients are connected via telecommunication or the internet, through which the patient can describe or show the symptoms, etc., and the healthcare professional can make a diagnosis and give support to the patient. The second item means that, e.g., health promoters in rural areas are trained via the mentioned channels in order to give the best advice to a patient, even if a professional physician is not available.

A field trip to a local healthcare center in the Bajo Lempa region in El Salvador in 2015 showed that the healthcare system does not work as efficient as it would be necessary to treat diseases caused by pesticides. The visited healthcare center is relatively new and was established in 2011 to treat people suffering from CKD or people having symptoms of a disease caused by

pesticides. Relatively modern equipment to perform dialysis is available. The goal is to teach people suffering from CKD how to do a dialysis for themselves with only a minimum visits to a hospital or healthcare center. The staff of the regional healthcare center have said that they cannot use the equipment because the materials needed to do so or to teach patients about dialysis is missing, such as tubes. In terms of the definition, access to the healthcare system is limited according to the availability of the needed resources.

Communication with staff of the local healthcare center showed another problem occurring especially in less-developed countries. The center has information about death rates, pesticides applied by the dead people, and so on. It would be possible to do statistical analyses on potential risk factors for the prevalence of a disease caused by pesticides, but the money is not available for staff to analyze the collected data. Such a risk factor analysis would be important for finding possible risk mitigation strategies. This fact can also be attributed to the availability in an indirect sense. There are not enough medical researchers available to investigate the causes of the mentioned pesticide-related disease. Therefore, effective countermeasures cannot be undertaken.

One of a possible risk mitigation strategy is to optimize the healthcare situation in the framework of the logistical optimization of medical resources and give spatial decision support to the people, e.g., where the next appropriate physician is in terms of acceptability, accommodation, accessibility, availability, and affordability to where the patient is located.

2.2 Pesticides and CKDu

2.2.1 Pests, pesticides, and alternative pest management strategies

Pesticides are one of the suspected risk factors for CKDu. In this thesis, a multidisciplinary approach for developing risk mitigation strategies is proposed for which stakeholders from different disciplines work together to develop individual fitting risk mitigation strategies. With the mentioned focus on pesticides, a contribution from the perspective of one of the involved disciplines, ecotoxicology, on the multidisciplinary problem solve strategy is performed. In the proposed user centered innovation approach (chapter 5), a problem description and mechanisms for finding risk mitigation strategies must be developed from each of the involved disciplines.

To understand the mechanisms of risk mitigation in the field of ecotoxicology, a theoretical introduction to the use and fate of pesticides is given in the following chapter.

There are many definitions available describing the characteristics of organisms that are called pests. The most have in common that an organism is regarded as a pest if it causes negative effects to humans, their animals and plants, or agricultural products. An organism labeled as a pest is always related to damage to humans or their properties. That leads to a species being considered a pest when it is located in an area used by humans with a given population density, but outside of that area the same species might be regarded as wanted species. Within this definition nearly every organism in interaction with humans can act as a pest. If the damage caused by a pest is higher than a threshold level, countermeasures must be taken. These countermeasures are then called pest management [Hil87].

As humans are in competition with other organisms for plants, intended for human use, humans must manage these organisms from the beginning of agriculture [THK07]. There are several ancient reports available describing the occurrence of organisms endangering plants intended for human use, for example, in the 2nd book of Moses in the Bible, a pest outbreak with locusts is described [Bib52]. Also, other historical sources describe the occurrence of organisms harming agricultural yields and the management of them, e.g., the Sumerians used sulfur as an insecticidal agent to repellent insect pests [Den00, THK07]. The management of organisms harming agricultural yields has followed human agriculture and related science up to today. Pesticides are substances that are used to control organisms, such as plants, insects, nematodes, mollusks, or microorganisms that can decrease the harvest or cause illness or other negative effects to humans and animals.

At the beginning of commercial agriculture, a change in the methods for pest control could be observed. In the area of subsistence agriculture, pest control was mainly achieved through the use of biological and mechanical pest control, the use of resistant varieties, and biological pesticides. At the beginning of the Green Revolution in the middle of the 20th century, increasing yields were achieved through the extensive use of agrochemicals like fertilizers and pesticides. Pest management was mainly performed through the use of pesticides. Due to the large impact on nature and health recognized during the last 60 years of intense pesticide use, a slight shift in pest management has been observed in which a method called Integrated Pest Management (IPM) is also used; therefore, pesticides are no longer the only method for pest management, and alternative pest management strategies are also employed [Mat15]. In the IPM approach, pesticides are regarded as the last option and are only used when an economic threshold caused by pest damage is exceeded. This strategy helps to reduce the use of pesticides with all its negative implications and also to save money in the agricultural production

process [Kog98]. Globally, in 2018 around 1.2% of the agricultural land was organic farmland [WLK18].

During the second half of the 20th century, pesticide induced negative effects on environment and human health, as described in chapters 2.2.6.3 and 2.2.6.4, became obvious. In commercial agriculture, the methods changed, and mainly synthesized pesticides were used for pest control. Synthesized pesticides had the advantage that they are cheap and effective [Den00]. However, pesticides are substances that have positive and negative effects: On the one side, they are tools that are important for productivity in agriculture. On the other side, they are responsible for environmental and human health problems [Fit02].

Despite the high usage of pesticides, pests are responsible for yield reduction. [Pim09] estimate a reduction of about 40% globally, a meta-analysis showed yield reductions between conventional and organic farming between 13% and 34%, dependent on the cultured crop [SRF12]. Pesticides are used in agriculture, the public health arena, and other applications in the household and landscape [fASC14]. In the United States of America (USA), agricultural pesticide use accounted for about 72% in 2007, private use in home and the garden for about 13% and the governmental and industrial use for about 15% [KDG$^+$04]. In regions with vector borne diseases, like malaria, west Nile virus or dengue fever, pesticides are also used to control these diseases by managing the vectors, which are, in these cases, mosquitoes of the genus Aedes [Ros01]. Additionally, highly toxic pesticides are used for vector control, for example, in India, Dichloro-Diphenyl-Trichloroethane (DDT) is allowed to be used for vector control despite its ban in agriculture [vdBZY$^+$12].

The aims of the use of pesticides are to gain the greatest possible harvest or to prevent illnesses and destruction on the crops [Ste04]. For pesticides, there are different classifications available. They can be classified according to the organisms they act against. The major groups of pesticides are insecticides, fungicides, herbicides, nematicides, molluscicides [Cou11], and rodenticides [Mat15]. Another possible classification is the categorization into biological and synthesized pesticides [Ste04]. The amount of pesticide used per area is dependent on the cultivated crop, environmental parameters, and pest pressure.

Pesticides are not sold as the active ingredient itself but as a formulation. A pesticide formulation can consist of at least one active ingredient, the chemical intended to act against a pest, and additional chemicals, called inert ingredients. Depending on the substance, the mechanism of an active ingredient can differ. An active ingredient can repeal, kill, mitigate, or prevent a pest. Other mechanisms of active ingredients are that they defoliate plants, act as

a plant regulator, or act as a desiccant. Inert ingredients are used to improve the effectiveness of the pesticide and to get specific product characteristics with the aim of improving product performance. Some examples of the function of inert ingredients are that they prolong the durability of the product, they help the active ingredient to get into the plant, or they make a substance more toxic to an organism. However, an inert substance can also be toxic to humans and must also be assessed by authorities [(US16]. Pesticides can be formulated as granular, dust, aerosols, liquids, or powder sprays. The type of the formulation has an influence on the exposure to the pesticides and on the product's characteristics, i.e., on the degradation time [FW05].

The following numbers about pesticide use are not current and are based on the years of the beginning 21st century to make them comparable with values calculated with a model described in chapter 11, in which the used input maps and values are also based.

In 2007, pesticides were sold with an estimated value of about US$39,443 million and an amount of 2.364 Mil t; the largest amount was used in the herbicides group (40% market share), followed by other pesticides, including nematicides, fumigants, and sulfur with a 33% market share. Insecticides and fungicides account for about 17% and 10%, respectively [GDKW11]. In the EU in 2014, pesticides about 0.3959 Mil t were sold, with fungicides accounting for about 43.7% , herbicides for 33.1%, insecticides for 5% and other pesticides for 17.8% [otEU16]. In the USA, herbicides account for 47%, insecticides for 8%, fungicides for 6%, and other for 39% . In the USA, the largest proportion of pesticides is used in agriculture (62%), followed by home and garden use (22%)[GDKW11]. Pesticide consumption in the USA and EU account together for about 38.5% of global pesticide consumption [otEU16, GDKW11]. In table 2.3, the applied overall pesticide amount, the agricultural area, and application rates are listed.

The application rates in the USA and Europe are higher than the average application rate in the rest of the world. According to Figure 2.2, an increase of about 70% in global pesticide use in agriculture for the overall pesticide amount can be observed between 2000 and 2024.

Regarding the annual applied pesticide amount in the USA, a decreasing trend between 1995 and 2015 can be observed for the classes of insecticides and multi-use substances; for herbicides an increasing trend is seen, and for fungicides there is also an increasing trend after a phase of reduction [SBP$^+$21].

Table 2.3: Pesticide use and agricultural area, sources: a [GDKW11] for the year 2007, b [otEU16] for the year 2014, c [FD16] for the year 2014, values without a source are calculated from values given in the table

Region	**Pesticide use**		**Agricultural area**		**Average application rate**
	tons per year	%	ha	%	kg per ha
United States	513920 [a]	21.7	99924162 [c]	8.3	5.143
Europe	395944[b]	16.8	170141510 [c]	14.2	2.327
Other regions	1453804	61.5	927926999 [c]	77.5	1.567
World total	**2363668**[a]	**100**	**1197992671** [c]	**100**	**1.973**

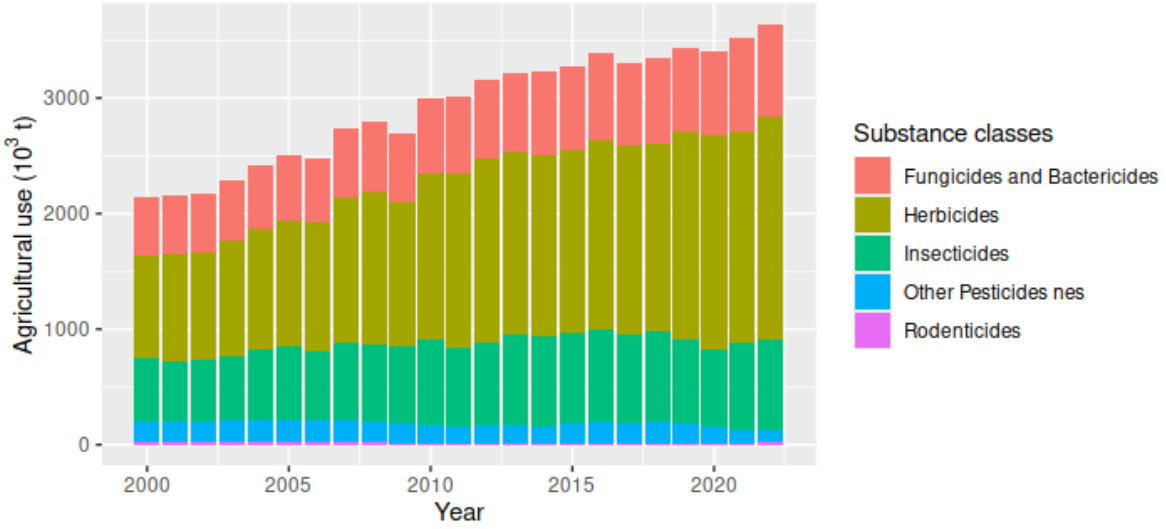

Figure 2.2: Global pesticide use in the agricultural sector [FD24a] (figure generated with *R*).

2.2.2 Pesticides in the environment, fate, and their behavior

Pesticides are substances with characteristics that can lead to negative effects to the environment. They are toxic to non-target organisms, persistent and can biomagnificate [Fen13]. Released into the environment, chemicals are subject of different transfer, transport, and transformation processes, with spatial distribution of the chemicals changing within or between compartments. In a transformation process, the chemical structure of the substance itself changes. The distribution and spread of chemicals in the environment is determined by the physical and chemical properties of the chemical and the ecosystem as well as by transport, transfer, transformation, and biogeochemical processes [Fen13]. According to [Fen13], the concentration of chemicals and distribution patterns can be attributed to the following factors:

- input sources, input quantity and input characteristics;
- physical and chemical properties of the substance (i.e., molecular structure, water and fat solubility, partition coefficients, vapor pressure);
- physical and chemical properties of the ecosystem (i.e., temperature, salinity, pH, suspended solids, sedimentation rate, nutrient cycling); and
- transformation processes (i.e., photolysis, hydrolysis, redox reactions and biotic degradation).

Important transport processes are, for example, advection, diffusion, dispersion, or transport on particles. Solution in water, sorption to particles, sediments and soil (sedimentation), votalization in the atmosphere, and atmospheric deposition are counted among the transfer processes. Transformation processes can be categorized into biotic and abiotic transformation processes. Abiotic transformation processes are for example photolysis, hydrolysis, and redox reactions. Aerobic and anaerobic degradation belong to the biotic transformation processes [Fen13].

An overview about the most important transfer, transport, and transformation processes is visualized in figure 2.3. Released into the environment, pesticides do not necessarily stay at

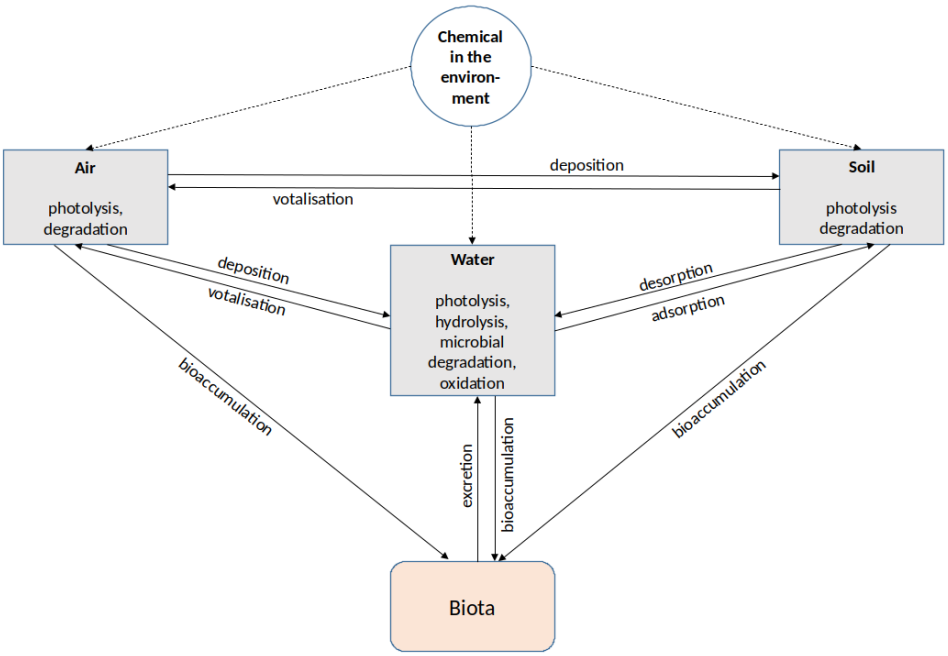

Figure 2.3: Fate of chemicals in the environment according to [Fen13] (figure generated with *LibreOffice Draw*).

the point they were entered. Nowadays, pesticides can be found nearly all over the world, even in regions like Antarctica, where they were never applied [PLN+20]. The reason for this global occurrence is the tendency of chemicals to spread. The tendency of a substance to leave a phase is called fugacity [SM98]. The longevity of a chemical in the environment is called persistence [Gre80].

In section 2.2.5, different uptake routes of pesticides into the body are mentioned, such as oral uptake due to drinking water or food, inhalation, or dermal absorption through the air. Due to the mentioned transfer and transport processes and the fugacity, pesticides do not stay at the location on which they were applied and are influenced by transformation processes. The tendency of substances to stay or to leave a phase can be assessed by equilibrium constants (explained later in this chapter). In a risk mitigation approach, like the one described in this thesis, and to mitigate or avoid exposure to these pesticides, it is necessary to know about potential concentrations in the mentioned uptake routes and compartments. As direct measurements of pesticide concentrations in the different media and uptake routes need special equipment and are expensive and time-consuming, they are not always affordable in a risk mitigation approach in less-developed countries. Therefore, it is necessary to estimate these concentrations with a temporal and spatial dimension through fate models, taking into account the different transfer, transport, and transformation processes as well as agricultural practice, pesticide application patterns, and environmental conditions like soil type and meteorological data related to time and space. In a multidisciplinary risk mitigation approach, this would be the task for stakeholders from the field of ecotoxicology.

The exchange of chemicals between the individual environmental compartments is controlled by the tendency of the chemicals to enter into phase transfer processes. The speed of transfer depends on the physical and chemical properties of the substance and on environmental conditions. These exchange processes can be described by equilibrium constants [PdCdM+16]. By dividing the environment into the main compartments of soil, sediment, air, water and biota, it is possible to describe the affinity of a chemical to the mentioned compartments with the following parameters: Henry's Law constant K_h, the n-octanol-water partition coefficient K_{OW}, the Organic Carbon-Water Partition Coefficient K_{OC}, the octanol-air partition coefficient K_{OA}, the water solubility S [VC93], and the solid-water partition coefficient K_p [Fen13]. There are also other substance dependent parameters like the water solubility

S, the molecular weight m, vapor pressure p, type of formulation, half-life $t_{\frac{1}{2}}$, or the GUS (Groundwater Ubiquity Score) index [PdCdM$^+$16].

Pesticides can also be uptaken by biota, concentrate in fatty tissue, and accumulate along the food chain. The uptake process of pesticides into biota and accumulation of the substance in a living organism is called bioaccumulation; the accumulation along the food chain is called biomagnification. Biomagnification is responsible for high concentrations of pesticides with the tendency to bioaccumulate being found in top predators [Kat10]. The mentioned parameters can be used to estimate the fate of pesticides and resulting environmental concentrations. In the following part, the meaning of the mentioned parameters is only explained as an example for two parameters to demonstrate the relevance and their use in a risk mitigation approach; for the other parameters, further details from [Fen13], [PdCdM$^+$16] can be used.

Water solubility describes the amount of a substance dissolved in water, expressed in $\frac{mg}{L}$ or ppm. Substances with a high water solubility reach deeper zones in the soil and the groundwater and tend to go into aquatic ecosystems, such as rivers and lakes [MSL06]. According to [PdCdM$^+$16], substances with a solubility equal or greater than 30 ppm tend to leach and have the potential to contaminate ground water.

The mentioned partition coefficients are equilibrium constants, which can be measured in standard laboratory tests by releasing a chemical into a two-phase system and measuring the concentration of the chemical in the two phases after establishing an equilibrium [Fen13]. Henry's Law constant K_h of a substance can be used, for example, to determine the tendency of a substance to votalize and therefore to transfer from the solid or liquid phase into a gaseous phase. It can be calculated, for example, by $K_h = \frac{C_g}{C_w}$, with C_g representing the concentration of the chemical in the gaseous phase and C_w the concentration in the aqueous phase at the equilibrium. Substances with a $K_h > 100 \frac{Pa}{molm^3}$ are votalizing fast from the water into the gaseous phase, while only slow amounts of substances votalize with a $K_h < 25$ [SM98].

Besides the mentioned substance related parameters, environmental parameters also play a role in the fate of chemicals:

Wind direction and strength determine and influence the transport of substances and can lead to unwanted transport to unintended areas, such as protected areas or from human-settled areas [MoARA14]. [MoARA14] describes a range of wind speed between 2 and 15 $\frac{km}{h}$ in which it is appropriate to spray pesticides. Conditions with a wind speed lower than 2 $\frac{km}{h}$ are too inconstant, with changing wind directions, and have a too long stay of the droplets within

the air, which is not wished; conditions with a wind speed higher than 15 $\frac{km}{h}$ lead to a too high drift of the substances. At conditions with higher wind speed, farmers should change the application patterns to minimize drift, e.g., by reducing the distance between nozzles and plant or by spraying directly into the canopy [MoARA14].

In addition, temperature and humidity have an influence on the fate and behavior of pesticides. [MoARA14] recommend not to spray when temperature is higher than 25 $\circ C$ and humidity is lower than 40% because of relatively high drift loss under these conditions. The high drift under these conditions is attributed to high evaporation and temperature inversion under these conditions. Evaporation leads to finer droplets of the agrochemicals.

Furthermore, precipitation has an effect on the fate of agrochemicals. Dependent on the substance properties, precipitation can lead to the transport of pesticides into surface or ground water. Precipitation after the application can lead to a wash of the pesticides off the plants. That means to have the same protection level, a higher amount of pesticides must be applied [MoARA14]. Additionally, spray-drift and runoff are regarded as important sources of pesticides in surface waters, with runoff having a higher potential [Sch01]. According to [CMV$^+$11], soil characteristics, such as organic matter content or pH, also have an influence on the fate of pesticides in the soil. Organic matter leads to adsorption of pesticides to soil particles, and the pH of the soil influences the adsorption of ionic chemicals. Agricultural practice, such as the type of application and the amount applied into the environment, and the type of formulation of the applied pesticides have an effect on the fate of pesticides in the environment [Car00].

Landscape also influences the distribution of chemicals in the environment. For example, vegetated field strips around farm land can help to reduce entry of chemicals into neighboring surface water by runoff or transport in the air to non-target areas around the farm land [HED$^+$05, SDH99]. [PdCdM$^+$16] mention temperature and the type of soil also as parameters influencing the fate of pesticides.

According to [HS01], the quantity of pesticide droplets in the air after application are mainly influenced by weather conditions (wind speed, direction, humidity, temperature), pesticide properties (viscosity, vapor pressure), application technique, and attitude, care and skill of the pesticide applicator.

Pesticide concentrations in the air can in general be attributed to the drift processes after application. [Mue15] differentiates between three main drift processes: drift of pesticide droplets evaporated directly after application, drift of vaporized amounts after the application,

and drift of pesticides bound on soil particles caused by wind.

The results of this chapter help to characterize chemicals and identify the fate and possible entry path into different ecosystems and environmental compartments, which are used or consumed by humans.

2.2.3 Human exposure on different levels and uptake routes

Humans are exposed to pesticides in different ways. A distinction is made between occupational and residential exposure. Occupational exposure happens when agricultural workers get in contact with pesticides when they mix or apply them or when they get in contact with products or plants with pesticide residues on them during their work. Residential exposure occurs when people are exposed to pesticides at their home, i.e., through breathing air with pesticide residues or eating or drinking contaminated products [FW05]. Human exposure also happens after pesticides are applied on non-agricultural areas, for example, when they are applied next to roads or on golf courses [AM14].

According to [AM14], important routes of human exposure are, inter alia, through air, water, food and soil and can be divided into oral uptake by ingestion, respiratory uptake by breathing, and dermal uptake through the skin. After taken into the body, pesticides are transported by the blood and can be distributed through the human organism; excretion through air, urine, or the skin can also be observed [DE11].

In the agricultural working process, dermal exposure is very common, especially for pesticide handlers who load, mix, and apply pesticides or clean pesticide spray equipment [FW05]. However, residues of pesticides on everyday equipment can also lead to uptake into the human body [BUH+14]. According to [MCK+13], the uptaken pesticide amount is influenced inter alia by humidity, temperature, type of pesticide formulation, the use of protective clothing, and how the skin is covered as well as by the duration of exposure and the amount of pesticides a person to which is exposed. In general, liquid pesticide formulations are uptaken faster than solid formulations. The uptaken amount is also related to the skin region exposed to the chemical. For example, the genital area and skin on the head have higher absorption rates than the forearms or palms [DLSA10].

Oral exposure to pesticides is mainly attributed to the uptake through food and water as well as through accidental uptake [DE11]. Especially in less-developed countries, groundwater is an important source for drinking water [NMS+05]. Several studies have shown the presence of pesticides in groundwater [Mou07, TGLR14, LAL+11] that also exceed the limits set in

quality standards as proposed in directives for groundwater of the EU [PotEU06].

This behavior was also observed during a field trip in 2015 to the Bajo Lempa region in El Salvador. Empty pesticide containers and fertilizer bags were not disposed of appropriately; we recognized some empty pesticide containers lying next to the field. In figure 2.4, an empty pesticide container formerly containing the substance glyphosate can be seen. Not obvious in figure 2.4 is that a small channel or ditch is running approximately 10 m away from the empty pesticide container. Personal communication with responsible people for the water supply showed that the community is not connected to the national water supply network, and drinking water is gained from surface or ground water that has potential contamination from the empty pesticide containers.

The uptaken amount of pesticides by food can in some cultures be higher than the uptaken amount due to water uptake or air inhalation [JACH07]. Accidental uptake can be often observed if pesticides are stored in unlabeled bottles or in empty bottles with a different label [GHS10], if bottles or containers with residues are used for drinking water, or if pesticide handlers do not follow security measures, like washing their hands after working with pesticides [DTT08]. [TCG$^+$03] reports that missing hand washing is responsible for the transfer of pesticide residues into the domestic environment.

A lot of pesticide formulations are liquids and sprayed in the agricultural working process, the pesticides are deliberately released into the environment. Therefore, respiratory uptake of pesticides plays an important role during pesticide application. Exposure increases with temperature and decreases with droplet size. Conventional spray equipment, as used in industrial agriculture, produces relatively large droplets [Ama14]. In several studies, pesticide residues were found in the urine [KTT$^+$14, CTS$^+$04] and blood samples of pesticide applicators [KSM$^+$10].

In the agricultural workflow, there are different tasks connected with pesticides, and each task has different potential for exposure. To estimate the exposure on a individual basis, the task of a person must be known. [FW05] defines different workflow scenarios: mixing, loading, application, flagging of fields when sprayed with airplanes, and other activities, such as cleaning equipment, etc.[FW05]. Within the group of pesticide handlers, pesticide mixers and loaders have highest risk because of contact with the undiluted substance, and applicators are under lower risk [Mat15]. [MA02] analyzed the risk literacy of farmers in Ethiopia and their knowledge about appropriate methods to avoid pesticide exposure. Only a small proportion of the farmers had knowledge about the importance of personal protective equipment, such as

Figure 2.4: Empty pesticide container disposed on a field in the Bajo Lempa region, El Salvador (source: M. Hieber-Ruiz)

gloves and goggles, partly because the used protective equipment did not fit or was damaged due to age. Additionally, pesticide application manuals were not read or not understood by the farmers. Reasons were, inter alia, illiteracy or low education in reading. Additionally, insufficient hygiene practices when working with pesticides were observed. The importance of medical treatment and medical checks were not obvious for the farmers as methods to mitigate the risk caused by pesticide poisonings [MA02].

In the agricultural pesticide handling process, exposure is relatively high compared with ex-

posure caused by environmental pesticide levels and exposure [DE11].

Yet, not only people working as pesticide handlers are affected by pesticide poisonings. Their relatives and dependents also have exposure to pesticides. One study analyzing pesticide exposure of children living in families with parents working in agriculture has demonstrated that pesticide residues in the household dust of these families are more than seven times higher than in homes of families without an agricultural background. Furthermore, pesticide metabolites in urine were significantly higher in children of this group than in the reference group. A comparison of the metabolite concentration of different pesticides between children living near pesticide treated orchards and children living far away from such orchards showed that metabolite concentrations were higher in the first group [LFSK00].

A meta study analyzing non-occupational pesticide exposure of people living near to agricultural land showed that people living near agricultural fields had higher exposure levels of pesticides than people living in non-agricultural areas. In addition, exposure is higher during periods when pesticides are applied [DFQD20].

Non-occupational exposure near to agricultural areas is mostly attributed to votalization and spray drift [FUL$^+$10a]. [DFH$^+$15] also mention take home pathways like residues in clothes as well as consuming products from treated fields.

Several studies have shown that after or during the application of pesticides on an orchard they can reach areas outside the treated orchard through deposition [MH86] [MCMT$^+$91] [FRB$^+$93]. When uptaken in the body pesticides can be detected via biomarkers. A biomarker is a measurement to detect interactions between a biological system and a environmental agent. Besides pesticides and other chemicals, such an agent might also be of physical or biological nature, such as light or a predator. With biomarkers, it is possible to estimate the exposure of an organism to this agent [Int93]. The metabolized substances and pesticides themselves can be detected in different organic materials with biological methods or analytical techniques: organic fluids, such as blood, urine, and breast milk, or in tissue samples, such as biological samples, fat, or serum [Anw97].

Biomarkers can, for example, be used in the proposed LL approach to evaluate if the developed method to estimate the exposure of humans to pesticides is in common with reality.

In general, [FRB11] mentions the following factors determining the prevalence of agrochemical related diseases: the number and frequency of applications, the type of the used formulation, and the toxicity of the used substance.

In literature, different methods are listed as mechanisms to reduce exposure. A study con-

ducted in South India showed the success of educational programs related to pesticide handlers. After participating in an educational program, knowledge, attitude, and practice were improved, resulting in a decreased number of pesticide poisoning cases [SAP+08]. The use of protective clothing as well as hygiene measures increases within farmers in Bolivia after they participate in a training course about the proper use of pesticides [JLH+14]. According to [KKC+13], the risk for pesticide poisonings of people working in agriculture increases significantly if people do not wear protective clothing, gloves, or masks and if people don't follow common safety instructions like following pesticide label instructions or spraying pesticide against the wind direction. Additionally, a correlation between the number of pesticide applications of a person and the risk is observed [KKC+13]. A comparison between a group of pesticide handlers working under the US Environmental Protection Agency Worker Protection Standard (WPS), using appropriate and clean working clothes and gloves as well as proper hand washing after pesticide application, with a group working without WPS methods demonstrated that the group following WPS had lower levels of pesticide metabolites in their urine than the other group. However, the analyzed pesticide metabolite levels overall were higher than in the general population [SBC+08]. According to [SMS+03], there is a relation between pesticide residues in the homes of people and the amount of pesticide applied in the surrounding region and the distance of the homes to agricultural fields. Therefore, reducing the applied pesticide amount and building new homes away from agricultural fields might be mechanisms to reduce pesticide exposure.

The preceding chapter has shown that there are different exposure routes in residential and occupational environments. Exposure is higher for people involved in the agricultural working process no matter how non-agricultural workers are exposed to pesticides. The results of this chapter can help to identify possible exposure routes and derived exposure reduction strategies.

2.2.4 Agriculture in less-developed countries, risk groups and factors influencing the risk

In the following chapter, the characteristics of agriculture and pesticide use in less-developed countries is analyzed. Additionally, observations that were made during a field visit to a agricultural community in El Salvador within the LLinES project are presented. The field trip to the community Ciudad Romero in the Bajo Lempa region in El Salvador took part in December 2015 (section 2.1.3).

In less-developed countries, most farmers are so-called small-scale farmers with highly labor-intensive working processes and less use of machines like tractors. Globally, 90% of all farms are managed by only one person or family, mainly under small-scale farming conditions. In less-developed countries a trend can be observed that farm sizes get smaller [Kwa01].

In general, groups with a high risk of being exposed in high doses to pesticides are pesticide formulators and people involved in the pesticide production process, people working on the agricultural fields, and people involved in the pesticide application process like pesticide mixers, loaders and sprayers. The risk for people in the pesticide production process is higher than that of those working in agriculture [ASC09].

According to [GLR99], the grade of education in general and the illiteracy rate are related to the probability for the occurrence of a agrochemical related negative health impact. This can be explained by the fact that illiterate farm workers are not able to read and therefore do not receive important information about the material safety data sheet (MSDS) of the used agrochemicals. Information given in the safety data sheets for agrochemicals contain information about health risks, correct storage of the chemical, protection issues, and application patterns, such as application rates. In general, they contain the information needed for safe use of the substance [Age15a].

[NGKD06] reported that Ghanaian cocoa farmers wear little or no protective clothing when applying pesticides. Money for protective clothing is often not available in poor farming communities. Additionally, protective clothing is not usable in less-developed countries in high temperature regions due to the temperature risks [Din93]. A lack of training and education in pesticide and agrochemical use can be observed in many less-developed countries [OCB$^+$02]. Observations during the field trip in the Bajo Lempa in El Salvador confirmed the mentioned facts. Agrochemical application workflow is characterized by a strongly labor-intensive agricultural practice with the use of little or no protective clothing. Figure 2.5 shows some campesinos applying pesticides on a young corn field. The farmers apply pesticides with a manual hand pump walking on a line over the field.

Some of the farmers do not have long-sleeved clothes; nobody wears a mask or a towel over their mouth and nose to protect themselves against the exposure through pesticides. Through the walk over the field and the use of manual hand pumps, the farm workers are exposed for a longer time to the pesticide dust than, e.g., when it is applied by a tractor. The same is reported in other studies, for example, from Pakistan [KSM$^+$10], Mexico [BML11] or Cameroon [SANN18].

Figure 2.5: Campesinos in the Bajo Lempa region, El Salvador, applying pesticides with manual hand pumps and without sufficient protective clothing (source: M. Hieber-Ruiz)

Something similar could be recognized at a field on which farm workers applied fertilizers on a corn field, as visualized in figure 2.6. They applied fertilizers during a walk over the field by hand on every single corn plant. They do not wear gloves or anything similar, meaning that they are exposed to the fertilizer directly through their skin.

[Dev10] found out that there is a general connection between the poverty level of a society and the misuse and wrong application of pesticides. This confirms the thesis that most cases of agrochemically induced negative effects can be observed in developing countries where, in general, the poverty rate is higher than in developed countries.
Other reasons for the high prevalence of agrochemical related diseases in less-developed countries are a observable lethargy in introducing innovations in terms of agrochemical use and safety in the agricultural process as well as a lack of legislative pesticide regulations in less-developed countries [Bul82]. According to [SGKJ21], health issues related to pesticide use in less-developed countries are often attributed to a wrong assessment of the risk caused by pesticides, the use of prohibited or illegal pesticides, a lack of pesticide regulation, and a lack of literacy needed to read and understand the safety data sheets of pesticides.
A lack of pesticide regulations results in partly highly toxic substances that are prohibited in

Figure 2.6: Campesinos in the Bajo Lempa region, El Salvador, applying fertilizer by hand on plants (source: M. Hieber-Ruiz)

countries belonging to the EU or the USA being still legal in some less-developed countries [Eco01].

However, policy options like subsidies for organically produced products, prohibitions of pesticides, and so on can lead to a decreasing use of pesticides [fECoO97]. In some regions of less-developed countries, pesticides are not used due to an incorrect assessment of the relation between disease and yield loss, e.g., Ghanaian peanut farmers do not use pesticides on peanuts because they lack awareness about diseases and yield losses. Pesticide application management is sometimes based on an crop calendar, e.g., Ghanaian cocoa farmers [NGKD06].

[Awu97] reported a relatively high effort yet ineffective fungicide management practice in the wet season of tomato growing due to a high wash off, which is attributed to the high amount of rain in the wet season. [MQL$^+$14] reports from exposure to pesticides caused by incorrect disposal of empty pesticide containers, mixing of different pesticides together in one application not intended to be mixable, and the use of highly toxic pesticides can be observed in less-developed countries.

Summarizing, agriculture in general as well as pesticide use, risk perception, and literacy differ between less- and highly developed countries. Therefore, existing risk and exposure strategies as used in industrialized agriculture are not directly applicable in less-developed countries and

must be adapted.

The results of the actual section and of section 2.2.1 can be used to identify risks caused by pesticides, find steps in the agricultural workflow with exposure to pesticides, and therefore to derive options of action for risk mitigation in the field of ecotoxicology and pesticides. During the mentioned field trip to the Bajo Lempa region in El Salvador, some of the potential exposure routes and problems with pesticide use in less-developed countries could be observed, such as lack of protective clothing and low risk literacy.

2.2.5 Models to estimate the fate of chemicals and human exposure

In the framework described in this thesis, a system is developed that can help people in making decisions, e.g., to avoid areas with pesticide contamination in the air or drinking water contaminated with agrochemicals. Since agrochemicals are released into the environment and continuous temporal measurements of pesticide concentrations are usually not affordable, fate and exposure models can help to estimate the temporal and spatial distribution of such parameters, which can be an appropriate decision support tool.

There are different models available to estimate pesticide concentrations in the environment or the vulnerability, for example, of ground-, soil, or surface water also used by authorities for regulatory risk assessment. Additionally, human exposure models are available and used in risk assessment. Some examples of environmental and human exposure models are introduced in the following chapter.

According to [ZP12], there are two types of environmental fate models: index-based and process-based models; both have advantages and disadvantages for their use in a risk mitigation context. Index-based models calculate a index describing the risk of, for example, a groundwater reservoir to be polluted by contaminants. In contrast to process-based models, index-based models need only few input parameters, and the indexes are easily calculated. However, only an index of risk is calculated with which it is possible to compare different compartments in relation to their vulnerability; however, these index-based models do not calculate concentrations in the regarded compartment. These index-based risk models do not incorporate the properties of the substances or chemicals and transformation or transport processes and relevant parameters for these processes, such as the mentioned chemical properties, e.g., determining if a substance has the tendency to bound on organic compounds or if a substance has the tendency to be dissolved in water. With the help of such indexes, it is possible to compare the risk or vulnerability of different locations [ZP12]. In a risk mitigation

approach, these index-based models can be used to identify locations and areas under high risk or can help to identify and prioritize water bodies or water reservoirs, where a contamination might be possible, therefore making further investigations, e.g., by direct measurements necessary.

The second type of models are process-based models. In these models, transformation, transport, and transfer processes and related parameters are also included [TBVdLV06]. There are process-based fate models for the different compartments available that calculate environmental concentrations in single compartments. In the following section, some examples of such models are presented. However, in a multidisciplinary risk mitigation approach, an evaluation of such models according to their usability in the regarded region must be performed by an ecotoxicologist.

For estimating concentrations in surface water, models like *PWC (Pesticide in Water Calculator)*, which calculates the concentrations in surface water bodies and groundwater after the application to land surfaces [You16], the *PFAM (Pesticides in Flooded Applications Model)*, a model to estimate concentrations in water bodies originated by pesticides used in flooded areas [You13], or the *TOXSWA (TOXic substances in Surface WAters)* model, a model to estimate concentrations in sediment and water of surface water courses located near to agricultural fields [THBVdB16], are available. Another important model to estimate risk for water basins caused by agricultural pesticide application is called *SWAT (Soil and Water Assessment Tool)* [ASMW98]. The recent version *SWAT+* has a temporal dimension of 1 day; landscapes and watersheds can be modeled by implementing the landscape's characteristics with spatial objects [BAR$^+$17]. Originally, both models were developed for medium and large watersheds but were also applied successfully on small watersheds [WGB$^+$24].

A model to estimate bioaccumulation and related concentrations in freshwater fish is the *KABAM (K_{OW} (based) Aquatic BioAccumulation Model)* model [Ger09]. A model to estimate deposition caused by different application scenarios is for example *AgDRIFT* [TBE$^+$02]. There are also multimedia models like *PEARL (Pesticide Emission Assessment at Regional and Local scales)*, which estimates pesticide concentrations for groundwater caused by leaching in surface water caused by drainage and the persistence of chemicals in the upper soil [VdBTBVdL16].

The mentioned models and methods are only some examples for existing models usable in a risk mitigation approach and applied by the *United States Environmental Protection Agency* (US EPA). However, there are also models available for other regions in the world,

such as models for the EU as proposed by *FOrum for the Co-ordination of pesticide fate models and their USe (FOCUS)*, an initiative of the European Commission to harmonize the calculation of predicted environmental concentrations (PEC) of active substances of plant protection products in the framework of the *EU Directive 91/414/EEC*. *FOCUS* is based on cooperation between scientists from regulatory agencies, academia, and industry [FOC21]. However, the *FOCUS* model tends to underestimate realistic fungicide [KMRS14] and insecticide [KSSS12] concentrations measured in field samples.

Models used in the EU context are, for example, *MACRO* for drainage modeling in cropped soil [BBJ01], *TOXSWA (TOXic substances in Surface WAters)* to estimate the fate of pesticides in surface waters located near to agricultural fields [THBVdB16], or *PRZM (Pesticide Root Zone Model)* for run-off modeling [Suá05]. *SWASH (Surface WAter Scenarios Help)* uses the results of the three models and manages model outputs to calculate pesticide exposure concentrations in surface waters in an agricultural context [TRvdBA$^+$15].

There are also multicompartment models available that combine models for single compartments and their outputs to a single model. [LSYS11] developed a multicompartment transport model consisting of different single compartment models for agricultural, soil, and atmospheric processes provided by the US EPA. The aim of the developed model is to estimate the temporal and spatial distribution of pesticides on a large scale. With the help of this model, the authors created, for example, a residue map of the pesticide toxaphene for the USA and Mexico with a resolution of $36 \times 36km$ grid [LSYS11].

However, the mentioned models partly use assumptions for the regions for which they were developed, for example, for the USA and the EU [FOC00]. For their use in less-developed countries, they must be adapted to the prevailing conditions in the region on which they should be applied.

In addition, the concentrations calculated with the described models do not necessarily represent concrete concentrations that are directly measurable in the environment [FOC00].

Besides the described environmental models, models directly related to human exposure and health were also developed:

For example, a model to estimate exposure for people located near to an area treated with soil fumigants is the *PERFUM (Probabilistic Exposure and Risk Model for FUMigants)* model [RG06]. *SHEDS (Stochastic Human Exposure and Dose Simulation Model)* are models to assess chemical exposure in a human population related to inhalation, dietary, and non-dietary intake as well as skin contact caused by everyday activities like dietary intake or near field

activities [IGE+14]. [GS05] review different methods for modeling pesticide concentrations in the air during the application process caused by spray drift.

Standard Operating Procedures for Residential Pesticide Exposure Assessment (Residential SOPs) are manuals, in which it is described how to model exposure related to residential pesticide applications like caused by pet treatment or by pesticide use in gardening [(USd].

Furthermore, there are also models available to estimate direct human exposure. For example, [FB15] provide a model to calculate dermal exposure. US EPA also provides methods to estimate exposure caused by occupational activities related to pesticides for pesticide handlers, like mixers, applicators, and flaggers [(US21a], as well as for people exposed by post-application activities like hand-harvesting, thinning, or scouting for pests [(US21b]. In 2014, [(EF14] published a guidance document on risk assessment about people involved in the pesticide application process and bystanders.

There are different models available to model environmental or human exposure. Each model has its own prerequisite for the use and uses different input parameters. In a risk mitigation approach, a stakeholder from the field of ecotoxicology must select an appropriate model with respect to the location where the model should be applied, the model prerequisites, the complexity of the model, and the available input data. Additionally, the adaptation of the mentioned models to the local environmental conditions in less-developed countries might be a task in a risk mitigation approach.

2.2.6 Effects for public health and ecosystems caused by pesticides

The last sections dealt with the use, fate, and exposure modeling of pesticides. In the following section, pesticide-induced effects on ecosystems and humans and how they are interconnected in the one health approach are explained.

2.2.6.1 One Health approach

The One Health concept is based on the idea that human health, animal health, and ecosystem health are interconnected. The *One Health Initiative* wants to bring together physicians, veterinarians, environmental researchers, and other health-related scientists to work together to improve the conditions for human health. The preponderance of publications on the theme of One Health deal with infectious diseases and their pathways from animals and wildlife to humans [Ass08].

Other environmental factors besides infectious diseases, like exposure to toxicants, heavy

metals, or pesticides are a main cause for diseases and death in developed and less-developed countries, and the number of people affected by these diseases or resulting death are higher in less-developed countries. The WHO sees the following as key factors for diseases and death in the sense of the One Health approach: lead exposure, polluted water, urban air pollution, indoor smoke produced by heating and cooking, climate change, and unintentional poisonings by toxic chemicals and pesticides. Unintentional poisonings are, according to the WHO, responsible for 355,000 death cases each year, with 60% of the cases occurring in less-developed countries. The reasons for the poisonings are inappropriate use and high exposure to these substances [Org15a]. Nevertheless, in this sense, less scientific work has been done about toxic exposure in animals and the consequences for human health in the past [But11].

There are many examples of how environmental health and contamination with chemicals or other toxins influences human health; for example, consumption of with methylmercury-contaminated food caused death and illnesses in several cases [MR06]. The first documented cases of illness and death caused by mercury-contaminated seafood was observed in Japan in 1956 [Aro05]. Several other cases are documented, e.g., in the 1990s, through the consumption of whale meat in the Faroe Islands [GWJ$^+$92].

Pesticides have an strong impact on the health of people. The preponderance of people affected by pesticide-related illness, poisoning, or death are agricultural workers or those involved in the agricultural production process who are exposed directly and acutely to pesticides. However, there are also indications that pesticides have an impact on people who are not indirectly exposed to pesticides and have a long-term chronic exposure [J$^+$90]. [WHY09], for example, found that there is a relation between the time of high pesticide concentrations in surface water bodies and the occurrence of birth defects.

These examples show how tightly the ecosystem health and human health are connected, and how important the One Health approach is. Thus, for the assessment of the risk to human health caused by pesticides, it is also important to examine the ecosystems and their contamination.

2.2.6.2 Hazardous effect of pesticides: toxicity

Risk caused by pesticides is dependent on the exposure and the toxicity [Fen13]. The word toxicology has its origin in the Greek language and is derived from the word “toxikon”. It is defined as a “science within human medicine that describes the effect of poisons on human”

[Ste04, p.9]. Toxicology analyses both the toxicokinetics and toxicodynamics. Toxicokinetics describes the type and velocity of the uptake of a substance, distribution in an organism, metabolism, and excretion of a substance, whereas toxicodynamics describes the interactions between the substance and molecular structures of the cells and the organism and the mechanism of the toxic effect [Fen13].

A poison or toxicant is defined as "any substance that causes a harmful effect when administered [...] to a living organism" [HL04, p.3]. Whether a substance can lead to harm for an organism and consequently is a poison depends nearly for all substances on the uptaken dose, as *Bombastus von Hohenheim* and *Paracelsus* formulated as "All substances are poisons; there is none which is not a poison. The right dose differentiates a poison from a remedy" [LK06, p.498]. For example, there are essential elements like copper, iron, or zinc that are toxic in high doses but are also essential in low doses for humans. This explains why the dose-response relationship is so important in toxicology [HL04].

Toxic effects can be reversible, e.g., through the regeneration of the damaged cells, or irreversible, e.g., through mutagenicity or mortality [Fen13]. Toxic effects can be lethal or sublethal, for example, behavioral effects or reduced reproduction or growth rate [DDD07].

Ecotoxicological endpoints are measured in standard laboratory tests in which organisms are exposed to different predefined substance concentrations, the resulting effects on the organisms are measured, and a resulting dose-response relationship is obtained. Toxicological endpoints, such as the No Observed Effect Concentration (NOEC) or LC_{50} can consequently be derived from the dose-response relationship [Rit10]. Toxicity to humans cannot be directly measured due to ethical concerns. Human toxicity values are often derived from toxicity values for mammals [(USc].

In reality, organisms are mostly exposed to more than one substance resulting in mixture toxicity. Substances can interact with each other and the organism with different effects: antagonistic, additive, or synergistic [Ced14]. However, risk assessment is based on single compounds. A short overview about human and ecotoxicological risk assessment is presented in section 3.2.

On the molecular level, there are about seven principal mechanisms in which pesticides react with biomolecules in an organism. According to [Ste04], the typical toxic biochemical mechanisms in a cell can be described with the following reactions:

- Inhibition of enzymes: the function of an enzyme or protein is inhibited by the substance.

- Disturbance of chemical signal systems: substance imitates a signal substance in the organism (agonist) or blocks the receptor for a signal substance.
- Generation of very reactive molecules: free radicals are produced which can destroy cellular components.
- Degradation of the pH gradient across membranes: production of energy in mitochondria is disturbed.
- Distortion of the physical structure of membranes: change of the physical characteristics of membranes.
- Disturbance of the electrolytic, osmotic balance or the pH.
- Destruction of tissue, proteins or Deoxyribonucleic Acid (DNA).

The toxicity of an agent is also dependent on the species exposed to the agent [Fen13]. e.g., some pesticides, such as pyrethroids, are very toxic to arthropods and fishes but less toxic to endotherms, such as mammals [Bör09].
This chapter demonstrated how substances like agrochemicals can act on organisms and which mechanisms lead to hazardous effects. This can be used, for example, to estimate the risk of agrochemicals in an LL approach.

2.2.6.3 Effects of pesticides on human health

Pesticides are substances not intended for direct contact with the human organism. However, due to the exposure pathways demonstrated in section 2.2.3, they can enter the human body and trigger different diseases or negative health effects, such as CKD. The following provides some examples of pesticide-induced negative health effects.
The type and the strength of the influence of pesticides on human health is substance specific. There are pesticides with a high impact on human health but also substances or groups with a low impact [KEKK21]. Not every human organism reacts in the same strength to the exposure to a pesticide. The reaction depends, e.g., on the metabolism rate of the organism or on the repair process of the DNA [Vai95].
There are different pesticide-caused effects discussed in literature: [VHS$^+$91] have reported detections of disorders of the skin, cardiovascular, respiratory, and nervous systems, as well as disorders of sensory organs, headache, sickness, disturbed electrical activity of the brain,

altered liver enzyme activity, and reduced lung function after exposure to insecticides. Increased abortions were observed in people exposed to pesticides [RRR87, RMD+90].

Neuronal diseases are also associated with the exposure to pesticides. [PC13] reported that the exposure to pesticides is a risk factor for the occurrence of Parkinson's disease. The carcinogenic characteristics of 12 pesticides allowed in the USA and Canada were reported by [WMC10]. An increase in the number of people who have lung cancer [ARB13, MMWZ88, GSMS11], skin cancer [FBH11], cancer of the lip and testis [WDHE89], and pancreatic cancer [AS12] is discussed with high pesticide exposure.

Glyphosate, a widely used herbicide, has an influence on kidney function through oxidative damage to the kidneys [LNLV12]. The relation between respiratory pesticide uptake and the presence of chronic obstructive pulmonary disease (COPD) was observed [YBMS13]. A review study from Brazilian farm workers who were occupationally exposed to pesticides showed different pesticide-related health symptoms like neurological and neurodegenerative effects, diabetes, DNA damage, or metabolic diseases [LFMBL+22]. However, not only can the active substances used for pest management harm the human organism. The active ingredients of pesticides are not sold on their own. The active ingredients are sold as formulations that means that other substances, such as emulsifiers, solvents, and carriers, are mixed with the active substance. These other substances can have also toxic characteristics for organisms [Anw97].

As mentioned earlier, pesticides may have adverse effects on human health. This means that in a capitalistic perspective, they lead to economic impacts. Depending on the strength of the negative health effect, people whose health is damaged may not able to work or may not have the same power to work like people without any limitations, thus they may not be as efficient as healthy people, get lower wages, and must pay for healthcare [ASJB11].

The examples listed in this chapter show that some substances belonging to the group of pesticides can harm the human organism with a strong impact. A reduction to the exposure of pesticides can help to increase the health of people – in cases other than of CKD and its possible relation to pesticides. Therefore, the focus of this thesis lies on pesticides and agrochemicals.

2.2.6.4 Ecological problems caused by pesticides

Pesticides are substances which act direct against living organisms like insects, plants, fungi, or other organisms. This effect is desirable on target sites, but they also have a strong effect

on non-target organisms. Several negative effects of pesticides on different ecological levels to aquatic and terrestrial ecosystems can be observed. In the following chapter, some examples of negative effects on biodiversity are demonstrated.

Besides their toxicity, some pesticides have characteristics that make them harmful for ecosystems, such as persistence and potential to bioaccumulate. Persistence means that they have a long half-life and remain a long time in the environment until they are degraded. Bioaccumulation means that these substances are fat soluble and accumulate along the food chain in fat tissues, leaving high concentrations in top predators putting them at the highest risk [MRB+13]. Pesticides that are persistent and bioaccumulating can also be detected in regions of the world where they were never applied [WM96]. Pesticide residues can be frequently found in air [SDTR+16, ZKPS+22] water [MRK+15, HA08], and soil samples [CHKW+17, VFV05].

In literature, there are several examples of negative effects on living biota caused by pesticides on every ecological level – from the decline of single species [HM08] up to the at least temporal destruction of parts of whole ecosystems [BH97]. *Rachel Carson's* book *Silent Spring* [Car62] brought the negative effects of pesticides on the environment to the attention of the general public for the first time. In her work, she describes the possible negative effects of pesticides giving the example of the DDT induced decline of griffin population.

Examples of negative effects can be found in both aquatic and terrestrial wildlife.

In general, insecticides used in agriculture can be a threat to surface waters [SS15] and pesticide-induced reduced biodiversity of aquatic invertebrates can be observed [BKSL13]. On the terrestrial side, a decline in biodiversity caused by insecticides and fungicides on European farmland can be observed [GBB+10]. Furthermore, the worldwide decline of the insect population can be attributed to pesticides, among other factors [SBW19]. The reduction of the breeding success of booted eagles is also attributed to pesticides [PBM+23]. Pesticide mixtures are also a risk for soil invertebrates [PvGV+22].

The effects of pesticides on species and ecosystems can be direct or indirect. Direct effects are aimed at the vital functions of the organism. They can be reversible, e.g., through repair processes, or they are irreversible, e.g., mortality. While local effects occur at the site of contact with the chemicals, systemic effects only occur after the pollutant has been distributed elsewhere. Indirect effects are only caused secondarily as a result of ecological interactions, e.g., predator-prey relationships, and ecosystem characteristics. They are also ultimately attributable to direct effects on organisms [Fen13].

To prevent negative effects on ecosystems caused by pesticides there are different approaches

for substance regulation. Persistent Organic Pollutants (POP), a group of substances with characteristics such as high bioaccumulation potential, tendency for being mobile, and high toxicity [RSF+95], are considered particularly hazardous pesticides and chemicals and are listed in the Stockholm Convention on Persistent Organic Pollutants, resulting in their restriction for production and use [otSC23]. The present chapter showed that pesticides enter ecosystems and have strong effect on them. Humans are interconnected with ecosystems, for example, they drink water from a river and eat fish or wildlife and fruits. Therefore, ecosystem health also plays an important role in the described LL approach.

2.2.6.5 Pesticides and their benefits

Despite the negative effects of pesticides mentioned in the last chapters, they also have beneficial effects.

Pesticides help to increase economic welfare by controlling pests and plant diseases and consequently minimizing yield loss. For example, [WHP22] analyzed the consequences of pesticide use on crop yield output of six main plants in USA and Canada. The mean increasing yield lies between 16 and 84%. According to [Pim97] each dollar spent on pesticides yields a $4 return through a protected or increased crop yield. On the other side, [LBB+14] have investigated the effect of pesticide use intensification on social, environmental, and economic factors in different cropping systems in France. They could not find a positive correlation between pesticide use intensity and profitability and productivity. However, economic analysis is more difficult because the costs for ecosystem and human health degradation must be implemented in an economic analysis to calculate the total cost of ownership for society.

Pesticide use results in higher agricultural production and lower food prices [Dam09]. Besides quantity, pesticides can also have an effect on the quality of the produced crops, e.g., crops are not damaged by an insect or fungi and can be sold to a higher price [SLZ+07].

Outside of agriculture, pesticides, and biocidal products are used in the private sector in home and garden, in the commercial sector as, for example, wood preservatives, municipal sectors to control traffic areas, or parks, and for vegetation control on railroad tracks and golf courses [Sat01]. Additionally, pesticides and mainly insecticides are used in many tropical and subtropical countries to control vectors that are responsible for vector-borne diseases like malaria, leishmaniasis, dengue, and other vector-borne diseases [vdBZY+12].

Summarizing, besides the negative effects described in chapters 2.2.6.3 and 2.2.6.4, pesticides are substances that have their benefits in agriculture and healthcare for vector control.

2.3 Chapter conclusion

In the present chapter, a characterization of CKDu was performed and potential risk factors were identified. One of the suspected risk factors is pesticides, although there is no clear evidence. The focus of this thesis is set on the risk factor pesticides because, besides CKDu, they are harmful to humans and ecosystems. Even if they are not the cause for CKDu, a reduction in pesticide exposure might help humans or ecosystems to resist other stressors better, which might be the cause for CKDu.

Additionally, the methods for determining the stage of CKD were explained, which in turn can be used in the described LL approach to determine the success of the implemented and applied risk mitigation strategies.

Furthermore, the general characteristics of less-developed countries as well as agriculture and pesticide use practices in these regions were elaborated. They are characterized by low economic and infrastructural resources, high manual-labor-intensive work, and a lack of education and risk literacy. These characteristics must be considered in different tasks in the used approach, e.g., for the adaptation of possible risk mitigation strategies (part III), the adaptation of the used research and development environment (chapter 5), or the adaptation of methods for generating a SDSS (chapters 15 and 18).

Pesticide use and fate, as well as fate and exposure modeling, and the risks and benefits caused by pesticides were analyzed. The results help to identify potential pathways where risk mitigation strategies can be implemented in the described approach with the given characteristics.

Summarizing, results of this chapter help to adjust methods necessary for the described approach to the characteristics of countries affected by CKDu.

3 Risk and risk assessment

One of the aims of this thesis is to develop risk mitigation strategies. To mitigate risk, it is essential to have a proper definition of risk and how it can be assessed.

3.1 The term risk as it relates to pesticides

Following some definitions of risk are introduced and discussed, including how they can be used in the developed framework.

3.1.1 Introduction to the term risk and some definitions

There are several definitions for the term risk. Nearly each scientific field has its own definition [Thy06].
The following is a definition for the term risk that is mainly used in disaster management:

Definition 3.1.1 (Risk) *Risk R is defined as the probability P of an adverse event E times the consequences or impact I if the event happens*

$$R(E) = P(E) \cdot I(E). \tag{3.1}$$

[FHLL05]

However, in natural disaster risk assessment, there are also risk definitions used that examine only the probability of a predefined impact or endpoint ("risk is defined as the probability that an event will occur") [Thy06] or on the impact of the event ("risk is defined as the expected number of lives lost, people injured, damage to property and disruption of economic activities due to a particular natural phenomenon") [Tie92]. Definitions given by [Thy06] and [Tie92] can be regarded as special cases of the definition given by [FHLL05]; the first one uses a predefined impact, for example, "number of people suffering from the disease" – the second negates the probability or sets it to one.

To demonstrate the outcome of the definition above, it is assumed that the risk of two volcanoes, A and B, on the forests around them should be assessed. With a GIS analysis, it is possible to estimate how many trees would be destroyed if both volcanoes erupt. The GIS analysis results show that volcano A would destroy 50 trees and volcano B 200 trees. A geologist assumes the probabilities of eruption in the next 10 years for volcano A about 0.5 and for volcano B about 0.1.
Therefore, the risk for both volcanoes can be calculated as follows:

$$R(A) = P(A) \cdot I(A) = 50 \cdot 0.5 = 25 \tag{3.2}$$

and

$$R(B) = P(B) \cdot I(B) = 200 \cdot 0.1 = 20. \tag{3.3}$$

The example shows that the risk for volcano A is higher than for volcano B, although fewer trees are destroyed by an eruption. This can be attributed to the much higher probability for an eruption.
Regarding human health risk, for example for CKD, the definitions above have some limitations. The risk of a person suffering from CKD could be, for example, expressed as the probability that the person will develop CKD. With this definition, the stage of CKD is not considered. On the other side, the personal impact could be expressed as the stage of CKD. There is the question if the risk values are comparable.
The interpretation of the calculated risk values and whether they can be used as a value to compare the risk of the person is a task that must be answered by a stakeholder from medical science in an LL approach. Additionally, the related impact must be selected by the stakeholders. Another possible impact parameter could be, for example, the amount of money necessary to treat the CKD stage in a medical facility. Using the risk definition proposed by [Tie92], risk could also be estimated by the number of people suffering from CKD per area.
Risk can be further differentiated into the individual risk and the social or population risk. The individual risk is related to the probability for a single organism to cause a defined adverse effect. The social or population risk is further associated with the number of organisms that suffer an adverse effect and the frequency with which this effect occurs [Ill02].
The individual risk in the framework of agrochemicals has an individual component that is influenced by personal health parameters and can be calculated individually for each person. The risk to suffer a given negative effect can also be related to a specified area, e.g., the risk for a community. This value describes the aggregated risk over the community population.

Definition 3.1.2 (Aggregated Risk) *The aggregated risk over an area* $Risk_{area}$ *with* $i = 1, ..., n$ *events is defined as the sum of the products of the probability* $P(E_i) \in [0; 1]$ *and the impact* $I(E_i)$*:*

$$Risk_{area} = \sum_{i=1}^{n} P(E_i) \cdot I(E_i) \tag{3.4}$$

The difference between individual and social or population risk can be illustrated with the following example:
A field was treated with pesticides by airplane. Directly after the application, substance residues are still in the air with a specific concentration, and people or farm workers are working for a specific duration on the field. In case A, only three people work in the field, in case B ten people work on the field for the same period of time as in case A. When ignoring the individual health history of the farm workers, in both scenarios the individual risk is the same because of the same concentration and the same exposure duration. The individual risk for a person working in the field to suffer an adverse effect can be described as, e.g., 30%. The social or population risk in case B is higher than in case A because more people are affected by the adverse effect.
The difference of these two risk concepts is also important for this thesis. A value for the individual risk is used in a personal SDSS, e.g., the change in the value for the individual risk for different routes with the aim of suggesting the user a route through a field with the lowest exposure and therefore with the lowest risk.
The social risk and derived maps are relevant parameters for logistic optimization or for authorities to determine in which regions intervention is necessary. To define the social risk, we must multiply the probability that a person suffers death or a illness times the people of living in this area.

As this thesis deals with the risk caused by pesticides or agrochemicals, an ecotoxicological risk definition is also presented. According to [Thy06], the chemical risk is mainly associated with the terms event, hazard, and probability.

Definition 3.1.3 (Hazard) *Hazard is defined as the "inherent property of an agent (e.g., pesticide) or situation having the potential to cause adverse effects when an organism, system, or (sub-) population is exposed to that agent or situation" [SFH+06, p. 2108].*

The hazards of pesticides are hazards for human and ecosystem health. The WHO uses the

Table 3.1: Classification of pesticides according to the WHO [Org10]

WHO class		**LD50 for rat (mg/kg body weight)**	
		Dermal	Oral
Ia	Extremely hazardous	<50	<5
Ib	Highly hazardous	50-200	5-50
II	Moderately hazardous	200-2000	50-2000
III	Slightly hazardous	>2000	>2000
U	Unlikely to present acute hazard	>5000	>5000

classification theme given in Table 3.1, based on the toxicity values for rats.

The hazard caused by chemicals for the different organisms can be expressed as the toxicity of the substance. The toxic hazard for an organism or an ecosystem caused by a specific chemical or pesticide is related to the exposure.

Definition 3.1.4 (Toxicity) *Toxicity is defined as the "capacity to cause injury to a living organism defined with reference to the quantity of substance administered or absorbed, the way in which the substance is administered and distributed in time (single or repeated doses), the type and severity of injury, the time needed to produce the injury, the nature of the organism(s) affected, and other relevant conditions" [SFH+06, p. 2143].*

Definition 3.1.5 (Exposure) *Exposure is defined as the "concentration or amount of a pesticide (or agent) that reaches a target organism, system, or (sub-)population in a specific frequency for a defined duration" [SFH+06, p. 2100].*

The exposure is defined by a certain amount of the chemical and the frequency and the duration the organism or ecosystem is exposed to the chemical.
In the sense of pesticides or chemicals, risk is defined as follows:

Definition 3.1.6 (Risk probability) *The risk is defined as the "probability of an adverse effect in an organism, system, or (sub-) population caused under specified circumstances by exposure to an agent"[SFH+06, p. 2134].*

This risk definition is again related to the definition above with a probability and an impact or adverse effect. However, when applying the definition proposed by [SFH+06] to a agrochemical framework, there might be the problem that it is hard to estimate the probability of an

adverse effect because it can only be done by medical experts with expertise in effects of agrochemicals on the structures of living organisms. Additionally, knowledge about previous health history and possible previous damage is also essential in order to make statements about the probability of negative health effects. This might be a difficult task in an LL approach in less-developed countries, where resources are limited.
A more appropriate definition for the agrochemical related risk is provided by the US EPA:

Definition 3.1.7 (Risk in risk assessment) *The ecotoxicological risk for organisms can be described by the following equation:*

$$Risk_{USEPA} = Toxicity \cdot Exposure \tag{3.5}$$

[(USa]

In this sense exposure is a value expressing the amount of a substance to which an organism is exposed, e.g., a fish in a river is exposed to 5 $\frac{\mu g}{l}$ to a substance for 10 minutes. The toxicity value is a value derived from laboratory, semi-field, or field studies, describing a concentration with a specific endpoint, e.g., how high the concentration is to kill 50 % of the test organisms in 10 min exposure or what is the concentration that does not have an effect on the organisms. According to definitions 3.1.4 and 3.1.5, toxicity is related to the substance S, exposure to the exposed amount d, the duration of exposure t and the frequency of exposure f. Therefore, Definition 3.1.7 can be rewritten as:

$$Risk_{USEPA}(S, d, t, f) = Toxicity(S) \cdot Exposure(d, t, f) \tag{3.6}$$

The equation makes it clear that the risk associated with agrochemicals depends on the substance used S and the quantity d, exposure duration t and frequency f of exposure. These are parameters which must be adjusted to mitigate the risk.
Ecotoxicological risk and toxicity assessment, methods to determine risk and toxicity, and their limitations for the described LL approach in developing countries are highlighted in more detail in the following section 3.2.

3.1.2 Discussion of the relation between the risk definitions

In the last section, different definitions of the term risk were presented. A risk definition used in disaster management relates the probability of an adverse effect to the impact of

the effect. However, there are also risk definitions related to disaster management in which the risk is expressed only with the probability of an adverse effect or the number of people suffering from the negative effect. [Tie92] also distinguish between the individual risk, e.g., the personal probability to suffer a negative health effect, and the social risk, e.g., how many people suffer a negative effect in a specific area. Another risk definition that was introduced in the last chapter was the ecotoxicological risk definition as proposed by the US EPA, whereby the agrochemical risk is related to exposure and toxicity.

It is difficult to handle the disaster risk definition for individual spatial decision support. The probabilities of an adverse effect can be only rated by medical experts with knowledge about personal health history. Maybe a framework that can be solved through the use of a expert system with fuzzy rules, as introduced in section 16.2.

However, definition for a social risk can be used, for example, by authorities or in the proposed LL approach to identify areas under high risk, for example, by estimating the probability of suffering from a disease or the number of people per area suffering from the disease.

In the field of an SDSS for personal support, the definition proposed by [(USa] is further used. In this definition, risk is determined by the product between exposure and toxicity. Therefore, to mitigate the individual risk, toxicity or exposure can be reduced. As toxicity is a value related to the substance, reducing toxicity means using a less toxic pesticide. As exposure is related to the amount of a substance an organism is exposed to as well as the exposure time and frequency, reducing exposure means reducing the amount, time, or frequency of exposure.

In the present chapter, the different risk definitions were discussed for their use in an SDSS in an LL approach. The following chapter deals with how the risk for humans and ecosystems can be assessed and managed.

3.2 Risk assessment and risk management

In the following chapters a general overview about human and ecological risk assessment of pesticides is presented. In the EU and USA, there is an authorization requirement for plant protection products. In order to place a chemical substance on the market, a risk assessment must be carried out that includes risk assessment for humans and ecosystems. Risk assessments for chemicals are subject to different authorities, such as the *European Food Safety Authority* (EFSA) or national regulatory authorities for countries belonging to the EU and US EPA for USA. An authorization is not granted until it is determined that there is no

unacceptable risk caused by the substance [GJRS24].

3.2.1 Human risk assessment

Part of the authorization process for pesticides is the risk assessment.

Definition 3.2.1 (Risk assessment) *Risk assessment is defined as a "process intended to calculate or estimate the risk to a given target organism, system, or (sub-) population, including the identification of attendant uncertainties, following exposure to a particular pesticide or agent of concern as well as the characteristics of the specific target system. It is the first component in a risk analysis process" [SFH+06, p. 2134].*

How legal risk assessment must be performed is established in different acts and regulations. For the EU, regulations about the assessment of whether a pesticide is safe or not can be found in *Regulation (EC) No 1107/2009* [Com09], and for the USA, it is found in the *Federal Insecticide, Fungicide, and Rodenticide Act* [Tou20].

Human risk assessment is used to estimate an acceptable dose of exposure, so that no negative effect on human health can occur. Before pesticides are used in the USA and EU, a risk assessment must be performed. Pesticides are not allowed to be used until their safe use is assured [oPR13]. That means that no adverse health effects should occur through the use of pesticides after permission. In reality, this is not the case, and despite the conducted risk assessment and compliance with application rules, pesticide-induced negative health effects are observed in the USA. Observed negative health effects can be reported to the US EPA, and pesticides can be reevaluated [Cen13].

According to the US EPA, the human health risk assessment is a process in which the probability of occurrence and the nature of negative health effects caused by the exposure to pesticides is estimated. Thereby, the type of health problems, the possibility of occurrence of different health effects dependent on the exposure level, threshold for acceptable doses or concentrations without risk, prediction of the exposure concentration or dose and the duration, vulnerability, and exposure probability of different groups of people are estimated and analyzed [(USc].

According to the US EPA, risk assessment for human health consists of four steps; for each step there are guidelines in which the procedures are described: hazard identification, dose-response assessment, exposure assessment, and risk characterization [oPR13, (USc]. An overview of the different steps and how they are connected with each other can be taken from figure 3.1

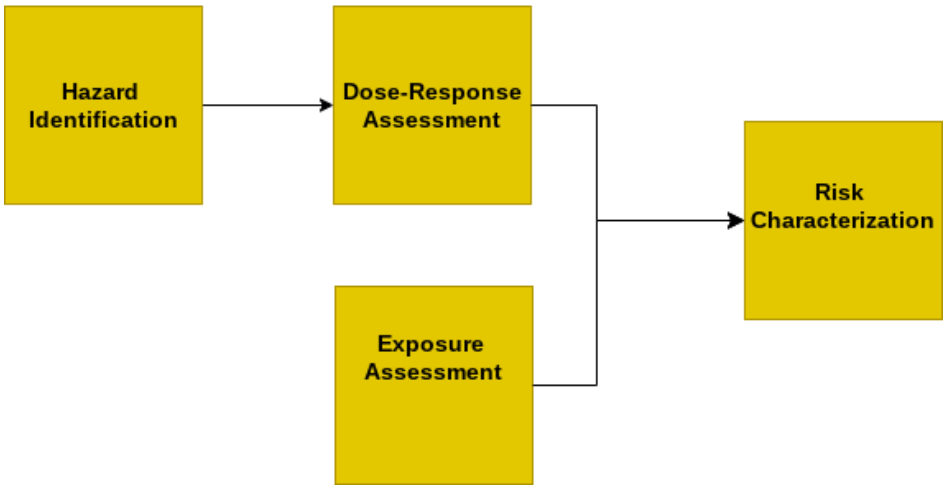

Figure 3.1: Overview of the different steps in the US EPA risk assessment for pesticides and human health. Source:[(US13b] (figure generated with *draw.io*).

In the hazard identification step, the toxic properties of the chemical are reviewed, e.g., the type and the strength of the negative health effect and if the exposure to these stressors can increase the incidence of these effects [oPR13]. Mostly, controlled clinical studies on humans are not available due to ethical concerns [(US13b]. Several studies have shown that chemicals that cause negative health effects in animals also cause such effects on human health [(US91]. Thus, animal laboratory toxicity studies (e.g., on rats, mice, dogs, and monkeys) are used. There are different guidelines available in which it is described how toxicity studies must be performed, for example, as provided by the OECD. These test guidelines describe the conditions under which the tests must be performed, e.g., temperature, duration, test organisms etc.

The expected exposure level at which predefined effects occur and the expected level where negative effects do not occur are also determined [oPR13]. Research about the toxicokinetics, the toxicodynamics, and the mode of action are involved in this process [(US13b]. This step is needed to clarify whether from an agent, harmful effects can be expected and the conditions under which these harmful effects can appear [(USc].

The second step is the dose-response assessment. For a non-linear dose-response correlation through animal laboratory studies, the highest dose where no significant effect against a non-treated control group occurs is determined, the No Observed Adverse Effect Level (NOAEL) [(US13a].

The use of animal toxicity data for human health risk assessment implies some uncertainties.

Therefore, a uncertainty factor of ten is mostly used to extrapolate the no effect dose from animals to humans. Additionally, the difference between human beings is considered, and it is assumed that the most sensitive human is ten times more sensitive than a standard human being. Thus, the overall uncertainty factor is 100 and is considered in the risk characterization. With the help of the NOAEL, which is derived from animal studies, the reference dose (RfD) or reference concentration (RfC) is calculated by dividing the NOAEL by the uncertainty factor. The RfD or RfC are defined as the "daily oral exposure to the human population (including sensitive groups, such as asthmatics, or life stages, such as children or the elderly) that is likely to be without an appreciable risk of deleterious effects during a lifetime" [(US13a].

In the third step, called exposure assessment, the magnitude, the way of exposure (e.g., inhalation, ingestion or dermal absorption), the frequency, and the duration (acute, chronic) of the human pesticide exposure is examined [oPR13]. It must be evaluated e.g., which groups of people are exposed to the pesticides, the number of people in these groups that are exposed, do the exposed groups exposed have higher vulnerability, and the type and frequency of the exposure for the different person groups. With this information, the human uptake over time is estimated [(US92].

For agricultural workers, it is assumed that they wear clothing with long sleeves, long trousers, shoes and stockings, and protective gloves [(US21a]. In relation to the application technique, the exposure assessment and the level of exposure is estimated according to different application scenarios, such as "open pour mixing/loading liquids," "open cab groundboom applications of liquids," or "Closed cab airblast applications" [(AH13].

In the last step of the risk assessment, the risk characterization, a quantified statement about the risk assessment can be given. The values gained from the dose-response assessment and from the exposure assessment are compared with each other. For example, in the risk assessment for a farmer who is using pesticides, a value for the expected dose to which he is exposed is compared with the No Observed Effect Level (NOEL) from an animal laboratory study by calculating the ratio between the NOEL and the expected exposure dose. This ratio is compared with an uncertainty factor (mostly 100) to be on the protective site that no negative effect can occur [oPR13].

The results of the risk assessment are surveyed from the relevant authority, and it is determined whether the risk is acceptable. If the risk is unacceptable, a risk management is started. In the risk management process, it should be analyzed whether and how the risk can

be mitigated, e.g., through the reduction of the exposure level [oPR13]. If the risk assessment shows that there is a too high risk, the authority can enforce rules about the handling or the application of the pesticide [(US92].

3.2.2 Ecotoxicological risk assessment

Besides the assessment of human health risks, an ecological risk assessment also must be performed in the EU and USA.

For example, in the EU, ecological risk assessment consists of aquatic and terrestrial animal and plant toxicity testing as well as environmental fate testing must be performed. This includes toxicity tests for birds, non-target insects, mammals, aquatic animals, and tests on non-target plants [Com09].

Toxicity tests are performed under standard conditions with predefined test parameters like duration, toxic endpoints, concentrations, test medium, and used organism. Test guidelines are, for example, available through the OECD [fECoO24], EU [Com24] or US EPA [(USb].

An example of a test guideline in the EU is the *Guidance document on aquatic ecotoxicology in the context of the Directive 91/414/EEC* [Com02]. According to this guideline, the risk for organisms caused by the application of pesticides is assessed by the comparison of the toxicity of the used component and the probable exposure concentration of an organism. The ratio between these values is called Toxicity Exposure Ratio (TER). For the assessment of the toxicity standard, toxicity tests, for example, laboratory tests, are used to determine toxicological endpoints like the lethal concentration where 50% of the individuals of the tested organisms die (Lethal Concentration $(LC)_{50}$), or the lethal dose where 50% of the individuals of the tested organisms die (Lethal Dose $(LD)_{50}$). The exposure, for example, for aquatic organisms, is determined by calculating a so-called Predicted Environmental Concentration (PEC).

$$TER = \frac{toxicity}{exposure} = \frac{LD_{50}}{PEC} = \frac{LD_{50}}{c_{waterbody}} \tag{3.7}$$

To assess if the application of a pesticide can cause an acute risk for a certain species, the ratio between the toxicity, in this case the LC_{50}, and the exposure (PEC), is compared with a safety factor. If the ratio between the toxicity and the exposure is less than the safety factor, there is an acute risk for the tested species; if it is equal or greater than the safety factor there is no acute risk [Com02]. According to [Com02], in the case of the risk assessment for

aquatic organisms, the safety factor is determined as 100:

$$TER < 100: \text{ acute risk} \tag{3.8}$$

and

$$TER \geq 100: \text{ no risk} \tag{3.9}$$

PEC is calculated with different models and scenarios and software tools. For example in the EU the model developed by the *FOrum for the Co-ordination of pesticide fate models and their USe (FOCUS)* are used to estimate the PEC for sediment and surface waters [Com11]. Toxicity testing and exposure assessment is a tiered approach in which the uncertainty and hence the safety factor can be decreased with more realistic studies, such as mesocosm and field studies, or more realistic exposure and fate models.

3.2.3 Mixture toxicity

In reality, mixtures of different pesticides can often be observed in the environment. These mixtures are not regarded during the legal pesticide risk assessment. However, in the proposed agrochemical risk mitigation framework in developing countries such mixtures might be observed in the environment.

To compare the effects of different pesticides and related concentrations in the environment, the concept of toxic units (TU) was introduced by [Spr70]. In this concept, the concentration of a pollutant i in the environment is related to the toxicity of the regarded substance:

$$TU_i = \frac{c_i}{EC_{xi}} \quad [\text{JBF}^+06] \tag{3.10}$$

In the equation above, c_i represents the concentration of substance i, and EC_{xi} represents a toxic endpoint of substance i, for example, the effect concentration EC_{50}, concentration where a 50% effect can be observed.

The combined toxicological strength of a mixture is calculated as:

$$TU_{mix} = \sum_{i=1}^{n} TU_i \tag{3.11}$$

whereby n is the number of toxicants in the mixture [JBF$^+$06].

The above equation uses the assumption of interaction additivity; however, antagonistic or synergistic effects can also happen when different substances are mixed [HBW21].

3.2.4 Pesticide legislation in developing countries

According to [SGKJ21] and [Eco01], health issues related to pesticide use in less-developed countries are attributed to the use of prohibited or illegal pesticides and missing pesticide regulation.

[SGKJ21] reports that the pesticide registration process in Kenya is one of the strictest in African countries. However, the authorization process is limited to analyses about the efficiency and purity of the substance; human and ecological risk assessment are not part of the registration process.

[ST12] analyzed the progress in the ratification of the *Rotterdam Convention* for 139 countries in relation to the countries' income group. In 2010, in the majority of the low- and low-middle-income countries, the use of substances listed in Annex III of the *Rotterdam Convention* was completely unrestricted. In more than 80% of the high income countries, restrictions were made for one or more of the listed substances.

Furthermore, ecological risk assessment has limitations in some of the less-developed countries. [CMW+14] analyzed the status of aquatic risk assessment in Latin American countries. Realistic exposure scenarios for Latin American countries are often missing. Additionally, in international standard test guidelines of the OECD, local species are not used as test organisms. Adaptation of these guidelines to the environmental conditions in these countries is necessary.

The pesticide registration process is often less strict in developing countries than in developed countries, resulting partly in a large number of registered substances [BP17].

3.3 Chapter conclusion

In the present chapter, different definitions of risk were introduced. In an LL approach, they must be selected according to the topic and the domain on which they work. For example, the social risk definition proposed by [Tie92], expressed as the number of people per area suffering from CKD, can be used in the LL approach to identify areas of concern.

However, this risk definition has some limitations in the framework of personalized decision support. For the SDSS developed in the following chapters, the risk definition used in legal risk assessment is more suitable. This definition relates risk to the toxicity of the substance and to the exposure level. As exposure is related to the amount of a substance an organism is exposed to as well as the exposure time and frequency, reducing exposure means reducing

the amount, time, or frequency of exposure. Reducing toxicity means to use a substance which is less harmful. Therefore, risk mitigation strategies in the proposed LL framework, as developed in part III, must work on these parameters.

Pesticide authorization and related risk assessment is well regulated in countries belonging to the EU and in the USA. However, limitations in the used models can also be observed and are discussed [KMRS14, BP17]. Adaptation of the used models and standard laboratory tests to the environmental characteristics in less-developed countries are missing but necessary. Additionally, human risk and exposure assessment as carried out in developed countries cannot be applied directly to developing countries. For example, the exposure scenarios used assume the use of protective clothing or tractors, which is not always the case in less-developed countries (section 2.2.4), resulting in higher exposure levels.

Therefore, involvement of stakeholders that are connected to the pesticide authorization process might be necessary in the LL approach.

4 Mathematical modeling and CKD

The following chapter introduces methods from mathematical modeling and shows how mathematical modeling is implemented in the proposed LL approach.

4.1 Mathematical modeling and system analysis

This thesis deals with the problem of a disease of unknown etiology in less-developed countries. One of the aims of this thesis is to describe methods with which it is possible to mitigate the risk caused by the disease. For different reasons, it is necessary to develop mathematical models, e.g., to predict the severity of a disease or to estimate the best fitting risk mitigation strategies. Therefore, principles of mathematical modeling and how they can be used in the context of an LL are described in the following sections.
As the etiology of the disease and the effectiveness of the risk mitigation strategies are not directly known and must be analyzed, a useful tool can be to work with models, whereby the models should describe CKDu with the hypothesized etiology and risk mitigation strategies. The following section gives an overview about how the described problem can be regarded as a system and how mathematical models can be derived from.
First, the term system must be defined.

Definition 4.1.1 (System) *A system S is a set of objects, also called system components or system variables, between which there are relations [IK13].*

A schematic representation of a system is visualized in figure 4.1. Relations between the objects are called inner relations. A system can consist of different subsystems which are again in relation to each other. The analysis of a system by the classification and analysis of the subsystems is called reductionism. Understanding the subsystems and their interactions can help to understand systems that are not completely understandable. System modeling has the aim of building simplified images or models of a complex system, whereby the main function of the system should be visible. That means that a system modeler has always to

walk on the edge between the complexity of a system and the oversimplification [IK13].
A system S consists of a set of different system variables $\{V_1, V_2, ...\}$, which are connected by inner relations. S has a boundary to the surrounding environment, which is called system boundary B. System boundary does not mean that there is no interaction between the environment and the system. Such interactions between the system and the environment are called outer relations. However, in most cases, the influence of the system to the environment is negligible, as the environment is much larger than the system. Overall, a system can be understood as a theoretical construct that helps to understand the world. There is no absolutely valid system boundary; the selection of the system boundary is related to the research question [IK13].
To analyze a system, it must be visualized. Such a visualization is called a model.

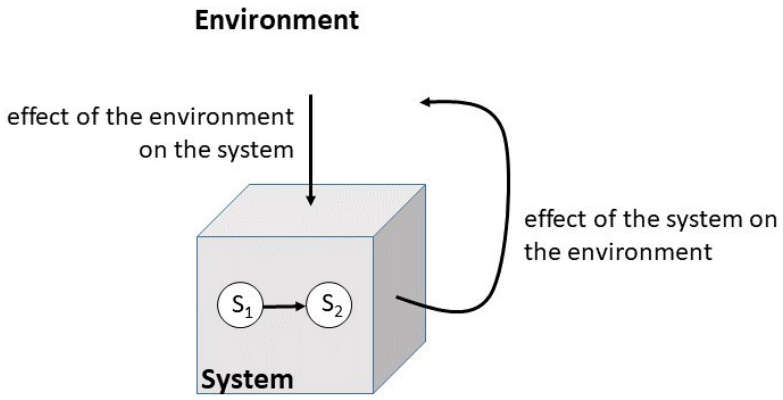

Figure 4.1: Schematic representation of a system according to [IK13] with system variables S_1 and S_2 , an inner relation between V_1 and V_2, and outer relations between the system and the environment as well as between the environment and the system. (figure generated with *LibreOffice Draw*).

Definition 4.1.2 (Model) *A model is a concept with which a system can be represented in a simplified way [IK13].*

The visualization of a system can be in different ways, for example, a town planner builds the model of a town with small objects, representing streets, buildings, trees, etc. Systems can be, for example, also visualized by words and sentences. The method used in this thesis is to visualize systems with mathematical tools with the aim of building a mathematical model.

In scientific literature, there are different definitions for the term mathematical model available.

Definition 4.1.3 (Mathematical model I) *A mathematical model is a projection or representation of a natural or artificial original that is limited in relation to the mathematical attributes that appear as relevant for the model builder. In this sense, mathematical models are subjective and cannot be assigned uniquely to the original [Rei16].*

Regarding this definition, a mathematical model is an abstraction of a problem taken from reality, whereby only factors or input variables are taken that are suspected to have a relevance for the outcome value of the model. The decision of whether a parameter is relevant is undertaken by the model builder and should satisfy scientific requirements. The process of developing a mathematical model is called modeling.
[GV14] gives a more precise definition in terms of mathematics:

Definition 4.1.4 (Mathematical model II) *A mathematical model E is an object (S, Q, M), whereby S is a system, Q a question regarding S and M a set of mathematical propositions $M = \{\Sigma_1, \Sigma_2, ..., \Sigma_n\}$, which can be used to answer Q [GV14].*

A model should help to understand the function of a system, whereby the complexity of the real-life system is reduced. A model helps to answer questions and to solve problems that exist in the system. With a model, it is possible to perform simulations of the modeled system. The best model is the easiest model that can solve the task or problem. This approach helps to reduce the complexity in a system and can help to understand the important mechanisms of a system [GV14].

Definition 4.1.5 (Model building) *The process of the implementation of a concrete problem or system from applied science into a well-defined mathematical system is called modeling. The aim of modeling is to find a meaningful problem formulation from which conclusions and solutions for the underlying problem can be derived [EGK08].*

Models are developed and improved in a dynamic process, called modeling cycle.

Definition 4.1.6 (Modeling cycle) *A modeling cycle is a possibility to describe the process that is executed during solving a modeling task [Rei16].*

In scientific literature, different types of modeling cycles are described, e.g., in [CAL11] ,[Gei11] or [GE11]. In this thesis, the modeling cycle described by [BL05], as visualized in Figure 4.2, will be further used. The modeling cycle consists of two parts, one belonging to the sphere of

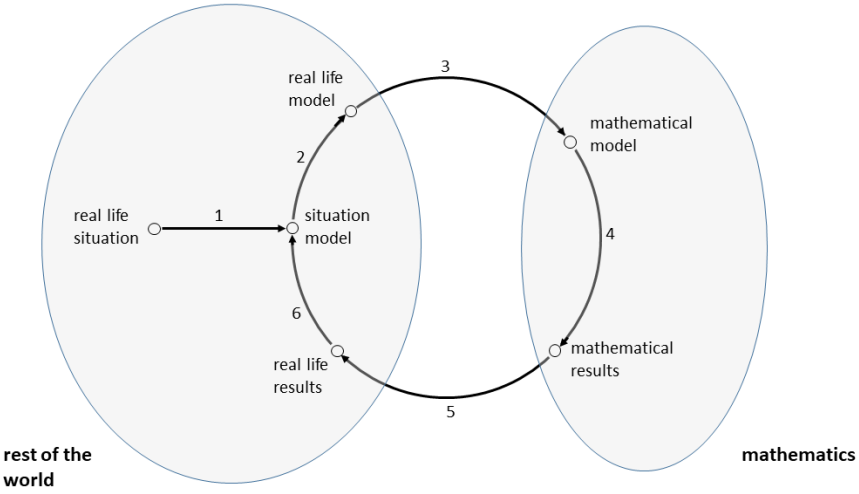

Figure 4.2: Modeling cycle according to [BL05]. 1: Problem understanding , 2: Simplifying/ Structuring, 3: Mathematization, 4: Working mathematically, 5: Interpreting, 6: Validating (figure generated with *LibreOffice Draw*).

mathematics and one belonging to the rest of the world. The rest of the world in this sense means, a situation outside of mathematics, e.g., in applied science, from where the problem origins and for which the model should be build. The two parts, mathematics and rest of the world, are connected in two directions. One connection describes the mathematization of a real world problem, whereby mathematical tools are used. The second connection between the rest of the world part and the mathematical part describes how results from the mathematical model are used, interpreted, and validated with results from the real world [BL05].

The modeling cycle starts with a real-life situation for which the modeler wants to build the model. The modeler must understand the problem and analyze the relevant factors to describe the real-life situation. First, a situation model is built that is only valid for a concrete situation. By structuring and simplifying, a real-life model is developed. In the next step, the real-life model is mathematized and translated into a mathematical model with which a modeler can work and simulate the system, and through which, mathematical results are gained. The mathematical results are then interpreted and compared and validated with real-life results. If the deviations between the mathematical results are too large as derived from

the model and real-life results, the modeling cycle starts again with the aim of improving the model. The modeling cycle is run through until the model has the desired characteristics. The modeling and simulation scheme proposed by [GV14] goes a little bit more into detail about how an initial model can be developed before it is improved during the modeling cycle. According to the modeling and simulation scheme proposed by [GV14], the initial phase of developing a model starts with definition phase in which the problem that must be solved or the question that must be answered must be defined. Additionally, the system or the segment of reality underlying the problem or question must be defined. The second task is system analysis, whereby the parts of the system which are relevant for the modeling approach are identified [GV14]. After that, the modeling cycle with the phases modeling, simulation, and validation starts.

According to [IK13], to develop a model, first the underlying system must be defined by its system variables, and the system boundary and by defining inner and outer relations.

4.2 Chapter conclusion

As described in chapter 2.1.1, this thesis deals with a disease of unknown etiology. The severity of the disease can be measured with medical tools. In the scientific literature, different parameters determining the risk and severity of the disease are described with different hypotheses of the etiology. However, direct causalities between the CKD stage and the mentioned suspected risk factors have not yet been found. CKD is a severe problem in some areas of less-developed countries with high mortality rates.

One of the aims of this thesis is to develop a SDSS for a disease with an unknown etiology. To give spatial decision support, it must be understood what the relevant parameters are that determine the grade and severity of the disease and how these parameters interact. The mathematical methods and algorithms that can be used to find such structures and interactions behind the disease are part of chapter 16.

As mentioned earlier, a model is related to a system. The stage of CKD, for example, expressed with parameters like the GFR, and related risk factors and related risk mitigation strategies can be understood as system variables V_i within a system S_{CKD}. The inner relations of the system must be approximated and described with mathematical methods like mapping or logical expressions. Inside the system boundary are all relevant parameters or system variables located that are necessary to describe the progress of CKD. Since until now

there is no direct causality found between risk factors and the progress of CKD, the system is not understood as a whole. Therefore, the system S_{CKD} is only a preliminary system that changes over time with possible new scientific findings. One way to identify the different version stages of the CKD system would be to label the system with a version number $S_{CKD_{ver}}$
To find mathematical expressions or mathematical models for inner relations in the system S_{CKD} as well as to adapt risk mitigation strategies, the modeling cycle described in the previous chapter can be used. Inner relations in this sense means mathematical expressions describing the relationship between the stage of CKD, possible risk factors, and risk mitigation strategies.
A research and development environment to find and evaluate such inner relations in a user-centered real-life approach is described in the following part.

II | Research and development environment and data sampling

5 A research and development environment in the described framework

5.1 Introduction

As described earlier, this thesis deals with a disease of unknown etiology in less-developed countries with a focus on the suspected risk factor of pesticides. In the previous chapters, the characteristics of pesticide use and challenges with risk assessment and pesticide regulation in less-developed countries were presented. The results show that risk mitigation strategies must be developed and adapted to the characteristics of people living in rural areas in less-developed countries. Therefore, a research and development environment usable in rural areas in less-developed countries in the described framework is required, and a workspace for innovation in the field of agrochemical related risk mitigation strategies and SDSS can be developed and evaluated in a collaborative, multidisciplinary, and open approach. In this environment, experts or stakeholders from different disciplines work together with local community members, with each of them delivering a own set of tools and knowledge related to possible risk mitigation strategies.

The research and development environment should contain the infrastructure for a workspace for an open and multidisciplinary group of experts and community members where innovations in the field of agrochemical-related risk mitigation strategies can take place. For example, it can be equipped with meeting rooms, offices, or workshops where stakeholders can meet or develop technical innovations, collect and identify possible risk mitigation strategies, test the application of the offered risk mitigation strategies, and improve them in a user centered approach.

Additionally, it might serve as an environment with the necessary IT infrastructure to collect, store, deliver, and process data related to risk mitigation strategies. In this framework, it is necessary to collect personal data about users as well as observations about agrochemical

use. With this information, it is intended to use an SDSS and propose the best-fitting risk mitigation strategy to the user, and the user can then give feedback about the applicability of the proposed risk mitigation strategy. This information is used to improve the decision-making process in a learning task. Overall, the environment should contain an IT infrastructure with which it is possible to collect, store, process, and deliver data from and to the user while keeping the last mile gap in mind [Sha06].

In order to increase the efficacy of a risk mitigation regime, it can help to support the decision making of the community members with an SDSS. In the framework of this thesis, an SDSS is developed that learns from the feedback of the users. In most cases, such an SDSS can be regarded as an algorithm and is implemented in an ICT system [Kee03]. Therefore, in the environment itself, computing units and related IT infrastructure, such as a data center, are required.

The environment should also serve as a place for sensors for measuring possible risk parameters and their temporal and spatial extent and to evaluate the success of possible risk mitigation strategies through user feedback or by measuring the physical parameters of the user. In the described framework, sensors are needed to determine the temporal and spatial dimension of risk parameters, to evaluate the success of new or further developed risk mitigation strategies, and to measure physical parameters related to the regarded disease.

Summarizing, a research and development environment is required in which stakeholders and community members can work together and bring their expertise into a research and development cycle, where risk mitigation strategies can be – similar to the described modeling cycle – developed, applied, evaluated and improved in a user-centered real-life approach. The effectiveness and environmental parameters necessary for the selection of the suggested risk mitigation strategies are tracked by sensors. Sensors can be, e.g., a physiological survey, such as blood parameters sampled by medical personnel, answers from the user about how helpful the risk mitigation strategy was, or measurements of environmental parameters like temperature or humidity to select the right time for a pesticide application.

In the following sections, requirements for such a research and development environment are determined, and a resulting environment for the described approach is presented and adapted.

5.2 Requirements for a research and development environment

A research and development environment with the overall aim of developing and validating solutions in order to mitigate the risk caused by agrochemicals in a rural area in a less-developed country is required for the approach described in this thesis. The present chapter deals with the environment itself, possible risk mitigation strategies and their adaptations are described in chapter 9, the development of a SDSS in order to inform the users about the best fitting risk mitigation strategy is described in chapter 18.
Derived from the previous chapters, the research and development environment needed for the described approach must meet the following requirements:

- **(R1.1) Infrastructure for a workplace and space for a research and development environment:** The research and development environment should be constructed in a way that risk mitigation strategies in the field of agrochemicals can be developed and scientifically evaluated that are usable in poor rural areas in less-developed countries. A digital SDSS is proposed to support the user's decision in selecting fitting risk mitigation strategies. They are proposed in relation to personal parameters such as health history and literacy and general parameters such as availability and applicability. In the environment, different experts work together with community members in a real-life environment, for example, in the community itself. Risk mitigation strategies are further developed in a collaboration between experts and community members and tested, evaluated, and adapted in the test environment. The workplace must be equipped with the necessary infrastructure to operate an LL, e.g., meeting rooms, workshops, and IT infrastructure.

- **(R1.2) Multidisciplinary approach with a network of experts:** This thesis deals with a disease of unknown etiology. To investigate the etiology and success of possible risk mitigation strategies and the disease's relations to potential risk factors, experts from different disciplines must work together. At the beginning of the research and development process, experts from different disciplines offer a set of tools for possible risk mitigation strategies from their disciplines in an initial knowledge base (part III). These risk mitigation strategies are then developed further during the research and development process in collaborative work between the experts and community members.

- **(R1.3) Low-cost methods:** CKDu is mostly recognized in rural areas in less-developed

countries. As described in section 2.1.2, financial resources and income are low in such regions [Lal16] . Therefore, the research and development environment itself must be constructed using a low-cost approach, and research about and development of risk mitigation strategies should be low-cost to make them affordable for poor rural community members.

- **(R1.4) Long-term approach:** This work is intended to investigate a chronic disease. Chronic effects are long-term effects [Fen13]. To investigate the success of possible risk mitigation strategies, it is necessary to monitor medical user parameters also for the long term. Therefore, duration of the research and development environment must also be set for the long term. The conditions for long-term collaboration must be manifested, e.g., through contracts or memoranda of understanding.

- **(R1.5) User-driven innovation process:** Users or community members have a good knowledge about which risk mitigation strategies are applicable for them, e.g., because of aesthetics and social restrictions, or they know how an existing risk mitigation strategy must be adapted to be applicable [SQB09]. Therefore, community members function as co-creators, i.e., they are directly involved in the development process of a risk mitigation strategy, and they give feedback about its applicability.

- **(R1.6) Open innovation process:** To increase the benefits, the created innovations and results should be open for everybody to make them also usable in other regions with a structurally equivalent problem. Therefore, open licenses should be used for the developed solutions and research outcomes. By applying this concept, the created solutions can be used for free and adapted to the needs in other regions. Open also means that the group of stakeholders is open for interested stakeholders who can contribute to the project goals and agree to the project agreements.

- **(R1.7) Situated in a real-life environment:** The agricultural production process and the use of agrochemicals affect most people living in rural areas in less-developed countries in their everyday lives and not merely the people involved in the agricultural production process [Bry03]. To test the success and applicability of the SDSS and invented or further developed risk mitigation strategies, research and development should be done under realistic conditions to obtain the best results in a real-life environment.

Besides the mentioned requirements on the used methodology, the following requirements

can be stated to the location or to a pilot region:

- **(R1.8) Political and governmental conditions:** In the proposed research and development environment, data about agrochemical use and about the health status of people living in a less-developed country are sampled and used. Analysis in project use-cases have shown that it is not always politically desirable, especially for foreign stakeholders, to sample data about a possible misuse of agrochemicals and a non-optimally operating healthcare system. However, the will to implement risk mitigation strategies and data from governmental stakeholders, e.g., the location of healthcare facilities, is required. Therefore, it might be necessary to involve political stakeholders in the innovation process.

- **(R1.9) Location in a safe area:** The research and development environment should be set in a real-life environment with realistic conditions in a pilot community. Problems with violent crime can be often observed in rural areas in less-developed countries [GN16]. A community in which a research and development environment can possibly be installed should be situated in an area in which stakeholders can work safely. The community must be also reachable for stakeholders that are not directly located in the rural area of the test community.

- **(R1.10) Community willing to work on the problem's solution:** The research and development environment should be located in a pilot community with an existing awareness of the problem and in which the community members are willing to work in the research and development process with external stakeholders. During the research and development process, community members must deliver personal data in order to improve and evaluate the success of risk mitigation strategies. Because of the use of sensitive personal data, trust between the stakeholders and community members must be created. During the innovation process, community members might have additional work and expense from data sampling and delivery and nonoptimal working prototypes [OLH+13]. Despite these additional expenses, community members should have the will to collaborate in the innovation process, e.g., through highlighting the benefits of the developed innovations [DBM10].

A possible research and development environment meeting the mentioned requirements is described in the following section.

5.3 A Living Lab as research and development environment in the proposed framework

In short, a research and development environment is needed in which solutions and innovations are developed in a real-life context by the involvement of different stakeholders together with community members as co-creators. Stakeholders deliver a knowledge base of tools and expert knowledge into a research and development cycle. A concept for a research and development environment that fits to the most requirements described in the previous chapter is called LL. In the following section, the theoretical background for the concept of an LL is introduced and its limitations in relation to the requirements are discussed.

5.3.1 The theory of a Living Lab

One possible methodology fitting the requirements listed in section 5.2 is called LL. The concept of an LL is based on the theory published by [HSBK10] and [ENK05]. According to them, an LL can be considered as a system and an environment. It is a concept that refers to research and development methods intended for the development and validation of services and products. This is done in cooperation between actors from different disciplines, henceforth called stakeholders, from areas such as industry, politics, academia, or the local population, who live under realistic conditions in the living and working environment [Gee11].

Stakeholders and community members deliver possible solutions to a problem from their discipline; these solutions are analyzed according to their applicability, acceptance, and effectiveness. If possible, the offered tools and solutions can be adapted and improved or new solutions can be developed in a research and development process. In the center of this process are the users – in the described framework local community members – for whom a solution should be developed. Users can decide which solution is applicable and affordable for them. They also have the best knowledge about the concrete problem and situation and can deliver relevant information about how the offered solutions can be improved. Developed solutions also must meet the demands of the users, such as comfort and aesthetics, and they must agree with their lifestyles [SQB09].

An LL is constructed in the environment in which the process or service that shall be improved or developed takes place, e.g., at home, in a community, or in a factory. An LL can be seen both as an environment where research is carried out and a research methodology [Gee11].

According to [ENK05], an LL can be understood as a user-centered method for the acquisi-

tion, prototyping, validation, and development of complex solutions in multiple and evolving real-life contexts. The consumer or user is involved in the development process as a co-creator. The development process in an LL is characterized as an open innovation process whereby development and research are organized in a collaborative way between the stakeholders. Furthermore, the participation of each stakeholder in an LL is voluntary and open for everybody. Therefore, an LL can be regarded as an open innovation network [BPD05]. Despite the open and voluntary character of an LL, one or more stakeholders can take on leadership positions in the innovation process, which is due to by a higher work intensity of a stakeholder. Despite the more active role of one or more stakeholders, all stakeholders have the same power and rights [LWN12].

The development process in an LL can be described in a so-called innovation cycle, which is visualized in figure 5.1.

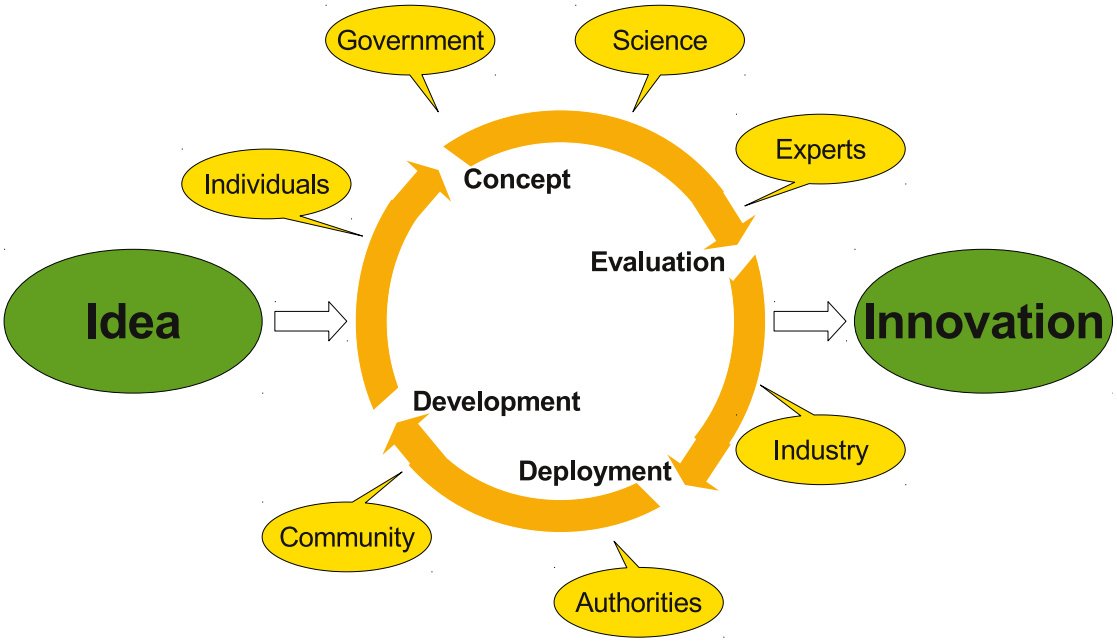

Figure 5.1: An example of an innovation cycle within an LL, according to Marlien Herselman, unpublished (figure generated with *LibreOffice Draw*).

The innovation cycle is a process in which possible innovations and tools are collected and evaluated according to the characteristics needed for the problem-solving strategy, e.g.,

applicability, financial effort, or effectiveness. In the next step, the strategies can be applied in a test environment and evaluated. After the evaluation, it is possible to discuss further improvements. In the following step, the improved methods are evaluated, and the cycle continues. The whole process takes part in the stakeholder group together with the community members.

Within the LL approach and the innovation process, sensing of parameters, e.g., change of human behavior or resource use or feedback from stakeholders, take a key role. Smart technologies, e.g., smartphones, can be used to quantify and to sense these issues to sample data and to validate the success of the task developed in the LL. The interaction between technology, change in human behavior and the environment, and its feedback loops can be studied under realistic conditions. Therefore, it is an innovation and development environment that is usable for investigating and developing methods for a more sustainable living, among others [vTK17]. According to [vTK17], smart techniques using augmented reality (AR), cloud-based services, and the network of connected physical devices have the highest potential to make living more sustainable. Besides data sampling, the use of ICT in the LL approach also has additional functions. It helps to organize and connect the different stakeholders and share common ideas.

[LWN12] mention four types of LLs that are differentiated by the most active role of a stakeholder in an LL: utilizer-driven, enabler-driven, provider-driven and user-driven LL. Each type has different characteristics related to organization, coordination, and structure. In a utilizer-driven LL, companies try to improve and develop products for the commercial business of a company; an enabler-driven LL is mostly organized by public-sector actors with the aim of finding solutions for societal needs. They are used to find solutions for social and structural problems in a specific region or urban area. They are often managed and structured by educational institutions and have a low influence of commercial companies, so that the development process takes place close to the user. The desired outcomes are solutions to improve everyday life for the community members as well as increasing cooperation between the stakeholders. A provider-driven LL is initiated by educational institutions with the goal of improving research methods and theories. A user-driven LL is initiated by the users or a user community with the goal of finding solutions for a problem in the users' everyday lives and less for the product development for a company. Such LLs are organized by the bottom-up principle in which leadership by one stakeholder is not obvious and the organization is informal. In this type of LL, the stakeholders deliver tools for the users to solve the existing

problem. Such tools can be financial resources, knowledge, or simply advice. Information about the users, their behavior, and everyday life are collected and used in the development process [LWN12].

In general, sensed and sampled data in the LL approach can be used to gain insights into the living patterns of the community members and validate the developed products to use them to make fast decisions and inform community members [vTK17]. According to [BKHS09], the success of an LL is dependent on five key elements: continuity, openness, realism, empowerment of users, and spontaneity.

Continuity in the work in the LL is relevant, as trust must be built between the external stakeholders and the LL inhabitants or community members. Community members give information about their daily lives, health, and life histories to the pool of stakeholders. This data contains sensitive and intimate data. To deliver it to a person outside of the circle of friends or well-known people, trust is needed. This trust can only be built by continuous work over time.

Openness means that the innovation cycle should be open for people from different disciplines to watch the process from different perspectives to develop solutions acceptable from different points of view. It is also important because people working in this field but not originally part of the project can also report their experiences to the LL.

Realism means that the work in the LL should be under realistic conditions to find real-life solutions suitable for everyday life.

Another relevant point is the empowerment of users. They must bring their experiences into the innovation process and must decide which solutions are suitable for them in their daily work. The involvement of users in the development process is a point in which the LL concept differs from other common concepts.

In the framework of an LL, spontaneity means that spontaneous user reactions are detected in order to involve the wishes and desires of the users, societal and social needs into the development process.

There are different motivations for participating in an LL. Using new technologies and changing behavior in the development process can be regarded as costs for the users or community members [OLH+13]. To accept these costs and inherent additional work and expenses, a benefit must be obvious to hold motivation in participation high. The highest motivation for participation can be found in projects where a win-win situation for both the stakeholders and the community members can be achieved. Such benefits can be, for example, gaining data

for scientists, improving a method or a product for industrial stakeholders, and improving a specific problem for political institutions and authorities. Some obvious benefits for the community members might be a possible improvement of the problem and monetary savings through an improved process or product [DBM10].

Several LL have been established. Research about and with LL started in the 1990s, with the first starting in an academic project by students to investigate problems in a real-life situation. Following this, the concept LL was often used to investigate smart homes. Currently, LLs are used in different tasks to prototype, validate, and refine developed solutions [LWN12].

In 2006, the *European Network of Living Labs (ENoLL)* was founded and now consists of over 300 accepted LL. *ENoLL* describes the concept of LL as a realistic test and experimentation environment where users and producers are co-developing innovations [oLLE24].

LLs in Southern Africa were combined to a network called *Living Labs in Southern Africa (LLISA)*. Part of the network *LLISA* is Siyakhula LL, located in the Eastern Cape, which can serve as an example of successful, established LL that operates in research, development, and training in the field of electronic telecommunications for rural communities [GTT+12].

LLs are designed for the investigation of information and communication technologies and their application in rural areas of less-developed countries is emerging and established [Her11]. Although LLs are currently widely used, a transfer of the LL concept to decision support with the aim of minimizing risks and where ecotoxicology, medicine, and agricultural production process are interconnected has not yet been carried out. LLs related to the agricultural production process exist [Haa10, HK+08]. However, the aim of these LLs is to develop new agricultural methods for industrialized agriculture but not with the goal of developing risk mitigation strategies related to the use of agrochemicals, as described in this thesis.

5.3.2 A Living Lab in the proposed framework and adaptations

The introduced concept LL meets large parts of the requirements listed in section 5.2. Requirements **(R1.2) multidisciplinary approach with a network of experts**, **(R1.7) situated in a real-life environment**, **(R1.5) user-driven innovation process**, and **(R1.6) open innovation process** are core parts of the LL concept [HSBK10, Gee11]. Therefore, the concept LL is used as a foundation for a research and development environment in the following scientific analysis for the described framework. The implementation and necessary adaptations to the case of a disease with the focus on the suspected risk factor pesticides are

discussed in the following sections.

5.3.2.1 Open community approach and open source

In the framework proposed in this thesis, risk mitigation strategies and a related research and development environment should be developed that are usable in rural areas in less-developed countries. In chapter 2.1.2, the socioeconomic situation in less-developed countries with limited financial resources were described. The used and developed solutions should meet requirement **(R1.3)** and should be free of charge to be affordable for people with limited financial resources. In the case of ICT solutions, there should be no license costs. Therefore, the used and developed ICT solutions are developed as open-source solutions to be free of license costs. This implies also the right to modify, redistribute, and use developed products like maps and IT solutions in other similar projects.

The research and development environment is accompanied by scientific research and scientific publications. To gain the greatest benefit, the outcomes of the scientific research should be available for other communities dealing with the same or a similar problem. Therefore, publications about the research and development process should be in the field of open access journals.

The research and development environment should work with a user-centered and open innovation approach **(R1.6)**. An open community is used because of two reasons: First, everybody interested in the development of risk mitigation strategies should have the ability to contribute in the open community to bring their experience into the development process and adapt to other application scenarios with different requirements and constraints. On the other side, an open community is also required in which community members work as co-creators together with a network of professional stakeholders (requirements **(R1.6)** and **(R1.2)**).

In the following section, the theoretical background of open source and the open community approach are highlighted.

One option to fulfill requirement **(R1.2)** with regard to the use and development of IT components and software can be the use and development of open-source software.

The *Open Source Initiative* is a public benefit and non-profit corporation and was funded as a result of a strategy meeting about open source in 1998. It provides a broadly accepted definition about how open-source software is defined and promotes the use of open-source software, supports in the building of open-source communities, and educates about open-source software. It also approves and reviews licenses if they adhere to the open-source definition

[Ini13].

According to the definition of the *Open Source Initiative* [Ini14], software is labeled as open source if the following conditions are fulfilled:

- There must be "Free Redistribution", meaning it can be freely shared.
- The source code of the software is available.
- The license must allow modifications of the software.
- "Integrity of The Author's Source Code".
- There can be no discrimination against people, groups, or specific working fields.
- The licenses attached to the software apply for every person using the software.
- The license must not be specific to a product, e.g., independent from a software distribution.
- There can be no restrictions on other software by the license.
- No provision of the license may be dependent on a specific technology or a specific type of interface.

The most important and widely used licenses approved by the *Open Source Initiative* are for example the *Apache License 2.0*, the *GNU General Public License* (GNU GPL) or the *Mozilla Public License 2.0*. Most of the software used for and in this thesis are under the GNU GPL. In section 17.4, possible open-source software solutions are listed that can be used for the operation of an LL in the described approach.

Both ethical and practical issues are mentioned as reasons for people to use and develop open-source software. There are many advantages of open-source software over free or commercial software [GBD00]. For the described approach, open-source software is used for the following reasons:

The software can be used for free. This helps to minimize operating costs for the LL and related software necessary for a SDSS or for risk mitigation methods. Also, software tools used in a crowdsourcing approach as described in chapter 6 can be used for free by the community members.

The software tools developed for risk mitigation strategies can be distributed for free and can

be used by other people who work on similar projects, e.g., on CKDu in another country. The available source code can be accessed and modified, which helps to save development costs.
Additionally, there is a large community working in the field of open-source software. For example, help and software support is often conducted via the internet and the open-source community itself. Costs for commercial support can therefore be saved.
Through the use of open-source software it is possible to let scientists and other people working in the field of risk maps or risk mitigation strategies participate in further development of the generated software tools. They can modify the software to their needs.
Overall, the use and creation of open-source software allows a greater number of people to participate in the risk mitigation approach described in this thesis.
The openness of the open-source concept can also be related to communities.
With the development of the internet, a new collaborative working kind was created: the virtual community. A virtual community consists of people connected and interacting via the internet. They are connected for social relationships, like in social media platforms, but also for information exchange [BD02].
The open community concept is more broadly open than a virtual community. The participants must not necessarily be connected via the internet. An open community is defined as follows:

Definition 5.3.1 (Open Community) *An "Open Community is an generalization of the concept of Open Source to other collaborative effort. The term "open" for an open community refers to the opportunity for anyone to join and contribute to the collaborative effort. The direction and goals are determined collaboratively by all members of the community. The resulting work is made available under a free license, so that other communities can adapt and build on them" [Ini15].*

The aim of an open community is to create a space for user-driven services and open innovations as well as to improve processes in terms of knowledge and information sharing. The origin of open community lies in the LL concept [WHL08].
Because of the following reasons, the open community approach is proposed in this thesis. Most of the problems in the topic of agrochemical-related risk mitigation are interdisciplinary problems, e.g., when analyzing or developing new possible risk mitigation strategies in the field of agrochemicals people with expert knowledge about biology, chemistry, ecotoxicology, mathematical modeling, meteorology, agriculture, or material science can contribute their expert knowledge. That means that the community must be open for people from different

disciplines.

An open community has further advantages that can be useful in the LL approach. One theory holds that when a lot of people with different points of view work together on a project, the results of the resulting collaborative work are better than if the same number of people work on the same topic independent of each other [Sur05]. This effect is called social intelligence and is described in more detail in section 6.2.

The open community approach is additionally used because of its openness to people who were originally not intended to participate in such a project. It should be also open for people who are not organized in a typical stakeholder group, e.g., volunteers not living in the LL community but with experience in the field of work. In a closed group of project members, it would not be possible to use these experiences.

The described determination of goals and objectives in a collaborative process are also important. When thinking about traditional projects, goals and objectives are defined by project leaders or a core group of project members. The open community approach allows everybody involved in the project to help to adjust the goals and objectives. Often also knowledge about traditional behavior and feelings of the community members are lacking. Through the collaborative effort, it is possible to introduce these experiences from different disciplines and different social groups into the process of determining goals and objective. Therefore, the goals and objectives are broader and more adapted to the different fields and the needs of the community members.

A major advantage and necessary for the sustainability of the project is the release of the developed products, e.g., maps, methods, or software tools, under a free license. The created maps and software solutions can be used for free. This is especially important in less-developed countries where a lot of people cannot afford money for maps and software.

The release of the product under a free license is also important for the sustainability of the risk mitigation strategies. The created software tools and maps can be used, modified, and extended by everybody to accomplish the collaboratively defined goals. This can help if, e.g., a funded project stops or if a project group responsible for the creation of the tools stops working. In these cases, it is possible that the tools and maps are updated and further developed by people not involved in the project.

The present section has shown that through the use of open-source software and the open community approach a benefit for people living in poor rural areas can be increased in the described context.

5.3.2.2 Selection of stakeholders

As mentioned in section 5.3.1, several stakeholders are involved in an LL. Despite the used open community concept at the beginning of an LL project, a core group of relevant stakeholders must be selected according to different criteria and project goals. This group can consist of a group of stakeholders necessary to achieve the described research and development goals of the project and of additional stakeholders in the sense of an open community, for example, volunteer citizen scientists or local communities.

Stakeholders should be interested in the project and want to work on it without a focus on commercial interests. It is aimed to deliver solutions for community members in poor rural areas for whom it is not possible to pay for services. The developed risk mitigation strategies should be low-cost.

Therefore, developed risk mitigation materials, scientific publications, and software tools should be released under an open license. Selected stakeholders must agree with the open community approach.

Stakeholders should deliver options or strategies to implement risk mitigation strategies or parts of them, help to improve them with their interdisciplinary expert knowledge, or help to implement risk mitigation strategies on a political level. As trust plays an important role in an open community approach, trust between the stakeholders may be another criterion in the stakeholder selection process.

In the case of an LL with the aim of reducing the exposure to agrochemicals in a SDSS approach, a core group of stakeholders might, for example, consist of stakeholders from the fields of healthcare authorities, agricultural communities and authorities, health, computer and natural science, industry, community members, healthcare workers, and political actors. They should be selected according to the project goals. Stakeholders may change over time and stakeholder selection is a process that can also be performed during the operating phase of the LL. Local stakeholders can be first implemented when the pilot region or area for a possible LL is determined. Stakeholders have the task of contributing their tools to the development and improvement of risk mitigation strategies, setting political conditions for their implementation and contributing their expertise in a multidisciplinary research and development approach.

According to requirement **(R1.4)**, a long-term approach is intended. Long-lasting cooperation can be ensured with contracts and memoranda of understanding through which participation and type of cooperation are described legally.

5.3.2.3 Pilot region

Requirements that are related to the selection of a pilot region for an LL are requirements **(R1.1) infrastructure for a workplace and space for a research and development environment**, **(R1.8) political and governmental conditions**, **(R1.9) located in a safe area**, and **(R1.10) community willing to work on the problem solution**.

(R1.1) dictates an existing minimal infrastructure which is necessary to operate the LL in the community, for example, regional stakeholders, meeting rooms, work spaces, and IT or telecommunication infrastructure to operate a SDSS and related systems and monitor the success of risk mitigation strategies, for example, with physical or environmental parameters. An example for an IT infrastructure to operate an LL in the described framework with a SDSS and related risk mitigation strategies is presented in chapter 17.

In addition, political and governmental conditions in the pilot region must be organized in such a way that there are no political obstacles for the implementation of an LL in the pilot community. Local political decision makers should support the concept **(R1.8)**.

(R1.9) is necessary for the assumption of a safe operation of the LL, in which, for example, foreign stakeholders can stay, move, and work without being exposed to criminal attacks and working equipment can be safely stored in the community. Criminal activity can be a problem for LL in poor rural areas.

A region in which community members are interested in participating in the risk mitigation approach should be selected as a pilot region. Community members are co-creators in the research and development cycle, give feedback about the proposed risk mitigation strategies, and deliver personal and crowd-sourced data **(R1.10)**.

5.3.2.4 Motivation for community members to participate

As described in section 5.3.1, the best working LL can be found where a win-win situation can be achieved for the stakeholders and users or community members. The benefits for stakeholders are obvious. Scientists, for example, gain new data collected in a real-life environment; industrial stakeholders can use innovations from the LL approach for new commercial development; agencies and authorities gain knowledge about possible risk mitigation strategies and their efficacy.

The greatest benefit for community members in an area with agrochemical-related diseases

might be that participants can help to improve the health conditions by not loosing economical disadvantages for themselves, friends, and relatives. Beside this non-monetary aspect, there are also monetary aspects. In the developing world, subsistence and small-scale agriculture is more widespread than in the developed world. [WZ11] describes that relatives and whole families work together to grow crops for their daily food or for sale. People with a disease are not as productive as healthy people because their manpower is missing in the agricultural production process and, therefore, additional workers must be paid.

An additional monetary benefit might be that expenses for medical goods and medical services can be decreased by a lower number of cases and strength of the diseases through the use of the proposed risk mitigation strategies. Money can also be saved by an adapted and reduced pesticide use. This can be achieved through different risk mitigation strategies, e.g., precision farming and educational programs, for example, to avoid overdosing by correct application rates.

These benefits are sometimes not directly obvious for all participants and should be communicated to the stakeholders to keep motivation high. Besides the described benefits of LL, inhabitants might perhaps, as described in Chapter 6, have fun being part of a scientific community, working with modern electrical devices, or helping to improve the SDSS and other risk mitigation tools.

5.4 Establishing a Living Lab

The process of establishing an LL in the described framework can start with a core group of stakeholders, as described in section 5.3.2.2. This core group can later be refined and opened to other stakeholders according to the used open community approach (section 5.3.2.1). To ensure long-lasting cooperation, contracts or memoranda of understanding can be established in which the stakeholders' type of participation and responsibilities are described.

The stakeholder group must define general project aims and objectives in an open community approach. This should include the theoretical implications of the LL approach, like the type of user involvement or legal issues as described in [Her17]. Additionally, the group of stakeholders must set the framework for the project.

A pilot region meeting the mentioned requirements must be selected (section 5.3.2.3); contacts with possible local stakeholders or associations can be closed, possible cooperation explored, and the group of stakeholders refined. In the pilot region existing work space, IT and mon-

itoring infrastructure must be identified and related to methods used in the selected risk mitigation approach adapted and extended. An example for an IT infrastructure usable in an LL approach with a digital SDSS is described in chapter 17.

Stakeholders deliver their initial collection of possible risk mitigation strategies to the innovation cycle. This collection will be named as toolbox for risk mitigation. In relation to regional available risk mitigation methods and materials, an initial set of tools of possible risk mitigation strategies, which can be implemented and further developed, must be defined by the stakeholders and community members. Examples for possible risk mitigation strategies in the field of ecotoxicological and environmental modeling can be found in part III.

During the operating phase, risk mitigation strategies are developed and refined in a user-centered open innovation approach with a research and development cycle. Scientific research, software, and created risk mitigation materials are, according to the open community approach (section 5.3.2.1), released under an open license.

Methodology LL with a development cycle can also be used to optimize the operation and structure of the LL itself.

5.5 Chapter conclusion

In the present chapter, a framework for the adaptation of the LL approach as a research and development environment related to a disease of unknown etiology in less-developed countries was developed. A research and development environment adapted to the needs of people living in rural areas and agrochemical related disease was designed.

The concept of an open community is proposed for the operation of the LL in which different stakeholders work together with community members on solutions in an open innovation approach to find risk mitigation solutions in the field of CKD. Different adaptations were necessary to let the concept LL fit to the case of CKD. The LL concept was combined with a open community approach. Open in this sense means the stakeholder group is both open for stakeholders from different disciplines and different professional levels to assure a multidisciplinary research view from different angles. Additionally, the term open is also related to the license properties of the developed risk mitigation strategies and products. They are released under a open license to make them affordable for people living in poor rural areas.

Additional requirements that are specific to the case of CKDu were worked out, e.g., require-

ments on the pilot region and involved stakeholders.

Summarizing, a research and development environment was developed that can be used to find risk mitigation strategies for an agrochemical-related disease in poor rural areas in developing countries.

6 Data sampling in a Living Lab in less-developed countries

For different steps in the approach used in this thesis, data must be collected from community members living in the LL. This can be personal data about the actual location, life and health history, surveys related to suspected risk factors, the fitness of a risk mitigation strategy, or data sampled in a crowdsourcing or citizen-science approach, e.g., environmental parameters or detection of pesticide application patterns as described in chapter 12.
In the following chapter, the theory about data sampling with a crowdsourcing or citizen-science approach is introduced and their limitations and application in an LL approach in less-developed countries are discussed.

6.1 Crowdsourcing

In the last few years, mass collaboration and the open community approach were widely used to solve problems or find new innovations, e.g., commercial companies used such methods to find new designs or a new product [Bra08a], develop wikis [Gas12], or analyze satellite images with the goal of finding a lost airplane [Hui22].
The term crowdsourcing was first introduced by [How06] and is described as the collaborative work of people connected via the internet and the potential of this principle of operation. Originally, the term crowdsourcing referred to an industrial process, for example, to develop a new design for a product. Crowdsourcing is a neologism that is formed by the words outsourcing and crowd. The meaning is that some tasks of the industrial production process are outsourced to the crowd [How06].
There are several definitions for the term crowdsourcing.

Definition 6.1.1 (Crowdsourcing) *Crowdsourcing is defined as an interactive form of ser-*

vice delivery that is organized collaborative or competitive and that includes a large number of extrinsically or intrinsically motivated actors with a different level of knowledge which use modern information and communication systems based on the Web 2.0 [How06].

Performance objects are products or services with different degrees of innovation that are developed reactively by a network of participants due to external stimuli or proactively by automatically identified needs, gaps, or opportunities [MLV08]. Other authors have a wider definition. [DRH11] define a crowdsourcing system "if it enlists a crowd of humans to help solve a problem defined by the system owners" [DRH11, p. 87].
[Bra08a] argues that there is a difference between the concepts crowdsourcing and open source. Solutions and ideas developed by crowdsourcing are always adopted and later owned by companies, but open-source solutions are always, within the meaning of the definition of the open-source community [Ini14], for free. Crowdsourcing solutions are mostly used for non-software products, e.g., for the creation of a new design for clothes. To use this solution, it is necessary that clothes are produced, imported, and printed; production costs must be covered, and there must be company sponsorship. [KVR08] understand crowdsourcing as a process where commercial firms outsource specific tasks to the public with the intention of using the result for a commercial product. The participants work for free or for much less money than professionals. Companies generate money with the developed solutions. The consumer acts as a coworker under the responsibility of a commercial company [KVR08]. Separation between the societal spheres of production and consumption is increasingly dissolved. The buying consumer gets transformed into the working consumer [Pap09].
The motivation of internet users to participate in a crowdsourcing project is less a monetary aspect and more the idea of participating in a project to have fun or be accepted in communities [LP07] or the improvement of the own skills [Bra08b].

6.2 Citizen science

Another method referred to as crowdsourcing but more suitable for the described approach is called citizen science.

Definition 6.2.1 (Citizen science) *The citizen-science model is defined as a model that "engages a dispersed network of volunteers to assist in professional research using methodologies that have been developed by or in collaboration with professional researchers. The public plays a role in data collection (...) across broad geographic regions (and often, over long*

periods of time), usually to address questions raised by researchers" [CDPB07, p. 2].

It is a commonly used method especially in environmental science and ecology [Sil09], e.g., to determine the temporal and spatial distribution of organisms [BLB05] or in water quality monitoring [CKR+20]. The concept integrates the engagement of people with scientific tasks and involves them in the scientific working process [Sil09]. Participants are engaged in the scientific process and create or deliver data; participation is active and goes beyond recognizing participants as research subjects or as supplier of computing power [WC11]. In contrast to the earlier mentioned method crowdsourcing, citizen science is not intended to gain commercial success.

There are several different citizen-science methods differing within the kind of the participation of the volunteers, e.g., if they are involved in the scientific study design process or in the analysis or interpretation of the data [WC11]. [BBJ+09] propose a classification within the citizen-science projects in contributory, collaborative, and co-created projects. In contributory projects, volunteer participants are only involved in the data collection process, for which, in collaborative projects, the study design is created by professional scientists, but the volunteers can be involved in design refinement or in data analysis. In co-created projects, the study is designed by both professional scientists and public participants. In addition, the public is involved in all steps of the study from design to publication [BBJ+09].

It is not necessary that citizen-science projects are only mediated by information and telecommunications technology. If they are, then citizen science can be seen as a part of crowdsourcing, adapted, and used in science. In contrast to open-source projects, citizen-science projects have a hierarchical structure, with professional scientists giving the framework and scientific methods. There is also a difference to open science because it is mostly not intended to publish the entire scientific process [WC11]. Furthermore, citizen-science projects are often not self-organized [CLW+07].

The voluntary participants are called citizen scientists. Their role in scientific studies can be described as unpaid field assistants. Mostly, they do not have a scientific education in the field in which the study occurs [Coh08]. The citizen scientists can help to deal with a great amount of data [Gal14]. Citizen science can be a useful tool for collecting large datasets with a temporal and spatial distribution [DWB10].

Collaboration between professional and citizen scientists results from the citizen scientists interest in the field of the study, being outdoors, and the problem to which they can help to find a solution. Analyzing the collected data and writing a paper about the findings is still

part of the work of the trained scientists who initiated the scientific research [Coh08].

In 2007, a citizen-science project called *Galaxy Zoo* was started [Gal14]. A website was released on which images of the galaxy from the *Sloan Digital Sky Survey* were published. Volunteers had the task of categorizing the pictured galaxies according to their morphology. Within the first year of the project more than 50 million galaxies were classified by the citizen scientists. The project is an example of a co-production of amateur astronomers as citizen scientists and professional science [MSS11]. It is now integrated in the project *Zooniverse* [Zoo14], an internet platform on which several different citizen-science projects in the field of biology, climate science, and astronomy are presented. Interested people get also information about how to participate in the different projects.

For the citizen-science project *Mückenatlas* [MÏ4], people can send trapped mosquitoes and a notice about the trap location to a zoological lab where the samples are categorized. Together with the coordinates of the trap location, distribution maps can be created.

The open participation approach is also used for emergency tasks. For example, volunteer participants analyzed satellite images after the Haitian earthquake in 2010 and helped to generate actual maps for the web-based mapping services, e.g., for aid agencies [ZGSG10]. Citizen science is also used to conduct clinical research studies, e.g., the participants can answer questionnaires, self-report data, or self-track device data [Swa12].

[CC12] report about a sea turtle monitoring project initiated by the state of North Carolina, USA, conducted with volunteer participants as field workers. They found that through the work in the project and the gained knowledge, the citizen scientists are able to participate in scientific debates about the topic and are able to argue for their opinions. Additionally, the volunteer participants get involved in scientific processes as well as in the decision-making processes of authorities and can gain understanding for the different opinions of the stakeholders. Authorities and the scientific community benefit from the citizen's work, thus a win-win situation arises [CC12].

The method called citizen science has the advantage that datasets can be created without a restriction of the spatial scale. Biases and uncertainties can occur and must be kept in mind when using this method [Sil09].

It is discussed that the reliability of the data is not as good as data collected by trained scientists. In some cases, the volunteer participants must be trained in how to collect the data. Another benefit of using this method is that the citizen scientists are educated in the field in which they work and become aware of the problem [Coh08]. Another advantage of

the citizen-science method is that a lot of studies suffer from a lack of funds with which to pay fieldworkers, especially when they need a special training. There, it is wiser to collect data with lower reliability or minor mistakes. Also, with volunteers, it is possible to enlarge a study to a longer time span and a larger spatial extent [Coh08].

As previously described, the citizen scientists are mostly not experts. Thus, the tasks must be chosen in a way that the citizen scientists can execute them with basic knowledge and perhaps a short introduction. However, the collected data must be checked for mistakes and reliability by experts [Coh08]. For example, in a study with a citizen-science approach in which citizen scientists should answer questions about the spatial distance of their living homes to agricultural fields treated with pesticides, errors were observed in the self-reported items' distance and the application practices. The authors attribute this to a lack of ability to estimate distances and experience in pesticide application techniques [RRS06].

One reason people participate is that they are proud to help in scientific tasks and they can benefit nature [Coh08]. In [MSS11], the motivation of amateur astronomers to participate in the citizen-science project *Galaxy Zoo* [Gal14] was evaluated. The participants were attracted for aesthetic reasons, community building and cooperation with professional scientists, and the interest in and wish to participate in professional science [MSS11]. Participants received scientific knowledge and it animated them to build interest in science as well as in nature and explore it [BLSF09].

Participation in the scientific working process can have a positive impact on the literacy of the citizen scientists. The participants also get knowledge and awareness of the topic of interest, and the motivation to participate in other public problem related initiatives is increased. Additionally, the scientific process skills of the participants can be improved [CJH+13]. Participants are also motivated to learn new facts about the task, and they have fun in the work [RBG+10]. [TBBC00] reported that participation in a citizen-science project leads to the start of scientific thinking processes. The participation of the public in citizen-science projects can lead to the understanding and acceptance of scientific processes and can be further seen as a kind of public science education [BLB05].

6.3 ICT in crowdsourcing and citizen-science projects

Citizen-science projects existed before the development of the internet and modern ICT systems, e.g., collection of data about birds or astronomical and meteorological observations

[Dro07].

Through the introduction of modern ICT and internet technologies, it is easily possible to establish public databases where citizen scientists can inform themselves about ongoing projects and scientists who want to start a citizen-science project can get information about how to start such a project. Additionally, the connection and coordination between professional and citizen scientists gets much easier [NWC+12].

Internet technology provides conditions for using this collective intelligence. People from all over the world can communicate in an open single system. People with different cultural backgrounds and skills can give their opinions about how to solve a problem. The internet also favors a kind of work in which the user is involved in the working process and where they can act or create and innovate new things [Bra08a].

With the so called *Web 2.0*, it is easy to communicate in both directions – from and to the participants, which makes it much easier for the volunteers to take part in the project and to use this method [KVR08]. Through the communication in both directions, not only researcher-driven, top-down research questions can be formulated but also bottom-up questions, formulated by the citizen scientists, can be integrated into the research program [DBB+09].

The widespread use of the internet and development of GIS-based web applications makes it easy for the participants to deliver the collected data to a central database. Smartphones with their integrated cameras and sensors are useful tools to sample data easily and cheaply [DSB+12]. New communication technologies like mobile internet and mobile devices have the potential and are the cornerstone for the broad and cheap use of citizen-science methods. They can be used in a wireless sensor network in which the mobile devices can be regarded as mobile sensors, which can determine different environmental parameters, or it is possible to transmit recognized phenomena to a central server and connected database. With the help of mobile devices, a lot of people can be reached; through their use, people can be motivated to participate in public volunteer work. Data collection can be broadened, the accuracy of models can be tested, and decision-making can be accelerated. Through the use of modern ICT infrastructure and inherent techniques, such as sensors and mobile apps, the amount and quality of the sampled data will increase [NWC+12]. This trend will go on because the upcoming ICT technology shows that, for example, normal mobile phones that were originally intended for communication are transformed into mobile personal instruments with which it is possible to be connected to a network and measure different parameters [PHH08].

Mobile apps can help to increase the quality of the collected data. [GHS11] reported from a

research project in which citizen scientists were to identify plants. Through the use of a mobile app for smartphones in connection with the integrated Global Positioning System (GPS) sensor, it was possible to exclude identifications that were not possible because of the user's location and the spatial distribution of the plant.

[KVSD10] showed that it is possible to add different plugins on a common smartphone with which the tracking of more challenging issues is possible, e.g., soil moisture or the electrocardiogram of humans.

Wireless data sensing can be regarded as a connection between the laboratory and nature [NWC+12]. Through the use of modern and mobile ICT, citizen scientists are integrated in data collection in a way that they can collect data on their everyday activities without any special effort [GHS11].

6.4 Data sampling and delivering in less-developed countries

In the LL approach described in this thesis, it is intended to sample data for different tasks in citizen-science projects. There are different methods to sample the needed data, e.g., through a paper-based survey, interview, or an app for mobile devices.

Data sampling and delivery differs between developing and developed countries. In developed countries, most people have access to mobile devices and are connected to the internet; thus, collecting data is much easier than in less-developed countries, where, in rural areas without the ability to use mobile devices, mostly paper surveys must be conducted. Such paper-based surveys have the disadvantage that there is a long delay between data collection and processing. Furthermore, a lot of human resources are needed to collect the data, and errors in data transcription can happen [AHB+09].

Although in less-developed countries the development of the ICT infrastructure lags behind developed countries, an exponential growth in the build up of infrastructure and the use of mobile phones can be recognized [AHB+09] – as well as an increasing use of the internet and computers [CF10]. There is not only a gap in the use of computers and the internet between developing and developed countries but also within the group of less-developed countries [CF10].

Some authors found out that there is a strong correlation between the use of ICT and the nominal income of the country and the inhabitants [QATRM03], whereas other studies found a relationship between ICT use with access to telecommunication and the costs for telecom-

munication technology and access [DLW01] and several other factors like education, telephone line density [CF10], or gender and cultural behavior.

In general, in less-developed countries, the era of desktop computers or laptops has jumped, and mobile devices like smartphones or tablet computers are more common [BSD+13]. Despite the increasing use of mobile devices and smartphones, a last-mile problem exists in less-developed countries. The last-mile problem is described as a problem that data delivery to the user is difficult or impossible due to transition problems caused by a lack of infrastructure. The last connection to the user is not established, or there is a bottleneck in the ICT infrastructure, i.e., a wired or wireless connection to the end user is not available or not affordable [Sha06]. The last-mile gap must be first closed before technological solutions in the delivery chain can be implemented.

As the use of mobile devices and access to the internet is increasing rapidly in less-developed countries [O3b14], the following chapters focus on digital data sampling.

In chapter 12, an example is presented of how citizen scientists can contribute with data sampling and delivery through mobile devices. The requirements for such data collection with mobile devices are a broad distribution of mobile devices and that the sampled data can be delivered to a central database. Thus, at least a temporal internet connection must be available to transfer the collected data to the database.

When using mobile devices for data sampling, it must be kept in mind that in some areas the illiteracy rate is really high [Loc12]. Apps for data sampling might be adapted to local characteristics, e.g., with speech recognition software or software reading out the questions or with adapted icons.

6.5 Requirements for an app for data sampling and delivering

As mentioned in the previous chapter, this thesis focuses on data sampling and delivering based on mobile electronic devices, whereby data is sampled by an app or a software tool. However, in section 6.7 an open-source software tool is demonstrated with which it is possible to create paper-based questionnaires with automated report generation. Requirements for an app or a software tool for data sampling in the described framework in less-developed countries are that it must:

- **(R2.1)** be free of charge,
- **(R2.2)** be adaptable to the user's needs, e.g., layout and language,

- **(R2.3)** not be permanently connected to an IT system,
- **(R2.4)** have interfaces for external sensors and software,
- **(R2.5)** allow processing of spatial data, and
- **(R2.6)** provide encryption.

The software should be free of charge with respect to the limited financial resources in less-developed countries, and the background must be usable by a maximum number of people. It should be adaptable to the user's preferences, e.g., with different languages or the possibility to switch between text and visual or audio mode for illiterate people. Due to limited infrastructure resources in the field of ICT, it is important that the tool can also be used in regions in which a permanent internet or phone connection does not exist. Another important requirement might be that the tool offers interfaces to external sensors and software. With external sensors it is possible to measure parameters like humidity or the location with a GPS sensor. Determined location coordinates can be attached to the collected data through which spatial data is obtained. As personal data also is collected in the described approach, the encryption of the sampled data is necessary.

6.6 Example for an open-source app for citizen-science projects in less-developed countries

A tool with the ability to build a survey or a data collection form, collect data on a mobile device, and aggregate the collected data is called an *Open Data Kit* (ODK). It is a collection of tools intended for easy use and modification for organizations with limited technical or monetary resources. ODK is developed by a user community for research conducted by the *Department of Computer Science and Engineering* of the *University of Washington* [Kit14]. In the past, it was used by different organizations in less-developed countries, e.g., for mHealth systems [RMA+12], decision support systems [ARP+12], or for general data sampling [Ful10b]. It runs on *Android* devices. It consists of different modular software tools that can be combined or run alone and are named after their intended use, e.g., *Build*, *Collect*, and *Aggregate*, to create forms, collect data, and aggregate the data in a way in which it can be analyzed. It is an open-source software tool; the source code can be used by an open community, and new modules can be created. It uses a standard text-based layout. As mentioned in section

2.1.2, in some less-developed countries, the illiteracy rate is high and text-based layout is not appropriate. Through its status as an open-source tool, it can be easily modified, e.g., to implement text-to-speech or text recognition modules. It might be possible to adapt the graphical user interface (GUI) to the user needs, e.g., as described in [Pla14], or to adapt it to other devices and operating systems.

ODK uses or generates files in a standard format so that the interface to other tools is ensured. The software is adapted to the situation in less-developed countries, e.g., asynchronous data transfer and several different ways to transfer data to the user are implemented. In addition, the user interface is structured in a user-friendly way, and many different data types can be implemented, e.g., pictures or coordinates. It is also possible to use integrated sensors with the package *Sensor* for data collection, such as the GPS sensor or a camera to collect pictures, GPS coordinates, or bar codes. It is also possible to collect data sampled by external sensors connected to the device via Universal Serial Bus (USB), Bluetooth, or Wireless Fidelity (WiFi). It is easy to create a questionnaire – even for people who are less-skilled in ICT and programming, e.g., questions can be created in a spreadsheet document and compiled to a questionnaire that can be used by ODK [BSD+13].

Through user and developer feedback, a redesign of the software structure was conducted, leading to the *ODK 2.0*. With *ODK 2.0*, it is possible to integrate a present survey or data into a previously collected database or to modify a present database directly from the mobile device. In addition, machine-readable paper surveys can be scanned in and processed by ODK. It is also possible to collect data without a mobile internet connection by Short Message Service (SMS). The data collected in a cloud or on a private server can be managed and exported to different formats. In the case of private or sensitive data, it is also possible to encrypt the collected data on the server as well as in the transfer chain. Collected data can be stored temporarily on the device if a mobile connection, e.g., mobile internet connection or connection to the phone network, is not available. With the tool *Submit*, it is possible to transmit data to the server when a connection, e.g., via SMS, WiFi, or General Packet Radio Service (GPRS), is available. In this way, the history of the transmitted data is stored on the mobile device. Thus, only data not transmitted in the past is transmitted, and non-essential traffic can be avoided. [BSD+13].

With respect to Section 6.5, ODK fulfills nearly all requirements listed there: ODK is an open-source tool and allows the sampling of data cheaply and easily **(R2.1)**. The software itself allows the implementation of different media formats, e.g., videos, images, or sound files.

Therefore, the created questionnaires can be adapted to the needs of the user, e.g., to use sound files or images for illiterate people **(R2.2)**. Where the infrastructural situation allows the use of a mobile device, the questionnaire can be used directly on the device itself, with the answers directly typed in the device. On the other side, where the infrastructure does not allow the use of a mobile device, e.g., because of a lack of connection to the power supply system or through the lack of a mobile device in a certain area, it is also possible to conduct the survey in a paper form, scan it in, and process it automatically **(R2.2)**. It is also possible to store data directly on the mobile device and transmit the data later when a connection to the server is established. It is also possible to transmit data without a internet connection via SMS to the server. These are important characteristics of ODK for use in less-developed countries that allow it to bridge the last-mile gap and overcome the lack of mobile internet connection **(R2.3)**. Different software and sensor interfaces are also implemented in ODK. Parameters measured with internal or external sensors can be integrated in questionnaires, for example, images, sound files, and coordinates **(R2.4)**. ODK has methods implemented that allow encryption of stored and transferred data **(R2.6)**. This is required when working with sensitive data, such as personal health data.

The processing of spatial data is not fully supported by ODK **(R2.5)**. It is possible to determine spatial coordinates with the help of the GPS sensor, but a webgis function is missing, e.g., for marking spatial areas to create geodata. As ODK is an open-source software, the source code can be adapted and possible interfaces to GIS software can be developed.

6.7 SDAPS as a sensor for complex and dynamic systems

Machine readable paper-based surveys are an option for data sampling in regions where digital devices or electrical power are not available, as in some regions in developing countries. An open-source software to create paper-based questionnaires with optical mark recognition of answered questions and automated report creation is called Scripts for data acquisition with paper-based surveys (SDAPS) [Ber14].

To be usable in SDAPS, the created questionnaire must have a special predefined format from the software. SDAPS is able to locate the position of questions and check-boxes on the paper. It is possible to create a SDAPS readable questionnaire with different software tools like *Libre Office* [Fou20e] or in the format of a *tex* file.

After the creation of the questionnaire, it is analyzed by SDAPS, and SDAPS recognizes the

location of the questions and the check-boxes and fields for free text on the paper. Following this, the questionnaire can be printed out and answered by the interviewees. After this step, the answered questionnaires are scanned in as a file in *tif* format. In the next step, SDAPS analyzes automatically which check-boxes of the scanned questionnaires are marked. To check if SDAPS has recognized the marked boxes correctly, a GUI for manual mark recognition is implemented in SDAPS. With this GUI, it is possible to correct check-boxes that were recognized as marked but that are not. SDAPS has the possibility to create a report or a spreadsheet with a basic statistical analysis about the answered questions.

Through the use of a optical mark recognition software, time can be saved that would be necessary for the manual digitization of the answered questionnaires and creation of a basic statistical report.

The use of SDAPS also has some difficulties. A detailed manual and description of the software is not available; therefore, time for learning how to use the software is required.

The following problems occurred during installation and the general use of SDAPS. For example, repositories were missing and not automatically installed, the length of the questionnaire is limited to 6 pages, and only uncompressed pictures in *tif* format and scanned in as black and white image with 300 dots per inch (dpi) are accepted by SDAPS.

Overall, despite the mentioned limitations of the software, it represents an alternative to software tools like ODK for which a digital device is required.

6.8 Citizen science in the described approach

Decentralized data sampling by non-professional volunteers like community members can be used in the LL approach for different tasks, for example, to sample data about individual health and pesticide use history or take surveys of suspected risk factors, environmental parameters, or detected pesticide application patterns. An example for an app with which it is possible to report pesticide application events is presented in chapter 12.

According to the classification of [BBJ+09] and the user-centered innovation process in the LL, a collaborative citizen-science project is proposed in which citizen scientists are not only data samplers but also are involved in the scientific study design process. Items or topics that must be tracked in a citizen-science approach can be chosen according to the risk factors in the spotlight and must be defined in the LL process according to the research questions, skills of the citizen scientists, and available sensors. A more mathematical view on the data needed

for an SDSS can be taken from section 18.2.

In this thesis, the citizen-science approach is used for different reasons. It is possible to sample large data sets like pesticide application events that would not be possible without the help of volunteers because of a lack of financial resources for professional workers. Additionally, involved community members or citizen scientists obtain knowledge about the topic, increased or improved scientific thinking, and risk awareness [CJH+13]. It might also be an option to combine data sampling with a risk mitigation framework. For example, a data sampling app or tool can be combined with open educational resources about pesticide related risk mitigation or materials to increase risk literacy. However, the citizen-science tasks in the LL process must be selected according to the available sensors and skills of the community members. Despite the mentioned advantages, data sets sampled in a citizen-science approach somewhat have the problem of low data quality.

Survey methods and software must be selected according to the last-mile gap and existing ICT infrastructure in the pilot region. For example, the tool ODK, which was presented in section 6.6, has many features for digital data sampling and questionnaire creation but can only be used with a digital device. A paper-based tool like SDAPS can be used in regions without internet connection.

6.9 Chapter conclusion

In the present chapter an approach for open and decentralized data sampling in a citizen-science approach was introduced. With the selected approach, it is possible to sample large data sets in a low-cost and open approach. Citizen science is a widely used method to sample large datasets [DSB+12]. Additionally, citizen-science methods help to increase the awareness of the citizen scientists of the examined topic. Adapted to the local characteristics in developing countries, two exemplary open-source tools for data sampling were presented: one for questionnaires and data sampling with a digital device and one for regions with low infrastructural conditions with a paper-based questionnaire with optical mark recognition and automated report creation. With respect to requirement **(R1.3)**, both tools belong to the group of open-source software.

ODK has many more features than the paper-based SDAPS tool, for example, multimedia elements can be used, and it provides an interface for external sensors. However, the tool must be selected according to the infrastructural conditions in the region: if a last-mile gap

exists for a digital questionnaire, then all the features of a software like ODK cannot be used at all. As digital devices get more and more common in less-developed countries, the focus of the thesis lies on data sampling and spatial decision support with digital devices, despite the last mile gap.

Further developments in the described data sampling approach with a digital device can be to include adaptive elements of the questionnaire as described in [SCL+11]. In an adaptive questionnaire, the structure of elements and questions is not fixed, and questions appear in relation to the answers of previously answered questions. Additionally, a data sampling app can be enriched with open educational resources to increase risk literacy and awareness. An example of an app for a citizen-science project with an adaptive GUI and enriched with information about risk mitigation strategies is presented in chapter 12.

The design of the citizen-science task must be developed in the innovation cycle in the LL and should be done in a way that data sampling can be integrated in everyday life and does not lead to an overload for the citizen scientists.

Despite the mentioned advantages of citizen-science projects, low data quality is partly reported for citizen-science projects. The following chapter deals with how to deal with data sets with missing items and how to estimate the spatial and temporal quality of data.

7 Dealing with incomplete data and quality of data

7.1 Introduction

In reality, it is often necessary to deal with incomplete, imprecise, or obsolete data, especially in less-developed countries [BL03] and citizen-science approaches [ADSZ+19]. Citizen-science projects show a wide variety in data quality [BMN+21]. Different errors in data collection and processing can lead to a reduction of the data quality. Additionally, the incorrect use of sensors and data sources for generating temporal and spatial data can lead to a diminished data quality [HB92].

A distinction is made between internal and external data quality. The previously described errors lead to a decrease of the internal data quality. The internal data quality describes the difference between the produced data and the data that should be produced. In the best case, the produced data should give an equation of reality with specified filters [HB92].

External data quality describes the appropriateness of data to the user's needs. The best case for decision support would be to have current data from every location of interest; however, in reality, this is not the case. Therefore, data must be used from different time points in the past, interpolated from different locations with known attributes. External data quality can be measured as the difference between the needed data and the available data [HB92]. Thus, external data quality is always in relation to the needs of the person using the data.

The following chapters deal with how to handle missing values and inter- and extrapolation and how data quality is influenced and can be measured.

7.2 Interpolation and extrapolation of temporal and spatial data

As everything in the world changes over space and time, all things and phenomena in the real world as well as in the described data sampling approach in an LL have a temporal and spatial dimension. To describe these phenomena, for example, concentrations of substances, location, and amount of resources, the spatial and temporal dimension must be incorporated. The spatial dimension can be, for example, visualized by maps or diagrams, and the temporal dimension can be visualized by a bundle of maps or diagrams for different time points.
In former years, cartographers could only create maps from well-known areas, e.g., to map the density of trees in a wood. With modern computer techniques and the use of GIS, it is easily possible to interpolate missing locations with known points or areas surrounding them or missing time points for a given location with values for neighboring time points. As a result, different techniques are used through most of which use nearby points in a geographical or temporal sense to estimate values for unknown areas or points in time.
For a mathematical description, the following introduces some definitions.

Definition 7.2.1 (Statistical unit) *A statistical unit or element ω_i, $(i = 1, ..., n)$ is the single object of a statistical analysis. It is the carrier of the information of interest in the investigation. The index i characterizes the statistical unit, and n denotes the total number of recorded elements [Koh05].*

In the case of the described approach elements can be, for example, locations at a given time point (x_i, y_i, t) or people to whom a risk value is assigned to at a given time point (p_i, t).

Definition 7.2.2 (Population) *The statistical mass or population Ω is the set of statistical units that fulfill a specified delimitation criteria:*

$$\Omega = \{\omega_i | criteria\ applies\ to\ \omega_i\} \tag{7.1}$$

[Koh05]

The population Ω might be, for example, all locations within a given area L or all people living in an LL community.

Definition 7.2.3 (Sample) *If only a part of the population Ω is recorded in a statistical study, then this part is called a sample [Koh05].*

In data collection with a spatial and or temporal dimension, mostly only samples can be investigated due to the finite characteristics of the measurement, i.e., only a finite number of locations can be investigated at a finite number of time points.
During a sampling approach, a subset of locations or time points are measured with the aim of estimating values for unsampled locations or time points [Goo97].

Definition 7.2.4 (Feature) *A property of a statistical unit that is of interest in a statistical analysis is called feature $X(\omega_i)$ or X [Koh05].*

Definition 7.2.5 (Feature space) *The set of all possible observation values is called the feature space A [Koh05].*

In this sense, a feature X can be regarded as a mapping that assigns a observation value from the feature space to each element or statistical unit ω_i:

$$X : \Omega \rightarrow A \tag{7.2}$$

In the case of an agrochemical-related LL, a feature X can be for example the stage of CKD with a feature space $A = \{1, 2, 3, 4, 5\}$, and the statistical units ω_i might be people living in the LL. In the sense of a mapping, the feature X assigns to each element ω_i or person living in the LL a value between 1 and 5.
In general, and also in the described LL approach, measurements are made in a finite number of statistical units, e.g., the number of people living in an LL is finite or measurements about agrochemical content in soil or water is only sampled at a finite number of locations. However, in reality, the unknown distribution is continuous. By using only a finite number of sampling points, the continuous but unknown distribution is estimated by a discrete distribution.

Definition 7.2.6 (Discrete feature) *If a feature can only have a finite number of values or a countably infinite number of values, it is called a discrete feature; if it can be all values within an interval, it is called a continuous feature. [Koh05]*

Regarding the feature location itself, expressed in coordinates, it is, according to the definition above, a continuous feature. However, with the boundary condition that an investigation has only limited resources and therefore only a finite number of sampling points can be used in the investigation, it gets a discrete feature.

Definition 7.2.7 (Discretization) *Discretization methods are procedures for the approximate solution of continuous problems. The solution is only determined at specified discretization points within the domain of function. [Wal16a]*

In reality, all spatial and temporal measurements are discrete because only a finite number of measurements can be made, for example, a subset of locations or a finite number of time points for one location.
Following the example of an investigation interested in the concentrations of an agrochemical, e.g., in soil in relation to the spatial location (x, y) and time point t_0, is regarded. The spatial and temporal distribution of the agrochemical is an unknown but continuous function, as visualized in figure 7.1. During the survey the unknown continuous distribution of the agrochemical should be estimated by using a finite set of discrete sampling points.
For such an approach, a sampling design is created, for example, an appropriate spatial distribution of sampling points and the regularity of the time points at which they are measured. Detailed information about sampling designs can be taken from [SE03], [WSGG12] or [ZS06]. In figure 7.1, the hypothetical but mostly unknown spatial distribution of the concentration of a substance together with potential sampling locations are visualized. For demonstration purposes, an irregular sampling grid is used; however, in reality, the selection of a fitting sampling grid is a task that must be performed by stakeholders, e.g., a scientist, within the LL approach.

During a survey, the goal is that the unknown distribution of the parameter of interest is modeled using a sample of locations and, for each location, different time points. A researcher might sample a discrete number of locations and obtain, e.g., a table in which each row represents a measurement, and columns represent parameters of interest. Such data might be stored in a database or vector format in a GIS, for example, each measurement with at least the coordinates x_i, y_i of the location i, time point t_j, and value of the measured data z stored as a tuple (x, y, t, z) for each measurement. This data can be for example visualized with a GIS in a map with marks in which the z-value or the color of the marks represents the value of the measured parameter z at the location (x, y) at a given time point t. Such a map is visualized in following figure 7.2.

In a measurement for each location of the subset or time point, a value for the parameter of interest is obtained. With the obtained values for the parameter of interest, the goal is to estimate the unknown function in relation to space and or time through which values for unsampled locations can be estimated. There are three main techniques to estimate missing

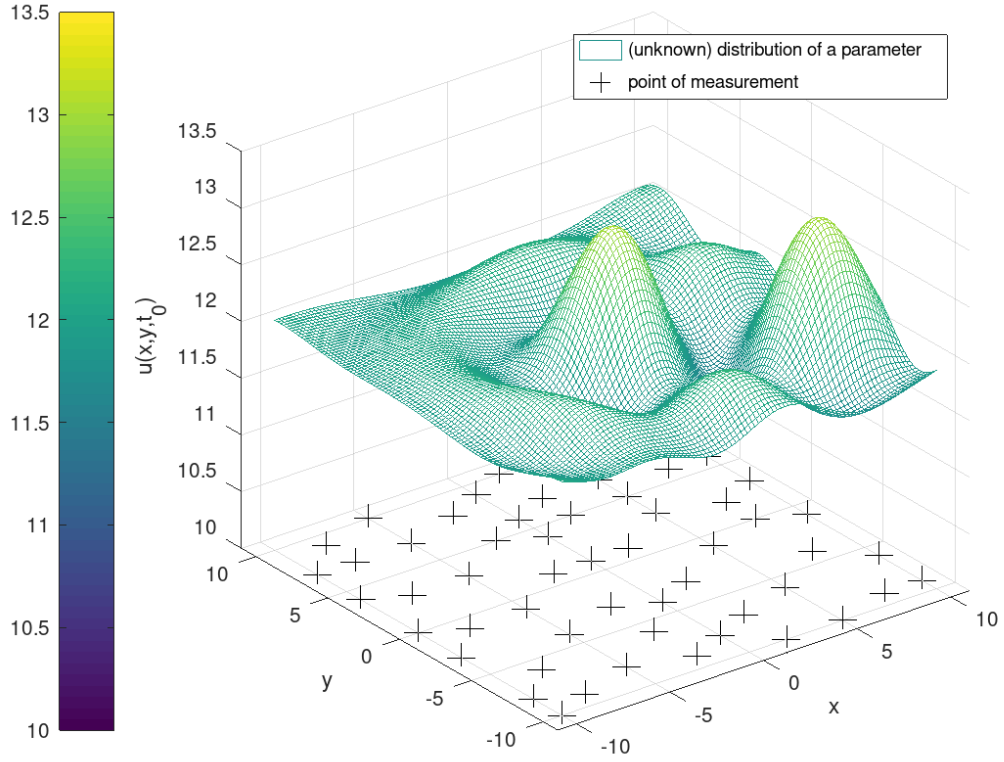

Figure 7.1: Example of a function $u_{t_0}(x, y)$ representing the unknown spatial distribution of a parameter of interest at time point t_0 together with a underlying sampling grid with possible sampling points (figure generated with *GNU Octave*).

functions by incorporating known points: extrapolation, interpolation, and function approximation.

Definition 7.2.8 (Interpolation and extrapolation) *Interpolation and extrapolation are defined as the estimation of values of a parameter at unsampled locations by using values of sampled locations; an estimation made within a sampled area is called interpolation, and one made outside of an area of sampled locations is called extrapolation [BML15].*

According to the definition above, extra- and interpolation differ in that point, i.e., that interpolation methods model values within a region of known points and extrapolation outside of a region with known points. However, methods used for spatial extrapolation are the same as used for spatial interpolation [LH08]. Extrapolation is used for temporal and spatial forecasting in an SDSS.

According to the dimensions of the problem, interpolation, extrapolation, and approximation can be temporal, spatial, or spatio-temporal.

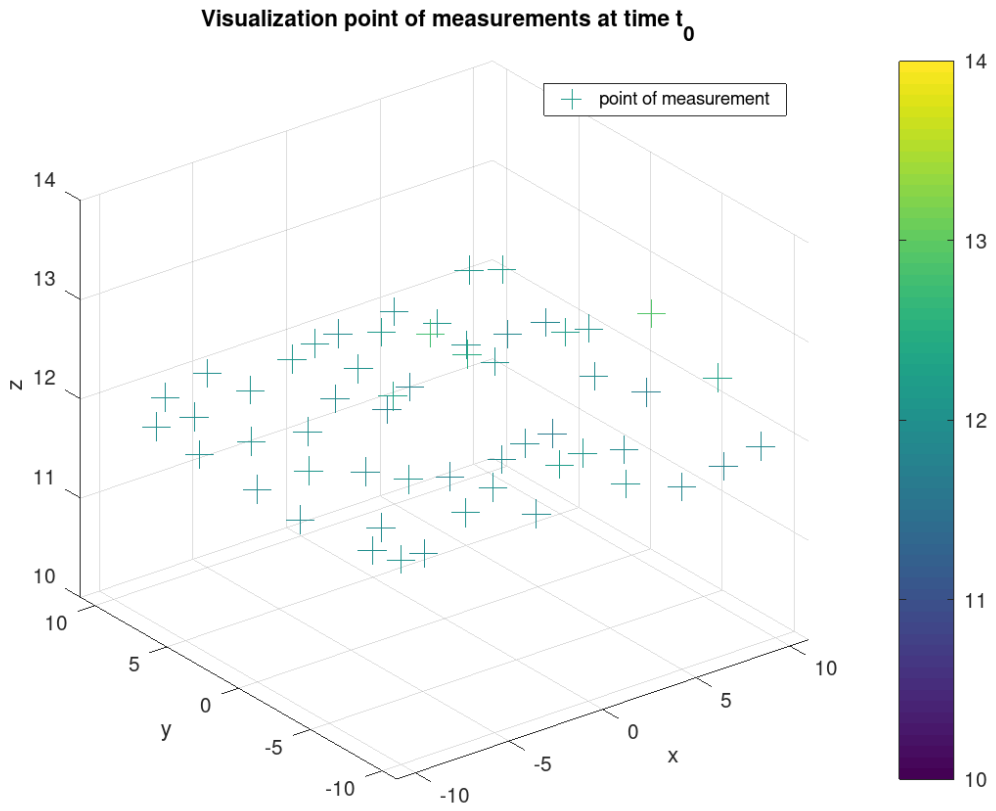

Figure 7.2: Example for the visualization of discrete point samples for a parameter of interest in a map at a time point t_0 (figure generated with *GNU Octave*).

The definition above is a definition used in a spatial or temporal context, but the following is a more mathematical one:

Definition 7.2.9 (Interpolation) *Interpolation is defined as construction of a function f that assumes predefined values z_j at a finite number of points x_j, the so-called sampling points with the following condition:*

$$f(x_j) = z_j \ for \ j = 1, ..., n \tag{7.3}$$

[Wal16b]

Interpolation as defined in definition 7.2.9, where the interpolation condition only regards function values x_j and z_j, is called *Lagrange interpolation* [Wal16b]. There exist a infinite number of equations that fit to this condition. Other conditions, like the type of the searched function or information about derivatives, can be used to refine and to reduce the number of possible functions. Other interpolation techniques, e.g., by implementing derivatives into the interpolation condition, are called *Hermite interpolation* [SX95] or *Birkhoff interpolation*

[LZ71]. However, for the task in an LL approach, where a function should be constructed from scattered data with known data points, *Lagrange interpolation* fits.
Similar to interpolation is the method of approximation. For approximation equation 7.3 is reduced to

$$f(x_j) \approx z_j \; for \; j = 1, ..., n \quad \text{[Fol87]} \tag{7.4}$$

meaning that in contrast to interpolation, the output values of the found function $f(x_j)$ at the known input values x_j have only to fit approximately to the output values z_j.
There are several temporal, spatial, and spatio-temporal interpolation and approximation methods described in the literature.
[LH14] divides available spatial interpolation methods into three categories. However, currently machine-learning interpolation methods, which were not included in [LH14] due to the date of the publication, have become increasingly popular, i.e.:

- non-geostatistical methods, like *nearest neighbor*, *inverse distance weighting*, *splines*, *regression trees* and models, *triangular irregular network related interpolations*[LH14],
- geostatistical methods, mainly *kriging*-related methods [LH14],
- machine learning methods,e.g., *support vector machine* [ZSX20], [SKTGMdS$^+$20], *neural networks* [ZSX20], [AGRK20], [SKTGMdS$^+$20], [NM23], *neuro-fuzzy networks* [KPUR23], [NM23], *random forest* [SKH$^+$20], *boosted decision trees* [RLS$^+$20],[TWK$^+$21], and *k-nearest neighbor* [ZSX20], [SKTGMdS$^+$20], and
- combined methods, where non-geostatistical, geostatistical, and machine learning methods are combined, like *linear mixed models*, *regression kriging* [LH14].

[LH14], [LH08] and [LH11] give a good overview about spatial interpolation techniques in environmental science.
Geostatistical interpolation methods incorporate, in contrast to non-geostatistical methods, values for the variance at unsampled locations together with a probability distribution of the variance, whereby a semivariogram is calculated and used in the interpolation process [BML15]. A semivariogram visualizes the relationship between a location and its surrounding locations with respect to the distance [SdSJH16]. These methods are also called *kriging*, named after one of the pioneers of geostatistics, *Danie G. Krige* [BML15]. *Kriging methods* are very common and often used, although they have a high computational resource need

[LH14].

According to [LH08], interpolation methods differ in their features, characteristics, and applicability, such as the geographic extent of incorporated data points, if they are interpolation or approximation methods, if they use a deterministic or stochastic procedure, if they produce a discrete or continuous output, if the interpolation is based on uni- or multivariate input data, and if they work on a regular or irregular network of points. [LH08] also give an overview about non-geostatistical, geostatistical, and combined methods, their features, limitations, computing load, and the suitability.

Research in the field of spatio-temporal models is much lower than in the only spatial field [SdSJH16]. [SdSJH16] review different spatio-temporal interpolation methods, for example, *triangular*, *inverse distance weighting*, *kriging*, and *distribution-based distance-weighting interpolation*. [XCC+21] divides spatio-temporal models into three categories; however, mixed models are also available. The following is a non-exhaustive list of examples of commonly used methods in spatio-temporal interpolation:

- Statistical methods: *geographical weighted regression* [FCY15], [DWZ+18], *spatio-temporal kriging* [XZCH20], [HLZ+20], [YH18], *Bayesian maximum entropy* [HK18], [HK17] , *ARIMA* [DRU+20],[MSI24], *timescape* [CCP+22], *dual kriging* [KCM24].
- Machine-learning methods: *support vector machine* [DCKBK24],[dSdLdS+21],[GRMI21], *neural networks* [NSR23],[SBD21], *random forest* [MDB+20], [Yeş20].
- Physical methods: *EPIC*, *SWAT*, *Weather Research and Forecasting model* [XCC+21].
- Mixed models.

According to [LH14], the quality of the interpolated data is dependant on the used model, used model parameter, characteristics, and quality of input data, such as spatial distribution and density of samples and correlations between input data.

Packages and commands for spatial and spatio-temporal interpolation are partly available in open-source GIS tools or other open-source software. For example, in *QGIS 3* [QGI24], there are packages for *inverse distance weighting* (*qgis:idwinterpolation*) and *triangulated irregular network* (*qgis:tininterpolation*) interpolation directly implemented. Plugins for *ordinary kriging* and *support vector machine* methods are available by the *Smart-Map Plugin* [PVQ+22]. Additionally, other commonly used open-source GIS have different interpolation methods implemented; in *GRASS GIS 7* [Tea15], there are, e.g., *bilinear* or *bicubic spline interpolation* (*v.surf.bspline*), *inverse distance weight interpolation* (*v.surf.idw*), and in *SAGA-GIS 7*

[CBB+15], e.g., *inverse distance weight interpolation*, *modified quadratic Shepard interpolation*, *bilinear spline interpolation* or different *kriging* algorithms. Additionally, in open-source tools for numerical computation different interpolation methods are implemented; for example, in *GNU Octave 5* [EBHW20], the methods *nearest neighbor* or *linear interpolation from nearest neighbor* are used.

Other indirectly implemented interpolation models can be manually implemented, e.g., through the python interface of *QGIS* or with programming skills directly in other open-source software like *GNU Octave*, *R* or *Python*.

In an LL approach, the appropriate interpolation method must be selected according to the mentioned features of the interpolation methods and how they fit to the data for which a interpolation should be performed. For an LL in a less-developed country with limited resources, there should also be a focus on whether the method is implemented in an open-source GIS or whether stakeholders are available who can implement the methods in the used software tools. Furthermore, the workload and required IT equipment in terms of central processing unit (CPU), random-access memory (RAM) and storage must be emphasized.

Regarding the example above and assuming that for each location (x_i, y_i) samples were taken at different time points t_j, the values for the parameter of interest $z_{i,j}$ were obtained. $(x_1, y_1, t_1)....(x_n, y_n, t_m)$ are further called interpolation points and can be described as a location with x and y coordinates at a given time point t. To each interpolation point, a corresponding value $z_{1,1}, ..., z_{n,m}$ is related that represent measured values of the parameter of interest at location i and time point j.

Regarding the example above, interpolation in general means that a function f has to be determined, for which the following property fits:

$$f(x_i, y_i, t_j) = z_{i,j} \quad for\ i = 1, ..., n\ and\ j = 1, ..., m \tag{7.5}$$

In other words, a function is searched for which the output value $f(x_i, y_i, t_j)$ for a given location and time is the same as the measured value $z_{i,j}$. The value for an unknown or unsampled location and time is then obtained and estimated by inserting the regarded values into the found expression.

For a map, representing the situation at a fixed time point t_0 the equation is reduced to:

$$f(x_i, y_i, t_0) = f_{t_0}(x_i, y_i) = z_{i,0} \quad for\ i = 1, ..., n \tag{7.6}$$

In the following figure 7.3, an example of a spatial interpolation is visualized, a *linear interpolation to the nearest neighbor* in each direction was used as interpolation method.

Interpolation was performed with the software tool *GNU Octave*. *GNU Octave* does not give

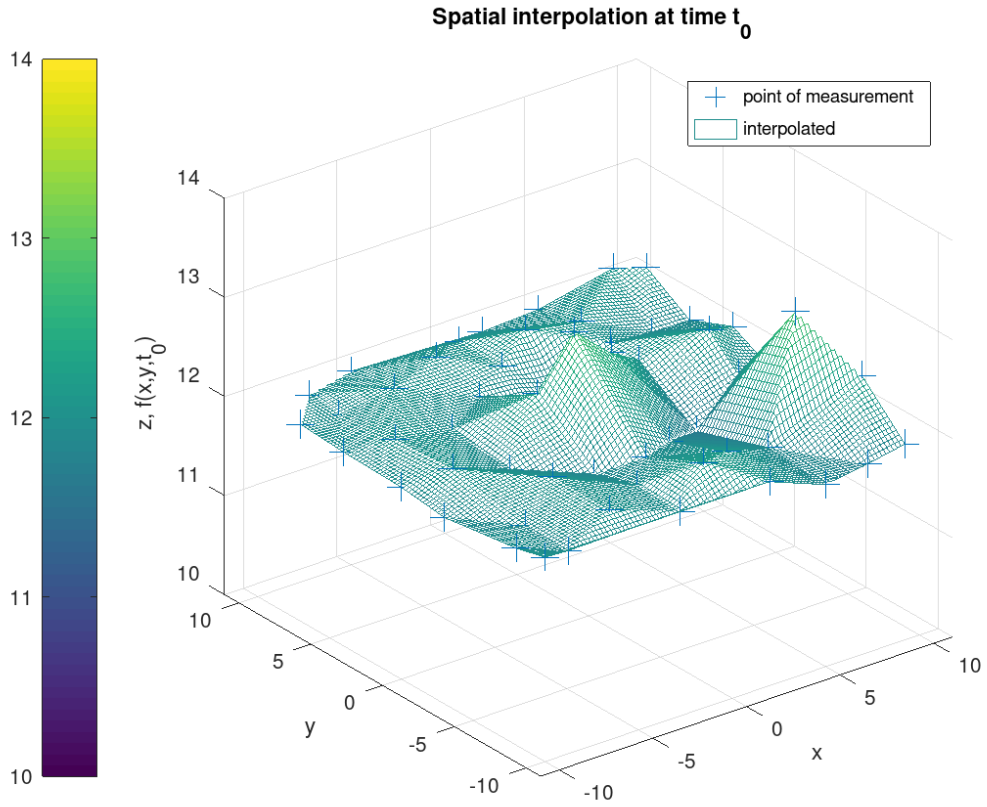

Figure 7.3: Example for the interpolation based on scattered discrete point samples for a parameter of interest in a map at a fixed time point t_0 (figure generated with *GNU Octave*).

an explicit equation for the interpolated function; however, the interpolated z values can be obtained by a query with respect to the location x and y within the *GNU Octave* code.
Figure 7.4 shows an example for a second measurement at time point t_1 at the same locations as measurement at time point t_0 together with a spatial interpolation, performed with the same *linear interpolation* method as used in figure 7.3: With these two measurements at different time points, it is possible to perform a temporal interpolation, e.g., to estimate the situation at any time point between t_k and t_l with $l > k$:
Temporal interpolation is demonstrated by applying a method called convex combination.

Definition 7.2.10 (Linear and convex combination) *Let K be a field and λ_i scalars with $\lambda_i \in K$, V a vector space and $v_i in V$. A vector $v \in V$ to which $a_1, ..., a_n \in K$ exist and with $v = \lambda_1 v_1 + ... + \lambda_n v_n$ is called linear combination of the vectors $v_1, ..., v_n$. A linear combination is called convex combination only if $\lambda_1, ..., \lambda_n \geq 0$ and $\lambda_1 + ... + \lambda_n = 1$ [Wal16b]*

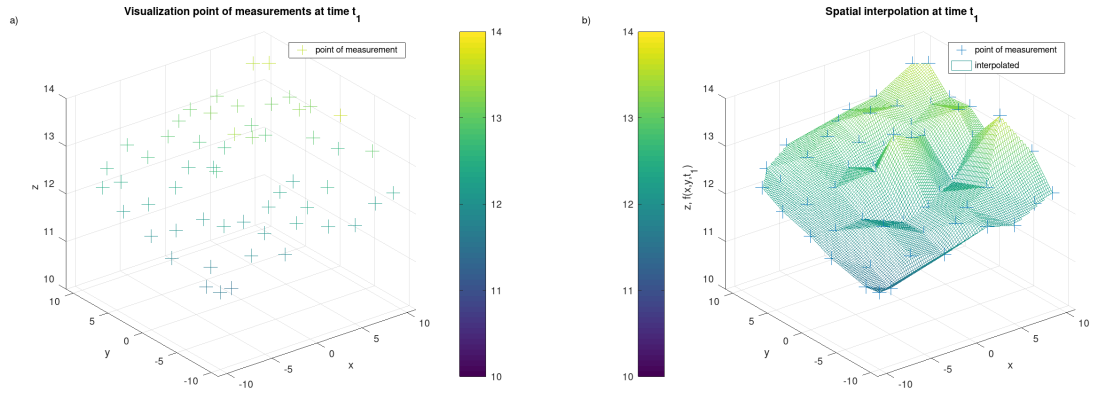

Figure 7.4: Example for the visualization of discrete point samples for a parameter of interest in a map at a time point t_1 together with a spatial interpolation (figure generated with *GNU Octave*).

This definition is following used to perform a linear interpolation of the z values of a unknown time point.

It is assumed that for every spatial location (x_i, y_i) with $i = 1, ..., n$ the corresponding $z-$ values $z_{i,k}$ and $z_{i,l}$ for time points t_k and t_l are known. The aim of the following temporal interpolation is to get $z-$ values for time point t_x with $k <= x <= l$, for which the corresponding $z-$ values $z_{i,x}$ were not measured during a survey.

Using the definition for a convex combination for two vectors, the equation from the definition 7.2.10 becomes:

$$v = \lambda_1 v_1 + \lambda_2 v_2 \tag{7.7}$$

with $\lambda_1 + \lambda_2 = 1$ equation 7.7 becomes:

$$v = (1 - \lambda_2) v_1 + \lambda_2 v_2 \tag{7.8}$$

and following:

$$v = (1 - \lambda) v_1 + \lambda v_2 \tag{7.9}$$

assuming a linear relationship and $k <= x <= l$, λ is defined as:

$$\lambda = \frac{x - k}{l - k} \tag{7.10}$$

resulting in:

$$v = \left(1 - \frac{x - k}{l - k}\right) v_1 + \frac{x - k}{l - k} v_2 \tag{7.11}$$

with v_1 and v_2 regarding as two vectors $v_1, v_2 \in \mathbb{R}^3$ and defined as:

$$v_1 = \begin{pmatrix} x_i \\ y_i \\ z_{i,k} \end{pmatrix} \text{ and } v_2 = \begin{pmatrix} x_i \\ y_i \\ z_{i,l} \end{pmatrix}$$

equation 7.11 becomes:

$$z_{i,x} = \left(1 - \frac{x-k}{l-k}\right) \begin{pmatrix} x_i \\ y_i \\ z_{i,k} \end{pmatrix} + \left(\frac{x-k}{l-k}\right) \begin{pmatrix} x_i \\ y_i \\ z_{i,l} \end{pmatrix} \tag{7.12}$$

By applying equation 7.12 on all points of measurement for two time points, a simple linear temporal interpolation between two time points can be performed.

Regarding the scatter plots visualized in figure 7.3 and figure 7.4 for time points t_0 and t_1, a map for time point t_x with $0 <= x <= 1$ can be created by using equation 7.12.

For example, for the location $i = 11$ with the coordinates $x = -8$ and $y = 7$, the following z-values were obtained by measurements at time point $t_0 = 0$ and $t_1 = 1$ with $z_{11,0} = 11.958$ and $z_{11,1} = 12.391$. Through temporal interpolation, the missing value at $t_{0.6}$ is obtained.

Therefore, the following parameters can be set: $k = 0$, $l = 1$, $x = 0.6$,

$$v_1 = \begin{pmatrix} x_i \\ y_i \\ z_{i,k} \end{pmatrix} = \begin{pmatrix} x_{11} \\ y_{11} \\ z_{11,0} \end{pmatrix} = \begin{pmatrix} -8 \\ 7 \\ 11.958 \end{pmatrix} \text{ and } v_2 = \begin{pmatrix} x_i \\ y_i \\ z_{i,l} \end{pmatrix} = \begin{pmatrix} x_{11} \\ y_{11} \\ z_{11,1} \end{pmatrix} = \begin{pmatrix} -8 \\ 7 \\ 12.391 \end{pmatrix}$$

and with equation 7.12 following in

$$z_{11,0.6} = \left(1 - \tfrac{0.6-0}{1-0}\right) \begin{pmatrix} x_1 1 \\ y_1 1 \\ z_{11,0} \end{pmatrix} + \left(\tfrac{0.6-0}{1-0}\right) \begin{pmatrix} x_1 1 \\ y_1 1 \\ z_{11,1} \end{pmatrix} = 0.4 \cdot \begin{pmatrix} -8 \\ 7 \\ 11.958 \end{pmatrix} + 0.6 \cdot \begin{pmatrix} -8 \\ 7 \\ 12.391 \end{pmatrix} =$$

$$\begin{pmatrix} 0.4 \cdot (-8) \\ 0.4 \cdot 7 \\ 0.4 \cdot 11.958 \end{pmatrix} + \begin{pmatrix} 0.6 \cdot (-8) \\ 0.6 \cdot 7 \\ 0.6 \cdot 12.391 \end{pmatrix} = \begin{pmatrix} -8 \\ 7 \\ 0.4 \cdot 11.958 + 0.6 \cdot 12.391 \end{pmatrix} = \begin{pmatrix} -8 \\ 7 \\ 12.218 \end{pmatrix}$$

Therefore, at location 11 with $(x = -8, y = 7)$ and time point $t_{0.6}$, the following z-value is obtained: $z_{11,0.6} = 12.218$.

This calculation is further performed for all $i = 1, ..., n$. In this manner, interpolated values are visualized in figure 7.5 a), a spatial interpolation was then performed, resulting in figure 7.5 b):

Figure 7.5 is the result of an easy implementable tempo-spatial interpolation performed with open-source software since it can be used in the described LL approach in a SDSS.

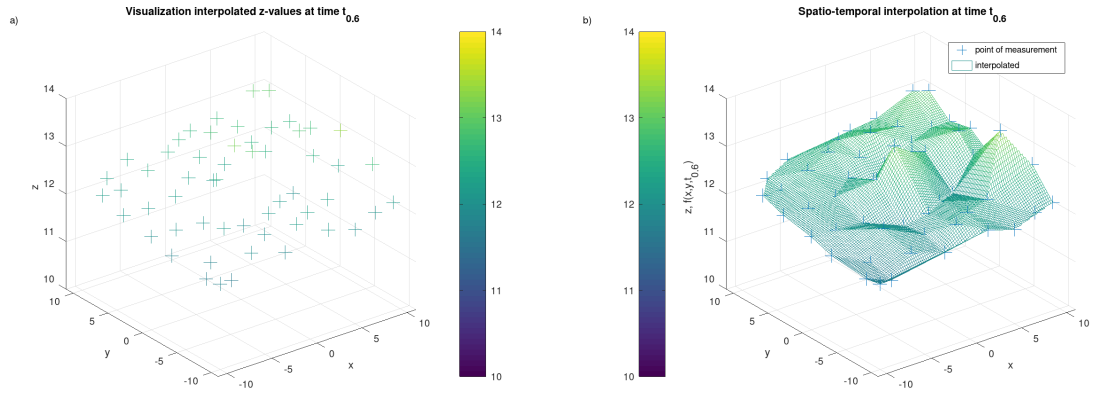

Figure 7.5: Visualization of discrete point samples generated by temporal interpolation with convex combination at time point $t_{0.6}$, together with a spatio-temporal interpolation (figure generated with *GNU Octave*).

7.3 Data density, interpolation quality and data quality

Interpolated maps, as visualized in figure 7.3, look as if precise measurements were performed at every location. Instead, missing data was modeled, and the quality of the modeled data is not the same as for locations for which concrete measurements are available.

When dealing with spatial and temporal data, it must be kept in mind that this data represents a part of the reality in a modeled way. In spatial and temporal data, items and attributes of the real situation are sometimes not considered, grouped, or simplified. They try to give a picture of reality, but they are not reality [DJ06]. The deviation or the probability of deviation between modeled and really measured data must be considered when this data is used for, e.g., spatial decision support, as described in this thesis.

In the data modeling framework one of the internal sources of error can be described as interpolation errors. In figure 7.1, the distribution of a parameter of interest is visualized. In reality, this distribution is unknown but should be modeled and estimated using a sample of measurements at predefined points of measurement, for example, as visualized in figure 7.2. The resulting interpolated map is presented in figure 7.3.

The comparison between the unknown distribution $u_{t_0}(x, y)$ (figure 7.1) and the modeled or interpolate distribution $f_{t_0}(x, y)$ (figure 7.3) shows that there is a difference between the two functions. As an error function, the squared error is used:

$$e_{f,t_0}(x, y) = (u_{t_0}(x, y) - f_{t_0}(x, y))^2 \tag{7.13}$$

The graph of e_{f,t_0} is visualized in following figure 7.6 The volume under the surface of the error function e_{f,t_0} is further used as a measurement for the overall error E_{f,t_0} of interpolation

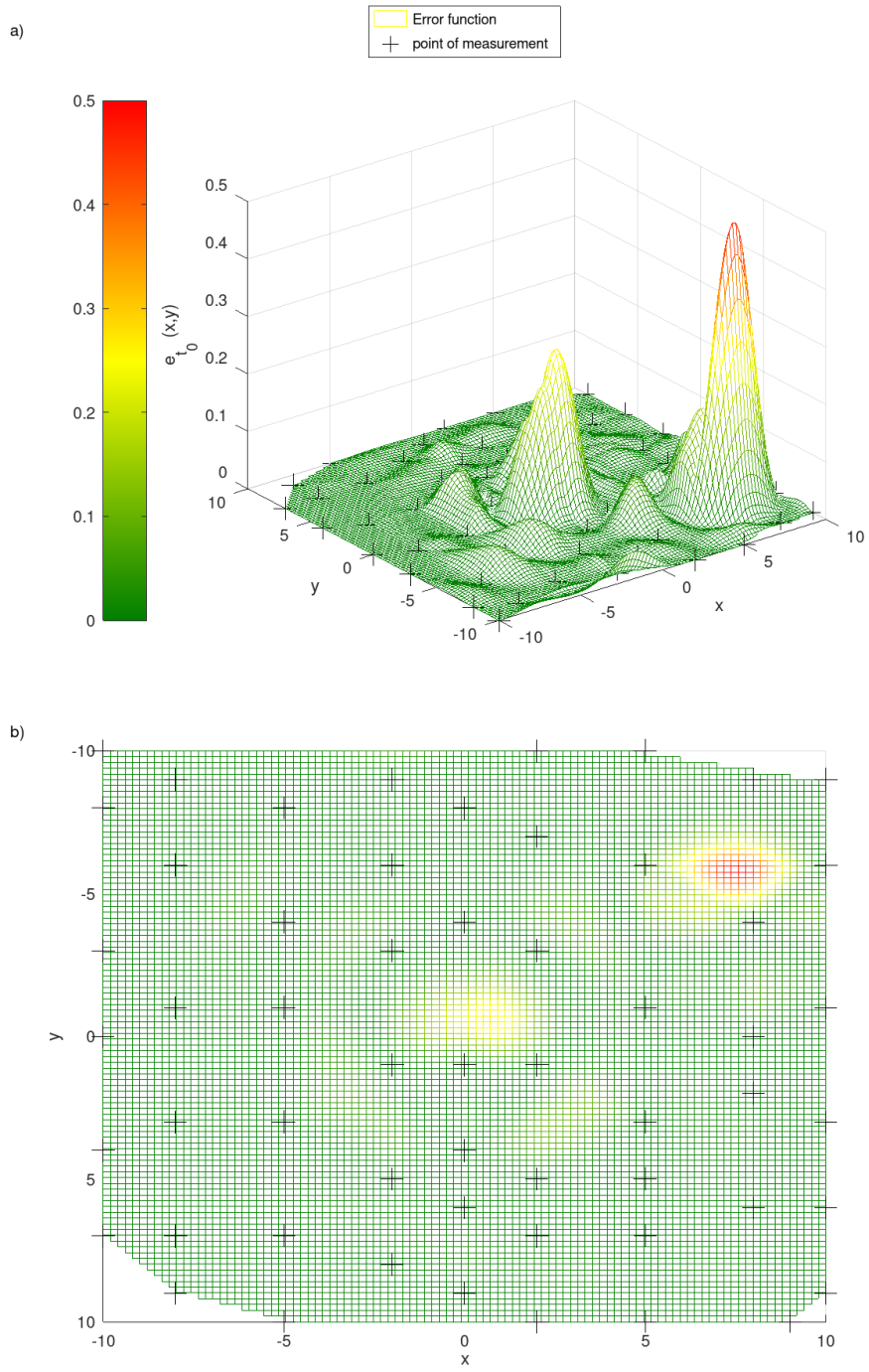

Figure 7.6: Graph of the error function e_{f,t_0} from two different angles of view (figure generated with *GNU Octave*).

function f_{t_0}.

$$E_{f,t_0} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e_{f,t_0}(x,y)\cdot I_G(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (u_{t_0}(x,y)-f_{t_0}(x,y))^2\cdot I_G(x,y)\,\mathrm{d}x\,\mathrm{d}y \quad (7.14)$$

However, E cannot be calculated in reality as u is recently unknown. In equation 7.14, the term $I_G(x, y)$ is an indicator function, which describes a bounded measurable set and is defined as follows.

Definition 7.3.1 (Indicator function) *Indicator function of an area G is defined as:*

$$I_G(x,y) = \begin{cases} 1 & , \ if \ (x,y) \in G \\ 0 & , \ else \end{cases} \tag{7.15}$$

As it is possible that the improper integral diverges to infinity, it is introduced to limit the integral to a finite value.

According to figure 7.6, there is a heterogeneous distribution of the error e_{t_0} on a global scale with a higher error near the peaks of the unknown function u_{t_0}. This might be explained by the locations and distribution of the points of measurement as well as with the shape of the error function e. Furthermore, the used linear interpolation method to the nearest neighbor might be responsible for the obtained error distribution.

Regarding definition 7.2.9, the error is 0 at the marked interpolation points and increases with the distance to the interpolation points. The error is lower in regions where more points are present per area or in other words, in regions with a higher point density.

Therefore, a density function d_i is introduced that is related to the distance of the location of a point of measurement (x_i, y_i) to its surrounding locations and decreases with increasing distance:

$$d_i(x,y) = p_i(x,y) \cdot I_G \tag{7.16}$$

with

$$p_i(x,y) = \frac{1}{1 + \left(\frac{x-x_i}{s}\right)^2 + \left(\frac{y-y_i}{s}\right)^2} \tag{7.17}$$

In equation 7.17, s is a value describing how far a single point of measurement has an influence on the surrounding locations. Higher values of s imply a greater influence of a specific measuring point on surrounding locations.

Regarding the measurement grid from the example above, an aggregated density function d_{spat,t_k} at time point t_k can be calculated as:

$$d_{spat,t_k}(x,y) = \sum_{i=1}^{n_k} d_i(x,y) \tag{7.18}$$

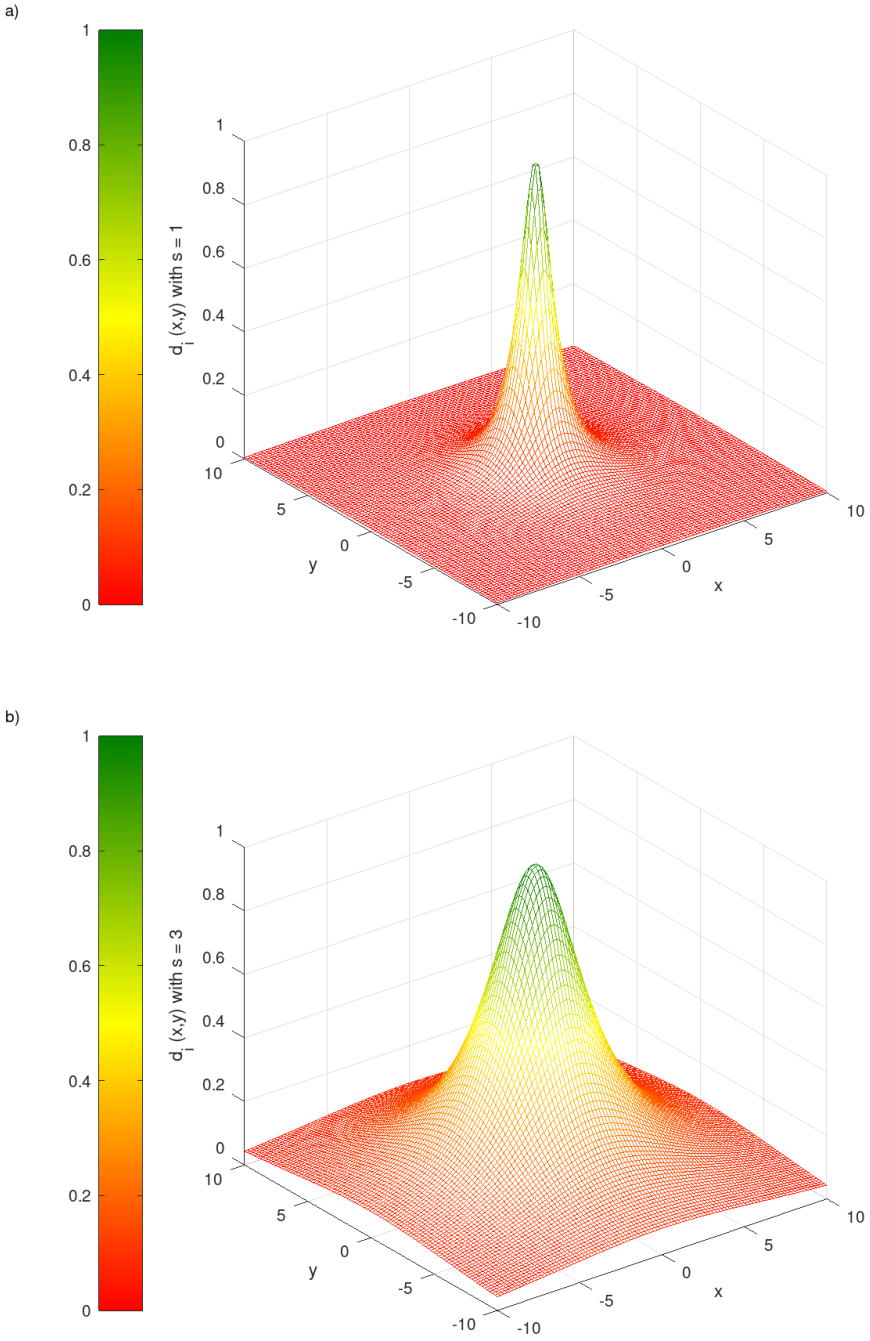

Figure 7.7: Graph of the density function d_i for a measuring point i at location $(0, 0)$ with s = 1 and s = 3 (figure generated with *GNU Octave*).

$$= \sum_{i=1}^{n_k} p_i(x, y) \cdot I_G \tag{7.19}$$

$$= \sum_{i=1}^{n_k} \frac{1}{1 + (\frac{x-x_i}{s})^2 + (\frac{y-y_i}{s})^2} \cdot I_G \tag{7.20}$$

n_k represents the number of sampling points for which measurements are available at time point t_k.

With respect to the sampling grid given in figure 7.1, aggregated density functions with $s = 1$ and $s = 3$ are visualized in the following figure 7.8.

The aggregated density function for $s = 3$ has a smother surface with a higher d_{spat} values than compared with the aggregated density function for $s = 1$, which can be expected by regarding the density functions d_i for single points and the method of summing up the density functions for every single sampling point. The influence of a sampling point to the neighboring locations is higher for $s = 3$, resulting in higher d_{spat} values than compared with $s = 1$.

Selection of a value for s can be regarded as fitting of a model parameter and must be performed in the LL approach according to the data used in the model and how far a value of an unknown location can be estimated by a value of a nearby lying sampling point.

Aggregated density function can be used to calculate the mean information density: $\overline{d_{spat,t_k}(A)}$ within an area A with $A \subset G$:

$$\overline{d_{spat,t_k}(A)} = \frac{Volume\ under\ d_{spat,t_k} with\ base\ area\ A}{Area\ of\ A} = \frac{\int\limits_A d_{spat,t_k}(x,y)\ \mathrm{d}x\ \mathrm{d}y}{\int\limits_A 1\ \mathrm{d}x\ \mathrm{d}y} \tag{7.21}$$

The value of $\overline{d_{spat,t_k}}$ can be used to compare, e.g., the mean information density and quality of two areas D and E, or to select areas with a high information density.

For the spatio-temporal interpolation, a function $q_{temp,t_k}(t)$ is introduced. $q_{temp,t_k}(t)$ describes the relationship between the temporal quality and the time span between time t and the time point t_k of a measurement. The temporal quality decreases with a longer time span between the measurement and a regarded time point. q_{temp,t_k} is 1 at time points at which concrete measurements are available and decreases with an increasing time span between the time point of measurement and the regarded time point. Such a quality function is introduced for every measurement and related time point t_k:

$$q_{temp,t_k}(t) = \frac{1}{1 + \left(\frac{t-t_k}{s}\right)^2} \tag{7.22}$$

In the following figure 7.9, two examples for $q_{temp,t_k}(t)$ with $t_k = 1$ and $s_t = 1$ or $s_t = 3$ are visualized.

Again, s_t is a parameter describing how fast the temporal quality decreases with an increasing time span. The following examples,

$$q_{temp,t_k=1,s_t=1}(0) = \frac{1}{1 + \left(\frac{0-1}{1}\right)^2} = 0.5 = q_{temp,t_k=1,s_t=1}(2) \tag{7.23}$$

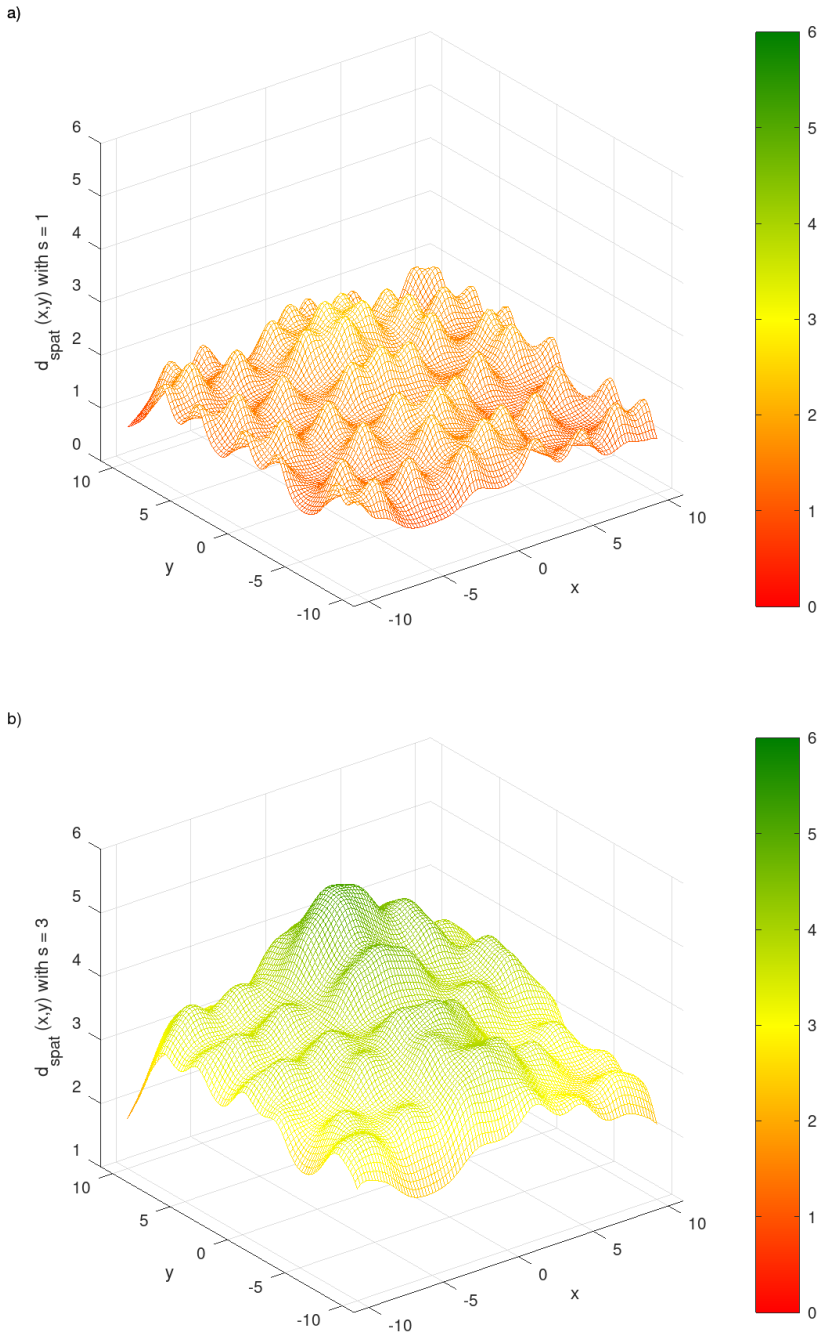

Figure 7.8: Graph of the aggregated spatial density function d_{spat,t_k} for all measuring points at location (x_i, y_i) with $i = 1, ..., n$, s = 1 and s = 3 (figure generated with *GNU Octave*).

and

$$q_{temp,t_k=1,s_t=3}(0) = \frac{1}{1 + \left(\frac{0-1}{3}\right)^2} = 0.9 = q_{temp,t_k=1,s_t=3}(2), \tag{7.24}$$

demonstrate that, for $s_t = 1$ the quality is reduced to 50% and for $s_t = 3$ to 90% for an interpolated time point, which is one time unit away from a time point for which a measurement

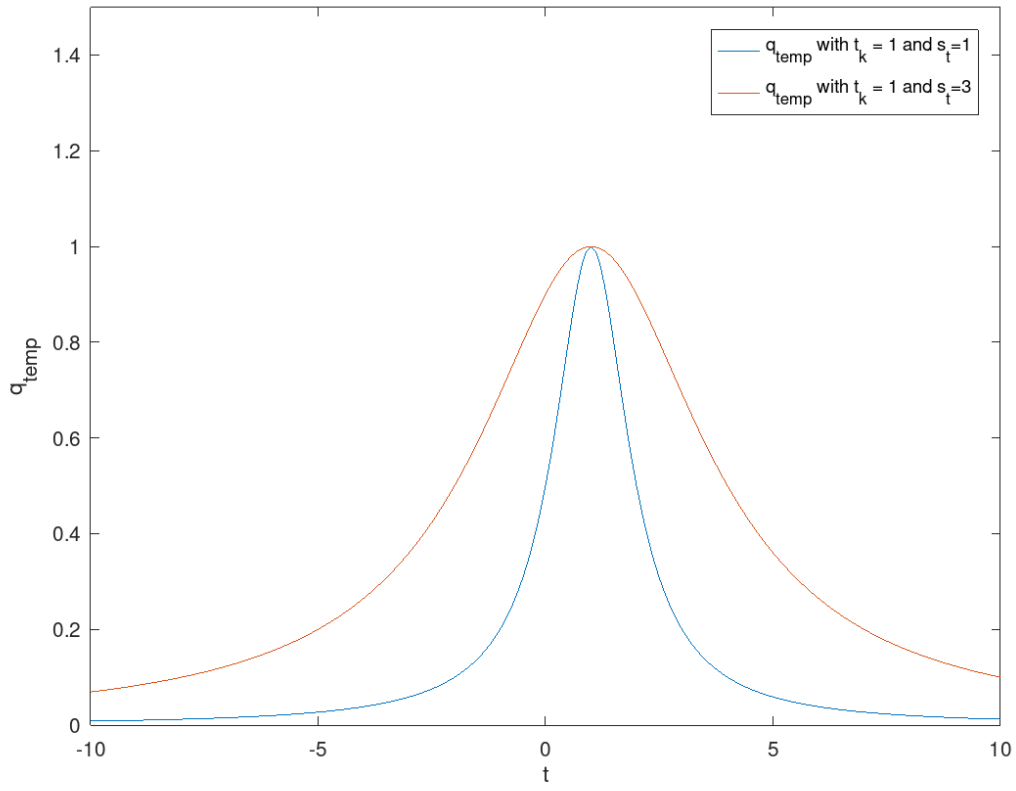

Figure 7.9: Example for a temporal quality function for a measurement at time point $t_k = 1$ and two different slope factors $s_t = 1$ and $s_t = 3$ resulting in $q_{temp,t_k=1,s_t=1}(t) = \frac{1}{1+\left(\frac{t-1}{1}\right)^2}$ and $q_{temp,t_k=1,s_t=3}(t) = \frac{1}{1+\left(\frac{t-1}{3}\right)^2}$ (figure generated with *GNU Octave*).

is available.

In the following, it is assumed that concrete measurements are available for time points t_j with $j = 1, ..., m$. For each time point, a spatial density function together with a temporal quality function $q_{temp}(t)$ is available.

The spatio-temporal density function $d_{spattemp,t_k}(x, y, t)$ for a single measurement at time point t_k is obtained by

$$d_{spattemp,t_k}(x, y, t) = d_{spat,t_k}(x, y) \cdot q_{temp,t_k}(t) \cdot I_G \tag{7.25}$$

$d_{spattemp,t_k}(x, y, t)$ is a function describing the spatio-temporal quality of a single measurement at time point t_k. For a fixed location (x_0, y_0), the value $d_{spattemp,t_k}(x_0, y_0, t)$ decreases with a increasing distance between t and t_k.

The overall spatio-temporal density function is finally calculated by:

$$d_{spattemp}(x, y, t) = \sum_{j=1}^{m} d_{spattemp,t_j}(x, y, t) \tag{7.26}$$

$$= \sum_{j=1}^{m} d_{spat,t_j}(x,y) \cdot q_{temp,t_j}(t) \cdot I_G \tag{7.27}$$

$d_{spattemp}(x,y,t)$ describes the spatio-temporal quality of location (x,y) at a time point t in relation to the spatial and temporal distribution of the measuring locations and time points of measurements.

The described interpolation methods and methods to estimate the spatial, temporal, or spatio-temporal quality of data points and sets are necessary for the application of a SDSS. For example, through interpolation methods, it is possible to estimate parameters required for a SDSS at locations where direct measurements of the parameter are not possible. The spatio-temporal quality of the used data is also a value that is necessary to estimate the robustness of the generated spatial support. The quality of the spatial support generated with data with low quality is not as high as a decision made with high quality input data.

7.4 Methods for dealing with missing data

In the last sections, methods were introduced that can be used to estimate unknown values of a parameter lying between existing data records in a spatial or temporal sense. The following chapter deals with methods how to handle missing data within data records.

In general, missing data values or whole parameters can appear due to, inter alia, non-responding [Rub04], errors in sensors [JE00], missing communication between a sensor and the electronic devise that records the sampled values [LG05], an accidental loss of values, or difficulties in measuring a parameter caused by high financial costs or special equipment needed for a measurement [LEDL$^+$05].

Besides missing data values, noisy data can often be regarded in surveys. Reasons are random errors or implicit errors occurring because of errors in a sensor or by using different sensors for measuring the same parameter [GLH$^+$15]. In less-developed countries in particular, data sampling of personal parameters is a difficult task. Personal experience has shown that available data sets are sometimes not complete or have missing values. However, for an SDSS, it is necessary to have complete data sets. If input data for an SDSS has missing values, this input data must be preprocessed in a way that algorithms can work with it and give appropriate decision support results.

Therefore, for an SDSS in the described LL approach a system is necessary that is able to handle missing values of a parameter in a data set. There are different methods available for dealing with missing data; [GLSGFV10] list the most important methods.

An easy and formerly often used method is called *complete data analysis*. In *complete data analysis*, record entries with missing values are deleted, and only record entries with a complete set of values are used. The method can be used if only a small percentage of the recorded entries have missing values. The method has the disadvantage that information are not used or lost [GLSGFV10], resulting in small data sets [LR19].

To impute missing values, statistical analysis and related values can also be used, for example, the mean or median of the missing parameter. In the method called *mean imputation*, the average value of the missing parameter is calculated by using existing values of the parameter of the complete records. However, correlations between the parameters and variability within the data set are not regarded [Sch97] by applying *mean* or *median imputation* the variance v of the data set changes.

Another method is called *regression imputation*. The regression method is used if a correlation between the missing parameter and a parameter with full records exists. A regression curve is calculated by using only the value pairs of the dependent and independent variable for which both values are available. Missing values are then calculated by using the obtained regression curve. However, regression analysis can only be used if the data set fulfills some conditions, such as independence of the predictor variables, homoscedasticity, or an uncorrelated error distribution [FHT15].

Hot deck imputation is another method to estimate missing values. Existing values in the incomplete vectors are compared with the values of the other vectors in the complete vector set. The vector in which the existing values have the highest similarity to the vector with the missing value is taken, and the missing value in the incomplete vector is replaced by the value from the vector with the highest similarity. With this method, global information in the whole area space is ignored, and only a single vector is responsible for the estimation. This results in a low quality of information [Sch97].

For model-based approaches like the *maximum likelihood* or *expectation-maximization algorithm*, knowledge about the distribution of the parameters is necessary [Sch97].

Another commonly used imputation method is called *multiple imputation* [KC07]. In contrast to *single imputation*, in the *multiple imputation* approach, imputation is repeated by incorporating probability distribution, resulting in many possible processed datasets. In a second phase, the different imputed datasets are finally combined into a single dataset [LSA15].

There are also different machine learning methods available [GLSGFV10] that are highlighted in Chapter 16:

- *K-Nearest-Neighbor (k-NN) imputation* [Zha12],
- *multilayer perceptron* [SRPMLC15],
- *self organizing maps* [FZCM15],
- *recurrent neural network imputation* [LR12],
- *support vector methods* [PDBSDM05], and
- *random forest imputation* [TI17].

[TR21] give a summary about imputation methods from the machine-learning field [RT20] for unsupervised imputation methods. A comparison of the different imputation techniques can be taken from [JPR19].

The different methods differ in computer load and accuracy. [JPR19] compared the performance of seven different common used imputation methods, whereby *k-NN imputation* imputation had the lowest root mean square error. However, in terms of computing load *k-NN imputation* is high.

They also differ in the applicability in terms of the statistical characteristics of the dataset, for example, if the parameters are statistically dependent or independent, e.g., for *mean* or *median imputation*, it does not matter if the variables are dependent; however, for regression, it is necessary.

Also a distinction is made between whether the values are missing at random, missing completely at random, or missing not at random. Completely at random means that there is no dependency between missing values and the variables; missing not at random means that there is a dependency, e.g., missing values are located in the upper scale of the variable [SWC+09].

Imputation methods are implemented in common open-source software tools like *R* with packages *mice* [VBGO11] for multivariate, *amelia* [HKB11] for *multiple*, *Hmisc* [HJHJ19] for *single*, or *missForest* [SB12] for *random forest imputation* and in *Python* with the classes *IterativeImputer* for multiple or *KNNImputer* for *k-NN imputation* from the *Scikit* package [PVG+11].

7.5 Chapter conclusion

In the present chapter methods were introduced with which it is possible to pre-process data in order that the datasets can be used in a SDSS. Datasets with missing data values can be

processed, e.g., by removing data records with missing values or by estimating missing data values with imputation methods. Such imputation methods are available for common open-source software tools, as proposed in the described LL approach in less-developed countries. Imputation methods from the field of artificial intelligence (AI) and machine learning (ML) are presented in more detail in the following chapter 16.

Additionally, interpolation methods were presented with which it is possible to estimate values at locations or time points for which concrete measurements are not available. In less-developed countries in particular, where resources to perform concrete measurements in a spatial or temporal narrow net of measurements are often not available, such interpolation methods are necessary. The introduced temporal, spatial, and spatio-temporal interpolation methods are also available for common open-source software tools or can be easily implemented.

However, when generating data by imputation or interpolation, a bias between imputed or interpolated data and real but unknown values can be observed. According to [LH14], the quality of interpolated data values are inter alia dependent on the density of the sampling sites. In section 7.3, methods were demonstrated with which it is possible to calculate the temporal, spatial, and spatio-temporal density. Density can be further used as a criterion to estimate the quality of the generated data and related decision support.

Summarizing, methods for data preprocessing were introduced that can be used in a SDSS system in the described LL approach.

During the research and development cycle in an LL, methods must be selected by incorporating the availability of the method and available algorithms in the used software tools and IT systems, the computational load of the used method, statistical requirements to which the method must fit, the characteristics of the dataset, and the accuracy of the method.

8 LLinES application and a Living Lab in El Salvador

Results of the last chapters were partly used for the LLinES project. Within the LLinES project, initial contacts between participating stakeholders were made at scientific conferences in 2012. A core group consisting of members of the *Instituto Nacional de Salud El Salvador*, *University of El Salvador*, *University Koblenz-Landau* and the *CSIR Meraka Institute* discussed in informal meetings possible research fields and solutions for the case of CKDu in El Salvador. As a result of the discussion, the concept LL was selected. The core group was extended with members of the *Embassy of El Salvador in Germany*. The final group of stakeholders and their experience are listed in table 8.1.

Project members from El Salvador had previously worked in the CKD-affected region Bajo Lempa with local farmer communities and had experience in this area. They had contacts with local farmers and grassroots cooperatives, such as *ACUDESBAL* and *Asociacion Mangle*, and were accepted by local politicians and community members. Local community members and stakeholders have shown that they were willing to participate in the project. Additionally, a minimum of an existing IT and working infrastructure was present in the region, for example, electrical power in the community, WiFi access, and meeting and working rooms in the community center. A small health center *Oscar Arnulfo Romero*, which was operated by the *Instituto Nacional de Salud El Salvador*, was also present in the community. Therefore, the region was selected as a pilot region for an LL.

In 2014, an application at the *German Federal Ministry of Education and Research* was successful, and the LLinES project was funded. The aim of the application was to build an infrastructure for the operation of an LL in El Salvador. The LL was designed in a way that it could be continued autonomously by the local project partners after the set-up phase. In addition, the transferability of the approaches and solutions developed within the LL to other regions was also a primary objective. The original work plan was divided into the following

steps: preparatory measures, interdisciplinary analysis of the current situation, development of a specification sheet for a selected risk mitigation strategy, implementation in existing resources in the LL, application of the risk mitigation strategy, and review and adaptation of workflow optimization in the LL. Milestones should ensure that the established LL was operational even without follow-up projects after the end of funding.

Funding included travel expenses for stakeholder meetings in El Salvador, promotional materials, and surveys to investigate existing infrastructure in the pilot region.

In 2015, two stakeholder meetings took place in the Bajo Lempa region. Several meetings with local stakeholders and community members were held. Figure 8.1 shows an example of a stakeholder meeting in the Bajo Lempa region.

The core group was finally extended with the local agricultural community *ACUDESBAL*

Figure 8.1: Stakeholder meeting in the Bajo Lempa region (source: Mäggi Hieber Ruiz).

and the *Asociacion Mangle*, a "grassroots community organization that works to strengthen capacities, build skills, and advance agricultural practices to improve the quality of life of the population in the Bay of Jiquilisco, El Salvador" [Man24].

Memoranda of understanding in which the participation in an open community approach was agreed were signed by the stakeholders. According to the open community approach, the group was still open for additional stakeholders and participants. The selection of these stakeholders made it possible to combine expertise from the fields of medicine, computer sci-

Table 8.1: Selected LL stakeholders in the LLinES application

Stakeholder	**Description and Expertise**
Instituto Nacional de Salud El Salvador	access to medical facilities
(*National Health Institute of El Salvador*)	scientific medical expertise
	Data about prevalence of CKD
	trusted stakeholder
	contact to local community
University of El Salvador	expertise in surveys
	physical sensors
Embassy of El Salvador in Germany	legal and political issues
University Koblenz-Landau	expertise in mathematical modeling
	ecotoxicology, teacher education, data science
Asociacion Mangle	grassroots community organization
	advocacy and citizen participation
	agroecological production
ACUDESBAL	local farmer community and association
CSIR Meraka	LL expertise
	ICT in rural areas

ence, ecotoxicology, environmental sciences, mathematical modeling, and educational science. Local stakeholders helped to gain access to the local infrastructure and communities and were representatives and accepted leaders of the local farming community. Additionally, they had knowledge about the situation in the pilot community, for example, about social and cultural restrictions or existing and applicable risk mitigation strategies and materials. They were also involved in the structure of the local community.

Communication with local community members showed a high level of motivation to participate in the LL and related data sampling approach. Motivation was explained by the grief caused by CKDu in the community and the high case and death numbers. The involvement of local and well known stakeholders like the *Instituto Nacional de Salud El Salvador* and *University of El Salvador* helped to increase confidence in the whole project group, which extended also to foreign stakeholders.

According to the classification made by [LWN12], a user-driven LL method was used. Representatives of the local stakeholders were also involved in the process of adjusting scientific

research goals and selection of scientific methods. However, due to the project specification in type of a fixed project application, where concrete milestones are written down, it was not possible to adjust the main project aims together with the local stakeholders, despite the open community approach.

During the project phase a brochure was designed and printed to inform local community members about the project, goals, methods, and ways to participate in the project. According to the open community approach, it was intended to use open-source software. Software analyses were conducted which resulted in the use of ODK.

In 2016, gang violence increased in El Salvador and in the Bajo Lempa. Gang violence was so high that travel warnings were issued for El Salvador and visits by stakeholders in the Bajo Lempa region were not possible anymore. Preparatory work was still continued, for example, surveys were designed.

However, due to the structural problems in the pilot region the project had to be terminated in 2018. Up to now, contact in the stakeholder group is still maintained; however, the political agenda in the health sector in El Salvador has changed away from CKDu.

III | Initial knowledge base for risk mitigation strategies

9 Risk mitigation strategies in an agrochemical related LL

The following chapters deal with the requirements for risk mitigation strategies that can be used in the described LL context and shows how an initial knowledge base for possible risk mitigation strategies related to spatial mathematical modeling and ecotoxicology that is usable at the beginning of a research and development cycle within the LL can be derived from the requirements.

9.1 Requirements for risk mitigation strategies in less-developed countries in an LL approach

According to the findings in chapter 2, a high level of exposure to toxic pesticides can be observed in less-developed countries. Additionally, the medical treatment of people affected by a disease is often insufficient. Therefore, risk mitigation strategies must be developed that help to reduce the exposure to these substances and the related impact or help to decrease the risk of suffering from a negative impact by improving the medical knowledge and treatment. In general, the situation in less-developed countries in terms of possible risk mitigation strategies is different than compared with that of developed countries.
Summarizing the results from chapter 2, the application of possible risk mitigation strategies for an LL approach in a less-developed country must be enacted under the following parameters:

- **(R3.1)** must be performable with limited resources, i.e., money, medical treatment, and infrastructure,
- **(R3.2)** must deal with illiteracy and low education,
- **(R3.3)** must fit to tropical climatic conditions,

- **(R3.4)** must fit to cultural, religious, and social conditions, and
- **(R3.5)** must have an obvious benefit.

The proposed risk mitigation strategies must fit to the situation in less-developed countries, which are different than those in developed countries. For example, due to limited financial resources, they must be constructed in a way that they can be also afforded by community members in poor rural areas.

The proposed risk mitigation strategies must be adapted to the available resources, e.g., in an area without a dialysis machine, it does not make sense to propose dialysis as a risk mitigation strategy. Therefore, spatial and temporal data about the availability of required resources might be implemented in the decision support process. Adaptation to available resources might be also in the sense that low-tech equipment is used, for example, as demonstrated in the VRT approach described in chapter 10.

Additionally, infrastructural limitations must be considered in the selection process of possible risk mitigation strategies. Digital education materials are, for example, not appropriate if digital devices are not available or if digital materials cannot be delivered to the user due to weak ICT infrastructure.

The proposed risk mitigation strategies should be also adapted to the users' knowledge, education, and skills. For an illiterate person, image- or sound-based risk materials might be more appropriate than text-based materials. Additionally, educational materials must be adapted to the users' existing knowledge.

Additionally, the behavior of the community members and existing cultural, religious, or social restrictions must be considered, e.g., if community members do not trust a special risk mitigation method due to cultural or religious beliefs, then this strategy will not be applied by the community members – even if it is proposed by an SDSS.

Most cases of CKDu are observed in less-developed countries with tropical or subtropical climate. Therefore, the proposed risk mitigation strategies must be also applicable in areas with high temperatures.

These requirements show how important community members are in the innovation cycle. A scientist does not necessarily know how hard it is to work the whole day with protective clothing in a tropical climate or which risk mitigation strategies are not applicable due to social restrictions. Only community members themselves have this experience and information; other LL stakeholders can only assist them with the available risk mitigation tools.

As described in section 5.3.1, in an LL approach risk mitigation strategies as well as the SDSS tool selecting the most appropriate risk mitigation strategies are developed in a research and development cycle in collaboration between the stakeholders and community members living in the LL. They must work actively in the development process, inter alia, by delivering personal data in a crowdsourcing approach. In order to motivate them to participate in the described LL approach, a winning situation for the community members must be achieved, which means that through the risk mitigation strategies a real benefit must be obvious for the participating community members.

9.2 Possible risk mitigation strategies

In general, the development of effective risk mitigation strategies in the described LL approach is a process through which stakeholders deliver possible risk mitigation strategies and tools from their own disciplines and a related pool of risk mitigation strategies to the community. They are further applied in the LL and evaluated and improved in a research and development cycle. Therefore, the type of risk mitigation strategies offered in an LL approach is strongly dependent on the involved stakeholders and might change over time.

In the following chapter, possible risk mitigation strategies are listed. There are many other conceivable risk mitigation strategies; the described risk mitigation strategies are the most obvious risk mitigation strategies related to GIS and SDSS. As the innovation process in the LL approach is an open innovation cycle in which it is possible to implement ideas of possible risk mitigation strategies, and because there is the possibility that during the innovation cycle new ideas about pest mitigation strategies are born, this list is never complete and can be updated during the innovation process.

Possible concrete risk mitigation strategies derived from the author's expertise in ecotoxicology and mathematical modeling and fulfilling the mentioned requirements are visualized in figure 9.1.

As the risk caused by pesticides is mainly related to toxicity and exposure, minimizing exposure and toxicity are the main components in the proposed risk mitigation approach. However, healthcare also plays a role in mitigating the risk – even if exposure to toxic pesticides has happened and a negative health effect can be observed.

Risk mitigation strategies listed in figure 9.1 can be regarded as an initial list that a stakeholder

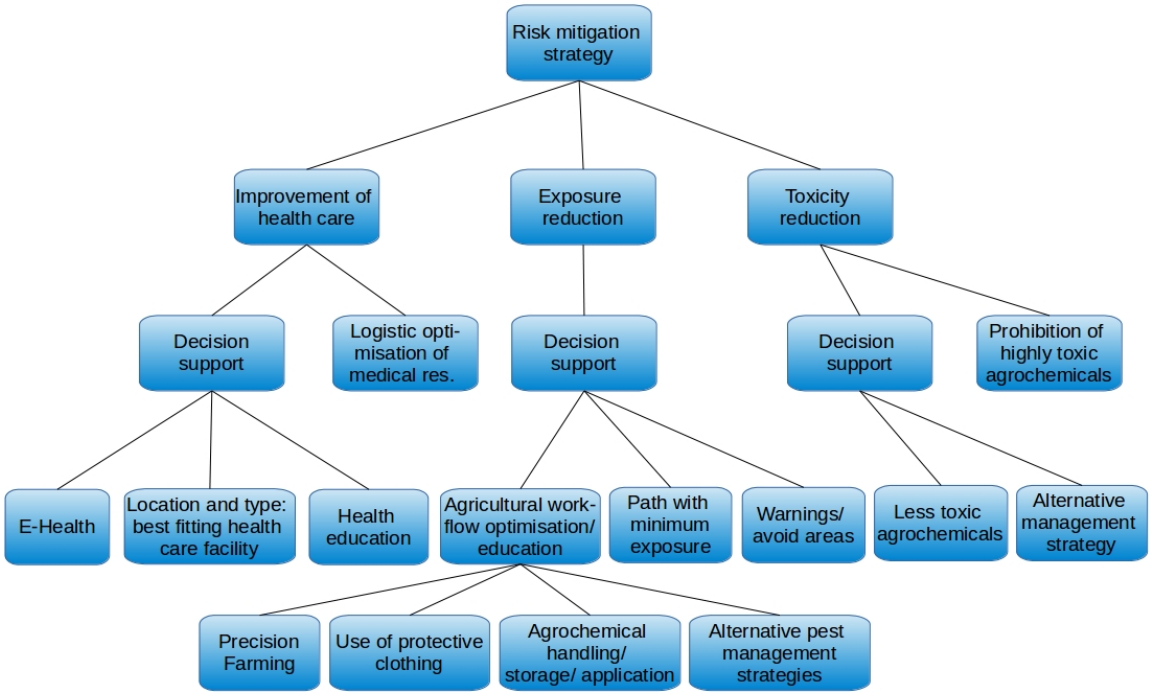

Figure 9.1: Possible risk mitigation strategies fitting to the described LL approach (figure generated with *R*).

from the disciplines ecotoxicology and mathematical modeling can offer to the pool of possible risk mitigation strategies at the start of an LL approach related to pesticides. The list of possible risk mitigation strategies in figure 9.1 is non-exhaustive and was derived from the results of chapter 2 together with the author's expert knowledge.

In this manner, all involved stakeholders might deliver tools and possible risk mitigation strategies from their own discipline and experience. During the operation of an LL and the innovation cycle, the offered risk mitigation strategies are applied, evaluated, and then removed or improved according to their usability and effectiveness.

The following chapters highlight possible low-cost risk mitigation strategies from the field of ecotoxicology and mathematical modeling in a more technical sense.

10 The adaptation of precision farming and VRT for less-developed countries

10.1 Introduction: precision farming and VRT

A system usable in an SDSS with the potential to reduce exposure to agrochemicals, especially when working with GIS and remote sensing methods, is called precision farming. A special concept of precision farming is called VRT. [GAT+11] define VRT as "a method of applying varying rates of inputs in appropriate zones throughout a field" [GAT+11, p. 2]. In contrast to traditional agriculture, where inputs like pesticides or fertilizers are applied at more or less predefined rates, the framework of VRT is used to adapt application rates to the actual requirements of the plants within an agricultural area or farm in order to increase the input efficiency [TB12].

There are two principal VRT methods for detecting the physiological status or requirements of plants: sensor-based or map-based VRT. In the map-based approach, the requirements of plants are determined by parameters visualized in maps, e.g., to determine the required nutrients according to the soil properties, whereby information about soil properties are available in a map. These maps are called prescription maps.

Within the sensor-based approach, demand is determined on the fly via a sensor that measures indirect parameters which determine the input demand and an actuator which for example adjusts application rates [CCC11]. Both approaches are connected with each other, as information visualized in a map must be previously measured, mostly with a sensor. In a tighter definition, the sensor-based approach is meant to be used when the necessary parameter is measured and the results are directly used in nearly real time to adjust the application rate. The visualization of data measured with a sensor is skipped.

[GAT+11] describe the map-based workflow with the following steps: The current position of the farmer or tractor is measured via GPS, and coordinates are sent to a processing unit or

computer with a running GIS. Then the input demand at the current position is calculated by using information from the prescription map and the actual location. Finally, application rate information is sent to the farmer or tractor to adjust the application rates. For the sensor-based approach, a GPS sensor is not necessary. Information measured by a sensor are sent to a computing unit, located with the farmer or on the tractor that computes the suggested application rates and sends this information directly back to the farmer or the tractor. Computing of the application rate must be performed in nearly real time to apply the determined rate at the measured location.

For the map-based approach a GPS sensor and WiFi connection between farmer or tractor and computing unit in a data center is required. On the other hand, sensors and computing units located at the farmer are expensive and every farmer must buy such a computing unit. For the map-based approach, different parameters describing crop health must be incorporated. In one approach, remotely sensed data gained by multispectral sensors are used to calculate a vegetation index describing the crop health and crop coverage. Non-optimal crop health can be attributed to one or more parameters not fulfilling the requirements of the plant [Pla01]. *Liebig's Law of the Minimum* states that the growth of plants is limited by the scarcest resource [Lie95]. Limiting resources can be missing nutrients and moisture, suboptimal pH, salinity, water supply, or competition with or damage by pests [Pla01]. With the knowledge of the actual limiting factor and the spatial variability of the crop health, it is possible to determine spatially tailored application rates of the limiting substance.

By comparing the actual crop physiology with situations in the past in which an ideal crop physiology is connected with a known input amount, it is possible to calculate parameters usable for tailored crop and input management. [BRP+01] describe a method that uses a vegetation index called *Normalized Difference Vegetation Index* (NDVI) and a crop model usable for tailored input management with satisfying results. [KSAKB21] review several approaches in which spectral images and related vegetation indices were connected with a machine-learning method called *support vector machine* to determine the site or plant specific needed nitrogen amount.

In practice, multispectral images are needed to calculate the NDVI. They are derived from satellite images (map-based approach) or from sensors located on the tractor. The second approach is more expensive since sensors, an automated application rate calculator, and an application-rate mixer on the tractor are necessary, resulting in related expenses. Therefore, it might be not appropriate for most farmers in less-developed countries. In contrast to most

freely available multispectral satellite images, the resolution and, therefore, the spatial quality is much higher. Additionally, the temporal quality of data is much higher than in the satellite-based approach, as the crop health data is used in nearly real time.

In the satellite-based approach there is a time shift between the date for which the crop health was calculated and the date when the data is used, resulting in a lower temporal quality (chapter 7.3). The time shift can be explained by the fact that satellite images must be processed before they are usable. Therefore, physiological parameters of the crop on which the calculated application rates are based are not up to date and a change in crop physiology might follow during the dates of crop health detection and the application date. Another advantage of the sensor-based approach is that data for crop health can be estimated at the desired date. The date for taking the satellite images is determined by the orbit of the satellite.

10.2 Vegetation indices derived by remote sensing

In literature there are several vegetation indices described that are derived by remote sensing, like NDVI or *Soil-adjusted Vegetation Index* (SAVI). The most common used is the NDVI [Hue88]. To estimate crop health or coverage, additional data, e.g., data about soil properties, are not necessary. Because of the common use and existing data over time and the ease of calculation, for which only free available satellite images are necessary, the NDVI is further used in an approach for a possible adaptation of a risk mitigation strategy for less-developed countries.

With the NDVI, something like the greenness of an area can be estimated by using multispectral reflectance data of the crop. A strong correlation between the NDVI and several parameters like green area coverage, photosynthesis activity, and plant productivity is observed [CR97]. [BMM18] found a strong correlation between the chlorophyll content in green leaves and the nitrogen supply.

Multispectral reflectance data of the crop can be obtained by multispectral radiometers. In practice, they can be located on a tractor, airplane, drone, or satellite.

Definition 10.2.1 (NDVI) *The NDVI is defined as:*

$$NDVI = \frac{r_{NIR} - r_{RED}}{r_{NIR} + r_{RED}} \tag{10.1}$$

whereby r_{NIR} and r_{RED} are the surface reflectance rate of the near-infrared and the red spectrum. [CR97]

As the NDVI changes over time and space, it also has a spatial and temporal dimension:

$$NDVI(\omega, t) = \frac{r_{NIR}(\omega, t) - r_{RED}(\omega, t)}{r_{NIR}(\omega, t) + r_{RED}(\omega, t)} \tag{10.2}$$

In equation 10.2, t represents the time point at which the multispectral image was taken, and ω the location of the plant. To calculate the NDVI, multispectral images of the vegetation surface are used. They are taken with a multispectral sensor for two spectral ranges, the near infrared band with a wavelength about $800nm$ and the red band with a wavelength about $600nm$. With these multispectral images, it is possible to analyze the reflection rate of light within the described spectral ranges from the earth surface, in the case of an agricultural area, from the plants and their leaves [CR97]. In general, healthy plants with green leaves have a low reflection (or high absorption) rate for light within the red spectral band, meaning the reflection rate of light within the near infrared light is high (absorption rate is low). This circumstance is characteristic for green vegetation. A schematic representation of typical reflection rates is visualized in figure 10.1.

The NDVI is a dimensionless parameter with possible values between -1 and 1, where high positive values are a result of green vegetation with high chlorophyll content and values near to 0 or negative show areas without green leaves [JT14]. For example, using the reflectance values from figure 10.1 and definition 10.2, NDVI values of $NDVI = \frac{0.5-0.1}{0.5+0.1} = 0.67$ for the green plant and $NDVI = \frac{0.4-0.3}{0.4+0.3} = 0.14$ for the plant with brown leaves are obtained.
Through this characteristic, it is possible to detect areas covered with vegetation using multispectral satellite images in a remote sensing approach. With a higher chlorophyll content, the reflection of the near infrared light increases. A high chlorophyll content indicates that the plant is productive and healthy [AN15a]. Within this approach, it is not possible to analyze whether the coverage with green leaves is in general low or if there is a high coverage with a low chlorophyll content. However, by comparing NDVI time series data, it is possible to analyze if there is a deviation against the normal state; therefore, it is possible to make assumptions about the general crop health [CR97].

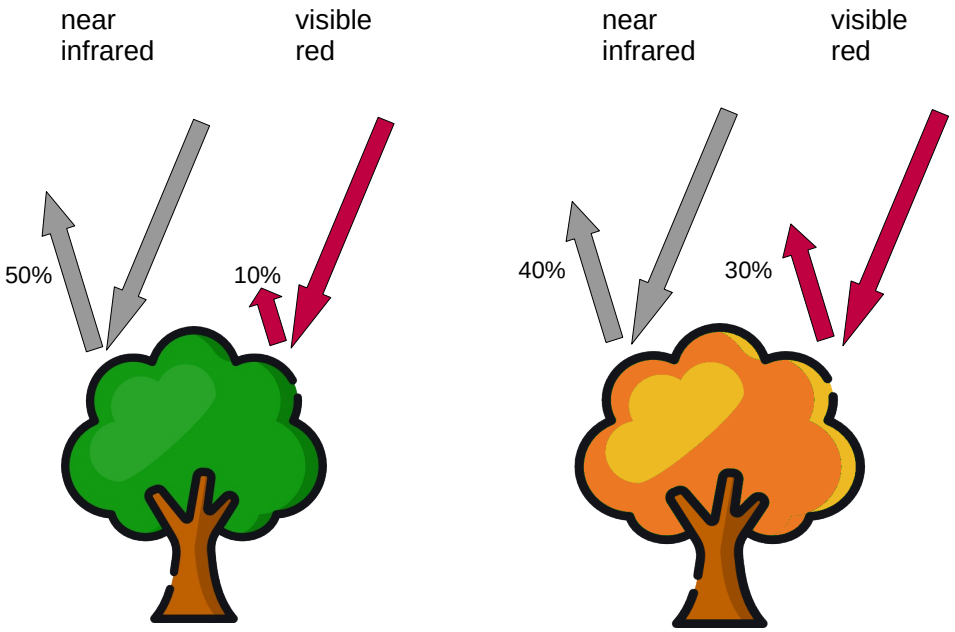

Figure 10.1: Typical reflectance rates of healthy green and brown vegetation (figure generated with *LibreOffice Draw*; image source `https://www.flaticon.com/free-icons/tree`, *Tree icons created by Freepik - Flaticon*).

10.3 Remote sensing-based VRT as a risk mitigation strategy in less-developed countries

In general, inputs in agriculture, like fertilizers or pesticides, are applied with the rule of the middle: an average application rate is applied over the whole field without considering spatial patterns of pest pressure or nutrient supply [LTH+20]. Through the implementation of VRT, it is possible to adjust the input of agrochemicals to the actual needed amount. In the best case, the amount applied to the crops and the exposure of applicators can be minimized.

Because of the mentioned exposure and toxicity mitigation effects resulting from the use of the proposed VRT system, it can be considered as a risk mitigation strategy that is implementable in the described LL approach. To use it in countries characterized as less-developed countries, some adaptations must be made in contrast to the systems that are described in the literature and that are appropriate for use in industrialized agriculture.

In the literature, several use cases are described – most of which are for use in industrial-

ized agriculture – in which tractors and high-tech equipment are used and the necessary IT infrastructure is available [WRL+14, SAB+13, RECO15]. As the agricultural workflow in less-developed countries is different than that in developed countries, such as decreased use of machines or tractors and limited IT infrastructure, the described approach must be adapted to the situation in less-developed countries.

[GT20] describe the economics of the precision farming approach. As used in industrialized agriculture, the approach is connected with investment for the upgrade of the agricultural equipment and for data generation, such as investment for sensors, aerial pictures, or soil property data. As mentioned earlier, the approach described in this thesis deals explicit with less-developed countries and the inherent characteristics, such as under-developed agriculture and limited monetary resources. Therefore, the approach must be adopted to less-developed countries. With respect to requirement R3.1 (Section 9.1), data generation methods that could be adopted for a VRT approach might be:

- the use of low-cost or free available satellite data, e.g., *Sentinel Online* [Age15b],
- the use of low-cost technical sensors, e.g., soil properties with mobile phone [ADCG16],
- the use of humans as sensors in a crowdsourcing approach, e.g., pest observation and crop health,
- the use of low-cost equipment for aerial pictures, e.g., drones [DDVG+19], and
- automated algorithms for application rate determination with open-source software.

These measures aim of reducing the expenses for data generation needed to estimate the tailored application rates. However, free available satellite images have a lower resolution than satellite images for commercial use. The lower resolution of the satellite images also means that the resolution for the generated application-rate recommendation maps is lower than those intended for commercial use. The use of human sensors, for example, in a citizen-science approach, incorporates the subjectivity of human beings. Sampling and monitoring methods must be standardized.

[GAT+11] and [JŠZB15] describe VRT approaches usable in industrialized agriculture. This approach is visualized in figure 10.2. In this approach, the modern ICT and agricultural equipment used is associated with high investment costs for the farmers, which might often not be affordable for farmers in less-developed countries.

Adaptations to the described approach to the situation in less-developed countries must be

in the field of reducing investment costs (requirement R3.1). Some of these adaptations for use in less-developed countries might be:

- making use of available systems for data delivery (e.g., mobile phone, paper-based data delivery) instead of visualization units on tractors,
- using manual mixing instead of an automated mixing unit, and
- proposing alternative pest management strategies or less-toxic pesticides.

Small-scale farmers often do not use machines like tractors; a lot of work is done by hand [OO15, AOOWKS10]. In these cases, it is not possible to implement a system with high-tech equipment, as visualized in figure 10.2. Figure 10.2 shows an example of the setup for the sensor-based approach. A sensor is located on the tractor, and information from the sensor is sent to a controller and processor unit to determine application rates on the fly. This information is next sent to a variable rate drive, which adjusts the nozzles in a way that the desired application rates are released to the plants. A low-cost approach with a *Raspberry Pi* as a computer unit and two cameras is described in [SVCCL23].

The setup of a tractor for the map-based approach is different. Instead of a sensor or camera, a GPS unit on the tractor and a WiFi connection to a server, on which the prescription maps are stored, are needed. The actual location of the tractor is measured via a GPS sensor on the tractor, and the actual location is then sent to the server unit. Within the server unit, information about proposed application rates is gained by the actual location of the tractor and the information stored in the prescription maps. The actual application rate is then sent to the processor unit, which adjusts the nozzles of the spraying equipment. Information about actual crop health is gained through remote sensing with multispectral images from satellites.

Even if tractors are available, such a system is connected with high investment costs, and therefore, it might not be appropriate for less-developed countries. Instead of a computing unit connected with the nozzles and an automated mixing unit, a system for less-developed countries must be adapted to use existing equipment for data delivery or to adjust application rates manually.

A possible solution for the use in a hand-application approach might be that the needed data, e.g., a map of the farmer's fields together with data about proposed application rates and other information needed and described in section 10.3, are delivered to the user via existing equipment, e.g., on a cell or mobile phone with graphical output, on a desktop computer with a printer, or, if no digital source is available, on paper.

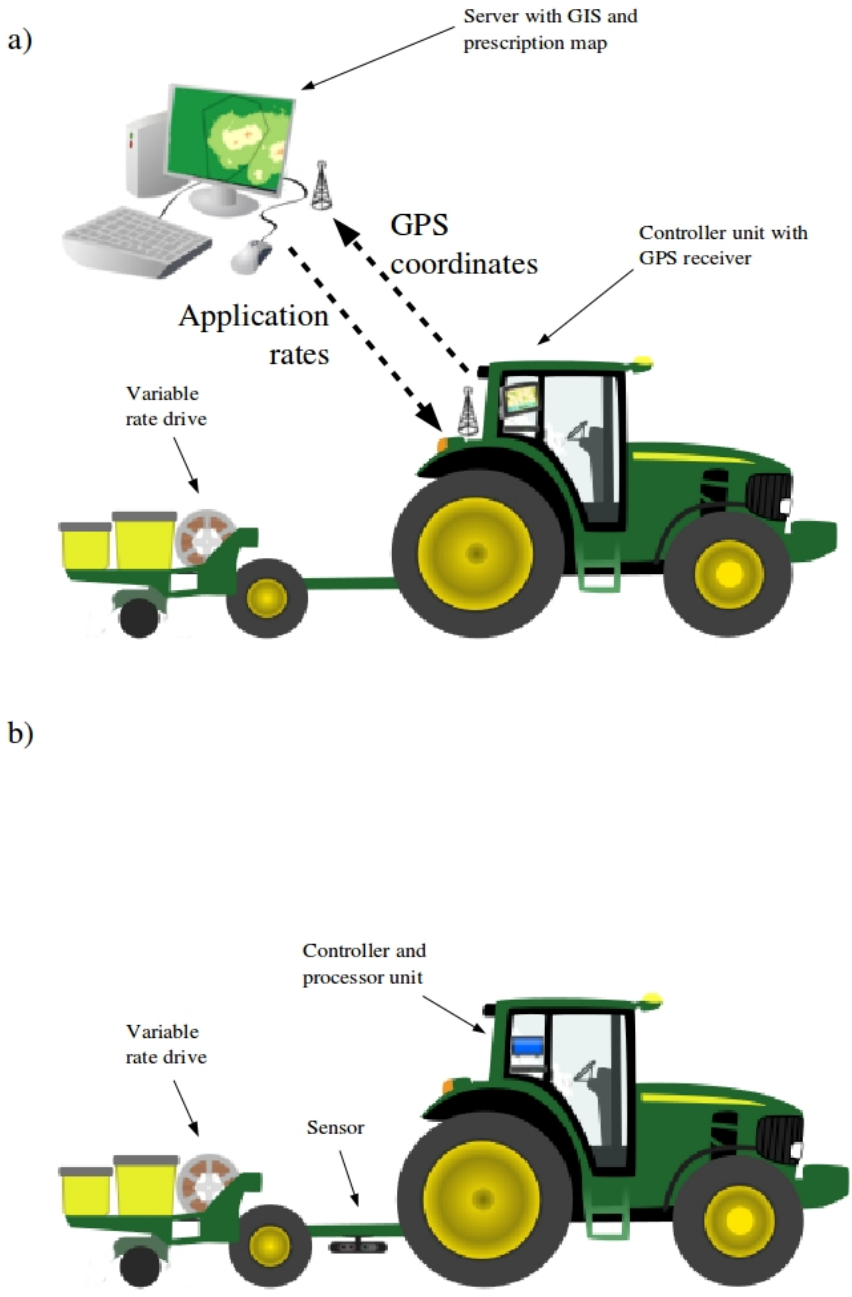

Figure 10.2: An example of high-tech VRT-systems on a tractor. Figure a) represents the map-based approach, and figure b) the sensor-based approach according to [GAT+11] (figure generated with *LibreOffice Draw*, image source: `https://freesvg.org/farm-tractor-with-planter-vector-graphics` *Free SVG*, `https://openclipart.org/detail/6024/cartoon-computer-and-desktop` *OpenClipart* created by *DTRave*).

Investment costs for an automated mixing unit can be saved by mixing the agrochemicals manually. There is a larger effort to adjust the application rate according to the spatial variation. Continuous variation, which is possible with an automated mixing and adjustment unit, is nearly impossible. Developing methods to adjust application rates at manual hand bumps might be an approach in the innovation cycle in the LL.

Overall, the benefit of the satellite-based approach is limited in comparison to the sensor-based approach; however, it is more likely that the satellite-based approach is affordable for people with limited financial resources.

10.4 Remote-sensed data derived from satellite images

Multispectral satellite images usable to calculate the NDVI are available for free, e.g., provided by *National Agriculture Imagery Program* (NAIP) [oAFSAF15] or the *Landsat* program [AN15b]. Commercial multispectral images, such as data from the *Advanced Very High Resolution Radiometer* (AVHRR) with a sensor located at *National Oceanic and Atmospheric Administration* (NOAA)'s *Polar Orbiting Environmental Satellites* (POES), have a higher frequency of image taking, but the data is not available for free.

Landsat is a joint program of *National Aeronautics and Space Administration* (NASA) and *United States Geological Survey* (USGS) aimed to generate satellite images of the Earth's surface. The first satellite of the program was launched in 1972; the most recent satellite within the *Landsat* program is the *Landsat 9* satellite [AN15b]. With images taken by *Landsat* satellites, it is possible to "track land use and to document land change due to climate change, urbanization, drought, wildfire, biomass changes (carbon assessments), and a host of other natural and human-caused changes" [Sur15c].

A low-cost VRT approach is demonstrated in the present chapter in which satellite images taken by *Landsat 8* are used. *Landsat 8* has sensors with which it is possible to get images from eight different spectral bands from thermal infrared up to the visible spectrum, with original resolutions of 15 $m \times 15$ m down to 100 $m \times 100$ m per pixel, depending on the sensors and the waveband.

Satellite images from two bands taken by the *Landsat 8* satellite are visualized in figure 10.3, a) represents Band 4 and b) represents Band 5. If data from *Landsat 8* is used, Bands 4 and 5 are needed to calculate the NDVI. Data is expressed in a unit called *digital number*. This approach is used because data from different sensors with different calibrations are taken, and

with this unit it is possible to combine them. The unit *digital number* can be easily converted into the usually used units *Top Of Atmosphere Radiance* or *Top Of Atmosphere Reflectance* [Sur15b].

A true-color satellite image merged of three bands representing the spectrum of the visible

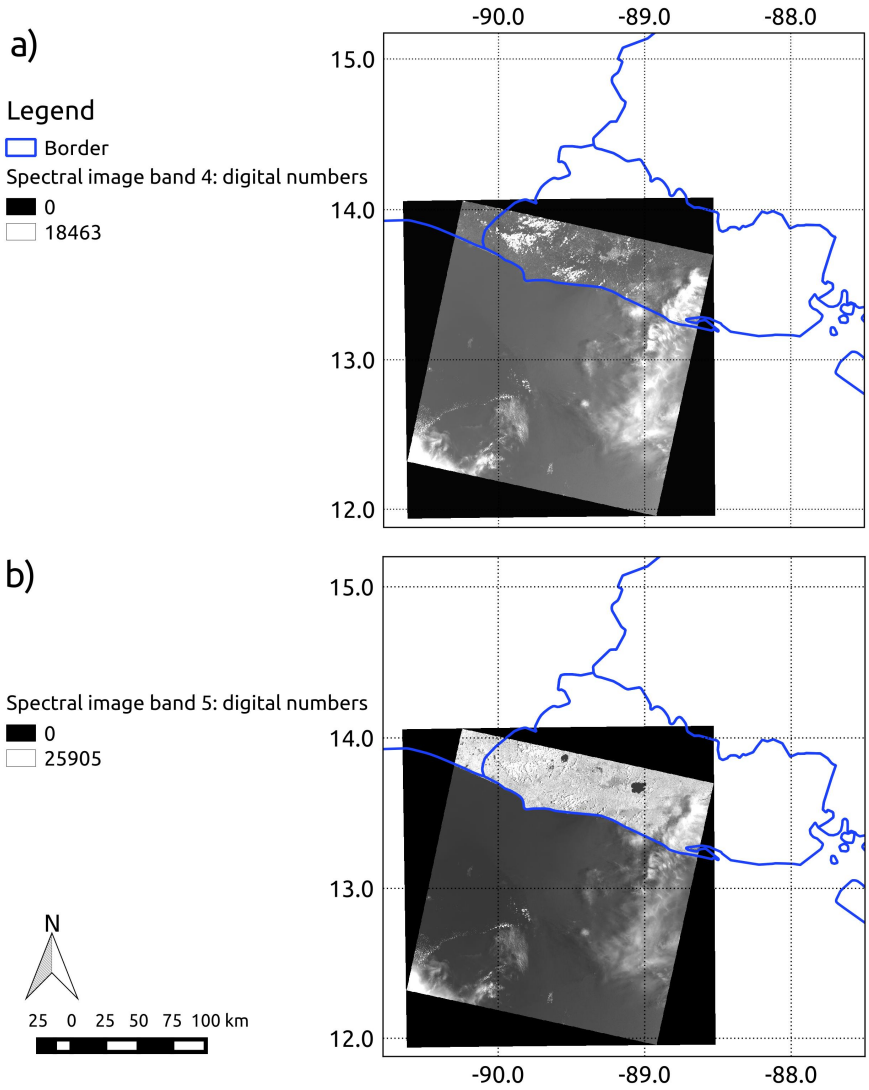

Figure 10.3: Spectral satellite images from *Landsat 8* of a part of Central America taken on August 29th 2015, a) Band 4, visible red, wavelength 0.64 - 0.67 μm, resolution 30 m b) Band 5, near-infrared, wavelength 0.85 - 0.88 μm, resolution 30 m. For better visualization, only values in the range of 0 - 98% of the maximum value are visualized. Values are expressed in the unit provided by the NASA called *digital numbers*. (figure generated with *QGIS*, Level 1 geoTIFF file, *Landsat Scene Identifier LC80190512015241LGN00*, source: [Exp15]).

light (Band 2: blue, 0.450 - 0.51 μm; Band 3: green, 0.53 - 0.59 μm; Band 4: red, 0.64 - 0.67 μm) is visualized in figure 10.4.

Landsat 8 covers the whole Earth and passes every location on the Earth's surface nearly ev-

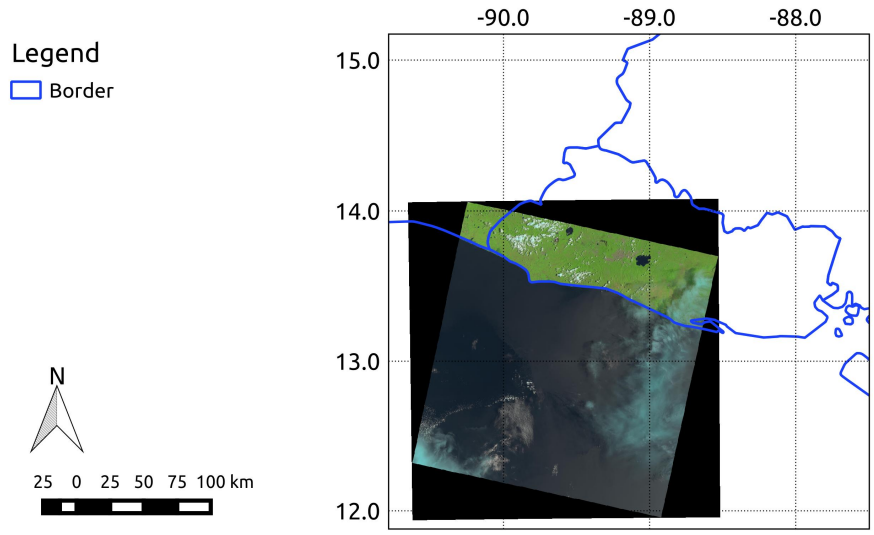

Figure 10.4: True-color satellite images from *Landsat 8* of a part of Central America, taken at August 29th 2015.
(figure generated with *QGIS*, Level 1 geoTIFF file, *Landsat Scene Identifier LC80190512015241LGN00*, source: [Exp15]).

ery 16 days. Within the *Landsat* program, a second satellite (*Landsat 7*) was still in operation with the same orbit as *Landsat 8* but shifted for eight days. Therefore, images derived from the *Landsat* program were only available with a gap of eight days [AN15b]. The quality of data estimated with the satellite-based approach is also dependent on atmospheric conditions, as clouds and aerosols are obstacles and disturb the measured reflected amount and lead to wrong assumptions [VSKD97]. However, there are different methods available to produce cloud-free images, as explained in more detail in section 10.5.

10.5 Processing of satellite images to derive vegetation indices

Depending on the source of the satellite images, it might be necessary that the used images be preprocessed to calculate vegetation indices such as the NDVI. These steps are described in the following section.

Satellite images as, e.g., provided by the NASA, must be processed for different reasons:

Data usable for calculating the NDVI must be expressed as the reflectance rate at ground. However, original data sampled by sensors located on a satellite or airplane do not represent

the reflectance rate or the radiance at ground because atmospheric conditions, such as molecular and aerosol scattering and gaseous absorption, influence the radiation before it is detected by the sensor, therefore, values sampled by airplanes or satellites describe a parameter called *top of atmosphere reflectance* or *top of atmosphere radiance* [VESJ02]. As a result, an atmospheric correction must be performed to calculate necessary indices. An additional reason for image processing is that multi-spectral data provided by NASA is, for example, expressed in a unit called *digital numbers* [Sur15b]. To calculate a vegetation index, these numbers must be converted into a value representing the *top of atmosphere radiance* or directly with a atmospheric correction into the surface reflectance.

There are different correction algorithms described for atmospheric correction. Some of these methods need additional data measured at the ground, e.g., the visibility at ground [VTD+97]. There are also methods available for which ground data is not necessary. These methods can estimate an offset of radiance decrease by recognizing the darkest pixel and calculating this offset back from the original data. These correction methods have a lower correction quality in general than the methods using ground data [DT17].

Which correction method is used depends on the available data. If ground data is available, it makes sense to use the more complex methods because of the higher quality of the results. Different algorithms for satellite data processing, e.g., to calculate the reflection at ground (with atmospheric correction) or to *top of atmosphere radiance*, are available in open-source GIS methods implemented that use original uncorrected and unprocessed *Landsat* satellite images expressed in *digital units* , e.g., in *GRASS GIS 7* (figure 10.3). An algorithm to convert information expressed as *digital numbers* without atmospheric correction into *top of atmosphere radiance* or *top of atmosphere reflectance* or with implemented atmospheric correction algorithm into a value describing the *surface reflectance* or *surface radiance* (command *i.landsat.toar*) are implemented. Within this command, several versions of a correction algorithm called *dark-object subtraction* (DOS) are implemented [Tea15]. For these correction algorithms, additional data is not needed, and the offset radiance or reflectance is determined by the assumption that in every satellite image with a high probability pixels are in with zero reflectance. The reflectance of the darkest pixel is, therefore, regarded as the offset radiance or reflectance [Cha88]. As mentioned earlier, these correction algorithms do not need additional ground data, which may not be always available especially in less-developed countries. Therefore, such correction algorithms can be used everywhere, keeping in mind the relative low quality.

Another *GRASS GIS* command called *i.atcorr* calculates the *surface reflectance* or *surface radiance* from *top of atmosphere reflectance* or *top of atmosphere radiance*. With this command, an atmospheric correction based on an algorithm called *Second Simulation of the Satellite Signal in the Solar Spectrum* (6S) [Tea15] is implemented. To perform the 6S algorithm, data measured at the ground is necessary, inter alia, the visibility at the ground, temperature, pressure or aerosol concentration [VTD+97].

Topographical and sun illumination correction are additional steps that must be performed. Topography has an influence on the parameters measured by multispectral sensors located on satellites. The terrain slope itself has an influence on the measured parameter, as a surface directed toward the sun appears brighter in contrast to an area directed away from the sun. Additionally, the position of the sun compared to the Earth's surface, as well as the location and geometry of the sensor and the surface influences the reflection rate in which the reflection rate influenced by the mentioned parameters, is also dependent on the terrain slope [HF83]. To minimize these influences and get information comparable with data for other areas and different topology, such a correction must be performed [Tei86]. However, some satellite data available for free are topographical corrected [Sur15a], such as the *Landsat Level 1* products used in this section. There are open-source methods available for topographical uncorrected data, such as the *GRASS GIS* command *i.topo.corr*, which performs a topographical correction on satellite images describing the reflectance, with an implemented sun illumination terrain model. Data needed for this correction method are a digital elevation model (DEM), and parameters describing the position of the sun against the Earth's surface, such as the solar zenith and the solar azimuth [Tea15]. Data describing a DEM are partly available for free in a GIS usable format, e.g., data from NASA's *Shuttle Radar Topography Mission* (SRTM) are published as open data and available for areas with a latitude between -60 °and + 60 °for free [NAS15b] or the *Global Digital Elevation Model* (GDEM) collected by NASAs *Advanced Spaceborne Thermal Emission and Reflection Radiometer* (ASTER) [NAS15a]. Data about the sun position at the date when the image was taken is either delivered by the image providers, e.g., for *Landsat* images in a metadata file or can be calculated with an open-source GIS, e.g., with the command *r.sunmask* in *GRASS GIS*[Tea15].

An additional image processing step might be the interpolation of values for areas covered with clouds. Clouds act as an obstacle; therefore, reflection or radiance values obtained from cloudy areas represent the reflection or radiance from the cloud surface. To obtain values for areas covered with clouds, it is necessary to identify clouds and to remove values from cloudy

areas or to mark them as *not available*.

In *GRASS GIS*, a command called *i.landsat.acca* is implemented that performs an automated cloud cover assessment based on an algorithm described in [IBGA06] for *Landsat* images. Clouds are identified by the fact that they are colder than the surface area. Therefore, a band representing the temperature is needed, as implemented in the *Landsat Level 1* datasets. The described *GRASS GIS* module must be adjusted for *Landsat 8* datasets, and band numbers are selected according to the file name pattern in which the band number is stored. However, band numbers related to their wavelength have changed between *Landsat 7* and *Landsat 8* data, and a manual band number selection is not implemented in the described module [Sur15b].

The module can be used principally for every multispectral satellite image dataset that contains a thermal band but must be adjusted to the described requirements for the file names. In each *Landsat 8* dataset, an image called quality assessment (QA) band is implemented. With this QA image, it is, inter alia, possible to identify areas covered with clouds or ice using a slightly modified algorithm in contrast to that described by [IBGA06] or [Sur15b]. Figure 10.5 represents a map of the QA band in which the included information are summarized into three categories: *clouds*, *maybe clouds*, and *no clouds*.

As clouds act as an obstacle, the reflectance values and therefore also related NDVI values of cloudy areas must be removed for interpolation tasks. There are different methods available to restore information lost through clouds, e.g., by time series of images and using *Fourier analysis* [RMV00] or by iterative interpolation methods [JS10]. However, besides standard interpolation methods, such modules are not implemented in the common open-source GIS.

Another method to obtain spatial NDVI information lost by cloud cover could be by combining multi-temporal images of one scene and estimating the NDVI of areas covered by clouds, e.g., by interpolation. However, at least two NDVI values from different time points are needed to interpolate the information. It would be also possible to use a spline interpolation for every cell with missing values to obtain the missing values.

In the following figure 10.6, a process pipeline for the described automated processing approach for NDVI maps is visualized.

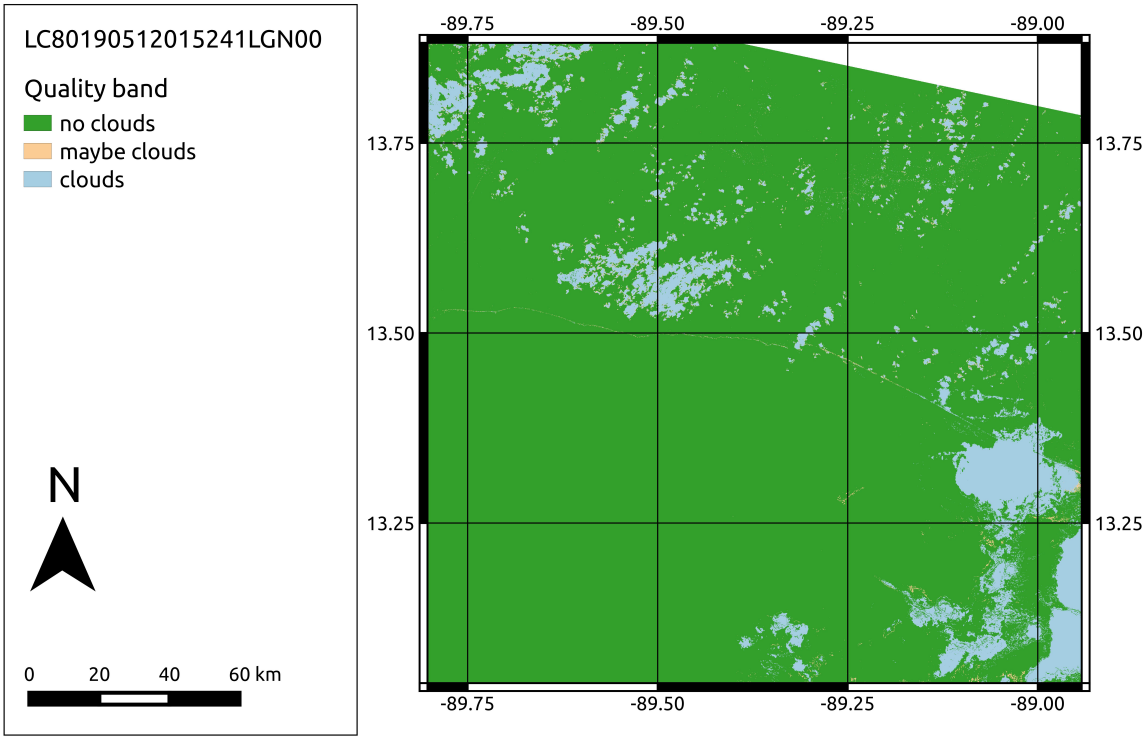

Figure 10.5: QA band categorized into areas with *clouds*, *maybe clouds*, and *no clouds*, taken on August 29th 2015.
(figure generated with *QGIS*, Level 1 geoTIFF file, *Landsat Scene Identifier LC80190512015241LGN00*, source: [Exp15]).

10.6 From vegetation indices to spatial decision support

With the steps described in the last section, information about the spatial distribution of the NDVI and related maps can be generated with open-source software. To gain concrete information about crop health, NDVI values of plants in the field must be compared with reference NDVI values of plants for which the crop health is known.

For each crop, a time series of NDVI values under optimal conditions can be generated in trials. In figure 10.7, the NDVI time series of corn plants under fertilized and unfertilized conditions are illustrated through which the change of the NDVI within a cropping season and between fertilized and unfertilized plants can be observed. By taking the NDVI values for the different growth stages of a crop into account, it is possible to compile a crop calendar. By comparing the NDVI values under optimal conditions at a given time point with NDVI values in the field, it can be estimated if a problem with crop health exists and countermeasures

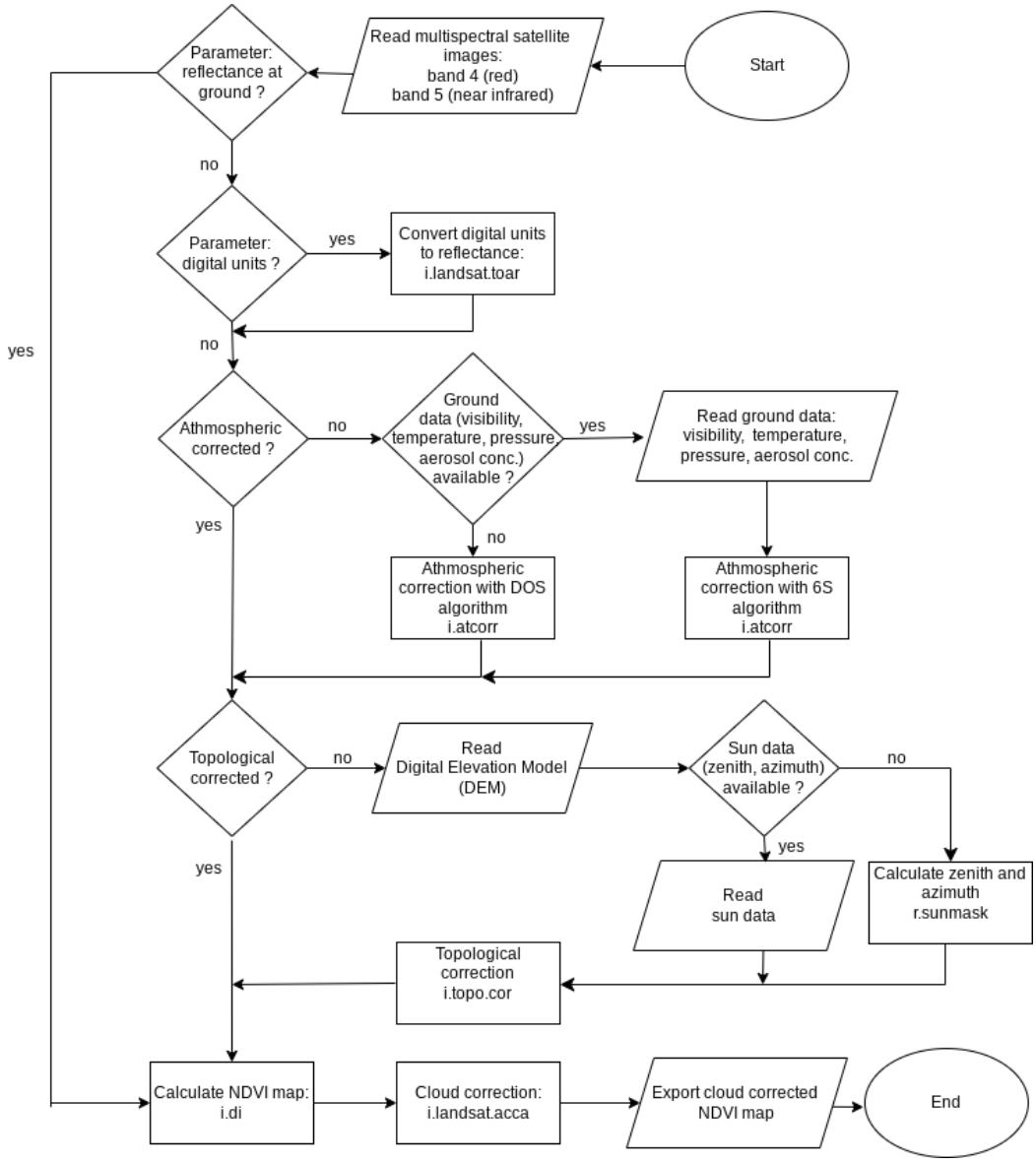

Figure 10.6: Flowchart of the creation process for NDVI maps with *GRASS GIS* based on satellite images (figure generated with *draw.io*).

must be taken [PO14].

However, plant physiology and the related crop calendar for optimal growth might also have a spatial and temporal dimension, as crop growth is also connected to environmental parameters, such as sunshine duration, precipitation etc., that have a temporal and spatial dimension. There might also be a shift in the crop calendar between different seasons and years. Therefore, it might be necessary to generate such a crop calendar for plants under optimal conditions, e.g., in a trial field located within the LL in an area with similar conditions for which the method should be used, and to adjust this calendar to the environmental weather

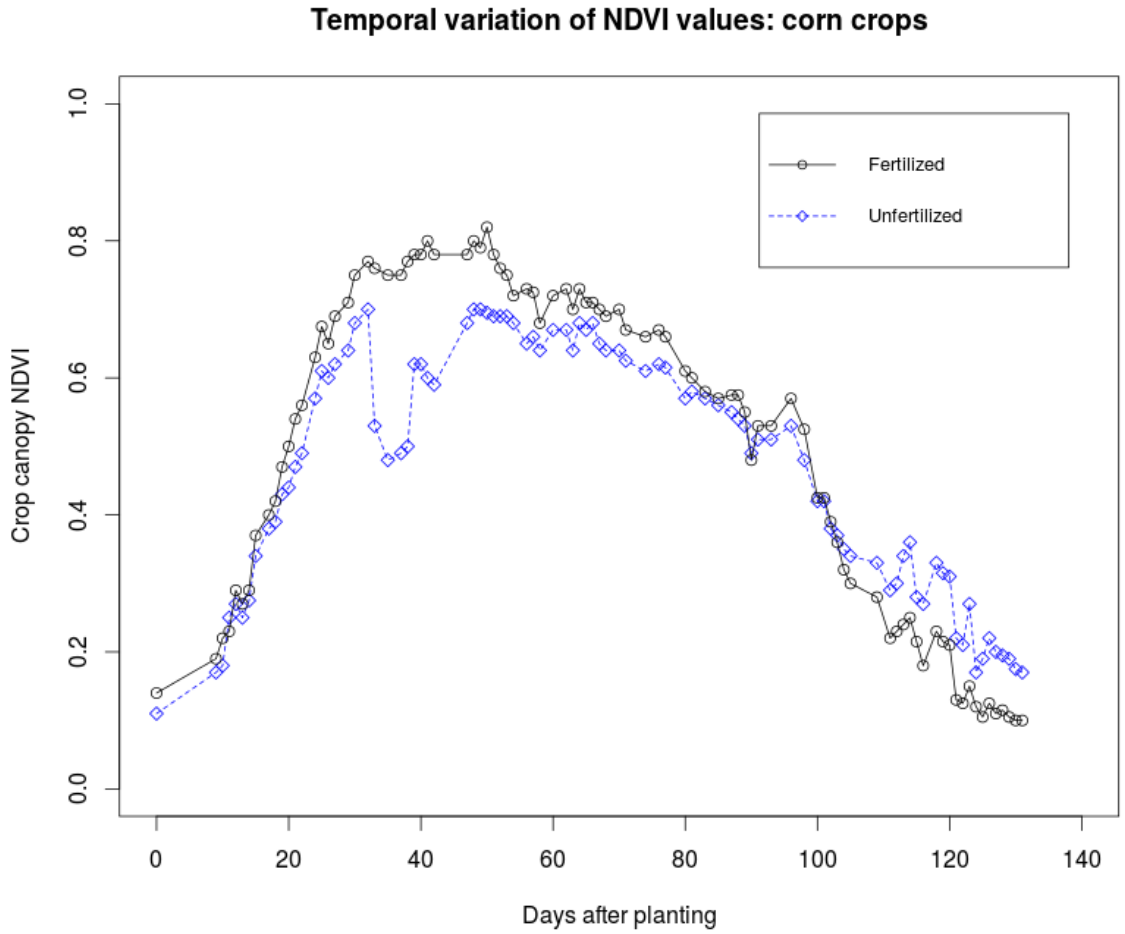

Figure 10.7: Temporal change of NDVI values for corn plants within a crop season and between fertilized and unfertilized plants (figure generated with *R*, source: [Zha15]).

conditions responsible for temporal shifts in plant physiology. It must be kept in mind that each crop type has its own characteristic growth stages and that this is the reason that a specific trial must be performed for each crop type.

It is also possible that the calibration is done by experts or stakeholders in the LL, e.g., that biologists identify areas within a field with an optimal crop stage, to determine the GPS coordinates of this area and identify this area and related NDVI values in the processed NDVI maps. However, experience and expert knowledge is needed to identify whether a crop is healthy.

Overall, the actual NDVI value $NDVI(\omega, t)$ for each location ω with agricultural crops at time t can be compared with the optimal NDVI value $\text{NDVI}_{opt}(A, crop, t)$, whereby A is subset of Ω, a set of locations with the same environmental characteristics responsible for plant growth, such as at the location for which spatial decision support is needed.

For each cell or pixel of the NDVI map, a value called $\Delta\text{NDVI}(\omega, t)$, describing the deviation between the actual and expected NDVI value under optimal conditions can be calculated:

$$\Delta\text{NDVI}(\omega, t) = NDVI(\omega, t) - \text{NDVI}_{opt}(A, crop, t) \tag{10.3}$$

However, to use the appropriate NDVI_{opt} value, it is necessary to have information about which crop type is present at each location and to which climatic region the regarded location belongs. Existing land-use data can be used as well as generated land-use data through, e.g., a citizen-science approach, with the local farmers as citizen scientists delivering information about the spatial distribution of their fields and the type of crop. Where the necessary IT infrastructure is available, the citizen-science approach (chapter 6) can be used, for example, with a mobile app through which farmers can identify the location of their fields, mark them in the app on a digital device, and connect them with the attributes needed, such as the type of the planted crop or treatment in the past. An example of an open-source web app to record pesticide application patterns in a citizen-science approach is highlighted in chapter 12.

If the treatment of a field and the environmental parameters responsible for plant growth together with related NDVI values from a same time point in the cropping season are known from the past, it is also possible to compare the actual NDVI values of a location with the NDVI values in the past and gain information about the actual grade of the plant physiology. In figure 10.8 NDVI, values calculated with images from a *Landsat 8* scene over El Salvador, taken on August 29, 2015, are visualized together with a land use map representing areas planted with sugarcane. Figure 10.9 shows a detail of the same map with a higher zoom level. By analyzing $\Delta\text{NDVI}(\omega, t)$, statements about crop health can be made for every location ω to inform the farmer which areas in the field must be observed with more attention or on which areas a treatment is not necessary.

To use this information for decision support, such as suggestion of agrochemical application rates in relation to the location, more information is needed, e.g., which is the factor responsible for non-optimal crop health. According to [MS13], non-optimal crop health can be attributed to several factors, e.g., a lack of water or light, nutrients, pH value, or damage or competition from a pest. The importance of identifying the limiting factor in order to increase the plant physiology was first described by the German chemist *Justus von Liebig* in his most notable work *"Die grundsätze der agricultur-chemie mit rücksicht auf die in England angestellten untersuchungen"* [vL55], published in the middle of the 19th century. The growth of plants is limited by the resource that is proportionally the most limited. Growth cannot be increased by adding additional resources despite the limitations. In the situation of equal

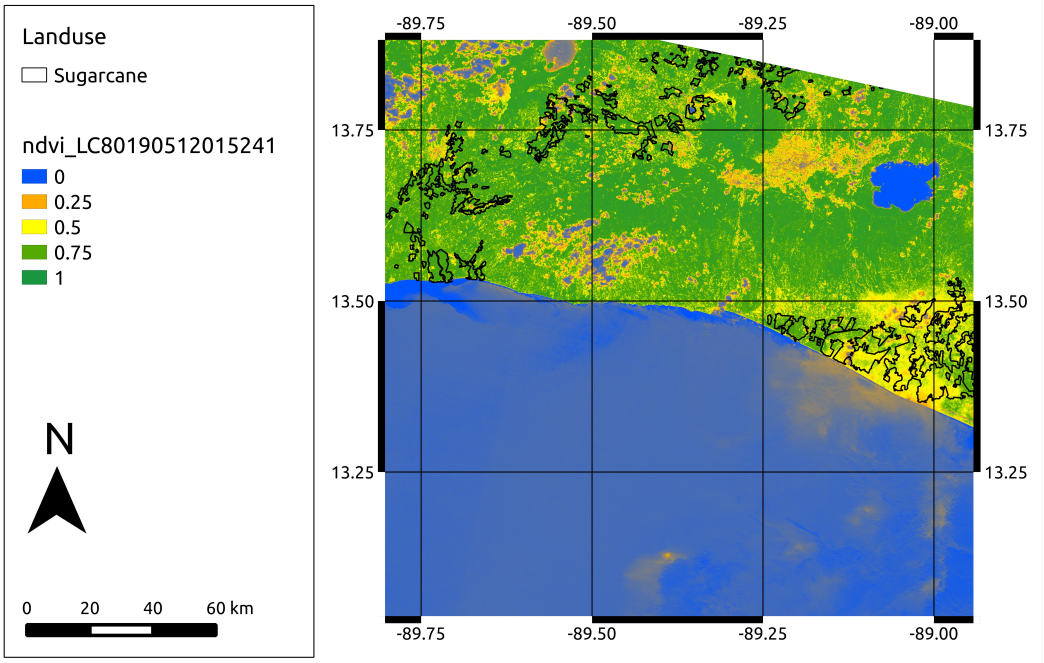

Figure 10.8: NDVI values calculated for a scene over El Salvador and a land-use map representing areas planted with sugarcane (figure generated with *QGIS*, Level 1 geoTIFF file, *Landsat Scene Identifier LC80190512015241LGN00*, source: [Exp15]; land-use map: *Instituto Nacional de Salud El Salvador*).

supply of atmospheric resources such as water or sun, growth is directly related to the supply of nutrients. *Liebig's law* is related to the supply of nutrients. However, for the described VRT approach, it is extended to all factors responsible for plant growth or, conversely, responsible for diminished plant growth. To adjust the agrochemical input to the needed amount, it is necessary to identify the factor responsible for the non-optimal growth.

To identify the limiting factor, sensors or expert knowledge are needed. The expert knowledge of educated farmers can be used, e.g., to detect pests that might be responsible for the non-optimal crop health or to analyze whether precipitation was enough for the crop-type of plants or if pest pressure is too high for the regarded crop. These factors are more or less visible and can be analyzed without a sensor; however, expert knowledge is needed and must be considered in the stakeholder selection process in the LL (section 5.3.2.2).

Other limiting factors, such as insufficient nutrient supply or an unsuitable pH value, can be only measured with sensors. Due to the characteristics of less-developed countries with limited financial resources, the sensors needed to identify the limiting factor must be in the range of low-cost sensors and usable with equipment, available in the LL.

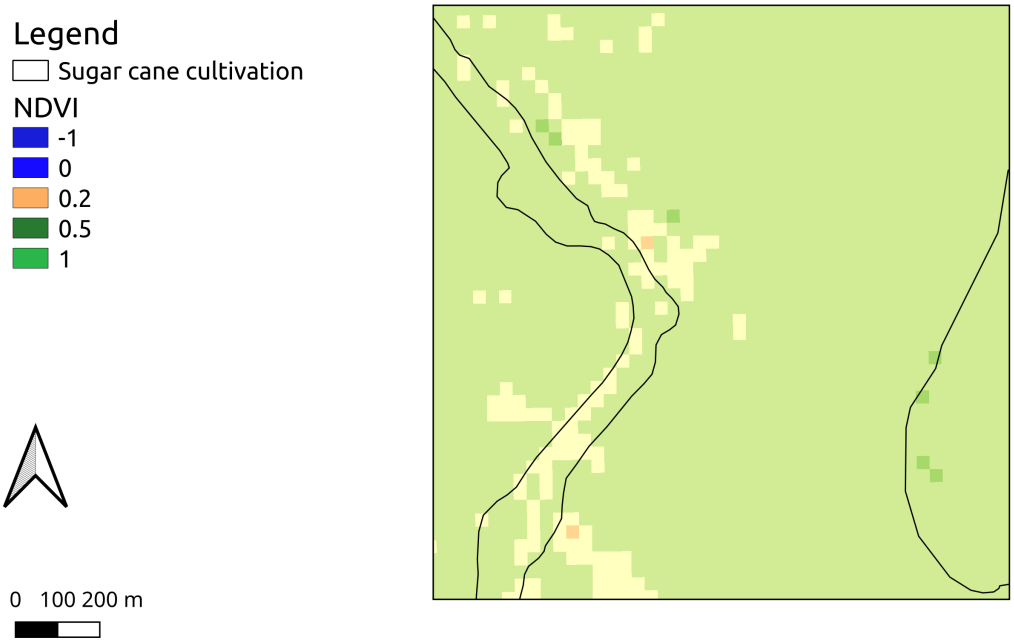

Figure 10.9: NDVI map for a scene over El Salvador and a land-use map representing areas planted with sugarcane with a high zoom level (figure generated with *QGIS*, Level 1 geoTIFF file, *Landsat Scene Identifier LC80190512015241LGN00*, source: [Exp15]; land-use map: *Instituto Nacional de Salud El Salvador*).

With modern machine learning methods, such as a *support vector machine*, it is possible to determine the site or plant-specific nutrient amount needed based on multispectral images [KSAKB21].

Precipitation can be measured on a daily basis, e.g., by installing a low-cost rain gauge as proposed in [Sen12]. In terms of a low-cost solution, such a rain gauge can be created by using an empty bottle from which the base has been removed and whose surface area and volume are known. By emptying the rain gauge at a given time interval and recording the volume, it is possible to estimate the precipitation rate and compare it with the needed amount of the plant.

The detection of pests must be done manually, e.g., by observing the plants and fields manually to detect possible pests, herbs, or fungi. Expert knowledge is needed to identify the pests, which might be a problem in less-developed countries where farmers sometimes have no or only little agricultural education. If IT infrastructure is available, tools for pest detection for electronic devices can be used, as proposed in a presentation by [RE14]. [PPG14] and [WC13] have described tools for mobile apps for pest identification. They are based on the

principle that images of the pest or affected plant material or a description of the pest are taken and sent to a specialist to identify the pest. However, such tools are not available for all pests and are not currently fully implemented.

[GRLRM+13] have described a smartphone app with which it is possible to estimate the soil color according to the *Mansell system*. It is possible to derive soil characteristics such as the humus content of a soil from soil color. The color is detected from soil images taken by a smartphone.

However, some of the parameters might not be measurable with a low-cost solution. Additionally, in some cases, it might not be possible to identify the factor responsible for spatial diversity in plant growth, or the reason could be multifactorial.

To give decision support in terms of proposed application rates, type of substance, and time point of application, stakeholders from the agricultural sector must be involved in the LL approach. Their expert knowledge can be implemented in an SDSS, for example, an FLC as described in section 16.2.4. Additionally, machine-learning methods with supervised-learning methods can be used to estimate the site-specific nutrient amount, as demonstrated by [KSAKB21].

An SDSS can help to adjust the field management to the needs of the farmer; however, additional information is needed. Personal health history decides which agrochemicals should be avoided according to the health status of the farmer, e.g., agrochemicals harming the kidneys should be avoided by farmers with CKD. The finances and time available for field management can help to determine the most appropriate method, e.g., hand weeding, a time-intensive method related with low expenses, can be recommended if it can be done by the owner of the field himself or for farmers with low monetary resources but a lot of free manpower.

Additionally, meteorological and behavioral data from the farmer is needed to determine the appropriate date of application, e.g., not during rainfall or when people are located next to the field on which the agrochemicals are applied. If it is not possible to give concrete application rates to the farmer, the farmer should be informed about fields with a abnormal plant physiology.

10.7 Interim conclusion

In the last section, the framework of VRT and precision farming was adapted to the conditions in less-developed countries. Data used for the described low-cost approach as well as necessary software tools for data processing and generation are available for free, and needed geospatial methods for satellite image processing are implemented in existing open-source GIS. This takes requirement **R3.1** (section 9.1) into account in that people living in less-developed countries must deal with limited financial resources.

Image processing to derive the NDVI was demonstrated with methods implemented in *GRASS GIS*. IT infrastructure needed for image processing, storage, and delivery should be available in an LL approach as discussed in chapter 17.

Additionally, the equipment required to use the information about appropriate application rates can be adapted to the characteristics in less-developed countries, e.g., it is not necessary to have a tractor with a integrated computing unit, an automated adjustable nozzles, and an application rate mixer, for which farmers with limited economic resources would hardly have use. Regarding **R3.2**, people must be trained to use the equipment and related software tools, for example a web app or a GIS tool.

The spatial and temporal quality of information within the described approach is lower than in approaches with high-tech solutions, e.g., with a sensor on the tractor it is possible to use nearly real-time information without the given limitations when using satellite images where a time shift between image recording and delivery can be observed, resulting in a low temporal quality. Additionally, the spatial resolution is much lower than in the tractor-located approach, analysis of the NDVI of a single plant is nearly impossible (figure 10.9).

Despite the lower temporal and spatial information quality, the information obtained by the NDVI is helpful in the sense of recognizing deviation patterns in crop health with the ability to initiate appropriate countermeasures or have an overview of the plant conditions in a given area and, therefore, adjust application rates to the real conditions of plants in contrast to a given regular predefined application rate. This helps to fulfill **R3.5** by having the obvious benefit that the used pesticide amount and therefore money can be reduced.

Implementing temporal data, such as weather conditions and daily human behavior, helps to minimize the exposure of farm workers, for example, by avoiding inefficient applications caused by unfitting weather conditions. To implement such models in an SDSS, expert knowledge in the field of agriculture must be available in the LL.

Problems might occur through bad quality or missing input data, e.g., it is hard to find land-use data with a high spatial quality of information. Figure 10.8 shows an example of a map with land-use data obtained from an agency in El Salvador. A spatial shift of the land use data is obvious and may be related to incorrect geospatial processing techniques. The example in figure 10.8 demonstrates that there are regions within the areas planted with sugarcane that have completely different NDVI values despite the fact that, according to the land cover map, all the marked areas should have an NDVI indicating that similar vegetation is there. Reasons for the deviation might be that the used land-cover map is from a different date than the satellite images and, in the meantime, other crops are used with different NDVI values, or that crops were destroyed in some areas due to pests pressure or other stressors.

However, the last-mile problem must be considered. The information that must be delivered to the farmer has a spatial component, and the different application rates must be visualized, e.g., on a map. Therefore, an electronic device with a display with the ability to visualize maps is required.

The described methods and components can be connected to a working system usable and implementable in an LL. The concrete realization, evaluation, and adjustment of the methods and models is a task for the research and development cycle in the LL.

.

11 Creation of pesticide application maps

To estimate the exposure to pesticides, for example, for ecosystems or organisms in the environment, it is necessary to have information about the released amount of pesticides into the environment. In the following section, a method is described for how pesticide application maps can be created through which the amount and type of pesticides released into the environment by agriculture can be estimated.
In an unpublished diploma thesis [Rap11], a similar method to create such maps for pesticides belonging to the substance group insecticides was used. Part of this thesis is the creation of such maps for substances belonging to the groups herbicides and fungicides. The method for how to create such maps is described in more detail in the mentioned diploma thesis.
The difference between the diploma thesis and the method described in this thesis are in the software used. For the diploma thesis, proprietary software was used, and for this thesis, exclusively open-source software is used. Additionally, a adapted model for African countries described in section 11.1.3 was exclusively used in the present thesis.

11.1 Methods

11.1.1 Software used

The software used for the following steps are released under free or open-source licenses.
For this part of the thesis, several open-source GIS were used: *GRASS GIS*, *QGIS* and *SAGA*. Most of the described steps for spatio-temporal calculations and for conversions between raster and polygon files were performed with *GRASS GIS* versions 6.4.3 up to 7.0.0. In addition, *SAGA* version 2.1.0 was used for some conversion steps that could not be performed with *GRASS GIS* because of missing or unknown conversion modules. The maps presented in this thesis were laid out with *QGIS* versions 2.2 and 2.4.
Statistical analysis and related figures presented in this chapter were performed and created with the software package *R* version 3.0.2.

The office software suits *Apache OpenOffice* and *LibreOffice*, specifically, the spreadsheet and database management applications *Calc* and *Base*, were used for the preparation of raw data and for steps that had to be performed with a database management tool.
Most of the steps described in the following parts had to be repeated several times with differing parameters. Scripts were used in which rules for the use of the different parameters for spatio-temporal calculations and conversions are described. *GRASS GIS* has an integrated shell for the programming language *Python*. The scripts used for the calculation and conversion steps were written in *Python* version 2.7.9.

11.1.2 Used external maps and their modification

For different steps, a world map of the type of a GIS layer was required. Because of origin of the data for the harvested area was from the year 2000, it was decided to use a world map with the states and geographic boundaries from around the year 2000. A world map in *shapefile* format provided by the *Geo Data Portal* of the *United Nations Environment Programme* (UNEP) [Pro11] was used. The geographic boundaries refer to the year 1998. The file consists of different polygons representing the different countries, and the name of the country is stored in a related attribute table.
To use a database for the application rates with a granularity down to the states of the USA without the loss of the spatial resolution the area of the USA, the world map used should consist of the different national states. Therefore, the polygon representing the USA was selected and deleted from the UNEP *shapefile*.
For the federal states of the USA, a *shapefile* provided by the *United States Census Bureau* was used [Bur11]. Both maps were loaded into a GIS and merged with the *GRASS GIS* command *v.overlay* whereby the logical operator *OR* was selected. Figure 11.1 shows the resulting world map. To create maps visualizing the spatial distribution of the agriculture-related applied pesticide amount with the following approach, data was needed that represented the spatial distribution of areas used for crop cultivation. A raster dataset, described in the publication [MRF08], was used as a base for determining areas of agricultural crops or crop groups. The dataset can be downloaded from the internet for free and originally consists of 175 raster layers for 175 crops or crop groups, with the attributes of the raster cells representing the ratio of the cell area on which the related crop was harvested in a resolution of 5 $arcmin \times 5$ $arcmin$, resulting in grids consisting of 2160 rows, 4320 columns, and 9331200 cells. An example of such a raster layer is illustrated in figure 11.2.

Figure 11.1: Created world map (figure generated with *QGIS* and *GRASS GIS*, sources: [Pro11] and [Bur11]).

This dataset was created based on data published by the *Food and Agriculture Organization of the United Nations* (FAO), e.g., from the project *Agro-MAPS*, for the years around the year 2000 for 206 countries. If data for the year 2000 was not available in the primary used FAO data, the authors took data from years adjacent to 2000 for the creation of the raster dataset [MRF08].

As the intention of the described work was to visualize the spatial distribution of pesticide application, layers related to crops without described pesticide use were not used in this work. Additionally, some crops could not be implemented in this work because of undescribed pesticide application patterns and rates. Overall, from the original dataset described by [MRF08], 129 crop layers were used for this work. In relation to pesticide use, the most important crops were integrated in this work.

For most calculations, raster layers representing the harvested area expressed as an area value in ha or km^2 were needed. Therefore, the original raster layers with cell attributes representing the ratio of the harvested area of a crop in relation to the cell area were transformed into layers, representing the area in ha. The original layers were multiplied with a raster in which

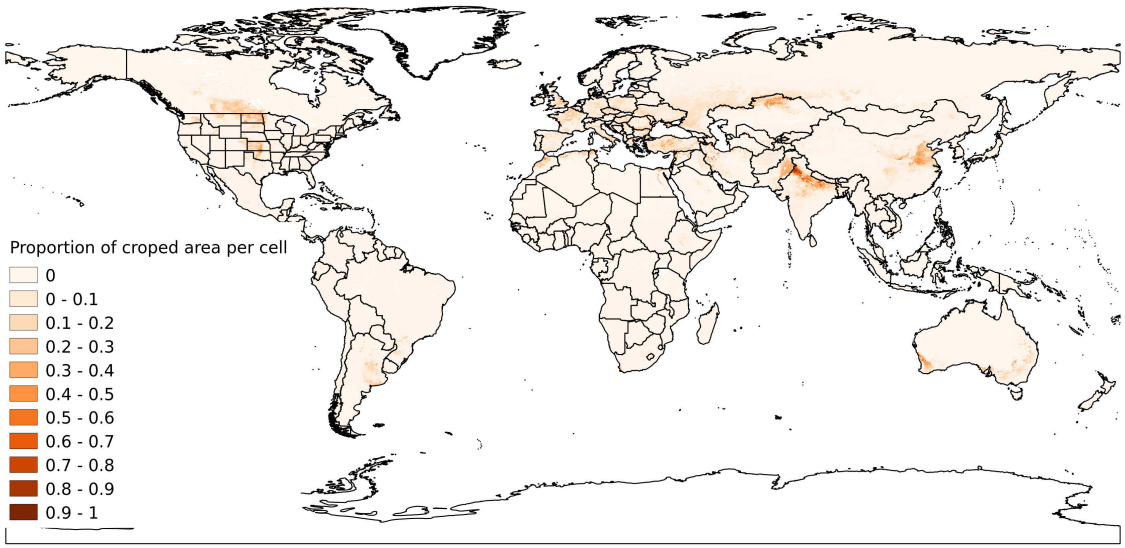

Figure 11.2: Sample map representing the proportion per cell of the area harvested with wheat (figure generated with *QGIS*, source:[MRF08]).

the cell values represent the overall cell area. For this work, a cell area grid was used that was provided by a working group of the *University Frankfurt* [Fra11], in which the cell area in *ha* was stored for the used geographical latitude / longitude projection.

The absolute crop area $a^M_{cropped\ absolute}(r, c, crop)$ for a cell in row r and column c for a *crop* can be calculated with the following equation:

$$a^M_{cropped\ absolute}(r, c, crop) = a_{cell}(r, c) \cdot q^M_{procentual}(r, c, crop) \tag{11.1}$$

whereby $a_{cell}(r, c)$ refers to the absolute cell area, stored in the cell area grid, and $q^M_{procentual}(r, c, crop)$ to the percentual crop area, as provided in the original [MRF08] dataset.

The cell area grid and the raster files in which the cell values represent the harvested area in percentages were multiplied with the *GRASS GIS* command *r.mapcalc*. This command was also used for the following calculation steps for raster files.

During the use of the [MRF08] dataset and comparison with additional data, e.g., *FAOSTAT (Food and Agriculture Organization Corporate Statistical Database)* [FD11a], it became obvious that in part, there were large deviations between the data used in the mentioned GIS

raster dataset and additional external data, e.g., statistical data published by *FAOSTAT* and other databases and reports published by national and international agencies.

To gain an overview of the deviations between the harvested area values used in the [MRF08] dataset and other, more reliable datasets, e.g., published by *FAOSTAT* [FD11a], the harvested area raster layers had to be processed to get comparable values. The data provided by [MRF08] did not contain data about the harvested area per crop and country or a cell attribute describing the name of the country to which the raster cells belong and with which it would be possible to calculate the harvested area per crop and country. As the *FAOSTAT* data [FD11a] is expressed in harvested area per crop, country, and year, the data provided by [MRF08] were also transformed in these units in order to compare the values.

The following procedure describes how to derive the harvested area per crop and country and was performed for each crop layer. The first step of the aggregation was performed by transforming the raster layers into vector files with the *SAGA* command *gridvaluestopoints* and the option *nodes* as type in which, for each crop, the raster layers were transferred into vector files consisting of points. In the attribute table of the created point vector files, the cell attribute value, in this case the absolute crop area in *ha*, was saved. Conversion from raster data into vector point data was necessary because of the characteristic of raster data, which, in contrast to vector data, only contain the values of one attribute. With vector data an attribute table is connected with the option to assign more than one attribute to an object. For the current work, the harvested area per cell and the name of the country the cell or point lies within was assigned to each point. The created vector files also contained points for which no value were available or for which the harvested area was 0. To reduce the needed data space and also the needed time for calculations, all points with no value or with a value of 0 were removed from the original vector files with the *QGIS* command *Extract by attribute*. In the next step the vector layer representing a world map with a connected attribute representing the country name of each polygon was loaded into *QGIS* and spatially joined with the created point vector files. This command is not included in *GRASS GIS*; therefore, the open-source GIS tool *SAGA* with the command *addpolygonattributestopoints* was used for this step. As a result of this step, the name of the country it is lying in was assigned to each point of the created vector file.

The aim of the following step was to sum up the attribute values representing the harvested area per cell or point in relation to the country in which it lies for each crop. As a result of this step, the harvested area per country and crop according to the data published by [MRF08]

was determined. This step was performed with the open-source database management tool *LibreOffice Base*. A copy of the database file in *dbf* format of each point *shapefile* created in the previous step for each crop was imported into the described database management tool. An *SQL*-query in the type

SELECT 'countryname', sum('cropedarea') FROM 'shapefiletabel' GROUP BY 'countryname'

was used to sum up the harvested area for each country. As this step was performed for all crops regarded in this thesis, the harvested area in relation to crop and country $a^M(crop, country)$ was obtained.

For 24 points or cells with an overall agricultural area of $26977ha$, a country name could not be assigned to due to the fact that these points, representing the center of the originally raster cells, do not lie within the borders of a polygon of the created world map, which represent countries. This may be due to different reasons, e.g., through the loss of accuracy and resolution in the conversion between polygon to raster files.

In the next step, the harvested area per crop according to the data published by [MRF08] $a^M(crop, country)$ was compared with values for the harvested area from external databases $a^{ref}(crop, country)$ for the year 2000. As the external database for countries besides the USA, a database published online by the FAO, which is called *FAOSTAT* [FD11a], was used. In this database, only crops for human consumption are implemented; the values for crops not implemented in the *FAOSTAT* database, such as food crops for animals, were not changed in the original raster files published by [MRF08]. As a reference value for the harvested area, the arithmetic mean of the harvested area values for the years between 1997 and 2003 for available crops and countries were calculated. For the correction procedure for states within the USA, harvested area values were used published in the *National Pesticide Use Database 2002* (NPUD 2002) for the year 1997 [CF]:

$$a^{ref}\begin{pmatrix} crop \\ country \end{pmatrix} = \begin{cases} a^{NPUD2002}\begin{pmatrix} crop \\ state \\ year \end{pmatrix} & \text{USA, year} = 1997 \\ \frac{1}{7}\sum_{year=1997}^{2003} a^{FAOSTAT}\begin{pmatrix} crop \\ state \\ year \end{pmatrix} & \text{, countries besides USA} \end{cases} \tag{11.2}$$

To derive the deviation between $a^{ref}(crop, country)$ and $a^{M}(crop, country)$ a value further called $\Delta a(crop, country)$ was calculated for each crop,country combination:

$$\Delta a(crop, country) = a^{ref}(crop, country) - a^{M}(crop, country) \tag{11.3}$$

If the value in the *Monfreda* files was smaller, the percentage was calculated, and the planted area was decreased, with the calculated amount in every cell where the crop was planted. For each crop layer, a set of cells C_{crop} can be defined for which the cell attribute, in this case the harvested area per cell, is greater than 0.

If the value in the external data was larger than that calculated from the *Monfreda* files, it was decided to distribute the difference between these values to all cells where, in general, planted areas from crops used in this work are located. More details are published in the diploma thesis.

$$a^{cor}\begin{pmatrix} r \\ c \\ crop \end{pmatrix} = (1 + \frac{\Delta a\begin{pmatrix} crop \\ country \end{pmatrix}}{a^{ref}\begin{pmatrix} crop \\ country \end{pmatrix}} \cdot a^{M}\begin{pmatrix} crop \\ r \\ f \end{pmatrix} \quad \forall\ crop \in C \wedge country \in D \tag{11.4}$$

In equation 11.4, C represents the set of crops and D the set of countries regarded in this step.

With the last step for each crop, a corrected raster file was created that represents the planted area per cell for the selected crops and is further used as the basis for spatial data for the used model.

The corrected raster files, a table with the implemented crops and data to correct the original files are provided in *S2* in the online *gitlab* repository under the following uniform resource locator (URL): `https://gitlab.rlp.net/jrapp1/dissertation`.

11.1.3 Data used to calculate yearly application rates per crop and country

In general, large parts of the data sources used to calculate yearly application rates were the same as those used in the diploma thesis. For states within the USA, application rates were calculated with values derived from the NPUD 2002 published for 1997 [CF]; for countries belonging to the EU, values published by *Eurostat* in the reports [otEU07] and [otEU14] were used. South Africa data, such as the yearly applied amount of pesticides in relation to substance group and crop and for the planted area, was obtained from an unpublished data source and provided by a market research institute.

For everyday crops for African countries (with the exception of South Africa), the application rates like those described in the publication [fIDU94] were used.
A detailed description of the used data and databases published by [fIDU94], [CF],[otEU14] and [otEU07] can be taken from the diploma thesis. The mentioned unpublished database for South Africa consisted of data about the applied substance amount for fungicidal, insecticidal, and herbicidal substances for 26 crops for 2009 and the harvested area of each crop.
To reduce the created data and have a better overview of the results, the following GIS-based calculations were performed with application rates not related to a single substance but with application rates for substance groups. Each substance group consists of a set of substances. Therefore, each substance was grouped into a substance group, as suggested in the classification scheme used in [otEU07]. However, in this classification scheme not all substances used in the databases are implemented, e.g., substances not allowed in the EU. These substances were classified according to the classification as suggested in [Cou11]. To derive the applied amount in relation to the substance group, all applied amounts of substances belonging to one substance group were totaled.
Data published in [fIDU94] was partly expressed as applied formulation amount and was available for Zimbabwe, Kenya, and the Ivory Coast. To derive the applied amount expressed in active ingredient (a.i.), conversion factors from free, available safety data sheets of the regarded formulations were used.
With the exception of the data source published in [otEU07], direct application rates expressed as applied amount per area, as needed for the calculations, were not available, but data were available to calculate this value. Data published in the mentioned publication described the applied amount $q(crop, substance, country)$ per crop, substance, and country or state. Additionally, with the exception of [fIDU94], the harvested area $ha(crop, country)$ of each crop was also provided in the database and was used to calculate the value for the applied substance group amount per area. To calculate this value for countries published in [fIDU94], harvested area data published by *FAOSTAT* [FD11a] was used.
With this data the application rates $p\begin{pmatrix} crop \\ country \\ substancegroup \\ year \end{pmatrix}$ expressed as a value in the format

applied amount per area can be in general calculated as follows:

$$p\begin{pmatrix} crop \\ country \\ substancegroup \\ year \end{pmatrix} = \frac{\sum\limits_{\substack{substances \in \\ substance\ group}} q\begin{pmatrix} crop \\ country \\ substance \\ year \end{pmatrix}}{a\begin{pmatrix} crop \\ country \\ year \end{pmatrix}} \tag{11.5}$$

For European countries outside the EU, the mean application rates of all EU countries; for Canada, the arithmetic mean of the application rates for states belonging to the USA were used. For countries outside Africa, the USA and Europe, the arithmetic mean of the EU and USA application rates were used.

The results of the diploma thesis for insecticides showed that the used model fits well when compared with published values for the yearly applied overall pesticide amount with the exception of African countries. In most African countries, agriculture is not as industrialized as it is in most of the rest of the world, e.g., characterized by high labor input and low yields [Mil15]. In general, due to financial limitations and market reasons, in African countries, the use of agrochemicals and pesticides are not as common as in the industrialized world [McD09]. Additionally, pesticides are mostly used for crops intended for export or so-called cash crops [NBL03].

Therefore, for African countries a different model for the calculation of the application rates was used. Application rates were adjusted in relation to the ratio of exported crops with the assumption that in the group of non-everyday crops only exported crops are treated with pesticides. For these crops, it was necessary to calculate the ratio between the exported u_{exp} and the produced amount of a crop u_{prod}. In this model, it is assumed that the ratio between the exported and the produced amount is the same as between the area in which the crops for export are and the overall area per crop and country.

To calculate the exported amount per country for all crops, two databases from *FAOSTAT* were used. In the first database, the produced amount of a crop

$$u_{prod}\begin{pmatrix} crop \\ country \\ year \end{pmatrix}$$

expressed in t or kg per country and year is published [FD11b]. In the second database, export quantities in kg or t

$$u_{exp}\begin{pmatrix} good \\ country \\ year \end{pmatrix}$$

of agricultural products are listed [FD11c]. With the help of the second database, the exported amount of each agricultural crop implemented in the described framework can be calculated. As also processed agricultural goods are exported, it was necessary to convert the amount of the exported agricultural good

$$g\begin{pmatrix} good \\ country \\ year \end{pmatrix}$$

into a value, describing the input amount of a crop which is necessary to produce the exported amount. If available, conversion factors $\gamma(crop, good)$ for an agricultural product and the weight of the output were used to calculate the input amount of the agricultural crop to produce the goods, e.g., a conversion factor between the weight of apples needed to produce one liter apple juice. The needed input value of a crop to produce the exported amount of the agricultural product can be calculated as the sum of the product between conversion factor and exported amount of a good with the following equation:

$$u_{exp}\begin{pmatrix} crop \\ country \\ year \end{pmatrix} = \sum_{\substack{goods \in \\ crop-good-list}} \gamma\begin{pmatrix} crop \\ good \end{pmatrix} \cdot g\begin{pmatrix} good \\ country \\ year \end{pmatrix} \quad with\ year = 2000 \tag{11.6}$$

For the export of the crop itself, the conversion factor is 1. However, such conversion factors are often not available in scientific literature; therefore, also non-scientific sources were used to gain such conversion factors. Such conversion factors could not be found for all products. They were not integrated in the calculation. A table of used conversion factors and their sources are listed in S5 in the online repository (`https://gitlab.rlp.net/jrapp1/dissertation`). In the next step, the ratio

$$\mu\begin{pmatrix} crop \\ country \\ year \end{pmatrix}$$

between the exported amount and the produced amount was calculated:

$$\mu \begin{pmatrix} crop \\ country \\ year \end{pmatrix} = \frac{u_{exp} \begin{pmatrix} crop \\ country \\ year \end{pmatrix}}{u_{prod} \begin{pmatrix} crop \\ country \\ year \end{pmatrix}} \quad with\ year = 2000 \tag{11.7}$$

In the used model, it is assumed that in African countries only crops produced for export are treated with agrochemicals; therefore, the application rates for African countries were corrected by multiplying the original application rates with the related ratio $\mu \begin{pmatrix} crop \\ country \\ year \end{pmatrix}$.

Summarizing, application rates $p \begin{pmatrix} crop \\ country \\ substancegroup \end{pmatrix}$ were calculated according to table A.1 in Appendix A.

This correction method has different limitations, e.g., it is assumed that the produced amount per area is the same for areas treated with agrochemicals and areas without agrochemical treatment.

The application rates per *ha* and *year* in relation to the crop, country, and substance were saved in a spreadsheet document in *LibreOffice Calc* and exported to a *character-separated values* (CSV) file. As it is intended to join the CSV table to a *shapefile* with polygons representing the different countries or states, it is necessary to have a common identifier for the countries in both the application rate table and the *shapefile* of the world map.

Next, the CSV-file and the world map *shapefile* were imported into *QGIS*. The CSV-tabel was joined to the world map *shapefile* and the *shapefile* with the existing joined file was saved as a new *shapefile* to make the merge permanent. The application rate values were following stored in the attribute table of the exported world map *shapefile*.

In the next step, global raster files representing the application rates for each substance-group-crop combination were created. Therefore, the current region and the resolution of the project was set with the *GRASS GIS* command *g.region* to obtain raster layers with the wished resolution. As it is intended to create the raster files for the whole Earth, the extent was set from $-180°$ to $+180°$ for the longitude and for the latitude from $-90°$ to $+90°$. The

resolution of the resulting raster maps was chosen according to the used area grids with a cell size of 5 $arcmin \times 5\ arcmin$; therefore, the number of columns and rows was set to 4320 and 2160.

Each column of the attribute table of the described world map *shapefile*, which represent the application rates per *ha* and *year* for one of the crop-substance-group combinations, was converted into a raster file with the *GRASS GIS* command *v.to.rast*. The cell attributes of the created raster layers represent the yearly applied amount of the regarded substance group per *ha* in relation to the crop and country.

11.1.4 Data used to validate the model

As described in section 11.1.3, data about application rates per crop, country and substance group were not available for a lot of countries. For these countries a simple model to estimate the application rates per crop and country, as explained in the last section 11.1.3, was used. Because of the bias generated with this model and in general to validate the used model and method, it is necessary to compare the obtained calculated results for the yearly applied amount per country and substance group $m_{calc}(substancegroup, crop, country)$ with values for the applied amount $m_{ref}(substancegroup, crop, country)$, as listed in the used databases. Data about the yearly applied amount per country in relation to the pesticide group was obtained from *FAOSTAT* [FD13] and for states within the USA from NPUD 2002 [CF]. The described *FAOSTAT* database consists of values reported by the countries themselves in a yearly questionnaire about the used pesticide amount separated into substance groups in agriculture. The used *FAOSTAT* database has some limitations, e.g., for some countries no values, only import, domestic production, or expressed as formulated products are available. In contrast to the procedure described in section 11.1.3, the USA was handled as one state for validation. Therefore, the overall applied pesticide amount in the USA related to the substance group was calculated by totaling the applied amount for all states. For countries belonging to the EU, an additional database was available and was provided by *Eurostat* [otEU15], the statistical office of the EU.

Data from *Eurostat* and *FAOSTAT* databases differ in small values for the same time span. It was decided to use data from *FAOSTAT* if it was available in both databases because it might be the case that the data sampling procedure differs between *FAOSTAT* and *Eurostat*. As most data was obtained from *FAOSTAT*, it was decided to use this database for countries

belonging to the EU also to have the same bias and sampling errors.

Overall, for herbicides, data for validation was available for 101 countries – 99 from the used *FAOSTAT* database [FD13], one value from the *Eurostat* database [otEU15], and one value from NPUD 2002 [CF]. For the validation of the calculated data for fungicide application, overall there were 98 values – 97 values were taken from the *FAOSTAT* database [FD13] and one value from NPUD 2002 [CF].

The available reference data $m_{ref}(substancegroup, crop, country)$ was crosschecked. If it was obvious that data for one or more countries were not valid, these data items were marked. In the next step, an online search for more valid values was performed. However, the online search showed that free available data about pesticide use per crop and country is very scarce besides the mentioned databases.

Additionally, a online research for outliers, countries with high deviation between the calculated and reference amount, was conducted. Outlier countries, for which there is evidence in the literature of generally reduced herbicide or fungicide use, have been marked with the label "reduced fungicide use" or "reduced herbicide use" in the following figures 11.7 and 11.8. References for publications where a reduced agrochemical use is mentioned are listed in the provided tables in S3 in the online repository provided in (`https://gitlab.rlp.net/jrapp1/dissertation`).

In general, if possible, the average yearly applied amount was calculated, if available, for the years between 1997 and 2003 for data obtained by *FAOSTAT* and *Eurostat* databases. As reference value for the USA, the applied amount for the year 1997 was taken from NPUD 2002. The reference values were stored in a CSV tabel.

11.1.5 Creation of application maps in GIS and validation analysis

Each application rate raster was multiplied with the related harvested area raster using the *GRASS GIS* command *r.mapcalc*.

In the resulting raster, the yearly applied pesticide amount per cell in relation to the substance group and the crop is visualized. In the next step, all applied amount rasters belonging to the same substance group were summed up by using the raster map calculator *r.mapcalc* in *GRASS GIS*.

With this step, raster maps visualizing the yearly applied pesticide amount in relation to the substance group were created. For some tasks, it is necessary to gain a value for the overall used amount in relation to the organisms the substance should act against, e.g., for herbicides

or fungicides. Therefore, it is necessary to sum up all raster files belonging to the regarded substances – again, with the command *r.mapcalc*.

To validate the calculations performed with GIS, it is necessary to compare the calculated pesticide amount per country $m_{calc}(majorgroup, country)$ with values published in reports or scientific publications $m_{ref}(majorgroup, country)$. To obtain the overall applied amount per country, the same method was used, converting the raster to a *shapefile*, spatially joining with a world map vector file, and totaling the applied amount per country, with the same software as described in section 11.1.2. As a result, a table in which for each country the calculated overall applied amount related to the substance group to which it was assigned was obtained. The calculated amounts per country were stored together with the related reference values (11.1.4) in a CSV table.

For the validation of the described calculation method, both variables were $log-transformed$, whereby the base 10 logarithm for each value and each country was calculated and used for further regression analysis. To validate the described model, linear regression models were used.

As a prerequisite to use linear regression a normal distribution of the data is necessary. This prerequisite was checked visually with a Quantile-Quantile-plot (Q-Q-plot), *R* commands *qqnorm()* and *qqline()*, with a histogram, *R* command *hist()*, and with the *Kolmogorov Smirnov* test, *R* command *ks.test()*. Data for both substance groups showed a non-normal distribution. Through the logarithmic transformation of both, the dependent and the independent variable, this prerequisite could be fulfilled. The coefficient for intercept, slope, standard errors, significance of the regression line, and other statistical values, such as R or R^2, were obtained with the *R* commands *summary(lm(y x)*. $X-Y$ plots visualized in section 11.2 and the regression line were printed with the *R* commands *plot(x, y)* and *abline(lm(y x)*.

11.2 Results

According to [FD24b] for 2000, an applied herbicide amount of 879138.94 t and fungicides with an amount of 520403.32 t were estimated globally for the agricultural sector. By applying the described models for fungicides, an overall applied amount of about 1542351.775 t, and for herbicides, an overall applied amount of about 1650192 t were calculated. This results in a global overestimation of about 2.96 times for fungicides and 1.88 times for herbicides by the

used model.

Table 11.1 gives an overview of the calculated and the reference applied amount divided into geographical regions. According to table 11.1, for fungicides (with the exception of

Table 11.1: Calculated applied fungicide and herbicide amount in *t a.i.* and reference values [FD24b] for 2000, divided into regions

Region	**Fungicides [*t a.i.*]**			**Herbicides [*t a.i.*]**		
	calculated	reference	$\frac{calculated}{reference}$	calculated	reference	$\frac{calculated}{reference}$
Europe	172460.995	196172.28	0.88	137612.316	163220.68	0.84
Asia	892284.472	178695.67	4.99	990648.232	219817.38	4.51
Oceania	13454.714	3320.67	4.05	21513.71	26088.7	0.82
North America	65674.518	23722.14	2.77	223904.753	228187.53	0.98
Latin America	267978.556	94252.3	2.84	238328.023	217439.84	1.1
Africa	130498.518	21637.78	6.03	38185.409	17927.57	2.13

European countries), the used model overestimates the applied amount in a range between 2.77 times (North American countries) and 6.03 (African countries). For European countries, the calculated amount is about 88% of the reference value. For herbicides, an underestimation of the used model can be observed for European, Oceanian, and North American countries, and for Asian, Latin American and African countries, an overestimation of the used model can be observed.

The resulting maps for the overall applied herbicide and fungicide amount are visualized in the following figures 11.3 and 11.4. Maps visualizing the applied amount per substance group and crop, per crop and per substance group are provided in S1 in the *gitlab* repository under the following URL: `https://gitlab.rlp.net/jrapp1/dissertation`

The applied amount per area is expressed in the unit *kg a.i. per cell ha* in the presented maps. The unit *per cell ha* means that the applied amount refers to the overall area located in the raster grid cell and not merely to the agricultural area located in the grid cell. This unit might provide a better visualization of the applied and present pesticide amount in the overall area located in the raster cell and not merely in the agricultural area. This might be a better risk indicator for a disease like CKD, which can be also observed in the population not directly working in the agricultural sector. The units can be easily converted by multiplying

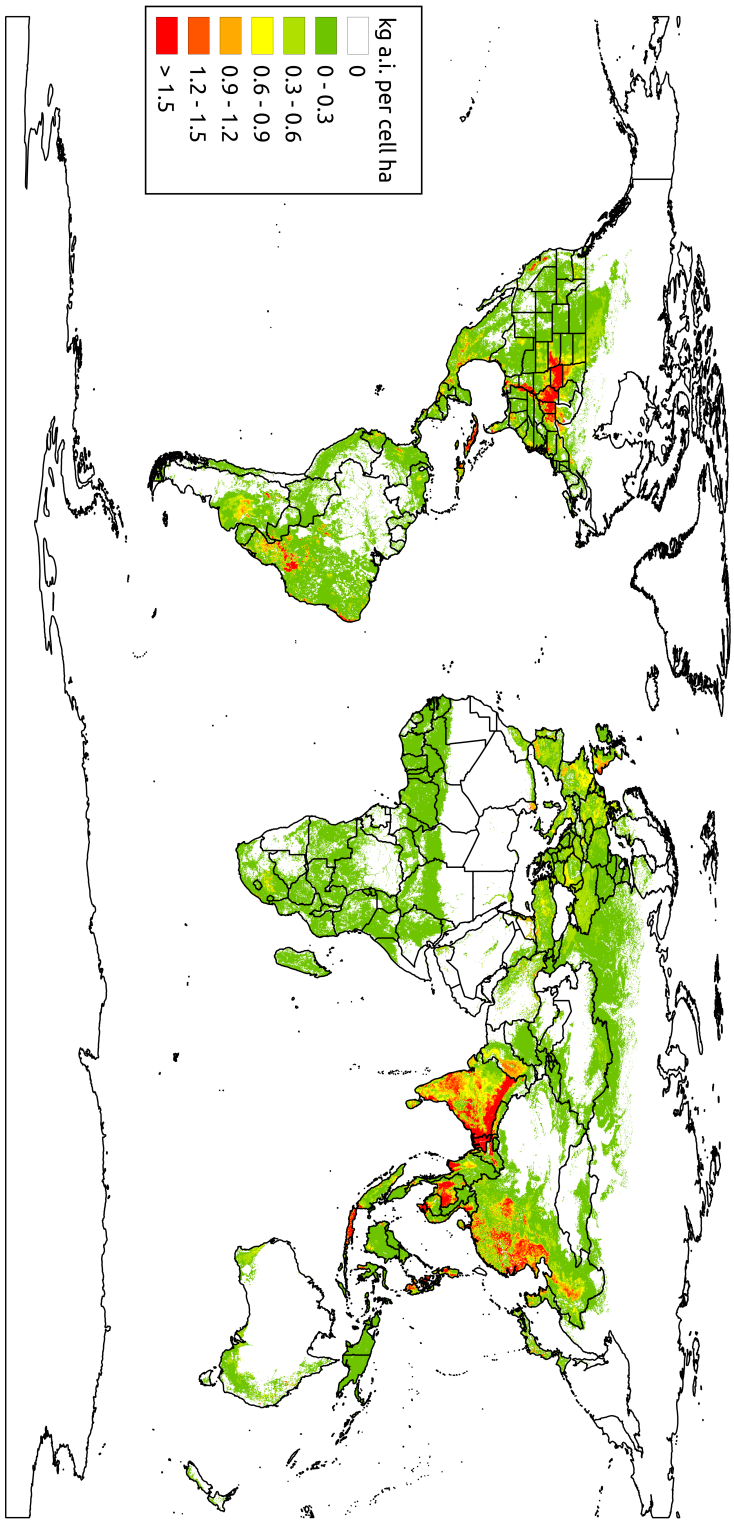

Figure 11.3: World map visualizing the applied amount [*a.i. per cell ha*] for herbicides (figure generated with *QGIS*).

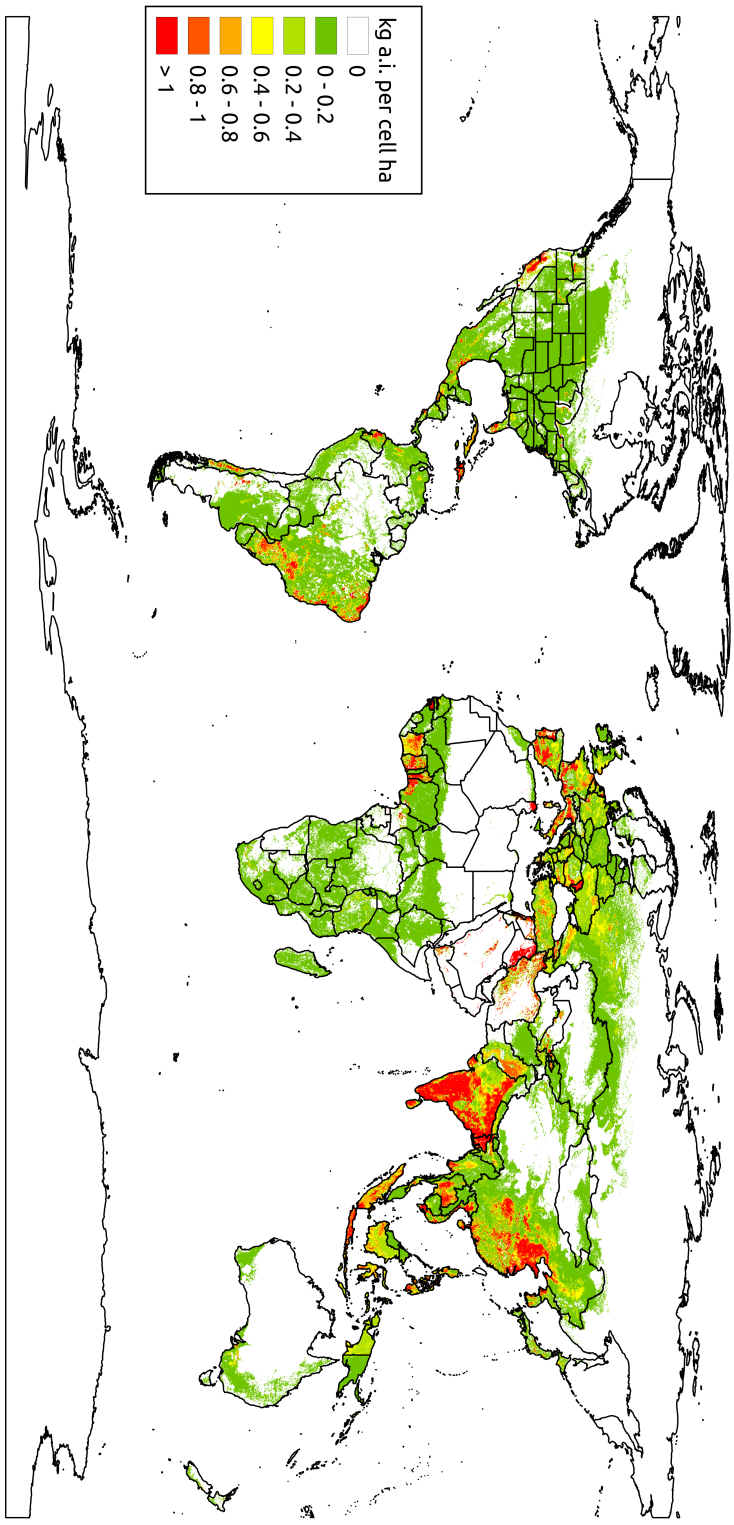

Figure 11.4: World map visualizing the applied amount [*a.i. per cell ha*] for fungicides (figure generated with *QGIS*).

the maps expressed in *kg a.i. per cell ha* with an area grid as described in section 11.1.2 and dividing the result by a map representing the sum of the agricultural area.

High amounts of herbicides applied in agriculture can be observed in southern Brazil, central states of the USA, and regions next to the Mississippi, as well as in Asian regions and countries like India, eastern parts of China and Southeast Asia.

Hot spots of fungicide application are more spread across the world. High application rates can be, for example, observed in California and Florida as well as in different smaller areas in different Latin American countries, however, they can also be observed in Western and Southern Europe, western African countries and in several tropical and subtropical countries in Asia.

As a measure for the deviation between the calculated and published reference values, the ratio $\frac{log(calc)}{log(ref)}$ was calculated for all countries for which a reference value was available. A table summarizing calculated and reference amount, as well as deviation values, can be found in S3 in the online repository `https://gitlab.rlp.net/jrapp1/dissertation`. In the following two maps, this ratio is visualized for herbicides (figure 11.5) and fungicides (figure 11.6):

Log-log regression between the calculated applied amount per country and reference values for all countries in the world are visualized in figures 11.7 and 11.8, resulting in adjusted R^2 values of 0.4864 for herbicides (n=101, intercept = 0.377, slope = 0.885, p = $3.25*10^{-16}$) and R^2 = 0.324 for fungicides (n = 98, intercept = 0.34, slope = 0.848, p = $5.82*10^{-10}$), both with high significance. In contrast to the general rule, the dependant variable is located on the x and the independant on the y axes.

As mentioned in section 11.1.4, for outlier countries with high deviation between reference and calculated amount a non scientific online research was performed.

Countries for which there are incidences for a reduced herbicide or fungicide use in contrast to the average global use are marked in red colour in figures 11.8 and 11.7. For herbicides, incidence for reduced herbicide use was found for 3 countries (Haiti, Laos, Myanamar), for fungicides for 8 countries (Bhutan, Ghana, Haiti, India, Iraq, Laos, Malawi and Mozambique).

Regression trend lines show an overestimation of the calculated applied pesticide amount in the range of the countries for which data was available for reference values for both substance groups, and overestimation of fungicides is higher than for herbicides.

The following figures 11.9 and 11.10 show the applied log-log-regression for countries separated into European, Oceanian, Asian, American, and African countries; the corresponding statistical values of the regressions are listed in tables 11.2 and 11.3.

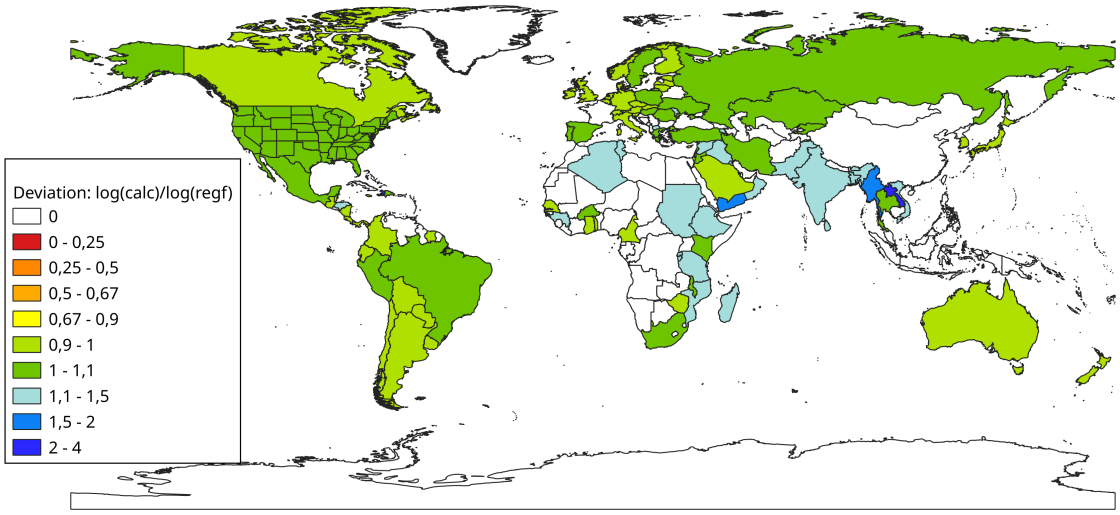

Figure 11.5: World map representing the deviation value $\frac{log(calc)}{log(ref)}$ for herbicides (figure generated with *QGIS*).

Besides the regression for fungicides and African countries, all other regressions are significant, with a used significance level of $\alpha = 0.05$

For herbicides, the highest adjusted R^2 can be observed for European countries (0.764) and the lowest for American countries (0.378). Adjusted R^2 values for fungicides vary between 0.798 for European countries and 0.01 for African countries, for which the mentioned non-significant regression is observed.

For fungicides, the trend line shows an overestimation of the model for Australia and Asian countries as well as for African countries in the range of the regarded countries. In the group of European and American countries, there is no such clear relationship obvious, and the trend line crosses the optimal line within the range of the values of the regarded countries. However, the majority of the values lie near the optimal line with the exception of the outliers for each of the two country groups.

The graphs for herbicides show in general a lower deviation between the trend and optimal line. Similar to the related fungicide graphs, an overestimation for African, Australian, and Asian countries and an unclear behavior for European and American countries can be ob-

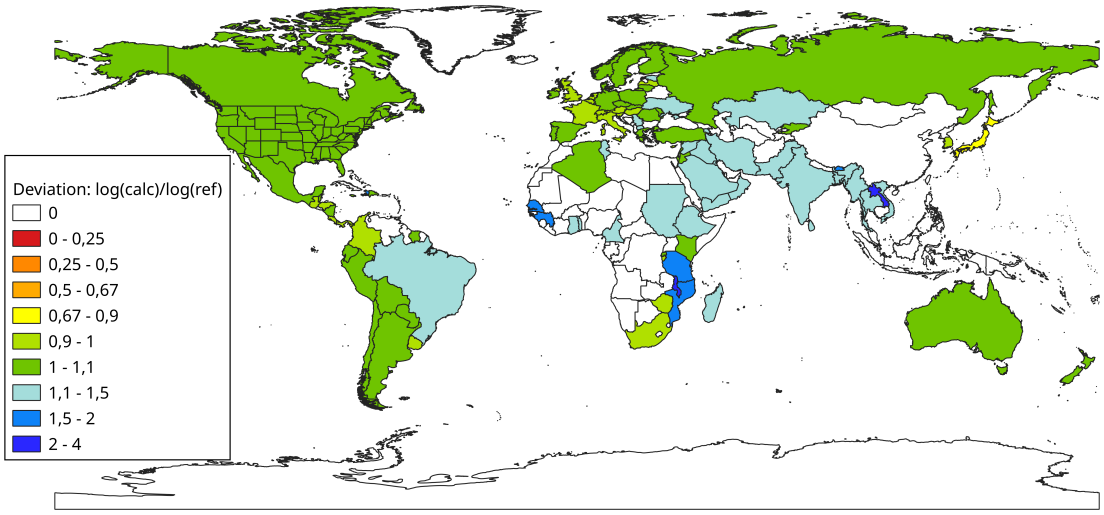

Figure 11.6: World map representing the deviation value $\frac{log(calc)}{log(ref)}$ for fungicides (figure generated with *QGIS*).

served. Additionally, the values lie nearer on the optimal line than the related values for fungicides, resulting in higher adjusted R^2 values (with an exception for American countries).

11.3 Discussion

The results for the regression for herbicides and fungicides are significant in a global view since the calculated applied amount is about 2.96 times higher for fungicides and 1.88 for herbicides. However, there are varying deviations for the different regarded spatial regions (table 11.1).

The used approach has some general limitations that partly explain the high deviations. The method for calculating application rates for countries for which concrete application rate data were not available might be not as correct as needed for detailed calculations. For these countries, different assumptions for calculating the application rates were used, e.g., for countries outside North America, Europe, and Africa, application rates were determined by the arithmetic mean between the European and USA application rates, as described in section 11.1.3.

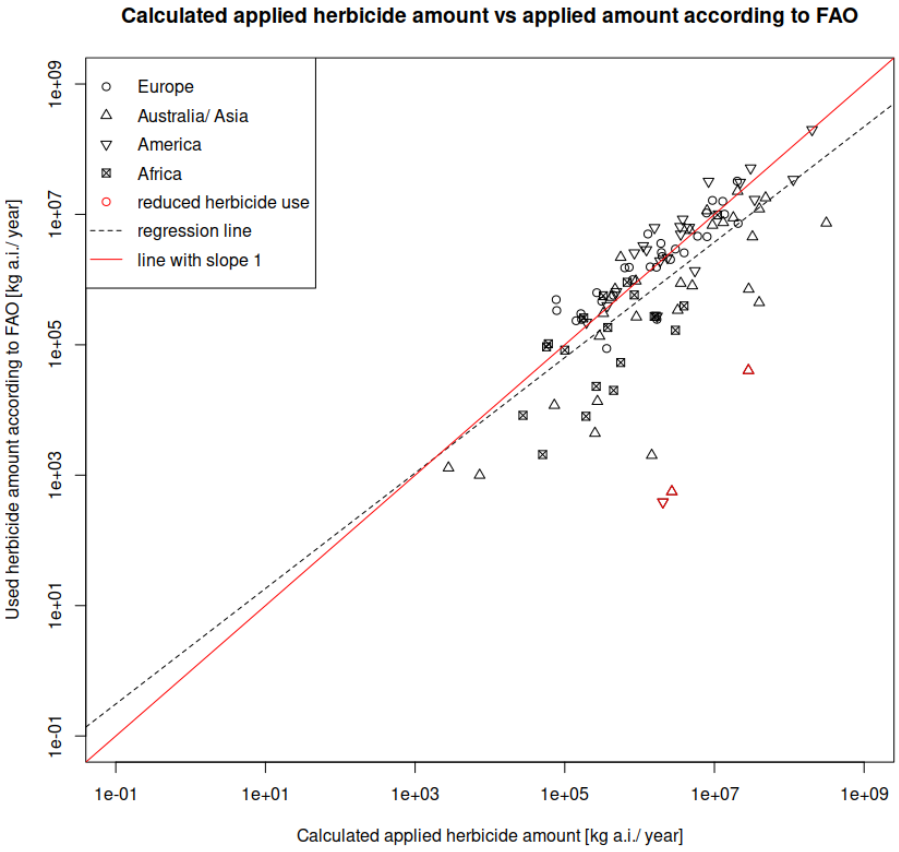

Figure 11.7: Regression between calculated amount and reference values for herbicides with n = 101, adjusted $R^2 = 0.4864$, intercept = 0.377, slope = 0.885, p = $3.25*10^{-16}$ (figure generated with *R*).

But application rates differ due to different environmental, economic and social variables, such as pest pressure, climatic conditions, availability and price of pesticides, labor price, and toxicity of the used pesticides [TDRW+21]. Therefore, the used values are only estimations and might differ to the used application rates in reality. This is also obvious for countries marked with reduced agrochemical use.

In the used approach for most crops, application rates for a single year were taken. However, analysis showed that application rates vary between different years, e.g., fungicide application rates for dates in California published in NPUD 2002 vary between $126.8 kg/ha$ in 1997 and $1.6 kg/ha$ for the year 2002 [CF]. Further analysis of application rates for date cultivation in California, which were published in a California specific pesticide use database, depict that the application rate for 1997 is an extreme value [oPR16]. However, such extreme years may be decreased by the used approach, which uses the average application rates of different

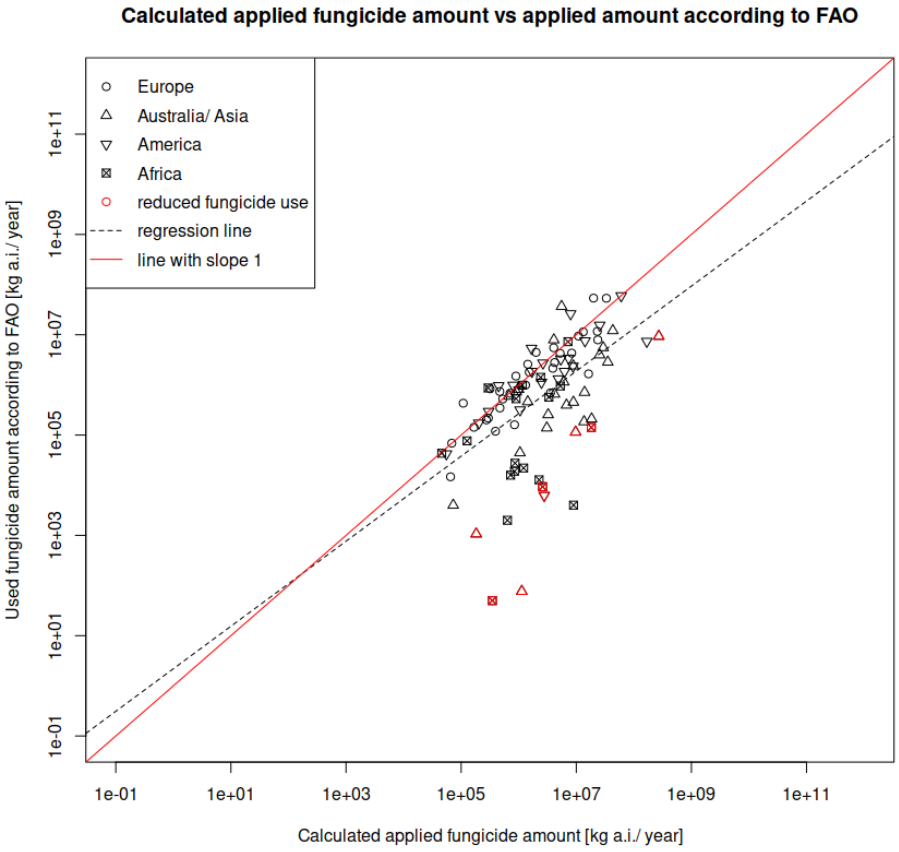

Figure 11.8: Regression between calculated amount and reference values for fungicides with n = 98, adjusted R^2 = 0.324, intercept = 0.34, slope = 0.848, p = $5.82*10^{-10}$ (figure generated with *R*).

spatially separated areas, such as that done using application rates from European countries and the USA, if the reason for the high application rate only has a small spatial dimension. However, differences in application rates for the same crop and for countries or states on the same continent can also be observed [TALS20].

Differences between the calculated and reference values might be also possible because of the use of pesticides with different toxicities in the regarded countries compared with the countries from which the application rates are taken. For example, when comparing two pesticides with different toxicities to a certain pest, the pesticide amount necessary to gain a distinct ecotoxicological effect is lower for a pesticide with a higher toxicity than for a pesticide with lower toxicity. Therefore, the used pesticide active ingredients and related toxicity also determine the necessary amount. In the used approach (with mean values), these differences are not directly implemented.

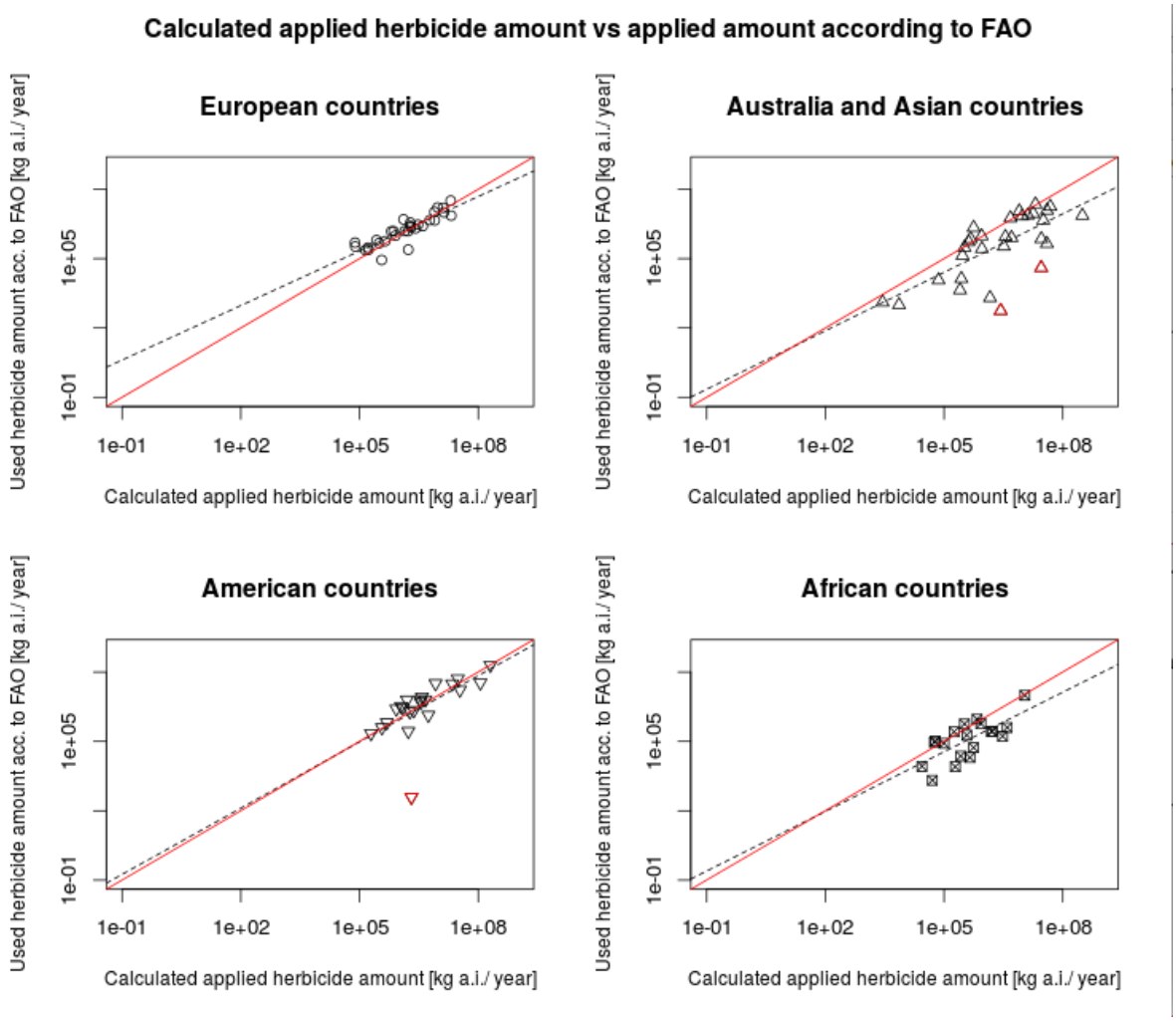

Figure 11.9: Regression between calculated amount and reference values for herbicides, countries grouped into continents or regions (figure generated with *R*).

Further investigations on the input data showed that there are several inconsistencies in the used statistical databases, e.g., for date plant cultivation in California for 1997, both used databases list an overall use of inorganic sulfur of about $323t$. However, the value for the area cultivated with dates differs between two databases [CF] and [oPR16], leading to different application rates.

Additionally, analysis of the used input maps described by [MRF08] showed that the values used for the planted area differ between the [MRF08] files and related reference values listed in [FD11a]. By applying the correction method described in Section 11.1.2, this error should be minimized.

To adapt and improve the used model, further analyses are necessary. For example, in countries with high deviations and overestimation, it would help to determine which of the crops

Table 11.2: Statistical values of the regression between calculated amount and reference values for herbicides, countries grouped into continents or regions.

	adjusted R^2	**intercept**	**slope**	**p**	**n**
Europe	0.764	1.402	0.786	$1.75*10^{-10}$	30
Australia and Asia	0.468	0.19	0.844	$1.86*10^{-5}$	30
America	0.378	0.202	0.995	0.00139	22
Africa	0.471	0.252	0.86	0.0007	19

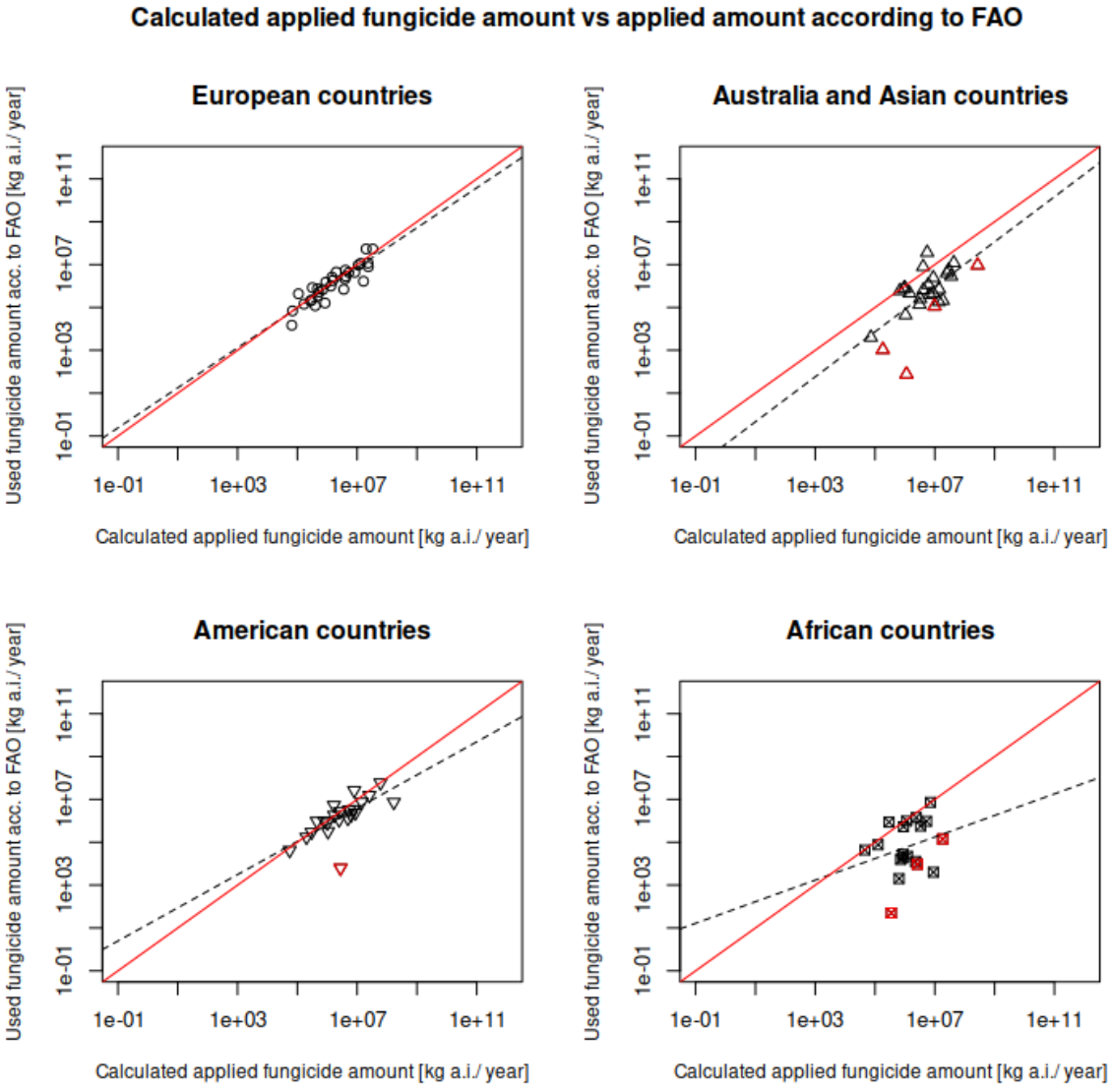

Figure 11.10: Regression between calculated amount and reference values for fungicides, countries grouped into continents or regions (figure generated with *R*).

Table 11.3: Statistical values of the regression between calculated amount and reference values for fungicides, with countries grouped into continents or regions.

	adjusted R^2	**intercept**	**slope**	**p**	**n**
Europe	0.798	0.314	0.934	$8.26*10^{-12}$	31
Australia and Asia	0.395	-1.389	1.05	0.00035	26
America	0.476	1.16	0.774	0.00028	22
Africa	0.013	1.721	0.503	0.28	19

are responsible for the high amount of applied pesticides in the used model. Maps visualizing the crops that are responsible for the highest applied pesticide amount per cell are visualized in Appendix B.

In general, the data quality of input and reference values might not as high as needed for high quality calculations. Comparing reference values from the same dataset at two time points approximately a decade apart shows that data for the same historical period may vary over time. This suggests that initial data may contain statistical inaccuracies, which are subsequently corrected as methodologies improve or more information becomes available.

The previously mentioned sources of error can describe deviations between calculated and reference values in general. However, the results in section 11.2 show possible patterns of deviations in countries for which application rates were available or not for the regarded substance groups and also for different spatial regions.

The model fits better for herbicides, resulting in a trend line for herbicides lying nearer to the optimal line than that for fungicides and in a higher R^2 value. However, both R^2 values are relatively low with values of about 0.4 and 0.2. The low R^2 value for fungicides is also obvious in the scatter plot (Figure 11.8), with more points lying away from the trend and optimal line than for herbicides (figure 11.7). The quality of the used model differs between the two substance groups but also for different regions within the substance groups (figures 11.9 and 11.9). The best results are obtained in both substance groups for European countries. For most European countries, concrete application rates were available in the used data sources. A small deviation can be also observed for the USA with $\frac{calculated}{reference} = 1.014$ for herbicides and 1.01 for fungicides. Therefore, the model fits better for countries with a higher quality of input data.

For African, Asian, and Oceanian countries, the trend line shows an overestimation of the

calculated value in the range of the regarded countries in both substance groups. Visual inspection of the spatial distribution of countries with an overestimation, visualized in figures 11.5 and 11.6, shows that there is a clustering in regions with a low latitude. To improve the used model by modifying the used mean values, further analyses are necessary that investigate the relationship between the deviations and parameters influencing pesticide application patterns, as described in [Eco01]. Possible parameters influencing pesticide application practice might be the agricultural structure (small- or large-scale farmers) in the regarded country, for example, expressed as the number of people involved in the agricultural sector, the economic structure in the regarded country (can people afford expenses for pesticides), or climatic conditions, such as the annual precipitation or mean temperature.

However, the results show that application rates are higher in Europe and the USA than in African and Asian countries. In general, in less-developed countries in comparison to industrialized countries, a much higher use of manual weed control can be observed. [Gia13] reports from several less-developed countries in which manual weed control plays a high role in weed management. [Mel94] reports that in the tropics a method called hoe weeding is often used for weed management in cassava production. It is a very labor-intensive, traditionally used method for weed control with a high manpower requirement. It is mostly used in countries in which cheap labor is available. Therefore, the expense for manual weed control is lower than for agrochemical use.

For African countries, a different approach was used with the assumption that pesticides are only applied on crops intended for export. The selected model fits relatively well for herbicides; however, it does not fit well for fungicides. Regression for fungicides and African countries was the only regression model that was not significant. The reason is unknown and must be further investigated. However, regression for the other country groups were also better for herbicides, but the regression for fungicides showed a lower but still significant regression.

The reason for the difference in the model quality between the substance groups is unknown because for both major groups the same input and reference values were used. Maybe there is a higher spatial and temporal variability in pest pressure or higher variability in used substances and toxicity values for fungicides, both leading to different application rates. In further research, the analysis of the used input data in terms of differences in the variability of pesticide use within the different crops and substance groups would be useful.

Summarizing, a simple model was developed with which it is possible to model the yearly applied herbicide and fungicide amount to a certain degree, with differences in the quality

according to the available input data and the regarded substance group.

11.4 Use of the described model and application maps for an SDSS in an LL

Maps developed in the present chapter have a relatively low spatial and temporal quality, as they represent yearly applied pesticide amounts with a resolution of $5arcmin \times 5arcmin$, about $10km \times 10km$ near the equator. For an SDSS with personally tailored support, where people are, for example, warned to temporarily stay away from an area with high pesticide amounts in the air, the resolution is too low.

However, the maps have another useful application: With these maps it is possible to identify regions with high releases of pesticides into the environment, as totaled over a year. Therefore, hot spots of pesticide application can be determined. Such hot spots might be areas where for example possible negative effects on biota, such as ecosystems in general or humans, can be investigated, e.g., for pesticide residue monitoring in ecosystems or humans or as pilot regions, e.g., for risk mitigation programs.

The method itself can also be applied to input data with a higher resolution, e.g., to maps representing the land cover on a municipal or regional level. Possible application fields are optimization problems, e.g., to find best-fitting locations for medical stations treating pesticide poisonings or, together with a watershed layer, to identify locations where drinking water can be taken from surface or ground water.

The described method has the disadvantage that only the applied amount is regarded without including toxicity of the substance. The applied amount is not directly an indicator for the risk caused by the amount of pesticides; risk is always related to toxicity. If input data is available on the a.i. level, the toxicity of the substances can be implemented by using the toxic-unit approach as proposed by [JBF+06], whereby the amount of the substance is related to its toxicity and amounts of pesticides can be calculated into toxic units and then be compared with amounts of other substances related to their toxicity. Additionally, mixtures of pesticides have different toxicity than the single substances [HGL17] and were not implemented in the used model.

The used open-source approach allows that the resulting maps can be published under an open license. This permits the maps to be used by everybody for free. Regarding the requirements

defined in section 9.1, **R3.1** is therefore fulfilled. To give SDSS in terms of supporting people to avoid areas with a possible high risk, spatial and temporal resolution must be increased. Additionally, models for transport and transfer processes can be implemented to include the fate of pesticides in the environment (section 2.2.5).

The described LL approach with a crowdsourcing framework offers the opportunity to sample data about the used pesticide amount per crop directly by participating farmers in a citizen-science approach but in a much higher temporal and spatial resolution and quality.

12 Generating pesticide application maps with crowdsourcing

12.1 Overview over the framework

In this section, a framework is described in which the idea of a citizen-science project for data collection (chapter 6) with an open-source approach (section 5.3.2.1) in an LL are combined. The suggested software for data collection, processing, and presentation is completely released under an open-source license. The aim of this framework is that citizen scientists report their observations or their knowledge about the temporal or spatial scheme of pesticide application to a central server with a computing unit where data is collected, processed, modeled, and sent back to the user, as visualized in figure 17.1.
In the used framework, citizen science is used for gaining information about the presence of pesticides in the environment. Every community member in the LL can participate in a collaborative effort. The citizen scientists can deliver information about the presence of pesticides in the environment and, if possible, about the characteristic of the pesticide application. The citizen scientists are people who have knowledge about the time point or the spatial character of a pesticide application, such as people who recognize that pesticides are applied, agricultural workers who are involved directly in the pesticide application process or have knowledge about the spatial or temporal application patterns, or people who plan pesticide application.

12.2 Transmitted data items by citizen scientists

Data collection and transmission can be, for example, electronically. The data is entered into a electronic device, in the best case with an adaptive GUI as described in [Pla14]. The used device is dependent on the user's needs. The devices can be mobile, like smartphones, tablets

or laptops, or static, like desktop computers or digital doorways. Mobile devices are relatively widespread in developing countries: more than 75% of the worldwide subscriber identification module (SIM) cards are run in less-developed countries [Pea13]. A study from 2006 analyzed the use of mobile devices among others in South Africa, with the result that in 2006, a relatively wide distribution was observed and was getting broader [Sut08]. It is necessary that the device has at least a temporal connection to the network, either WiFi or with a cable, and the data must be transmitted to the server and the calculation unit. Then, it must be further processed by applying fate models and delivered to the users to give spatial decision support.

Decision support generated by an SDSS should have a spatial and temporal dimension. Therefore, sampled information about pesticide application should also have a temporal and spatial dimension as well as information about the used substance and the crop on which pesticides were applied.

The basic information needed to characterize the pesticide application and create related risk maps may contain the following items, as listed in Table 12.1:

To gain knowledge about the existing risk, it is necessary to measure the temporal and spatial

Table 12.1: Objects and related items for which information are collected to create risk maps

Object	**Items**
Area of application	Coordinates X_n, Y_n
	Polygon
Time span of application	Start time of application
	End time of application
	Time point of application
Applied pesticide	Substance type and formulation
Application method	Application type
	Equipment type
Crop	Crop type
	Crop stage
Application rate	Application rate per ha
	Applied amount per area

extent of risk. In terms of agrochemicals, it is necessary to measure the temporal and spatial

dimension of an initial agrochemical concentration in the environment. This measurement can be conducted directly, by monitoring the concentration of the agrochemical in a compartment, as described in [BGK14] or [KPHW+21], or indirectly, by describing the temporal and spatial dimension of the agrochemical application. The results can be visualized in static maps. Temporal dimension and dynamic maps can be obtained by implementing fate models.

As listed in table 12.1, the items needed to characterize the pesticide application and thus the

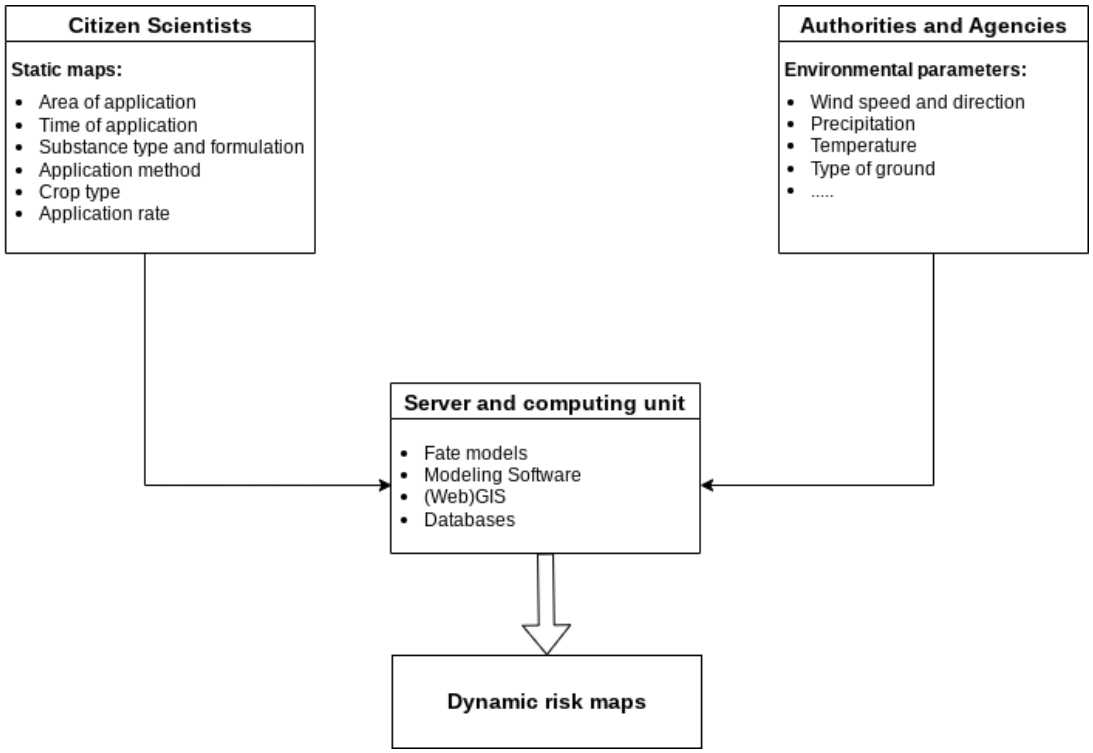

Figure 12.1: Different sources of data to create risk maps with citizen sciences methods (figure generated with *draw.io*).

impact are application area, time span of the application, the kind of the applied substance, the application method, crop, and the applied amount or the application rate.

In figure 12.2, an example for an open-source app with an adaptive GUI for collecting pesticide application observations in an citizen-science approach is visualized. The app was created as a web app with the *R* packages *shiny* [CCA+20], *leaflet* [CSKX24], *RMySQL* [OJD+23] and *DBI* [RWM24] and runs on a *shiny server* with the URL `https://shiny.projects.rptu.de/rapp/LLsampling/`.

To use the app, internet access and a digital device with a browser are required. Observers or citizen scientists can enter items describing the characteristics of the observed pesticide applications as listed in table 12.1, for example, time span of the application, type of the applied

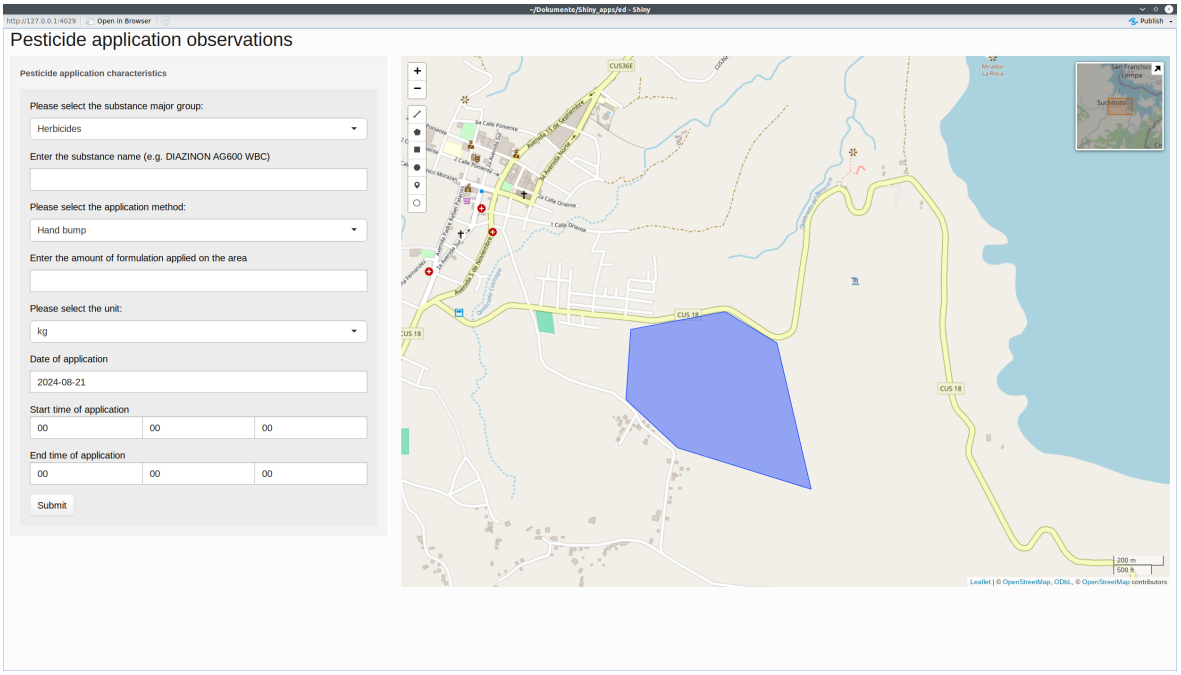

Figure 12.2: Example for the use of an open-source app to sample temporal and spatial information of pesticide application observation (screenshot of a *shiny* app).

substance, application method, and applied amount. Additionally, the area of application can be marked by creating a polygon in a window where a physical map layer is visualized. Entered items and coordinates of the polygon are stored in a database. The GUI of the app is adaptive and responsive. Adaptive means, that GUI elements are presented in relation to previous answered questions, responsive means, that the layout of the GUI changes with the used devices and available size and resolution of the display.

12.3 From static to dynamic risk maps: fate models

In the preceding chapter, an approach with the aim of creating static maps of pesticide use was described. However, pesticides underlie fate with transport and transfer processes (section 2.2.2). With models describing these processes, it is possible to create maps visualizing pesticide contamination with a temporal and spatial dimension.

In section 2.2.5, different models were described with which it is possible to estimate the temporal and spatial fate of pesticides in different environmental compartments. In an LL approach, the model must be selected according to the regarded environmental compartment and to the temporal and spatial dimension on which it operates and which is necessary to give spatial decision support. However, models described in section 2.2.5 have a relatively low

temporal resolution, for example, *SWAT+* has a temporal resolution of one day [BAR+17]. Each model has its own set of input parameters, an app as the one described in the last chapter must be modified according to the needed information necessary for the used model. The open-source concept allows such easy modifications of available source code, e.g., add new items in the app.

12.4 Results and discussion: generating risk maps with crowdsourcing

In the present chapter, an approach was described whereby crowdsourcing or citizen science was used to get information from community members in an LL about pesticide application practices. The approach uses only open-source software with which information is sampled by a web app for which internet communication is necessary. The limitations in paperless communication and the last-mile problem must be kept in mind.

[MOM18] developed an citizen-science application for Ugandan small-scale farmers where they can send pictures of cassava plants with possible diseases or pests and additional geographic information via mobile phone to an ad hoc surveillance system. They have reported several obstacles in the crowdsourcing approach in a rural community in a less-developed country, such as a lack of technical knowledge with mobile devices, lack of internet coverage and GPS access, and lost or stolen equipment [MOM18]. A similar approach for Indian farmers is described in [AKIK21]. Nevertheless, due to the high number and increasing trend in numbers of people using mobile internet, the described approach and risk mitigation strategies are intended for use with mobile devices.

The gained information about pesticide application with a citizen science approach has, in comparison with the approach described in chapter 11, the advantage that the gained data has a higher spatial and temporal quality. However, problems with data gained by a citizen-science project as described in section 6.2 also may happen.

The described approach produces static maps. By using fate models as described in section 2.2.5, it is possible to create temporal forecasts, e.g., of possible pollution in a watershed to avoid drink water usage of contaminated water or of pesticide residues in the air.

13 Calculating the individual impact induced by a walk through a contaminated area

In the last chapters, methods were described with which it is possible to create pesticide release maps showing the released amount of pesticides into the environment. A possible application can be to model, for example, the temporal and spatial distribution of pesticide concentration in the air, as visualized in figure 13.1, and consequently give spatial decision support to community members.

In the present chapter, a method is introduced that can be used to generate spatial decision support for people who want to move from a point A to a defined point B through an area with a toxic substance in the environment, e.g., to determine a way whereby the exposure to a contaminant is minimized. The decision support should be assembled in a way that a path is calculated in which the person suffers a minimum of pesticide-induced impact when moving along this path. This path is then delivered to the user.

In the following parts of this chapter, the mathematical methods and data needed for the described approach are highlighted.

13.1 Needed data to estimate the individual impact

According to definition 3.1.7, the risk caused by pesticides is determined by the toxicity of a substance and by the amount of a substance to which an organism is exposed. As an indicator for the toxicological strength, in the following framework, the so-called toxic-unit approach, as defined in section 3.2.2 and introduced by [Spr70], is further used. In this approach, the concentration of the toxic substance in the environment is related to a toxicological effect concentration.

Another indicator for estimating the risk related to exposure might not only be the amount to which a person is exposed but also the exposure time. Negative effects to organisms increase with a longer exposure time [PS91]. However, there are more complex exposure

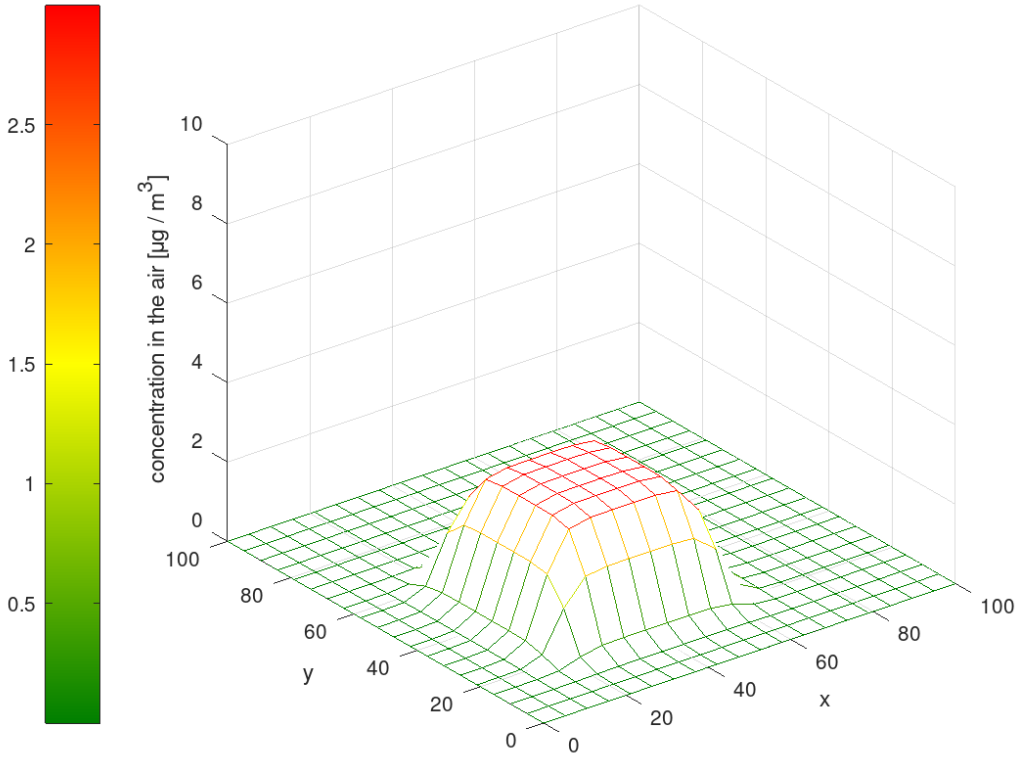

Figure 13.1: Sample spatial visualization of pesticide concentrations in the environment with a underlying street map (figure generated with *GNU Octave*, source of the map: `https://www.openstreetmap.org`).

estimation models for humans working in the agricultural sector available [DAR+02, MPBV08, MRRA+19].

Pesticide concentrations in the environment also have a temporal and spatial component. They are temporal in the sense that concentrations in the environment change over time through different transport and transfer processes [Fen13].

To estimate the negative impact on human health caused by pesticides by walking through an area with pesticides in the environment, it is necessary to have the following information and data:

- path network of the area,
- mathematical functions of the concentration of the regarded substance in the air, water and sediment in relation to the location ω and time t,
- individual information about exposure reduction: use and type of protection clothes, type of used clothes (long-sleeved, shorts etc.),

- information about the path: speed of movement or mean of transportation, start and endpoint of the path, time available to move from A to B, and
- information about the substance: type, toxicity, uptake routes, and rates acceptable daily intake (ADI).

With this information processed for a GIS and with different mathematical methods, such as the path integral, graph theory, and optimization algorithms, paths can be determined in a way that permits a minimum of impact on human health.
The needed data and information can be collected in different ways. Individual information, e.g., about the used protective clothing, mean of transportation, start and endpoint of the planned path, time available, etc., can be individually sampled. This can be conducted, e.g., with a app for mobile devices, similar to the app demonstrated in section 12.2, or with a paper-based survey. As the use of mobile devices is increasing and as the use of mobile devices for delivering risk to the user has advantages against the use of a paper-based SDSS, the following framework is intended for the use of a mobile device. An app integrated or connected to the SDSS can collect the necessary information, e.g., the user must answer relevant questions.
Additionally, a network of possible paths that can be taken to move from point A to point B is required, e.g., as a graph network. In reality, there are natural and anthropogenic obstacles that cannot be passed or must be circumvented, such as rivers, hills, national borders, etc. It is also assumed that people who want to move from a point A to a certain point B will move along existing paths and will not move cross-country. That means that people only walk on predefined paths. To find possible paths, a route map on which roads and ways are illustrated can be used. In figure 13.1, a route map for a community in El Salvador is visualized. As a mathematical structure of possible ways a graph network is used.

13.2 Graph theory

Graph theory can be regarded as part of discrete mathematics. A first publication about graph theory was published in 1736 by *Leonhard Euler*. In this publication, *Euler* described the problem of the seven bridges of Königsberg. At that time in Königsberg, seven bridges crossed the Pregel River. He wanted to find a way to walk through Königsberg where each bridge was passed once and only once. In this publication, he described the necessary and sufficient conditions for how a finite and undirected graph can be walked through in one turn in a way that each edge is passed only once a time [BL95].

Today, graph theory is an important field in mathematics, with several applications in natural science, computer science, and economic science. Graphs are often used to model circumstances and systems of the real world. In this application, the nodes often represent objects of the real world whereby the edges represent the relations between these objects [BL95].

Definition 13.2.1 (Graph) *Let Q be a nonempty set, A a set disjoint from Q and I a mapping that links to each element of A a pair of elements of Q, which can be ordered or unordered. Then the triple $G = (Q, A, I)$ is defined as a graph, I is called incidence mapping, and the elements of A are called edges and the elements of Q nodes of G [BR12].*

For a network of roads, a graph can be used to model the sequence of ways and roads, or in other words, which road can follow after another road was passed.
To find a way of minimum exposure, only a part of the graph is needed, and an algorithm selects a part of the graph, called a path.

Definition 13.2.2 (Edge sequence, trail and path) *Let $G = (Q, A, c)$ be a directed graph with $a_1, ..., a_p \in A$ and $q_i, q_{i-1} \in Q$ and $c(a_i) = (q_{i-1}, q_i)$ for $i = 1, ..., p$. Under these conditions, $(\tilde{a}_1, ..., \tilde{a}_p)$ with $\tilde{a}_1, ..., \tilde{a}_p \in A$ is called edge sequence or walk from q_0 to q_p with the length p. A edge sequence is called trail if all edges are pairwise disjunct. A trail with pairwise disjunct nodes is called a path [Vol13].*

In a directed path as described in the definition above, q_0 is called the beginning node and q_p the end node of the path. The weight of a path is defined as the sum of the edge weights of all edges included in the path [Die96].
Using graph theory for a network of roads, Q represents a set of j crossroads with $Q = \{q_1, q_2, ..., q_j\}$. A represents a set of connections between the crossroads or roads. In a graph, connections can be directed or undirected. If a graph consists of directed edges or connections, it is called a directed graph or digraph [BR12]. In undirected graphs, the connection exists in both directions, e.g., between crossroad q_l and crossroad q_k. That means that the road can be walked in both directions. As there exist one-way streets that can be only passed in one direction, a directed graph is proposed for modeling a street network.
In a directed graph, the set A consists of all existing ordered connections c_m between crossroads, expressed as tuples. For example, if crossroad q_l follows after crossroad q_k, a directed connection from q_k to q_l exists. Mathematically, a directed connection $c_{l,k}$ can be expressed

as $c_{l,k} = (q_k, q_l)$:

$$A \subseteq \{(q_k, q_l) \in Q \times Q\} \tag{13.1}$$

The incidence map I associates to each connection the related nodes connected by the connection with:

$$I : A \to Q \times Q. \tag{13.2}$$

Edges and nodes in a graph can be weighted or unweighted. Graphs with weighted edges are called edge-weighted graphs.

Definition 13.2.3 (Edge-weighted graph) *Let G be a graph. To each edge $c \in A$ of G, a number $w(c)$, called its edge weight, is associated with*

$$w : A \to \mathbb{R} \tag{13.3}$$

$$c \mapsto w(c). \tag{13.4}$$

The resulting graph is called a edge-weighted graph [BR12].

In edge-weighted graphs, a value w_{q_k,q_l} is connected to each edge $c_{l,k} = (q_k, q_l)$:

$$W : Q \times Q \to \mathbb{R} \tag{13.5}$$

Graphs can be visualized in diagrams, whereby the points represent nodes and the lines represent edges.
In figure 13.2, an sample digraph with weighted edges is visualized. The weights of the edges can be stored in an adjacency matrix.

Definition 13.2.4 (Adjacency matrix) *Let $D = (Q_G, A_G)$ be a digraph with $n = |Q_G| \geq 1$ nodes $q_1, q_2, ..., q_n$, and $0 \leq m = |A_G| \leq n^2$ edges. The $n \times n$ matrix $E(D) = (e_{i,k})$, $1 \leq i, k \leq n$, is called adjacency matrix of D if and only if the elements $e_{i,k}$ of $E(D)$ are calculated as*

$$e_{i,k} = \begin{cases} 1 & , if\ (q_i, q_k) \in A \\ 0 & , else. \end{cases} \tag{13.6}$$

[BL95]

Definition 13.2.4 is valid if loops are allowed. If loops are not allowed, the number of edges $|A_G|$ is limited to $0 \leq m = |A_G| \leq n(n-1)$. In edge-weighted graphs, instead of 1 for a connection and 0 for a missing connection between two nodes, the adjacency matrix consists

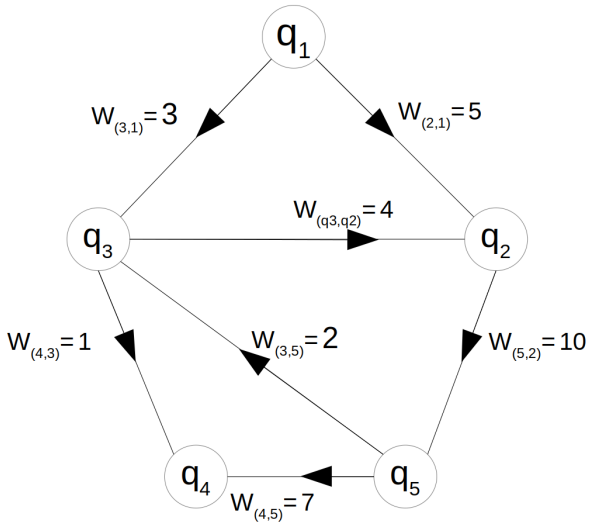

Figure 13.2: Exemplary weighted digraph with 5 nodes and 7 edges (figure generated with *Libre Office Draw*.

of the weights w of the edges.

For an edge-weighted graph G with edge weights $w_{(q_i,q_k)}$, the adjacency matrix $E = (e_{i,k})$ is defined over its entries as

$$e_{i,k} = \begin{cases} w_{(q_i,q_k)} & ,if\ (q_i,q_k) \in A \\ 0 & ,else. \end{cases} \tag{13.7}$$

[BL95] The following matrix represents the adjacency matrix for the graph, visualized in Figure 13.2.

$$E = \begin{pmatrix} 0 & 5 & 3 & 7 & 0 \\ 0 & 0 & 0 & 0 & 10 \\ 0 & 4 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 7 & 0 \end{pmatrix}$$

In the used approach, an edge represents a way segment and can be walked in both directions. A node N represents an object on which an edge starts or ends, the direction of the path changes, or a path segment ends or starts.

In a first approach, it is assumed that velocity is the same for every edge. The development of functions determining the time needed to move from point A to B is not described in this thesis; additionally, conditions influencing velocity, such as slope of the way segment, are not

implemented in this approach.

A graph network based on streets is illustrated in figure 13.3.

The graph visualized in the previous figure is a so-called digraph. That means that it is

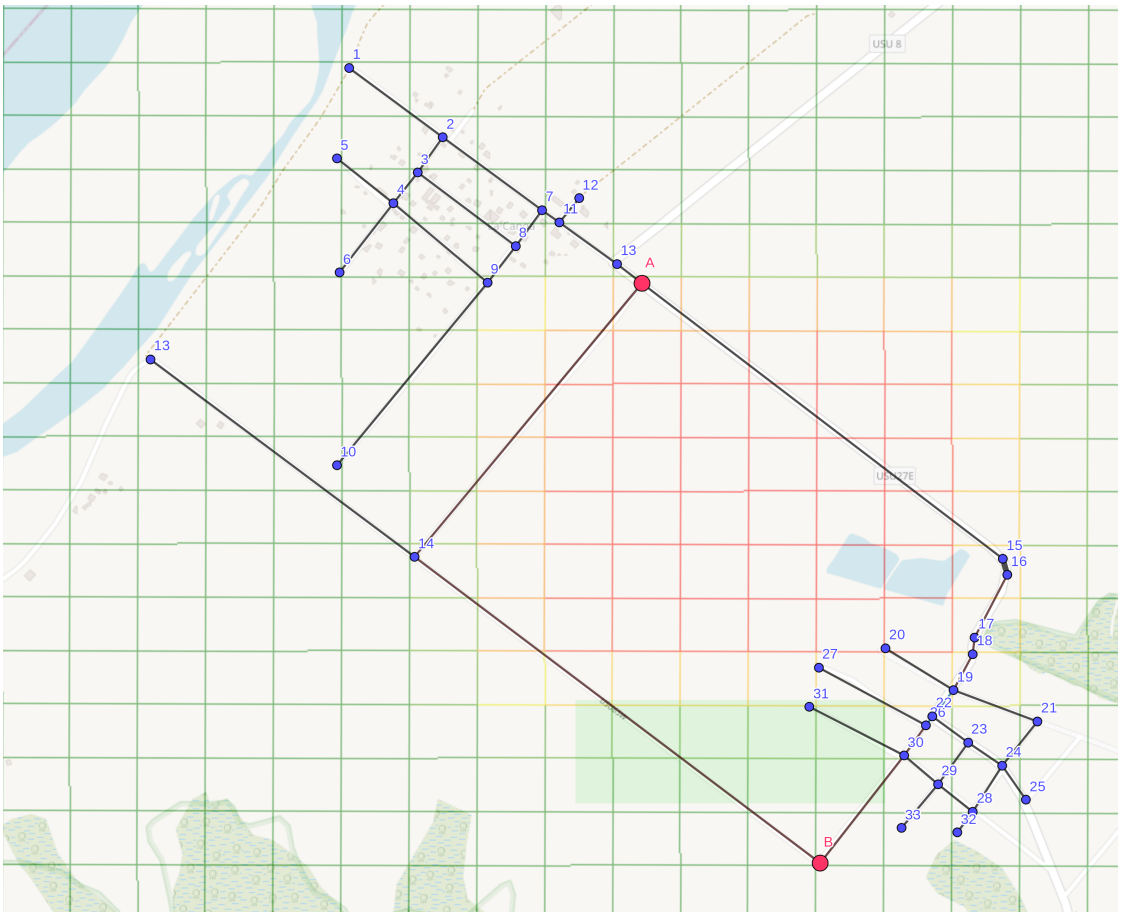

Figure 13.3: Graph network based on a route map from *OpenStreetMap* (OSM) with a overlying air concentration layer for a community in El Salvador (figure generated with *GNU Octave*, source of the map: `https://www.openstreetmap.org`).

possible that a route can be walked in both directions, e.g., from node A to B as well as in the opposite direction from node B to A. However, it might also be possible that a way segment from one node to another can only be passed in one direction, e.g., if there is a one-way road. With this route network, it is possible to model all possible combinations of edges to move from a starting point A to an end point B. Regarding the path network visualized in figure 13.3, a person who wants to walk from node A to node B can take different routes, e.g., route γ_1, moving along the edges $e_{A,14}$, $e_{14,B}$, or route γ_2, moving along the edges $e_{A,15}$, $e_{15,16}$,$e_{16,17}$ $e_{17,18}$, $e_{18,19}$, $e_{19,22}$, $e_{22,26}$, $e_{26,30}$ and $e_{30,B}$.

The combination of edges with the lowest impact on health is delivered to the user or the user

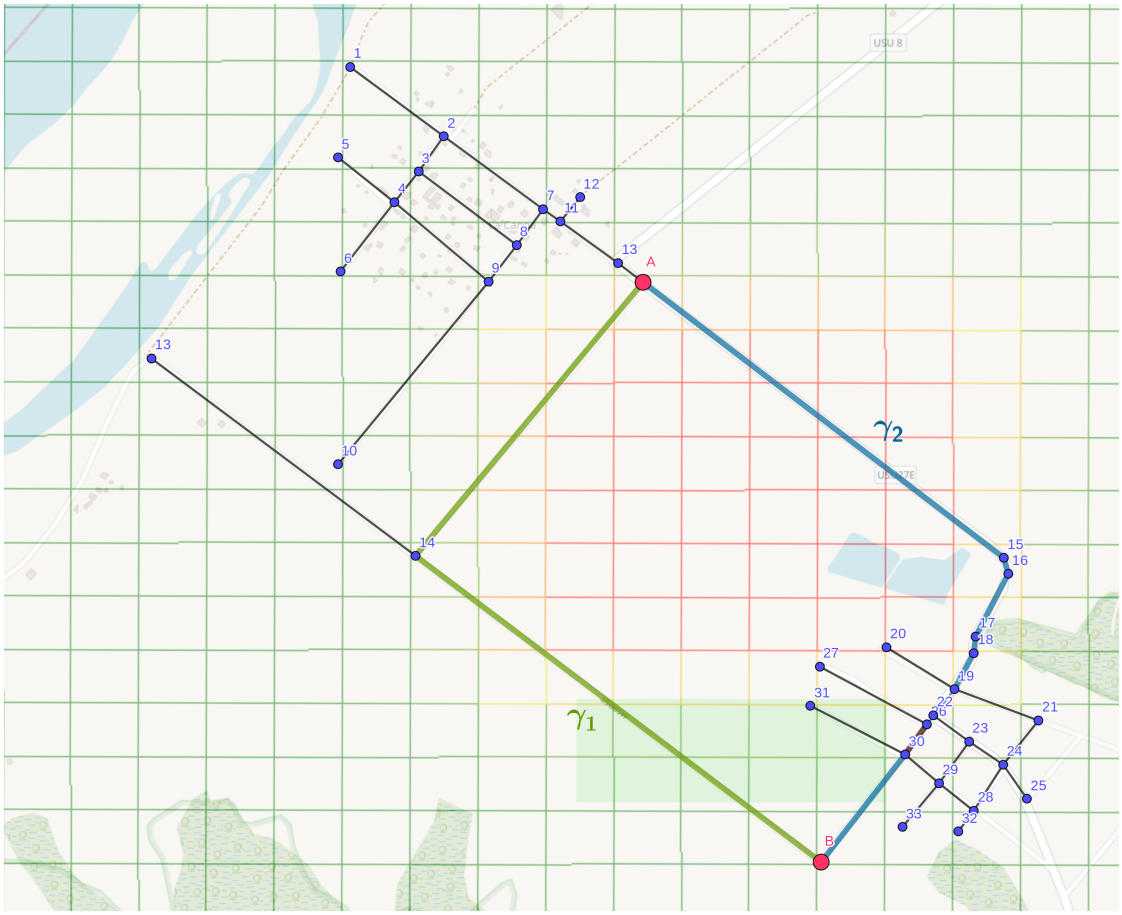

Figure 13.4: Route map from OSM with a path network for a community in El Salvador and possible routes γ_1 (green path) and γ_2 (blue path) (figure generated with *GNU Octave*, source of the map: `https://www.openstreetmap.org`).

is warned, e.g., that he or she should not move at the moment to the chosen destination. Overall, a route from starting point A to end point B is sought for which the impact on human health is minimized.

13.3 Concentration and uptake reduction function

To estimate the impact on human health it is necessary to have information about the uptaken amount of a pesticide or the amount to which a person is exposed. The uptaken amount is related to different factors, such as the concentration of a substance in the regarded compartment, the uptake rate of the substance in this compartment, vapor pressure, the use of equipment to decrease the uptake rate, etc. [DE11].

The use of protective clothing can lead to a decreased uptake that means that, e.g., through the use of a cloth over the mouth, the pesticide amount uptaken by inhalation can be de-

creased. With the individual information about exposure reduction, a value called reduction coefficient r, describing how high the uptake rate is decreased through the use of protective clothing, can be calculated. For example, if a person walks without protective clothes through an exposed area the reduction coefficient is 1, and the uptake rate is not decreased.
The second important information is the knowledge of the spatial and temporal distribution of concentrations of the substance in the different environmental compartments. Figures 13.1 and 13.3 show examples for the visualization of such a concentration function $c_{air}(x, y)$ with a spatial dimension, with green areas representing a low concentration and red a high concentration of the substance in the environment. The function shown is only an artificial example for such a concentration function in the air. The selection of appropriate functions and fate models is a task that must be performed during the research and development cycle in an LL.

13.4 Integral over a path

Additionally, values are needed with which it is possible to model the exposure of a person moving along a path in an area with pesticide concentrations in the air. A mathematical method with which it is possible to calculate a measure to estimate the exposure to pesticides in the air when walking along a path is called path integral.

Definition 13.4.1 (Integration over a path) *A piecewise continuously differentiable mapping $\gamma : [a, b] \mapsto \mathbb{C}$ is called the integration path. The image set $[\gamma] = \gamma([a, b])$ is called carrier, and $[a, b] \subset \mathbb{R}$ is the parameter interval of the path. If $f : [\gamma] \mapsto \mathbb{C}$ is continuous, then*

$$\int_{\gamma} f(z)\mathrm{d}z := \int_{a}^{b} f(\gamma(t))\ dt \tag{13.8}$$

is defined as the integration of f over a path γ.
(Definition adapted from [Bor16], p. 21).

In contrast to the original definition published in [Bor16], the integral defined in definition 13.4.1 is dependent from the way.
The defined integral can be interpreted as the area between a function $f(x, y)$ and a path γ, if $f(x, y) \geq 0$ for all (x, y). If $f(x, y)$ describes the intensity of radiation or pesticide concentration, then the path integral along a way γ can be interpreted as a value for the

exposure to radiation or pesticides.

In the used approach, it is assumed that people only walk along existing ways or streets. Mathematically, ways or streets and their connections are further modeled as a graph. Nodes represent corners, starting, or ending points of streets, and edges represent ways or streets that are called way segments.

For the used approach, a value is assigned to each edge that represents the exposure to a substance if a person walks along this way segment. In other words, a value is assigned to each way segment to which the path integral at a given time point is assigned. The higher this number or edge weight, the higher the exposure to the regarded substance.

By using the concept of linear combination (definition 7.2.10), a route that connects points P_1 and P_2 in a straight line can be described with the path

$$\gamma(t) = (1-t) \cdot P_1 + t \cdot P_2 \tag{13.9}$$

and $t \in [0,1]$, and the route is expressed as $[P_1, P_2]$.

The integral over a path with start and end nodes P_1 and P_2 is calculated as

$$\int\limits_{[P_1,P_2]} f(z)\mathrm{d}z = \int\limits_0^1 f((1-t) \cdot P_1 + t \cdot P_2)\ dt \tag{13.10}$$

$$= \int\limits_0^1 f((1-t) \cdot P_1 + t \cdot P_2) dt \tag{13.11}$$

With this method, it is possible to calculate the uptaken amount of pesticides caused by a defined walk along a path through a area with a given contamination function $c(x, y)$, and it is possible to compare different routes according to the exposure.

For each edge, the defined path integral is calculated according to Definition 13.4.1; this value is further used as the edge weight $w_{j,i}$ for an edge $e_{j,i}$ between two point P_i and P_j, connected over a straight line:

$$w_{i,j} = \int\limits_0^1 f((1-t) \cdot P_i + t \cdot P_j)\ dt \tag{13.12}$$

With the equation above, a value representing the amount to which a person is exposed can be assigned to each way segment.

Summarizing, the described approach can be modeled with a directed weighted graph $G = (Q, A, I)$, whereby the edges A of the graph represent way segments and the weights a value describing the exposure of a person walking along the way segment. If a way segment cannot

be walked in one direction, then the weight in the regarded direction is set to ∞.

13.5 Finding a way with a minimum of exposure

After reducing the approach to a directed graph with weighted edges, finding a way with minimum of exposure can be attributed to a graph optimization problem where a path is sought between a start point A and an end point B for which the sum of the weights along the path is minimized. The described optimization approach is called shortest-path problem [Sch12].

In optimization theory, a problem with a starting point A and a finishing point B and non-negative edge weights is called a single-pair shortest-path problem [ES11]. A review of available shortest-path algorithms can be taken from [MAR+17].

Mathematically, the shortest-path problem for a directed weighted graph can be described as follows:

A directed graph

$$G = (Q, A, I) \tag{13.13}$$

with edge weights

$$w : A \mapsto \mathbb{R}_0^+ \tag{13.14}$$

is given.

The weight of a path

$$p = (P_0, P_1, ..., P_k) \tag{13.15}$$

with

$$P_0, P_1, ..., P_k \in Q \tag{13.16}$$

can be calculated as

$$w(p) = \sum_{i=1}^{k} w_{(i-1,i)} \tag{13.17}$$

A path $p^*_{i,k}$ from node P_i to node P_k is called shortest path if

$$w(p^*_{i,k}) \leq w(p_{i,k}) \tag{13.18}$$

for all paths $p_{k,i}$ from P_i to P_k.

Classical applications of the shortest-path problem are, for example, traffic navigation systems to find the shortest routes in a street network [NMST08] or algorithms in internet-related IT

systems to find the shortest path for data packages between two digital devices [KK10].

An easily implementable algorithm to solve the shortest path problem is the *Dijkstra algorithm* [Dij59]. Other shortest path algorithms are, among others, *Floyd-Warshall* [Flo62], *Bellmann-Ford* [Bel58], and *Genetic algorithm* [LKS16]. They differ, inter alia, in their performance in finding a shortest-path solution [GMNZ14].

An example of the most famous shortest-path algorithms is highlighted. The basic idea of the *Dijkstra algorithm* is always to follow the edge that promises the shortest route section from the start node. Other edges are only followed once all shorter route sections have been considered. This procedure ensures that when a node is reached, no shorter path to it can exist. Once a distance has been calculated between the start node and a visited node, it is saved. However, the summed distances to nodes that have not yet been processed can change, i.e., decrease, during the course of the algorithm. This procedure is continued until the distance of the target node has been calculated [JAE+12].

Dijkstra and *Bellmann-Ford algorithm* are implemented in common open-source software packages like *R* package *igraph* [CN06] or the *GnuOctave* package *grShortestPath*.

13.6 Discussion and conclusion

In the present chapter, a method was developed based on graph theory with which it is possible to find a walk through an area with contaminants in the air with a minimized exposure to the contaminant. The described approach has some limitations but can be regarded as a base for the research and development cycle in an LL.

The limitations of the described approach are highlighted in the following:

In the described approach, the value to estimate the possible negative impact on human health is based on the concentration of a contaminant to which a person is exposed and to the path's distance. However, the relevant factor in ecotoxicological risk assessment is also exposure time [Fen13]. The time a person needs for a way segment changes with the speed. However, in the described approach, it is assumed that each way segment is taken with the same speed, disregarding individual differences in speed but also natural characteristics like slope or type of ground. Therefore, during the research and development cycle a parameter describing the speed should be implemented to make the approach more realistic.

There are also limitations with the used concentration function used as a function to estimate

the negative impact. In the described approach, a static concentration function was used. However, the fate of pesticides and related processes lead to varying concentrations over time and the concentration function becomes a function with a temporal and spatial component $c(x, y, t)$. If c has a temporal dimension, then also the way segments' weights change over time and have a temporal dimension.

In the described approach, the environmental concentrations are taken as a measure for exposure. However, pesticides must come in contact with biological entities and must be uptaken to lead to a negative effect. With personal protective equipment, it is possible to reduce the uptaken amount. Therefore, in a further step during the research and development cycle, the approach should implement personal protective equipment or other methods to reduce the uptaken amount, e.g., by a uptake reduction factor r on an individual base.

The optimization problem described in section 13.5 is only related to the pesticide concentration in the environment. But concentration or exposure alone is not enough to estimate the risk caused by pesticides; toxicity also plays an important role in risk assessment (section 3.2.1). If the described approach is extended to different pesticides with varying toxicity, it would make sense to use a toxic-unit function, as described in [Spr70], whereby the amount is related to a toxicological endpoint. A toxic-unit function would be a more realistic function to estimate the negative impact than a concentration function alone, especially if there are more than one pesticides in the environment in which a person wants to walk.

A similar approach was used by [Alz10], whereby the shortest paths through a field of radiation were calculated with *Dijkstra* and *Bellmann-Ford* algorithms. Another application of a shortest-path problem for radiation reduction for walks in a power plant building can be found in [LXYY21]. In contrast to the more general approach described in this thesis, a discrete step function was used as irrigation function.

Summarizing, the described approach can help to reduce exposure to pesticides when a person wants to find a way through an area with contaminants in the air and can be used in an SDSS. During the research and development cycle in the LL, algorithms to solve the shortest-path problem must be selected according to the performance of the algorithms and the available hardware. There are still ways, which were mentioned in this chapter, to improve the described method in the research and development cycle.

14 Part conclusion

This part deals with risk mitigation strategies that are adapted to the characteristics in less-developed countries and to the described LL approach. They can be regarded as low-cost methods and are free of charge when they operate with open-source software that is adapted to the characteristics of rural communities in less-developed countries.

They can help to minimize the release of pesticides into the environment (precision farming approach, section 10) and exposure of community members to pesticides (path integral, section 13) or to identify areas with high pesticide use for monitoring programs (11).

The developed open-source risk mitigation strategies are an example for a possible tool set for risk mitigation strategies that can be brought into the research and development cycle of an LL from the perspective of a geoscience- and modeling-related stakeholder. In the research and development cycle, they can be, for example, adjusted, further developed, and improved based on the needs of the community.

The described risk mitigation strategies can be regarded as a base for a repository of possible risk mitigation strategies in a LL with a low cost framework. However, there are also different other risk mitigation strategies thinkable, also from other scientific disciplines or fields not mentioned in this thesis. For example the application of AI based chat bots as risk mitigation tool comes more and more common [KHN+24].

However, not every risk mitigation strategy fits to the different people in an LL community in the same way. For example, the VRT approach described in chapter 10 fits more for people involved in the agricultural working process than for people not directly involved in pesticide application practice. The following part deals with methods with which it is possible to determine the fitness of risk mitigation strategies according to, e.g., personal, environmental, and economic parameters in an SDSS .

IV | SDSS

15 SDSS and requirements in the described LL approach

One of the aims of the thesis is to develop an SDSS that can act as a countermeasure against the prevalence of CKD in less-developed countries with an unknown etiology. Decision support might help to improve the situation in CKD affected areas. With the mathematical methods described in the following chapters, it should be possible to build an SDSS system adapted to the characteristics of a rural community in less-developed countries and that can be developed and used in an LL approach.
According to [Kee03], an SDSS is a system that helps to make decision making easier by using a computer when the decision has a spatial component. Mostly, GIS is a system used with an SDSS in order to store, access, and manipulate spatial data.
The following chapter combines results from the previous chapters and determines requirements for a possible solution for an SDSS in the described framework.

15.1 Problem, aim and background of an SDSS in the described framework

People endangered by a disease of unknown etiology caused by agrochemicals might not have complete knowledge about existing risk mitigation strategies or of their spatial and temporal availability and effectiveness. They need a tool that can help them make better decisions in choosing a appropriate risk mitigation strategy in relation to parameters with a spatial dimension, with the overall aim of mitigating the risk of a disease caused by agrochemicals. Such a tool can be an SDSS.

Definition 15.1.1 (SDSS) *"SDSS are integrated computer systems that support decision*

makers in addressing semistructured or unstructured spatial problems in an interactive and iterative way with functionality for handling spatial and nonspatial databases, analytical modeling capabilities, decision support utilities such as scenario analysis, and effective data and information presentation utilities" [SD10, p. 14].

According to this definition, an SDSS is more than merely a system that helps make decisions on spatial and temporal data, but the term also implies tools for data handling and visualization utilities.

Research on systems for decision support without a spatial component, further called Decision Support System (DSS), started in the 1970s. The main applications of DSS were in economics and business [Kee03]. As spatial data has large data processing and manipulation requirements, the application of DSS with a spatial components started with the commercial use of computer systems in the middle of the 1980s. In the first SDSS approaches, DSS were combined with GIS tools to expand decision making to spatial topics [CWP95]. Currently, SDSS have wide applications, e.g., in the fields of land-use planning [YDT$^+$12, CDT12], biodiversity conservation topics [BCD$^+$13, PSK$^+$11], agriculture [ZMBM20, RZJ$^+$20], or health-related topics [KTVC12, GAD$^+$22]. Common methods to generate decision support are, among others, *weighted linear combination*, *cellular automata*, *agent-based models*, ANN, and *fuzzy modeling* [SD10].

In the literature, there are few SDSS for application in agriculture and related to pesticides and fertilizer input. For example, [AAS14] develop a *fuzzy logic*-related method for decision support in fertilizer management based on soil nutrition maps and cropping dates. [RZJ$^+$20] describe an SDSS that gives site specific decision support in terms of fertilizer management in maize cultivation. Through the application of the SDSS, an increasing yield could be achieved with slightly decreased inputs. An ANN-based approach to predict the temporal development of plant virus diseases and insect pests in relation to meteorological parameters can be found in [LYGM06].

In the present thesis, an approach is developed for how such an SDSS can be constructed with available mathematical methods and open-source software to inform people exposed to risks or who suffer from the disease about the best-fitting risk mitigation strategies. The decision making in the described approach is done by incorporating different parameters that determine the fitness of risk mitigation strategies tailored to the user and environmental parameters. The SDSS should be designed in a way that it learns by the users' feedback about the proposed risk mitigation strategy to improve decision making and that the decision mak-

ing process should be also adapted to changes in environmental parameters. Requirements for such a system are determined in the following chapter.

15.2 Requirements for an SDSS in the described framework

In the following section, requirements and constraints for an SDSS system in the described framework are determined.

(R4.1) Tailored decision support: The proposed SDSS should give decision support that is tailored to the personal characteristics of the user, available risk mitigation resources, and environmental parameters with a spatial and temporal dimension. Expressed mathematically, a system is needed that performs a mapping to the grade of fitness of different risk mitigation strategies in relation to predefined input parameters. Therefore, the decision support, e.g., proposing the fitness of different risk mitigation strategies, must be adopted to the users' characteristics and the area profile, e.g., environmental or social-economic characteristics, to achieve the highest risk mitigation benefit.
For example, it does not make sense to propose a tutorial in text form about the usage of pesticides to an illiterate person or a person not working with pesticides. Additionally, for a person in an early stage of the disease other risk mitigation strategies are required than for a person suffering from CKD at a higher stage.
The grade of fitness of a risk mitigation strategy must also be adapted to environmental parameters, such as precipitation or mean temperature. Environmental parameters can be used, e.g., to suggest the right application time to minimize the used agrochemical amount or to suggest strategies applicable only in a specific temperature.
Social parameters must be considered, e.g., if some risk mitigation strategies are not accepted through cultural or religious restrictions. Additionally, the spatial and temporal availability of resources needed for a distinct risk mitigation strategy must be considered, e.g., if protective clothing is not available at a location, this risk mitigation strategy should not be suggested.
The proposed risk mitigation strategy must also be adapted to the existing risk, e.g., to the amount of a agrochemical in the environment. Overall, the system must be able to suggest risk mitigation strategies with the highest benefit for the user based on the available ressources.
(R4.2) Learning system: To achieve the highest benefit from decision support, the developed system must be organized in a way that it is able to learn from existing cases in order

to improve decision support through feedback and learning. The system should be able to deal with feedback of the user if the proposed risk mitigation strategy is compatible with his or her needs and constraints and adjust the decision support in following cases to the user's feedback. Through this learning procedure, the decision making process should be improved in similar cases that follow. An on the fly learning procedure is required in which new data delivered by the users is used to improve the decision algorithm.

(R4.3) Dealing with incomplete and noisy data: The system must deal with noisy and incomplete input data that might come from surveys in regions with low infrastructural resources. This is the case if the results of surveys are used in which some data items are missing or where data generation has methodological errors resulting in noisy data.

(R4.4) Fast data delivery: Because of the high toxicity of the regarded substances and fast fate of agrochemicals, it is necessary to have a system that delivers data about the grade of risk or grade of quality quickly and with high quality and accuracy.

(R4.5) Data heterogeneity: Data from different sources with different data quality describing the same topic must be usable, e.g., if data about the personal exposure is not directly available, maps visualizing application rates or the applied substance amount per year should be used as a proxy variable instead.

(R4.6) Adaptable and free software: The system should be usable in less-developed countries as a possible low-cost risk mitigation strategy. That means that the used software and algorithms must be free. Additionally, adaptation of the software components of the SDSS during the research and development cycle or to structurally equivalent problems must be possible, resulting in the use of open-source software.

(R4.7) Fuzzy logic and rules: In the example of CKD in El Salvador, for some necessary input parameters, crisp values were not available, but fuzzy values were. In this case, the system must be able to work with fuzzy values.

Rules to determine the fitness of different risk mitigation strategies can be expressed in logical expressions with which it is possible to calculate the grade of fitness. As, inter alia, fuzzy values are used, the system should be able to work with *fuzzy logic*.

(R4.8) Rule extraction: In the field of a disease with unknown etiology such as CKD in particular, relations between possible risk factors and the impact and risk are not exactly clear. In the best case, the system should be able to extract rules from presented input-output value patterns with which human experts can, for example, generate new hypotheses of the etiology of the disease.

The overall aim of the SDSS is to evaluate the grade of fitness for all risk mitigation strategies in relation to personal, social, and environmental parameters with the aim of determining the personal fitness of each risk mitigation strategy adapted to the user's situation and parameters.

In the following chapters, methods and software implementations are introduced with which it is possible to create an SDSS fulfilling the mentioned requirements. In chapter 18, the requirements are taken up again, and a possible solution for an SDSSin the described framework is developed in which the different chapters of the thesis are brought together.

16 Fuzzy sets, fuzzy logic, and ANN and their use in an SDSS

In an LL in less-developed countries, as proposed in this thesis, it is assumed that data necessary to give decision support or monitor the success of risk mitigation strategies might be fuzzy. With the mathematical concepts fuzzy sets, fuzzy logic, and fuzzy rules, it is possible to deal with such fuzzy data. The concept of ANN can be used, inter alia, to give personalized decision support, as proposed in chapter 15. Therefore, in this chapter the mathematical basics about fuzzy sets 16.2, fuzzy logic 16.2.4, and ANN 16.3 are introduced. The theory of fuzzy sets and fuzzy logic are necessary for dealing with non-crisp fuzzy data mathematically.

16.1 Introduction: the use of fuzzy sets and fuzzy logic in the described LL framework

The first chapters of this part of the thesis deal with fuzzy sets and fuzzy logic. This mathematical theory is used frequently in the following thesis, e.g., to combine two maps logically or also for a *fuzzy logic controller* in section 16.2.4, to determine the best-fitting risk mitigation strategies. These mathematical concepts are used because a lot of data used in the described LL approach, e.g., the grade of illiteracy or the grade of CKD, is fuzzy. Through the use of fuzzy sets and fuzzy rules it is possible to combine, deal, and operate logically on such fuzzy data. As this concept is essential for the following parts, the first chapters of this part deal with this concept.

In the chapters after the following, this framework is extended, e.g., how to deal with incomplete data, how to use mathematical methods to work with this data, and how to improve the used fuzzy rules with personalized feedback.

16.2 Fuzzy set theory

In this section, fuzzy set theory is introduced and applied in the context of agrochemicals and their related risks.

16.2.1 Introduction to fuzzy set theory

One of the most important concepts in mathematics to describe objects and how they can be aggregated is the so-called set theory. The axiomatic set theory, as used nowadays, was established by the German mathematician *Georg Cantor* in his paper "Beiträge zur Begründung der transfiniten Mengenlehre" in 1895 [Can95].

Definition 16.2.1 (Set) *A set is any aggregation M of certain well-differentiated objects m of our view or thought, which are called the elements of M, into a whole [Can95].*

In classical set theory, a set is characterized by specifying its elements. If two sets A and B have the same elements, then they are called equal. If an element b belongs to a set B, it is symbolized with $b \in B$. According to the previous definition, an element does not have to be a number; the elements of a set can be any object, including sets themselves or names. In ordinary set theory, there are also operators, such as union $\cup$ or intersection $\cap$, with which it is possible to operate on sets or combine them with logical expressions [AHK+15].
According to the classical set theory, an object can be either an element of a given set or not. However, in reality, it is often hard to make such a crisp differentiation.
For example, regarding a set F, which consists of all temperatures in which a human being would say that he or she has a fever, it is clear that a temperature of 37 is not an element of F ($37 \notin F$), however a temperature of 40 would be an element of F ($40 \in F$). But what about a temperature of 38? Does it belong to F or not? Here, it is hard to decide if it is an element of the set or not. It belongs somehow to the temperatures associated with fever but not really. With classical set theory as proposed by *Cantor*, it is not possible to describe such objects that belong, put simply, partly to an aggregation or set.
Another example for the use of *fuzzy set theory* and the vagueness in human language is the statement of whether a human being is large. Without doubt, everybody would say that a man with a body height of 2 m is large. But this categorization has a smooth transition. What about a man with a body height of 1.80 m? Is he large? With fuzzy set theory, the grade of largeness can be described mathematically; for example, a man with a body height

of 1.80 m might have a largeness of 0.6 .

Nowadays, *fuzzy set theory* has many practical applications. The first applications were in control theory and engineering and for the development of artificial intelligence and neuronal networks [Low96]. Currently, *fuzzy set theory* is used in several different fields, and its application is very widespread, such as in profitability assessment [WDB13] and in the analysis of human behavior [ZT13] or decision making [HC00].

In the described approach of an agrochemical-related LL in a less-developed country, the gained information, e.g., about the used amount of agrochemicals, might be imprecise and fuzzy or only describable with linguistic expressions because of missing monitoring or concrete measuring. For such fuzzy data, the classical set theory and related operators, as proposed by *Cantor*, do not work. [Zad65] proposed a mathematical theory called *fuzzy set theory* and *fuzzy logic* with which it is possible to operate on such fuzzy data and sets.

Fuzzy set theory is a theory which describes and links fuzzy sets and was first described in [Zad65].

Definition 16.2.2 (Membership function) *A fuzzy set is characterized by a mapping μ from an underlying set Ω to the real unit interval:*

$$\mu : \Omega \longrightarrow [0, 1] \tag{16.1}$$

$$\omega \longmapsto \mu(\omega) \tag{16.2}$$

[Bie97]

The function μ is called membership function, and its values are called grades of membership. The membership function can be described by a functional equation, a functional graph, or the pairs of value of the elements and their grades of membership. A membership function can be determined by empirical methods or manually. For the manual determination, expert knowledge of the problem is needed [Bie97].

Figure 16.1 illustrates a sample membership function, mapping from $\mathbb{R}$ to $[0, 1]$: In the framework of an agrochemical-related LL, *fuzzy set theory* is, inter alia, used to visualize tempo-spatial tasks, such as the grade of risk related to space and time. For spatial or tempo-spatial tasks, Ω gets a spatial and or temporal dimension, for example, by incorporating latitude, longitude, and time. Thus, *fuzzy logic* can also be used in data with a spatial and temporal task.

Regarding the whole Earth, Ω can be for example expressed as

$$\Omega_s := [-90°, +90°] \times]-180°, +180°] \tag{16.3}$$

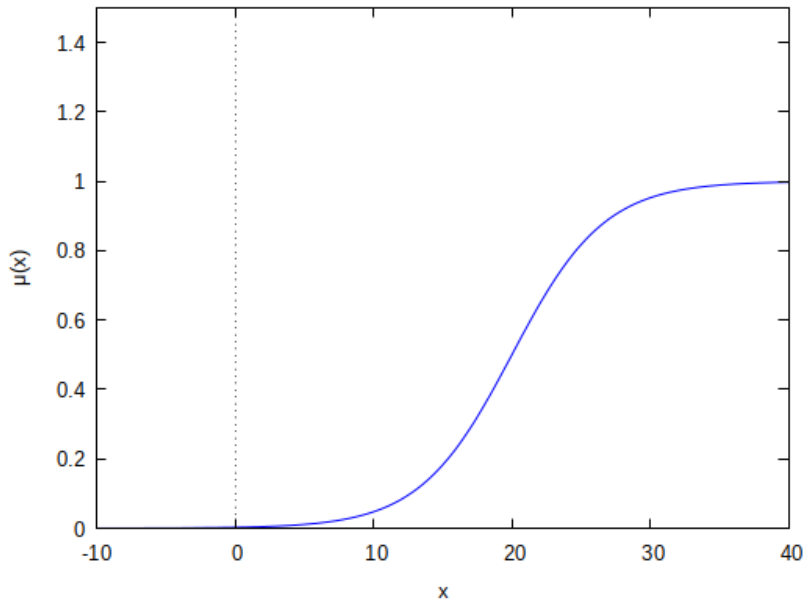

Figure 16.1: Graph of a sample membership function $\mu(x) = \frac{1}{1+e^{-0.3\cdot(x-20)}}$ (figure generated with *GNU Octave*).

or by incorporating space and time as

$$\Omega_{st} := [-90°, +90°] \times]-180°, +180°] \times N_t \tag{16.4}$$

with $N_t \subseteq \mathbb{R}$. In figure 16.2, an example for a map with a mapping $[-10, 10] \times [-10, 10] \rightarrow [0, 1]$ is visualized. Data with a spatial and temporal dimension can be visualized as a sequence of maps, with each map in the sequence representing the situation at one time point.

The fuzzy sets that can be built over the fundamental set Ω are symbolized with $\Phi(\Omega)$.

In contrast to ordinary *set theory*, where objects can be only part of a set or not, in *fuzzy set theory*, objects can partly belong to a set, described as a grade of membership. In the concept of binary logic and classical set theory, the membership of an element to a set can only be described as true (1) or false (0). In *fuzzy set theory*, each element of a set can be characterized with grades of membership between 0 and 1 [Zad65].

Ordinary sets can be considered as special fuzzy sets, where all objects belonging to a set have a grade of membership of 1.

Additionally, *fuzzy set theory* is a concept that enables the handling of linguistically formulated knowledge. *Zadeh* postulated that in reality, classes often do not have precise membership criteria [Gra95]. The concept of fuzzy theory was introduced to model and deal with linguistic uncertainty and vagueness mathematically. The mentioned uncertainty is not a result of the

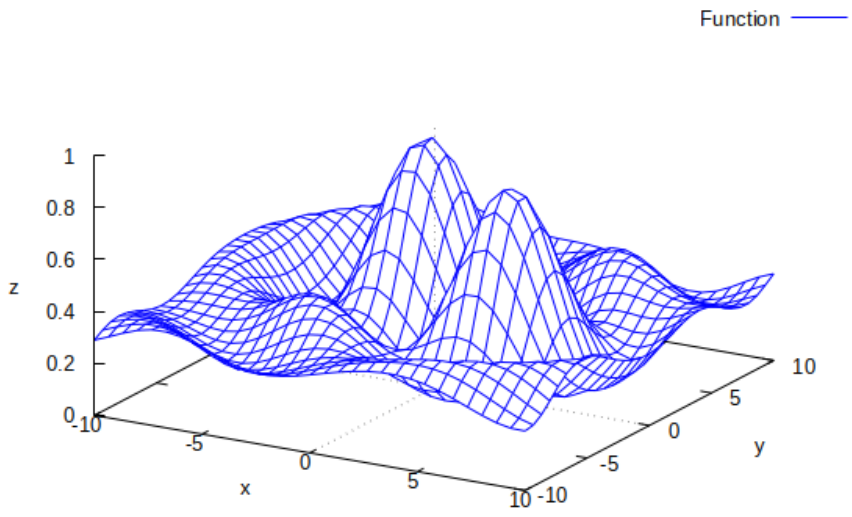

Figure 16.2: A sample map with fuzzified values (figure generated with *GNU Octave*).

lack of knowledge about whether an event will occur or not, as in the probability theory. It is a result of the problem that precise definitions with proper thresholds are not possible or meaningful.

Fuzzy sets can be used to model fuzzy probability statements such as if the occurrence of an an event is very probable [Bie97].

For some tasks in the LL, data must be transformed into fuzzy data. This process is called fuzzification. Fuzzification is also a method used when working with linguistic variables. Within this thesis, a distinction is made between two concepts. In the first concept, further called fuzzification without term set, a crisp value is transformed into a grade of membership related to a single variable, e.g., the variable risk. To understand the concept of fuzzy logic, linguistic variables must be defined and described:

Definition 16.2.3 (Linguistic variable) *"A linguistic variable is characterized by a quintuple $(x, T(x), U, G, M)$ in which x is the name of the variable, $T(x)$ is the term set of x that is the set of names of linguistic values of x with each value being a fuzzy number defined on U; G is a syntactic rule for generating the names of values of x; and M is a semantic rule for associating with each value its meaning."[LEE90, p. 406].*

The application of this definition is illustrated with the term risk in the ecotoxicological risk assessment (3.2.2). To describe the risk caused by pesticides as a linguistic variable, x

represents the term *risk*, the term set of risk can be, e.g., $T(risk) \in \{low, medium, high\}$, dependent on the choice how fine the risk shall be rated. The risk caused by pesticides is rated through the calculated TER, which can be in the interval $U = [0, +\infty]$. G is a rule, which describes the transformation from a crisp value x_0, e.g., TER = 100 means "medium risk" to a given grade, whereby the rules M describes the process of defuzzification to generate a crisp output.

Definition 16.2.4 (Fuzzification) *"The fuzzification is a process of transforming the crisp value into a grade of membership using a membership function of the associated fuzzy set." [Nan12, p. 159]*

In this sense, fuzzification can be regarded as a mapping F from the domain $U \neq \emptyset$ to $[0, 1]$, which transforms a crisp output into a grade of membership:

$$F : U \to [0, 1] \tag{16.5}$$

$$t \mapsto F(t) \tag{16.6}$$

Regarding the membership function in Figure 16.1

$$\mu(x) = \frac{1}{1 + e^{-0.3 \cdot (x-20)}} \tag{16.7}$$

a crisp value $x = 15$ can be fuzzified or transferred into a grade of membership of

$$\mu(15) = \frac{1}{1 + e^{-0.3 \cdot (15-20)}} \approx 0.182. \tag{16.8}$$

In a second concept of fuzzification, a crisp output is transformed into a grade of membership related to a term of a variable, e.g., the grade of membership to medium risk.
The following Figure 16.3 gives an example with three membership functions for the terms

$$low: \ \mu_l(x) = e^{-5 \cdot x^2}, \tag{16.9}$$

$$medium: \ \mu_m(x) = e^{-5 \cdot (x-0.75)^2} \tag{16.10}$$

and

$$high: \ \mu_h(x) = e^{-5 \cdot (x-0.75)^2}. \tag{16.11}$$

Fuzzifying, in the sense of the second concept, would mean that an value of $x = 1$ represents the following fuzzified values

$$A(1, low) = e^{-5 \cdot 1^2} \approx 0.007, \tag{16.12}$$

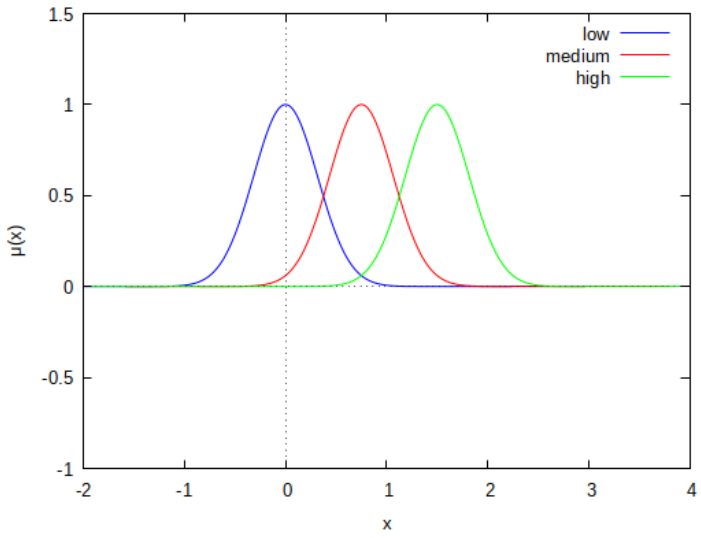

Figure 16.3: Graph of three membership functions, representing the terms low, medium, and high (figure generated with *GNU Octave*).

$$A(1, medium) = e^{-5\cdot(1-0.75)^2} \approx 0.732, \tag{16.13}$$

$$A(1, high) = e^{-5\cdot(1-0.75)^2} \approx 0.287. \tag{16.14}$$

In other words, a crisp value of $x = 1$ belongs to a grade of 0.007 to the fuzzy set *low*, a grade of 0.732 to the fuzzy set *medium*, and a grade of 0.287 to the fuzzy set *high* in relation to the given membership functions.

Through fuzzification, concrete information can be lost, e.g., a sharp value. However, through the transformation into grade of memberships, comprehensiveness can be gained. The use of fuzzy sets and their advantages in contrast to the crisp set theory is demonstrated in the following example.

An example of the use of fuzzy logic might be the ecotoxicological risk assessment for pesticides. If it is defined that a TER that causes a risk has a grade of membership of 1 and a TER that causes no risk has a grade of membership of 0, this can be described with the following function expression:

$$c : \{x \mid x \in \mathbb{R}, 0 \leq x \leq \infty\} \rightarrow \{0, 1\} \tag{16.15}$$

$$x \mapsto \begin{cases} 0 & , \; if \; x \geq 100 \\ 1 & , \; else \end{cases} \tag{16.16}$$

Equation 16.18 gives an example for a membership function by using fuzzy set theory:

$$f : \{x \mid x \in \mathbb{R}, 0 \leq x \leq \infty\} \rightarrow \{0, 1\} \tag{16.17}$$

$$x \mapsto 1 - \frac{1}{1 + e^{-x+105}} \tag{16.18}$$

According to [Com02] and equation 3.9 a TER of 99.9 means an acute risk for the tested aquatic organism, while a TER of 100 means no risk for the organism although the calculated TER is nearly the same. By considering deviances in the determination of toxicological endpoints for the same species under the same test conditions in different laboratories [BBB+91], it will be clear why a sharp threshold between the values which mean a risk and no risk makes little sense. Figure 16.4 shows the relation between the TER values and the grade of risk. The red line shows the grade of risk with a crisp threshold; the blue line shows the grade of risk in the sense of fuzzy set theory. According to 16.18, 16.4, and 16.16, with classical set

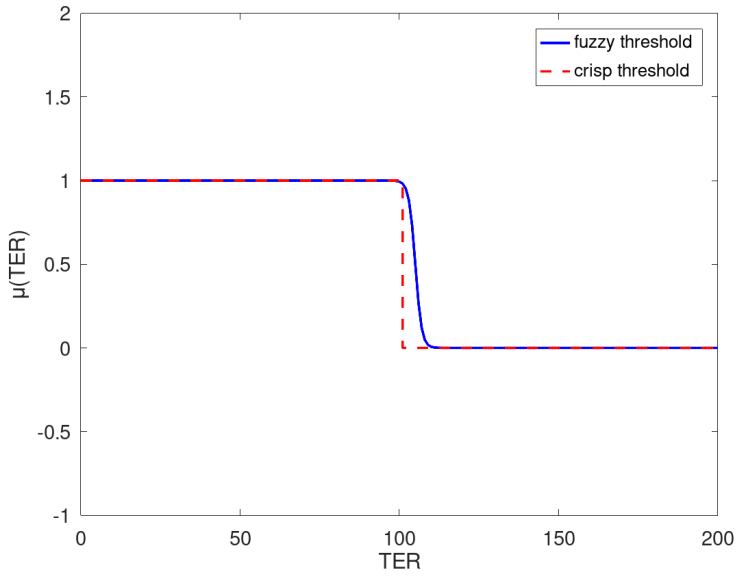

Figure 16.4: Risk and the TER of a certain compound: crisp threshold according to [Com02] (red) and in the sense of fuzzy set theory (blue) (figure generated with *GNU Octave*).

theory as proposed by *Cantor*, a TER of 105 would mean that there is no risk, $c(105) = 0$, and a TER of 95 would mean that there is a risk, $c(95) = 0$. By using fuzzy set theory, a TER of 105 means that there is a risk with a degree of 0.5, $f(105) = 0.5$, and with a TER of 95, there is a risk with degree near to 1. The last step in this example is a fuzzification according to definition 16.2.4 Within this chapter, the basic principles of fuzzy sets and their application were introduced. The following chapter gives an overview of the characteristics of

fuzzy sets.

16.2.2 Characteristics of fuzzy sets

As mentioned above, this thesis deals with risk maps and grades of risk for human and ecosystem health. In the following section characteristics of fuzzy sets and their membership functions are introduced and examples are given how these characteristics can be used in the described LL approach for example for generating risk maps. The characteristics introduced in this section are only a part of the available characteristics, which are necessary in the framework of this thesis.

Definition 16.2.5 (Complement of fuzzy set) *The complement of a fuzzy set A, which is characterized by the membership function so that it is true for all* $x \in \Omega$ *is:*

$$\mu_{\bar{A}}(x) = 1 - \mu_A(x). \tag{16.19}$$

[Bie97]

The complement of fuzzy sets can be used, for example, if there is a map and data available showing the grade of risk, and if derived from this data, locations should be determined where there is no risk.

Definition 16.2.6 (Cardinality of finite fuzzy set) *The cardinality of a finite fuzzy set A over* Ω *(*$|A|$*) is equal to the mathematical sum of the grades of membership of its elements:*

$$\mid A \mid = \sum_{x \in \Omega} \mu_A(x) \tag{16.20}$$

[Bie97]

By comparing the cardinality of two fuzzy sets, e.g., of the elements of risk maps for substance A and for substance B, the average grade of memberships, e.g., the grade of risk for human health, can be compared.

Definition 16.2.7 (Height of a fuzzy set) *The height of a fuzzy set A (hgt(A)) is the supremum of its grades of memberships.*

$$hgt(A) = sup\ \mu_A(x) \tag{16.21}$$

[Bie97]

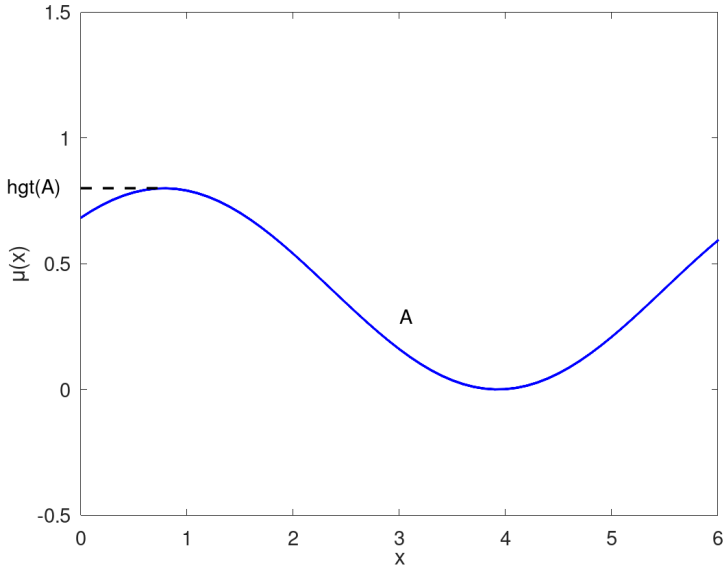

Figure 16.5: Height of a fuzzy set A (figure generated with *GNU Octave*).

The height of the elements of a risk map gives the maximum of the grade of risk existing in a considered area.

Definition 16.2.8 (Normalized fuzzy set) *A fuzzy set A is normalized if the height of A is equal to 1:*

$$hgt(A) = 1 \tag{16.22}$$

[Bie97]

If the height of a fuzzy set is unequal to 1, it is called subnormal.
Sometimes it makes sense to normalize a subnormal fuzzy set, e.g., the grade of memberships of the elements of a risk map. Thus, the relative risk compared to the maximum risk can be calculated, and every grade of risk can be simply related to the supremum of the grade of risk without the knowledge about the accurate supremum value.
With the help of a scale transformation, every subnormal fuzzy set A_s can be transformed into a normalized fuzzy set A_n. The transformation function is called *norm*.

$$A_n = norm(A_s) \tag{16.23}$$

Its membership function $\mu_{A_n}(x)$ can be calculated with the following equation:

$$\mu_{A_n}(x) = \frac{\mu_{A_s}(x)}{hgt(A_s)} \tag{16.24}$$

In the following figure 16.6, a normalized fuzzy set is visualized.

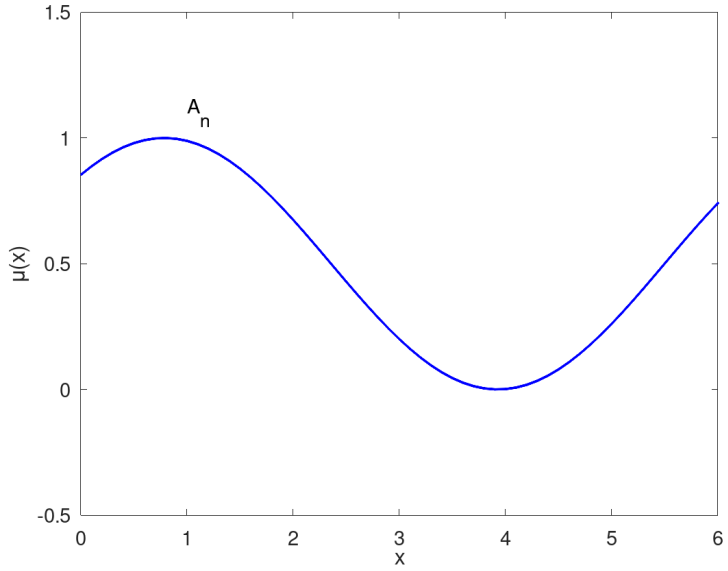

Figure 16.6: Graph of a normalized fuzzy set A_n with a height of 1 (figure generated with *GNU Octave*).

Definition 16.2.9 (Support of a fuzzy set) *The support of a fuzzy set A over Ω, written as $supp(A)$, is the set of all elements of Ω whose grades of membership are greater than 0:*

$$supp(A) = \overline{\{x \ \in \Omega \mid \mu_A(x) > 0\}} \tag{16.25}$$

[Bie97]

Definition 16.2.10 (Core of a fuzzy set) *The core of a fuzzy set A over Ω, written as $core(A)$, is the set of all elements of A whose grades of membership are equal to 1:*

$$core(A) = \{x \ \in \Omega \mid \mu_A(x) = 1\} \tag{16.26}$$

[Bie97]

Definition 16.2.11 (α-cut of a fuzzy set) *The α-cut of a fuzzy set A over Ω, written as $A^{\geq\alpha}$, with $\alpha \in [0, 1]$, is the set of all elements of A whose grades of membership are greater than or equal to the threshold level α:*

$$A^{\geq\alpha} = \{x \ \in \Omega \mid \mu_A(x) \geq \alpha\} \quad ; \ \alpha \in [0; 1] \tag{16.27}$$

[Bie97]

Analogous to definition 16.2.11, the *strong $\alpha - cut$* can be defined as follows:

Definition 16.2.12 (Strong α-cut) *The strong α-cut of a fuzzy set A over Ω, written as $A^{>\alpha}$, with $\alpha \in [0,1]$, is the set of all elements of A whose grades of membership are greater than the threshold level α:*

$$A^{>\alpha} = \{x \ \in \Omega \mid \mu_A(x) > \alpha\} \quad ; \ \alpha \in [0;1] \tag{16.28}$$

[Bie97]

In terms of risk maps, the support of a fuzzy set contains all elements in which risk occurs, the core of a normalized fuzzy set contains all elements with a maximum grade of risk, the *strong* $\alpha - cut$ all elements with a grade of risk above a threshold level, e.g., above an acceptable risk, and the $\alpha - cut$ all elements with a grade of risk equal to or greater than the acceptable risk.
With the help of the characteristics mentioned in this chapter, it is possible to describe and to characterize single fuzzy sets. In the following section 16.2.3, the relations of fuzzy sets are introduced.

16.2.3 Relations of fuzzy sets

This thesis deals, inter alia, with maps displaying the grade of risk for human health caused by pesticides at a specific location, expressed by the grade of membership. As previously mentioned, risk maps and their elements can be described as elements of fuzzy sets. To compare, overlay, or intersect these risk maps and their elements, it is necessary to relate these fuzzy sets to each other. The following section describes how fuzzy sets can be related to each other.

Definition 16.2.13 (Equality of fuzzy sets) *Two fuzzy sets, A and B over Ω, are equal if and only if their membership functions are equal:*

$$A = B \Leftrightarrow \mu_A(x) = \mu_B(x) \quad \forall x \in \Omega \tag{16.29}$$

[Bie97]

With this definition, the inequality of two fuzzy sets can be defined.

Definition 16.2.14 (Inequality of fuzzy sets) *Two fuzzy sets A and B are unequal if and only if one element of the fundamental set Ω exists, for which its grades of membership are unequal:*

$$A \neq B \Leftrightarrow \mu_A(x) \neq \mu_B(x) \quad \exists x \in \Omega \tag{16.30}$$

[Bie97]

According to these definitions, two risk maps are equal if at every location the grade of risks are the same. Two risk maps are unequal if at least at one location the grade of risks is unequal.

Definition 16.2.15 (Subset of a fuzzy set) *Let A and B be two fuzzy sets over Ω. A is a subset of B ($A \subseteq B$) if and only if for all elements of Ω, the grades of membership of A do not exceed the grades of membership of B:*

$$A \subseteq B \Leftrightarrow \mu_A(x) \leq \mu_B(x) \quad \forall x \in \Omega \tag{16.31}$$

[Bie97]

In several human health problems, an acceptable risk is defined by a threshold value at which the occurrence of an event is very improbable [Org97]. This acceptable risk can have a spatial dimension; thus, it can vary at different geolocations, e.g., because of the higher sensitivity of children, the acceptable risk where a kindergarten is located is below the acceptable risk at a location where only adults are located.

To gain knowledge about whether the grade of risk for humans in an specific area is acceptable, e.g., to use it as a settlement area, you can check whether the grade of membership of the risks for humans is a subset of the grade of membership of the acceptable risk.

Similar to classical logic, the union and intersection of fuzzy sets must be defined. Functions that fulfill minimal requirements for a intersection operator are called t-norms; functions for the union are called t-conorm. The operators defined in 16.2.2 are those suggested by *Zadeh*. However, in some cases, others' operators are also usable. To use them, they must fulfill conditions called t-norms and t-conorms. T-norms are usable for calculating the intersection of two fuzzy sets, and t-conorms for the union of the fuzzy set.

Definition 16.2.16 (T-norm) *A t-norm is a function $\top$: [0,1] $\times$ [0,1] $\longrightarrow$ [0,1] that fulfills for all a,b,c,d $\in$ [0,1] the following conditions:*

(i)	$\top(a,1) = a$	*identity element*	*[Bie97]*
(ii)	$a \leq b \wedge c \leq d \Rightarrow \top(a,c) \leq \top(b,d)$	*monotonicity*	
(iii)	$\top(a,b) = \top(b,a)$	*commutativity*	
(iv)	$\top(a,\top(b,c)) = \top(\top(a,b),c)$	*associativity*	

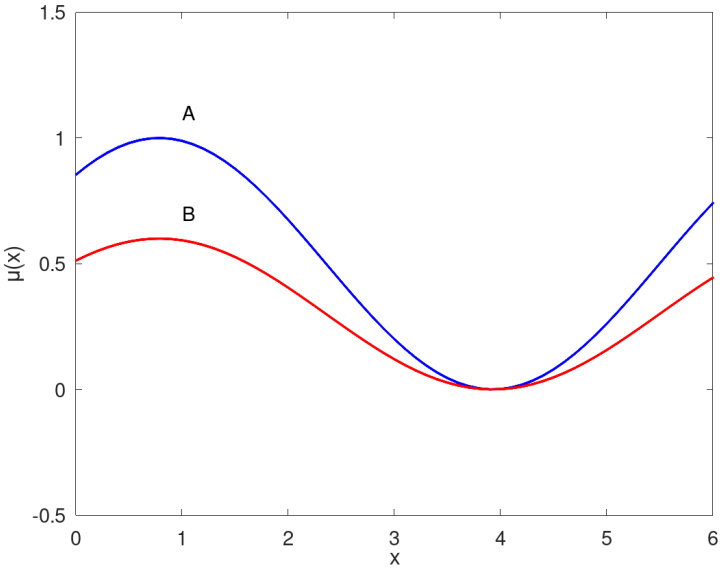

Figure 16.7: Example for a subset B of the set A: $B \subseteq A$ (figure generated with *GNU Octave*).

Definition 16.2.17 (T-conorm) *A t-conorm (s-norm) is a function* $\perp$*: [0,1]* $\times$ *[0,1]* $\longrightarrow$ *[0,1] that fulfills for all a,b,c,d* $\in$ *[0,1] the following conditions:*

(i)	$\perp(0, a) = a$	*identity element*	
(ii)	$a \leq b \wedge c \leq d \Rightarrow \perp(a, c) \leq \perp(b, d)$	*monotonicity*	*[Bie97]*
(iii)	$\perp(a, b) = \perp(b, a)$	*commutativity*	
(iv)	$\perp(a, \perp(b, c)) = \perp(\perp(a, b), c)$	*associativity*	

It is important for such operators that they have as much as possible in common with the classical set operators.

Through the conditions (iii) and (iv), the t-norm and t-conorm operators are such that the results of the union or intersection are independent from the order of the fuzzy sets. Condition (ii) is responsible for the fact that if the grade of membership of one operand increases, the result of the operation does not decrease.

The first three axioms of the t-norm and t-conorm definition lead to the following facts for the logical endpoints 0 and 1:

$$\top(0, 1) = \top(1, 0) = \top(0, 0) = 0, \quad \top(1, 1) = 1 \tag{16.32}$$

and

$$\perp(0, 1) = \perp(1, 0) = \perp(1, 1) = 1, \quad \perp(0, 0) = 0 \tag{16.33}$$

The previous equations make clear that on the endpoints 0 and 1, the t-norms and t-conorms act like the classical logical operators AND and OR.

According to [NKK94], t-norms and t-conorms can be transferred into each other:

$$\perp(a,b) = 1 - \top(1-a, 1-b) \quad \forall a,b \in [0,1] \tag{16.34}$$

and

$$\top(a,b) = 1 - \perp(1-a, 1-b) \quad \forall a,b \in [0,1] \tag{16.35}$$

Regarding definitions 16.2.16 and 16.2.17, all functions fulfilling the given axioms can be used to calculate the intersection or union of fuzzy sets.

Often used t-norms $\top$ and t-conorms $\perp$ are listed in table 16.1: According to [Roj96], by

Table 16.1: Often used t-norms and t-conorms according to [NKK94]

$a \cap b$	$a \cup b$
$\top_{min}(a,b) = min\{a,b\}$	$\perp_{min}(a,b) = max\{a,b\}$
$\top_{Luka}(a,b) = max\{0, a+b-1\}$	$\perp_{Luka}(a,b) = min\{0, a+b-1\}$
$\top_{prod}(a,b) = a \cdot b$	$\perp_{prod}(a,b) = a+b-a \cdot b$

using $\perp_{min}(a,b)$, the union can be extended to the following equation:

$$\mu_{A \cup B}(x) = \mu_A(x) \vee \mu_B(x) = max\{\mu_A(x), \mu_B(x)\} \tag{16.36}$$

If $\perp_{min}$ is used as the the union operator of two fuzzy sets A and B ($A \cup B$) with the membership functions $\mu_a : \Omega \to [0,1]$ and $\mu_b : \Omega \to [0,1]$, the resulting fuzzy set has the membership function

$$\mu_{A \cup B} : \Omega \to [0,1], \tag{16.37}$$

so that for all $x \in \Omega$ the function

$$\mu_{A \cup B}(x) = max\{\mu_A(x), \mu_B(x)\} \tag{16.38}$$

is true.

The union operator can be used if there are two risk maps that are independent from each other to combine them, e.g., if one risk map visualizes the grade of risk for species A and the other risk map the grade of risk for species B. If both species do not interact with each other, the union operator can be used to determine the risk for the biocoenosis.

In figure 16.8, the graph of the union of two fuzzy sets is visualized in which $\perp_{min}$ was used as union operator.

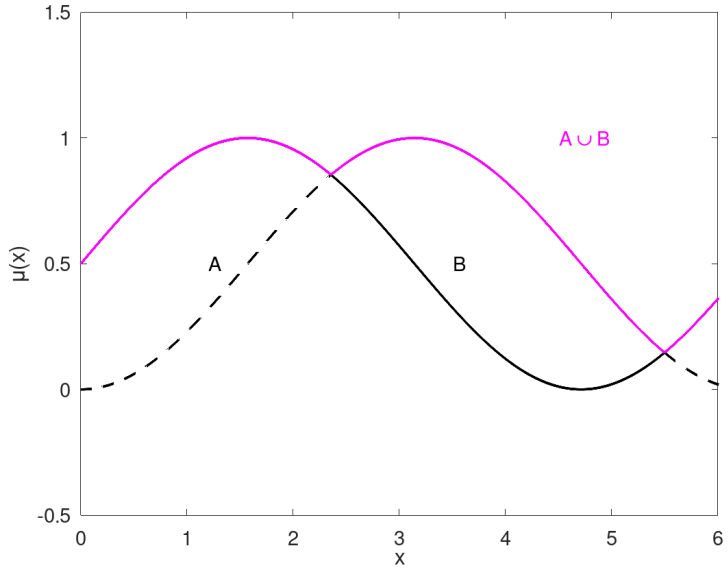

Figure 16.8: Example for the union of two fuzzy sets A and B by using $\perp_{min}$ (figure generated with *GNU Octave*).

According to [Roj96], by using $\top_{min}(a, b)$ the intersection be extended to the following equation:

$$\mu_{A \cap B}(x) = \mu_A(x) \wedge \mu_B(x) = min\{\mu_A(x), \mu_B(x)\} \tag{16.39}$$

If $\top_{min}$ is used as intersection operator of two fuzzy sets A and B ($A \cap B$) with the membership functions $\mu_A : \Omega \rightarrow [0, 1]$ and $\mu_B : \Omega \rightarrow [0, 1]$, the resulting fuzzy set has the membership function

$$\mu_{A \cap B} : \Omega \rightarrow [0, 1], \tag{16.40}$$

so that for all $x \in \Omega$ the function

$$\mu_{A \cap B}(x) = min\{\mu_A(x), \mu_B(x)\} \tag{16.41}$$

is true.

The intersection operator can be used in the framework of risk maps if a risk only occurs when two risk risk factors are present at the same location and time to determine areas and times where the risk occurs.

Figure 16.9 gives an example of the fuzzy intersection by using the $\top_{min}$ operator: The *min* and *max* operators are operators belonging to classical logic. For some tasks, it is necessary to use other operators such as the product or bounded sum. To use such operators, it is necessary that they fulfill the mentioned conditions for t-norm and t-conorm. In other cases,

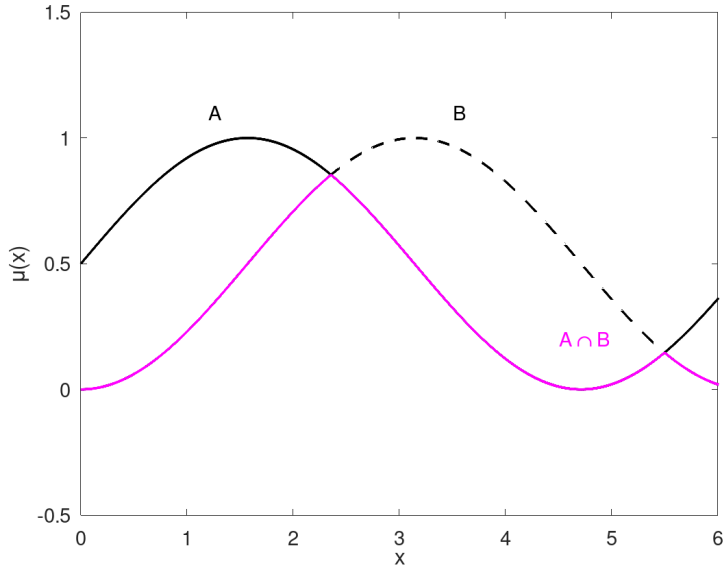

Figure 16.9: Example for the intersection of two fuzzy sets A and B (figure generated with *GNU Octave*).

it might be necessary to define new operators, e.g., if how the output of a process looks and the related input variables are known. To gain an operator, connecting the input parameters with the output new operators can be defined that fulfill the mentioned conditions. This can be regarded as an optimization problem.

Definition 16.2.18 (Bounded sum) *The bounded sum of two fuzzy sets A and B with the membership functions μ_A:$\Omega \rightarrow$[0, 1] and μ_B:$\Omega \rightarrow$[0, 1] is a fuzzy set with the membership function $\mu_{A \oplus B}$:$\Omega \rightarrow$[0, 1]:*

$$\mu_{A \oplus B}(x) = min\{1, \mu_A(x) + \mu_B(x)\}, \quad \forall x \in \Omega \tag{16.42}$$

[Gra95]

The bounded sum can be used if there are two risk maps for substances A and B and if they interact in an additive way. This means that the resulting grade of risk is the sum of the grade of risk of substance A and of substance B. If it is defined that a grade of risk of 1 means the highest risk with a defined endpoint like death, then it is clear why the resulting grade of risk cannot be higher than 1. For several tasks within an agrochemical related LL, fuzzy sets must be related to each other. In mathematics, relations can be understood as subsets of the Cartesian product of related crisp or fuzzy sets [Bie97].

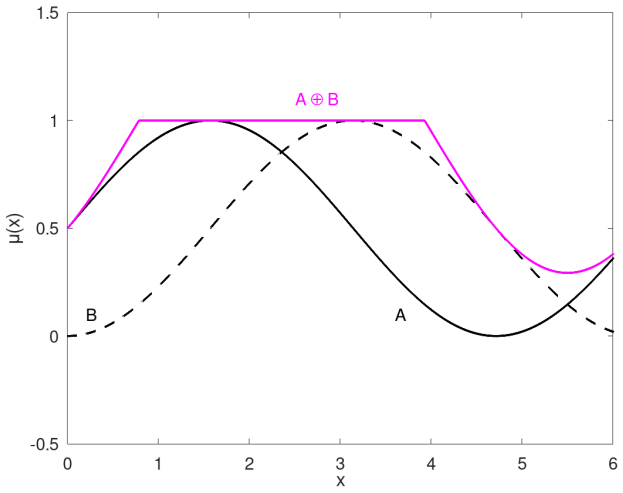

Figure 16.10: Bounded sum of two fuzzy sets A and B (figure generated with *GNU Octave*).

Definition 16.2.19 (Fuzzy relation) *Fuzzy relations are sets whose universe is a Cartesian product $X \times Y$, i.e., a fuzzy relation is a mapping*

$$\rho : X \times Y \to [0, 1]. \tag{16.43}$$

The value $\rho(x, y)$ indicates, how strong x is in relation to y.
[NKK94]

For example, let A be a set with 4 different active ingredients, $A = \{$AI1, AI2, AI3, AI4 $\}$ and B be the set of organs that can be damaged by the active ingredients, $B = \{$ liver, kidney, stomach $\}$. A fuzzy relation $\rho : A \times B \to [0, 1]$ would, for example, describe the relation "it damages the following organs". For example, $\rho(AI2, liver) = 0.9$ would indicate that the substance $AI1$ leads to a damage of the liver with a grade of 0.9 .
As mentioned in definition 16.2.19, fuzzy relations are calculated by the Cartesian product of fuzzy sets:

Definition 16.2.20 (Cartesian product of fuzzy sets) *The Cartesian product of n fuzzy sets $A_1, ..., A_n$ on the underlying sets $\Omega_1, \Omega_2, ..., \Omega_n$, expressed as $A_1 \times ... \times A_n$, is the fuzzy set of all ordered n-tuples (x_1,..., x_n) in the product space $\Omega_1 \times ... \times \Omega_n$ with the membership function:*

$$\mu_{A_1 \times ... \times A_n}(x) = min_i\{\mu_{A_i}(x_i) | x = (x_1, ..., x_n), x_i \in \Omega_i\}. \tag{16.44}$$

[Bie97]

According to definition 16.2.20, the Cartesian product is defined in such a way that the membership of a tuple in the product set is equal to the minimum of the degrees of membership of these elements in the sets from which the product set is formed.
With the help of definitions 16.2.19 and 16.2.20, it is possible to calculate fuzzy relations. For some tasks within the described LL, it is necessary to relate two fuzzy relations with a mathematical technique called composition.

To understand the operation methods of a fuzzy controller it is also important to introduce the composition of two fuzzy relations. To composite the grade of membership of the tuples of two fuzzy relations a union or intersection operator is needed. Such an operator is needed because the composition is not an injective mapping. Thus, an operator such as the *min* or *max* operator are needed to choose the right grade of membership of the tuple. For the conjunction in general, t-norms are used. Such compositions with t-norms are called sup-star-compositions. [Bie97]

Definition 16.2.21 (Sup-star-composition) *The result of the sup-star-composition of two fuzzy relations $R_1 \subseteq A_1 \times A_2$ and $R_2 \subseteq A_2 \times A_3$ is the fuzzy set R_3 of all ordered pairs*

$$\{(x,z)|\exists y \in A_2((x,y) \in supp(R_1) \wedge (y,z) \in supp(R_2))\} \tag{16.45}$$

with the membership function:

$$\mu_{R_3}(x,z) = sup_y \; t(\mu_{R_1}(x,y), \mu_{R_2}(y,z)). \tag{16.46}$$

[Bie97]

In the working field of pesticides and ecotoxicology, such compositions can be used, e.g., if there are two relations, e.g., one relation between the toxicity and the number of killed organisms in a laboratory experiment and a second relation describing the dependence between the number of killed *Daphnias* to the decrease in algae development. With a composition between these two fuzzy relations, it is possible to gain the grade of membership of the resulting relation between the toxicity and the decrease in algae development.
The mechanism behind a *fuzzy logic controller* (section 16.2.4) can be explained with rules based on *IF ... THEN ...* requests. Such *IF ... THEN ...* statements can be regarded as implicational relations [Bie97].

Definition 16.2.22 (Implication) *Let p and q be two propositions. In classical logic, the implication $p \Rightarrow q$ is a conjunction of two propositions that is only false if p is true and q is false. In other words, this issue can be expressed as if p, then q [AHK+15].*

In classical logic, p and q can have the truth values true (1) or false (0). Therefore the classical binary implication operates on $\{0,1\} \times \{0,1\} \rightarrow \{0,1\}$.
In table 16.2, the truth table of the implication in classical two-valued logic is visualized: However, the mentioned concept and definition 16.2 only work for binary logic. As mentioned

Table 16.2: Truth table of the implication $p \Rightarrow q$ for classical two-valued logic

p	q	$p \Rightarrow q$
1	1	1
1	0	0
0	1	1
0	0	1

above, processed data within the LL is sometimes fuzzy. For fuzzy logic, there is also an implication operator, called fuzzy implication, which makes it possible to work logically on multi-valued logic, as fuzzy sets and fuzzy values are processed.
In fuzzy logic, truth values are in the interval $[0,1]$. Therefore, a fuzzy implication operator operates on $[0,1] \times [0,1] \rightarrow [0,1]$. There are different operators available to model a fuzzy implication. As binary logic is a special case of fuzzy logic, the fuzzy implication operators should be such that the classical implication with binary truth values should also fit for these operators [OB87].
There are different operators available that fulfill the mentioned requirements for a fuzzy implication [OB87, Fod91]. In the following table, often used fuzzy implication operators are given in which $\mu_A(x)$ and $\mu_B(y)$ represent the grade of truth of x and y and $\mu_R(x,y)$ the truth value of the implication [CCBC04]:

It is obvious that all fuzzy implication operators listed in table 16.3 lead to the same result for the extreme values $\mu_A(x) = 0$, $\mu_A(x) = 1$ and $\mu_B(y) = 0$, $\mu_B(y) = 1$, as the classical implication for classical two-valued logic. However, for truth values or grade of memberships in the interval $]0,1[$, the truth values for $\mu_R(x,y)$ differ for the listed implication operators. More about the characteristics of the different fuzzy implication operators can be found in

Table 16.3: Some often used fuzzy implication operators [CCBC04]

Operator name	$\mu_R(x,y)$
Lukasiewicz	$min\{1\ , 1-\mu_A(x)+\mu_B(y)\}$
Kleene-Dienes	$max\{1-\mu_A(x), \mu_B(y)\}$
Zadeh	$max\{min\{\mu_A(x), \mu_B(y)\}, 1-\mu_A(x)\}$
Goguen	$\begin{cases} 1 & , \ if\ \mu_A(x) < \mu_B(y) \\ \frac{\mu_B(y)}{\mu_A(x)} & , \ else \end{cases}$

[CCBC04], [Fod91] or [SM87].

In classical binary logic, $p \Rightarrow q$ is equivalent to $\neg p \vee q$ [AHK+15]:

$$(p \Rightarrow q) \Leftrightarrow (\neg p \vee q) \tag{16.47}$$

In fuzzy logic, the truth value of p is represented by a grade of membership $\mu_A(x)$, and the truth value of q is represented by a grade of $\mu_B(y)$. Therefore, $\neg p \vee q$ can be written as

$$\neg\mu_A(x) \vee \mu_B(y). \tag{16.48}$$

According to definition 16.2.5, $\neg\mu_A(x)$ can be written as $1-\mu_A(x)$, resulting in

$$(1-\mu_A(x)) \vee \mu_B(y) \tag{16.49}$$

and by using max as the $\vee$ operator (equation 16.36):

$$max(1-\mu_A(x), \mu_B(y)). \tag{16.50}$$

Transferring $\neg p \vee q$ into fuzzy logic leads to the *Kleene-Dienes* implication operator.

For the other implication operators listed in table 16.3, there are similar reasons they are used as fuzzy implication operators.

With the definitions and results of the last two chapters, the basics about fuzzy logic and how fuzzy sets can be combined logically were introduced. In the following chapter, these findings are necessary to understand the principles of a concept based on fuzzy logic, called fuzzy logic control.

16.2.4 Fuzzy Logic Controller theory

Fuzzy logic control is based on the theory of *Lotfi Asker Zadeh* called fuzzy set theory [Zad65] (section 16.2). The first application of fuzzy control was described by *Mamdani et al.* in 1975

[MA75]. They used a fuzzy logic controller to control the valves of a steam engine. In the 1970s, the research and development of fuzzy controlled systems increased. At the beginning of the 1980s, fuzzy controllers had their first usage in industrial processes [LEE90]. In 1987, the control of an automated underground tram for Sendai (Japan) was developed with fuzzy control. Today, fuzzy control has a wide usage in industrial processes as well as in our everyday lives, such as in a lot of consumer goods, e.g., the automotive industry [PH11]. Today, there are many applications of fuzzy logic control, such as in the controls for wind turbines [SHA+18], photovoltaic systems for achieving maximum power output [LWHJ19], or urban drainage systems [Li20].

Fuzzy control can principally be used everywhere for processes that can be controlled by trained human professionals. It is not necessary that the mechanism or a control algorithm is exactly known [LEE90]. In classical control and feedback control systems, the control is realized through the use of concrete mathematical and natural science models, e.g., expressed as functions like differential equations. In a classical control unit, a process is controlled by measuring a process variable that is compared with a setpoint. The difference between the measured variable and the setpoint is fed back to the system and influences the system. The aim of such a control with feedback is to minimize the difference between the process variable and the setpoint. To determine a function of the setpoints in relation to the process variables it is necessary to have a mathematical model of the control loop. To develop the model for such a controller, special experience is often needed [Trö11].

If it is not possible to find a mathematical model based on the knowledge of the physical structure behind a process and the related theory, it is possible to find such relations with experimental methods. However, such control units often have restricted usability and the quality is mostly limited [Bie97].

However, for many tasks, it is not possible or too difficult to find concrete equations or realize a classical controller with which the control of a system is possible. Missing explicit mathematical equation can be for example attributed to a lack of data, if calculations are computational expensive or if it is wanted to limit the model to a specific class of functions for which solving necessary equations is not possible. Nevertheless, for some of these problems, humans have the ability to describe the control of this process in linguistic expressions and conditions. The control of these processes can be realized with a fuzzy logic controller [Bie97].

The mode of operation of a fuzzy controller is based on a set of several linguistic rules, which

are expressed in *IF ... THEN ...* conditions or implications, for example:

IF inhalation of a toxic substance is high, THEN the fitness of risk mitigation X is high.

The *IF* part in an implication is called antecedent, and the *THEN* part consequent.

These conditions or implications are subsequently evaluated together with the input parameters and the fuzzy controller generates an output value [LEE90].

In general, a fuzzy controller consists of the processes fuzzification, decision-making logic based on the knowledge base, and, in the last step, the defuzzification [LEE90]. The schematic function of the fuzzy controller is visualized in figure 16.11.

The most used fuzzy controller are the control systems described by *Mamdani* [MA75] and

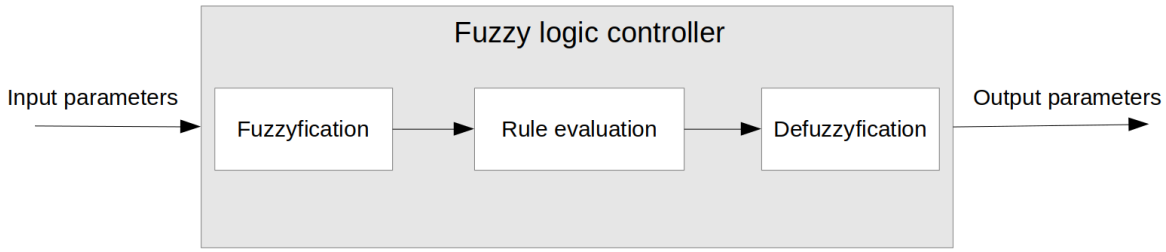

Figure 16.11: Schematic function of a fuzzy controller according to [LEE90] (figure generated with *GNU Octave*).

the system described by *Takagi and Sugeno* [TS83].

Definition 16.2.23 (Mamdani-fuzzy-system) *A Mamdani-fuzzy-system MF_R is a mapping*

$$MF_R : X \to Y, \tag{16.51}$$

with $X = X_1 \times \cdots \times X_n \subseteq \mathbb{R}^n$ named as the domain and $Y = Y_1 \times \cdots \times Y_m \subseteq \mathbb{R}^m$ named as the co-domain. The vectors $x = (x_1, \ldots, x_n) \in X$ and $y = (y_1, \ldots, y_m) \in Y$ are input and output vectors, respectively.

$\mathcal{R}$ is called fuzzy rule base, which determines the structure of the fuzzy system:

$$\mathcal{R} = \{R_1, \ldots, R_r\} \tag{16.52}$$

Each rule $R_k \in \mathcal{R}$ is a tuple of fuzzy sets

$$R_k = (\mu_k^{(1)}, \ldots, \mu_k^{(n)}, \nu_k^{(1)}, \ldots, \nu_k^{(m)}), \tag{16.53}$$

in which $\mu_k^{(i)}$ is a fuzzy set over the range of the input variable x_i and $\nu_k^{(j)}$ a fuzzy set over the range of the output variable y_j.

We define

$$MF_R(x) = y = (y_1, \ldots, y_m), \tag{16.54}$$

with

$$y_j = defuzz\left(\underset{R_k \in \mathcal{R}}{\perp}\{\hat{\nu}_k^{(j)}\}\right) := defuzz\left(\perp\left(\bigcup_{k=1}^{r}\{\hat{\nu}_k^{(j)}\}\right)\right), \tag{16.55}$$

with

$$\hat{\nu}_k^{(j)} : Y_j \to [0,1], \quad y_j \mapsto \top\{\tau_k, \nu_k^{(j)}\}, \tag{16.56}$$

with

$$\tau_k = \top\{\mu_k^{(1)}(x_1), \ldots, \mu_k^{(n)}(x_n)\}. \tag{16.57}$$

In this sense, τ_k is the degree of fulfillment of the fuzzy rule R_k and defuzz a defuzzifying method

$$defuzz : \mathcal{F} \to \mathbb{R}, \tag{16.58}$$

which transforms an output fuzzy set $\hat{\nu}_k^{(j)}$ into a crisp output value.

[BKKN03]

The rule base consists of different *IF ... THEN ...* implications or rules R_r, for example

$$R_r : IF\ x_1\ is\ \mu_a^{(1)}\ and\ ....\ x_n\ is\ \mu_b^{(n)}\ THEN\ y\ is\ \nu_c. \tag{16.59}$$

The definition above is a very basic definition based on the original concept as proposed by *Mamdani*, whereby in the rules the statements of the consequent and antecedent part of the fuzzy implication are only connected with the fuzzy *AND* operator, mathematically expressed as a t-conorm $\perp$. For $\top$ and $\perp$, each of the operators described in Table 16.1 can be used. However the single statements in the *IF ... THEN ...* rules can also be connected with the fuzzy *OR* operator or with a t-norm.

The fuzzy controller of *Mamdani* type can be described in the following way:

For each of the sets of values for the input variables $X_1, ...X_n$ and the output variable Y, appropriate partitions and suitable linguistic terms such as small, medium or large are defined using fuzzy sets, and each of these fuzzy sets is associated with a linguistic value.

The rule base is created by an expert or by a algorithm as described in the following section 16.4.3. Each rule in the rule base can be interpreted as a piece of a piecewise defined function, and each rule defines a fuzzy value of the output or control function. In the operating phase,

the rules are fed by data. The rules are then evaluated, and the degree of fulfillment τ_k is calculated. Then, the output membership functions of the different rules are combined to a single output membership function, and consequently, a crisp output value is determined via a defuzzification method.

Defuzzification means that the grade of the output variable must be defuzzified and, therefore, transformed from a fuzzy into a crisp value. For defuzzification, there are different methods described in literature, such as *Center-of-Area Method* (COA), *Mean-of-Maxima Method* (MOM) [BKKN03], *Center-of-Maximum Method* (COM) [Bie97] and several others [Run96], [JSD22]. The methods differ in their usability. For example, some can only be used with symmetrical membership functions, they can differ in the quality of calculated crisp output or in the method how a crisp output value is derived from an output fuzzy set and in the complexity of the calculation. A detailed description of the different methods for defuzzification can be taken from the cited literature.

One of the most often used methods with low requirements, e.g., on the output membership functions, is COA, which will be briefly described in the following: In the COA, the x-coordinate of the centroid of the area between the x-axis and the output fuzzy set is used as the crisp output value y_i of the FLC [BKKN03].

Definition 16.2.24 (Centroid of a function) *The x-coordinate of the centroid x_s of a function $f(x)$ within the borders a and b is calculated by*

$$x_s = \frac{\int_a^b f(x) \cdot x \; dx}{\int_a^b f(x) \; dx}, \tag{16.60}$$

$\int_a^b f(x) \; dx \neq 0$
[BKKN03]

Therefore, the crisp output value y_c of a FLC of *Mamdani* type is calculated as the x-value of the centroid of the calculated output membership function $\mu_c(y)$

This leads to the following formula to calculate y_c with the COA method:

$$y_c = \frac{\int \mu_c(y) \cdot y \; dy}{\int \mu_c(y) \; dy} \tag{16.61}$$

$\int \mu_c(y)\ dy \neq 0$. Often, especially in FLC like described in 16.2.4, triangle functions in the form of

$$\mu_t(x) = \begin{cases} 1 - \mid \frac{m-x}{d} \mid & ,if\ m-d \leq x \leq m+d \\ 0 & ,else. \end{cases} \tag{16.62}$$

or functions representing a Gaussian distribution, such as

$$\mu_g(x) = e^{-a(x-m)^2} \quad ,a > 0, m \in \mathbb{R} \tag{16.63}$$

are used as membership functions [NKK94].

The application of an FLC of *Mamdani* type in the context of risk mitigation in an LL is described in the following chapter.

The second type of fuzzy controller is called *Sugeno-fuzzy-system.*

Definition 16.2.25 (Sugeno-Fuzzy-System) *A Sugeno-Fuzzy-System SF_R is a fuzzy system that uses a special type of fuzzy rules. Each rule R_k of the rule base $\mathcal{R}$ is a tuple*

$$R_k = (\mu_k^{(1)}, \ldots, \mu_k^{(n)}, f_k^{(1)}(x_1, \ldots, x_n), \ldots, f_k^{(m)}(x_1, \ldots, x_n)). \tag{16.64}$$

$\mu_k^{(i)}$ is a fuzzy set over the range of the input variable x_i and $f_k^{(j)} : X \to Y_i$ a function over the input variables to determine the output variable y_j. With $\vec{x} \in X$, $SF_R(\vec{x})$ is determined by

$$SF_R(\vec{x}) = \vec{y} = (y_1, \ldots, y_m), \tag{16.65}$$

with

$$y_j = \frac{\sum\limits_{r \in \mathcal{R}} \prod\limits_{i=1}^{n} \mu_r^{(i)}(x_i) \cdot f_r^{(j)}(x_1, \ldots, x_n)}{\sum\limits_{r \in \mathcal{R}} \prod\limits_{i=1}^{n} \mu_r^{(i)}(x_i)} \tag{16.66}$$

[Bie97]

The rule base usable for a controller described by *Takagi and Sugeno* consists of rules in following way:

$$IF\ \xi 1\ is\ A_{j1r}^{(1)}\ and\ ....\ and\ \xi n\ is\ A_{jnr}^{(n)}\ THEN\ \eta\ is\ f(\xi_1, ...\xi_n)[\text{Bie97}] \tag{16.67}$$

In most cases, f is a linear combination of the input value with weighted parameters p. Therefore equation 16.67 can be written as:

$$IF\ \xi 1\ is\ A_{j1r}^{(1)}\ and....and\ \xi n\ is\ A_{jnr}^{(n)}\ THEN\ \eta = p_0 + p_1\xi_1 + ... + p_n\xi_n[\text{Bie97}] \tag{16.68}$$

The two systems differ in that the system described by *Mamdani* uses the knowledge base of an operator, whereby the operator can describe their activities depending on the input values. The system described by *Takagi and Sugeno* is used in systems in which controllers are able the control a system, but they cannot describe it with linguistic expressions [Bie97]. The detailed processes behind a fuzzy controller are not part of this work.
In a fuzzy logic controller as proposed by *Takagi and Sugeno*, the rules can be weighted with truth values c, which determines how high the weight of rule is and, therefore, how greatly it influences the overall result. However, the subsequent parts are linear functions of input variables. Unlike the *Mamdani*-type controller, the output parameter is a crisp value. That means that last step called defuzzification is not necessary.
In the following chapter, an application of the described FLC concepts to a risk mitigation task in an LL is described.

16.2.5 The application of fuzzy logic for generating SDSS in an LL approach

One of the aims of this thesis is to develop a framework to give decision support in terms of risk mitigation strategies and to evaluate the individual risk in an agrochemical-related LL. In general, the SDSS should help to define which of the risk mitigation strategies involved in the LL approach and the related research and development cycle have the highest fitness for a person in the LL. The fitness of a risk mitigation strategy is determined for an LL inhabitant in relation to personal, social, health, workflow, and spatial-related parameters by the SDSS. The necessary parameters of the SDSS user are surveyed within the LL approach, e.g., by digital questionnaires. The SDSS determines the fitness of each of the n risk mitigation strategies under survey in the LL and the best fitting are proposed to the user. After the risk mitigation strategy was applied by the user, the fitness of the proposed risk mitigation strategy can be rated with methods that are either direct, through the response of the user, or indirect, through a measurement of the medical parameters of the user, such as the GFR or the stage of CKD.
In an LL approach, this can be described in the following processes: There are only linguistic rules available to describe the fitness of the risk mitigation strategy but no concrete mathematical expression. The parameters of the membership functions used in the linguistic expressions are known or can be determined by an expert from the group of LL stakeholders. This task within the LL approach could be solved with a modified fuzzy logic controller. As

the consequence part of the rules can be mostly expressed as a linguistic variable with a related fuzzy set, a modified FLC as proposed by *Mamdani* would make more sense than an FLC as described by *Sugeno*:

For example, 'go to the next hospital' can be considered a risk mitigation strategy. A stakeholder or expert in the field of medical treatment might have enough knowledge to set up rules, which can be evaluated with an FLC. This task must be done by one or more stakeholders in the LL with special expertise, such as a medical doctor or an expert in logistics.

In the next step, a stakeholder from the field of mathematics or computer science must set up the FLC and preprocessing of data with the following tasks:

- transforming parameters from a nominal or ordinal into an interval or ratio scale,
- selecting software with necessary modules for fuzzy logic like *R* or *GNU Octave*,
- finding appropriate membership functions for the mentioned linguistic variables,
- if necessary, adjusting the available software modules, and
- setting up the FLC with the selected software and modules.

For this approach, the FLC described by *Mamdani* or *Sugeno* must be modified because defuzzification is not necessary in the described approach, as the output value should be a grade of membership or the grade of fitness of the risk mitigation strategy.

To demonstrate the application of an FLC of *Mamdani* type for risk evaluation, the following example is taken:

A value describing the spatial risk for a person in relation to the grade of a disease, e.g., CKD a person is suffering from, expressed in a typical parameter describing the progress of a disease like GFR or the quantity of a biomarker, and to the spatial distance to a medical station that can treat the disease. In the following part, a hypothetical disease is used, as the process described below with finding appropriate membership functions, etc. must be performed with different experts within the research and development cycle in an LL and was not performed in reality for a real disease like CKD.

This can be mathematically expressed as follows:

A mapping M is searched that maps the input values F and D, representing the grade of a disease expressed with a characteristic parameter describing the progress of the disease, such as GFR for CKD, and the distance expressed in km between the person's home and a medical doctor to a spatial risk value S for a person suffering from the regarded disease, for example,

expressed in a value between 0 and 100. For the hypothetical disease, a parameter FR is used to describe the progress of the disease, analog to the GFR:

$$M : F \times D \to S \tag{16.69}$$

$$(f, d) \mapsto s = M(f, d) \tag{16.70}$$

Such a spatial risk value S could be, for example, used to create a map with favorable locations for the home of a person suffering from CKD in relation to the distance to a medical treatment station.

An expert, in the case of an LL, a stakeholder, for example, a medical doctor, might have the following thoughts or has found them in scientific publications:

A person with a medium or high stage of the disease measured in a medium or low filtration rate needs a high number of medical treatments. If the distance to a medical center is great, then the treatment might not be taken as often as needed, and there is a high risk that the disease increases. However, if the person has only a low stage of the disease or high filtration rate, then they can treat themselves, and a high distance leads to a only low risk.

These thoughts can be used for example to determine a individual risk value or to create a risk map with the concept of a FLC.

One of the steps in setting up a FLC is to set up a rule base. According to the described example above, the following rule base is selected:

$$R_1 : \ if\ F\ is\ low\ AND\ D\ is\ high\ THEN\ S\ is\ high \tag{16.71}$$

$$R_2 : \ if\ F\ is\ medium\ AND\ D\ is\ high\ THEN\ S\ is\ high \tag{16.72}$$

$$R_3 : \ if\ F\ is\ high\ AND\ D\ is\ high\ THEN\ S\ is\ low \tag{16.73}$$

One of the tasks in the preparation of a FLC system is to determine the type of FLC to be used. As the consequent part of the rules can be expressed as a fuzzy statement with membership functions, a FLC of *Mamdani* type is used.

To define the different membership functions of the input and output values and related linguistic expressions, different stakeholders like mathematicians, medical doctors, and logistic experts must work together to define the type of the membership functions, e.g., Gaussian or triangle like described in 16.62 or 16.63, and find appropriate partitions of the input space and parameters of the selected membership functions.

Selecting the appropriate membership functions, partitions, function parameters, and t-norms

or t-conorms are tasks that must be performed by experts according to the task the FLC should perform, the performance of the FLC, and the real-life situation the FLC should model. During the research and development cycle in an LL approach, these model properties and parameters might change in order to improve the output of the FLC. In section 16.4, methods are described for the ways partitions, parameters of the membership functions, and a rule base can be developed with machine learning methods from real-world data.

In the following example, it is assumed that the stakeholders and experts decided to use Gaussian distributed membership functions, as described with equation 16.63, with appropriate function parameters and found partitions of the input and output space.

For example, each of the variables of the input and output space is partitioned with three membership functions representing the linguistic expressions *low*, *middle*, and *high* with appropriate function parameters. For the parameter distance D, the following membership functions were defined:

$$\mu_{D_{low}}, \mu_{D_{medium}}, \mu_{D_{high}} : \mathbb{R} \rightarrow [0,1] \tag{16.74}$$

$$d \mapsto \mu_{D_{low}}(d) = e^{-0.1 \cdot d^2} \tag{16.75}$$

$$d \mapsto \mu_{D_{medium}}(d) = e^{-0.1 \cdot (d-5)^2} \tag{16.76}$$

$$d \mapsto \mu_{D_{high}}(d) = \begin{cases} e^{-0.1 \cdot (d-10)^2} & , if\ d \leq 10 \\ 1 & , else. \end{cases} \tag{16.77}$$

with the following graphs:

The distance d is a value related to the location ω_u of the user and the location of the nearest healthcare facility ω_h.

The membership functions for the parameter F representing the filtration rate were defined as follows:

$$\mu_{F_{low}}, \mu_{F_{medium}}, \mu_{F_{high}} : \mathbb{R} \rightarrow [0,1] \tag{16.78}$$

$$f \mapsto \mu_{F_{low}}(f) = e^{-0.0005 \cdot f^2} \tag{16.79}$$

$$f \mapsto \mu_{F_{medium}}(f) = e^{-0.0005 \cdot (f-60)^2} \tag{16.80}$$

$$f \mapsto \mu_{F_{high}}(f) = \begin{cases} e^{-0.0005 \cdot (f-120)^2} & , if\ f \leq 120 \\ 1 & , else. \end{cases} \tag{16.81}$$

and for the output parameter S:

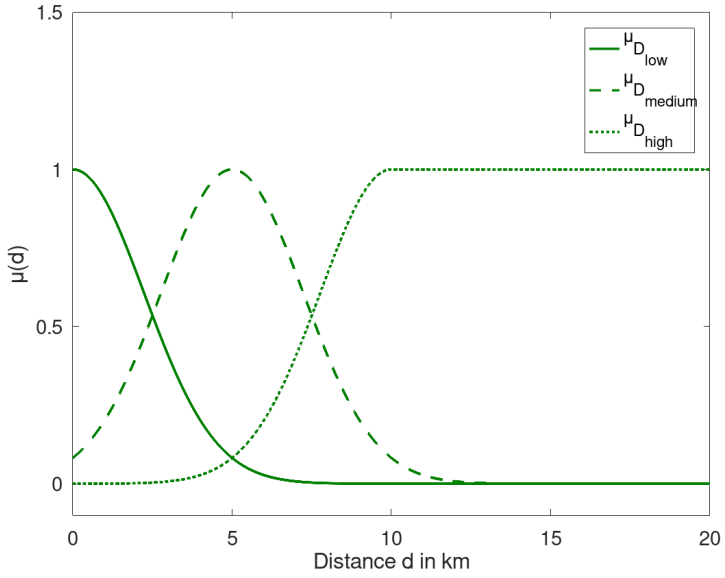

Figure 16.12: Graphs of the membership functions for the parameter Distance D (figure generated with *GNU Octave*).

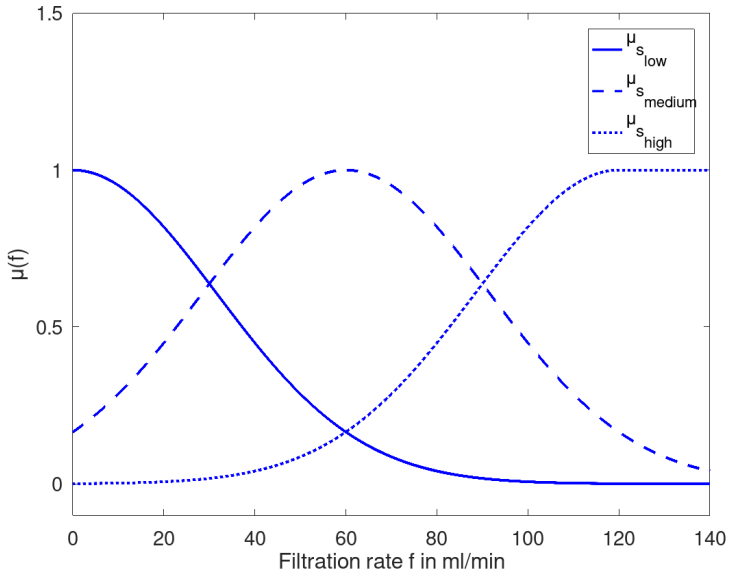

Figure 16.13: Graphs of the membership functions for the parameter Distance F (figure generated with *GNU Octave*).

$$\mu_{S_{low}}, \mu_{S_{medium}}, \mu_{S_{high}} : [0, 120] \to [0, 1] \tag{16.82}$$

$$s \mapsto \mu_{S_{low}}(s) = e^{-0.0007 \cdot s^2} \tag{16.83}$$

$$s \mapsto \mu_{S_{medium}}(s) = e^{-0.0007 \cdot (s-45)^2} \tag{16.84}$$

$$s \mapsto \mu_{S_{high}}(s) = \begin{cases} e^{-0.0007 \cdot (s-90)^2} & , if \ s \leq 90 \\ 1 & , else. \end{cases} \tag{16.85}$$

The next step in setting up a FLC is the selection of appropriate norms and conorms as

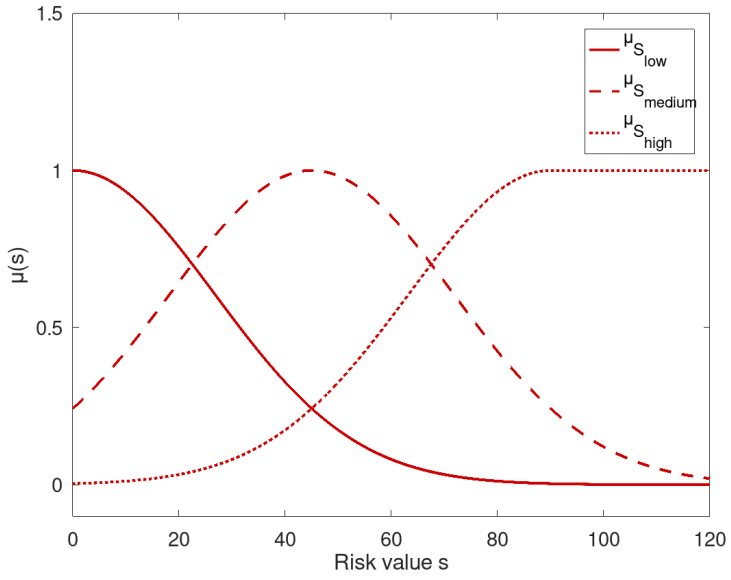

Figure 16.14: Graphs of the membership functions for the output parameter S (figure generated with *GNU Octave*).

described in sections 16.2.16 and 16.2.17. The selection of the operator is performed by stakeholders according to the mathematical requirements for the operator, e.g., for some methods described in section 16.4, differentiable operators are needed and for the desired calculated output values and performance of the FLC.

For this example, it is assumed that the $\top_{min}(a,b) = min\{a,b\}$ and $\perp_{min}(a,b) = max\{a,b\}$ are used as operators. An FLC with these operators is called *Max-Min-* FLC and is a common *Mamdani* type FLC [LZCC05].

To demonstrate the process of deriving an output value with the described FLC, an example is used in which a person has a FR of $f_1 = 60\frac{ml}{min}$ and the distance to the next medical treatment station is about $d_1 = 7\ km$.

The described example can be extended to a spatial task. For example, a map with the locations of medical stations. With a GIS and a raster map, it is possible to calculate the minimum distance of a raster cells middle point to a station. With a given FR, the risk value S can be calculated for each raster cell, leading to a map

Regarding R_1 with

$$R_1 : \; if \; F \; is \; low \; AND \; D \; is \; high \; then \; S \; is \; high \tag{16.86}$$

and

$$\tau_k = \top_{min}\{\mu_k^{(1)}(x_1), \ldots, \mu_k^{(n)}(x_n)\} \tag{16.87}$$

can be calculated by evaluating the membership functions $\mu_{F_{low}}(f) = e^{-0.0005 \cdot f^2}$ for *F is low* and

$$\mu_{D_{high}}(d) = \begin{cases} e^{-0.1 \cdot (d-10)^2} & , if \; d \leq 10 \\ 1 & , else. \end{cases} \tag{16.88}$$

for *D is high* at the crisp values $f_1 = 60$ and $d_1 = 7$.

Using $\top_{min}$, this leads to a degree of rule fulfillment for rule 1 of

$$\begin{aligned} \tau_1 &= \top_{min}\{\mu_{F_{low}}(60), \mu_{D_{high}}(7)\} && (16.89)\\ &= min\{\mu_{F_{low}}(60), \mu_{D_{high}}(7)\} && (16.90)\\ &= min\{e^{-0.0005 \cdot 60^2}, e^{-0.1 \cdot (7-10)^2}\} && (16.91)\\ &= min\{0.165, 0.407\} = 0.165 && (16.92) \end{aligned}$$

Next, the output fuzzy set for rule 1

$$\hat{\mu}_{S_1}(s) \tag{16.93}$$

with

$$s \mapsto \top_{min}\{\tau_1, \mu_{S_{high}}(s)\} \tag{16.94}$$

and

$$\mu_{S_{high}}(s) = \begin{cases} e^{-0.0007 \cdot (s-90)^2} & , if \; s \leq 90 \\ 1 & , else. \end{cases} \tag{16.95}$$

leads to the following fuzzy output for rule R_1

$$\begin{aligned} s \mapsto \nu_1(s) &= \top_{min}\{0.165, \mu_{S_{high}}(s)\} && (16.96)\\ &= min\{0.165, \mu_{S_{high}}(s)\} && (16.97)\\ & && (16.98) \end{aligned}$$

This output fuzzy set ν_1 for rule R_1 is visualized in figure 16.15.

In the same way, an output fuzzy set ν_k can be calculated for every rule R_k.

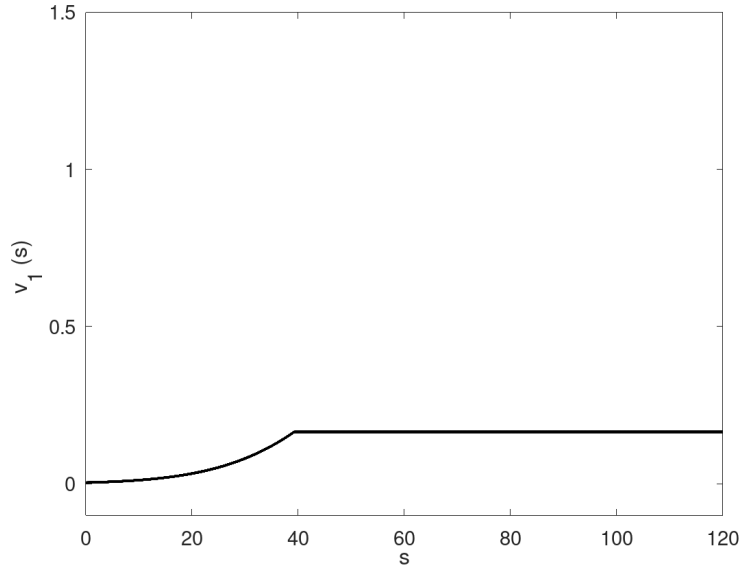

Figure 16.15: Graph of the output membership function ν_1 for the rule R_1 (figure generated with *GNU Octave*).

For rule R_2, the following values are obtained:

$$\begin{aligned} \tau_2 &= \top_{min}\{\mu_{F_{med}}(60), \mu_{D_{high}}(7)\} && (16.99)\\ &= min\{1, 0.407\} = 0.407 && (16.100) \end{aligned}$$

resulting in the following output fuzzy set for rule R_2:

$$\begin{aligned} s &\mapsto \nu_2(s) = \top_{min}\{0.407, \mu_{S_{high}}(s)\} && (16.101)\\ &= min\{0.407, \mu_{S_{high}}(s)\} && (16.102)\\ & && (16.103) \end{aligned}$$

and for rule R_3

$$\begin{aligned} \tau_3 &= \top_{min}\{\mu_{F_{high}}(60), \mu_{D_{high}}(7)\} && (16.104)\\ &= min\{0.165, 0.407\} = 0.165 && (16.105) \end{aligned}$$

and the resulting output fuzzy set for R_3

$$\begin{aligned} s \mapsto \nu_3(s) &= \top_{min}\{0.165, \mu_{S_{low}}(s)\} && (16.106)\\ &= min\{0.165, \mu_{S_{low}}(s)\} && (16.107)\\ & && (16.108) \end{aligned}$$

The resulting output fuzzy sets ν_1, ν_2 and ν_3 are visualized in figure 16.16.

According to definition 16.2.23 and equation 16.55, the resulting output fuzzy set is calculated

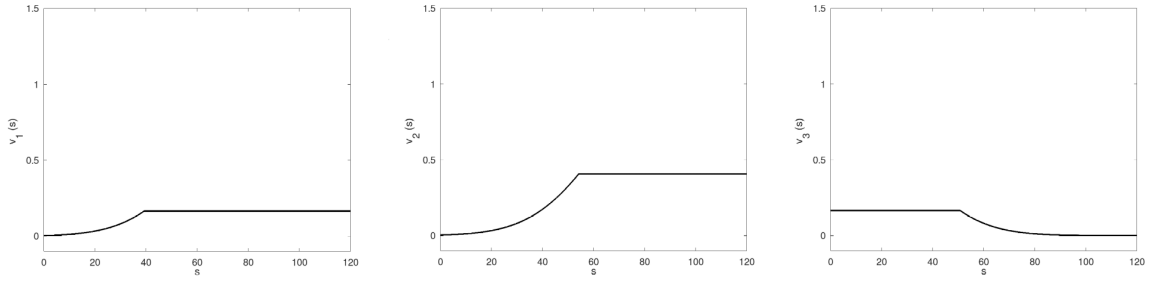

Figure 16.16: Graph of the output membership functions ν_1, ν_2 and ν_3 for the rules R_1, R_2 and R_3 (figure generated with *GNU Octave*).

by

$$\underset{R_k \in \mathcal{R}}{\perp} \{\hat{\nu}_k^{(j)}\} \tag{16.109}$$

whereby $\perp$ represents the *max* operator in the used approach. The resulting output fuzzy set

$$\nu = max\{\nu_1, \nu_2, \nu_3\} \tag{16.110}$$

is visualized in figure 16.17.

The last step in an FLC is the defuzzification through which a crisp output value is obtained.

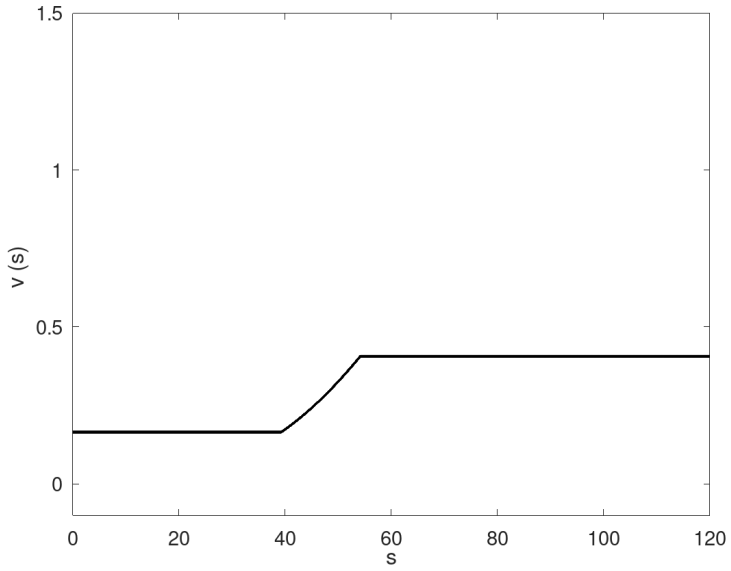

Figure 16.17: Graph of the combined output membership function ν for the rules R_1, R_2, and R_3 (figure generated with *GNU Octave*).

For defuzzification, the COA method is used (section 16.2.4), resulting in the crisp value $Y = s_c$ with

$$y = s_c = \frac{\int_0^{120} \nu(s) \cdot s \; ds}{\int_0^{120} \nu(s) \; ds} \approx 71.1 \tag{16.111}$$

The evaluation of an FLC system described in this thesis for input values of $f_1 = 60\frac{ml}{min}$ and $d_1 = 7 \; km$ results in a crisp output value of $s_c \approx 71.1$.

In the LL approach with an SDSS that evaluates the fitness of different risk mitigation strategies, the output value might be the fitness itself for a risk mitigation strategy. As d is related to the spatial location of the user ω_u, the resulting fitness also has a spatial component. To evaluate the fitness of several risk mitigation strategies, for every risk mitigation strategy an FLC system with a rule base and fuzzy operators must be defined according to the example in this section.

16.2.6 Limitations in the application of a standalone FLC

In the example described in the last chapter, further called ***Case 1***, a method using an FLC of *Mamdani* type was used to describe the spatial fitness of a risk mitigation strategy that can be described with expert knowledge by linguistic *IF .. THEN ...* rules and known membership functions of fuzzy sets.

This method can be also used to develop an SDSS for different risk mitigation strategies through the evaluation of the grade of fitness for different risk mitigation strategies so that the user of the SDSS might receive a ordered list of proposed risk mitigation strategies.

However, in reality there might be difficulties, as expert knowledge is not available to construct a rule base or select appropriate membership functions of the fuzzy sets or related function parameters. This can be summarized in the following cases:

Case 2: *SDSS for risk mitigation strategies that can be described with linguistic rules but unknown membership functions of fuzzy sets*

In this scenario, the linguistic rules are available from the literature but the membership functions of the fuzzy sets cannot be set by experts or stakeholders. The types of the membership functions of the fuzzy sets are known, but their parameters must be determined with mathematical methods.

Case 3: *SDSS for risk mitigation strategies that can be only partly described with linguistic rules*

For a part of the surveyed risk mitigation strategies, a causality between potential risk factors and the fitness were not found until now; there are only evidence-based hypotheses. However, how strong their influence is cannot be determined. If possible, a valid rule base should be extracted from the evidence based hypothetical rules with mathematical methods.

Case 4: *SDSS for risk mitigation strategies that cannot be described with linguistic rules:*
The relationship between the sampled input parameters and the fitness of the risk mitigation strategies cannot be described with linguistic expression. Mathematical methods other than an FLC are needed to perform a mapping.

With learning methods as proposed in the next section, it is possible to find appropriate parameters for linear functions in the consequence part of an FLC of the *Sugeno* type. However, for the SDSS, there are additional requirements, such as being adaptive and a learning system, which is the reason a fuzzy logic controller alone in its fundamental way would not work.

The proposed decisions made by the SDSS might not always be correct. Therefore, the processes generating decision support should be such that the SDSS is able to learn through the direct or indirect feedback of the user such that the decision is better and optimized if the user asks for a decision the next time under same conditions. Learning in the sense of a *Mamdani* type FLC would mean that the parameters of the membership functions are optimized.

But there is another reason why a traditional FLC would not fit. For example, users preferences, their knowledge, and other parameters might change over time. That includes that the process of selecting the best-fitting risk mitigation strategy should also change over time; therefore, the SDSS should be adaptive.

The use of rules of implication together with adaptivity can be achieved by the combination of fuzzy logic and ANN. Additionally, the tasks described in Cases 2 - 4 can be achieved by the combination of FLC with ANN methods, for example, finding the appropriate parameters of the used membership functions (Case 2), finding a rule base (Case 3), or by an ANN alone (Case 4).

16.3 Artificial neural networks

In this section, the concept and basics of ANN are introduced. Through the combination of FLC and ANN, it is possible to make the FLC adaptive, as required in the described LL approach, and optimize the decision finding process. As ANN consist of artificial neurons, the first part of this chapter deals with the biological mode of operation of a neuron and the

concept of artificial neurons.

16.3.1 Artificial neurons and biological background

ANN have an increasing popularity in the scientific discussion and community. ANN consist of connected artificial neurons. The following chapter gives an overview of the behavior and mode of operation of biological neurons and how these concepts are used in artificial neurons. Originally with ANN, the function of the biological nervous system was modeled [Roj96]. Nervous systems consist, inter alia, of neurons or neural cells. Neurons are a type of specialized cells that can be found in most animals belonging to the group of *Metazoa*. In general, neurons process signals in the nervous system. They can receive a signal, and they can give a response related to the incoming signal. They are specialized cells performing saltatory conduction and neurotransmission [And95].

A neuron consists mainly of three types of structures, the soma (cell body) and two types of cellular extensions: dendrites and axons. Dendrites receive signals from other cells; axons transmit signals to other cells. The end of the axons and dendrites are interconnected by synapses. A synapse is a region in which the signal is transmitted from one cell to the other. A neuron can be connected by more than one synapse to other cells. The neuron that sends a signal is called a presynaptic cell, and the neuron receiving the signal is called postsynaptic cell. Cells are not connected in a material sense; there is a small gap between the synapse and the cell to which it is connected, which is called synaptic gap. Communication between cells is performed by electrical signals with the help of neurotransmitters. The direction of signal transmission is defined [CRM06].

If a neuron is excitated, it releases neurotransmitter into the synaptic gap. The connected cell receives the neurotransmitter and a signal is transmitted. Neurotransmitters lead to a depolarization of the connected cell. If the depolarization is higher than a threshold value – called the activation threshold – an action potential is created. However, if the synaptic gap is too wide, electrical signals cannot be transmitted directly [Roj96]. The biological details of neurons are not part of this thesis, as they are not necessary to understand the fundamentals of ANNs. In other words, the action potentials are totaled. If the overall sum is higher than the excitation threshold, the cell is excitated and sends a signal.

The mode of operation of artificial neurons are originally based on the described mode of operation of biological neurons. There are different concepts for artificial neurons described that differ in whether the incoming signal is processed by multiplying with a weight factor,

by the type of information they can process, and by the type of integration and activation function. For example, the *McCulloch-Pitts-Neuron*, the first developed artificial neuron, has unweighted edges and only binary signals can be processed [MP43]. Another important concept for artificial neurons is the *perceptron*, as described in section 16.3.2.

Simple ANNs consist mathematically of nodes and directed or undirected edges. The nodes in a simple ANN are called computing units or artificial neurons. The edges can be regarded as cables that transmit signals x_i from neuron n_i to other connected neurons. A incoming signal can be excitatory or inhibitory. In general, a computing unit n_i is connected with another computing unit n_j by edges through which the signals $x_1, ..., x_n$ are transmitted.

The inside of a artificial neuron can be regarded as divided into two parts. On the one side, there is an integrating function k. With the help of k, the incoming signals are transformed into one value, e.g., by totaling the incoming signals [Roj96].

There are different types of artificial neurons. In most cases and in modern ANN, the input signals $x_1, ..., x_i$ are multiplied with weight factors $w_1, ..., w_i$ and can be described as:

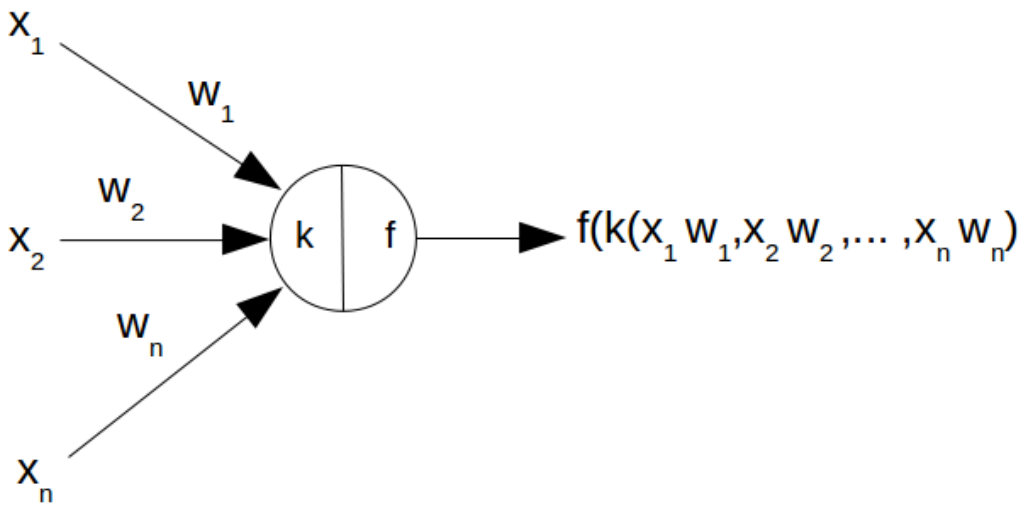

Figure 16.18: Example of a simple neuron with weighted edges (figure generated with *LibreOffice Draw* according to [Roj96]).

$$x = (x_1, x_2, ..., x_n)^T \text{ and } w = (w_1, w_2, ..., w_n)^T \tag{16.112}$$

$$k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \tag{16.113}$$

$$(w, x) \mapsto k(w, x) = \langle w, x \rangle = \sum_{i=1}^{n} x_i \cdot w_i \tag{16.114}$$

On the other side, there is an monotonically increasing activation function f, which uses the single value produced by the integration function, also called net input, to calculate to overall output value

$$y = f(k(x, w)) \tag{16.115}$$

with

$$f : \mathbb{R} \to \mathbb{R} \tag{16.116}$$

$$\lambda \mapsto f(\lambda) \tag{16.117}$$

and the resulting mapping

$$f \circ k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \tag{16.118}$$

$$(x, w) \mapsto (f \circ k)(x, w) = f(k(w, x)) \tag{16.119}$$

In most cases, the integration function k calculates the sum of the incoming signals. Then, the resulting output value y can be calculated with the following formula:

$$y = f\left(\sum_{i=1}^{n} w_i x_i\right). \tag{16.120}$$

[Roj96].

The activation function can be of different function types in relation to the task the neurons should perform. Examples for the activation function are the linear function, the step function, the sigmoid function, or *tanh*.

For some learning algorithms, like the backpropagation algorithm described in section 16.4.1.2, a continuous differentiable activation function is needed. For such purposes, e.g., a sigmoid function, *tanh,* or *arc-tan*, can be used.

In figure 16.19, an example for a logistic or sigmoid function is illustrated:

In general, an artificial neuron can be regarded as a mapping machine through which the input values are transformed into a defined output value.

$$N : \mathbb{R}^n \to \mathbb{R} \tag{16.121}$$

$$x \mapsto N(x) = f(k(x, w)) \tag{16.122}$$

In this chapter, the general mode of operation of a artificial neuron was introduced. For the described SDSS task, a special type of artificial neuron is used, the *perceptron*.

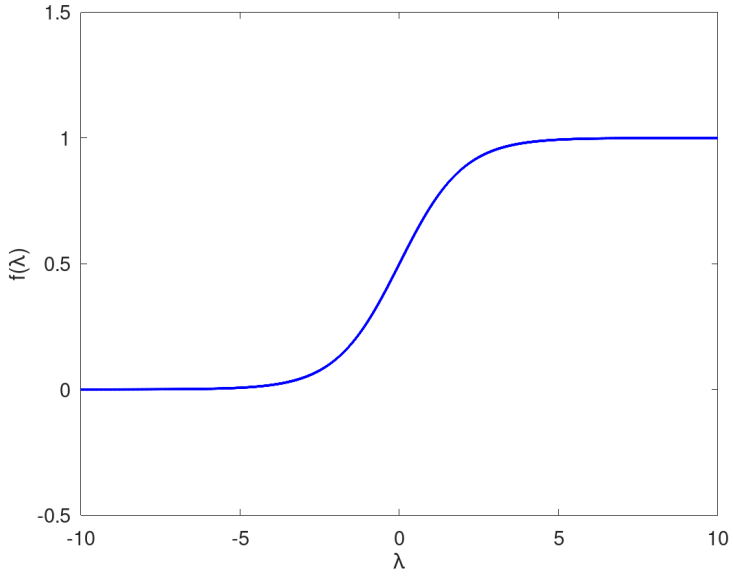

Figure 16.19: Example of a continuous differentiable activation function, here sigmoid function (figure generated with *GNU Octave*).

16.3.2 The perceptron

A *perceptron* is an ANN consisting of only one neuron with weighted edges and a threshold value Θ. A representation of the signal flow in a *perceptron* is visualized in Figure 16.20. A standard single *perceptron*, as proposed by *Rosenblatt* with an integration function calculating the linear combination of the incoming signals and a hard limiter, can be used for solving linear separable problems. With such a hard limiter, the output of the *perceptron* can have only two states. Common states are 0 and 1 or -1 and 1, depending on the used activation function φ [Hay04]. According to figure 16.20, the *perceptron* calculates v and y as

$$v = \sum_{i=1}^{p} x_i \cdot w_i - \Theta \tag{16.123}$$

and

$$y = f(v). \tag{16.124}$$

In the original concept of *Rosenblatt*, a step function is used as activation function f.

$$x = (x_1, x_2, ..., x_n)^T \text{ and } w = (w_1, w_2, ..., w_n)^T \tag{16.125}$$

$$k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \tag{16.126}$$

$$(w, x) \mapsto k(w, x) = \langle w, x \rangle = \sum_{i=1}^{n} x_i \cdot w_i \tag{16.127}$$

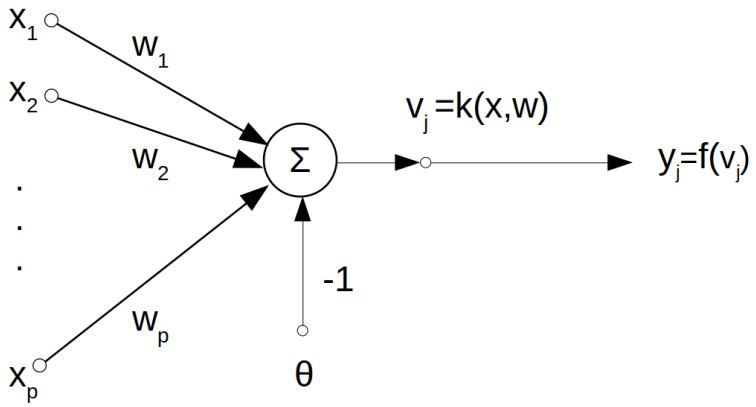

Figure 16.20: Signal flow in a *perceptron* according to [Hay04]. $x_1, ..., x_p$ represent input signals, Θ represents a threshold value, $v = \sum_{i=1}^{p} x_i \cdot w_i - \Theta$, the output after passing the integration function k; f, the activation function; $w_1, ..., w_p$, the edge weights; and y, the output of the *perceptron* (figure generated with *LibreOffice Draw*).

On the other side, there is an activation function f, which uses the single value produced by the integration function to calculate to overall output value

$$y = f(k(x, w)). \tag{16.128}$$

$$f : \mathbb{R} \to \mathbb{R} \tag{16.129}$$

$$\lambda \mapsto f(\lambda) \tag{16.130}$$

with the resulting mapping:

$$f \circ k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \tag{16.131}$$

$$(w, x) \mapsto (f \circ k)(w, x) \quad = \quad y = f(k(w, x)) = f\left(\sum_{i=1}^{n} x_i \cdot w_i\right) \tag{16.132}$$

$$= \quad \left\{ \begin{array}{ll} 1 & , \, if \sum_{i=1}^{n} x_i \cdot w_i > \Theta \\ 0 & , else. \end{array} \right\} \tag{16.133}$$

[Roj96].

Another concept might be to regard the threshold value Θ as the weight of an incoming signal $w_0 = \Theta$ with an incoming signal $x_0 = -1$ itself. In this concept, an incoming signal can be defined as

$$x = (-1, x_1, x_2, ..., x_n)^T \; and \; w = (\Theta, w_1, w_2, ..., w_n)^T \tag{16.134}$$

$$k : \mathbb{R}^{n+1} \times \mathbb{R}^{n+1} \to \mathbb{R} \tag{16.135}$$

$$(w, x) \mapsto k(w, x) = \sum_{i=0}^{n} x_i \cdot w_i \tag{16.136}$$

and

$$f : \mathbb{R} \to \mathbb{R} \tag{16.137}$$

$$\lambda \mapsto f(\lambda) \tag{16.138}$$

with the resulting mapping:

$$f \circ k : \mathbb{R}^{n+1} \times \mathbb{R}^{n+1} \to \mathbb{R} \tag{16.139}$$

$$(w, x) \mapsto (f \circ k)(w, x) \quad = \quad y = f(k(w, x)) = f\left(\sum_{i=0}^{n} x_i \cdot w_i\right) \tag{16.140}$$

$$= \quad \left\{ \begin{array}{ll} 1 & , if \sum\limits_{i=0}^{n} x_i \cdot w_i \geq 0 \\ 0 & , if \sum\limits_{i=0}^{n} x_i \cdot w_i < 0. \end{array} \right\} \tag{16.141}$$

In the described concepts, the *perceptron* has an activation function f which is not linear. The output of f can have only two values, 0 and 1 or -1 and 1, dependent on the used step function. Through the use of a single *perceptron* it is possible to categorize sets of input stimuli x^t into two categories depending on the values of the input stimuli x^t and the calculated output y [And95]. Overall, a *perceptron* as described in this chapter fulfills the task to define two decision rooms ζ_1 and ζ_2, dependent on the output value y of the *perceptron* with $y \in \{0, 1\}$ or $y \in \{-1, 1\}$ separated by the hyperplane

$$\sum_{i=0}^{n} x_i \cdot w_i, \tag{16.142}$$

whereby the weights w_i are adapted by an error-correction rule called the *perceptron* converge algorithm. Such a *perceptron* can categorize a specific input pattern x^t into two linearly separable classes if the input patterns are such that they can be linearly separated. This is a limitation of the classical *perceptron* because many real-world problems are not linearly separable and need more complex decision algorithms [And95].

With an example data set that is linearly separable, a *perceptron* can be trained with supervised learning such that the weights w_i are adapted in a finite time. The *perceptron* can categorize a random data entry of the example data set into the right category. In other words, the *perceptron* will find a solution if a solution exists. A proof of the convergence of the *perceptron* convergence algorithm for linearly separable problems can be found, for example, in [And95] , [Roj96] or [Hay04].

16.3.3 Introduction to ANNs

The last two sections dealt with the basic elements of ANN. The following section gives an overview of how artificial neurons can be connected in nets and about their application.
The first scientific research in the fields of neuronal networks was intended to gain information about the characteristics of the human brain. The first adaptive ANN, the *perceptron*, was described in 1958 by *Frank Rosenblatt* [Ros58]. The *perceptron* has some restrictions that made its use in this time not applicable. Further scientific investigations in the field of artificial neuronal networks began 20 years later after the description of newer, more powerful learning algorithms [NKK94].
ANN is currently an important concept in mathematics and computer science and can perform different tasks like clustering, function approximation, and regressions analysis. This leads to a wide application of ANN, e.g., for pattern recognition purposes [TWPZ12] or in predicting and forecasting issues for groundwater levels [TCS12], optimization problems [MSV18], speech [ZGP+18], text recognition [LZJW20], and many other fields. Against other concepts, ANN have the ability to learn from examples without an explicit algorithm or function that describes the relation between input and output. FLC stand in contrast to the described characteristics for which a concrete rule base is needed, resulting in the non-black-box behavior.
The task that an ANN can perform is determined by the type of neurons used and the way the neurons are connected.
ANN "are systems that are able to modify their internal structure in relation to a function objective. [...] The base elements of the ANN are the nodes, also called processing elements (PE), and the connections" [GB07, p. 1046]. The processing elements of an ANN, also called artificial neurons, are connected with other neurons or with themselves. The topology of an ANN is determined by the connections between the neurons [Roj96]. Each connection or edge between the neurons is related with a number called weight $w_{i,j}$. During the learning phase of an ANN, the weights can change resulting in a changed behavior of the ANN. Signals flowing through an edge are modified through the weights, for example, the incoming signal is multiplied with the related weight. In an artificial neuron, the (modified) incoming signals are totaled, and the neuron's output signal is transformed by a nonlinear function in relation to the sum of the incoming signals [Roj96].
In short, neural networks consist of connected units or neurons. In the neurons, calculations are performed.

In an ANN, neurons are connected mathematically in graphs, often with weighted edges. A detailed introduction to graph theory can be found in section 13.2. ANN can be regarded as a directed and weighted graph.

Definition 16.3.1 (Directed graph) *A directed graph or digraph D is an ordered triple (Q, A, I) consisting of a set Q of vertices or nodes and of a set $A_D \subseteq Q \times Q$ of edges and incidence map I. It is defined that an edge $c = (q_u, q_v) \in A$ with $q_u, q_v \in Q$ is directed from node q_u to node q_v.*
[BKKN03]

In ANN, normally weighted graphs are used. Graphs in general can be either edge- or node-weighted. The edge weights can be used, e.g., to multiply the strength of the incoming signal times the edge weight.
Derived from definition 13.2.3, the structure of a neural net can be defined.

Definition 16.3.2 (Artificial neural net) *An artificial neural net is a directed graph $G = (U, C)$. The nodes $u \in U$ are called neurons or units, and the edges $c \in C$ are called connections. The set U of nodes is divided in the set U_{in} of input neurons, the set U_{out} output neurons, and the set U_{hidden} hidden neurons, with*

$$U = U_{in} \cup U_{hidden} \cup U_{out} \tag{16.143}$$

and

$U_{in} \neq \emptyset$, $U_{out} \neq \emptyset$, $U_{hidden} \cap (U_{in} \cup U_{out}) = \emptyset$.
To each connection $(u, v) \in U$ is a weight $w_{u,v}$ assigned to and to each neuron $u \in U$ three state variables: the network input net_u, the activation act_u, and the output out_u. Each input neuron $u \in U_{in}$ also has a fourth state variable, the external input ext_u. To each neuron $u \in U$, three functions are assigned to:

- *network input function or integrating function,*
- *activation function, and*
- *output function,*

with which the network input net_u, the activation act_u, and the output out_u can be calculated [BKKN03].

The weights $w_{u,v}$ of edges directed from neurons v to neurons u net can be stored in a weight matrix W with

$$W = \begin{pmatrix} w_{1,1} & w_{1,2} & w_{1,3} & ... & w_{1,r} \\ w_{2,1} & w_{2,2} & w_{2,3} & ... & w_{2,r} \\ . & & & & . \\ . & & & & . \\ . & & & & . \\ w_{r,1} & w_{r,2} & w_{r,3} & ... & w_{r,r} \end{pmatrix} \text{ [BKKN03]} \tag{16.144}$$

The characteristics of the weight matrix W is dependent on the network architecture. It can be, for example, symmetric for undirected graphs.

There are two types of ANN differing in whether the incoming signal is transmitted directly into a neuron body or transformed by multiplying the incoming value with a weight factor w: unweighted and weighted ANN. For the task described in this thesis, weighted ANNs are used. In a weighted ANN, the outgoing signal x_i from neuron n_i to a connected neuron n_j is multiplied with a weight factor $w_{i,j}$. However, weighted and unweighted ANNs are equivalent and can be transformed into each other, but the learning process is different. In weighted ANN, learning is performed by adjusting the weights of the input signals [And95].

The weight of each edge $w_{j,i}$ between the output of neuron i and the input of neuron j represents biologically the strength of the synaptic connection of the neuron. One definition interprets learning as a strengthening of the synaptic connection or the value of the weight $w_{j,i}$ between neurons i and j is increased [Roj96].

The graph and the weights of an ANN can be described in an adjacency matrix. In general, ANN can, according to the connections, be divided into feed-forward and recurrent networks. In feed-forward ANN, the signals are transmitted from the input side to the output side without any loops in the connection. In a recurrent network, there are loops, e.g., a neuron in which the signal output is used as the input of the same neuron [Roj96].

Another classification of ANN is synchronous and asynchronous ANN. In synchronous ANN, the output of each neuron in a layer is calculated at the same time. In asynchronous ANN, the output values of the neurons are calculated independently of the other neurons at stochastically determined times [Roj96].

ANNs can be also categorized by supervised or unsupervised learning methods. In a supervised learning task, there is a set of connected input and output vectors available, called the training set. The ANN is trained with this set of connected input and output vectors where

the output of the ANN is compared with the output it should produce, and the deviation between desired output and produced output is used to adjust the parameters of the ANN such that the produced output comes closer to the desired output. There are different algorithms for how such a supervised learning mechanism can be performed. In section 16.4.1.2, a learning algorithm called *error backpropagation* is highlighted in more detail. After an ANN with a supervised learning task is trained, it can be used to predict an output vector for a given input vector. Learning in this sense means that the parameters in an ANN are adjusted in a way that the overall error between desired and created output is minimized. Therefore, a error function is used [Roj96].

In unsupervised learning tasks, the output the ANN should perform is not available. The ANN must organize itself such that similar input vectors are associated with a similar output [Roj96].

According to the task the neural network should perform, the integration function, the activation function, the input space, and the network typology must be selected. By connecting artificial neurons, a network of primitive functions can be created. The type of function the network should perform is determined by the connection type of the artificial neurons, e.g., how they are interconnected, if they process information synchronously or asynchronously, or if there are loops in the connection pattern [Roj96].

ANN can be visualized similar to graphs, an example is illustrated in figure 16.21:

In 1994, the psychologist *Donald O. Hebb* introduced a theory of how learning is performed in the nervous system. He postulated that "[w]hen an axon of cell A is near enough to excite cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased" [Heb49, p. 62]. This means that the more often two neurons are excited together, the more preferentially they react to each other. In other words, neurons that fire together wire together. In a biological way, learning means that the strength of the synaptic connection between neurons is increased. If neuron A is stimulated, neuron B is reacting. In other words, neuron B reacts in relation to the stimulation of neuron A [And95].

Another definition of learning is provided by [Hay04]:

Definition 16.3.3 (Learning of ANN) *"Learning is a process by which the free parameters of a neural network are adapted through a continuing process of stimulation by the environment in which the network is embedded. The type of learning is determined by the manner in which the parameter changes take place" [Hay04, p. 45].*

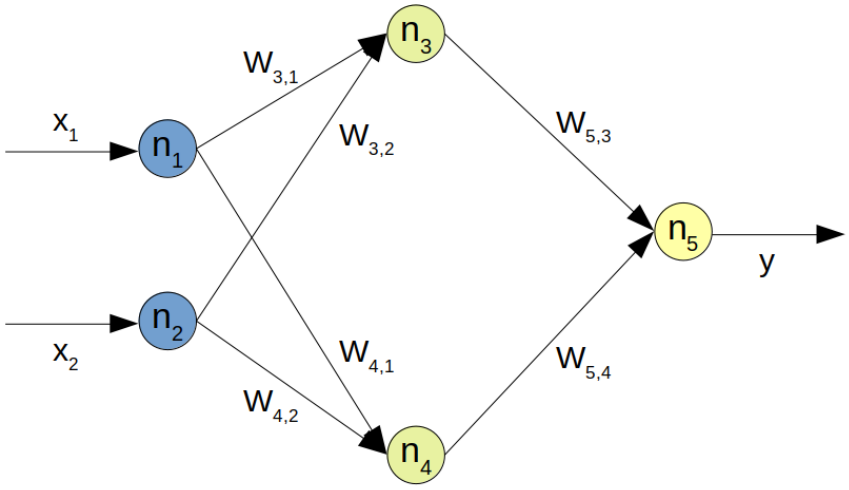

Figure 16.21: An example of a simple neural net with two input units $U_{in} = \{n_1, n_2\}$, two hidden units $U_{hidden} = \{n_3, n_4\}$, and one output unit $U_{out} = \{n_5\}$ that performs a mapping from two input variables x_1 and x_2 to a output variable y. The edges between the neurons are directed, visualized through the arrows at the end of the edges, and weighted, as symbolized by the weights $w_{i,j}$ (figure generated with *LibreOffice Draw*).

According to this definition, learning is divided into the following three steps: Input signals are presented to the ANN (stimulation). Then, this stimulation leads to an change in the ANN. If a new stimulation is presented to the ANN, the ANN reacts in a different way than before, and the stimulation has led to a change in the ANN structure [Hay04].

In computer science, there are different types of learning algorithms, only parts are based on the biological mode of learning. For example, the *backpropagation algorithm*, an often-used technique based on an error-minimization problem to adjust weights during the learning phase in feed-forward layered networks, has no biological analogy.

The stimulation of the ANN leads to a change in the weights of the edges connecting the neurons. When regarding the edge between the neurons i and j, a learning task between two time points or iteration steps t and $t+1$ in a learning algorithm L can be described as the change of the weight $w_{i,j}$, as follows:

$$L(w_t) = w_{t+1} \tag{16.145}$$

$$L : \mathbb{R}^{n^2} \to \mathbb{R}^{n^2} \tag{16.146}$$

$$w \mapsto L(w) \tag{16.147}$$

$$w_{i,j}(t+1) = w_{i,j}(t) + \Delta w_{i,j}(t) \tag{16.148}$$

The learning and the change of $w_{i,j}(t)$ can be expressed by regarding the *rule of Hebb* in the following way:

$$\Delta w_{i,j}(t) = \eta y_i(t) x_j(t) \tag{16.149}$$

whereby η is a positive constant called learning rate, x_i and y_j are input and output signals [Hay04].

If the target response r_k and the real or actual response y_k of a neuron n_k at time t is known, an error signal $e_k(t)$ for neuron k at time point t can be calculated as follows:

$$e_k(t) = r_k(t) - y_k(t) \text{[Hay04]} \tag{16.150}$$

In general, an ANN performs the following task:

$$N : \mathbb{R}^n \mapsto \mathbb{R}^m \tag{16.151}$$

whereby n input values are mapped to m output values.

In the literature, there are different types of ANN mentioned, such as *single* or *multilayer perceptrons* (section 16.3.2), *self-organizing maps*, or *Hopfield nets*, differing in the used neurons, learning methods, net architecture, and the task they can perform [NKK94]. Some of them, usable in the described LL approach, are described in the following chapters.

16.4 The application of ANN in combination with FLC in an LL framework

Mathematical methods to determine the individual risk and the grade of fitness of risk mitigation strategies are required. In section 16.2.4, an approach to determine the fitness of risk mitigation strategies by using an FLC was described. However, this approach has some limitations, such as:

- Linguistic rules are unavailable for some risk mitigation strategies.
- There are unknown membership functions of the used fuzzy sets.
- The rule base is not complete.
- Only an evidence-based hypothesis is available.

Through the use of an ANN or by combining an ANN with an FLC, it is possible to reduce these limitations.

Both ANN and fuzzy logic methods can be used to model expert behavior. However, both techniques have advantages and disadvantages compared with each other, as listed in table 16.4.

Both methods have the advantage that a concrete mathematical process model is not nec-

Table 16.4: Advantages and disadvantages of neural nets and fuzzy systems according to [BKKN03]

Neural net	Fuzzy system
Advantages	
• explicit definition of the mathematical function (input-output behaviour) is not necessary • rule knowledge not necessary • in-output-behaviour determined by different learning algorithms	• explicit definition of the mathematical function (input-output behaviour) is not necessary • a priori rule knowledge usable • easy interpretation and implementation
Disadvantages	
• black box behavior • rule knowledge not extractable • heuristic choice of net parameters • adaption to altered parameters can be difficult and it might be necessary to repeat the learning process • a priori knowledge not usable • learning process does not converge in all cases	• rule knowledge must be available • no learning ability • no formal methods for tuning • semantic problems when interpreting tuned systems • adaption to altered parameter might be difficult • tuning trial might be without success

essary. However, with an FLC, linguistic rules, which can be expressed in *IF ... THEN ...* statements, can be used, which is not possible in an ANN. ANN have the ability to learn from examples, while FLC do not have such an ability. By using an ANN, a training set consisting of tuples of input and related output values is used to train the system in a way that it performs the desired task. A fuzzy system can be easily interpreted and also implemented in which an ANN behaves as a black box, and after the learning process, the rule knowledge is not extractable [BKKN03]. By combining FLC and ANN, the advantages of both systems

can be utilized, the learning ability of ANN and the ability of a system with interpretable rules [VDM04].

Systems that consist of both ANN and FLC methods are called *neuro-fuzzy systems* [BKKN03, VDM04]. *Neuro-fuzzy systems* have a higher performance than simple ANN [LL+91].

[VDM04] and [Abr01] differentiate three types of *neuro-fuzzy systems*: *cooperative*, *concurrent*, and *hybrid neuro-fuzzy systems*. In a *cooperative neuro-fuzzy system*, the ANN and FLC can be regarded as two subsystems in which the ANN is responsible for a preprocessing task, such as finding appropriate rules in a clustering approach or membership functions from training data. After the rules or membership functions are determined, they are used in an FLC in a second phase [VDM04, Abr01].

In a *concurrent neuro-fuzzy system*, the workflow of both components is continuous, with the ANN processes inputting or outputting the data of the FLC [VDM04]. In this approach, the parameters of the fuzzy system are not optimized, and the controller behavior of the whole system is improved through the use of the ANN. The learning ability of the system is only caused by the ANN, and the parameters of the FLC are not changed during learning. However, the interpretability of the results is not given due to the black-box behavior of the ANN [BKKN03].

In contrast to a *cooperative neuro-fuzzy system*, in a *hybrid neuro-fuzzy system*, the ANN and FLC cannot be regarded as two subsystems. A *hybrid neuro-fuzzy system* has a continuous architecture, consisting of both elements of ANN and FLC in which the communication between two subsystems is not necessary. Therefore, learning can be performed in offline as well as online mode. Learning in a *hybrid neuro-fuzzy system* means a change in the network architecture, leading to the adaptation of weights (fuzzy sets) or the deletion or addition of elements (variables or fuzzy rules). Therefore, through learning, explicit knowledge can be gained and interpreted that can be used as a rule base in an FLC. Learning in a *hybrid neuro-fuzzy system* takes place as a supervised or reinforcement learning task. ANN determines, e.g., parameters of membership functions or fuzzy sets with a learning algorithm, that are used within the FLC [BKKN03].

Cooperative and *concurrent neuro-fuzzy systems* have the disadvantage that the results, e.g., the determined fuzzy rules, cannot or cannot fully be interpreted. The advantage of *hybrid neuro-fuzzy systems* lies in the interpretable results, e.g., fuzzy rules and membership functions, the ability to learn from given examples of input, and related output samples [Abr01]. Today, they play a more and more important role in neuro-fuzzy research [dCS20].

In figure 16.22, a schematic representation of the three types of *neuro-fuzzy systems* is visualized. In the LL framework, one of the tasks is to determine a appropriate system to give

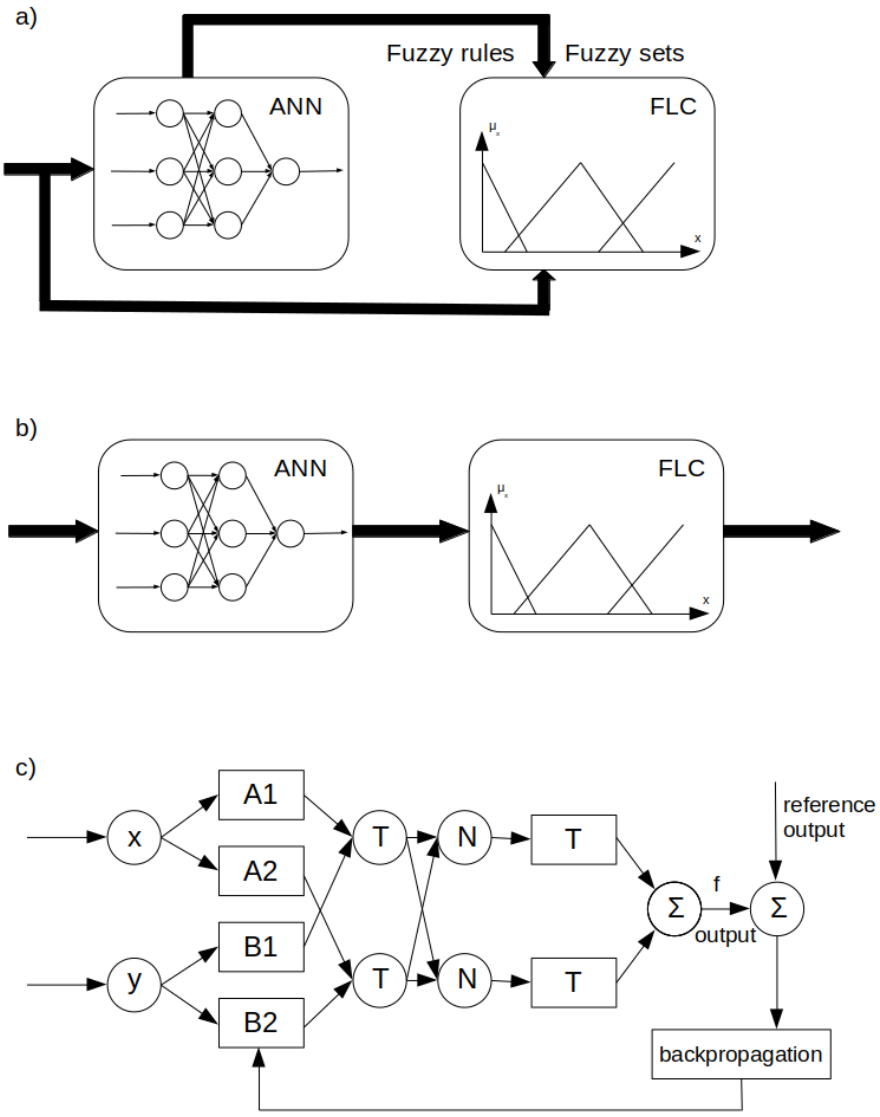

Figure 16.22: Schematic representation of a) *cooperative* [VDM04], b) *concurrent* [BKKN03], and c) a *hybrid neuro-fuzzy system* [EDDSM15]. The visualized *hybrid neuro-fuzzy system* is an *Adaptive Neuro-Fuzzy Inference System* (ANFIS) with a network architecture in which $A_1, ..., B_2$ represent fuzzy sets, $\perp$ t-norms of the antecedent, and the consequent part of a rule, N a normalizer, and Σ the sum function for calculating the output f and a back-propagated error signal (figure generated with *LibreOffice Draw*).

decision support. In general, the use of FLC is favored over the use of an ANN because a priori knowledge can be used in an FLC and because of its easy implementation.

When linguistic rules are available, a combination of a FLC and ANN can be used in which the response of the SDSS user is used to optimize (Case 1) or to find (Case 2) appropriate parameters of the membership functions by the ANN. If there are only evidence based hypotheses without the knowledge of to what extent the rules are valid or if a rule base must be found, a combination of an FLC and an ANN can be used for clustering purposes in which the ANN is responsible for finding a rule base that can be processed by the FLC (Case 3). If there is no a priori knowledge in the type of a linguistic rule base available (Case 4), an ANN alone can be used without an FLC to perform the desired mapping; however, they all come with disadvantages in contrast to an FLC.

Besides Cases 1 to 4, it might be necessary to pre-process input or output data, e.g., because in the input data some items are missing or because the data is noisy. For this task, the use of a *concurrent neuro-fuzzy system* is intended e.g., if datasets gained during surveys in the LL are not complete, whereby the ANN is responsible for completing missing data,e.g., by an *autoassociative neural net*.

The different cases and suitable combinations of ANN and FLC are discussed in the following sections.

16.4.1 A concurrent neuro-fuzzy system to give SDSS with incomplete or noisy data

One task to deal with in the described LL approach is the handling of missing parameters or noisy values in a part of the dataset that are used as input values for an SDSS. In chapter 7, methods to deal with missing data were introduced. In the following chapter, a method is described in more detail that is based on ANN methods and can be used in a *neuro-fuzzy system*.

In the described approach, a method called *autoassociative neural network* is used because of the following reasons: For model-based and regression approaches, different assumptions on the distribution of the data are necessary but cannot be guaranteed. By using *single* and *multiple imputation methods*, too much information about the input space is lost.

That means that a method from the group called imputation based on machine learning is used. k-NN has a high computational cost, as all datasets must be compared with the given input [TCS+01]. Therefore, a method called *autoassociative neural network* is used and

demonstrated in the following chapter.

16.4.1.1 Data completion and noise reduction with autoassociative networks

Autoassociative neural networks are a kind of ANN in which each input vector x^i is associated with itself. A trained autoassociative neural net is able to eliminate noise in the input vector or estimate missing values in the input vector.

Definition 16.4.1 (Autoassociative neural net) *Autoassociative neural network are nets that are trained and with which it is possible to generate an approximated identity mapping for the trained patterns by using a learning algorithm [Kra92].*

Regarding definition 16.4.1, approximated identity mapping means that noisy input data converge towards a trained pattern to reduce the noise.
There are different types of autoassociative networks, characterized by their network architecture and learning algorithms, e.g., based on *Hopfield networks* [RSL$^+$20, HT18], and *multilayer perceptrons* in an *encoder-decoder architecture* [Kra92, BGMS18].
The following section focuses on the use of an *autoassociative memory* for noise reduction and estimating missing values of data gained in surveys in the LL approach. At the beginning of the research and development cycle in an LL approach in less-developed countries, an easily implementable algorithm for the implementation is recommended.
Multilayer networks with one or more hidden layers have a more difficult network architecture and also a higher time demand for network optimization and pruning in contrast to the second type of *autoassociative neural nets*, *Hopfield networks*. Neurons in a discrete *Hopfield net*, as described by [Hop82], [Roj96], and [Hay04], can only have binary or bipolar states like 0 and 1 or -1 and 1. A model with continuous states of the neurons was introduced by *Hopfield* in 1984, the *continuous Hopfield model* with a sigmoid activation function and resulting values in $[0, 1]$ [Hop84]. The decision for which of the ANN types is used as an *autoassociative neural network* in the LL approach lies in the hands of the involved stakeholders.
In the following section, the use of a *multilayer perceptron* is demonstrated because this type of ANN has the broadest application of all available ANN types and can be also used, e.g., for not linear separable classifying problems or in speech and image recognition problems or deep learning [NKK94].

Overall, an *autoassociative neural network* performs the following mapping f:

$$N: \ [0,1]^n \to [0,1]^n \tag{16.152}$$

$$x \mapsto \tilde{x} = N(x) \tag{16.153}$$

whereby x is a disturbed or incomplete input vector and $\tilde{x}$ the undisturbed or completed vector.

16.4.1.2 Autoassociative ANN with supervised learning: multilayer perceptron and the backpropagation algorithm

For the described framework of an *autoassociative neural network*, an ANN with a specific network architecture is used, called *multilayer perceptron.*

Definition 16.4.2 (Multilayer perceptron) *A r-layered perceptron is a neural net with a graph $G = (U, C)$ with*

$$U_{in} \cap U_{out} = \emptyset \tag{16.154}$$

$$U_{hidden} = U^1_{hidden} \cup U^2_{hidden} \cup ... \cup U^{r-2}_{hidden}, \tag{16.155}$$

$$U^i_{hidden} \cap U^j_{hidden} = \emptyset, 1 \leq i < j \leq r-2 \tag{16.156}$$

$$C \subseteq (U_{in} \times U^1_{hidden}) \cup (\bigcup_{i=1}^{(r-3)} U^{(i)}_{hidden} \times U^{(i+1)}_{hidden}) \cup (U^{(r-2)}_{hidden} \times U_{out}) \ for \ r > 2 \tag{16.157}$$

or

$$C \subseteq U_{in} \times U_{out}. \tag{16.158}$$

The net input function v of each hidden and output neuron is the sum of the connection weights multiplied by the input values:

$$\forall u \in U_{hidden} \cup U_{out} : v_u = \sum_{v \in pred(u)} w_{u,v} \cdot out_v \tag{16.159}$$

The activation function f of each hidden neuron is a differentiable and monotonic increasing function:

$$f : \mathbb{R} \to [0,1] \quad with \quad \lim_{x \to -\infty} f(x) = 0 \quad and \quad \lim_{x \to \infty} f(x) = 1 \tag{16.160}$$

The activation function of each output neuron can be a function of the same type as the activation function of a hidden neuron or also a simple linear function.
[BKKN03, NKK94]

In general, a *multilayer perceptron* with supervised learning ability consists of different *perceptrons*, structured in one input layer, one output layer, and one or many hidden layers between them. A *perceptron* is connected with every *perceptron* of the following layer, and there are no cycles in the connections. That means that the signals are transferred successively from one layer to the next until the last layer is reached. This ANN is used in a feed-forward way; input values are presented on the one side, and output values are received on the output side of the network. This type of multilayer networks can be used for universal function approximation [NKK94, Roj96]. An example of a *multilayer perceptron* is visualized in figure 16.21.
To describe the typology and the mode of operation of a *multilayer perceptron*, the terms predecessor and successor of a neuron must be defined:

Definition 16.4.3 (Predecessor and successor) *Let $G = (U, C)$ be a directed graph and $u \in U$ be a node. Nodes of the set*

$$pred(u) = \{v \in U \mid (v, u) \in C\} \tag{16.161}$$

are called predecessors of the node u and the nodes of the set

$$succ(u) = \{v \in U \mid (u, v) \in C\} \tag{16.162}$$

are called the successors of the node u. [BKKN03]

In a *multilayer perceptron*, a learning algorithm called *backpropagation* is used. It is based on gradient descent in which the derivation of the activation function is used to adjust the weights of the ANN in a way that the error of the *multilayer perceptron* is minimized. Therefore, it is necessary that the used activation functions are differentiable. An example for usable activation functions is the sigmoid function, such as the logistic function $f_{log}(x) = \frac{1}{1+e^{-a(x-x_0)}}$.
Finding and optimizing the network architecture, meaning the number of hidden layers, is called pruning [CC11]. To find a fitting network architecture, the scientific modeler needs knowledge of both the scientific discipline to which the data belongs and the scientific discipline of neural networks, e.g., the number of used input and output parameters and the number of neurons and hidden layers are connected [DS06]. The goal of network architecture optimization is to find a network architecture that avoids underfitting and overfitting.
Underfitting means that the ANN has too few neurons and is not able to learn the underlying problem, leading to poor predictions. An overfitting ANN consists of too many neurons and learns unimportant details in the data such as outliers but not the underlying relations that

lead to a good performance on entries of the learning dataset but not on data being noisy or differing from the data set used in the training. Pruning the network architecture should lead to an ANN making appropriate predictions, as well as minimized convergence and training time and avoidance of network over- or underfitting [ABA05].

There are different methods for finding an appropriate network architecture. [CC11] mention a simple trial and error method by adding or removing layers or neurons and testing the performance of the ANN in each trial using constructive and destructive methods, starting with an ANN architecture that is known to have too few or too many neurons and adding or deleting neurons until the desired behavior and performance is achieved. However, there are also learning algorithms available that are able to find a network architecture that fits [TTL95, BS95].

Finding the appropriate network architecture is a task that must be performed by one or more stakeholders in an LL approach.

A *multilayer perceptron* is trained by a set of p different input-output pairs $(x_i, r_i) \in \mathbb{R} \times \mathbb{R}$ with $i \in \{1, ..., p\}$. After an input signal x_i is presented to the network, it is propagated feedforward through the layers, and an output signal y_i is received. This output signal is compared with the target output value r_i for which an error value representing the deviation between r_i and y_i can be propagated back through the net, whereby the weights $w_{u,v}$ of the connections change in a way that the error E of the network is reduced. The error function is dependent on the weights of the network

$$E(W) = E(w_1, w_2, ..., w_n). \tag{16.163}$$

The weights of the network should be adapted in a way that the error of the network is minimized. Therefore, the *backpropagation* algorithm can be related to an optimization problem of the error function. Finding optimal weights for the network means finding a minimum in the error function. In the following, an approach is demonstrated for how a method or algorithm called gradient descent is used for the optimization problem.

Definition 16.4.4 (Gradient) *The gradient of a partial differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ at point x is defined as the vector of the partial derivatives at point x:*

$$\nabla f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{pmatrix} \tag{16.164}$$

[AHK+15]

The gradient is a vector in the domain oriented into the greatest increase of a function, and the negative of the gradient into the steepest descent. *Gradient descent method* is based on the fact that the variables $x_1, x_2, ..., x_n$ are changed in a way that the negative gradient is followed step by step until a local minimum of the function is reached and, in the case of ANN, the minimum of an error function with respect to the weights of the ANN.

In the literature, there are different modifications of the *perceptron learning algorithm*. The following derivation is based on [BKKN03] and [NKK94] .

Overall, for an output neuron, the error signal can be calculated as follows: t represents the number of training pairs presented to the ANN (e.g., the error-signal when the t^{th} training set was presented). The target response r_k and the real or actual response y_k of an output neuron $u_k \in O$; the error signal $e_k^{(t)}$ is given by the following equation:

$$e_k^{(t)} = (r_k^{(t)} - y_k^{(t)})^2 \text{[Hay04]} \tag{16.165}$$

The output signal $y_k^{(t)}$ of a output neuron k can be calculated by

$$y_k^{(t)} = f_k(v_k^{(t)}), \tag{16.166}$$

whereby $v_k^{(t)}$ represents the value after the incoming signals $x_1, ..., x_p$ were processed with the related weights by the integration function, and f represents the activation function of the neuron.

In a *backpropagation algorithm* with *gradient descent*, the change of the weights W can be expressed as

$$\Delta W = -\eta \nabla E(W) \tag{16.167}$$

whereby η is called learning rate. The change of the weights is proportional to the negative gradient and learning rate.

The change of each weight $w_{i,j}$ is calculated as

$$\Delta w_{i,j} = -\eta \frac{\partial}{\partial w_{i,j}} E(W) \tag{16.168}$$

The error function of the overall error of the network can be expressed as the sum of errors of all patterns and can be, for example, expressed as the mean-squared error:

$$E(W) = \sum_{t=1}^{p} E_t(W) \tag{16.169}$$

with

$$E_t(W) = \frac{1}{2} \sum_{u_j \in U_{out}} (r_{t,j} - y_{t,j})^2. \tag{16.170}$$

E_t represents the error of pattern or input-output pair t, $r_{t,j}$ and $y_{t,j}$ the target and the calculated output of output neuron u_j.

The input $v_{t,j}$ of output neuron u_j and pattern t is calculated with:

$$v_{t,j} = y_{t,i} \cdot w_{i,j} \tag{16.171}$$

for all $v \in pred(u)$ with $y_{t,i}$ and $w_{i,j}$ as the output of predecessor neuron i and the weight of the edge (i, j) from neurons $u_i \in succ(u)$ to u_j.

The output of neuron u_j can be expressed as

$$y_{t,j} = f(v_{t,j}) \tag{16.172}$$

with f representing the activation function of the neuron.

Following $\Delta w_{i,j}$ with

$$\Delta w_{i,j} = \sum_{all\ patterns} \eta \frac{\partial}{\partial w_{i,j}} E_t(W) \tag{16.173}$$

is regarded in more detail. As E_t is dependent from $y_{t,j} = f(v_{t,j})$ and by applying the chain rule, it follows:

$$\frac{\partial E_t(W)}{\partial w_{i,j}} = \frac{\partial E_t(W)}{\partial v_{t,j}} \cdot \frac{\partial v_{t,j}}{\partial w_{i,j}}. \tag{16.174}$$

With $v_{t,j} = y_{t,i} \cdot w_{i,j} = in_{t,j} \cdot w_{i,j}$ results for the second term to

$$\frac{\partial v_{t,j}}{\partial w_{i,j}} = \frac{\partial}{\partial w_{i,j}} (y_{t,i} \cdot w_{i,j}) = y_{t,i} = in_{t,j} \tag{16.175}$$

The term $\delta_{t,j} = -\frac{\partial E_t(W)}{\partial v_{t,j}}$ is defined as error signal and can be further expressed by applying the chain rule:

$$\delta_{t,j} = -\frac{\partial E_t(W)}{\partial v_{t,j}} = -\frac{\partial E_t(W)}{\partial y_{t,j}} \cdot \frac{\partial y_{t,j}}{\partial v_{t,j}}. \tag{16.176}$$

With $y_{t,j} = f(v_{t,j})$, the second term $\frac{\partial y_{t,j}}{\partial v_{t,j}}$ can be written as

$$\frac{\partial y_{t,j}}{\partial v_{t,j}} = \frac{\partial}{\partial v_{t,j}} f(v_{t,j}). \tag{16.177}$$

After defining $\delta_{t,j}$ and defining how it can be calculated, the focus is again on

$$\Delta w_{i,j} = \sum_{t=1}^{p} \eta \frac{\partial}{\partial w_{i,j}} E_t(W) = \sum_{t=1}^{p} \eta \frac{\partial E_t(W)}{\partial v_{t,j}} \cdot \frac{\partial v_{t,j}}{\partial w_{i,j}}. \tag{16.178}$$

With

$$\delta_{t,j} = -\frac{\partial E_t(W)}{\partial v_{t,j}} \tag{16.179}$$

and

$$\frac{\partial v_{t,j}}{\partial w_{i,j}} = y_{t,i} \tag{16.180}$$

follows for batch mode:

$$\Delta w_{i,j} = \eta \sum_{t=1}^{p} y_{t,j} \delta_{t,j} \tag{16.181}$$

or for one pattern k:

$$\Delta_k w_{i,j} = \eta y_{k,j} \delta_{k,j} \quad (batch\ mode) \tag{16.182}$$

For the calculation of

$$\delta_{k,j} = -\frac{\partial E_t(W)}{\partial v_{t,j}} = -\frac{\partial E_t(W)}{\partial y_{t,j}} \cdot \frac{\partial y_{t,j}}{\partial v_{t,j}} \tag{16.183}$$

a case distinction must be performed: the regarded neuron can be an output neuron or a non-output neuron.

If u_j is an output neuron, then:

$$-\frac{\partial E_t(W)}{\partial y_{t,j}} = -\frac{\partial}{\partial y_{t,j}} E_t(W) = -\frac{\partial}{\partial y_{t,j}} \frac{1}{2} \sum_{u_j \in U_{out}} (r_{t,j} - y_{t,j})^2 = r_{t,j} - y_{t,j} \tag{16.184}$$

and therefore

$$\delta_{k,j} = -\frac{\partial E_t(W)}{\partial v_{t,j}} = -\frac{\partial E_t(W)}{\partial y_{t,j}} \cdot \frac{\partial y_{t,j}}{\partial v_{t,j}} = (\frac{\partial}{\partial v_{t,j}} f(v_{t,j})) \cdot (r_{t,j} - y_{t,j}) \tag{16.185}$$

and

$$\Delta_k w_{i,j} = \eta y_{k,j} (\frac{\partial}{\partial v_{t,j}} f(v_{t,j})) \cdot (r_{t,j} - y_{t,j}) \tag{16.186}$$

In the second case, let u_j be a neuron in a hidden layer $u_j \in U_{hidden}$ and $u_i \in succ(u_j)$ a neuron in the following layer with connections to u_j. The deviation $-\frac{\partial E_t}{\partial y_{t,j}}$ can be only calculated indirectly by applying chain rule and by totaling over all successor neurons u_i of u_j:

$$-\frac{\partial E_t(W)}{\partial y_{t,j}} = -\sum_{u_i \in succ(u_j)} \frac{\partial E_t(W)}{\partial v_{t,i}} \cdot \frac{\partial v_{t,i}}{\partial y_{t,j}}. \tag{16.187}$$

With

$$v_{t,i} = \frac{\partial}{\partial y_{t,j}} (y_{t,j} \cdot w_{i,j}) \tag{16.188}$$

follows

$$-\frac{\partial E_t(W)}{\partial y_{t,j}} = -\sum_{u_i \in succ(u_j)} \frac{\partial E_t(W)}{\partial v_{t,i}} \cdot \frac{\partial}{\partial y_{t,j}}(y_{t,j} \cdot w_{i,j}) = -\sum_{u_i \in succ(u_j)} \frac{\partial E_t(W)}{\partial v_{t,i}} \cdot w_{i,j}. \quad (16.189)$$

With

$$-\frac{\partial E_t(W)}{\partial v_{t,i}} = \delta_{k,i} \quad (16.190)$$

follows

$$-\frac{\partial E_t(W)}{\partial y_{t,j}} = \sum_{u_i \in succ(u_j)} \delta_{k,i} \cdot w_{i,j} \quad (16.191)$$

resulting in

$$\delta_{k,j} = (\frac{\partial}{\partial v_{t,j}} f(v_{t,j})) \sum_{u_i \in succ(u_j)} \delta_{k,i} \cdot w_{i,j} \quad (16.192)$$

and, therefore,

$$\Delta_k w_{i,j} = \eta y_{k,j} (\frac{\partial}{\partial v_{t,j}} f(v_{t,j})) \sum_{u_i \in succ(u_j)} \delta_{k,i} \cdot w_{i,j}. \quad (16.193)$$

Regarding the formula above, it is obvious that the change of the weights connected with neurons in a hidden unit can be calculated recursively backwards from the output layer to the input layer, and the weights are adjusted in a way that the error of the ANN is minimized.

The described algorithm is the basis of *backpropagation with gradient descent algorithm*. This algorithm has the disadvantage that a found solution can be a local minima. In practice, there are different improvements and modifications of the described algorithm, such as normalization of the gradient, a variable step size or learning rate, a predefined termination condition, e.g., a maximum of iteration steps, and the variation of different starting points [Rud16].

A disadvantage of the *backpropagation with gradient descent algorithm* as proposed in this chapter is that the algorithm converges in local minima [Bri97]. In the literature, there are different modifications of the algorithm listed that permit avoiding being trapped in a local minima or optimum, such as the global descent algorithm, through the implementation of the repeller effect[CBB93], or a method in which the training of the ANN is performed several times with different randomly selected starting points of the weights [IR99, AS07] and local minima are avoided stochastically. The *backpropagation algorithm* is implemented in several open-source software tools, for example, the package *neuralnet* [FGW19] for *R* or the module *tensorflow* in *Python* [PNW20].

16.4.1.3 Interim conclusion: a concurrent neuro-fuzzy to pre-process incomplete or noisy data

In the last section, a feedforward layered ANN called multilayer perceptron together with the base of a learning algorithm called backpropagation with gradient descent was introduced. In the framework of an SDSS, this type of ANN is connected with a FLC with a methodology called a *concurrent neuro-fuzzy system* in which the ANN is used to pre-process noisy or incomplete input values and the preprocessed input values without noise or missing values deal as the input values for an FLC. The FLC then gives decision support by evaluating *IF ... THEN ...* rules. The described concurrent neuro-fuzzy system can be used for SDSS if the set of linguistic *IF ... THEN ...* rules are available for a risk mitigation strategy – with the constraint that the input parameters to give decision support are noisy or incomplete.

The described approach is only one example of how to give SDSS in the described framework, even if the input data is incomplete or noisy, and can be used as a starting point for an SDSS within the research and development cycle within the LL approach. The initial selection of an appropriate algorithm is a part performed during the research and development cycle within the LL by a stakeholder. During the research and development cycle, the performance of the selected SDSS system should be monitored and the selected algorithm can be improved by adding new modifications of the algorithm or by selecting more performative algorithms.

The described *multilayer perceptron* ANN architecture with the described *backpropagation with gradient descent algorithm* can be used in the LL approach as an *autoassociative neural network* to correct data for example sampled in a citizen-science approach (chapter 6) or in surveys with missing data items or with noise in the sampled dataset. There are two modifications of the algorithm, and it can be used in an online or offline mode. In offline mode, the ANN is trained with a training samples before it is used. In online mode, the weights of the neural nets are adjusted after each representation of a dataset during the operation mode. The mode of operating must be selected in an LL approach according to the availability of an appropriate training data set.

The algorithm described in this chapter has the advantage over the algorithms introduced in Section 7.4 that it can be efficient when combined with an FLC to pre-process data efficiently in a *concurrent neuro-fuzzy system.*

16.4.2 A neuro-fuzzy system to determine parameters for a FLC

There are several methods described to find fuzzy rules and fuzzy sets from example data, e.g., *adaptive fuzzy controllers*. Early approaches as demonstrated by [PM79], [DHR13], or [QZHS92] propose methods for self-organizing FLC, which have in common that the initial set of fuzzy rules is changed by a non-ANN-based algorithm in a way that the interpretation of the rule base is not possible anymore.

More modern approaches to determine the parameters of fuzzy sets and of the consequent part of the rules for their use in FLC are based on ANN methods. Adaptivity of the system and system parameters is particularly important because the described system should be able to be used in different regions of the world with different spatial characteristics. *Hybrid neuro-fuzzy systems* with the ability to be adaptive and find parameters for the membership functions are called ANFIS [Jan93]. The ANFIS model is used to determine membership functions in a supervised learning approach, for which the rule base for a *Takagi-Sugeno* type FLC must be available [BKKN03]. ANFIS can work with rules R_r of the type

$$R_r : If\ x_1\ is\ A_{j1} \wedge ... \wedge x_n\ is\ A_{jn}\ Then\ y = \alpha_0^{(r)} + \alpha_1^{(r)}\ x_1 + ... + \alpha_n^{(r)}\ x_n \tag{16.194}$$

Today, ANFIS models have a broad application in many different disciplines, inter alia, to predict crop harvest yield [NAO+12], in vehicle steering and control [Sto18], prediction of photovoltaic power [KBK+21] [KNNK22], or in an early warning system for flood disasters [FMI+21].

In an ANFIS, the components of a FLC are ordered in a net. In the original concept proposed by [Jan93], ANN-based learning algorithms are used to find appropriate membership functions. However, improvements of the ANFIS model in the last years were developed in which the structure of the net can also be changed [KK19]. The connections in an ANFIS are unweighted [Aza10]. Parts of the nodes of an ANFIS are adaptive, meaning that the parameters of node functions, e.g., parameters of membership functions, can change by applying a learning algorithm. For ANFIS models, several open-source software implementations are available, e.g., for *R* the package *FuzzyR* [CGR21] or in *Python* the module *anfis* [Meg15].

At the beginning of the research and development cycle in an LL, a model should be used that can be easily implemented. As the mentioned implementations of ANFIS work only by the adjustment of parameters and not with structural adaptions, this concept is highlighted in the following section. However, the adaption of available ANFIS package with the aim of finding a rule base could be a task during the research and development cycle in an LL.

In the original concept of [Jan93], an ANFIS consists of five layers, as visualized in figure 16.23.

The sample ANFIS visualized in figure 16.23 consists of two rules:

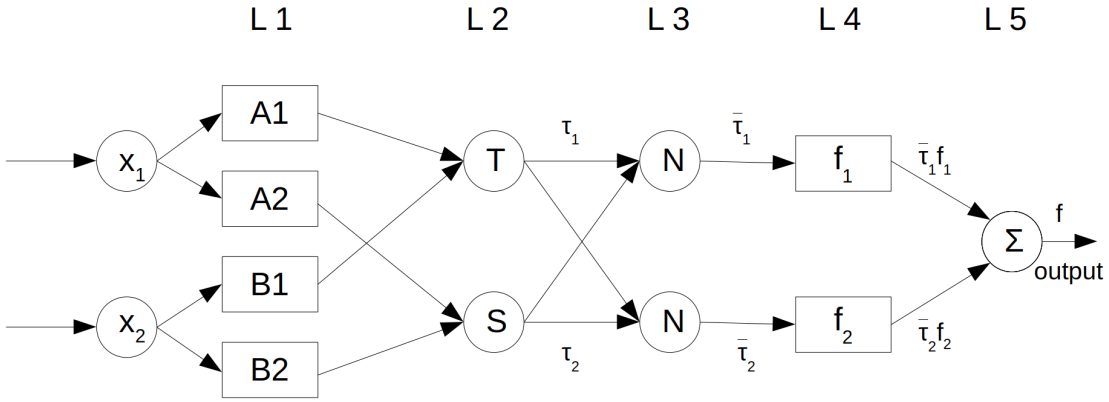

Figure 16.23: Representation of an ANFIS with two inputs, two rules, and five layers $L1$, ... , $L5$

$$R_1 : If\ x_1\ is\ A_1\ AND\ x_2\ is\ B_1\ THEN\ f_1 = p_1x_1 + q_1x_2 + o_1 \tag{16.195}$$

and

$$R_2 : If\ x_1\ is\ A_2\ OR\ x_2\ is\ B_2\ THEN\ f_2 = p_2x_1 + q_2x_2 + o_2 \tag{16.196}$$

with

$x_1, x_2, p_1, q_1, o_1, p_2, q_2, o_2 \in \mathbb{R}$ and A_1, A_2, B_1, B_2 representing membership functions.

The visualized rules belong to a first order fuzzy system, as only polynomials from grade one are in the consequent part of the fuzzy rule. However, for more complex problems higher polynomials can also be used.

Layer 1 L_1 is called fuzzification layer. Each node in this layer represents a differentiable parameterized membership function, e.g., a bell function. $x_1, ..., x_n$ are input values, and $A1, ..., B2$ are linguistic labels. The output of the nodes represent the degree to which the input values belong to the connected labels. In these nodes, learning and the adaption of the system are performed by adjusting the parameters of the membership functions such that the deviation between the output f and the desired or target output $\tilde{f}$ is minimized. For example, regarding node i in Layer L_1, a bell membership function and input x_j, and linguistic label B_i, the output of node i

$$O^1_{B_i} = \mu_{B_i}(x_j) = \frac{1}{1 + |\frac{x_j - c_i}{a_i}|^{2b_i}}, \tag{16.197}$$

whereby a_i, b_i and c_i are the premise parameters, which are adjusted during learning. However, every piecewise differentiable and continuous function with a minimum at 0 and a maximum

at 1 can be taken as a membership function, like the Gaussian normal distribution curve. However, the membership function used in this example has the great advantage that the antiderivative can be calculated explicitly. With other membership functions, there might also be other parameters. The output of these nodes is the grade of membership of the actual input value in relation to the connected linguistic label [Jan93].

Layer 2 L_2 consists of nodes representing the corresponding t-norm or s-norm related to a rule r [BKKN03]. As mentioned in section 16.2.3, t-norms are used for fuzzy *AND* and s-norms for fuzzy *OR*. For some specific learning algorithms, e.g., *backpropagation with gradient descent* (section 16.4.1.2), the used t-norms and s-norms must be piecewise differentiable and continuous. For example, regarding rule R_2 means that the corresponding node in L_2 is connected to an s-norm. A continuous and piecewise differentiable s-norm would be, for example, $\perp_{prod}$ which is defined as $\perp_{prod}(a,b) = a + b - a \cdot b$ (Section 16.2.3). By using this s-norm, the antecedent part of rule R_2 is evaluated as

$$O_2^2 = \tau_2(\mu_{A_2}(x_1), \mu_{B_2}(x_2)) = \perp_{prod}(\mu_{A_2}(x_1), \mu_{B_2}(x_2)) = \mu_{A_2}(x_1) + \mu_{B_2}(x_2) - \mu_{A_2}(x_1) \cdot \mu_{B_2}(x_2)). \tag{16.198}$$

The output of each node in this layer $O_r^2 = \tau_r$ can be interpreted as the grade of fulfillment of rule r [BKKN03] .

Layer 3 L_3 has as many nodes as the system has rules. Each node in this layer calculates the relative grade of fulfillment of this rule $\overline{\tau}_r$ in relation to the sum of all grades of fulfillment in this layer:

$$O_r^3 = \overline{\tau}_r = \frac{\tau_r}{\sum_i \tau_i}. \tag{16.199}$$

The output of these nodes is called the normalized grade of fulfillment. With the normalized grade of fulfillment, it is possible to make conclusions about the overall meaning of a rule in relation to all other rules. An interpretation of a absolute grade of fulfillment is not possible without information about the other grade of fulfillment of the other rules of the regarded ANFIS system [BKKN03].

Layer 4 L_4 consists of nodes, and each node calculates the output, for example, for node r as

$$O_r^4 = \overline{\tau}_r \cdot f_r = \overline{\tau}_r \cdot (o_r + p_r x_1 + q_r x_2) \tag{16.200}$$

with $\overline{\tau}_i$ representing the output of neuron i, and the predecessor in Layer 3 of node j. $p_r, q_r j$ and o_r are called consequent parameters of node r [BKKN03].

In Layer 5, the output $O^5 = f$ is calculated as the sum over the output signals of layer 4:

$$O^5 = f = \sum_{i=1}^{k} \overline{\tau}_i f_i = \sum_{i=1}^{k} \overline{\tau}_i (o_i + p_i x_1 + q_i x_2) \tag{16.201}$$

To find a solution for the premise and consequent parameters, an error function must be defined that should be minimized through the training algorithm. During the learning phase, a training dataset with P entries are presented to the ANFIS. For example, the sum of the quadratic deviation between target output $\tilde{f}$ and computed output f over all propagated training patterns can be regarded as an appropriate error measure that should be minimized:

$$E(a_i, b_i, c_i, p_r, q_r, o_r) = \frac{1}{2} \sum_{p=1}^{P} (\tilde{f_p} - f_p)^2 \text{[BKKN03]} \tag{16.202}$$

The standard learning algorithm in an ANFIS, for example, as proposed by [Jan93], is performed in two phases, a forward and a backward phase:

In the forward phase, the signal is transmitted from Layer 1 to Layer 5, whereby the premise parameters are treated as constants, and only the consequent parameters are adapted. In the backward phase, the premise parameters are adjusted, and the consequent parameters are treated as constants [BKKN03].

[Jan93] proposed an algorithm in which the estimation of the consequent parameters in the forward phase is attributed to a solving problem of a system of linear equations in relation to the consequent parameters p_r, q_r, o_r, with each training pattern related to one linear equation within the system.

Let N be a matrix containing a row in the type

$$(1, \overline{\tau}_1^{(l)}, \overline{\tau}_2^{(l)}, ..., \overline{\tau}_k^{(l)}) \tag{16.203}$$

for each training pattern, whereby $\overline{\tau}_i^{(p)}$ represents the normalized grade of fulfillment for rule i after the $p - th$ training pattern was presented. Let T be a column vector containing the target values f_p for all training patterns and

$$A = (o_1, p_1, q_1, ..., o_k, p_k, q_k)^T \tag{16.204}$$

the column vector containing the consequence parameters for all rules.

Writing the system of linear equations in matrix form, leads to

$$T = N\ A \tag{16.205}$$

As the linear equation system consists of equations derived from real-world data with noise and measuring faults, and as in practice the number of training patterns is in the most cases

higher than the number of unknown consequent parameters, a single solution for the system is not achievable. Therefore, a method to find an approximate solution A^* for A for an overdetermined system of linear equations is needed. This can be formulated as:

$$min\ ||NA - T||^2 \tag{16.206}$$

There are different methods to solve an overdetermined system of linear equations like *least square estimation method* with a Gaussian normal equation and the pseudo inverse[KK19]. A solution for solving this overdetermined system of linear equations with *least-square estimation* is, according to [BKKN03], given by:

$$A^* = (N^T N)^{-1} NT \tag{16.207}$$

A derivation of this solution for an overdetermined system of linear equations can be taken from [HB09].

The determination of the premise parameters of the membership functions is performed with gradient descent and error backpropagation, similar to the described method in section 16.4.1. Nowadays, there are different modifications of the learning algorithm available, improving the performance of the algorithm and the problems related to *backpropagation* like high computing costs and trapping in local minima [RP16]. An overview of solving methods and algorithms for an ANFIS is presented in [RP16]. With the described approach in this chapter, it is possible to find the parameters of membership functions in a supervised learning approach. This approach can be used for example in an LL approach to estimate the grade of fitness of a risk mitigation strategy where the rule base must be available but with unknown membership functions. The system is adaptive; therefore, it can be used for different use cases with different realities with their own (spatial) characteristics. In contrast to the use of an ANN, the use of an ANFIS has the advantage that the determined parameters and related membership functions can be used in further approaches in the research and development cycle without a black-box behavior as when an ANN is used.

After introducing methods to deal with noisy or missing parameters and to find appropriate parameters for membership functions and the consequent parameters of a first order *Takagi-Sugeno* type FLC, the following chapter deals with how to find a unknown rule base in a supervised learning approach.

16.4.3 A neuro-fuzzy system to determine rules for a FLC

In the last section, a method to determine parameters for an FLC in an adaptive approach was described. However, approaches determining the structure of an FLC in terms of a rule base are also needed, e.g., if an expert cannot describe the system, in the case of the described LL approach to describe the fitness of a risk mitigation strategy with *IF ... THEN ...* rules, but when numerical training data is available in the form of input-output pairs, e.g., obtained by a survey. Using the input-output pairs, it would be possible to train an ANN with the aim of getting a trained system for forecasting the fitness of different risk mitigation strategies in relation to personal general (tempo-spatial) parameters. However, ANN have the advantages described in section 16.4 that they behave, for example, like a black box from which no knowledge can be obtained from the trained ANN.

In the LL approach, described in this thesis knowledge about how the fitness of risk mitigation strategies can be estimated is necessary for further applications, e.g., to increase the risk literacy of the LL inhabitants in educational programs. Therefore, a method is needed with which it is possible to obtain a interpretable rule base for the estimation of the fitness of risk mitigation strategies that is usable in an FLC as well outside of an FLC. As the obtained rule base should be interpretable by humans, it should be as easy as possible. In short, a method is needed with which it is possible to generate and extract a interpretable rule base and related fuzzy sets that can be used in an FLC.

In the literature, there are different methods described for generating fuzzy rules and related membership functions from example data. Data driven rule generation is needed because of the large amount of possible rule bases, and it is not possible because of time and performance constraints to evaluate them all [BKKN03].

One method for obtaining fuzzy rules and membership functions from example data is called *fuzzy clustering*. In contrast to *crisp clustering*, where each data entry only belongs to one cluster, in *fuzzy clustering* data items can belong to more than one cluster, and the grade to which a data entry belongs to a specific cluster is expressed in a grade of membership. *Fuzzy clustering* is an unsupervised technique with the aim of finding patterns in the input space of the data [LL16]. In *fuzzy clustering*, the data items are assigned to a predefined number of clusters in a way that a function dependent, inter alia, on the distance of each data entry to the center of the clusters is minimized. The idea behind *fuzzy clustering* in the context of finding fuzzy rules and fuzzy sets is that each cluster represents a fuzzy rule that projects

the input space into the output coordinate space. Related fuzzy sets can be obtained by enveloping the projected data items with a fuzzy set [WM92]. The granularity of the fuzzy sets and rule base can be generated at the same time. However, the rule base obtained and the associated fuzzy sets are difficult or impossible to interpret, as each generated fuzzy rule is assigned to individual fuzzy sets, which is why they cannot be associated with linguistic terms [BKKN03]. However, in the described LL approach, the generated rule way should be extractable and interpretable in a way that it can be used in other contexts of the LL, such as increasing risk literacy.

16.4.3.1 Generating fuzzy rules with the FCM algorithm

Finding fuzzy rules is often based on clustering algorithms. The theory of *fuzzy clustering* and the detailed description of the algorithms can be taken from [LL16].

Definition 16.4.5 (Clustering) *Clustering is a process or method with which different multidimensional data items are grouped into subsets or clusters based on a measure regarding the similarity of the different data items [OES07].*

To generate fuzzy rules from example data there are different clustering algorithms available such as evolving clustering method (ECM) [KA15], FCM, and *K-means* [GD13] or other clustering algorithms. These algorithms group data items into the same cluster if they have a similarity such that all data items in a cluster are homogeneous, while the clusters themselves are heterogeneous [WWG$^+$21].
To find fuzzy rules from multidimensional data, algorithms based on FCM clustering are often used.
Following a basic description of the FCM algorithm as proposed by [IA11]:

Definition 16.4.6 (FCM) *Let $Y = \{y_1, y_2, ..., y_n\}$, $y_1, y_2, ..., y_n \in \mathbb{R}$ be the set of n data items that should be grouped into c fuzzy clusters, with $1 < c < n$, and let $Z = \{z_1, z_2, ..., z_c\}$ be the set of centroids or cluster centers. The result of the fuzzy clustering algorithm can be described with a matrix*

$$\mu = \begin{pmatrix} \mu_{1,1} & \dots & \mu_{1,c} \\ \vdots & \ddots & \vdots \\ \mu_{n,1} & \dots & \mu_{n,c}, \end{pmatrix} \tag{16.208}$$

consisting of c columns and n rows. $\mu_{i,j}$, with $i \in \{1, 2, .., n\}$ and $j \in \{1, 2, .., c\}$ describing the matrix' element in row i and column j with $\mu_{i,j} \in [0, 1]$ and representing the degree to

which the $i-th$ data item belongs to the $j-th$ cluster.

Additionally, μ has the following characteristics:

$$\sum_{j=1}^{c} \mu_{i,j} = 1, \forall i \in \{1, 2, ..., n\} \tag{16.209}$$

and

$$0 \leq \sum_{i=1}^{n} \mu_{i,j} \leq n, \forall j \in \{1, 2, ..., c\}. \tag{16.210}$$

Fuzzy clustering with FCM is performed by minimizing the following energy function

$$J_m = \sum_{j=1}^{c} \sum_{i=1}^{n} \mu_{i,j}^m d_{i,j} \tag{16.211}$$

with

$$d_{i,j} = ||y_i - z_j||, \tag{16.212}$$

whereby m is a value with $m > 1$ and $d_{i,j}$ the Euclidean distance between the data item y_i and the centroids of cluster j.

The centroid z_j of cluster j is calculated by

$$z_j = \frac{\sum\limits_{i=1}^{n} \mu_{i,j}^m y_i}{\sum\limits_{i=1}^{n} \mu_{i,j}^m}, \tag{16.213}$$

the degree $\mu_{i,j}$ to which data item i belongs to cluster j by

$$\mu_{i,j} = \frac{1}{\sum\limits_{k=1}^{c} \left(\frac{d_{i,j}}{d_{i,k}}\right)^{\frac{2}{m-1}}} \tag{16.214}$$

A solution of a FCM is found by minimizing J_m.

[IA11]

In the definition above, m is called the weighting exponent or fuzzifier and describes how crisp the data items are assigned to a cluster. Side condition $\sum\limits_{j=1}^{c} \mu_{i,j} = 1$ describes the fact that the sum of the grade of memberships to the c different clusters is summed up to 1. The second side condition $0 < \sum\limits_{i=1}^{n} \mu_{i,j} < n$ assures that the clusters are not empty. The distance function is not necessarily Euclidean distance.

[Bez73] proposed the following algorithm to perform FCM:

Definition 16.4.7 (FCM algorithm) *1) Determine matrix U for a given preselected m.*

2) Determine cluster centroids z_j with $\forall j \in \{1, 2, ..., c\}$.

3) Determine distances $d_{i,j}$.

4) Determine grades of memberships $\mu_{i,j}$.

5) Check selected convergence criteria, if not fulfilled then go to step (2).

Convergence criteria might be, for example, if a change in the cluster center is smaller than a predefined value ϵ.
[Bez73]

An implementation of the FCM algorithm and other soft clustering algorithms can be found, for example, in the *R* package *ppclust* [Ceb19] or the *fuzzy-c-means* module in *Python* [Dia19] and can therefore be used for free or in a low-cost LL.
The described FCM algorithm is used in several applications for *fuzzy clustering* in classification problems. However, up to today, there are also other fuzzy clustering algorithms like *fuzzy particle swarm optimization* for *fuzzy clustering* [LZKX12] or modified FCM algorithms [LYWL08] or combination of both [SFPSO15].
To date, there are different algorithms using fuzzy clustering techniques for finding fuzzy rules. [BMR20], and [ASA14] use a technique called *subtractive fuzzy clustering* to find fuzzy rules from numerical data. In *subtractive fuzzy clustering*, each data item is related to a value called potential P in which a high value of P means that the data item has a high density of neighboring data points and therefore has a high potential to be a cluster center. After a cluster center is found, potential P are recalculated for all data items, and the next cluster center is found. A detailed description of the algorithm can be found in [Chi94].
The described fuzzy clustering algorithms differ only in how the clusters are found. [CWGK07] describe an algorithm how to obtain fuzzy rules from input data based on the described FCM algorithm. Each cluster in the output space represents a linguistic term, and the grade to which a point belongs to that cluster represents the grade the point belongs to the linguistic term. For example, if a point belongs to cluster C_j, then the conclusion of the rule results in *then y is* C_j. Further, FCM is used to cluster all input variables from which membership functions are then generated and to which their linguistic meaning is assigned, allowing the antecedent part of the *IF ... THEN ...* rules to be derived from the found clusters and related membership functions and their linguistic meaning.
The described fuzzy clustering algorithms have a large disadvantage because for every rule individual fuzzy sets are used, even if they have the same meaning. Therefore, the interpretability is relatively low [BKKN03].

A fuzzy system is needed that fulfills a regression task and, therefore, a system finding fuzzy rules in the type of *Mamdani*.

16.4.3.2 Generating fuzzy rules with the Wang and Mendel algorithm

[WM92] developed an algorithm for finding fuzzy rules from example data, whereby the found fuzzy rules and related fuzzy sets are of *Mamdani* type, which are interpretable [BKKN03]. With the help of this algorithm, it is possible to create a rule base derived from example data with input-output pairs (x_i, y_i) [WM92] and to use this rule base for a FLC. [WM92] prove that with the described method, it is possible to generate an FLC that can approximate any nonlinear continuous function on a compact set with a predefined accuracy. The FLC with the developed rule base is then used to calculate the grade of fitness for different risk mitigation strategies in an LL approach.
It is assumed that in the LL approach, data is sampled in which attributes are supposed to be related with an increasing risk for CKD or the fitness of a risk mitigation strategy. This data is stored in a vector x_j, which is used as the input values for the SDSS.
x_j describes the input values for the jth data item in the training set S. The training set S consists of all data items consisting of input-output pairs; therefore, $S = \{(x_1, y_1), (x_2, y_2), ..., (x_p, y_p)\}$.
y_j represents the grade of fitness of a given risk mitigation strategy, rated by a expert in the LL or by the LL inhabitant, to which the input data belong. The aim of this task in the LL approach is to find extractable and interpretable fuzzy rules from numerical data expressed as input-output pairs (x_j, y_j), which can be used in risk literacy programs and as a rule base in an FLC working as an SDSS also for people living outside an LL without experts who can rate the fitness of different risk mitigation strategies related to input parameters x.

Definition 16.4.8 (Wang and Mendel algorithm) *The Wang and Mendel algorithm is structured into five steps. The following description is based on [WM92]:*

- *Step 1: All input and output variables are partitioned into equally distributed overlapping fuzzy membership functions. Each membership function is linked to a linguistic label, such as small, medium or large. The number of used membership function can differ for every variable; however, it should be a uneven number.*
- *Step 2: Fuzzy rules are generated. This step is performed for every input-output pair*

(x_i, y_i) and for every variable in the input output space. The grade of membership is calculated for the given crisp value of the regarded input variables x_i for the different partitions and membership functions generated in Step 1. The label of the partition with the highest degree of membership is used to generate a fuzzy rule. For example, if variable x_1 has the highest degree of membership for the membership function with the label medium, then an antecedent part of the generated rule might be

$$IF\ x_1\ is\ medium\ THEN\ .... \tag{16.215}$$

This step is performed for all input and output variables and for all input-output pairs (x, y) – meaning that for every input-output pair, a rule is generated. The result is an if then rule whereby the different input variables in the antecedent part are connected with a t-norm.

- *Step 3: A weight factor for each generated rule is calculated by multiplying the grade of memberships of the regarded parameters and related values of the used membership functions . For example, let R_k be a fuzzy rule obtained by input-output pair (x_k, y_k) with*

$$R_k:\ IF\ x^1\ is\ A_1\ and\ x^2\ is\ A_2\ and\ ....\ THEN\ y\ is\ B_1. \tag{16.216}$$

The weight factor G of rule R_k is calculated by

$$G(R_k) = \mu_{A_1}(x_k^1) \cdot \mu_{A_2}(x_k^2) \cdot ... \cdot \mu_{B_1}(y_k). \tag{16.217}$$

The weight factor G is used if conflicting rules occur, for example, with the same antecedent but different consequent parts.

- *Step 4: A combined rule base is created. For the combined rule base, rules and their weights obtained in Steps 2 and 3 together with existing rules from human experts are combined to a unique rule base. The rules from human experts also must be weighted by the expert. In this step, conflicting rules are regarded, for example, with the same antecedent but with different consequent parts. The conflicting rule with the higher weight G is used for the final rule base.*

- *Step 5: The final defuzzification is performed from a fuzzy into a crisp value y, for example, with the center of gravity method, and is described in detail in [WM92].*

With the algorithm described in [WM92], it is possible to extract *IF ... THEN ...* rules usable in an FLC and also understandable by humans and implement them in the *R* package

frbs [RBHB15]. However, the described algorithm has the disadvantage that the membership functions are fixed. That means that they are set at the beginning of the algorithm in Step 1 and are not further adjusted in terms of minimizing the error function. The application of this algorithm therefore lies in the area of whether the membership functions and the associated parameters can be determined by the knowledge of an expert or whether they are known. However, in an LL approach in a less-developed country, it is conceivable that there is no expert knowledge about the membership functions used. This task is handled in the following section.

16.4.4 An algorithm for generating fuzzy rules and membership function optimization

The described LL approach should be used in cases of substance-related diseases with unknown etiology often occurring in less-developed countries. In this approach, expert knowledge, e.g., for determining rules and membership functions, might not be available. Therefore, a system is needed that generates an interpretable fuzzy rule base with fitting related membership functions.

In the literature, there are different methods described that fulfill at least partly the mentioned tasks. An overview of some of the common *neuro-fuzzy systems* for rule generating is provided in [VK15].

Dynamic evolving neural-fuzzy inference system (DENFIS) is a cluster-based approach to generate *Takagi- and Sugeno*-based fuzzy rules in online and offline mode [KS02]. The DENFIS approach is implemented in a *R* package [RBHB15]. However, the interpretability of the generated rule base is low because of the nature of *Takagi- and Sugeno*-based FLC.

Fuzzy Adaptive Learning Control Network (FALCON) is hybrid *neuro-fuzzy system* and consists of a *multilayer network* with nodes representing elements of a *Mamdani* type FLC. The initial structure of the network is needed. This structure can change during the learning phase, for example, nodes can be added or deleted. Learning is performed in two phases: During the first phase, an unsupervised algorithm using a self-organizing method is applied to adapt the internal structure of the system. Adaptation of the parameters of membership functions is performed during the second phase in a supervised-learning approach [LL+91].

Another hybrid *neuro-fuzzy algorithm* to determine rule base and membership function is called *NEuro Fuzzy function apPROXimation* (NEFPROX) [NK98]. It has a direct standalone software implementation for *Linux* usable in an LL approach. NEFPROX can generate rules

for *Mamdani* and *Takagi and Sugeno* systems. Easily interpretable rule structure is obtained in which the used algorithm assures that the rule base consists only of a limited number of rules and of interpretable fuzzy sets with shared weights. Through the use of shared weights of fuzzy sets, fuzzy sets with the same linguistic meaning but in different fuzzy rules have the same membership function, and the interpretability remains high [AHM03]. However, besides the standalone implementation, there is no implementation for an open-source programming language like *R* or *Python*.

Self-cOnstructing Neural Fuzzy Inference Network (SONFIN) is also a hybrid approach based on ANFIS with the ability to change the network structure. Learning without a predefined network structure is possible; however, if a apriori knowledge is available, it cannot be used in the initial network structure [JL98]. An improved SONFIN method with higher performance is described in [PLL$^+$15].

Influential Rule Search Scheme (IRSS) is a *neuro-fuzzy system* to find membership functions and associated fuzzy rules and acts as a nonlinear function approximator. Membership functions and their partitions for each input and output variable are generated by an FCM clustering algorithm in which the number of membership functions or clusters must be predefined. The type of membership function is not predefined but derived from the training data used. For the rule generation, an algorithm called *initial fuzzy rule base construction* is applied. This algorithm first defines all rules possible with the related partitions and membership functions and calculates for all possible rules a strength of rule value for all input-output pairs. Rules with a high strength of rules values are held in the rule base. A detailed description of the algorithm can be found in [CRS06]. However, there is no direct implementation of the algorithm for open-source tools like *R* or *Python*. Other mentioned algorithms for generating a fuzzy rule base are, for example, *mEFuNNs* [KS99] or *CURE* [KAA$^+$02].

A system mitigating the disadvantages of the described systems with the ability to generate an interpretable fuzzy rule base of *Mamdani* type, to tune membership functions, and with a direct implementation in *R* is called *hybrid neural fuzzy inference system* (HyFIS) [KK99]. The system is able to find fuzzy rules of *Mamdani* type from numerical learning data combined with a *neuro-fuzzy system* that optimizes the parameters of membership functions with *error backpropagation*. The described system is able to do both tasks necessary: finding interpretable fuzzy rules and tuning related fuzzy sets or membership functions. It is also implemented in an open-source *R* package called *frbs* [RBHB15] and can therefore also be used in a low-cost LL approach as proposed in this thesis.

In the approach described by [KK99], learning is performed in two phases: Fuzzy rules are generated in the first phase with the *Wang and Mendel algorithm* [WM92] (section 16.4.3.2). In the second phase, the corresponding membership functions are tuned by the adjustment of the parameters of the used membership functions by applying a *neuro-fuzzy model* based on a *multilayer perceptron* with *error backpropagation* (section 16.4.1.2). A similar technique to that used in the second phase was described in the previous chapter called ANFIS; however, the *multilayer perceptron* used for the HyFIS has a modified structure when compared with ANFIS. If new training data is available, the generated fuzzy rules can be easily found and integrated into the existing rule base [KK99]. This is a requirement necessary in the described LL approach, as the learning phase of the system should continue during the application of the system as an SDSS. The system working in the first phase is further called *rule finding module*, and the system in the second phase *membership function tuning module*. A schematic representation of a HyFIS system is visualized in figure 16.24.

The membership function tuning module consists of a five layered *multilayer perceptron* with

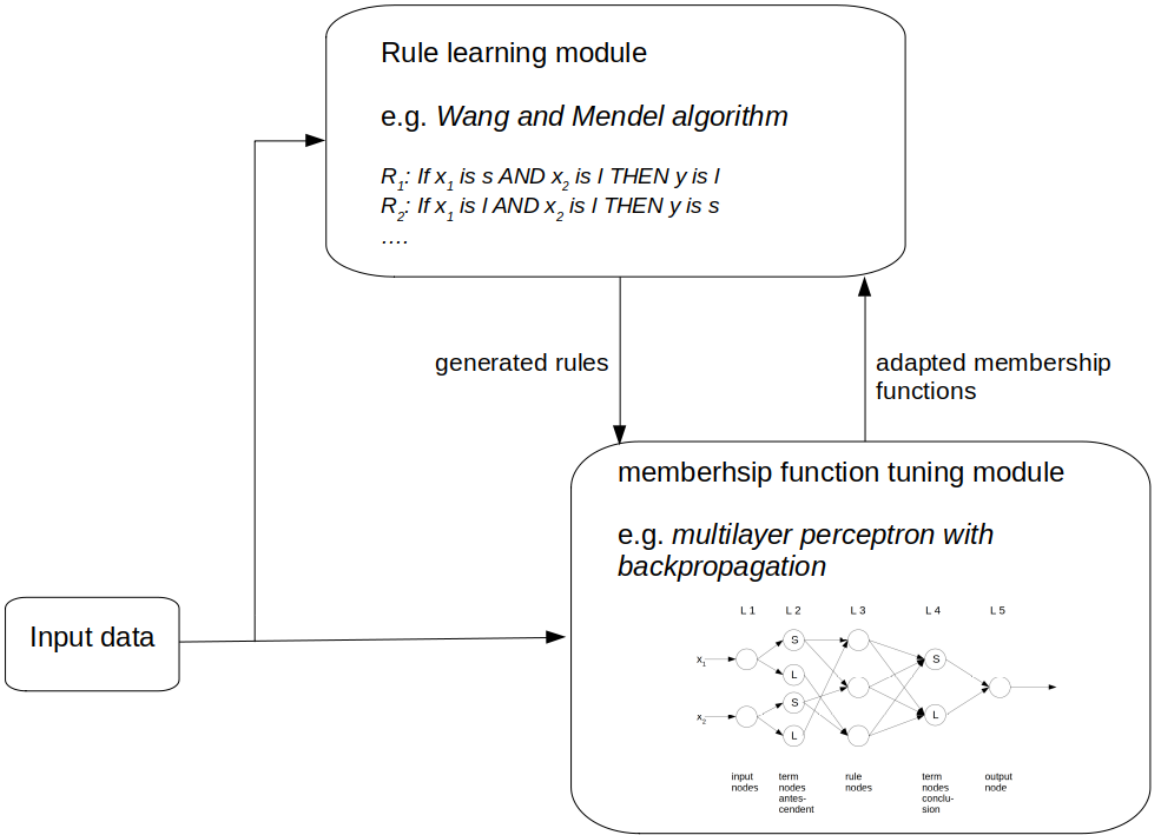

Figure 16.24: Schematic representation of a HyFIS system according to [KK99] (figure generated with *LibreOffice Draw*).

a supervised learning algorithm using *error backpropagation* with the following structure as proposed by [KK99]:

- Layer 1 consists of input nodes transmitting crisp input values to the second layer. Each node of Layer 2 is connected only with the nodes in Layer 2, which are used in the rules.

- Nodes in Layer 2 are called term nodes and consist of nodes representing membership functions and calculating the degrees of membership for the different input linguistic labels present in the antecedent of the used fuzzy rules. The parameters of the used membership functions are not fixed and can be changed during the learning phase. [WM92] and [KK99] originally used overlapping bell-shaped membership functions describing *Gaussian distribution* of the type

$$y_i^{(2)}(x_h) = \mu_i(x_h) = e^{-\frac{(x_h - m)^2}{s^2}} \tag{16.218}$$

  whereby $y_i^{(2)}$ represents the output of node i in layer 2 and m and s are the parameters adjusted during the learning phase. However, every function with free parameters fulfilling the characteristic of a membership function can be used. The outputs of the nodes in Layer 2 are grades of membership with respect to the input value x_h and the regarded linguistic term h. Connection weights to this layer are set to 1.

- Nodes in Layer 3 represent the antecedent part of a *IF ... THEN ...* rule, whereby the different statements of the antecedent part are connected with a fuzzy *AND* or t-norm. [KK99] suggest the *min* operator as t-norm $\top_{min}$; however, every other t-norm can be used as the operator within these nodes. Connection weights to nodes within this layer are set to 1. By using $\top_{min}$ as a t-norm, the output $y_j^{(3)}$ of a node j in layer 3 can be calculated as

$$y_j^{(3)} = \underset{i \in K_j}{min}(y_i^{(2)}). \tag{16.219}$$

  K_j is a set consisting of the indices of nodes of the previous layer, which have a connection to unit j.

- Layer 4 and related nodes represent the conclusion or *THEN* part of a rule. The operators used in nodes of layer 4 are t-conorms $\perp$ and the fuzzy *OR* operator. A node in this layer is related to a linguistic label of the output space like *small*, *medium*

or *large*. The t-conorm is used to integrate rules with the same conclusion part but different antecedent part. The nodes of Layer 4 are connected with all nodes in the previous layer. Connections between a node k in Layer 4 and a node j of the previous layer are weighted with weights $w_{k,j}$. The weights $w_{k,j}$ are also called certainty factor and can be interpreted as the root of a degree to which a certain rule influences the output after different rules were inferred. During the parameter learning phase, the weights $w_{k,j}$ are adjusted. At the beginning of the process, weights $w_{k,j}$ are randomly set to $w_{k,j} \in [-1, 1]$. The output of node k in layer 4 is consequently calculated by

$$y_k^{(4)} = \underset{j \in K_k}{max}(y_j^{(3)} w_{k,j}^2). \tag{16.220}$$

Again, K_k is a set consisting of the indices of nodes of the previous layer, which have a connection to unit k, and $w_{k,j}^2$ is the degree to which the corresponding rule is activated to the overall output.

- Nodes in Layer 5 fulfill defuzzification into a crisp value. Weights are set to 1 in the connections to these nodes. A defuzzification method described in section 16.2.4 can be used to generate a crisp output.

The architecture of the described *membership function tuning module* is visualized in figure 16.25.

The described HyFIS system has the advantage that it can be used in cases where both rules and parameters for membership functions are unknown and can be used in pairs of input and output data to generate a fuzzy rule base. However, interpretability gets lost to a certain degree since due to membership function adjustment, there are different membership functions for the same meaning and the rule base quickly becomes very large [KK99]. Nevertheless, extraction of the rule base is possible.

There are several improved algorithms based on the HyFIS method available, for example, [TQG09] and [TQG12] improved this method for noise-corrupted data. However, at the moment there is no direct implementation of the improved algorithms usable in an open-source LL approach. Implementing such improved methods would be a part within the research and development cycle in an LL.

Summarizing, the HyFIS system is a *neuro-fuzzy system* that provides a model with which it is possible to generate a rule base with adapted membership functions. Learning is performed in a supervised online mode in which rule generation and membership function parameter

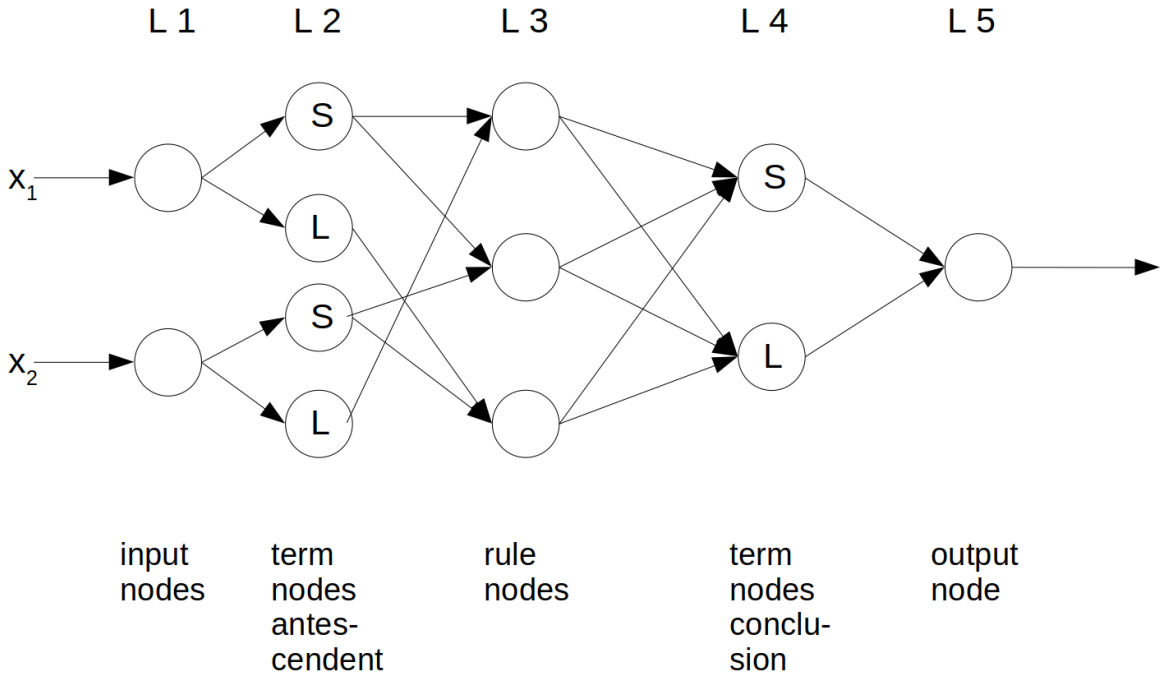

Figure 16.25: Representation of the *membership function tuning module* of a HyFIS system with two inputs, three rules, and five layers $L1$, ... , $L5$ [KK99] (figure generated with *LibreOffice Draw*).

adaption is performed during the operating phase. The generated rule base can be extracted but has a limited interpretability.

16.4.5 Interim conclusion

In the last sections, different methods were introduced with which it is possible to combine a FLC with ANN methods. With such *neuro-fuzzy systems*, it is possible to pre-process incomplete or noisy input data (section 16.4.1), find parameters for membership functions of fuzzy sets (section 16.4.2), or generate fuzzy rules from example data (section 16.4.3) or both (section 16.4.4).

With the described characteristics of CKDu with the unknown etiology of the disease and the described approach of data sampling in a citizen-science approach in developing countries, the described tools can help to achieve an FLC usable in an SDSS operable, even if the input data is not such that it can be directly used in an FLC or if the rule base is not available for an FLC.

16.5 Chapter conclusion: fuzzy logic and ANN

In the last chapters, mathematical methods were introduced, that can be used in an SDSS that is able to work with implicational fuzzy rules and fuzzy data (**requirement R4.7**, section 15.2). In the case of an LL for agrochemical-related risk mitigation strategies, the fitness of risk mitigation strategies is described with *IF ... THEN ...* rules, which are evaluated together with personal, logistic, and environmental parameters from an FLC, which is combined with an ANN to a *neuro-fuzzy system*. In an LL, for each risk mitigation strategy an own system can be set up, used, and trained before and during the operating phase of an LL.

Setting up the parameters of an FLC, for example, parameters of the membership functions or appropriate t-norms, is a hard task that must be performed by experts but can also be determined with algorithms as described in section 16.4.2. There are also algorithms available that can generate an extractable fuzzy rule base. The algorithm described in section 16.4.3 produces such an extractable rule base (**R4.8**); however, it has relatively low interpretability. Possibly, however, a rule base with a low interpretability can help to find patterns in mitigating the risk of CKDu and can be used for further risk mitigation approaches, such as increasing risk literacy in the population.

The described algorithms are based on the combination of FLC and ANN methods into a *neuro-fuzzy system*. The combination of FLC and ANN with inherent learning ability (**R4.2**) makes the resulting *neuro-fuzzy system* adaptive. Adaptivity helps to make the use of the SDSS possible in different regions of the world, where people may make different decisions, for example, due to differing social or cultural characteristics, and to adjust the decision making according to the behavior and feedback of the user.

The proposed algorithms are all available as open-source packages or modules and can be modified and used for free (**R4.6**). The method described in section 16.4.1 allows the pre-processing of noisy or incomplete data such that it can be used in the proposed SDSS (**R4.3**). The selected algorithms allow both online and offline modes. In online mode, the system is learning during the operating phase of the SDSS. In offline mode, the neuro-fuzzy system used as an SDSS can be trained with example data before its use in an LL approach.

Overall, with the described methods it is possible to create decision support with the aim of proposing the best fitting risk mitigation strategy in an SDSS approach with the characteristics of CKD, less-developed countries, and the selected LL approach.

17 Data delivering and needed ICT infrastructure

The following chapter deals with possible hardware and software solutions in the described LL framework.

17.1 Media to deliver spatial support

To use the described SDSS system, it is necessary to deliver information about the supported decisions to the user and get information about personal characteristics, the location of a person, and logistical or crowdsourced data to the SDSS. Different IT architectures for such an information delivering system are realizable.

As described in section 12.2, the use of mobile devices gets increasingly widespread in less-developed countries, although mobile internet connectivity lags behind the use of mobile devices.

Server-client concept as, for example, described in [Sin92], is further used as a method for data sampling, delivery, and processing. In this sense, one or more servers are used for data storing, processing, generating, and delivering, and a client which can be regarded as a information receiver GUI or in some cases, as a tool for data sampling and communication. The main parts of the software to generate decision support and related software tools are running on the server, and it takes the main computational load. Clients can be, for example, digital devices like mobile phones, laptops, or desktop computers. However, it is also possible to give spatial decision support via paper-based information or information delivered via radio or television (TV). The use of a radio or TV has the disadvantage that communication is only possible into one direction – to the users – but without a feedback from the user. Additionally, it is hardly or not possible to collect observations via the described citizen-science approach (chapter 6) for which a connection from the user to the SDSS is also needed.

The use of mobile devices as a medium for delivering personal adapted risk mitigation strate-

gies and to collect data has large advantages in contrast to the use of other mediums, such as radio, TV, or paper-based decision support. It is possible to give personally tailored advice, which is difficult with mass media like TV or radio; data can be sampled, processed, and delivered much faster than in a paper-based framework, where data must be delivered manually, e.g., information about observations must be transported, digitalized, processed, and information about decision support must be transported to the user. The use of a mobile electronic device also has the advantage that communication between the SDSS and the user is more intensive and faster. With faster information delivery by and to the user, the temporal and spatial quality of the delivered information increases.
However, the last-mile problem described in section 6.4 must be kept in mind. Despite the increasing number of mobile phone users and people with mobile devices, there are still regions without mobile internet connectivity.

17.2 Online and offline mode

A SDSS for a mobile device can be used in both offline and online mode. Offline mode means that no permanent internet connectivity and no connection via SMS is necessary or is only necessary for a short time, e.g., to update information or to upload or download data. Updates and downloads are only performed when a connection is available, e.g., WiFi at a public location or at home. Offline mode is used because mobile devices and sending SMS or a permanent mobile internet connection is too expensive or because of infrastructural limitations, such as a lack of internet connectivity. In offline mode, the software for generating SDSS often runs directly on the digital device itself, and large data amount must be stored on the device, e.g., maps or databases.
Online mode means that there is a permanent connection between the user's device and a central processing unit, e.g., in a data center. The information for decision support is generated on the central processing unit in the data center, e.g., a server and delivered to the user with a fast velocity. The client or user's device is used for information sampling, receiving and visualization. Online mode has the advantage that more actual information can be used and therefore the temporal quality of information is higher (section 7.3). In a permanent online mode, nearly real-time information can be used.
Offline mode has the disadvantages that a device with specified requirements in terms of CPU,

RAM, and storage is necessary to run the SDSS software tools and related applications, and the system requirements are higher for devices used in offline mode. Additionally, since information is only updated at irregular time points, e.g., when the user is in an area with WiFi access, the temporal quality of information is lower than in the online approach in which data is continuously updated.

Online mode can be also used with a device with low computing power, e.g., through a web application for which only a browser is needed on the user's device and everything related to SDSS and risk mitigation is computed on a central server. However, the personal data of the user might be processed or stored on a central server. Additionally, the hardware requirements for a server that processes and generates the spatial support queries for all connected client devices are higher in online mode. Additionally, an internet connection is necessary to use it. For the online mode, a permanent internet connection is needed. Data and maps are stored on a server, and SDSS related and additional software are installed and run on the server.

In figure 17.1, a sample information flow in a server-client approach in the LL is visualized. The different stakeholders are connected to the data center and receive information they need for their work in the field of risk mitigation strategies and deliver information to the data center that can be used by other stakeholders in the LL approach. The SDSS system operates on the hardware in the data center and delivers spatial support to the community members of the LL.

17.3 Hardware

For the described IT tasks in the LL, it is necessary to operate an own data center in an LL structure or rent the necessary IT infrastructure to run the necessary software solutions and deliver spatial support or information about risk mitigation strategies. In the following, the described server and client concept for data processing, generating, and delivery is used and described.

The hardware must be rented or purchased according to:

- selected information delivering mode,
- used media for information delivery,
- requirements for the selected software,
- available expertise of stakeholders,

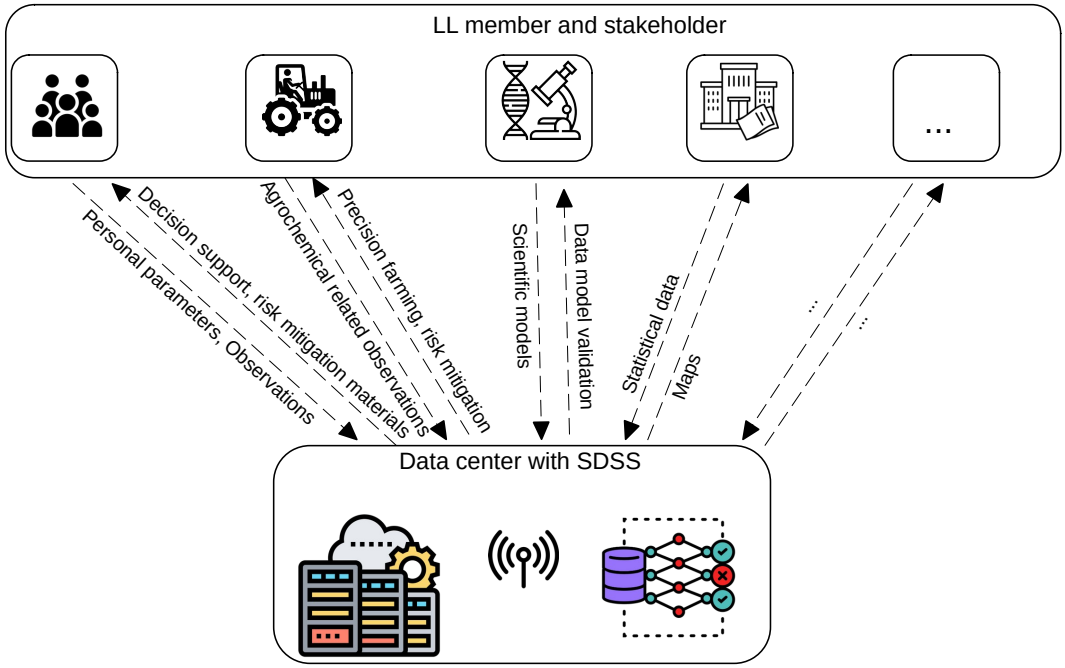

Figure 17.1: Example for the information flow between different stakeholders and a centralized data center in the LL approach (figure generated with *LibreOffice Draw*, image source `https://www.flaticon.com`, creators: *Freepik, Gajah Mada, Eucalyp, Karyative, fajarestuu*).

- available financial and infrastructural resources, and
- compatibility with the used open-source software.

In the LL approach, hardware must be selected according to available financial resources and requirements of the selected software tools. However, the used mode – online or offline – also has an influence on the requirements for the selected hardware. In offline mode, the different software tools of the SDSS run on the client itself; therefore, less computing power is needed on the centralized servers but more on the devices.

The available expertise of the stakeholders also determines which hardware equipment can be used for the SDSS. For example, if the group of stakeholders has only limited expertise in IT related topics, the operation of an own data center in the LL will be an obstacle.

To generate spatial decision support, personalized and health-related data must be stored and processed in an own operated data center or on rented IT equipment. For personalized and health-related data, there are special minimum requirements for data storage and handling

[IMKA20]. The decision of whether an own data center can be operated in the LL or if the IT infrastructure must be rented is also dependent on the knowledge of the involved stakeholders, for example, if data security specialists are involved to ensure the necessary IT security requirements on the operated IT systems.
Media used for information delivery also influences the requirements for hardware. Delivering spatial support via TV has other requirements for hardware than delivering information via the internet does.
Additionally, the compatibility of the hardware with the used open-source software must be regarded during the hardware selection process. Infrastructural conditions, e.g., available power supply, must also be emphasized.

17.4 Possible software solutions for the described LL approach

The described LL approach with the aim of finding and improving risk mitigation strategies for a disease with unknown etiology is located in less-developed countries with poor economic resources and the characteristics of a less-developed country, as described in section 2.1.2. Low economic resources in general may come along with low economic resources in the LL and its stakeholders and community members, as well as with IT components with a low performance. Stakeholders involved in the LL approach have a multidisciplinary background and are not in all cases IT specialists, therefore IT-related processes must be as easy as possible, e.g., the use and installation of software components. The developed software tools should be such that they can be used in a structural similar problem and can be adapted and developed further; therefore, they should be subject to an open license. The used and developed software tools should also run on devices with low performance, as such are often used in less-developed countries. The developed and used software tools should have a easy usability, as stakeholders and community members might not be IT experts and might have a low education in using IT components and related IT.
Summarizing, the software tools used in the LL approach should meet the following requirements:

- **(R4.1):** free of charge or low-cost,
- **(R4.2):** must work on IT components with partly low performance,

- **(R4.3):** easy usability and installation process, and
- **(R4.4):** products should be adaptable for development and use (open license).

These requirements fit to the concept of open-source software, as described in section 5.3.2.1. Open source software is free of charge **(R4.1)**, developed software solutions and products are under an open license and can therefore be modified and adapted **(R4.4)**, and they need IT resources with a lower performance than commercial software. For example, there are open-source operating systems available whose operating system distributions only need IT components with a relatively low performance, e.g., the operating system *Lubuntu* [hom20] **(R4.2)**.

In the past, open-source software solutions were mostly developed by specialists for their own use. Therefore, open-source solutions often had a low usability for other users. As increasingly more open-source software is developed for other users, the aspect of user-friendliness plays an ever greater role in the development of open-source software but often lags behind proprietary software [Ant16] **(R4.3)**.

Despite the lower usability, using a open-source operating system like *Ubuntu* [Can24] or other *Linux* distributions that allow to create customized live devices for installation with predefined software packages that must be installed and with installation scripts for automated installation is advisable. This has the advantage that all stakeholders involved in the LL using the live device for installation have a common software pool, helping to increase data integrity for the exchange between the different stakeholders as a result of a common data format. The implemented install scripts make the installation process as easy as possible. There are different live device creator tools available, e.g., for *Linux* the tool *LiveCDCustomization* (`https://help.ubuntu.com/community/LiveCDCustomization`).

The use of software under the open-source software concept allows also its modification and adaptation, e.g., for other projects with a structural similarity. Appendix C lists possible open-source software solutions for every task in an LL approach: components of a data center to run servers and core software like virtualization software, firewall, user management software, and e-mail back- and front-end, and open-source tools for office work of the stakeholders and LL administration (chapter 5), software tools for data delivery and collection (chapter 6), as well as for data processing and manipulation (chapter 7), and software tools for mathematical modeling (chapter 3) and generating spatial decision support (chapter 15), and risk mitigation materials (part III).

This pool of open-source software tools can be regarded as an initial base of software solutions in an LL approach that can be modified during the research and development cycle and the operating phase of the LL.

17.5 Chapter conclusion

In the LL approach, the mode of operation in terms of online and offline mode must be selected according to the available resources (data center and client devices) and available mobile internet access. For example, in areas with low mobile internet connectivity, an offline mode would make more sense because the SDSS components can run at least partly on the user's devices and internet access is only occasionally necessary. In addition, the use of paper-, TV-, or radio-based SDSS should be considered in regions with low infrastructural conditions.

The selection of the used hardware and software influences each other: used software has specific software requirements to the available hardware. However, there are often several software tools with different software requirements. The selection of software for the LL approach should be made according to the performance of the available hardware equipment, and the selection of hardware should be made according to available financial resources and selected software tools.

The proposed software tools are completely open source. They can be used for free and adapted to the special characteristics of the LL project. As personal and health-related data is processed and stored on the LL-related IT components, dedicated security measures must be considered in the area of IT security and data protection.

18 Development of an SDSS in the described LL framework

Requirements for an SDSS in the described framework were defined in section 15.2. In chapters 6, 7 and 16, methods were introduced with which it is possible to create an initial SDSS with the mentioned requirements as well as methods to sample data in a crowdsourcing approach in an LL. The present chapter summarizes results from the previous chapters together and proposes a model SDSS usable in an LL approach with a agrochemical related disease of unknown etiology.

18.1 Mathematical description of the adaptive mapping machine in the SDSS

In short, a system is needed that performs a mapping from a n-dimensional vector to a m-dimensional vector, where n parameters determine the grade of fitness of m risk mitigation strategies. In this vector, the grade of fitness for each risk mitigation strategy is stored as a value in an interval $[0, 1]$.

The values of n input parameters $x_1, x_2, \cdots, x_n$ that are stored in the vector $x = (x_1, x_2, \cdots, x_n)^T$ with $n \in \mathbb{N}$ can be in general of different types, e.g., expressed as a real value $\mathbb{R}$ or as a grade of membership in the interval $[0, 1]$. The domain K of the mapping F can be expressed as a subset of $\mathbb{R}^n$, $K \subset \mathbb{R}^n$. The m values of the output vector $y = (y_1, \cdots, y_m)^T$ with $m \in \mathbb{N}$ represent the grade of fitness of m possible risk mitigation strategies, expressed as a grade of membership. Therefore, in general, the SDSS should perform the following mapping F:

$$F: \ K \rightarrow [0, 1]^m \tag{18.1}$$

$$(x_1, \cdots, x_n)^T \mapsto F((x_1, \cdots, x_n)^T) = (y_1, \cdots, y_m)^T \tag{18.2}$$

The output vector y and the grade of fitness for the different risk mitigation strategies is proposed by the SDSS. However, the proposed values are often not the grades of fitness

as in reality. The realistic values can be, e.g., measured by the user's feedback, the values determined by the user's feedback are stored in a vector $\tilde{y} = (\tilde{y_1}, \cdots, \tilde{y_m})^T$.

The SDSS should be a system that learns from the user's feedback by taking the deviation between y and $\tilde{y}$ into account. That means that F is not a static mapping; in online-mode, the mapping F changes each time the SDSS is used and is adapted in relation to the deviation between y and $\tilde{y}$, with the aim that the deviation becomes smaller if the same vector is presented to the SDSS the next time.

Therefore, a parameter $k \in \mathbb{N}$ is introduced, describing the iteration step number the SDSS was used, and the mapping after k iteration steps is called F_k. The initial state is set to $k = 0$, the initial mapping before it is used is therefore called F_0.

This can be describe as follows:

$$F_k \xrightarrow{\text{learning}} F_{k+1} \tag{18.3}$$

In the $k - th$ iteration step, the mapping F_{k-1} is used, the input vector used in this step is defined as $x^k = (x_1^k, x_2^k, \cdots, x_n^k)^T$, the calculated output vector $y^k = (y_1^k, y_2^k, \cdots, y_m^k)^T$ and the grade of fitness determined by the user's feedback $\tilde{y^k} = (\tilde{y_1^k}, \tilde{y_2^k}, \cdots, \tilde{y_m^k})^T$, resulting in a change of the mapping from F_{k-1} to F_k .

$$F_k : \ K \to [0,1]^m \tag{18.4}$$

$$(x_1^k, \cdots, x_n^k)^T \mapsto F((x_1, \cdots, x_n)^T) = (y_1, \cdots, y_m)^T \tag{18.5}$$

Possible solutions that are able to perform such an adaptive and learning mapping for vague and fuzzy data are, for example, FLC, ANN, or *neuro-fuzzy systems* as introduced in chapter 16.

18.2 Domain and co-domain of the proposed SDSS

The situation in the analyzed LL framework can be described as a set of people or community members who want to mitigate the risk to suffer from a agrochemical-related disease. Risk mitigation is performed with the help of an SDSS. To generate SDSS, a mapping machine is used that maps from an input vector x to a output vector y. In the input vector, different values for parameters determining the fitness of the proposed risk mitigation strategies are stored.

Each user has an own risk profile and values of parameters that can be used to describe the

fitness of different risk mitigation strategies. In the user's profile, the value for each parameter determining the risk and fitness, such as the literacy rate, health history, constraints in the use of risk mitigation strategies due to religious or social aspects, availability of mobile devices, etc., can be stored.

Additionally, environmental and logistical parameters, such as temperature, precipitation, pesticide application patterns, and availability of risk mitigation resources also determine the fitness of risk mitigation strategies. They can be partly recorded with sensors or determined from available databases connected to the SDSS, for example, temperature, precipitation, or maps visualizing the availability of risk mitigation resources.

Mathematically the values of the user and environmental profile can be stored in a j-dimensional tuple, which can be regarded as an input vector for the SDSS. The elements incorporated in the input tuple must be determined by scientists, physicians, and community members and can be determined partly in an LL approach, the values of the parameters in the user profile in interviews, questionnaires, or medical investigations. To get an initial set of minimal data elements for the input vector, a similar framework to that used by [TBG+06] to identify minimum data elements for care of patients is proposed. [GÏ4] used this approach to identify minimal data elements for CKD patients. The minimal data elements defined in [GÏ4] can be used as a initial framework. However, the granularity is not as fine as needed and must be adjusted over time.

18.3 Developed SDSS

Derived from the results of this thesis and the requirements defined in section 15.2, the SDSS visualized in figure 18.1, which consists of different sub-components, is proposed as an initial base for use in a research and development cycle in the described LL framework. In general, the SDSS consists of different sub-modules: modules for data preparation with abilities for incomplete data handling and interpolation tasks and a data quality estimator module and adaptive mapping machine with the ability to learn from users' feedback. Examples for possible mathematical methods and related software solutions which can be used for the sub-modules are described in chapters 7 and 16.

According to requirement **R4.6** (section 15.2), the software components of the SDSS should be free for use in less-developed countries and with the opportunity to modify related software

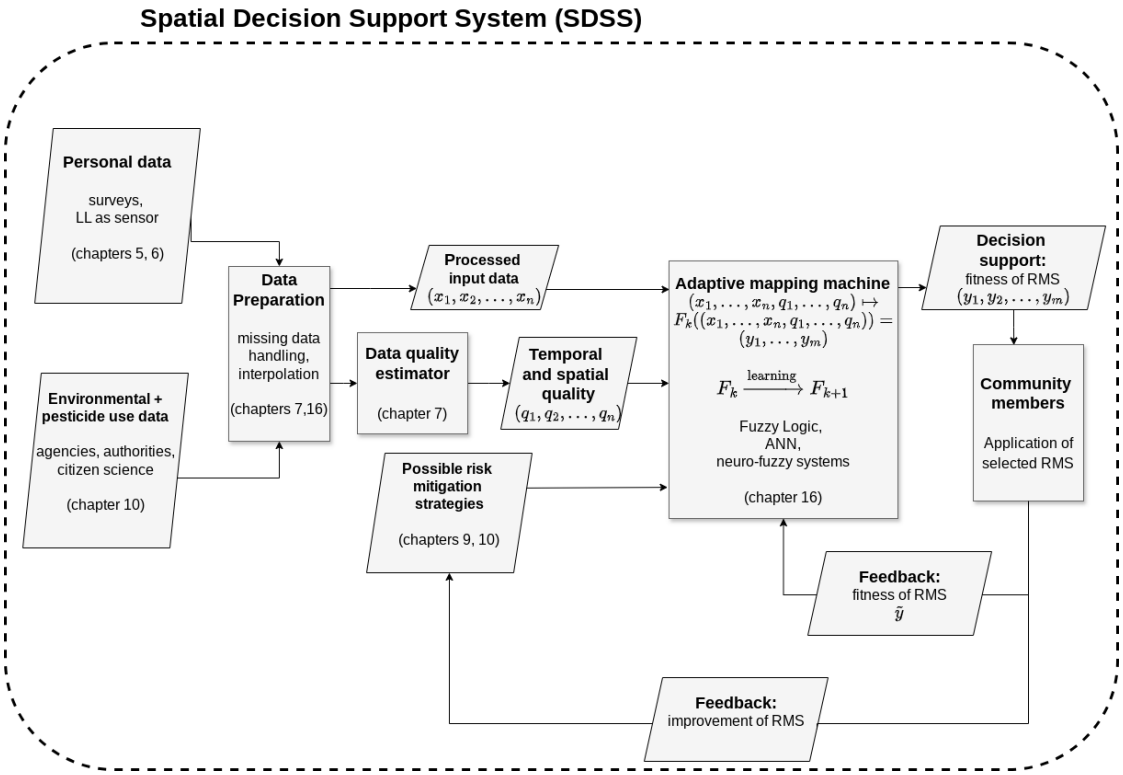

Figure 18.1: Flowchart of an SDSS in the proposed framework (figure generated with *draw.io*).

code. The concept of open-source software (section 5.3.2.1) with its different offered licenses fulfills this requirement. Therefore, the software solutions for the sub-modules of the SDSS in this thesis are completely open source. A list of the proposed open-source software modules is provided in Appendix C. However, the proposed software modules are only an example for a base SDSS; during the research and development cycle of the SDSS in an LL approach, the selected methods and related software modules might change.

Input data as described in section 18.2 are parameters helping to model the fitness of different risk mitigation strategies. They are derived from scientific literature or expert knowledge and can be grouped into personal data gained, for example, through surveys performed in the LL community as well as environmental and pesticide use data delivered by agencies and authorities or collected in a citizen-science approach (chapter 6). By incorporating personal data, decision support tailored to the user can be obtained, fulfilling requirement **R4.1**. These parameters are not a closed set; they might change over time, e.g., if new scientific or expert knowledge about the topic is gained.

Following this, input data is processed by a data preparation module. The proposed data preparation module performs missing data handling within datasets and interpolation tasks of the input data in order to be usable by the adaptive mapping machine or to obtain data

for missing data items. Possible methods to deal with missing data items and implemented open-source software tools can be found in section 7.4 and for interpolation tasks in Section 7.2. Missing data handling is necessary because some of the methods proposed as adaptive mapping methods, e.g., ANN, are based on complete data items. Interpolation in the described task is used to obtain new, unknown data between data items, e.g., in a discrete mesh of sampling points. Processed input data is further used as input data in the adaptive mapping machine. This sub-module is related to requirement **R4.4**, which holds that the system should be able to deal with incomplete and noisy data. However, there are also methods available that can work directly with noisy or missing data, such as *concurrent neuro-fuzzy systems* (section 16.4.1).

Connected to the data preparation module is a data quality estimator that is able to determine temporal and spatial data quality of the input data. The quality of the input data also determines the quality of the generated decision support, which is information necessary to estimate the validity of the decision support. Methods to estimate temporal and spatial data quality are described in section 7.3. As stated in requirement **R4.5**, the SDSS should be also able to work with aggregated data and with data heterogeneity. The use of a measure for data quality also allows the use of different data sources with different data quality, resulting in different decision support and related quality measures. For example, in chapter 11, a method was described for the way nearly real-time information about the presence of pesticides in the environment can be gained with a citizen-science approach. Data obtained with this approach has a relatively high temporal and spatial quality. If such precise data is not available, a fallback to a proxy variable with a lower temporal or spatial quality, for example, pesticide application maps averaged over a year, should be possible. However, decision support in terms of avoiding areas with pesticides in the environment generated with application maps over a year has a lower temporal quality than the same decision support generated with nearly real-time data.

Processed input data and temporal and spatial quality are further used in the adaptive mapping machine to generate decision support, for example, the grade of fitness for different in the LL risk mitigation strategies offered. A system with learning ability (requirement **R4.2**) is necessary to improve decision making and adapt it to the user's needs or different regions. Methods with a learning ability are, for example, *neuro-fuzzy systems* with parameter tuning (section 16.4.2) or ANN (section 16.3.3). In the case of CKDu in El Salvador, the system should be able to work with fuzzy logic and rules (**R4.7**) and if possible, be able to extract

logical rules (**R4.8**). Methods fulfilling requirement **R4.7** are, for example, all *neuro-fuzzy systems* introduced in section 16.4 or algorithms working on fuzzy logic with the ability to extract fuzzy rules are FCM or *Wang and Mendel algorithm* (section 16.4.3).

Decision support generated by the adaptive mapping machine, for example, a vector with the fitness of different risk mitigation strategies, is further delivered to the user of the SDSS, a person living in the LL, and the best-fitting risk mitigation strategy is applied by the user. The user can give feedback about the personal fitness of the proposed risk mitigation strategy, which is then delivered back to the adaptive mapping machine, resulting in the adaption of the parameters of the mapping machine, for example, weights of a digraph, in order to improve decision making in the next iteration step.

The user has a second path for feedback; they can give feedback about the risk mitigation strategy itself in the research and development cycle to improve the risk mitigation strategy or used methodologies within the research and development cycle in the LL together with the involved stakeholders. Examples for low-cost risk mitigation strategies usable in the described framework can be taken from part III.

The SDSS is implemented in the research and development cycle of an LL (section 5.3). Sub-components are selected, evaluated, and improved or adapted by the users and stakeholders in the LL approach. Sub-components can also be, for example, removed or added and evaluated in the LL with the aim of improving the decision-making process.

Figure 18.1 gives only an example about how such an SDSS can look at the start of the research and development cycle and that fulfills the requirements defined in section 15.2. However, during the research and development cycle, the structure and sub-components of the SDSS are also under permanent evaluation and might change over time.

18.4 Chapter discussion

In the previous chapter, an SDSS was developed that is based on the requirements defined in section 15.2 and derived from the characteristics of a disease with an unknown etiology related to agrochemicals in less-developed countries.

The resulting SDSS can deal with fuzzy or vague temporal and spatial data and gives decision support. It can be adapted to the needs of people living in rural areas in less-developed countries and is completely open source and free of charge. It is not only restricted to the case

of CKD in El Salvador, but through the open source and learning approach, it can be also adapted to other regions with similar diseases. In contrast to existing SDSS, the developed SDSS is not a static one; it is modified over time within the research and development cycle in an LL in order to improve decision making.

In contrast to the SDSS described in [RZJ+20], where a crop model called *Quantitative Evaluation of the Fertility of Tropical Soils* (QUEFTS) was used to adjust the fertilizer amount to site specific conditions, the SDSS described in this thesis has a broader application with different risk mitigation strategies to minimize the released agrochemical amount or to minimize the exposure to agrochemicals when walking through a contaminated areas. Other risk mitigation strategies can be also implemented in the SDSS described in this thesis.

[ZMBM20] identified several challenges for SDSS applied in agriculture: low usability and accessibility of GUI, missing functionalities for decision making over different time spans (short, mid and long term decisions), missing adaption to uncertainty and dynamic changes, implementing expert knowledge to adjust faulty decision support, implement historical information to improve the decision support, implement decision correction mechanisms, and allow forecasting for future decisions.

A SDSS with a fuzzy interference system to give decision support related to irrigation is described in [GML15] in which the system shows good performance. However, the SDSS described in this thesis is not restricted to a special decision support method due to its modular conception. The chosen LL approach allows the testing or adjusting of different decision support methods within the research and development cycle.

19 Discussion and conclusion

As described in section 1.2, the overall aim of this thesis is to address user-driven innovation for risk mitigation under the given local and regional requirements and constraints. The evaluated structure should be able to adapt to these local, regional, or even individual requirements and constraints to give decision support to a person in the field of risk mitigation related to exposure to hazardous agricultural substances with low-cost methods to be usable in a less-developed country. To achieve this aim, four research hypotheses were formulated in section 1.2:

H_1: A user-centered research and development environment can be adapted with the aim of developing risk mitigation strategies related to agrochemicals or CKDu in less-developed countries.

H_2: It is possible to generate an initial base repository of risk mitigation strategies for a user-centered research and development environment.

H_3: Mathematical methods can be used to develop an SDSS to find the best fitting risk mitigation strategies tailored to personal characteristics and available resources with fuzzy data.

H_4: Related IT systems and risk mitigation strategies can be developed with open source and low-cost methods.

In order to test hypotheses H_1 the characteristics of developing countries and agriculture practiced under these local and regional requirements and constraints were analyzed in section 2.2.4. Derived from these characteristics, a requirement and constraints analysis for the user-centered research and development environment was developed in section 5.2. The requirement and constraints analysis showed that the concept LL fits as a research and development environment. In contrast to the general application fields of LL, where often a

commercial market interest is the main driver for the LL application, in section 5.3, the LL approach was adapted to a low-cost approach in which spatial and temporal data can be processed. LL is used as a research and development environment and as a sensor through which data is collected by community members for example by surveys or in a citizen-science approach (chapter 6), and low-cost risk mitigation strategies (chapter 9) can be further developed and evaluated in the LL approach. Additionally, the LL itself is designed as a low-cost method, where open-source software is used to operate the LL (section 17.4). Therefore, H_1 can be accepted. Low-cost methods play an important role for upscaling successful LL approaches because the costs for adaptation and reusability are reduced to the licensing.

Regarding hypothesis H_2 in chapter 9, requirements for possible low-cost risk mitigation strategies in the described framework were defined. According to these requirements, different exemplary low-cost risk mitigation strategies were developed in part III that can be regarded as an initial base repository of risk mitigation strategies usable in the LL approach. In contrast to existing strategies to minimize, for example, agrochemical inputs, the proposed risk mitigation strategies work on the software side with open source and free of charge software tools and use free available data sources. The developed risk mitigation strategies can be applied and further developed by the user's feedback in an LL approach (section 18.3). Therefore, H_2 can be also accepted.

Requirements for an SDSS in the described framework related to H_3 were defined in chapter 15. Mathematical methods usable for an SDSS and which are at the same time implemented in open-source software like FLC, ANN or *neuro-fuzzy systems* were introduced in chapter 16. The driver for neuro-fuzzy systems are the adaptivity of the SDSS and the representation of expert knowledge to leverage information even if the source is fuzzy. Technical and software-side implementation were discussed in chapter 17. As quality of the generated spatial decision support is always connected to the quality of the input parameters that are used to generate spatial decision support, methods to rate the temporal and spatial quality of different parameters were developed in chapter 7. In section 18.3, an exemplary structure of an SDSS derived from the mentioned requirements, consisting of different sub-components and usable in the research and development environment LL, were developed. Summarizing, H_3 can be also accepted.

To check hypothesis H_4, it was analyzed if all necessary software components related to the SDSS and LL are available as open-source software, if they are interoperable, and if risk mitigation strategies can be developed as low-cost risk mitigation strategies. For all tasks

necessary to operate an LL and to generate and deliver spatial decision support, open-source software solutions can be found (chapter 17). Additionally, an example of different low-cost risk mitigation strategies usable in the described LL were developed (part III). These are the decision options of the SDSS. H_4 can be also accepted.

Summarizing, the aims of this thesis as defined in section 1.2 were achieved and research hypotheses were accepted. A concept for a research and development environment for an agrochemical disease of unknown etiology in a less-developed country was worked out together with a low-cost SDSS and risk mitigation strategies. Requirements and constraints are heterogeneous and a successful risk mitigation with method $A1$ at location $L1$ might not be successful at location $L2$ and vice versa. Therefore, in addition to a classical scientific approach to show which mitigation is in general the most helpful one, the presented approach in this thesis integrates the specific regional and local requirements and constraints into adaptation of response options for communities or specific social, cultural, or economic settings.

The benefit of the thesis in contrast to other scientific research based on risk mitigation strategies lies in the user-centered approach with its low-cost components and adaptivity in an agrochemical-related context. Through the combination of ecotoxicological methods with mathematical modeling and the LL in an open community approach, this could be achieved. An adaptive SDSS with the ability to learn from existing cases is the core of this concept. Through the user-centered innovation approach in the LL, the knowledge of community members can be used. The used low-cost approach allows the use in less-developed countries. This is a benefit in contrast to other pesticide reduction approaches, which focus mostly on high tech methods for industrialized agriculture.

The theoretical results of this thesis have not been implemented in a productive system to date. The results of the thesis were originally intended to be implemented into the concrete development of an LL as part of the LLinES project related to CKDu in El Salvador. However, the LLinES project had to be terminated due to security risks and gang crime in the Bajo Lempa pilot region and changes in health policy objectives in El Salvador. As mentioned in section 2.1.1, CKDu has a large spatial distribution over the whole world. It would be desirable that a new pilot region can be found to implement the developed concept in a CKDu affected area or region with other agrochemical related disease with a temporal and spatial dimension in a less-developed country.

However, the chosen multidisciplinary research and development approach in an LL with a modular open-source SDSS can be adapted to other topics and fields. The selected LL ap-

proach with low-cost methods can be further used for tasks outside of agrochemical-related topics in less-developed countries or in regions with low economic areas, and the proposed software modules are open source and can be adapted to other topics. Transfer to other applications seems to be reasonable whenever heterogeneous local requirements and constraints have an impact on the risk management options.

The proposed SDSS with the ability to handle temporal and spatial data might be usable in other contexts with a temporal and spatial dimension, for example, in an educational context. The availability of learning resources also has a temporal and spatial dimension. An SDSS might be helpful to assist, for example, pupils or students with learning difficulties to suggest learning materials tailored to the user's needs and learning difficulties and the temporal and spatial availability of learning materials. Through the use of open-source software and free available software tools, such a system can also be used in regions with low economic resources or a weak educational system. The adaptation to local and regional requirements and constraints is a key element and underlying principle of this thesis. The use of open source supports the freedom of communities to adapt a given SDSS to the requirements and constraints of the communities.

The rise of AI-based methods and especially chatbots opens new applications of AI-based tools in the described risk mitigation approach. AI and ANN are not only used to estimate the fitness of risk mitigation strategies, but they can also be used as risk mitigation strategies themselves. For example, chatbots based on *large language model* (LLM) can be used as a risk mitigation strategy. Community members can ask questions about pesticide reduction or application and get answers from the chatbot. LLM based chatbots have limitations, for example, hallucinations, where the chatbot produces good sounding answers but with wrong or inaccurate content [HMS24]. Therefore, its use in a health-related context must be done under high caution.

The first step in the general application of the theoretical results in a prototype system would be the implementation of the SDSS as an open-source software system with a user friendly GUI and defined interfaces between the used software tools. In this thesis, several building blocks and sub-modules were identified that were combined into a base SDSS prototype for further applications and adaptations and can be used for first demonstration purposes.

The developed low-cost risk mitigation strategies for a base repository (part III) can also be used and evaluated outside of the proposed LL approach related to CKDu. Saving necessary pesticide amounts and exposure to pesticides influences expenses on agricultural inputs and

might have a positive influence on ecosystem and human health in general.

Appendices

A Equations for the calculation of application rates

Table A.1: Equations for the calculation of application rates

Application rate p(crop, country, substgroup)=	**Region**	**Sources**
partly direct available or calculated for year 2000 $\frac{\sum_{substances \in substancegroup} q\begin{pmatrix} crop \\ country \\ substance \\ year \end{pmatrix}}{a\begin{pmatrix} crop \\ country \\ year \end{pmatrix}}$	countries in the EU	[otEU07] [otEU14]
$\frac{1}{\lvert EU \rvert} \sum_{countries \in EU} \frac{\sum_{substances \in substancegroup} q\begin{pmatrix} crop \\ country \\ substance \\ year \end{pmatrix}}{a\begin{pmatrix} crop \\ country \\ year \end{pmatrix}}$	European countries without EU states	[otEU07] [otEU14]
$\frac{\sum_{substances \in substancegroup} q\begin{pmatrix} crop \\ country \\ substance \\ year \end{pmatrix}}{a\begin{pmatrix} crop \\ country \\ year \end{pmatrix}}$	states in the USA	[CF]
$\frac{1}{\lvert USA \rvert} \sum_{states \in USA} \frac{\sum_{substances \in substancegroup} q\begin{pmatrix} crop \\ country \\ substance \\ year \end{pmatrix}}{a\begin{pmatrix} crop \\ country \\ year \end{pmatrix}}$	Canada	[CF]

Application rate p(crop, country, substgroup)=	**Region**	**Sources**
$\frac{\sum_{\substack{substances \in \\ substancegr}} q\begin{pmatrix} crop \\ country \\ substance \\ year \end{pmatrix}}{a\begin{pmatrix} crop \\ country \\ year \end{pmatrix}}$	South Africa	unpublished source
$\frac{\sum_{\substack{substances \in \\ substancegroup}} q\begin{pmatrix} crop \\ country \\ substance \\ year \end{pmatrix} \cdot \mu\begin{pmatrix} crop \\ country \end{pmatrix}}{a\begin{pmatrix} crop \\ country \\ year \end{pmatrix}}$	Zimbabwe, Kenya, Ivory Coast	[fIDU94], [fIDU94] [FD11b] , [FD11c]
$\frac{1}{3} \sum_{\substack{Zimbabwe, Kenya, \\ IvoryCoast}} \frac{\sum_{\substack{substances \in \\ substancegroup}} q\begin{pmatrix} crop \\ country \\ substance \\ year \end{pmatrix} \cdot \mu\begin{pmatrix} crop \\ country \end{pmatrix}}{a\begin{pmatrix} crop \\ country \\ year \end{pmatrix}}$	rest of Africa	[fIDU94], [fIDU94] [FD11b] , [FD11c]
$\frac{1}{\lvert EU \rvert + \lvert USA \rvert} \left(\sum_{\substack{countries \in \\ EU}} \frac{\sum_{\substack{substances \in \\ substancegroup}} q\begin{pmatrix} crop \\ country \\ substance \\ year \end{pmatrix}}{a\begin{pmatrix} crop \\ country \\ year \end{pmatrix}} + \frac{\sum_{\substack{substances \in \\ substancegroup}} q\begin{pmatrix} crop \\ country \\ substance \\ year \end{pmatrix}}{a\begin{pmatrix} crop \\ country \\ year \end{pmatrix}} \right)$	other countries	[otEU07], [otEU14] [CF]

B | Maps visualizing the crop responsible for the highest applied substance amount

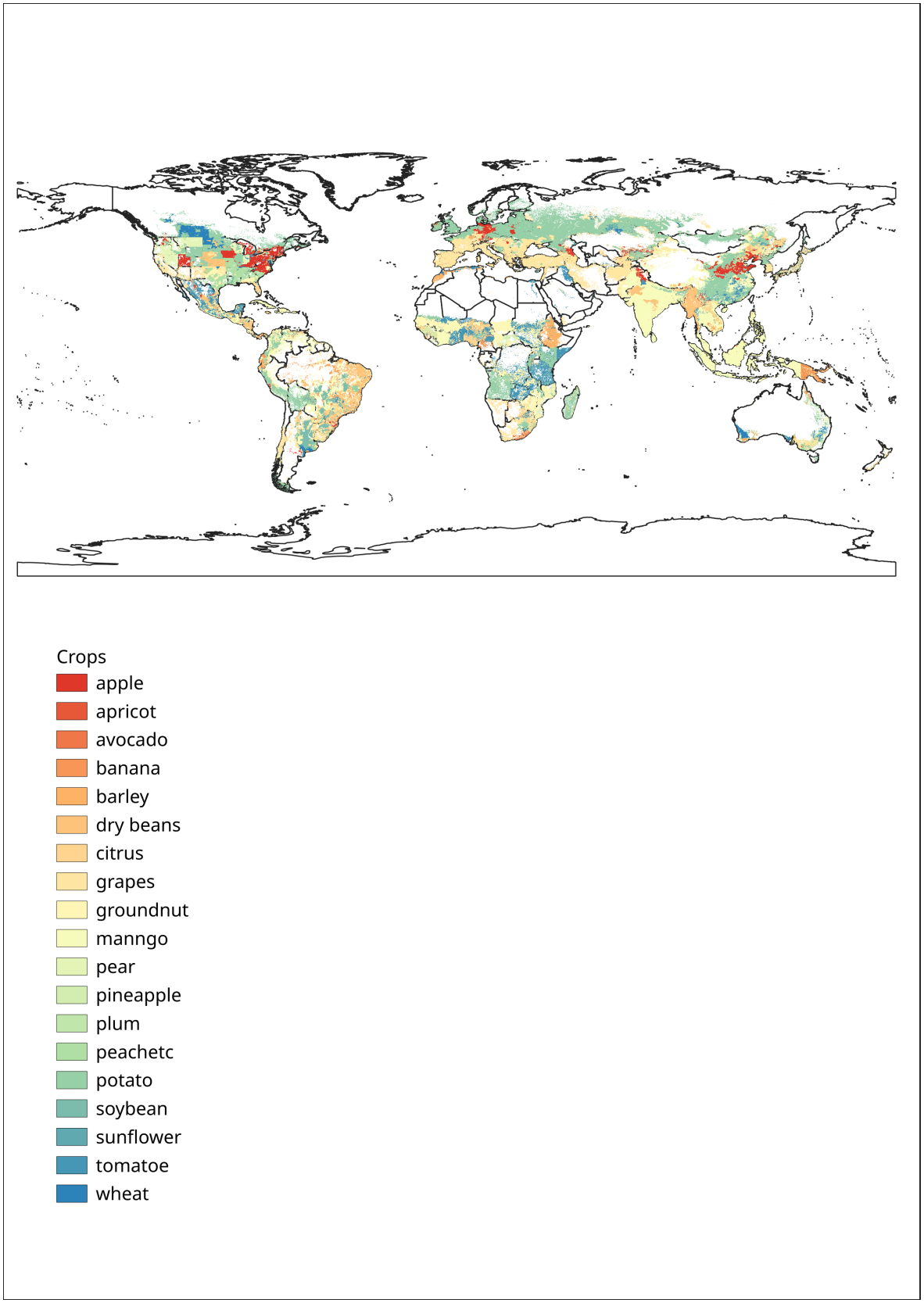

Figure B.1: Map visualizing the crop responsible for the highest applied fungicide amount per raster cell (figure generated with *QGIS*).

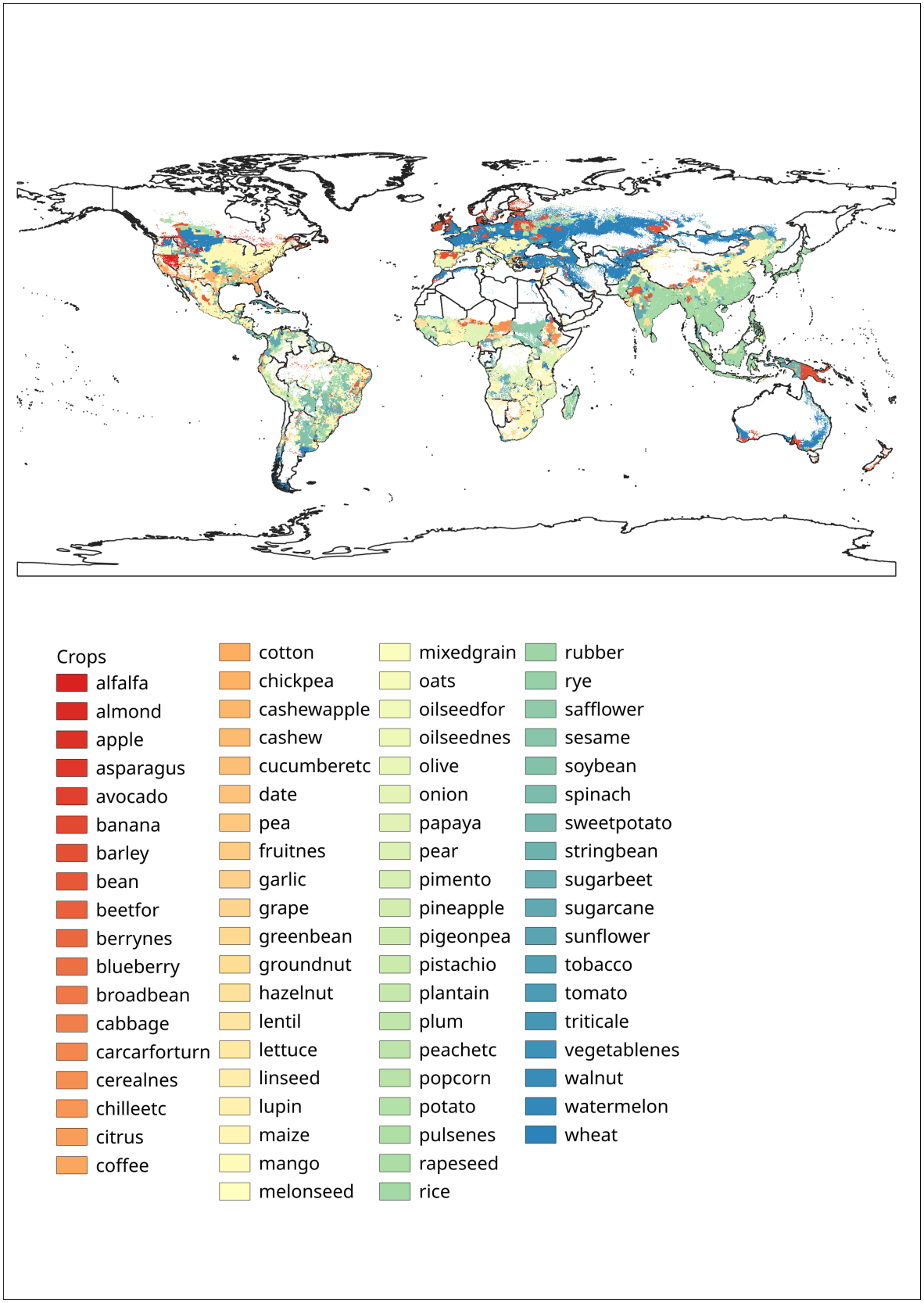

Figure B.2: Map visualizing the crop responsible for the highest applied herbicide amount per raster cell (figure generated with *QGIS*).

C | Table tasks in the LL and possible open-source software solutions

Table C.1: Tasks in the LL and possible open-source software solutions

General office work	
Task	**Possible open-source software solution**
Operating system	Linux, e.g., Lubuntu [Fou20f]
Coordination of the LL approach, project management	Weekan [Oja20], OpenProject
Communication via internet	Thunderbird [Fou20d], Firefox [Fou20c]
Generation of documents: working documents, reports and public relations	LibreOffice [Fou20e] , Apache OpenOffice [Fou20b], Atom TextEditor [Git21]
Generation of scientific publications	MiKTeX [Sch21], R [R C20], R-Studio [RSt20]: knitr [Xie20], rticles [AXR+20]
Software for data delivery and data collection	
tools to collect and track personal data from the community members (historical, behavior, medical and related to the given SDS) and of environmental parameters	R [R C20], R-Studio [RSt20]: shiny [CCA+20], shiny-server [RSt21], geoloc [Fay21] OpenDataKit [BOTS19], SDAPS [Ber14], Mozilla DeepSpeech [text-to-speech] [Cor20]
deliver SDS, early warning and risk mitigation materials from the server to the user (client)	Apache HTTP Server [Fou20a], ShinyServer [RSt21], MySQL relational database mngmnt system [Cor21], Firefox (client) [Fou20c], MediaWiki [Fou20g], Ampache media server [Tea20]

Software for data storage and manipulation	
Generate, access and manipulate databases for tempospatial and nontempospatial data	MySQL relational database mngmnt system [Cor21], R [R C20], R-Studio [RSt20], GNU Octave [EBHW20] , PostgreSQL [PGDG08], PostGIS [Pos24]
content collaboration platform, cloud	Nextcloud [Nex24]
imputation	mice [VBGO11],amelia [HKB11], Hmisc [HJHJ19], missForest [SB12]
Software for mathematical modeling	
generate and to work with maps with a temporal and spatial dimmension related to risk and ressource maps	R [R C20], R-Studio [RSt20]: sf [Peb18], sp [PB05], maps [BWB$^+$18], raster [Hij23], stars [PB23] , QGIS [QGI24], GRASS GIS [Tea15] SAGA-GIS [SAG24], Python [VRD09] GNU Octave [EBHW20], terra [Hij24]
generate algorithms for SDSS and early warning	R, R-Studio: nnet [VR02], neuralnet [FGW19], RSNNS [BB12], FuzzyR [CGR21], lfl [BŠ21] deepnet [Ron22], GNU Octave [EBHW20]
Software for generation of materials for risk mitigation	
create materials for risk mitigation: visualisation (digital and paper based)	Blender [Com18], GIMP [The19], office tools, GIS tools, A-Frame [MMN24], ARToolKit [CD16]
create digital materials for risk mitigation: audio and video	OpenShot [OS24] , OBS [Bai24], Audacity [Tea24], Libre Translate , MaryTTS (TextToSpeech) [ST03]

Data center	
Operating system for server	Linux, e.g., Ubuntu [Can24]
virtualisation platform	Proxmox [Pro24]
user management	OpenLDAP [Chu07]
E-Mail frontend	Sogo,
E-Mail backend	Postfix, Dovecot
Firewall	OPNsense

Bibliography

[AAS14] Ather Ashraf, Muhammad Akram, and Mansoor Sarwar. Type-II Fuzzy Decision Support System for Fertilizer. *The Scientific World Journal*, 2014(1):9, 2014.

[ABA05] Mohammed Attik, Laurent Bougrain, and Frédéric Alexandre. Neural network topology optimization. In *International Conference on Artificial Neural Networks*, pages 53–58. Springer, 2005.

[Abr01] Ajith Abraham. Neuro fuzzy systems: State-of-the-art modeling techniques. In *International Work-Conference on Artificial Neural Networks*, pages 269–276. Springer, 2001.

[ACR+17] Enriqueta Anticó, Sergi Cot, Alexandre Ribó, Ignasi Rodríguez-Roda, and Clàudia Fontàs. Survey of Heavy Metal Contamination in Water Sources in the Municipality of Torola, El Salvador, through In Situ Sorbent Extraction. *Water*, 9(11):877, 2017.

[ADCG16] Matt Aitkenhead, David Donnelly, Malcolm Coull, and Richard Gwatkin. Estimating soil properties with a mobile phone. *Digital soil morphometrics*, pages 89–110, 2016.

[ADSZ+19] Christine Anhalt-Depies, Jennifer L Stenglein, Benjamin Zuckerberg, Philip A Townsend, and Adena R Rissman. Tradeoffs and tools for data quality, privacy, transparency, and trust in citizen science. *Biological Conservation*, 238:108195, 2019.

[AF04] David Ansell and Terry Feest. Uk renal registry report 2004. Technical report, UK Renal Registry, 2004.

[Age15a] European Chemicals Agency. Material safety data sheet, 2015. Accessed: 2015-09-01. URL: `http://echa.europa.eu/regulations/reach/safety-data-sheets`.

[Age15b] European Space Agency. Sentinel Data Access Description, 2015. Accessed: 2015-10-06. URL: `https://sentinel.esa.int/web/sentinel/sentinel-data-access-description;jsessionid=742EC183E4E35CD7282BC8B114E0C74B`.

[AGRK20] Federico Amato, Fabian Guignard, Sylvain Robert, and Mikhail Kanevski. A novel framework for spatio-temporal prediction of environmental data using deep learning. *Scientific reports*, 10(1):22243, 2020.

[(AH13] Agricultural Handler Exposure Task Force (AHETF). AHETF Accomplishments, 2013. Accessed: 2013-09-27. URL: `http://www.exposuretf.com/Home/AHETF/AHETFAccomplishments/tabid/82/Default.aspx`.

[AHB+09] Yaw Anokwa, Carl Hartung, Waylon Brunette, Gaetano Borriello, and Adam Lerer. Open source data collection in the developing world. *Computer*, 42(10):97–99, 2009.

[AHK+15] Tilo Arens, Frank Hettlich, Christian Karpfinger, Ulrich Kockelkorn, Klaus Lichtenegger, and Hellmuth Stachel. *Mathematik*. Springer-Verlag, 2015.

[AHM03] Arif S Al-Hammadi and Robert H Milne. A neuro-fuzzy approach for student performance modeling. In *10th IEEE International Conference on Electronics, Circuits and Systems, 2003. ICECS 2003. Proceedings of the 2003*, volume 3, pages 1078–1081. IEEE, 2003.

[AKIK21] Poonam Adhikari, Ritesh Kumar, SR Iyengar, and Rishemjit Kaur. What a million Indian farmers say?: A crowdsourcing-based method for pest surveillance. In *KDD Workshop on Data-driven Humanitarian Mapping, 27th ACM SIGKDD Conference,*, 2021.

[Alz10] Abdul-Qadim Alzalloum. *Application of shortest path algorithms to find paths of minimum radiation dose.* PhD thesis, University of Illinois at Urbana-Champaign, 2010.

[AM14] Stacey E Anderson and B Jean Meade. Potential health effects associated with dermal exposure to occupational chemicals. *Environmental health insights*, 8:EHI–S15258, 2014.

[Ama14] André FS Amaral. Pesticides and asthma: challenges for epidemiology. *Frontiers in public health*, 2:6, 2014.

[Amm20] Adriano Luiz Ammirati. Chronic kidney disease. *Revista da Associação Médica Brasileira*, 66:03–09, 2020.

[AN15a] National Aeronautics and Space Administration (NASA). Measuring vegetation (NDVI and EVI), 2015. Accessed: 2015-10-14. URL: `http://earthobservatory.nasa.gov/Features/MeasuringVegetation/measuring_vegetation_1.php`.

[AN15b] National Aeronautics and Space Administration (NASA). The Landsat Program, 2015. Accessed: 2015-10-08. URL: `http://landsat.gsfc.nasa.gov/`.

[And95] James A Anderson. *An introduction to neural networks*. Massachusetts Institute of Technology, 1995.

[Ant16] Kaisa Anttila. Views toward Usability in Open Source Software Projects: a Longitudinal Case Study, 2016. University of Oulu.

[Anw97] Wagida A Anwar. Biomarkers of human exposure to pesticides. *Environmental health perspectives*, 105(4):801, 1997.

[AOOWKS10] Adijah M Ali-Olubandwa, Dolphine Odero-Wanga, NJ Kathuri, and WA Shivoga. Adoption of improved maize production practices among small scale farmers in the agricultural reform era: The case of Western Province of Kenya. *Journal of International Agricultural and Extension Education*, 17(1):21–30, 2010.

[ARB13] Michael CR Alavanja, Matthew K Ross, and Matthew R Bonner. Reply to Increased cancer burden among pesticide applicators and others due to pesticide exposure. *CA: a cancer journal for clinicians*, 63(5):366–367, 2013.

[Aro05] SM Aronson. The dancing cats of Minamata Bay. *Medicine and health, Rhode Island*, 88(7):209, 2005.

[ARP+12] Yaw Anokwa, Nyoman Ribeka, Tapan Parikh, Gaetano Borriello, and Martin C Were. Design of a phone-based clinical decision support system for resource-limited settings. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development*, pages 13–24. ACM, 2012.

[AS07] Akarachai Atakulreka and Daricha Sutivong. Avoiding local minima in feedforward neural networks by simultaneous learning. In *Australasian Joint Conference on Artificial Intelligence*, pages 100–109. Springer, 2007.

[AS12] Gabriella Andreotti and Debra T Silverman. Occupational risk factors and pancreatic cancer: A review of recent findings. *Molecular carcinogenesis*, 51(1):98–108, 2012.

[ASA14] Mohammed Al-Shammaa and Maysam F Abbod. Automatic generation of fuzzy classification rules from data. In *Proceedings of the 2014 International Conference on Neural Networks-Fuzzy Systems (NN-FS 14), Venice*, 2014.

[ASC09] Wasim Aktar, Dwaipayan Sengupta, and Ashim Chowdhury. Impact of pesticides use in agriculture: their benefits and hazards. *Interdisciplinary Toxicology*, 2(1):1–12, 2009.

[ASJB11] Kishor Atreya, Bishal K Sitaula, Fred H Johnsen, and Roshan M Bajracharya. Continuing issues in the limitations of pesticide use in developing countries. *Journal of Agricultural and Environmental Ethics*, 24(1):49–62, 2011.

[ASMW98] Jeffrey G Arnold, Raghavan Srinivasan, Ranjan S Muttiah, and Jimmy R Williams. Large area hydrologic modeling and assessment part I: model development 1. *JAWRA Journal of the American Water Resources Association*, 34(1):73–89, 1998.

[Ass08] American Veterinary Medical Association. One health: A new professional imperative. *One Health Initiative Task Force Final Report*, 2008.

[Awu97] RT Awuah. An evaluation of some systemic fungicides for the control of septoria leaf spot of tomato. *Ghana Journal of Agricultural Science*, 30(1):71–78, 1997.

[AXR+20] JJ Allaire, Yihui Xie, R Foundation, Hadley Wickham, Journal of Statistical Software, Ramnath Vaidyanathan, Association for Computing Machinery, Carl Boettiger, Elsevier, Karl Broman, Kirill Mueller, Bastiaan Quast, Randall Pruim, Ben Marwick, Charlotte Wickham, Oliver Keyes, Miao Yu, Daniel Emaasit, Thierry Onkelinx, Alessandro Gasparini, Marc-Andre Desautels, Dominik Leutnant, MDPI, Taylor and Francis, Ouzhan Oreden, Dalton Hance, Daniel Nuest, Petter Uvesten, Elio Campitelli, John Muschelli, Alex Hayes, Zhian N. Kamvar, Noam Ross, Robrecht Cannoodt, Duncan Luguern, David M. Kaplan, Sebastian Kreutzer, Shixiang Wang, Jay Hesselberth, and Christophe Dervieux. *rticles: Article Formats for R Markdown*, 2020. R package version 0.15. URL: `https://CRAN.R-project.org/package=rticles`.

[Aza10] Ahmad Taher Azar. Adaptive neuro-fuzzy systems. *Fuzzy systems*, 42(11):85–110, 2010.

[Bai24] Lain Bailey. Open broadcaster software, 2024. 2024. URL: `https://docs.obsproject.com/`.

[Ban16] The World Bank. World Bank Open Data: Employment in agriculture, 2016. Accessed: 2016-03-29. URL: `http://data.worldbank.org/indicator/SL.AGR.EMPL.ZS/countries?page=3&display=default`.

[Ban19a] The World Bank. The World Bank in El Salvador, 2019. Accessed: 2019-09-25. URL: `https://www.worldbank.org/en/country/elsalvador/overview#1`.

[Ban19b] The World Bank. World Bank Open Data: GDP per capita (current US$), 2019. Accessed: 2019-09-30. URL: `http://data.worldbank.org/indicator/NY.GDP.PCAP.CD`.

[Ban19c] The World Bank. World Bank Open Data: Life expectancy at birth, 2019. Accessed: 2019-09-30. URL: `https://data.worldbank.org/indicator/SP.DYN.LE00.IN?view=chart`.

[Ban19d] The World Bank. World Bank Open Data: Literacy rate, adult total, 2019. Accessed: 2019-09-25. URL: `https://data.worldbank.org/indicator/se.adt.litr.zs`.

[Ban19e] The World Bank. World Bank Open Data: Mortality rate, under-5 (per 1000 live births), 2019. Accessed: 2019-09-25. URL: `https://data.worldbank.org/indicator/SH.DYN.MORT?view=chart`.

[Ban19f] The World Bank. World Bank Open Data: Population, total, 2019. Accessed: 2019-09-25. URL:`https://data.worldbank.org/indicator/SP.POP.TOTL`.

[Ban19g] The World Bank. World Development Indicators: Land area, 2019. Accessed: 2019-09-21. URL: `https://databank.worldbank.org/reports.aspx?source=2&series=AG.LND.TOTL.K2&country=`.

[Bar02] Rashad S Barsoum. Overview: End-Stage Renal Disease in the Developing World. *Artificial organs*, 26(9):737–746, 2002.

[Bar06] Rashad S Barsoum. Chronic kidney disease in the developing world. *The New England journal of medicine*, 354(10):997–999, 2006. PMID: 16525136.

[BAR+17] Katrin Bieger, Jeffrey G Arnold, Hendrik Rathjens, Michael J White, David D Bosch, Peter M Allen, Martin Volk, and Raghavan Srinivasan. Introduction to SWAT+, a completely restructured version of the soil and water assessment tool. *JAWRA Journal of the American Water Resources Association*, 53(1):115–130, 2017.

[BB12] Christoph Bergmeir and José M. Benítez. Neural Networks in R Using the Stuttgart Neural Network Simulator: RSNNS. *Journal of Statistical Software*, 46(7):1–26, 2012.

[BBB+91] Donald J Baird, Ian Barber, Mairead Bradley, Amadeu MVM Soares, and Peter Calow. A comparative study of genotype sensitivity to acute toxic stress using clones of Daphnia magna straus. *Ecotoxicology and Environmental Safety*, 21(3):257–265, 1991.

[BBJ01] Sabine Beulke, Colin D Brown, and Nicholas J Jarvis. MACRO: a preferential flow model to simulate pesticide leaching and movement to drains. In *Modelling of environmental chemical exposure and risk*, pages 117–132. Springer, 2001.

[BBJ+09] Rick Bonney, Heidi Ballard, Rebecca Jordan, Ellen McCallie, Tina Phillips, Jennifer Shirk, and Candie C Wilderman. Public Participation in Scientific Research: Defining the Field and Assessing Its Potential for Informal Science Education. A CAISE Inquiry Group Report. Technical report, Center for Advancement of Informal Science Education (CAISE), Washington, D.C., 2009.

[BCD+13] Marta Bottero, Elena Comino, Marco Duriavig, Valentina Ferretti, and Silvia Pomarico. The application of a Multicriteria Spatial Decision Support System (MCSDSS) for the assessment of biodiversity conservation in the Province of Varese (Italy). *Land use policy*, 30(1):730–738, 2013.

[BD02] Richard P Bagozzi and Utpal M Dholakia. Intentional social action in virtual communities. *Journal of interactive marketing*, 16(2):2–21, 2002.

[Bel58] Richard Bellman. On a routing problem. *Quarterly of applied mathematics*, 16(1):87–90, 1958.

[Ber14] Benjamin Berg. SDAPS - Scripts for data acquisition with paper-based surveys, 2014. Accessed: 2014-06-02. URL: `http://sdaps.org/`.

[Bez73] James Christian Bezdek. *FUZZY-MATHEMATICS IN PATTERN CLASSIFICATION.* Cornell University, 1973.

[BGK14] Mirco Bundschuh, Willem Goedkoop, and Jenny Kreuger. Evaluation of pesticide monitoring strategies in agricultural streams based on the toxic-unit concept - experiences from long-term measurements. *Science of the Total Environment*, 484:84–91, 2014.

[BGMS18] Debjani Bhowmick, Deepak K. Gupta, Saumen Maiti, and Uma Shankar. Deep Autoassociative Neural Networks for Noise Reduction in Seismic data. *ArXiv*, abs/1805.00291, 2018.

[BH97] Astrid Boos-Hersberger. Transboundary water pollution and state responsibility: the Sandoz Spill. *Annual Survey of International & Comparative Law*, 4:103, 1997.

[Bib52] Holy Bible. Revised standard version. *Reference Edition, New York: Thomas Nelson & Sons*, 1952.

[Bie97] Benno Biewer. *Fuzzy-methoden.* Springer, 1997.

[BKHS09] BHMSA Bergvall-Kareborn, M Hoist, and A Stahlbrost. Concept design with a living lab approach. In *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on*, pages 1–10. IEEE, 2009.

[BKKN03] Christian Borgelt, Frank Klawonn, Rudolf Kruse, and Detlef Nauck. *Neuro-fuzzy-systeme: Von den Grundlagen künstlicher neuronaler Netze zur Kopplung mit Fuzzy-systemen.* Springer, 2003.

[BKS09] Birgitta Bergvall-Kåreborn and Anna Ståhlbröst. Living Lab: an open and citizen-centric approach for innovation. *International Journal of Innovation and Regional Development*, 1(4):356–370, 2009.

[BKSL13] Mikhail A Beketov, Ben J Kefford, Ralf B Schäfer, and Matthias Liess. Pesticides reduce regional biodiversity of stream invertebrates. *Proceedings of the National Academy of Sciences*, 110(27):11039–11043, 2013.

[BL95] Rainer Bodendiek and Rainer Lang. Lehrbuch der Graphentheorie., 1995.

[BL03] Paul P Biemer and Lars E Lyberg. *Introduction to survey quality.* John Wiley & Sons, 2003.

[BL05] W Blum and D Leiss. Modellieren mit der tanken-aufgaben. *Mathematic Lehren (128)*, 18421, 2005.

[Bla15] Amy J Blatt. Collaborative mapping. *Health, Science, and Place*, pages 63–75, 2015.

[BLB05] Dominique Brossard, Bruce Lewenstein, and Rick Bonney. Scientific knowledge and attitude change: The impact of a citizen science project. *International Journal of Science Education*, 27(9):1099–1121, 2005.

[BLSF09] Philip Bell, Bruce Lewenstein, Andrew W Shouse, and Michael A Feder. *Learning Science in Informal Environments: People, Places, and Pursuits.* National Academies Press, 2009.

[BML11] Julia Blanco-Munoz and Marina Lacasana. Practices in pesticide handling and the use of personal protective equipment in Mexican agricultural workers. *Journal of agromedicine*, 16(2):117–126, 2011.

[BML15] Peter A Burrough, Rachael A McDonnell, and Christopher D Lloyd. *Principles of geographical information systems*. Oxford University Press, USA, 2015.

[BMM18] Denis Bassi, Marcelo Menossi, and Lucia Mattiello. Nitrogen supply influences photosynthesis establishment along the sugarcane leaf. *Scientific reports*, 8(1):2327, 2018.

[BMN+21] Bálint Balázs, Peter Mooney, Eva Nováková, Lucy Bastin, and Jamal Jokar Arsanjani. Data quality in citizen science. *The science of citizen science*, 139(10.1007):978–3, 2021.

[BMR20] Khalid Bahani, Mohammed Moujabbir, and Mohammed Ramdani. Linguistic fuzzy rule learning through clustering for regression problems. *International Journal of Intelligent Engineering and Systems*, 13(3):80–89, 2020.

[Bör09] H Börner. *Pflanzenkrankheiten und Pflanzenschutz*. Springer, 2009.

[Bor16] Folkmar Bornemann. *Funktionentheorie*. Springer, 2016.

[BOTS19] Patrick Loola Bokonda, Khadija Ouazzani-Touhami, and Nissrine Souissi. Open Data Kit: Mobile Data Collection Framework For Developing Countries. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 8(12):4749–4754, 2019.

[BP17] Arnaud Boivin and Véronique Poulsen. Environmental risk assessment of pesticides: state of the art and prospective improvement from science. *Environmental Science and Pollution Research*, 24:6889–6894, 2017.

[BPD05] P Ballon, J Pierson, and S Delaere. Open innovation platforms for broadband services: benchmarking european practices. In *16th European Regional Conference*, Portugal, 2005.

[BPR+18] Boris Bikbov, Norberto Perico, Giuseppe Remuzzi, et al. Disparities in chronic kidney disease prevalence among males and females in 195 countries: analysis of the Global Burden of Disease 2016 Study. *Nephron*, 139:313–318, 2018.

[BR12] Rangaswami Balakrishnan and Kanna Ranganathan. *A Textbook of Graph Theory*. Springer Science & Business Media, 2012.

[Bra08a] Daren C Brabham. Crowdsourcing as a model for problem solving an introduction and cases. *Convergence: the international journal of research into new media technologies*, 14(1):75–90, 2008.

[Bra08b] Daren C Brabham. Moving the crowd at istockphoto: The composition of the crowd and motivations for participation in a crowdsourcing application. *First Monday*, 13(6), 2008.

[Bri97] Joan L Brierton. Techniques for avoiding local minima in gradient-descent-based ID algorithms. In *Radar Sensor Technology II*, volume 3066, pages 130–135. International Society for Optics and Photonics, 1997.

[Bri19] Encyclopaedia Britannica. El Salvador, 2019. Accessed: 2019-09-21. URL: `https://www.britannica.com/place/El-Salvador#ref276676`.

[BRP+01] B Basso, JT Ritchie, FJ Pierce, RP Braga, and JW Jones. Spatial validation of crop models for precision agriculture. *Agricultural Systems*, 68(2):97–112, 2001.

[Bry03] Erik Bryld. Potentials, problems, and policy implications for urban agriculture in developing countries. *Agriculture and human values*, 20(1):79–86, 2003.

[BS95] Jörg Bruske and Gerald Sommer. Dynamic cell structure learns perfectly topology preserving map. *Neural computation*, 7(4):845–865, 1995.

[BŠ21] Michal Burda and Martin Štěpnička. lfl: An R package for linguistic fuzzy logic. *Fuzzy Sets and Systems*, 2021.

[BSD+13] Waylon Brunette, Mitchell Sundt, Nicola Dell, Rohit Chaudhri, Nathan Breit, and Gaetano Borriello. Open data kit 2.0: expanding and refining information services for developing regions. In *Proceedings of the 14th Workshop on Mobile Computing Systems and Applications*, page 10. ACM, 2013.

[BUH+14] John D Beard, David M Umbach, Jane A Hoppin, Marie Richards, Michael CR Alavanja, Aaron Blair, Dale P Sandler, and Freya Kamel. Pesticide exposure and depression among male private pesticide applicators in the agricultural health study. *Environmental health perspectives*, 122(9):984–991, 2014.

[Bul82] David Bull. *A growing problem: pesticides and the Third World poor*. Oxfam, 1982.

[Bur11] United States Census Bureau. Shapefile of the USA, 2011. Accessed: 2011-11-28. URL: `http://www2.census.gov/geo/tiger/TIGER2009/tl_2009_us_state.zip`.

[But11] Danielle E Buttke. Toxicology, environmental health, and the One Health concept. *Journal of Medical Toxicology*, 7(4):329–332, 2011.

[BWB+18] Richard A Becker, Allan R Wilks, Ray Brownrigg, Thomas P Minka, and Alex Deckmyn. maps: Draw geographical maps. *R package version*, 3(0):2018, 2018.

[CAL11] Susana Carreira, Nélia Amado, and Filipa Lecoq. Mathematical Modelling of daily life in adult education: focusing on the notion of knowledge. In *Trends in Teaching and Learning of Mathematical Modelling*, pages 199–209. Springer, 2011.

[Can95] Georg Cantor. Beiträge zur Begründung der transfiniten Mengenlehre. *Mathematische Annalen*, 46(4):481–512, 1895.

[Can24] Ubuntu Canonical. Ubuntu webpage, 2024. Accessed: 2024-04-18. URL: `https://ubuntu.com/`.

[Car62] Rachel Carson. *Silent spring*. Houghton Mifflin Company, 1962.

[Car00] Andrée Carter. How pesticides get into water and proposed reduction measures. *Pesticide Outlook*, 11(4):149–156, 2000.

[CBB93] Bedri C Cetin, Joel W Burdick, and Jacob Barhen. Global descent replaces gradient descent to avoid local minima problem in learning with artificial

neural networks. In *IEEE International Conference on Neural Networks*, pages 836–842. IEEE, 1993.

[CBB+15] Olaf Conrad, Benjamin Bechtel, Michael Bock, Helge Dietrich, Elke Fischer, Lars Gerlitz, Jan Wehberg, Volker Wichmann, and Jürgen Böhner. System for automated geoscientific analyses (SAGA) v. 2.1. 4. *Geoscientific model development*, 8(7):1991–2007, 2015.

[CC11] Silvia Curteanu and Hugh Cartwright. Neural networks applied in chemistry. I. Determination of the optimal topology of multilayer perceptron neural networks. *Journal of Chemometrics*, 25(10):527–549, 2011.

[CC12] Myriah L Cornwell and Lisa M Campbell. Co-producing conservation and knowledge: Citizen-based sea turtle monitoring in North Carolina, USA. *Social Studies of Science*, 42(1):101–120, 2012.

[CCA+20] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. *shiny: Web Application Framework for R*, 2020. R package version 1.5.0.

[CCBC04] Jorge Casillas, Brian Carse, Larry Bull, and Brian Carse. Fuzzy XCS: an accuracy-based fuzzy classifier system. In *Proceedings of the XII Congreso Espanol sobre Tecnologia y Logica Fuzzy (ESTYLF 2004)*, pages 369–376. Citeseer, 2004.

[CCC11] Raffaele Casa, Andrea Cavalieri, and Benedetto Lo Cascio. Nitrogen fertilisation management in precision agriculture: a preliminary application example on maize. *Italian Journal of Agronomy*, 6(1):5, 2011.

[CCP+22] Marco Ciolfi, Francesca Chiocchini, Rocco Pace, Giuseppe Russo, and Marco Lauteri. Timescape: a novel spatiotemporal modeling tool. *Earth*, 3(1):259–286, 2022.

[CD16] Edwin Francis Cárdenas and Max Suell Dutra. An augmented reality application to assist teleoperation of underwater manipulators. *IEEE Latin America Transactions*, 14(2):863–869, 2016.

[CDPB07] Caren B Cooper, Janis Dickinson, Tina Phillips, and Rick Bonney. Citizen science as a tool for conservation in residential ecosystems. *Ecology & Society*, 12(2), 2007.

[CDT12] Maria Cerreta and Pasquale De Toro. Urbanization suitability maps: A dynamic spatial decision support system for sustainable land use. *Earth System Dynamics*, 3(2):157–171, 2012.

[Ceb19] Zeynel Cebeci. Comparison of Internal Validity Indices for Fuzzy Clustering. *Journal of Agricultural Informatics*, 10(2):1–14, 2019.

[Ced14] Nina Cedergreen. Quantifying synergy: a systematic review of mixture toxicity studies within environmental toxicology. *PloS one*, 9(5):e96580, 2014.

[Cen13] National Pesticide Information Center. Regulating Pesticides through Risk Assessment, 2013. Accessed: 2013-09-27. URL: `http://npic.orst.edu/reg/risk.html`.

[CF] Crop Protection Research Institute CropLife Foundation. National Pesticide Use Database 2002. Accessed: 2011-03-10. URL: `http://www.croplifefoundation.org/Documents/PUD/NPUD%202002/NPUD2002%20Complete%20Excel.zip`.

[CF10] Menzie D Chinn and Robert W Fairlie. Ict use in the developing world: an analysis of differences in computer and internet penetration. *Review of International Economics*, 18(1):153–167, 2010.

[CGR21] Chao Chen, Jon Garibaldi, and Tajul Razak. *FuzzyR: Fuzzy Logic Toolkit for R*, 2021. R package version 2.3.2.

[Cha88] Pat S Chavez. An improved dark-object subtraction technique for atmospheric scattering correction of multispectral data. *Remote sensing of environment*, 24(3):459–479, 1988.

[Chi94] Stephen L Chiu. Fuzzy model identification based on cluster estimation. *Journal of Intelligent & fuzzy systems*, 2(3):267–278, 1994.

[CHKW+17] Aurea C Chiaia-Hernandez, Armin Keller, Daniel Wächter, Christine Steinlin, Louise Camenzuli, Juliane Hollender, and Martin Krauss. Long-term persistence of pesticides and TPs in archived agricultural soil samples and comparison with pesticide application. *Environmental science & technology*, 51(18):10642–10651, 2017.

[Chu07] Howard Chu. OpenLDAP 2.4 Highlights Features of the Upcoming Release, 2007.

[CJH+13] Alycia W Crall, Rebecca Jordan, Kirstin Holfelder, Gregory J Newman, Jim Graham, and Donald M Waller. The impacts of an invasive species citizen science training program on participant attitudes, behavior, and science literacy. *Public Understanding of Science*, 22(6):745–764, 2013.

[CKR+20] Anna San Llorente Capdevila, Ainur Kokimova, Saunak Sinha Ray, Tamara Avellán, Jiwon Kim, and Sabrina Kirschke. Success factors for citizen science projects in water quality monitoring. *Science of the Total Environment*, 728:137843, 2020.

[CLW+07] Kevin Crowston, Qing Li, Kangning Wei, U Yeliz Eseryel, and James Howison. Self-organization of teams for free/libre open source software development. *Information and software technology*, 49(6):564–575, 2007.

[CMV+11] Véronique Chaplain, Laure Mamy, Laure Vieublé, Christian Mougin, Pierre Benoit, and Sylvie Nelieu. *Fate of pesticides in soils: Toward an integrated approach of influential factors*. InTech, 2011.

[CMW+14] Pedro Carriquiriborde, Paula Mirabella, Andrea Waichman, Keith Solomon, Paul J Van den Brink, and Steve Maund. Aquatic risk assessment of pesticides in Latin America. *Integrated environmental assessment and management*, 10(4):539–542, 2014.

[CN06] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

[Coh08] Jeffrey P Cohn. Citizen science: Can volunteers do real research? *BioScience*, 58(3):192–197, 2008.

[Com02] European Commission. Guidance document on aquatic ecotoxicology in the context of the Directive 91/414/EEC. *Sanco/3268/2001 rev*, 4:1–62, 2002.

[Com09] European Commission. Regulation (EC) No 1107/2009 of the European Parliament and of the Council of 21 October 2009 concerning the placing of plant protection products on the market and repealing Council Directives

79/117/EEC and 91/414/EEC. *Official Journal of the European Union*, 309:1–50, 2009.

[Com11] FOCUS Steering Committee. FOCUS Surface Water Scenarios in the EU Evaluation Process under 91/414/EEC. *Report of the FOCUS Working Group on Surface Water Scenarios, EC Document Reference SANCO/4802/2001*, page 245, 2011.

[Com18] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018.

[Com24] European Commision. Guidelines and supporting documents on Active Substances and Plant Protection Products, 2024. Accessed: 2024-12-13. URL: `https://webgate.ec.europa.eu/dyna2/pgd/`.

[Cor20] Mozilla Corporation. Deepspeech [speech-to-text engine], 2020. Accessed: 2020-07-06. URL: `https://github.com/mozilla/DeepSpeech/`.

[Cor21] Oracle Corporation. Mysql 8.0.23 [relational database management system], 2021. Accessed: 2021-10-16. URL: `https://www.mysql.com/de/`.

[Cou11] British Crop Protection Council. *The pesticide manual: a world compendium*. Tomlin, CDS, Ed, 2011.

[CR97] Toby N Carlson and David A Ripley. On the relation between NDVI, fractional vegetation cover, and leaf area index. *Remote sensing of Environment*, 62(3):241–252, 1997.

[CRM06] Neil A Campbell, Jane B Reece, and J Marl. Biologie. 8., aktualisierte auflage. *München (ua): Pearson Studium 2006*, 2006.

[CRMT11] William G Couser, Giuseppe Remuzzi, Shanthi Mendis, and Marcello Tonelli. The contribution of chronic kidney disease to the global burden of major noncommunicable diseases. *Kidney international*, 80(12):1258–1270, 2011.

[CRS06] Amitava Chatterjee, Anjan Rakshit, and Patrick Siarry. Generalised influential rule search scheme for fuzzy function approximation. *Soft Computing*, 10(8):631–642, 2006.

[CRWJ14] Ricardo Correa-Rotter, Catharina Wesseling, and Richard J Johnson. CKD of unknown origin in Central America: the case for a Mesoamerican nephropathy. *American Journal of Kidney Diseases*, 63(3):506–520, 2014.

[CSKX24] Joe Cheng, Barret Schloerke, Bhaskar Karambelkar, and Yihui Xie. *leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library*, 2024. R package version 2.2.2.

[CTS+04] Gloria D Coronado, Beti Thompson, Larki Strong, William C Griffith, and Ilda Islas. Agricultural task and exposure to organophosphate pesticides among farmworkers. *Environmental health perspectives*, 112(2):142–147, 2004.

[CWGK07] Alex Chong, Kok Wai Wong, Tamás D Gedeon, and László T Kóczy. Sparse Fuzzy System Generation By Cluster Estimation: A Projection Based Approach. *Australian Journal of Intelligent Information Processing Systems*, 8(2):82–93, 2007.

[CWP95] Martin D Crossland, Bayard E Wynne, and William C Perkins. Spatial decision support systems: An overview of technology and a test of efficacy. *Decision support systems*, 14(3):219–235, 1995.

[Dam09] Christos A Damalas. Understanding benefits and risks of pesticide use. *Scientific Research and Essays*, 4(10):945–949, 2009.

[DAR+02] Mustafa Dosemeci, Michael CR Alavanja, Andrew S Rowland, David Mage, Shelia Hoar Zahm, Nathaniel Rothman, Jay H Lubin, Jane A Hoppin, Dale P Sandler, and Aaron Blair. A quantitative approach for estimating exposure to pesticides in the Agricultural Health Study. *Annals of Occupational Hygiene*, 46(2):245–260, 2002.

[DBB+09] Finn Danielsen, Neil D Burgess, Andrew Balmford, Paul F Donald, Mikkel Funder, Julia PG Jones, Philip Alviola, Danilo S Balete, TOM Blomley, Justin Brashares, et al. Local participation in natural resource monitoring: a characterization of approaches. *Conservation Biology*, 23(1):31–42, 2009.

[DBM10] Benoît Dutilleul, Frans AJ Birrer, and Wouter Mensink. Unpacking european living labs: analysing innovation's social dimensions. *Central European journal of public policy*, 4(1):60–85, 2010.

[DCKBK24] Sunshine A De Caires, Ali Keshavarzi, Eduardo Leonel Bottega, and Fuat Kaya. Towards site-specific management of soil organic carbon: Comparing support vector machine and ordinary kriging approaches based on pedo-geomorphometric factors. *Computers and Electronics in Agriculture*, 216:108545, 2024.

[dCS20] Paulo Vitor de Campos Souza. Fuzzy neural networks and neuro-fuzzy networks: A review the main techniques and applications used in the literature. *Applied soft computing*, 92:106275, 2020.

[DDD07] Nicolas Desneux, Axel Decourtye, and Jean-Marie Delpuech. The sublethal effects of pesticides on beneficial arthropods. *Annual Review of Entomology*, 52(1):81–106, 2007.

[DDVG+19] Pasquale Daponte, Luca De Vito, Luigi Glielmo, Luigi Iannelli, Davide Liuzza, Francesco Picariello, and Giuseppe Silano. A review on the use of drones for precision agriculture. In *IOP conference series: earth and environmental science*, volume 275, page 012022. IOP Publishing, 2019.

[DE11] Christos A Damalas and Ilias G Eleftherohorinos. Pesticide exposure, safety issues, and risk assessment indicators. *International journal of environmental research and public health*, 8(5):1402–1419, 2011.

[Den00] David Dent. *Insect pest management*. CABI, 2000.

[Dev10] Indira Devi. Pesticides in agriculture–a boon or a curse? A case study of Kerala. *Economic and Political Weekly*, 45(26&27):199–207, 2010.

[DFH+15] Nicole C Deziel, Melissa C Friesen, Jane A Hoppin, Cynthia J Hines, Kent Thomas, and Laura E Beane Freeman. A review of nonoccupational pathways for pesticide exposure in women living in agricultural areas. *Environmental health perspectives*, 123(6):515–524, 2015.

[DFQD20] Clémentine Dereumeaux, Clémence Fillol, Philippe Quénel, and Sébastien Denys. Pesticide exposures for residents living close to agricultural lands: A review. *Environment international*, 134:105210, 2020.

[DHR13] Dimiter Driankov, Hans Hellendoorn, and Michael Reinfrank. *An introduction to fuzzy control*. Springer Science & Business Media, 2013.

[Dia19] Madson Luiz Dantas Dias. fuzzy-c-means: A simple python implementation of fuzzy c-means algorithm., May 2019. URL: `https://git.io/fuzzy-c-means`.

[Die96] Reinhard Diestel. Graphentheorie. *Springer-Verlag Heidelberg*, 2000:2006, 1996.

[Dij59] EW Dijkstra. A Note on Two Problems in Connexion with Graphs. *Numerische Mathematik*, 1:269–271, 1959.

[Din93] Barbara Dinham. *The pesticide hazard: a global health and environmental audit*. Zed Books, 1993.

[DJ06] Rodolphe Devillers and Robert Jeansoulin. *Fundamentals of spatial data quality*. ISTE Publishing Company, 2006.

[DLSA10] Leslie K Dennis, Charles F Lynch, Dale P Sandler, and Michael CR Alavanja. Pesticide use and cutaneous melanoma in pesticide applicators in the agricultural heath study. *Environmental health perspectives*, 118(6):812–817, 2010.

[DLW01] Susmita Dasgupta, Somik Lall, and David Wheeler. *Policy reform, economic growth, and the digital divide: An econometric analysis*, volume 2567. World Bank Publications, 2001.

[dNeHAdES10] Asociacion de NefrologÃa e Hipertension Arterial de El Salvador. Recomendaciones del Primer Taller de Salud Renal al Ministerio de Salud Publica y Asistencia Social de El Salvador, 2010.

[DRH11] Anhai Doan, Raghu Ramakrishnan, and Alon Y Halevy. Crowdsourcing systems on the world-wide web. *Communications of the ACM*, 54(4):86–96, 2011.

[Dro07] Sam Droege. Just because you paid them doesn't mean their data are better. In *Citizen Science Toolkit Conference. Cornell Laboratory of Ornithology*, pages 13–26, 2007.

[DRU+20] Muhammad Dawood, Atta-ur Rahman, Sami Ullah, Shakeel Mahmood, Ghani Rahman, and Kamran Azam. Spatio-statistical analysis of rainfall fluctuation, anomaly and trend in the Hindu Kush region using ARIMA approach. *Natural Hazards*, 101:449–464, 2020.

[DS06] Manojit Dam and Deoki N Saraf. Design of neural networks using genetic algorithm for on-line property estimation of crude fractionator products. *Computers & chemical engineering*, 30(4):722–729, 2006.

[DSB+12] Janis L Dickinson, Jennifer Shirk, David Bonter, Rick Bonney, Rhiannon L Crain, Jason Martin, Tina Phillips, and Karen Purcell. The current state of citizen science as a tool for ecological research and public engagement. *Frontiers in Ecology and the Environment*, 10(6):291–297, 2012.

[dSdLdS+21] Cecilia Cordeiro da Silva, Clarisse Lins de Lima, Ana Clara Gomes da Silva, Eduardo Luiz Silva, Gabriel Souza Marques, Lucas Job Brito de Araújo, Luiz Antonio Albuquerque Junior, Samuel Barbosa Jatobá de Souza, Maíra Araújo de Santana, Juliana Carneiro Gomes, et al. Covid-19 dynamic monitoring and real-time spatio-temporal forecasting. *Frontiers in public health*, 9:641253, 2021.

[DT17] Esthi Kurnia Dewi and Bambang Trisakti. Comparing atmospheric correction methods for Landsat OLI data. *International Journal of Remote Sensing and Earth Sciences (IJReSES)*, 13(2):105–120, 2017.

[DTT08] Christos A Damalas, Georgios K Telidis, and Stavros D Thanos. Assessing farmers' practices on disposal of pesticide waste after use. *Science of the total environment*, 390(2-3):341–345, 2008.

[DWB10] Vincent Devictor, Robert J Whittaker, and Coralie Beltrame. Beyond scarcity: citizen science programmes as useful tools for conservation biogeography. *Diversity and Distributions*, 16(3):354–362, 2010.

[DWZ+18] Zhenhong Du, Sensen Wu, Feng Zhang, Renyi Liu, and Yan Zhou. Extending geographically and temporally weighted regression to account for both spatiotemporal heterogeneity and seasonal variations in coastal seas. *Ecological Informatics*, 43:185–199, 2018.

[EBHW20] John W. Eaton, David Bateman, Søren Hauberg, and Rik Wehbring. *GNU Octave version 5.2.0 manual: a high-level interactive language for numerical computations*, 2020. URL: `https://www.gnu.org/software/octave/doc/v5.2.0/`.

[Eco01] Donald J. Ecobichon. Pesticide use in developing countries. *Toxicology*, 160(13):27–33, March 2001.

[EDDSM15] Otilia Elena Dragomir, Florin Dragomir, Veronica Stefan, and Eugenia Minca. Adaptive neuro-fuzzy inference systems as a strategy for predicting and controling the energy produced from renewable sources. *Energies*, 8(11):13047–13061, 2015.

[(EF14] European Food Safety Authority (EFSA). Guidance on the assessment of exposure of operators, workers, residents and bystanders in risk assessment for plant protection products. *EFSA Journal*, 12(10):3874, 2014.

[EGK08] Christof Eck, Harald Garcke, and Peter Knabner. *Mathematische Modellierung*, volume 2. Springer, 2008.

[ENB05] A Meguid El Nahas and Aminu K Bello. Chronic kidney disease: the global challenge. *The Lancet*, 365(9456):331–340, 2005.

[ENK05] Mats Eriksson, Veli-Pekka Niitamo, and Seija Kulkki. State-of-the-art in utilizing Living Labs approach to user-centric ICT innovation-a European approach. *Lulea: Center for Distance-spanning Technology. Lulea University of Technology Sweden: Lulea. Online under: http://www. cdt. ltu. se/main. php/SOA_ LivingLabs. pdf*, 2005.

[ES11] Stefan Edelkamp and Stefan Schrödl. *Heuristic search: theory and applications*. Elsevier, 2011.

[ET11] Philip D Evans and Maarten W Taal. Epidemiology and causes of chronic kidney disease. *Medicine*, 39(7):402–406, 2011.

[Exp15] United States Geological Survey (USGS) Earth Explorer. Landsat archive dataset: Landsat 8 OLI/TIRS, 2015. Accessed: 2015-10-12. URL: `http://earthexplorer.usgs.gov/`.

[fASC14] Council for Agricultural Science and Technology (CAST). The Contributions of Pesticides to Pest Management in Meeting the Global Need for Food Production by 2050. *Issue Paper*, 5:1–28, 2014.

[Fay21] Colin Fay. *geoloc: Geolocation in Shiny*, 2021. R package version 0.0.0.9500. URL: `https://github.com/ColinFay/geoloc`.

[FB15] Camilo Lesmes Fabian and Claudia R Binder. Dermal exposure assessment to pesticides in farming systems in developing countries: comparison of models. *International journal of environmental research and public health*, 12(5):4670–4696, 2015.

[FBH11] G Frost, T Brown, and A-H Harding. Mortality and cancer incidence among British agricultural pesticide users. *Occupational Medicine*, page kqr067, 2011.

[FCY15] A Stewart Fotheringham, Ricardo Crespo, and Jing Yao. Geographical and temporal weighted regression (GTWR). *Geographical Analysis*, 47(4):431–452, 2015.

[FD11a] Food and Agriculture Organization Corporate Statistical Database. Crop Area Database, 2011. Accessed: 2011-07-28. URL: `http://faostat.fao.org/site/567/default.aspx#ancor`.

[FD11b] Food and Agriculture Organization Corporate Statistical Database. Database production: crops, 2011. Accessed: 2011-03-27. URL: `http://faostat.fao.org/site/567/default.aspx#ancor`.

[FD11c] Food and Agriculture Organization Corporate Statistical Database. Database trade: crops and livestock products, 2011. Accessed: 2011-03-27. URL: `http://faostat.fao.org/site/535/default.aspx#ancor`.

[FD13] Food and Agriculture Organization Corporate Statistical Database. Database on Pesticides Use, 2013. Accessed: 2013-08-11. URL: `http://faostat.fao.org/site/424/default.aspx#ancor`.

[FD16] Food and Agriculture Organization Corporate Statistical Database. Database production: crops, 2016. Accessed: 2016-12-01. URL: `http://www.fao.org/faostat/en/#data/QC`.

[FD24a] Food and Agriculture Organization Corporate Statistical Database. Database on Pesticides Use, 2024. Accessed: 2024-11-26. URL: `https://www.fao.org/faostat/en/#data/RP`.

[FD24b] Food and Agriculture Organization Corporate Statistical Database. Database: Pesticide Use - Agricultural use, 2024. Accessed: 2024-08-06. URL: `https://www.fao.org/faostat/en/#data/RP`.

[fDCP18] Centers for Disease Control and Prevention. Chronic Kidney Disease Surveillance System, 2018. Accessed: 2018-03-12. URL: `https://nccd.cdc.gov/ckd/default.aspx`.

[fDP18] Social Council. Committee for Development Policy. *Handbook on the Least Developed Country Category: Inclusion, Graduation, and Special Support Measures*, volume 3. United Nations Publications, 2018.

[fECG19] Federal Ministry for Economic Cooperation and Development Germany. Encyclopedia entry: developing country, 2019. Accessed: 2019-09-25. URL: `https://www.bmz.de/de/service/glossar/E/entwicklungsland.html`.

[fECoO97] Organisation for Economic Co-operation and Development (OECD). *Agriculture, Pesticides and the Environment - Policy Options*. OECD Publications, 1997.

[fECoO19] Organisation for Economic Co-operation and Development (OECD). DAC List of Official Development Assistance Recipients, 2019. Accessed: 2019-10-02. URL: `http://www.oecd.org/dac/financing-sustainable-development/development-finance-standards/daclist.htm`.

[fECoO24] Organisation for Economic Co-operation and Development (OECD). Guidelines for the Testing of Chemicals, 2024. Accessed: 2024-12-13. URL: `https://www.oecd.org/en/topics/sub-issues/testing-of-chemicals/test-guidelines.html`.

[Fen13] Karl Fent. *Ökotoxikologie: Umweltchemie-Toxikologie-Ökologie*. Georg Thieme Verlag, 2013.

[FGW19] Stefan Fritsch, Frauke Guenther, and Marvin N. Wright. *neuralnet: Training of Neural Networks*, 2019. R package version 1.44.2.

[FHLL05] R Fell, KKS Ho, S Lacasse, and E Leroi. State of the Art Paper 1-A framework for landslide risk assessment and management. In *Proceedings of the International Conference on Landslide Risk Management, Vancouver, Canada*, volume 31, 2005.

[FHT15] Ludwig Fahrmeir, Alfred Hamerle, and Gerhard Tutz. *Multivariate statistische Verfahren*. Walter de Gruyter GmbH & Co KG, 2015.

[fIDU94] United States Agency for International Development (USAID). Pesticides and the Agrichemical Industry in Sub-Saharan Africa. Technical report, U.S. Agency for International Development, 1994. Project No. 698-0510.

[Fit02] Michael S Fitzner. Three decades of federal integrated pest management policy. In *Pesticides in Agriculture and the Environment*. CRC Press, 2002.

[Flo62] Robert W Floyd. Algorithm 97: Shortest path. *Communications of the ACM*, 5(6):345–345, 1962.

[FMI+21] Amrul Faruq, Aminaton Marto, Nadia Karima Izzaty, Abidemi Tolulope Kuye, Shamsul Faisal Mohd Hussein, and Shahrum Shah Abdullah. Flood disaster and early warning: application of ANFIS for river water level forecasting. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, 6(1):1–10, 2021.

[FOC00] FOCUS. FOCUS groundwater scenarios in the EU review of active substances. Technical report, Report of the FOCUS Groundwater Scenarios Workgroup, EC Document Reference Sanco/321/2000 rev.2, 202pp, 2000.

[FOC21] FOCUS DG SANTE. Overview of focus dg sante, 2021. Accessed: 2021-06-01. URL: `https://esdac.jrc.ec.europa.eu/projects/focus-dg-sante`.

[Fod91] Janos C Fodor. On fuzzy implication operators. *Fuzzy sets and systems*, 42(3):293–300, 1991.

[Fol87] Th A Foley. Interpolation and approximation of 3-D and 4-D scattered data. *Computers & Mathematics with Applications*, 13(8):711–740, 1987.

[FotUN19] Food and Agriculture Organization of the United Nations. FAO and family farming: The case of El Salvador , 2019. Accessed: 2019-09-25. URL: `http://www.fao.org/3/a-as175e.pdf`.

[Fou02] National Kidney Foundation. *Clinical practice guidelines for chronic kidney disease: evaluation, classification and stratification.* National Kidney Foundation, 2002.

[Fou20a] Apache Software Foundation. Apache HTTP Server, 2020. Accessed: 2020-01-11. URL: `https://httpd.apache.org/`.

[Fou20b] Apache Software Foundation. Apache OpenOffice 4.1.9, 2020. Accessed: 2020-01-11. URL: `https://www.openoffice.org/`.

[Fou20c] Mozilla Foundation. Mozilla Firefox, 2020. Accessed: 2020-07-06. URL: `https://www.mozilla.org/en-US/firefox/`.

[Fou20d] Mozilla Foundation. Mozilla Thunderbird, 2020. Accessed: 2020-07-06. URL: `https://www.thunderbird.net`.

[Fou20e] The Document Foundation. LibreOffice 7.1.0, 2020. Accessed: 2020-01-11. URL: `https://www.libreoffice.org/`.

[Fou20f] Ubuntu Foundation. Lubuntu 20.04 Focal Fossa, 2020. Accessed: 2020-01-11. URL: `https://lubuntu.me/`.

[Fou20g] Wikimedia Foundation. Mediawiki [wiki engine], 2020. Accessed: 2020-01-11. URL: `https://www.mediawiki.org`.

[Fra11] Universität Frankfurt. Cell-Area Grid, Resolution 5 min X 5 min., 2011. Accessed: 2011-10-03. URL: `ftp://ftp.rz.uni-frankfurt.de/pub/uni-frankfurt/physische_geographie/hydrologie/public/data/MIRCA2000/cell_area_grid/cell_area_ha_05mn.asc.gz`.

[FRB+93] RD Fox, DL Reichard, RD Brazee, CR Krause, and FR Hall. Downwind residues from spraying a semi-dwarf apple orchard. *Transactions of the ASAE*, 36, 1993.

[FRB11] Giuseppe Feola, E Rahn, and CR Binder. Suitability of pesticide risk indicators for less developed countries: a comparison. *Agriculture, ecosystems & environment*, 142(3):238–245, 2011.

[FUL+10a] Allan S Felsot, John B Unsworth, Jan BHJ Linders, Graham Roberts, Dirk Rautman, Caroline Harris, and Elizabeth Carazo. Agrochemical spray drift; assessment and mitigation-a review. *Journal of Environmental Science and Health Part B*, 46(1):1–23, 2010.

[Ful10b] Sherrilynne Fuller. Tracking the global express: New tools addressing disease threats across the world. *Epidemiology*, 21(6):769–771, 2010.

[FW05] Claire Franklin and John Worgan. *Occupational and residential exposure assessment for pesticides*, volume 9. John Wiley & Sons, 2005.

[FZCM15] Laura Folguera, Jure Zupan, Daniel Cicerone, and Jorge F Magallanes. Self-organizing maps for imputation of missing data in incomplete data matrices. *Chemometrics and Intelligent Laboratory Systems*, 143:146–151, 2015.

[GÏ14] Sarah Günther. An approach to deriving Minimum Data Elements related to Chronic Kidney Disease caused by non-traditional risk factors. Master's thesis, University of Koblenz-Landau, 2014.

[GAD+22] Guillermo A García, Brent Atkinson, Olivier Tresor Donfack, Emily R Hilton, Jordan M Smith, Jeremías Nzamío Mba Eyono, Marcos Mbulito Iyanga, Liberato Motobe Vaz, Restituto Mba Nguema Avue, John Pollock, et al. Real-time, spatial decision support to optimize malaria vector control: The case of indoor residual spraying on Bioko Island, Equatorial Guinea. *PLOS Digital Health*, 1(5):e0000025, 2022.

[Gal14] Galaxyzoo. Galaxyzoo homepage, 2014. Accessed: 2014-04-01. URL: `http://www.galaxyzoo.org/`.

[Gas12] Oliver Gassmann. *Crowdsourcing-Innovationsmanagement mit Schwarmintelligenz:-Interaktiv Ideen finden-Kollektives Wissen effektiv nutzen-Mit Fallbeispielen und Checklisten.* Carl Hanser Verlag GmbH Co KG, 2012.

[GAT+11] Robert D Grisso, Marcus M Alley, Wade E Thomason, David L Holshouser, and Gary T Roberson. *Precision farming tools: variable-rate application.* Virginia Cooperative Extension, 2011.

[GB07] Enzo Grossi and Massimo Buscema. Introduction to artificial neural networks. *European journal of gastroenterology & hepatology*, 19(12):1046–1054, 2007.

[GBB+10] Flavia Geiger, Jan Bengtsson, Frank Berendse, Wolfgang W Weisser, Mark Emmerson, Manuel B Morales, Piotr Ceryngier, Jaan Liira, Teja Tscharntke, Camilla Winqvist, et al. Persistent negative effects of pesticides on biodiversity and biological control potential on European farmland. *Basic and Applied Ecology*, 11(2):97–105, 2010.

[GBD00] JM Gonzalez-Barahona and C Daddara. Free Software / Open Source: Information Society Opportunities for Europe? *Working group on Libre Software, http://eu. conecta. it/paper/cathedral_ bazaar. html*, 2000.

[GD13] Soumi Ghosh and Sanjay Kumar Dubey. Comparative analysis of k-means and fuzzy c-means algorithms. *International Journal of Advanced Computer Science and Applications*, 4(4), 2013.

[GDKW11] Arthur Grube, David Donaldson, Timothy Kiely, and La Wu. Pesticides industry sales and usage. *Washington, DC: Office of Prevention, Pesticides and Toxic Substances, United States Environment Protection Agency*, 2011.

[GE11] Boris Girnat and Andreas Eichler. Secondary teachers? beliefs on modelling in geometry and stochastics. In *Trends in teaching and learning of mathematical modelling*, pages 75–84. Springer, 2011.

[Gee11] Guido L Geerts. A design science research methodology and its application to accounting information systems research. *International Journal of Accounting Information Systems*, 12(2):142–151, 2011.

[Gei11] Vince Geiger. Factors Affecting Teachers' Adoption of Innovative Practices with Technology and Mathematical Modelling. In *Trends in teaching and learning of mathematical modelling*, pages 305–314. Springer, 2011.

[Geo02] Georg Westermann Verlag, editor. *Diercke Weltatlas*. Westermann, Braunschweig, 2002. ISBN 978-3-14-100700-8.

[Ger09] K Gerber. User's Guide and Technical Documentation KABAM Version 1.0 (K OW (based) Aquatic Bioaccumulation Model). *Report of Environmental Fate and Effects Division. Office of Pesticide Programs, United States Environmental Protection Agency, Washington, DC*, 2009.

[GHS10] Robyn C Gilden, Katie Huffling, and Barbara Sattler. Pesticides and health risks. *Journal of Obstetric, Gynecologic & Neonatal Nursing*, 39(1):103–110, 2010.

[GHS11] Eric A Graham, Sandra Henderson, and Annette Schloss. Using mobile phones to engage citizen scientists in research. *EOS, Transactions American Geophysical Union*, 92(38):313–315, 2011.

[Gia13] Leonard P Gianessi. The increasing importance of herbicides in worldwide crop production. *Pest management science*, 69(10):1099–1105, 2013.

[Git21] Github. Atom texteditor v1.54.0, 2021. Accessed: 2021-12-08. URL: `https://atom.io/`.

[GJRS24] Luisa Gensch, Kerstin Jantke, Livia Rasche, and Uwe A Schneider. Pesticide risk assessment in european agriculture: Distribution patterns, ban-substitution effects and regulatory implications. *Environmental Pollution*, 348:123836, 2024.

[GLH+15] Salvador García, Julián Luengo, Francisco Herrera, Salvador García, Julián Luengo, and Francisco Herrera. Dealing with noisy data. *Data preprocessing in data mining*, pages 107–145, 2015.

[GLR99] J Gomes, OL Lloyd, and DM Revitt. The influence of personal protection, environmental hygiene and exposure to pesticides on the health of immigrant farm workers in a desert country. *International archives of occupational and environmental health*, 72(1):40–45, 1999.

[GLSGFV10] Pedro J García-Laencina, José-Luis Sancho-Gómez, and Aníbal R Figueiras-Vidal. Pattern classification with missing data: a review. *Neural Computing and Applications*, 19(2):263–282, 2010.

[GML15] Elisabetta Giusti and Stefano Marsili-Libelli. A fuzzy decision support system for irrigation and water conservation in agriculture. *Environmental Modelling & Software*, 63:73–86, 2015.

[GMNZ14] Mariusz Głabowski, Bartosz Musznicki, Przemysław Nowak, and Piotr Zwierzykowski. Review and performance analysis of shortest path problem solving algorithms. *International Journal On Advances in Software*, 7:20–30, 2014.

[GN16] Ulrike Grote and Frank Neubacher. Rural crime in developing countries: theoretical framework, empirical findings, research needs. *ZEF working paper series*, 2016.

[Goo97] Pierre Goovaerts. *Geostatistics for natural resources evaluation.* Oxford University Press, USA, 1997.

[Gra95] Adolf Grauel. *Fuzzy-Logik: Einführung in die Grundlagen mit Anwendungen.* BI Wissenschaftsverlag, 1995.

[Gre80] R Greenhalgh. Definition of persistence in pesticide chemistry. *Pure and Applied Chemistry*, 52:2565–2566, 1980.

[GRLRM+13] Luis Gómez-Robledo, Nuria López-Ruiz, Manuel Melgosa, Alberto J Palma, Luis Fermín Capitán-Vallvey, and Manuel Sánchez-Marañón. Using the mobile phone as Munsell soil-colour sensor: An experiment under controlled illumination conditions. *Computers and electronics in agriculture*, 99:200–208, 2013.

[GRMI21] Mir Reza Ghaffari Razin, Amir Reza Moradi, and Samed Inyurt. Spatio-temporal analysis of TEC during solar activity periods using support vector machine. *GPS Solutions*, 25(3):121, 2021.

[Gro09] Kidney Disease: Improving Global Outcomes (KDIGO) CKD-MBD Work Group. KDIGO clinical practice guideline for the diagnosis, evaluation, prevention, and treatment of Chronic Kidney Disease-Mineral and Bone Disorder (CKD-MBD). *Kidney international. Supplement*, 1(113):S1, 2009.

[GS05] Y Gil and C Sinfort. Emission of pesticides to the air during sprayer application: A bibliographic review. *Atmospheric Environment*, 39(28):5183–5193, 2005.

[GSMS11] B Ganesh, S Sushama, S Monika, and P Suvarna. A case-control study of risk factors for lung cancer in Mumbai, India. *Asian Pacific Journal of Cancer Prevention*, 12(2):357–362, 2011.

[GSS+17] Rishila Ghosh, Manushi Siddarth, Neeru Singh, Vipin Tyagi, Pawan Kumar Kare, Basu Dev Banerjee, Om Prakash Kalra, and Ashok Kumar Tripathi. Organochlorine pesticide level in patients with chronic kidney disease of unknown etiology and its association with renal function. *Environmental Health and Preventive Medicine*, 22(1):49, 2017.

[GT20] Terry Wayne Griffin and LaVona Traywick. The role of variable rate technology in fertilizer usage. *Journal of Applied Farm Economics*, 3(2):6, 2020.

[GTT+12] Sibukele Gumbo, Hannah Thinyane, Mamello Thinyane, Alfredo Terzoli, and Susan Hansen. Living Lab Methodology as an Approach to Innovation in ICT4D: The Siyakhula Living Lab Experience. In *IST-Africa 2012 Conference Proceedings*, Dar es Salaam, Tanzania, 2012.

[GV14] Marco Günther and Kai Velten. *Mathematische Modellbildung und Simulation: Eine Einführung für Wissenschaftler, Ingenieure und Ökonomen*. John Wiley & Sons, 2014.

[GWJ+92] P Grandjean, P Weihe, PJ Jørgensen, T Clarkson, E Cernichiari, and T Viderø. Impact of maternal seafood diet on fetal exposure to mercury, selenium, and lead. *Archives of Environmental Health: An International Journal*, 47(3):185–195, 1992.

[GZ14] Ane Garay Zarraga. La minería transnacional en centroamérica: lógicas regionales e impactos transfronterizos. el caso de la mina cerro blanco, 2014. Madrid: Paz con Dignidad.

[HA08] A Hussain and MR Asi. Pesticides as water pollutants. In *Groundwater for Sustainable Development*, pages 119–126. CRC Press, 2008.

[Haa10] HES Haapala. Living lab as usability development platform for agricultural technologies. In *International Conference on Agricultural Engineering-AgEng 2010: towards environmental technologies, Clermont-Ferrand, France, 6-8 September 2010*. Cemagref, 2010.

[Hay04] Simon Haykin. *Neural Networls: A comprehensive foundation*. Prentice Hall, 2004.

[HB92] Gary J Hunter and Kate Beard. Understanding error in spatial databases. *Australian surveyor*, 37(2):108–119, 1992.

[HB09] Martin Hanke-Bourgeois. *Grundlagen der numerischen Mathematik und des wissenschaftlichen Rechnens*. Springer, 2009.

[HBW21] Scott LJ Hepditch, Oana Birceanu, and Michael P Wilkie. A Toxic Unit and Additive Index Approach to Understanding the Interactions of 2 Piscicides, 3-Trifluoromethyl-4-Nitrophenol and Niclosamide, in Rainbow Trout. *Environmental Toxicology and Chemistry*, 40(5):1419–1430, 2021.

[HC00] Dug Hun Hong and Chang-Hwan Choi. Multicriteria fuzzy decision-making problems based on vague set theory. *Fuzzy sets and systems*, 114(1):103–113, 2000.

[HCB12] William H Hallenbeck and Kathleen M Cunningham-Burns. *Pesticides and human health*. Springer Science & Business Media, 2012.

[Heb49] Donald Olding Hebb. *The Organizations of Behavior: a Neuropsychological Theory*. John Wiley & Sons, 1949.

[HED+05] Matthew J Helmers, Dean E Eisenhauer, Michael G Dosskey, Thomas G Franti, Jason M Brothers, and Mary Carla McCullough. Flow pathways and sediment trapping in a field-scale vegetative filter. *Transactions of the ASAE*, 48(3):955–968, 2005.

[Her11] Marlien Herselman. Living Labs in Southern Africa network. PowerPoint presentation at 3rd Annual LLiSA Workshop, 2011.

[Her17] Natalia Romero Herrera. The emergence of Living Lab methods. In *Living Labs*, pages 9–22. Springer, 2017.

[HF83] H Hugli and W Frei. Understanding anisotropic reflectance in mountainous terrain. *Photogrammetric Engineering and Remote Sensing*, 49:671–683, 1983.

[HGL17] Antonio F Hernández, Fernando Gil, and Marina Lacasaña. Toxicological interactions of pesticide mixtures: an update. *Archives of toxicology*, 91:3211–3223, 2017.

[Hij23] Robert J. Hijmans. *raster: Geographic Data Analysis and Modeling*, 2023. R package version 3.6-26.

[Hij24] Robert J. Hijmans. *terra: Spatial Data Analysis*, 2024. R package version 1.7-78.

[Hil87] Dennis S Hill. *Agricultural insect pests of temperate regions and their control*. CUP Archive, 1987.

[HJHJ19] Frank E Harrell Jr and Maintainer Frank E Harrell Jr. Package hmisc. *CRAN2018*, 2019:235–236, 2019.

[HK+08] H Haapala, S Kankaanpää, et al. Agro Living Lab Seinäjoki. In *Agricultural and biosystems engineering for a sustainable world. International Conference on Agricultural Engineering, Hersonissos, Crete, Greece, 23-25 June, 2008*. European Society of Agricultural Engineers (AgEng), 2008.

[HK17] Marjan Hosseini and Reza Kerachian. A Bayesian maximum entropy-based methodology for optimal spatiotemporal design of groundwater monitoring networks. *Environmental monitoring and assessment*, 189:1–24, 2017.

[HK18] Junyu He and Alexander Kolovos. Bayesian maximum entropy approach and its applications: a review. *Stochastic Environmental Research and Risk Assessment*, 32:859–877, 2018.

[HKB11] James Honaker, Gary King, and Matthew Blackwell. Amelia II: A program for missing data. *Journal of Statistical Software*, 45:1–47, 2011.

[HKE04] Zelee Hill, Betty Kirkwood, and Karen Edmond. Family and community practices that promote child survival, growth and development. *Geneva: World Health Organization*, 2004.

[HL04] Ernest Hodgson and Patricia E Levi. *A textbook of modern toxicology*. Wiley Online Library, 2004.

[HLZ+20] Zhonghua He, Liping Lei, Yuhui Zhang, Mengya Sheng, Changjiang Wu, Liang Li, Zhao-Cheng Zeng, and Lisa R Welp. Spatio-temporal mapping of multi-satellite observed column atmospheric co2 using precision-weighted kriging method. *Remote Sensing*, 12(3):576, 2020.

[HM08] Mauro Hernández and Antoni Margalida. Pesticide abuse in Europe: effects on the Cinereous vulture (Aegypius monachus) population in Spain. *Ecotoxicology*, 17(4):264–272, 2008.

[HM11] Jessica Harris and Andrew McCartor. The World's Worst Toxic Pollution Problems. Report 2011. The Top Ten of the Toxic Twenty. Technical report, Blacksmith Institute, 2011.

[HMS24] Timothy R Hannigan, Ian P McCarthy, and André Spicer. Beware of botshit: How to manage the epistemic risks of generative chatbots. *Business Horizons*, 67(5):471–486, 2024.

[hom20] Lubuntu homepage. About lubuntu, 2020. Accessed: 2020-11-13. URL: `https://lubuntu.net/about/`.

[Hop82] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

[Hop84] John J Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the national academy of sciences*, 81(10):3088–3092, 1984.

[How06] Jeff Howe. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4, 2006.

[HS01] Vern Hofman and Elton Solseng. Reducing spray drift, 2001. North Dakota State University.

[HSBK10] Marita Holst, Anna Ståhlbröst, and Birgitta Bergvall-Kåreborn. Openness in Living Labs–Facilitating Innovation. In *IRIS33–Information Systems Research Conference in Scandinavia, at Aalborg, Denmark*, volume 556, 2010.

[HT18] Christopher J Hillar and Ngoc M Tran. Robust Exponential Memory in Hopfield Networks. *The Journal of Mathematical Neuroscience*, 8(1):1–20, 2018.

[Hue88] Alfredo R Huete. A soil-adjusted vegetation index (SAVI). *Remote sensing of environment*, 25(3):295–309, 1988.

[Hui22] Jennifer Yang Hui. *Crowdsourcing for national security*. JSTOR, 2022.

[IA11] Hesam Izakian and Ajith Abraham. Fuzzy C-means and fuzzy swarm for fuzzy clustering problem. *Expert Systems with Applications*, 38(3):1835–1838, 2011.

[IBGA06] Richard R Irish, John L Barker, Samuel N Goward, and Terry Arvidson. Characterization of the Landsat-7 ETM+ automated cloud-cover assessment (ACCA) algorithm. *Photogrammetric Engineering & Remote Sensing*, 72(10):1179–1188, 2006.

[iES20] LLinES (Living Lab in El Salvador). LLinES: The project, 2020. Accessed: 2020-08-04. URL: `http://llines.weebly.com/the-project.html`.

[IGE+14] Kristin K Isaacs, W Graham Glen, Peter Egeghy, Michael-Rock Goldsmith, Luther Smith, Daniel Vallero, Raina Brooks, Christopher M Grulke, and Haluk Ozkaynak. SHEDS-HT: an integrated probabilistic exposure model for prioritizing exposures to chemicals with near-field and dietary sources. *Environmental science & technology*, 48(21):12750–12759, 2014.

[IK13] Dieter Imboden and Sabine Koch. *Systemanalyse: Einführung in die mathematische Modellierung natürlicher Systeme*. Springer-Verlag, 2013.

[Ill02] H Paul A Illing. *Toxicity and risk: context, principles and practice*. CRC Press, 2002.

[IMKA20] Leila Ismail, Huned Materwala, Achim P Karduck, and Abdu Adem. Requirements of health data management systems for biomedical care and research: scoping review. *Journal of medical Internet research*, 22(7):e17508, 2020.

[Ini13] Open Source Initiative. About the open source initiative, 2013. Accessed: 2013-09-03. URL: `http://opensource.org/about`.

[Ini14] Open Source Initiative. The open source definition, 2014. Accessed: 2014-03-30. URL: `http://opensource.org/osd`.

[Ini15] Action Team 6 Follow Up Initiative. Definition Open Community, 2015. Accessed: 2015-01-15. URL: `http://at6fui.weebly.com/open-community-approach.html`.

[Int93] International Programme on Chemical Safety (IPCS), World Health Organization (WHO). Environmental health criteria 155, biomarkers and risk assessment: Concept and principles. *World Health Organization, Geneva*, 1993.

[IR99] Mahesh S Iyer and R Russell Rhinehart. A method to determine the required number of neural-network training repetitions. *IEEE Transactions on Neural Networks*, 10(2):427–432, 1999.

[J+90] J Jeyaratnam et al. Acute pesticide poisoning: a major global health problem. *World Health Stat Q*, 43(3):139–144, 1990.

[JACH07] R Juraske, A Antón, F Castells, and MAJ Huijbregts. Human intake fractions of pesticides via greenhouse tomato consumption: Comparing model estimates with measurements for Captan. *Chemosphere*, 67(6):1102–1107, 2007.

[JAE+12] Nadira Jasika, Naida Alispahic, Arslanagic Elma, Kurtovic Ilvana, Lagumdzija Elma, and Novica Nosovic. Dijkstra's shortest path algorithm serial and parallel execution performance analysis. In *2012 proceedings of the 35th international convention MIPRO*, pages 1811–1815. IEEE, 2012.

[Jan93] J-SR Jang. Anfis: adaptive-network-based fuzzy inference system. *IEEE transactions on systems, man, and cybernetics*, 23(3):665–685, 1993.

[Jay14] Saroj Jayasinghe. Chronic kidney disease of unknown etiology should be renamed chronic agrochemical nephropathy. *Medicc Review*, 16(2):72–74, 2014.

[JBF+06] Marion Junghans, Thomas Backhaus, Michael Faust, Martin Scholze, and LH Grimme. Application and validation of approaches for the predictive hazard assessment of realistic pesticide mixtures. *Aquatic toxicology*, 76(2):93–110, 2006.

[JE00] Chuanyi Ji and Anwar Elwalid. Measurement-based network monitoring: missing data formulation and scalability analysis. In *Information Theory, 2000. Proceedings. IEEE International Symposium on*, page 78. IEEE, 2000.

[JGS14] Channa Jayasumana, Sarath Gunatilake, and Priyantha Senanayake. Glyphosate, hard water and nephrotoxic metals: are they the culprits behind the epidemic of chronic kidney disease of unknown etiology in Sri Lanka? *International journal of environmental research and public health*, 11(2):2125–2147, 2014.

[JL98] Chia-Feng Juang and Chin-Teng Lin. An online self-constructing neural fuzzy inference network and its applications. *IEEE transactions on Fuzzy Systems*, 6(1):12–32, 1998.

[JLC+04] Cynda Ann Johnson, Andrew S Levey, Josef Coresh, Adeera Levin, Joseph Lau, and Garabed Eknoyan. Clinical practice guidelines for chronic kidney disease in adults: Part I. Definition, disease stages, evaluation, treatment, and risk factors. *American family physician*, 70(5):869–876, 2004.

[JLH+14] Erik Jørs, Flemming Lander, Omar Huici, Rafael Cervantes Morant, Gabriel Gulis, and Flemming Konradsen. Do Bolivian small holder farmers improve and retain knowledge to reduce occupational pesticide poisonings after training on Integrated Pest Management? *Environmental health*, 13(1):1–9, 2014.

[JPA+15] Channa Jayasumana, Priyani Paranagama, Suneth Agampodi, Chinthaka Wijewardane, Sarath Gunatilake, and Sisira Siribaddana. Drinking well water and occupational exposure to Herbicides is associated with chronic kidney disease, in Padavi-Sripura, Sri Lanka. *Environmental Health*, 14(1):6, 2015.

[JPR19] Anil Jadhav, Dhanya Pramod, and Krishnan Ramanathan. Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, 33(10):913–933, 2019.

[JS10] Yves Julien and José A Sobrino. Comparison of cloud-reconstruction methods for time series of composite NDVI data. *Remote Sensing of Environment*, 114(3):618–625, 2010.

[JSD22] Devender Jain, Shiv Kumar Sharma, and Pooja Dhiman. Comparative Analysis of Defuzzification Techniques for Fuzzy Output. *JOURNAL OF ALGEBRAIC STATISTICS*, 13(2):874–882, 2022.

[JŠZB15] Mladen Jurišić, Luka Šumanovac, Domagoj Zimmer, and Željko Barač. Technical and technological aspects in plant protection in the precision farming system. *POLJOPRIVREDA*, 21(1):75–81, 2015.

[JT14] Bernt Johansen and Hans Tømmervik. The relationship between phytomass, NDVI and vegetation communities on Svalbard. *International Journal of Applied Earth Observation and Geoinformation*, 27:20–30, 2014.

[KA15] Dmitry Kangin and Plamen Angelov. Evolving clustering, classification and regression with TEDA. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2015.

[KAA+02] Mehmet Kaya, Reda Alhajj, Ahmet Arslan, et al. Efficient automated mining of fuzzy association rules. In *International Conference on Database and Expert Systems Applications*, pages 133–142. Springer, 2002.

[Kat10] Toshiyuki Katagi. Bioconcentration, bioaccumulation, and metabolism of pesticides in aquatic organisms. *Reviews of environmental contamination and toxicology*, pages 1–132, 2010.

[KBK+21] Mosbeh R Kaloop, Abidhan Bardhan, Navid Kardani, Pijush Samui, Jong Wan Hu, and Ahmed Ramzy. Novel application of adaptive swarm intelligence techniques coupled with adaptive network-based fuzzy inference system in predicting photovoltaic power. *Renewable and Sustainable Energy Reviews*, 148:111315, 2021.

[KC07] Michael G Kenward and James Carpenter. Multiple imputation: current perspectives. *Statistical methods in medical research*, 16(3):199–218, 2007.

[KCM24] Chalida Kongsanun, Nawinda Chutsagulprom, and Sompop Moonchai. Spatio-Temporal Dual Kriging with Adaptive Coefficient Drift Function. *Mathematics*, 12(3):400, 2024.

[KDG+04] Timothy Kiely, David Donaldson, Arthur Grube, et al. Pesticides industry sales and usage: 2000 and 2001. *Washington, DC: Office of Prevention, Pesticides and Toxic Substances, United States Environment Protection Agency*, page 16, 2004.

[Kee03] Peter B Keenan. Spatial decision support systems. In *Decision-making support systems: Achievements and challenges for the new decade*, pages 28–39. IGI Global, 2003.

[KEKK21] Valeriya P Kalyabina, Elena N Esimbekova, Kseniya V Kopylova, and Valentina A Kratasyuk. Pesticides: formulants, distribution pathways and effects on human health–a review. *Toxicology reports*, 8:1179–1192, 2021.

[KFP+09] Chandrasekharan Nair Kesavachandran, Mohammad Fareed, Manoj Kumar Pathak, Vipin Bihari, Neeraj Mathur, and Anup Kumar Srivastava. Adverse health effects of pesticides in agrarian populations of developing countries. *Reviews of environmental contamination and toxicology*, 200:33–52, 2009. PMID: 19680610.

[KHN+24] Moh Heri Kurniawan, Hanny Handiyani, Tuti Nuraini, Rr Tutik Sri Hariyati, and Sutrisno Sutrisno. A systematic review of artificial intelligence-powered (ai-powered) chatbot intervention for managing chronic illness. *Annals of Medicine*, 56(1):2302980, 2024.

[Kit14] Open Data Kit. Open Data Kit - about, 2014. Accessed: 2014-06-23. URL: `http://opendatakit.org/about/`.

[KK99] Jaesoo Kim and Nikola Kasabov. HyFIS: adaptive neuro-fuzzy inference systems and their application to nonlinear dynamical systems. *Neural networks*, 12(9):1301–1319, 1999.

[KK10] Rakesh Kumar and Mahesh Kumar. Exploring genetic algorithm for shortest path optimization in data networks. *Global Journal of Computer Science and Technology*, 10(11):8–12, 2010.

[KK19] Dervis Karaboga and Ebubekir Kaya. Adaptive network based fuzzy inference system (ANFIS) training approaches: a comprehensive survey. *Artificial Intelligence Review*, 52(4):2263–2293, 2019.

[KKC+13] Ji-hyun Kim, Jaeyoung Kim, Eun Shil Cha, Yousun Ko, Doo Hwan Kim, and Won Jin Lee. Work-related risk factors by severity for acute pesticide poisoning among male farmers in South Korea. *International journal of environmental research and public health*, 10(3):1100–1112, 2013.

[KMRS14] Anja Knäbel, Karsten Meyer, Jörg Rapp, and Ralf Schulz. Fungicide Field Concentrations Exceed FOCUS Surface Water Predictions: Urgent Need of Model Improvement. *Environmental Science and Technology*, 48(1):455–463, 2014.

[KNNK22] Pascal Kuate Nkounhawa, Dieunedort Ndapeu, and Bienvenu Kenmeugne. Artificial neural network (ANN) and adaptive neuro-fuzzy inference system (ANFIS): application for a photovoltaic system under unstable environmental conditions. *International Journal of Energy and Environmental Engineering*, pages 1–9, 2022.

[Kog98] Marcos Kogan. Integrated pest management: historical perspectives and contemporary developments. *Annual review of entomology*, 43(1):243–270, 1998.

[Koh05] Wolfgang Kohn. *Statistik: Datenanalyse und Wahrscheinlichkeitsrechnung*. Springer-Verlag, 2005.

[KPHW+21] Maren Kruse-Plaß, Frieder Hofmann, Werner Wosniok, Ulrich Schlechtriemen, and Niels Kohlschütter. Pesticides and pesticide-related products in ambient air in germany. *Environmental Sciences Europe*, 33:1–21, 2021.

[KPUR23] Mohammad Amin Kazemi, Mary Pa, Mohammad Nasir Uddin, and Mashallah Rezakazemi. Adaptive neuro-fuzzy inference system based data interpolation for particle image velocimetry in fluid flow applications. *Engineering Applications of Artificial Intelligence*, 119:105723, 2023.

[Kra92] Mark A Kramer. Autoassociative neural networks. *Computers & chemical engineering*, 16(4):313–328, 1992.

[KS99] Nikola K Kasabov and Qun Song. *Dynamic Evolving Fuzzy Neural Networks with" m-out-of-n" Activation Nodes for On-line Adaptive Systems*. Department of Information Science, University of Otago, 1999.

[KS02] Nikola K Kasabov and Qun Song. DENFIS: dynamic evolving neural-fuzzy inference system and its application for time-series prediction. *IEEE transactions on Fuzzy Systems*, 10(2):144–154, 2002.

[KS17] Jacob Kumaresan and Ruwanika Seneviratne. Beginning of a journey: unraveling the mystery of chronic kidney disease of unknown aetiology (CKDu) in Sri Lanka. *Globalization and health*, 13(1):43, 2017.

[KSAKB21] Zhi Hong Kok, Abdul Rashid Mohamed Shariff, Meftah Salem M Alfatni, and Siti Khairunniza-Bejo. Support vector machine in precision agriculture: a review. *Computers and Electronics in Agriculture*, 191:106546, 2021.

[KSM+10] Dilshad Ahmed Khan, Saira Shabbir, Mahwish Majid, Tatheer Alam Naqvi, and Farooq Ahmed Khan. Risk assessment of pesticide exposure on health of Pakistani tobacco farmers. *Journal of exposure science & environmental epidemiology*, 20(2):196–204, 2010.

[KSSS12] Anja Knäbel, Sebastian Stehle, Ralf B Schäfer, and Ralf Schulz. Regulatory FOCUS surface water models fail to predict insecticide concentrations in the field. *Environmental science & technology*, 46(15):8397–8404, 2012.

[KTT+14] Michalis Koureas, Andreas Tsakalof, Manolis Tzatzarakis, Elena Vakonaki, Aristidis Tsatsakis, and Christos Hadjichristodoulou. Biomonitoring of organophosphate exposure of pesticide sprayers and comparison of exposure levels with other population groups in Thessaly (Greece). *Occupational and Environmental Medicine*, 71(2):126–133, 2014.

[KTVC12] Gerard C Kelly, Marcel Tanner, Andrew Vallely, and Archie Clements. Malaria elimination: moving forward with spatial decision support systems. *Trends in Parasitology*, 28(7):297–304, 2012.

[KVR08] Frank Kleemann, G Günter Voß, and Kerstin Rieder. Un (der) paid innovators: The commercial utiliza-tion of consumer work through crowdsourcing. *Science, Technology & Innovation Studies*, 4(1):PP–5, 2008.

[KVSD10] Ye-Sheng Kuo, Sonal Verma, Thomas Schmid, and Prabal Dutta. Hijacking power and bandwidth from the mobile phone's audio interface. In *Proceedings of the First ACM Symposium on Computing for Development*, page 24. ACM, 2010.

[Kwa01] Aileen Kwa. *Agriculture in Developing Countries: Which Way Forward?* South Centre, 2001.

[LAL+11] Marc Lamers, Maria Anyusheva, Nguyen La, Van Vien Nguyen, and Thilo Streck. Pesticide Pollution in Surface- and Groundwater by Paddy Rice Cultivation: A Case Study from Northern Vietnam. *Clean–Soil, Air, Water*, 39(4):356–361, 2011.

[Lal16] Sanjaya Lall. *Developing countries in the international economy: selected papers*. Springer, 2016.

[LBB+14] Martin Lechenet, Vincent Bretagnolle, Christian Bockstaller, François Boissinot, Marie-Sophie Petit, Sandrine Petit, and Nicolas M Munier-Jolain. Reconciling pesticide reduction with economic and environmental sustainability in arable farming. *PloS one*, 9(6):e97922, 2014.

[LBL+99] Andrew S Levey, Juan P Bosch, Julia Breyer Lewis, Tom Greene, Nancy Rogers, and David Roth. A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. *Annals of internal medicine*, 130(6):461–470, 1999.

[LC01] PK Li and Kai Ming Chow. The cost barrier to peritoneal dialysis in the developing world–an Asian perspective. *Peritoneal Dialysis International*, 21(Suppl 3):S307–S313, 2001.

[LC12] Andrew S Levey and Josef Coresh. Chronic kidney disease. *The Lancet*, 379(9811):165–180, 2012.

[LC17] Chen-Nan Liao and Ying-Ju Chen. Farmers' Information Management in Developing Countries - A Highly Asymmetric Information Structure. *Production and Operations Management*, 26(6):1207–1220, 2017.

[LEDL+05] Peng Liu, Elia El-Darzi, Lei Lei, Christos Vasilakis, Panagiotis Chountas, and Wei Huang. An Analysis of Missing Data Treatment Methods and Their Application to Health Care Dataset. In *International Conference on Advanced Data Mining and Applications*, pages 583–590. Springer, 2005.

[LEE90] CC LEE. Fuzzy logic in control systems: fuzzy logic controller-part I. *IEEE Trans. Syst., Man, Cybern.*, 20(2):404–418, 1990.

[LFMBL+22] Monica Lopes-Ferreira, Adolfo Luis Almeida Maleski, Leticia Balan-Lima, Jefferson Thiago Gonçalves Bernardo, Lucas Marques Hipolito, Ana Carolina Seni-Silva, Joao Batista-Filho, Maria Alice Pimentel Falcao, and Carla Lima. Impact of pesticides on human health in the last six years in Brazil. *International journal of environmental research and public health*, 19(6):3198, 2022.

[LFSK00] Chensheng Lu, Richard A Fenske, Nancy J Simcox, and David Kalman. Pesticide exposure of children in an agricultural community: evidence of household proximity to farmland and take home exposure pathways. *Environmental research*, 84(3):290–302, 2000.

[LG05] Mihail Halatchev Le Gruenwald. Estimating missing values in related sensor data streams. In *Advances in Data Management 2005, Proceedings of the Eleventh International Conference on Management of Data.*, 2005.

[LH08] Jin Li and Andrew D Heap. A review of spatial interpolation methods for environmental scientists, 2008. Geoscience Australia Canberra.

[LH11] Jin Li and Andrew D Heap. A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors. *Ecological Informatics*, 6(3-4):228–241, 2011.

[LH14] Jin Li and Andrew D Heap. Spatial interpolation methods applied in the environmental sciences: A review. *Environmental Modelling & Software*, 53:173–189, 2014.

[Li20] Jiada Li. A data-driven improved fuzzy logic control optimization-simulation tool for reducing flooding volume at downstream urban drainage systems. *Science of the Total Environment*, 732:138931, 2020.

[Lie95] Georg Liebscher. Untersuchungen über die Bestimmung des Düngerbedürfnisses der Ackerböden und Kulturpflanzen. *Journal für Landwirtschaft*, 43:49–125, 1895.

[LK06] Loralie J Langman and Bhushan M Kapur. Toxicology: then and now. *Clinical biochemistry*, 39(5):498–510, 2006.

[LKS16] Zhendong Liu, Yawei Kong, and Bin Su. An improved genetic algorithm based on the shortest path problem. In *2016 IEEE international conference on information and automation (ICIA)*, pages 328–332. IEEE, 2016.

[LL+91] Chin-Teng Lin, C. S. George Lee, et al. Neural-network-based fuzzy logic control and decision system. *IEEE Transactions on computers*, 40(12):1320–1336, 1991.

[LL16] Jiamin Li and Harold W Lewis. Fuzzy clustering algorithms - review of the applications. In *2016 IEEE International Conference on Smart Cloud (SmartCloud)*, pages 282–288. IEEE, 2016.

[LNLV12] K Larsen, R Najle, A Lifschitz, and G Virkel. Effects of sub-lethal exposure of rats to the herbicide glyphosate in drinking water: glutathione transferase enzyme activities, levels of reduced glutathione and lipid peroxidation in liver, kidneys and small intestine. *Environmental Toxicology and Pharmacology*, 34(3):811–818, 2012.

[Loc12] M Lockheed. The condition of primary education in developing countries. *Effective schools in developing countries*, pages 20–40, 2012.

[Low96] Robert Lowen. *Fuzzy set theory: basic concepts, techniques and bibliography.* Kluwer Academic Publishers, 1996.

[LP07] Karim R Lakhani and Jill A Panetta. The principles of distributed innovation. *innovations*, 2(3):97–112, 2007.

[LR12] Noel Lopes and Bernardete Ribeiro. Handling missing values via a neural selective input model. *Neural Network World*, 22(4):357, 2012.

[LR19] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

[LRB+13] Adeera Levin, Claudio Rigatto, Brendan Barrett, François Madore, Norman Muirhead, Daniel Holmes, Catherine M Clase, Mila Tang, Ognjenka Djurdjev, CanPREDDICT Investigators, et al. Biomarkers of inflammation, fibrosis, cardiac stretch and injury predict death but not renal replacement therapy at 1 year in a Canadian chronic kidney disease cohort. *Nephrology Dialysis Transplantation*, 29(5):1037–1047, 2013.

[LS10] Andrew S Levey and Lesley A Stevens. Estimating GFR using the CKD epidemiology collaboration (CKD-EPI) creatinine equation: more accurate GFR estimates, lower CKD prevalence estimates, and better risk predictions. *American Journal of Kidney Diseases*, 55(4):622–627, 2010.

[LSA15] Peng Li, Elizabeth A Stuart, and David B Allison. Multiple imputation: a flexible tool for handling missing data. *Jama*, 314(18):1966–1967, 2015.

[LSLS12] Trevor Lewis, Christina Synowiec, Gina Lagomarsino, and Julian Schweitzer. E-health in low-and middle-income countries: findings from the Center for Health Market Innovations. *Bulletin of the World Health Organization*, 90(5):332–340, 2012.

[LSYS11] Rong Li, M Trevor Scholtz, Fuquan Yang, and James J Sloan. A multimedia fate and chemical transport modeling system for pesticides: I. Model development and implementation. *Environmental Research Letters*, 6(3):034029, 2011.

[LTH+20] Wye-Hong Leong, Shu-Yi Teh, Mohammad Moshaddeque Hossain, Thiyagar Nadarajaw, Zabidi Zabidi-Hussin, Swee-Yee Chin, Kok-Song Lai, and Swee-Hua Erin Lim. Application, monitoring and adverse effects in pesticide use: The importance of reinforcement of Good Agricultural Practices (GAPs). *Journal of Environmental Management*, 260:109987, 2020.

[LWHJ19] Xingshuo Li, Huiqing Wen, Yihua Hu, and Lin Jiang. A novel beta parameter based fuzzy-logic controller for photovoltaic MPPT application. *Renewable Energy*, 130:416–427, 2019.

[LWN12] Seppo Leminen, Mika Westerlund, and Anna-Greta Nyström. Living Labs as open-innovation networks. *Technology Innovation Management Review*, 2(9), 2012.

[LXYY21] Mengkun Li, Zhihui Xu, Jun Yang, and Ming Yang. A graph representation model for radiation exposure prediction and path-planning in nuclear power plants. *Annals of Nuclear Energy*, 156:108196, 2021.

[LYGM06] Gang Liu, Xuehong Yang, Yinbing Ge, and Yuxin Miao. An artificial neural network-based expert system for fruit tree disease and insect pest diagnosis. In *2006 IEEE International Conference on Networking, Sensing and Control*, pages 1076–1079. IEEE, 2006.

[LYWL08] Hsiang-Chuan Liu, Jeng-Ming Yih, Der-Bang Wu, and Shin-Wu Liu. Fuzzy C-Mean Clustering Algorithms Based on Picard Iteration and Particle Swarm Optimization. In *2008 international workshop on education technology and training & 2008 international workshop on geoscience and remote sensing*, volume 2, pages 838–842. IEEE, 2008.

[LZ71] Georg Gunther Lorentz and KL Zeller. Birkhoff interpolation. *SIAM Journal on Numerical Analysis*, 8(1):43–48, 1971.

[LZCC05] Han-Xiong Li, Lei Zhang, Kai-Yuan Cai, and Guanrong Chen. An improved robust fuzzy-PID controller with optimal fuzzy reasoning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 35(6):1283–1294, 2005.

[LZJW20] Canjie Luo, Yuanzhi Zhu, Lianwen Jin, and Yongpan Wang. Learn to Augment: Joint Data Augmentation and Network Optimization for Text Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13746–13755, 2020.

[LZKX12] Chaoshun Li, Jianzhong Zhou, Pangao Kou, and Jian Xiao. A novel chaotic particle swarm optimization based fuzzy clustering algorithm. *Neurocomputing*, 83:98–109, 2012.

[MÏ14] Mückenatlas. Mückenatlas - Deutschland kartiert die Stechmücken, 2014. Accessed: 2014-04-02. URL: `http://www.mueckenatlas.de/`.

[MA75] Ebrahim H Mamdani and Sedrak Assilian. An experiment in linguistic synthesis with a fuzzy logic controller. *International Journal of Man-Machine Studies*, 7(1):1–13, 1975.

[MA02] Yalemtsehay Mekonnen and Tadesse Agonafir. Pesticide sprayers' knowledge, attitude and practice of pesticide use on agricultural farms of Ethiopia. *Occupational Medicine*, 52(6):311–315, 2002.

[Man24] Asociacion Mangle. Community-led development in the Bajo Lempa and Bay of Jiquilisco, 2024. Accessed: 2024-11-22. URL: `https://asociacionmangle.org/`.

[MAR+17] Amgad Madkour, Walid G Aref, Faizan Ur Rehman, Mohamed Abdur Rahman, and Saleh Basalamah. A survey of shortest-path algorithms. *arXiv preprint arXiv:1705.02044*, 2017.

[Mat15] Graham Matthews. *Pesticides: health, safety and the environment*. John Wiley & Sons, 2015.

[McD09] P McDougall. Facts and figures: The status of global agriculture. CropLife International, 2009.

[MCK+13] Ewan MacFarlane, Renee Carey, Tessa Keegel, Sonia El-Zaemay, and Lin Fritschi. Dermal Exposure Associated with Occupational End Use of Pesticides and the Role of Protective Measures. *Safety and Health at Work*, 4(3):136–141, 2013.

[MCMT+91] J Marshall Clark, Jacques R Marion, Daniel M Tessier, Matthew W Brooks, and William M Coli. Airborne drift residues collected near apple orchard environments due to application of insecticide mixtures. *Bulletin of Environmental Contamination and Toxicology*, 46(6):829–836, 1991.

[MDB+20] Utkarsh Mital, Dipankar Dwivedi, James B Brown, Boris Faybishenko, Scott L Painter, and Carl I Steefel. Sequential Imputation of Missing Spatio-Temporal Precipitation Data Using Random Forests. *Frontiers in Water*, 2:20, 2020.

[Meg15] Timothy Meggs. Python package anfis, 2015. Accessed: 2015-09-06. URL: `https://github.com/twmeggs/anfis`.

[Mel94] AA Melifonwu. Weeds and their control in cassava. *African Crop Science Journal.*, 2(4):519–530, 1994.

[MH86] James D MacNeil and Mitsuru Hikichi. Phosmet residues in an orchard and adjacent recreational area 1. *Journal of Environmental Science & Health Part B*, 21(5):375–385, 1986.

[Mil15] Andrew Mills. Agriculture in Africa - Transformation and Outlook. Technical report, New Partnership for Africa's Development (NEPAD), 2015.

[MIS+16] Isra Mahmood, Sameen Ruqia Imadi, Kanwal Shazadi, Alvina Gul, and Khalid Rehman Hakeem. Effects of pesticides on environment. In *Plant, Soil and Microbes*, pages 253–269. Springer, 2016.

[MLV08] Nicole Martin, Stefan Lessmann, and Stefan Voß. Crowdsourcing: Systematisierung praktischer Ausprägungen und verwandter Konzepte. In *Multikonferenz Wirtschaftsinformatik*, pages 273–274, 2008.

[MMN24] Diego Marcos, Don McCurdy, and Kevin Ngo. A-frame, 2024. Accessed: 2024-06-12. URL: `https://aframe.io/`.

[MMWZ88] Brian MacMahon, Richard R Monson, Helen H Wang, and Tongzhang Zheng. A second follow-up of mortality in a cohort of pesticide applicators. *Journal of Occupational and Environmental Medicine*, 30(5):429–432, 1988.

[MoARA14] Food Ministry of Agriculture and Ontario Rural Affairs. How Weather Conditions Affect Spray Applications, 2014. Accessed: 2024-10-08. URL: `https://www.ontario.ca/page/how-weather-conditions-affect-spray-applications`.

[MOM18] Daniel Mutembesa, Christopher Omongo, and Ernest Mwebaze. Crowdsourcing real-time viral disease and pest information: A case of nation-wide cassava disease surveillance in a developing country. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 6, pages 117–125, 2018.

[Mou07] Christophe Mouvet. Pesticides in European Groundwaters: Biogeochemical Processes, Contamination Status and Results from a Case Study, 2007.

[MP43] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[MPBV08] R McKinlay, JA Plant, JNB Bell, and N Voulvoulis. Calculating human exposure to endocrine disrupting pesticides via agricultural and non-agricultural exposure routes. *Science of the Total Environment*, 398(1-3):1–12, 2008.

[MPW+99] Rob McConnell, Feliciano Pacheco, Kåre Wahlberg, Willy Klein, Omar Malespin, Ralph Magnotti, Malin Åkerblom, and Douglas Murray. Subclinical Health Effects of Environmental Pesticide Contamination in a Developing Country: Cholinesterase Depression in Children. *Environmental Research*, 81(2):87–91, 1999.

[MQL+14] Roberto Mejía, Edgar Quinteros, Alejandro López, Alexandre Ribó, Humberto Cedillos, Carlos M Orantes, Eliette Valladares, and Dina L López. Pesticide-handling practices in agriculture in El Salvador: an example from 42 patient farmers with chronic kidney disease in the Bajo Lempa region. *Occupational Diseases and Environmental Medicine*, 2(03):56, 2014.

[MR06] Dariush Mozaffarian and Eric B Rimm. Fish intake, contaminants, and human health. *JAMA: the journal of the American Medical Association*, 296(15):1885–1899, 2006.

[MRB+13] Ezra J Mrema, Federico M Rubino, Gabri Brambilla, Angelo Moretto, Aristidis M Tsatsakis, and Claudio Colosio. Persistent organochlorinated pesticides and mechanisms of their toxicity. *Toxicology*, 307:74–88, 2013.

[MRF08] Chad Monfreda, Navin Ramankutty, and Jonathan A Foley. Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000. *Global Biogeochemical Cycles*, 22(1), 2008.

[MRK+15] Ursula S McKnight, Jes J Rasmussen, Brian Kronvang, Philip J Binning, and Poul L Bjerg. Sources, occurrence and predicted aquatic impact of legacy and contemporary pesticides in streams. *Environmental Pollution*, 200:64–76, 2015.

[MRRA+19] Stefan Mandic-Rajcevic, Federico Maria Rubino, Eugenio Ariano, Danilo Cottica, Sara Negri, and Claudio Colosio. Exposure duration and absorbed dose assessment in pesticide-exposed agricultural workers: Implications for risk assessment and modeling. *International Journal of Hygiene and Environmental Health*, 222(3):494–502, 2019.

[MS13] Hans Mohr and Peter Schopfer. *Lehrbuch der Pflanzenphysiologie*. Springer-Verlag, 2013.

[MSI24] Md Moniruzzaman Monir, Subaran Chandra Sarker, and Md Nazrul Islam. Assessing the changing trends of groundwater level with spatiotemporal scale at the northern part of Bangladesh integrating the MAKESENS and ARIMA models. *Modeling Earth Systems and Environment*, 10(1):443–464, 2024.

[MSL06] Donald Mackay, Wan-Ying Shiu, and Sum Chi Lee. *Handbook of physical-chemical properties and environmental fate for organic chemicals*. CRC press, 2006.

[MSS11] Trent A Mankowski, Stephanie J Slater, and Timothy F Slater. An Interpretive Study Of Meanings Citizen Scientists Make When Participating In Galaxy Zoo. *Contemporary Issues in Education Research*, 4(4), 2011.

[MSV18] K Muralitharan, Rathinasamy Sakthivel, and R Vishnuvarthan. Neural network based optimization approach for energy demand prediction in smart grid. *Neurocomputing*, 273:199–208, 2018.

[Mue15] Thomas C Mueller. Methods to measure herbicide volatility. *Weed Science*, 63(SP1):116–120, 2015.

[Nan12] Arup Kumar Nandi. Ga-fuzzy approaches: Application to modeling of manufacturing process. In *Statistical and Computational Techniques in Manufacturing*, pages 145–185. Springer, 2012.

[NAO+12] Leila Naderloo, Reza Alimardani, Mahmoud Omid, Fereydoon Sarmadian, Payam Javadikia, Mohammad Yaser Torabi, and Fatemeh Alimardani. Application of ANFIS to predict crop yield based on different energy inputs. *Measurement*, 45(6):1406–1413, 2012.

[NAS15a] NASA Jet Propulsion Laboratory. Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER), 2015. Accessed: 2015-11-02. URL: `https://asterweb.jpl.nasa.gov/`.

[NAS15b] NASA Jet Propulsion Laboratory. Shuttle radar topography mission (srtm), 2015. Accessed: 2015-11-02. URL: `http://www2.jpl.nasa.gov/srtm/`.

[Nat19a] United Nations. World population prospects 2019 - population density, 2019. Accessed: 2019-09-21. URL: `https://population.un.org/wpp/Download/Files/1_Indicators%20(Standard)/EXCEL_FILES/1_Population/WPP2019_POP_F06_POPULATION_DENSITY.xlsx`.

[Nat19b] United Nations. World population prospects 2019 - total population, 2019. Accessed: 2019-09-21. URL: `https://population.un.org/wpp/Download/Files/1_Indicators%20(Standard)/EXCEL_FILES/1_Population/WPP2019_POP_F01_1_TOTAL_POPULATION_BOTH_SEXES.xlsx`.

[NBL03] Peter Neuenschwander, Christian Borgemeister, and Juergen Langewald. *Biological control in IPM systems in Africa*. CABI, 2003.

[Nex24] Nextcloud. Nextcloud homepage, 2024. Accessed: 2024-03-24. URL: `https://nextcloud.com/`.

[NGKD06] William J Ntow, Huub J Gijzen, Peter Kelderman, and Pay Drechsel. Farmer perceptions and pesticide use practices in vegetable production in ghana. *Pest Management Science*, 62(4):356–365, 2006.

[NK98] Detlef Nauck and Rudolf Kruse. A neuro-fuzzy approach to obtain interpretable fuzzy systems for function approximation. In *1998 IEEE International Conference on Fuzzy Systems Proceedings. IEEE World Congress on*

Computational Intelligence (Cat. No. 98CH36228), volume 2, pages 1106–1111. IEEE, 1998.

[NKK94] Detlef Nauck, Frank Klawonn, and Rudolf Kruse. Neuronale Netze und Fuzzy-Systeme. *Vieweg, Wiesbaden*, 194, 1994.

[NM23] Vaishali Navale and Sumedh Mhaske. Artificial Neural Network (ANN) and Adaptive Neuro-Fuzzy Inference System (ANFIS) model for Forecasting Groundwater Level in the Pravara River Basin, India. *Modeling Earth Systems and Environment*, 9(2):2663–2676, 2023.

[NMS+05] RT Nickson, JM McArthur, B Shrestha, TO Kyaw-Myint, and D Lowry. Arsenic and other drinking water quality issues, Muzaffargarh District, Pakistan. *Applied Geochemistry*, 20(1):55–68, 2005.

[NMST08] Sara Nazari, M Reza Meybodi, M Ali Salehigh, and Sara Taghipour. An Advanced Algorithm for Finding Shortest Path in Car Navigation System. In *2008 First International Conference on Intelligent Networks and Intelligent Systems*, pages 671–674. IEEE, 2008.

[Noh98] Dieter Nohlen. *Lexikon Dritte Welt: Länder, Organisationen, Theorien, Begriffe, Personen.* Rowohlt, 1998.

[NSR23] Pratik Nag, Ying Sun, and Brian J Reich. Spatio-temporal DeepKriging for interpolation and probabilistic forecasting. *Spatial Statistics*, 57:100773, 2023.

[NWC+12] Greg Newman, Andrea Wiggins, Alycia Crall, Eric Graham, Sarah Newman, and Kevin Crowston. The future of citizen science: emerging technologies and shifting paradigms. *Frontiers in Ecology and the Environment*, 10(6):298–304, 2012.

[NWG+17] Nishanthe Nanayakkara, AWM Wazil, Lishanthe Gunerathne, Sewmini Dickowita, Robert Rope, Charaka Ratnayake, Anjali Saxena, and Shuchi Anand. Tackling the Fallout From Chronic Kidney Disease of Unknown Etiology: Why We Need to Focus on Providing Peritoneal Dialysis in Rural, Low-Resource Settings. *Kidney International Reports*, 2(1):1–4, 2017.

[O3b14] O3b Networks. Homepage of O3b Networks, 2014. Accessed: 2014-06-02. URL: `http://www.o3bnetworks.com/`.

[oAFSAF15] United States Department of Agriculture Farm Service Agency (FSA). National Agriculture Imagery Program (NAIP), 2015. Accessed: 2015-10-08. URL: `http://www.fsa.usda.gov/programs-and-services/aerial-photography/imagery-programs/naip-imagery/`.

[OB87] Kyung Whan Oh and Wyllis Bandler. Properties of fuzzy implication operators. *International Journal of Approximate Reasoning*, 1(3):273–285, 1987.

[OCB+02] Deogracias Ortiz, Jaqueline Calderón, Lilia Batres, Leticia Carrizales, Jesús Mejía, Lourdes Martínez, Edelmira García-Nieto, Fernando Díaz-Barriga, et al. Overview of human health and chemical mixtures: problems facing developing countries. *Environmental Health Perspectives*, 110(Suppl 6):901, 2002.

[O'D07] Owen O'Donnell. Access to health care in developing countries: breaking down demand side barriers. *Cadernos de Saúde Pública*, 23(12):2820–2834, 2007.

[OES07] Mahamed GH Omran, Andries P Engelbrecht, and Ayed Salman. An overview of clustering methods. *Intelligent Data Analysis*, 11(6):583–605, 2007.

[OHA+11] Carlos M Orantes, Raúl Herrera, Miguel Almaguer, Elsy G Brizuela, Carlos E Hernández, Héctor Bayarre, Juan C Amaya, Denis J Calero, Patricia Orellana, Rosa M Colindres, Maria E Velazquez, Sonia G Nunez, Veronica M Contreras, and Bertha E Castro. Chronic kidney disease and associated risk factors in the Bajo Lempa region of El Salvador: Nefrolempa study, 2009. *MEDICC review*, 13(4):14–22, 2011. PMID: 22143603.

[OHL+15] CM Navarro Orantes, R Valdés Herrera, MA López, DJ Calero, J de Morales Fuentes, NP Ascencio Alvarado, XF Parada Vela, SM Quezada Zelaya, DV Castro Granados, and P de Figueroa Orellana. Epidemiological characteristics of chronic kidney disease of non-traditional causes in women of agricultural communities of El Salvador. *Clinical Nephrology*, 83(7 Suppl 1):24–31, 2015.

[Oja20] Lauri Ojansivu. Wekan: The Open Source kanban, 2020. Accessed: 2020-04-23. URL: `https://github.com/wekan`.

[OJD+23] Jeroen Ooms, David James, Saikat DebRoy, Hadley Wickham, and Jeffrey Horner. *RMySQL: Database Interface and 'MySQL' Driver for R*, 2023. R package version 0.10.27.

[OLH+13] Corinna Ogonowski, Benedikt Ley, Jan Hess, Lin Wan, and Volker Wulf. Designing for the Living Room: Long-Term User Involvement in a Living Lab. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1539–1548. ACM, 2013.

[oLLE24] European Network of Living Labs (ENoLL). Homepage of the European Network of Living Labs (ENoLL), 2024. Accessed: 2024-12-06. URL: = https://enoll.org/.

[OMAB+14] Carlos M Orantes, MD Miguel Almaguer, Elsy G Brizuela, Lilian Núñez, Juan Carlos Amaya, Denis J Calero, Xavier F Vela, Susana M Zelaya, Delmy V Granados, and Patricia Orellana. Epidemiology of Chronic Kidney Disease in Adults of Salvadoran Agricultural Communities. *MEDICC review*, 16(2):24, 2014.

[ONHVAL+16] Carlos M Orantes-Navarro, Raul Herrera-Valdes, Miguel Almaguer-Lopez, Elsy G Brizuela-Diaz, Nelly P Alvarado-Ascencio, E Jackeline Fuentes-de Morales, Hector D Bayarre-Vea, Denis J Calero-Brizuela, Xavier F Vela-Parada, and Susana M Zelaya-Quezada. Enfermedad renal cronica en ninos y adolescentes en las comunidades agricolas de El Salvador: Estudio NefroSalva Pediatrico (2009-2011). *MEDICC Review*, 18(1-2), 2016.

[ONHVAL+17] Carlos Manuel Orantes-Navarro, Raúl Herrera-Valdés, Miguel Almaguer-López, Laura Lopez-Marin, Xavier Fernando Vela-Parada, Marcelo Hernandez-Cuchillas, and Lilly M Barba. Toward a Comprehensive Hypothesis of Chronic Interstitial Nephritis in Agricultural Communities. *Advances in Chronic Kidney Disease*, 24(2):101–106, 2017.

[OO15] SO Oyewole and OA Ojeleye. Factors influencing the use of improved farm practices among small-scale farmers in Kano State of Nigeria. *Net Journal of Agricultural Science*, 3(1):1–4, 2015.

[oPR13] California Department of Pesticide Regulation. Assessing the health risk of pesticides, 2013. Accessed: 2013-09-25. URL: `http://www.cdpr.ca.gov/docs/dept/factshts/artic12.pdf`.

[oPR16] California Department of Pesticide Regulation. California Pesticide Information Portal, 2016. Accessed: 2016-02-18. URL: `http://calpip.cdpr.ca.gov/main.cfm`.

[Org97] World Health Organisation. *Guidelines for drinking-water quality: Surveillance and control of community supplies*, volume 3. World Health Organization, 1997.

[Org10] World Health Organization. The WHO Recommended Classification of Pesticides by Hazard and Guidelines to Classification 2009, 2010. Geneva: World Health Organization.

[Org15a] World Health Organisation. Environment and health in developing countries, 2015. Accessed: 2015-09-02. URL: `http://www.who.int/heli/risks/ehindevcoun/en/`.

[Org15b] World Health Organization. Trade, foreign policy, diplomacy and health - E-Health, 2015. Accessed: 2015-09-03. URL: `http://www.who.int/trade/glossary/story021/en/`.

[Org18] World Health Organisation. Noncommunicable diseases and their risk factors: STEPwise approach to surveillance (STEPS) , 2018. Accessed: 2018-03-12. URL: `http://www.who.int/ncds/surveillance/steps/en/`.

[OS24] LLC OpenShot Studios. *OpenShot Video Editor Documentation*, 2024. Release 3.1.1-dev. URL: `https://cdn.openshot.org/static/files/user-guide/OpenShotVideoEditor.pdf`.

[oSLOw18] Presidential Task Force of Sri Lanka Official website. Presidential Task Force on Chronic Kidney Disease Prevention, 2018. Accessed: 2018-03-10. URL: `http://www.kidney.presidentialtaskforce.gov.lk/`.

[otEU07] Eurostat Statistical Office of the European Union. *The use of plant protection products in the European Union: Data 1992 - 2003*. Office for Official Publications of the European Communities, Luxemburg, 2007.

[otEU14] Eurostat Statistical Office of the European Union. Pesticide use in agriculture, 2014. Accessed: 2014-09-24. URL: `https://ec.europa.eu/eurostat/databrowser/view/aei_pestuse/default/table?lang=en&category=agr.aei.aei_pes`.

[otEU15] Eurostat Statistical Office of the European Union. Database about sales of pesticides, 2015. Accessed: 2015-03-12. URL: `http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=aei_fm_salpest&lang=en`.

[otEU16] Eurostat Statistical Office of the European Union. Pesticide sales by major groups, 2014, 2016. Accessed: 2015-03-12. URL: `http://ec.europa.eu/eurostat/statistics-explained/index.php/File:Pesticide_sales_by_major_groups,_2014_(Tonnes).png`.

[otSC23] United Nations Environment Programme (UNEP) Secretariat of the Stockholm Convention. Stockholm Convention on persistent organic pollutants (POPs) - text and annexes, 2023. Accessed: 2024-11-24. URL: `https://www.pops.int/Portals/0/download.aspx?d=UNEP-POPS-COP-CONVTEXT-2023.English.pdf`.

[OTW+10] Julie K O'Donnell, Matthew Tobey, Daniel E Weiner, Lesley A Stevens, Sarah Johnson, Peter Stringham, Bruce Cohen, and Daniel R Brooks. Prevalence of and risk factors for chronic kidney disease in rural Nicaragua. *Nephrology Dialysis Transplantation*, 26(9):2798–2805, 2010.

[(PA13] Pan American Health Organization (PAHO). Resolution CD51.R10: Chronic Kidney Disease in Agricultural Communities in Central America, 2013.

[(PA17] Pan American Health Organization (PAHO). Epidemic of Chronic Kidney Disease in Agricultural Communities in Central America. Case definitions, methodological basis and approaches for public health surveillance., 2017.

[Pap09] Christian Papsdorf. *Wie Surfen zu Arbeit wird: Crowdsourcing im Web 2.0.* Campus Verlag, 2009.

[PB05] Edzer J. Pebesma and Roger Bivand. Classes and methods for spatial data in R. *R News*, 5(2):9–13, November 2005.

[PB23] Edzer Pebesma and Roger Bivand. *Spatial Data Science: With applications in R*. Chapman and Hall/CRC, London, 2023.

[PBM+23] Andrea Peris, R Baos, A Martínez, F Sergio, F Hiraldo, and Ethel Eljarrat. Pesticide contamination of bird species from Doñana National Park (south-western Spain): Temporal trends (1999–2021) and reproductive impacts. *Environmental Pollution*, 323:121240, 2023.

[PC13] Gianni Pezzoli and Emanuele Cereda. Exposure to pesticides or solvents and risk of Parkinson disease. *Neurology*, 80(22):2035–2041, 2013.

[PDBSDM05] Kristiaan Pelckmans, Jos De Brabanter, Johan AK Suykens, and Bart De Moor. Handling missing values in support vector machine classifiers. *Neural Networks*, 18(5-6):684–692, 2005.

[PdCdM+16] Vanderley José Pereira, João Paulo Arantes Rodrigues da Cunha, Tâmara Prado de Morais, João Paulo Ribeiro de Oliveira, and João Batista de Morais. Physical-chemical properties of pesticides: concepts, applications, and interactions with the environment. *Bioscience Journal*, 32(3), 2016.

[Pea13] Katy E Pearce. Phoning it in: Theory in mobile media and communication in developing countries. *Mobile Media & Communication*, 1(1):76–82, 2013.

[Peb18] Edzer Pebesma. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10(1):439–446, 2018.

[PGB+08] David H Peters, Anu Garg, Gerry Bloom, Damian G Walker, William R Brieger, and M Hafizur Rahman. Poverty and access to health care in developing countries. *Annals of the New York Academy of Sciences*, 1136(1):161–171, 2008.

[PGDG08] PostgreSQL Global Development Group. PostgreSQL, 2008. URL: `http://www.postgresql.org`.

[PH11] Radu-Emil Precup and Hans Hellendoorn. A survey on industrial applications of fuzzy control. *Computers in Industry*, 62(3):213–226, 2011.

[PHH08] Eric Paulos, RJ Honicky, and Ben Hooker. Citizen Science: Enabling Participatory Urbanism. *Handbook of Research on Urban Informatics*, pages 414–436, 2008.

[Pim97] David Pimentel. *Techniques for Reducing Pesticide Use: Economic and Environmental Benefits*. John Wiley and Sons, 1997.

[Pim09] David Pimentel. Pest Control in World Agriculture. *Agricultural Science*, 2:272–293, 2009.

[Pla01] Richard E Plant. Site-specific management: the application of information technology to crop production. *Computers and Electronics in Agriculture*, 30(1):9–29, 2001.

[Pla14] Melanie Platz. *Mathematical Modelling of GIS Tailored GUI Design with the Application of Spatial Fuzzy Logic*. PhD thesis, Universität Koblenz-Landau, 2014.

[PLL+15] Mukesh Prasad, Chin-Teng Lin, Dong-Lin Li, Chao-Tien Hong, Wei-Ping Ding, and Jyh-Yeong Chang. Soft-Boosted Self-Constructing Neural Fuzzy Inference Network. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 47(3):584–588, 2015.

[PLN+20] Joanna Potapowicz, Dimitra Lambropoulou, Christina Nannou, Krystyna Kozioł, and Żaneta Polkowska. Occurrences, sources, and transport of organochlorine pesticides in the aquatic environment of Antarctica. *Science of the Total Environment*, page 139475, 2020.

[PM79] Tom J Procyk and Ebrahim H Mamdani. A linguistic self-organizing process controller. *Automatica*, 15(1):15–30, 1979.

[PNW20] Bo Pang, Erik Nijkamp, and Ying Nian Wu. Deep Learning With TensorFlow: A Review. *Journal of Educational and Behavioral Statistics*, 45(2):227–248, 2020.

[PO14] Jayesh H Patel and Markand P Oza. Deriving crop calendar using NDVI time-series. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 1:869–873, 2014.

[Pos24] PostGIS. PostGIS homepage, 2024. Accessed: 2024-04-13. URL: `http://postgis.net/`.

[PotEU06] European Parliament and Council of the European Union. Directive 2006/118/EC of the European Parliament and of the Council of 12 December 2006 on the protection of groundwater against pollution and deterioration. *Official Journal of the European Union, L*, 372:19–31, 2006.

[PPG14] Santasriya Prasad, Sateesh K Peddoju, and Debashis Ghosh. Energy efficient mobile vision system for plant leaf disease identification. In *Wireless Communications and Networking Conference (WCNC), 2014 IEEE*, pages 3314–3319. IEEE, 2014.

[Pro11] United Nations Environment Programme. Vector map: Sub Country Administrative Units 1998, 2011. Accessed: 2011-04-16. URL: `http://geodata.grid.unep.ch/mod_download/download_geospatial.php?selectedID=290&newFile=download/admin98_li_shp.zip`.

[Pro19a] United Nations Development Programme. Climate Change Adaption El Salvador, 2019. Accessed: 2019-09-25. URL: `https://www.adaptation-undp.org/explore/el-salvador`.

[Pro19b] United Nations Development Programme. Human Development Reports - Trends in the Human Development Index, 1990-2014, 2019. Accessed: 2019-09-25. URL: `http://hdr.undp.org/en/composite/trends`.

[Pro24] Proxmox Server Solutions GmbH. Proxmox homepage, 2024. Accessed: 2024-06-19. URL: `https://www.proxmox.com/`.

[PS91] JT Parsons and GA Surgeoner. Effect of exposure time on the acute toxicities of permethrin, fenitrothion, carbaryl and carbofuran to mosquito larvae. *Environmental Toxicology and Chemistry: An International Journal*, 10(9):1219–1227, 1991.

[PSK+11] Konstantinos Poirazidis, Stefan Schindler, Vassiliki Kati, Aristotelis Martinis, Dionissios Kalivas, Dimitris Kasimiadis, Thomas Wrbka, and Aristotelis C Papageorgiou. Conservation of Biodiversity in Managed Forests:

Developing an Adaptive Decision Support System. In *Landscape Ecology in Forest Management and Conservation*, pages 380–399. Springer, 2011.

[PVG+11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

[PvGV+22] Speranza C Panico, Cornelis AM van Gestel, Rudo A Verweij, Magali Rault, Colette Bertrand, Carlos A Menacho Barriga, Michaël Coeurdassier, Clémentine Fritsch, Frédéric Gimbert, and Céline Pelosi. Field mixtures of currently used pesticides in agricultural soil pose a risk to soil invertebrates. *Environmental Pollution*, 305:119290, 2022.

[PVQ+22] Gustavo Willam Pereira, Domingos Sárvio Magalhães Valente, Daniel Marçal de Queiroz, André Luiz de Freitas Coelho, Marcelo Marques Costa, and Tony Grift. Smart-Map: An Open-Source QGIS Plugin for Digital Mapping Using Machine Learning Techniques and Ordinary Kriging. *Agronomy*, 12(6):1350, 2022.

[QATRM03] Muhammad Ghulam Quibria, Shamsun N Ahmed, Ted Tschang, and Mari-Len Reyes-Macasaquit. Digital divide: determinants and policies with special reference to Asia. *Journal of Asian Economics*, 13(6):811–825, 2003.

[QGI24] QGIS Development Team. *QGIS Geographic Information System*. Open Source Geospatial Foundation, 2024. URL: `http://qgis.osgeo.org`.

[QRM+17] Edgar Quinteros, Alexandre Ribó, Roberto Mejía, Alejandro López, Wilfredo Belteton, Aimee Comandari, Carlos M Orantes, Ernesto B Pleites, Carlos E Hernández, and Dina L López. Heavy metals and pesticide exposure from agricultural activities and former agrochemical factory in a Salvadoran rural community. *Environmental Science and Pollution Research*, 24(2):1662–1676, 2017.

[QZHS92] Wu Zhi Qiao, Wang Pei Zhuang, Teh Hoon Heng, and Song Shou Shan. A rule self-regulating fuzzy controller. *Fuzzy Sets and Systems*, 47(1):13–21, 1992.

[R C20] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[RACR19] Francisco Rosales-Ayala and Rooel Campos-Rodriguez. Gestión de las aguas residuales en la ciudad La Libertad, El Salvador. *Revista Tecnología en Marcha*, pages ?g–43, 2019.

[Rap11] Jörg Rapp. Abschätzung der global eingesetzten Insektizidwirkstoffmengen mithilfe eines geografischen Informationssystems. diplomathesis, University of Koblenz-Landau, 2011.

[RBG+10] M Jordan Raddick, Georgia Bracey, Pamela L Gay, Chris J Lintott, Phil Murray, Kevin Schawinski, Alexander S Szalay, and Jan Vandenberg. Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers. *Astronomy Education Review*, 9(1):010103, 2010.

[RBHB15] Lala Septem Riza, Christoph Bergmeir, Francisco Herrera, and José M Benítez. frbs: Fuzzy Rule-Based Systems for Classification and Regression in R. *Journal of Statistical Software*, 65:1–30, 2015.

[RE14] Jörg Rapp and Niehaus Engelbert. How to mitigate risk through the use of insect identification tools and augmented reality software, 2014. Presentation: International Expert Meeting of the Action Team 6 Follow-Up Initiative.

[RECO15] José F Reyes, Wilson Esquivel, Daniel Cifuentes, and Rodrigo Ortega. Field testing of an automatic control system for variable rate fertilizer application. *Computers and Electronics in Agriculture*, 113:260–265, 2015.

[Rei16] Xenia-Rosemarie Reit. *Denkstrukturen in Lösungsansätzen von Modellierungsaufgaben: eine kognitionspsychologische Analyse schwierigkeitsgenerierender Aspekte*. Springer-Verlag, 2016.

[RG06] Richard Reiss and John Griffin. A probabilistic model for acute bystander exposure and risk assessment for soil fumigants. *Atmospheric Environment*, 40(19):3548–3560, 2006.

[Rit10] Christian Ritz. Toward a unified approach to dose–response modeling in ecotoxicology. *Environmental Toxicology and Chemistry*, 29(1):220–229, 2010.

[RJLJ+15] C Roncal-Jimenez, MA Lanaspa, T Jensen, LG Sanchez-Lozada, and RJ Johnson. Mechanisms by Which Dehydration May Lead to Chronic Kidney Disease. *Annals of Nutrition and Metabolism*, 66(Suppl. 3):10–13, 2015.

[RLS+20] Syahidah Izza Rufaida, Jenq-Shiou Leu, Kuan-Wu Su, Azril Haniz, and Jun-Ichi Takada. Construction of an indoor radio environment map using gradient boosting decision tree. *Wireless Networks*, 26:6215–6236, 2020.

[RMA+12] Zeshan A Rajput, Samuel Mbugua, David Amadi, Viola Chepng'eno, Jason J Saleem, Yaw Anokwa, Carl Hartung, Gaetano Borriello, Burke W Mamlin, Samson K Ndege, et al. Evaluation of an Android-based mHealth system for population surveillance in developing countries. *Journal of the American Medical Informatics Association*, pages amiajnl–2011, 2012.

[RMD+90] Mauricio Restrepo, Nubia Muñoz, Nicholas E Day, José E Parra, Laura de Romero, and Xuan Nguyen-Dinh. Prevalence of adverse reproductive outcomes in a population occupationally exposed to pesticides in Colombia. *Scandinavian Journal of Work, Wnvironment & Health*, pages 232–238, 1990.

[RMV00] GJ Roerink, M Menenti, and W Verhoef. Reconstructing cloudfree NDVI composites using Fourier analysis of time series. *International Journal of Remote Sensing*, 21(9):1911–1917, 2000.

[Roj96] Raúl Rojas. *Neutral Networks: A Systematic Introduction*. Springer, 1996.

[Ron22] Xiao Rong. *deepnet: Deep Learning Toolkit in R*, 2022. R package version 0.2.1.

[Ros58] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[Ros01] Robert I Rose. Pesticides and Public Health: Integrated Methods of Mosquito Management. *Emerging Infectious Diseases*, 7(1):17, 2001.

[RP16] Parit Raja and Batu Pahat. A review of training methods of ANFIS for applications in business and economics. *International Journal of u-and e-Service, Science and Technology*, 9(7):165–172, 2016.

[RPH+12] Dana Rotman, Jenny Preece, Jen Hammock, Kezee Procita, Derek Hansen, Cynthia Parr, Darcy Lewis, and David Jacobs. Dynamic changes in motivation in collaborative citizen-science projects. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pages 217–226, 2012.

[RRR87] P Rita, PP Reddy, and S Venkatram Reddy. Monitoring of workers occupationally exposed to pesticides in grape gardens of Andhra Pradesh. *Environmental research*, 44(1):1–5, 1987.

[RRS06] Rudolph P Rull, Beate Ritz, and Gary M Shaw. Validation of self-reported proximity to agricultural crops in a case–control study of neural tube defects. *Journal of Exposure Science & Environmental Epidemiology*, 16(2):147–155, 2006.

[RSF+95] L Ritter, KR Solomon, J Forget, M Stemeroff, and C O'leary. A Review of Selected Persistent Organic Pollutants. *International Programme on Chemical Safety (IPCS). PCS/95.39. Geneva: World Health Organization*, 65:66, 1995.

[RSL+20] Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020.

[RSt20] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, PBC, Boston, MA, 2020.

[RSt21] RStudio Team. Shiny server open source 1.5.16.958, 2021. URL: `https://rstudio.com/products/shiny/shiny-server/`.

[RT20] PS Raja and KJSC Thangavel. Missing value imputation using unsupervised machine learning techniques. *Soft Computing*, 24(6):4361–4392, 2020.

[Rub04] Donald B Rubin. *Multiple Imputation for Nonresponse in Surveys*, volume 81. John Wiley & Sons, 2004.

[Rud16] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[Run96] Thomas A Runkler. Extended defuzzification methods and their properties. In *Proceedings of IEEE 5th international fuzzy systems*, volume 1, pages 694–700. IEEE, 1996.

[RWM24] R Special Interest Group on Databases (R-SIG-DB), Hadley Wickham, and Kirill Müller. *DBI: R Database Interface*, 2024. R package version 1.2.3.

[RZJ+20] Jairos Rurinda, Shamie Zingore, Jibrin M Jibrin, Tesfaye Balemi, Kenneth Masuki, Jens A Andersson, Mirasol F Pampolino, Ibrahim Mohammed, James Mutegi, Alpha Y Kamara, et al. Science-based decision support for formulating crop fertilizer recommendations in sub-saharan africa. *Agricultural Systems*, 180:102790, 2020.

[SAB+13] N Saglam, B Aydogdu, K Belliturk, E Kesici, A Urusan, B Akdemir, and MG Ungor. Development of a Precision Farming System for Turkish Farmers. In *International Conference on Agricultural Engineering: New Technologies for Sustainable Agricultural Production and Food Security 1054*, pages 301–308, 2013.

[Sac01] Jeffrey Sachs. *Macroeconomics and Health: Investing in Health for Economic Development*. World Health Organization, 2001.

[SAG24] SAGA User Group Association. Saga - system for automated geoscientific analyses, 2024. URL: `https://saga-gis.sourceforge.io/`.

[Sam24] Veronika Samborska. How much have temperatures risen in countries across the world? *Our World in Data*, 2024. URL: `https://ourworldindata.org/temperature-anomaly`.

[SANN18] Jean Sonchieu, Edouard Nantia Akono, Cheche Tanwi Ngwamitang, and Benoït Martin Ngassoum. Heath risk among pesticide sellers in Bamenda (Cameroon) and peripheral areas. *Environmental Science and Pollution Research*, 25(10):9454–9460, 2018.

[SAP+08] Kishore Gnana Sam, Hira H Andrade, Lisa Pradhan, Abhishek Pradhan, Shashi J Sones, Padma GM Rao, and Christopher Sudhakar. Effectiveness of an educational program to promote pesticide safety among pesticide

handlers of South India. *International Archives of Occupational and Environmental Health*, 81(6):787–795, 2008.

[Sat01] Robert Sattelberger. *Einsatz von Pflanzenschutzmitteln und Biozid-Produkten im nicht-land-und forstwirtschaftlichen Bereich.* Umweltbundesamt, 2001.

[SB12] Daniel J Stekhoven and Peter Bühlmann. Missforest - non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

[SBC+08] Alicia L Salvatore, Asa Bradman, Rosemary Castorina, José Camacho, Jesús López, Dana B Barr, John Snyder, Nicholas P Jewell, and Brenda Eskenazi. Occupational Behaviors and Farmworkers' Pesticide Exposure: Findings From a Study in Monterey County, California. *American Journal of Industrial Medicine*, 51(10):782–794, 2008.

[SBD21] K Samal, Korra Babu, and Santos Das. Spatio-temporal Prediction of Air Quality using Distance Based Interpolation and Deep Learning Techniques. *EAI Endorsed Transactions on Smart Cities*, 5(14), 2021.

[SBP+21] Ralf Schulz, Sascha Bub, Lara L Petschick, Sebastian Stehle, and Jakob Wolfram. Applied pesticide toxicity shifts toward plants and invertebrates, even in GM crops. *Science*, 372(6537):81–84, 2021.

[SBW19] Francisco Sánchez-Bayo and Kris AG Wyckhuys. Worldwide decline of the entomofauna: A review of its drivers. *Biological conservation*, 232:8–27, 2019.

[SCA+10] Scott L Sanoff, Luis Callejas, Carlos D Alonso, Yichun Hu, Romulo E Colindres, Hyunsook Chin, Douglas R Morgan, and Susan L Hogan. Positive association of renal insufficiency with agriculture employment and unregulated alcohol consumption in Nicaragua. *Renal Failure*, 32(7):766–777, 2010. PMID: 20662688.

[Sch97] Joseph L Schafer. *Analysis of Incomplete Multivariate Data.* Chapman and Hall/CRC, 1997.

[Sch01] Ralf Schulz. Comparison of spray drift-and runoff-related input of azinphos-methyl and endosulfan from fruit orchards into the Lourens River, South Africa. *Chemosphere*, 45(4):543–551, 2001.

[Sch12] Alexander Schrijver. On the History of the Shortest Path Problem. *Documenta Mathematica*, 17(1):155–167, 2012.

[Sch21] Christian Schenk. Miktex 21.2 [tex distribution], 2021. URL: `https://miktex.org/`.

[SCL+11] Barry Schouten, Melania Calinescu, Annemieke Luiten, et al. *Optimizing quality of response through adaptive survey designs.* Citeseer, 2011.

[SD10] Ramanathan Sugumaran and John Degroote. *Spatial decision support systems: Principles and practices.* Crc Press, 2010.

[SDH99] T_J Schmitt, MG Dosskey, and KD Hoagland. Filter Strip Performance and Processes for Different Vegetation, Widths, and Contaminants. Technical report, Wiley Online Library, 1999.

[SdSJH16] Ferry Susanto, Paulo de Souza Jr, and Jing He. Spatiotemporal Interpolation for Environmental Modelling. *Sensors*, 16(8):1245, 2016.

[SDTR+16] Joanna Socorro, Amandine Durand, Brice Temime-Roussel, Sasho Gligorovski, Henri Wortham, and Etienne Quivet. The persistence of pesticides in atmospheric particulate phase: An emerging air quality issue. *Scientific Reports*, 6(1):33456, 2016.

[SE03] Alfred Stein and Christien Ettema. An overview of spatial sampling procedures and experimental design of spatial studies for ecosystem comparisons. *Agriculture, Ecosystems & Environment*, 94(1):31–47, 2003.

[Sen12] Kevin Sene. *Flash Floods: Forecasting and Warning.* Springer Science & Business Media, 2012.

[SFH+06] Gerald R Stephenson, Ian G Ferris, Patrick T Holland, Monica Nordberg, et al. Glossary of terms relating to pesticides (IUPAC Recommendations 2006). *Pure and Applied Chemistry*, 78(11):2075–2154, 2006.

[SFPSO15] Telmo M Silva Filho, Bruno A Pimentel, Renata MCR Souza, and Adriano LI Oliveira. Hybrid methods for fuzzy clustering based on fuzzy c-means and improved particle swarm optimization. *Expert Systems with Applications*, 42(17-18):6315–6328, 2015.

[SGKJ21] Swagata Sarkar, Juliana Dias Bernardes Gil, James Keeley, and Kees Jansen. *The use of pesticides in developing countries and their impact on health and the right to food*. European Union, 2021.

[Sha06] Haresh C Shah. The last mile: earthquake risk mitigation assistance in developing countries. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 364(1845):2183–2189, 2006.

[SHA+18] Mahmoud A Soliman, Hany M Hasanien, Haitham Z Azazi, Elwy E El-Kholy, and Sabry A Mahmoud. An Adaptive Fuzzy Logic Control Strategy for Performance Enhancement of a Grid-Connected PMSG-Based Wind Turbine. *IEEE Transactions on Industrial Informatics*, 15(6):3163–3173, 2018.

[Sil09] Jonathan Silvertown. A new dawn for citizen science. *Trends in ecology & evolution*, 24(9):467–471, 2009.

[Sin92] Alok Sinha. Client-server computing. *Communications of the ACM*, 35(7):77–98, 1992.

[SKH+20] Aleksandar Sekulić, Milan Kilibarda, Gerard BM Heuvelink, Mladen Nikolić, and Branislav Bajat. Random Forest Spatial Interpolation. *Remote Sensing*, 12(10):1687, 2020.

[SKTGMdS+20] Lucas Silveira Kupssinskü, Tainá Thomassim Guimarães, Eniuce Menezes de Souza, Daniel C. Zanotta, Mauricio Roberto Veronez, Luiz Gonzaga Jr, and Frederico Fabio Mauad. A Method for Chlorophyll-a and Suspended Solids Prediction through Remote Sensing and Machine Learning. *Sensors*, 20(7):2125, 2020.

[SLW+10] Peter Soderland, Shachi Lovekar, Daniel E Weiner, Daniel R Brooks, and James S Kaufman. Chronic kidney disease associated with environmental toxins and exposures. *Advances in chronic kidney disease*, 17(3):254–264, 2010. PMID: 20439094.

[SLZ+07] Steven E Sexton, Zhen Lei, David Zilberman, et al. The Economics of Pesticides and Pest Control. *International Review of Environmental and Resource Economics*, 1(3):271–326, 2007.

[SM87] Philippe Smets and Paul Magrez. Implication in Fuzzy Logic. *International Journal of Approximate Reasoning*, 1(4):327–347, 1987.

[SM98] Gerrit Schüürmann and Bernd Markert. *Ecotoxicology: Ecological Fundamentals, Chemical Exposure, and Biological Effects*, volume 113. Wiley-Interscience, 1998.

[SMS+03] Karen Semchuk, Helen McDuffie, Ambikaipakan Senthilselvan, James Dosman, Allan Cessna, and Donald Irvine. Factors associated with detection of bromoxynil in a sample of rural residents. *Journal of Toxicology and Environmental Health Part A*, 66(2):103–132, 2003.

[Spr70] John B Sprague. Measurement of pollutant toxicity to fish. II. Utilizing and applying bioassay results. *Water Research*, 4(1):3–32, 1970.

[SQB09] Kakee Scott, Jaco Quist, and Conny Bakker. Co-design, social practices and sustainable innovation: involving users in a living lab exploratory study on bathing. In *Proceedings of Paper for the Joint Actions on Climate Change? Conference, Aalborg, Denmark*, pages 8–9, 2009.

[SRF12] Verena Seufert, Navin Ramankutty, and Jonathan A Foley. Comparing the yields of organic and conventional agriculture. *Nature*, 485(7397):229–232, 2012.

[SRPMLC15] Esther-Lydia Silva-Ramírez, Rafael Pino-Mejías, and Manuel López-Coello. Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbours for monotone patterns. *Applied Soft Computing*, 29:65–74, 2015.

[SS15] Sebastian Stehle and Ralf Schulz. Agricultural insecticides threaten surface waters at the global scale. *Proceedings of the National Academy of Sciences*, 112(18):5750–5755, 2015.

[ST03] Marc Schröder and Jürgen Trouvain. The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology*, 6:365–377, 2003.

[ST12] Pepijn Schreinemachers and Prasnee Tipraqsa. Agricultural pesticides and land use intensification in high, middle and low income countries. *Food Policy*, 37(6):616–626, 2012.

[Ste04] Jørgen Stenersen. *Chemical Pesticides Mode of Action and Toxicology*. CRC press, 2004.

[STM+12] DR Sharma, Resham Bahadur Thapa, HK Manandhar, SM Shrestha, and SB Pradhan. Use Of Pesticides In Nepal And Impacts On Human Health And Environment. *Journal of Agriculture and environment*, 13:67–74, 2012.

[Sto18] Mirko Stojčić. Application of ANFIS model in road traffic and transportation: a literature review from 1993 to 2018. *Operational Research in Engineering Sciences: Theory and Applications*, 1(1):40–61, 2018.

[Str03] Roger Strasser. Rural health around the world: challenges and solutions. *Family practice*, 20(4):457–463, 2003.

[Suá05] LA Suárez. PRZM -3, A Model for Predicting Pesticide and Nitrogen Fate in the Crop Root and Unsaturated Soil Zones: Users Manual for Release 3.12.2. *United States Environmental Protection Agency (EPA), Washington, DC*, 2005.

[Sur05] James Surowiecki. *The wisdom of crowds*. Random House LLC, 2005.

[Sur15a] United States Geological Survey. Landsat 8 OLI (Operational Land Imager) and TIRS (Thermal Infrared Sensor), 2015. Accessed: 2015-10-29. URL: `https://lta.cr.usgs.gov/L8`.

[Sur15b] United States Geological Survey. Manual: Using the USGS Landsat 8 Product, 2015. Accessed: 2015-10-12. URL: `http://landsat.usgs.gov/Landsat8_Using_Product.php`.

[Sur15c] United States Geological Survey. What is the landsat satellite program and why is it important?, 2015. Accessed: 2025-02-04. URL: `https://www.usgs.gov/faqs/what-landsat-satellite-program-and-why-it-important`.

[Sut08] Ewan Sutherland. Counting mobile phones, sim cards & customers. *Sim Cards & Customers (September 5, 2009)*, 2008.

[SVCCL23] John D Stamford, Silvere Vialet-Chabrand, Iain Cameron, and Tracy Lawson. Development of an accurate low cost NDVI imaging system for assessing plant health. *Plant Methods*, 19(1):9, 2023.

[Swa12] Melanie Swan. Crowdsourced Health Research Studies: An Important Emerging Complement to Clinical Trials in the Public Health Research Ecosystem. *Journal of Medical Internet Research*, 14(2), 2012.

[SWC+09] Jonathan AC Sterne, Ian R White, John B Carlin, Michael Spratt, Patrick Royston, Michael G Kenward, Angela M Wood, and James R Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, 338, 2009.

[SX95] Thomas Sauer and Yuan Xu. On multivariate Hermite interpolation. *Advances in Computational Mathematics*, 4(1):207–259, 1995.

[TALS20] Mengya Tao, Paul R Adler, Ashley E Larsen, and Sangwon Suh. Pesticide application rates and their toxicological impacts: why do they vary so widely across the us? *Environmental Research Letters*, 15(12):124049, 2020.

[TB12] Yeong Sheng Tey and Mark Brindal. Factors influencing the adoption of precision agricultural technologies: a review for policy implications. *Precision Agriculture*, 13(6):713–730, 2012.

[TBBC00] Deborah J Trumbull, Rick Bonney, Derek Bascom, and Anna Cabral. Thinking scientifically during participation in a citizen-science project. *Science education*, 84(2):265–275, 2000.

[TBE+02] Milton E Teske, Sandra L Bird, David M Esterly, Thomas B Curbishley, Scott L Ray, and Steven G Perry. Agdrift®: A model for estimating near-field spray drift from aerial applications. *Environmental Toxicology and Chemistry: An International Journal*, 21(3):659–671, 2002.

[TBG+06] William M Tierney, Eduard J Beck, Reed M Gardner, Beverly Musick, Mark Shields, Naomi M Shiyonga, and Mark H Spohr. A Pragmatic Approach to

Constructing a Minimum Data Set for Care of Patients with HIV in Developing Countries. *Journal of the American Medical Informatics Association*, 13(3):253–260, 2006.

[TBVdLV06] A Tiktak, JJTI Boesten, AMA Van der Linden, and Marnik Vanclooster. Mapping ground water vulnerability to pesticide leaching with a process-based metamodel of europearl. *Journal of environmental quality*, 35(4):1213–1226, 2006.

[TCG+03] Beti Thompson, Gloria D Coronado, Julia E Grossman, Klaus Puschel, Cam C Solomon, Ilda Islas, Cynthia L Curl, Jeffry H Shirai, John C Kissel, and Richard A Fenske. Pesticide take-home pathway among children of agricultural workers: study design, methods, and baseline findings. *Journal of Occupational and Environmental Medicine*, 45(1):42–53, 2003.

[TCS+01] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[TCS12] Riccardo Taormina, Kwok-Wing Chau, and Rajandrea Sethi. Artificial neural network simulation of hourly groundwater levels in a coastal aquifer system of the Venice lagoon. *Engineering Applications of Artificial Intelligence*, 25(8):1670–1676, 2012.

[TDRW+21] Muyesaier Tudi, Huada Daniel Ruan, Li Wang, Jia Lyu, Ross Sadler, Des Connell, Cordia Chu, and Dung Tri Phung. Agriculture Development, Pesticide Application and Its Impact on the Environment. *International Journal of Environmental Research and Public Health*, 18(3):1112, 2021.

[Tea15] Geographic Resources Analysis Support System (GRASS) Development Team. GRASS GIS 7.0.2svn Reference Manual, 2015. Accessed: 2015-10-17. URL: `https://grass.osgeo.org/grass70/manuals/`.

[Tea20] Ampache Team. Ampache media server 4.2.3, 2020. Accessed: 2020-01-13. URL: `http://ampache.org/`.

[Tea24] The Audacity Team. *Audacity 3.5 Reference Manual*, 2024. URL: `https://manual.audacityteam.org/`.

[Tei86] PM Teillet. Image correction for radiometric effects in remote sensing. *International Journal of Remote Sensing*, 7(12):1637–1651, 1986.

[TGLR14] Patricia L Toccalino, Robert J Gilliom, Bruce D Lindsey, and Michael G Rupert. Pesticides in Groundwater of the United States: Decadal-Scale Changes, 1993–2011. *Groundwater*, 52(S1):112–125, 2014.

[THBVdB16] MMS Ter Horst, WHJ Beltman, and F Van den Berg. The TOXSWA model version 3.3 for pesticide behaviour in small surface waters: description of processes. Statutory Research Tasks Unit for Nature & the Environment. Technical report, WOt-technical report 84, 2016.

[The19] The GIMP Development Team. Gimp, 2019. version: 2.10.12. Accessed: 2019-06-12. URL: `https://www.gimp.org`.

[THK07] EL Taylor, A Gordon Holley, and Melanie Kirk. Pesticide Development - A Brief Look at the History. *Athens, GA: Southern Regional Extension Forestry*, 2007.

[Thy06] Katharina Thywissen. Core Terminology of Disaster Reduction. *Measuring Vulnerability to Natural Hazards: Towards Disaster Resilient Societies, United Nations University Press, Hong Kong*, 2006.

[TI17] Fei Tang and Hemant Ishwaran. Random Forest Missing Data Algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 10(6):363–377, 2017.

[Tie92] Herbert Tiedemann. *Earthquakes and Volcanic Eruptions: A Handbook on Risk Assessment*. Swiss Re, 1992.

[Tou20] LW Touart. The Federal Insecticide, Fungicide, and Rodenticide Act. In *Fundamentals of Aquatic Toxicology*, pages 657–668. CRC Press, 2020.

[TQG09] Sau Wai Tung, Chai Quek, and Cuntai Guan. T2-HyFIS-Yager: Type 2 Hybrid Neural Fuzzy Inference System Realizing Yager Inference. In *2009 IEEE International Conference on Fuzzy Systems*, pages 80–85. IEEE, 2009.

[TQG12] Sau Wai Tung, Chai Quek, and Cuntai Guan. SoHyFIS-Yager: A self-organizing Yager based Hybrid neural Fuzzy Inference System. *Expert Systems with Applications*, 39(17):12759–12771, 2012.

[TR21] Tressy Thomas and Enayat Rajabi. A systematic review of machine learning-based missing value imputation techniques. *Data Technologies and Applications*, 55(4):558–585, 2021.

[Trö11] Fritz Tröster. *Steuerungs-und Regelungstechnik für Ingenieure*. Oldenbourg Verlag, 2011.

[TRvdBA+15] JA Te Roller, F van den Berg, PI Adriaanse, A de Jong, and WHJ Beltman. Surface water scenario help (swash) version 5.3: technical description. Technical report, Statutory Research Tasks Unit for Nature & the Environment (WOT Natuur & Milieu), 2015.

[TS83] Tomohiro Takagi and Michio Sugeno. Derivation of Fuzzy Control Rules from Human Operator's Control Actions. In *Proceedings of the IFAC symposium on fuzzy information, knowledge representation and decision analysis*, volume 6, pages 55–60, 1983.

[TTL95] Gerald Tesauro, David S Touretzky, and Todd Leen. *Advances in neural information processing systems 7*, volume 7. MIT press, 1995.

[TWK+21] Andrius Tamosiunas, Hans A Winther, Kazuya Koyama, David J Bacon, Robert C Nichol, and Ben Mawdsley. Investigating cosmological GAN emulators using latent space interpolation. *Monthly Notices of the Royal Astronomical Society*, 506(2):3049–3067, 2021.

[TWPZ12] Stewart G Trost, Weng-Keen Wong, Karen A Pfeiffer, and Yonglei Zheng. Artificial Neural Networks to Predict Activity Type and Energy Expenditure in Youth. *Medicine and Science in Sports and Exercise*, 44(9):1801, 2012.

[(USa] United States Environmental Protection Agency (USEPA). Assessing Health Risks from Pesticides. Accessed: 2013-09-06. URL: `http://www.epa.gov/pesticides/factsheets/riskassess.htm`.

[(USb] United States Environmental Protection Agency (USEPA). Final Test Guidelines for Pesticides and Toxic Sub-

stances. Accessed: 2024-12-13. URL: `https://www.epa.gov/test-guidelines-pesticides-and-toxic-substances/final-test-guidelines-pesticides-and-toxic`.

[(USc] United States Environmental Protection Agency (USEPA). Overview of Risk Assessment in the Pesticide Program. Accessed: 2013-09-25. URL: `http://www.epa.gov/pesticides/about/overview_risk_assess.htm`.

[(USd] United States Environmental Protection Agency (USEPA). Standard operating procedures for residential pesticide exposure assessment. 2012. Office of Pesticide Programs Washington (DC).

[(US91] United States Environmental Protection Agency (USEPA). Guidelines for developmental toxicity risk assessment. *Federal Register*, 56:63798–63826, 1991.

[(US92] United States Environmental Protection Agency (USEPA). Guidelines for exposure assessment. *Federal Register*, 57(104):22888–22938, 1992.

[(US13a] United States Environmental Protection Agency (USEPA). Dose-Response Assessment, 2013. Accessed: 2013-09-26. URL: `http://epa.gov/riskassessment/dose-response.htm`.

[(US13b] United States Environmental Protection Agency (USEPA). Hazard Identification, 2013. Accessed: 2013-09-26. URL: `http://epa.gov/riskassessment/hazardous-identification.htm`.

[(US16] United States Environmental Protection Agency (USEPA). Label Review Manual, 2016. Accessed: 2016-11-30. URL: `https://www.epa.gov/pesticide-registration/label-review-manual`.

[(US21a] United States Environmental Protection Agency (USEPA). Occupational Pesticide Handler Exposure Data , 2021. Accessed: 2021-05-31. URL: `https://www.epa.gov/pesticide-science-and-assessing-pesticide-risks/occupational-pesticide-handler-exposure-data`.

[(US21b] United States Environmental Protection Agency (USEPA). Occupational Pesticide Post-application Exposure Data,

2021. Accessed: 2021-05-31. URL: `https://www.epa.gov/pesticide-science-and-assessing-pesticide-risks/occupational-pesticide-post-application-exposure`.

[Vai95] Harri Vainio. Molecular approaches in toxicology: change in perspective. *Journal of occupational and environmental medicine*, 37(1):14–18, 1995.

[VBGO11] Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45:1–67, 2011.

[VC93] Marco Vighi and Davide Calamari. Prediction of the environmental fate of chemicals. *Annali dell Istituto Superiore di Sanita*, 29:209–209, 1993.

[VdBTBVdL16] F Van den Berg, Aaldrik Tiktak, JJTI Boesten, and AMA Van der Linden. PEARL model for pesticide behaviour and emissions in soil-plant systems. Technical report, Statutory Research Tasks Unit for Nature & the Environment, 2016.

[vdBZY+12] Henk van den Berg, Morteza Zaim, Rajpal Singh Yadav, Agnes Soares, Birkinesh Ameneshewa, Abraham Mnzava, Jeffrey Hii, Aditya Prasad Dash, and Mikhail Ejov. Global trends in the use of insecticides to control vector-borne diseases. *Environmental Health Perspectives*, 120(4):577, 2012.

[VDM04] Jose Vieira, F Morgado Dias, and Alexandre Mota. Neuro-Fuzzy Aystems: A Survey. In *5th WSEAS NNA international conference on neural networks and applications, Udine, Italia*, pages 87–92, 2004.

[VDWBZVV09] Jacobus S Van Der Walt, AA Buitendag, Jan J Zaaiman, and Joey Jansen Van Vuuren. Community Living Lab as a Collaborative Innovation Environment. *Issues in Informing Science and Information Technology*, 6(1):421–436, 2009.

[VESJ02] Eric F Vermote, Nazmi Z El Saleous, and Christopher O Justice. Atmospheric correction of MODIS data in the visible to middle infrared: first results. *Remote Sensing of Environment*, 83(1):97–111, 2002.

[VFV05] A Belmonte Vega, A Garrido Frenich, and JL Martínez Vidal. Monitoring of pesticides in agricultural water and soil samples from andalusia by liquid

chromatography coupled to mass spectrometry. *Analytica Chimica Acta*, 538(1-2):117–127, 2005.

[VHS+91] H Vainio, E Heseltine, L Shuker, D McGregor, and C Partensky. Meeting report: Occupational exposures in insecticide application and some pesticides. *European Journal of Cancer and Clinical Oncology*, 27(3):284–289, 1991.

[VK15] Zs J Viharos and Krisztián Balázs Kis. Survey on Neuro-Fuzzy systems and their applications in technical diagnostics and measurement. *Measurement*, 67:126–136, 2015.

[vL55] Justus Freiherr von Liebig. *Die grundsätze der agricultur-chemie mit rücksicht auf die in England angestellten untersuchungen.* F. Vieweg und Sohn, 1855.

[VLdSW17] Mathieu Valcke, Marie-Eve Levasseur, Agnes Soares da Silva, and Catharina Wesseling. Pesticide exposures and chronic kidney disease of unknown etiology: an epidemiologic review. *Environmental Health*, 16(1):49, 2017.

[VLOR14] Darcy R VanDervort, Dina L López, Carlos M Orantes, and David S Rodríguez. Spatial Distribution of Unspecified chronic kidney disease in El Salvador by Crop Area cultivated and ambient temperature. *MEDICC review*, 16(2):32, 2014.

[Vol13] Lutz Volkmann. *Fundamente der Graphentheorie.* Springer-Verlag, 2013.

[VR02] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S.* Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[VRD09] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual.* CreateSpace, Scotts Valley, CA, 2009.

[VSKD97] E Vermote, N El Saleous, YJ Kaufman, and E Dutton. Data pre-processing: Stratospheric aerosol perturbing effect on the remote sensing of vegetation: Correction method for the composite NDVI after the Pinatubo eruption. *Remote Sensing Reviews*, 15(1-4):7–21, 1997.

[VTD+97] Eric F Vermote, Didier Tanré, Jean Luc Deuzé, Maurice Herman, and J-J Morcette. Second Simulation of the Satellite Signal in the Solar Spectrum, 6S: An Overview. *Geoscience and Remote Sensing, IEEE Transactions on*, 35(3):675–686, 1997.

[vTK17] Arjan van Timmeren and David V Keyson. Towards Sustainable Living. In *Living Labs*, pages 3–7. Springer, 2017.

[Wag04] Adam Wagstaff. *The Millennium Development Goals for Health : Rising to the Challenges*. World Bank Publications, 2004.

[Wal16a] Guido Walz. *Lexikon der Mathematik: Band 1: A bis Eif*. Springer-Verlag, 2016.

[Wal16b] Guido Walz. *Lexikon der Mathematik: Band 3: Inp bis Mon*. Springer-Verlag, 2016.

[WC11] Andrea Wiggins and Kevin Crowston. From Conservation to Crowdsourcing: A Typology of Citizen Science. In *System Sciences (HICSS), 2011 44th Hawaii International Conference on*, pages 1–10. IEEE, 2011.

[WC13] Yi Wu and Klarissa TT Chang. An empirical study of designing simplicity for mobile application interaction. In *Proceedings of the Nineteenth Americas Conference on Information Systems*, pages 1–8, 2013.

[WCB05] Catharina Wesseling, Marianela Corriols, and Viria Bravo. Acute pesticide poisoning and pesticide registration in Central America. *Toxicology and Applied Pharmacology*, 207(2 Suppl):697–705, September 2005. PMID: 16153991.

[WDB13] Daniel G Wright, Prasanta K Dey, and John G Brammer. A fuzzy levelised energy cost method for renewable energy technology assessment. *Energy Policy*, 2013.

[WDHE89] Kerstin Wiklund, J Dich, LE Holm, and G Eklund. Risk of cancer in pesticide applicators in Swedish agriculture. *British Journal of Industrial Medicine*, 46(11):809–814, 1989.

[WGB+24] Anne-Kathrin Wendell, Björn Guse, Katrin Bieger, Paul D Wagner, Jens Kiesel, Uta Ulrich, and Nicola Fohrer. A spatio-temporal analysis of environmental fate and transport processes of pesticides and their transformation products in agricultural landscapes dominated by subsurface drainage with SWAT+. *Science of The Total Environment*, page 173629, 2024.

[WHL08] Zhiyi Wang, M Hamalainen, and Zhangxi Lin. An Open Community Approach to Emergency Information Services during a Disaster. In *Information Science and Engineering, 2008. ISISE'08. International Symposium on*, volume 1, pages 649–654. IEEE, 2008.

[WHP22] Nicole Washuck, Mark Hanson, and Ryan Prosser. Yield to the data: some perspective on crop productivity and pesticides. *Pest Management Science*, 78(5):1765–1771, 2022.

[WHY09] Paul D Winchester, Jordan Huskins, and Jun Ying. Agrichemicals in surface water and birth defects in the United States. *Acta Paediatrica*, 98(4):664–669, 2009.

[WLK18] Helga Willer, Julia Lernoud, and Laura Kemper. The world of organic agriculture 2018: Summary. In *The World of Organic Agriculture. Statistics and Emerging Trends 2018*, pages 22–31. Research Institute of Organic Agriculture FiBL and IFOAM-Organics International, 2018.

[WM92] L-X Wang and Jerry M Mendel. Generating fuzzy rules by learning from examples. *IEEE Transactions on systems, man, and cybernetics*, 22(6):1414–1427, 1992.

[WM96] Frank Wania and Donald Mackay. Tracking the Distribution of Persistent Organic Pollutants. *Environmental Science & Technology*, 30(9):390A–396A, 1996.

[WMC10] Scott Weichenthal, Connie Moase, and Peter Chan. A review of pesticide exposure and cancer incidence in the agricultural health study cohort. *Environmental health perspectives*, 118(8):1117–1125, 2010.

[WRL+14] Pattarawan Watcharaanantapong, Roland K Roberts, Dayton M Lambert, James A Larson, Margarita Velandia, Burton C English, Roderick M Re-

jesus, and Chenggang Wang. Timing of precision agriculture technology adoption in US cotton production. *Precision Agriculture*, 15(4):427–446, 2014.

[WSGG12] Jin-Feng Wang, A Stein, Bin-Bo Gao, and Yong Ge. A review of spatial sampling. *Spatial Statistics*, 2:1–14, 2012.

[WWG+21] Thomas Weiber, Rolf Weiber, Sonja Gensler, Bernd Erichson, and Klaus Backhaus. *Multivariate Analysis: An Application-Oriented Introduction*. Springer, 2021.

[WZ11] Lucia Wegner and Gine Zwart. Who will feed the world? the production challenge. *Food Chain*, 1(2):187–205, 2011.

[XCC+21] Lei Xu, Nengcheng Chen, Zeqiang Chen, Chong Zhang, and Hongchu Yu. Spatiotemporal forecasting in earth system science: Methods, uncertainties, predictability and future directions. *Earth-Science Reviews*, 222:103828, 2021.

[Xie20] Yihui Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2020. R package version 1.30.

[XZCH20] Haiping Xiao, Zhenchao Zhang, Lanlan Chen, and Qimin He. An Improved Spatio-Temporal Kriging Interpolation Algorithm and Its Application in Slope. *IEEE Access*, 8:90718–90729, 2020.

[yAS18] Programa Salud Trabajo y Ambiente (SALTRA). Homepage of the Programa Salud, Trabajo y Ambiente (SALTRA), 2018. Accessed: 2018-03-10. URL: `http://www.saltra.una.ac.cr/`.

[YBMS13] Ming Ye, Jeremy Beach, Jonathan W Martin, and Ambikaipakan Senthilselvan. Occupational pesticide exposures and respiratory health. *International Journal of Environmental Research and Public Health*, 10(12):6442–6471, 2013.

[YDT+12] Boris Yatsalo, Vladimir Didenko, Alexander Tkachuk, Sergey Gritsyuk, Oleg Mirzeabasov, Valeria Slipenkaya, Alexey Babutski, Irina Pichugina, Terry Sullivan, and Igor Linkov. Multi-Criteria Spatial Decision Support System

DECERNS: Application to Land Use Planning. In *Societal impacts on information systems development and applications*, pages 255–273. IGI Global, 2012.

[Yeş20] Cafer Mert Yeşilkanat. Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm. *Chaos, Solitons & Fractals*, 140:110210, 2020.

[YH18] Jing Yang and Maogui Hu. Filling the missing data gaps of daily MODIS AOD using spatiotemporal interpolation. *Science of the Total Environment*, 633:677–683, 2018.

[You13] Dirk F Young. Pesticides in Flooded Applications Model (PFAM): Conceptualization, Development, Evaluation, and User Guide. *United States Environmental Protection Agency, Washington, DC*, 61, 2013.

[You16] Dirk F Young. Pesticide in Water Calculator User Manual for Versions 1.50 and 1.52. *United States Environmental Protection Agency, Washington, DC*, 2016.

[Zad65] Lotfi A Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, 1965.

[ZGP+18] Zixing Zhang, Jürgen Geiger, Jouni Pohjalainen, Amr El-Desoky Mousa, Wenyu Jin, and Björn Schuller. Deep Learning for Environmentally Robust Speech Recognition: An Overview of Recent Developments. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 9(5):1–28, 2018.

[ZGSG10] Matthew Zook, Mark Graham, Taylor Shelton, and Sean Gorman. Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake. *World Medical & Health Policy*, 2(2):7–33, 2010.

[Zha12] Shichao Zhang. Nearest neighbor selection for iteratively knn imputation. *Journal of Systems and Software*, 85(11):2541–2552, 2012.

[Zha15] Qin Zhang. Control of Precision Agriculture Production. *Precision Agriculture Technology for Crop Farming*, page 103, 2015.

[ZKPS+22] Johann G Zaller, Maren Kruse-Plaß, Ulrich Schlechtriemen, Edith Gruber, Maria Peer, Imran Nadeem, Herbert Formayer, Hans-Peter Hutter, and Lukas Landler. Pesticides in ambient air, influenced by surrounding land use and weather, pose a potential threat to biodiversity and humans. *Science of the Total Environment*, 838:156012, 2022.

[ZMBM20] Zhaoyu Zhai, José Fernán Martínez, Victoria Beltran, and Néstor Lucas Martínez. Decision support systems for agriculture 4.0: Survey and challenges. *Computers and Electronics in Agriculture*, 170:105256, 2020.

[Zoo14] Zooniverse. Zooniverse homepage, 2014. Accessed: 2014-04-01. URL: `https://www.zooniverse.org/`.

[ZP12] YY Zhao and YS Pei. Risk evaluation of groundwater pollution by pesticides in China: a short review. *Procedia Environmental Sciences*, 13:1739–1747, 2012.

[ZS06] Zhengyuan Zhu and Michael L Stein. Spatial sampling design for prediction with estimated parameters. *Journal of Agricultural, Biological and Environmental Statistics*, 11:24–44, 2006.

[ZSX20] Mo Zhang, Wenjiao Shi, and Ziwei Xu. Systematic comparison of five machine-learning models in classification and interpolation of soil particle size fractions using different transformed data. *Hydrology and Earth System Sciences*, 24(5):2505–2526, 2020.

[ZT13] Yong Zhou and Ling Tian. Applying Fuzzy Set Theory to Analysis Driving Maneuver. *Applied Mechanics and Materials*, 361:2244–2248, 2013.

List of Figures

1.1 Overview about the scientific main disciplines involved in this thesis: the content of the thesis lies in the intersection between ecotoxicology, mathematical modeling, and the LL approach (figure generated with *LibreOffice Draw*, image source: Bozhin Karaivanov and Ilham Wicaksono via Unsplash `https://unsplash.com/`, Herselman unpublished). 5

2.1 Map of the pilot region Bajo Lempa in El Salvador (figure generated with *QGIS*). 21
2.2 Global pesticide use in the agricultural sector [FD24a] (figure generated with *R*). 28
2.3 Fate of chemicals in the environment according to [Fen13] (figure generated with *LibreOffice Draw*). 29
2.4 Empty pesticide container disposed on a field in the Bajo Lempa region, El Salvador (source: M. Hieber-Ruiz) 35
2.5 Campesinos in the Bajo Lempa region, El Salvador, applying pesticides with manual hand pumps and without sufficient protective clothing (source: M. Hieber-Ruiz) 39
2.6 Campesinos in the Bajo Lempa region, El Salvador, applying fertilizer by hand on plants (source: M. Hieber-Ruiz) 40

3.1 Overview of the different steps in the US EPA risk assessment for pesticides and human health. Source:[(US13b] (figure generated with *draw.io*). 59

4.1 Schematic representation of a system according to [IK13] with system variables S_1 and S_2 , an inner relation between V_1 and V_2, and outer relations between the system and the environment as well as between the environment and the system. (figure generated with *LibreOffice Draw*). 66

4.2 Modeling cycle according to [BL05]. 1: Problem understanding , 2: Simplifying/ Structuring, 3: Mathematization, 4: Working mathematically, 5: Interpreting, 6: Validating (figure generated with *LibreOffice Draw*). 68

5.1 An example of an innovation cycle within an LL, according to Marlien Herselman, unpublished (figure generated with *LibreOffice Draw*). 79

7.1 Example of a function $u_{t_0}(x, y)$ representing the unknown spatial distribution of a parameter of interest at time point t_0 together with a underlying sampling grid with possible sampling points (figure generated with *GNU Octave*). . . . 110

7.2 Example for the visualization of discrete point samples for a parameter of interest in a map at a time point t_0 (figure generated with *GNU Octave*). . . . 111

7.3 Example for the interpolation based on scattered discrete point samples for a parameter of interest in a map at a fixed time point t_0 (figure generated with *GNU Octave*). 115

7.4 Example for the visualization of discrete point samples for a parameter of interest in a map at a time point t_1 together with a spatial interpolation (figure generated with *GNU Octave*). 116

7.5 Visualization of discrete point samples generated by temporal interpolation with convex combination at time point $t_{0.6}$, together with a spatio-temporal interpolation (figure generated with *GNU Octave*). 118

7.6 Graph of the error function e_{f,t_0} from two different angles of view (figure generated with *GNU Octave*). 119

7.7 Graph of the density function d_i for a measuring point i at location $(0, 0)$ with s = 1 and s = 3 (figure generated with *GNU Octave*). 121

7.8 Graph of the aggregated spatial density function d_{spat,t_k} for all measuring points at location (x_i, y_i) with $i = 1, ..., n$, s = 1 and s = 3 (figure generated with *GNU Octave*). 123

7.9 Example for a temporal quality function for a measurement at time point $t_k = 1$ and two different slope factors $s_t = 1$ and $s_t = 3$ resulting in $q_{temp,t_k=1,s_t=1}(t) = \frac{1}{1+\left(\frac{t-1}{1}\right)^2}$ and $q_{temp,t_k=1,s_t=3}(t) = \frac{1}{1+\left(\frac{t-1}{3}\right)^2}$ (figure generated with *GNU Octave*). 124

8.1 Stakeholder meeting in the Bajo Lempa region (source: Mäggi Hieber Ruiz). . 130

9.1 Possible risk mitigation strategies fitting to the described LL approach (figure generated with *R*). 138

10.1 Typical reflectance rates of healthy green and brown vegetation (figure generated with *LibreOffice Draw*; image source `https://www.flaticon.com/free-icons/tree`, *Tree icons created by Freepik - Flaticon*). 143

10.2 An example of high-tech VRT-systems on a tractor. Figure a) represents the map-based approach, and figure b) the sensor-based approach according to [GAT+11] (figure generated with *LibreOffice Draw*, image source: `https://freesvg.org/farm-tractor-with-planter-vector-graphics` *Free SVG*, `https://openclipart.org/detail/6024/cartoon-computer-and-desktop` *OpenClipart* created by *DTRave*). 146

10.3 Spectral satellite images from *Landsat 8* of a part of Central America taken on August 29th 2015, a) Band 4, visible red, wavelength 0.64 - 0.67 μm, resolution 30 m b) Band 5, near-infrared, wavelength 0.85 - 0.88 μm, resolution 30 m. For better visualization, only values in the range of 0 - 98% of the maximum value are visualized. Values are expressed in the unit provided by the NASA called *digital numbers*. (figure generated with *QGIS*, Level 1 geoTIFF file, *Landsat Scene Identifier LC80190512015241LGN00*, source: [Exp15]). 148

10.4 True-color satellite images from *Landsat 8* of a part of Central America, taken at August 29th 2015. (figure generated with *QGIS*, Level 1 geoTIFF file, *Landsat Scene Identifier LC80190512015241LGN00*, source: [Exp15]). 149

10.5 QA band categorized into areas with *clouds*, *maybe clouds*, and *no clouds*, taken on August 29th 2015. (figure generated with *QGIS*, Level 1 geoTIFF file, *Landsat Scene Identifier LC80190512015241LGN00*, source: [Exp15]). . . 153

10.6 Flowchart of the creation process for NDVI maps with *GRASS GIS* based on satellite images (figure generated with *draw.io*). 154

10.7 Temporal change of NDVI values for corn plants within a crop season and between fertilized and unfertilized plants (figure generated with *R*, source: [Zha15]). 155

10.8 NDVI values calculated for a scene over El Salvador and a land-use map representing areas planted with sugarcane (figure generated with *QGIS*, Level 1 geoTIFF file, *Landsat Scene Identifier LC80190512015241LGN00*, source: [Exp15]; land-use map: *Instituto Nacional de Salud El Salvador*). 157

10.9 NDVI map for a scene over El Salvador and a land-use map representing areas planted with sugarcane with a high zoom level (figure generated with *QGIS*, Level 1 geoTIFF file, *Landsat Scene Identifier LC80190512015241LGN00*, source: [Exp15]; land-use map: *Instituto Nacional de Salud El Salvador*). 158

11.1 Created world map (figure generated with *QGIS* and *GRASS GIS*, sources: [Pro11] and [Bur11]). 164
11.2 Sample map representing the proportion per cell of the area harvested with wheat (figure generated with *QGIS*, source:[MRF08]). 165
11.3 World map visualizing the applied amount [*a.i. per cell ha*] for herbicides (figure generated with *QGIS*). 177
11.4 World map visualizing the applied amount [*a.i. per cell ha*] for fungicides (figure generated with *QGIS*). 178
11.5 World map representing the deviation value $\frac{log(calc)}{log(ref)}$ for herbicides (figure generated with *QGIS*). 180
11.6 World map representing the deviation value $\frac{log(calc)}{log(ref)}$ for fungicides (figure generated with *QGIS*). 181
11.7 Regression between calculated amount and reference values for herbicides with n = 101, adjusted $R^2 = 0.4864$, intercept = 0.377, slope = 0.885, p = $3.25*10^{-16}$ (figure generated with *R*). 182
11.8 Regression between calculated amount and reference values for fungicides with n = 98, adjusted $R^2 = 0.324$, intercept = 0.34, slope = 0.848, p = $5.82*10^{-10}$ (figure generated with *R*). 183
11.9 Regression between calculated amount and reference values for herbicides, countries grouped into continents or regions (figure generated with *R*). 184
11.10 Regression between calculated amount and reference values for fungicides, countries grouped into continents or regions (figure generated with *R*). 185

12.1 Different sources of data to create risk maps with citizen sciences methods (figure generated with *draw.io*). 192
12.2 Example for the use of an open-source app to sample temporal and spatial information of pesticide application observation (screenshot of a *shiny* app). . 193

13.1 Sample spatial visualization of pesticide concentrations in the environment with a underlying street map (figure generated with *GNU Octave*, source of the map: `https://www.openstreetmap.org`). 196
13.2 Exemplary weighted digraph with 5 nodes and 7 edges (figure generated with *Libre Office Draw.* 200
13.3 Graph network based on a route map from OSM with a overlying air concentration layer for a community in El Salvador (figure generated with *GNU Octave*, source of the map: `https://www.openstreetmap.org`). 201
13.4 Route map from OSM with a path network for a community in El Salvador and possible routes γ_1 (green path) and γ_2 (blue path) (figure generated with *GNU Octave*, source of the map: `https://www.openstreetmap.org`). 202

16.1 Graph of a sample membership function $\mu(x) = \frac{1}{1+e^{-0.3\cdot(x-20)}}$ (figure generated with *GNU Octave*). 219
16.2 A sample map with fuzzified values (figure generated with *GNU Octave*). . . 220
16.3 Graph of three membership functions, representing the terms low, medium, and high (figure generated with *GNU Octave*). 222
16.4 Risk and the TER of a certain compound: crisp threshold according to [Com02] (red) and in the sense of fuzzy set theory (blue) (figure generated with *GNU Octave*). 223
16.5 Height of a fuzzy set A (figure generated with *GNU Octave*). 225
16.6 Graph of a normalized fuzzy set A_n with a height of 1 (figure generated with *GNU Octave*). 226
16.7 Example for a subset B of the set A: $B \subseteq A$ (figure generated with *GNU Octave*). 229
16.8 Example for the union of two fuzzy sets A and B by using $\perp_{min}$ (figure generated with *GNU Octave*). 231
16.9 Example for the intersection of two fuzzy sets A and B (figure generated with *GNU Octave*). 232
16.10 Bounded sum of two fuzzy sets A and B (figure generated with *GNU Octave*). 233
16.11 Schematic function of a fuzzy controller according to [LEE90] (figure generated with *GNU Octave*). 238
16.12 Graphs of the membership functions for the parameter Distance D (figure generated with *GNU Octave*). 246

16.13Graphs of the membership functions for the parameter Distance F (figure generated with *GNU Octave*). 246
16.14Graphs of the membership functions for the output parameter S (figure generated with *GNU Octave*). 247
16.15Graph of the output membership function ν_1 for the rule R_1 (figure generated with *GNU Octave*). 249
16.16Graph of the output membership functions ν_1, ν_2 and ν_3 for the rules R_1, R_2 and R_3 (figure generated with *GNU Octave*). 250
16.17Graph of the combined output membership function ν for the rules R_1, R_2, and R_3 (figure generated with *GNU Octave*). 250
16.18Example of a simple neuron with weighted edges (figure generated with *LibreOffice Draw* according to [Roj96]). 254
16.19Example of a continuous differentiable activation function, here sigmoid function (figure generated with *GNU Octave*). 256
16.20Signal flow in a *perceptron* according to [Hay04]. $x_1, ..., x_p$ represent input signals, Θ represents a threshold value, $v = \sum_{i=1}^{p} x_i \cdot w_i - \Theta$, the output after passing the integration function k; f, the activation function; $w_1, ..., w_p$, the edge weights; and y, the output of the *perceptron* (figure generated with *LibreOffice Draw*). 257
16.21An example of a simple neural net with two input units $U_{in} = \{n_1, n_2\}$, two hidden units $U_{hidden} = \{n_3, n_4\}$, and one output unit $U_{out} = \{n_5\}$ that performs a mapping from two input variables x_1 and x_2 to a output variable y. The edges between the neurons are directed, visualized through the arrows at the end of the edges, and weighted, as symbolized by the weights $w_{i,j}$ (figure generated with *LibreOffice Draw).* 263
16.22Schematic representation of a) *cooperative* [VDM04], b) *concurrent* [BKKN03], and c) a *hybrid neuro-fuzzy system* [EDDSM15]. The visualized *hybrid neuro-fuzzy system* is an ANFIS with a network architecture in which $A_1, ..., B_2$ represent fuzzy sets, $\perp$ t-norms of the antecedent, and the consequent part of a rule, N a normalizer, and Σ the sum function for calculating the output f and a back-propagated error signal (figure generated with *LibreOffice Draw*). . . . 267
16.23Representation of an ANFIS with two inputs, two rules, and five layers $L1$, ... , $L5$ 279

16.24Schematic representation of a HyFIS system according to [KK99] (figure generated with *LibreOffice Draw*). 291
16.25Representation of the *membership function tuning module* of a HyFIS system with two inputs, three rules, and five layers $L1$, ... , $L5$ [KK99] (figure generated with *LibreOffice Draw*). 294

17.1 Example for the information flow between different stakeholders and a centralized data center in the LL approach (figure generated with *LibreOffice Draw*, image source `https://www.flaticon.com`, creators: *Freepik, Gajah Mada, Eucalyp, Karyative, fajarestuu*). 299

18.1 Flowchart of an SDSS in the proposed framework (figure generated with *draw.io*).306

B.1 Map visualizing the crop responsible for the highest applied fungicide amount per raster cell (figure generated with *QGIS*). 319
B.2 Map visualizing the crop responsible for the highest applied herbicide amount per raster cell (figure generated with *QGIS*). 320

List of Tables

2.1 Potential risk factors for CKD and uptake routes 14
2.2 Stages of CKD, classification and GFR according to [Amm20] 16
2.3 Pesticide use and agricultural area, sources: a [GDKW11] for the year 2007, b [otEU16] for the year 2014, c [FD16] for the year 2014, values without a source are calculated from values given in the table 28

3.1 Classification of pesticides according to the WHO [Org10] 55

8.1 Selected LL stakeholders in the LLinES application 131

11.1 Calculated applied fungicide and herbicide amount in *t a.i.* and reference values [FD24b] for 2000, divided into regions 176
11.2 Statistical values of the regression between calculated amount and reference values for herbicides, countries grouped into continents or regions. 185
11.3 Statistical values of the regression between calculated amount and reference values for fungicides, with countries grouped into continents or regions. 186

12.1 Objects and related items for which information are collected to create risk maps 191

16.1 Often used t-norms and t-conorms according to [NKK94] 230
16.2 Truth table of the implication $p \Rightarrow q$ for classical two-valued logic 235
16.3 Some often used fuzzy implication operators [CCBC04] 236
16.4 Advantages and disadvantages of neural nets and fuzzy systems according to [BKKN03] 265

A.1 Equations for the calculation of application rates 316

C.1 Tasks in the LL and possible open-source software solutions 322

Erklärung

Ich versichere, dass ich die eingereichte Dissertation selbstständig verfasst habe. Alle von mir für die Arbeit benutzten Hilfsmittel und Quellen wurden, mit Außnahme der nachfolgend aufgelisteten KI basierten Hilfsmittel, in der Arbeit angegeben.

Es sind keine weiteren Autoren und Autorinnen zu nennen, da keine anderen Personen Teile dieser Dissertation verfasst haben.

Ich habe entgeltliche Hilfe in Form der Studi-Lektor GmbH in Form eines englischen Lektorats in Anspruch genommen. Dabei wurde die Rechtschreibung, Interpunktion, Grammatik geprüft sowie sprachliche Unebenheiten geglätet.

Ich habe die Dissertation nicht in gleicher oder ähnlicher Form als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüufung im In- oder Ausland eingereicht. Ich habe keine gleiche oder eine andere Abhandlung in einem anderen Fachbereich oder einer anderen wissenschaftlichen Hochschule als Dissertation eingereicht. Die in Kapitel 11 beschriebene Methode zur Erstellung von Pestizidapplikationskarten wurde, exklusive des Modells zur Anpassung der Applikationsraten für Afrikanische Staaten, in meiner im Jahr 2011 veröffentlichten Diplomarbeit "Abschätzung der global eingesetzten Insektizidmengen mithilfe eines geografischen Informationssystems" für eine andere Stoffgruppe (Insektizide) angewendet.

Mir ist bewusst, dass ein Verstoß gegen einen der vorgenannten Punkte den Entzug des Doktortitels bedeuten und gegebenenfalls auch weitere rechtliche Konsequenzen haben kann.

Landau in der Pfalz, den 19.05.2025

Jörg Rapp
Institut für Mathematik
FB Natur- und Umweltwissenschaften
RPTU Kaiserslautern-Landau
Fortstraße 7
D-76829 Landau

Danksagung

Danken möchte ich in erster Linie meinen Betreuern Bert Niehaus und Ralf Schulz, die es mir ermöglichten eine wissenschaftliche Laufbahn einzuschlagen und mich stets mit Rat und Tat unterstützt haben. Der fortlaufende Input mit neuen Ideen hat diese Arbeit erst zu dem gemacht was sie nun ist. Vielen Dank euch beiden !

Danke Bert, dass Du mit deiner herzlichen und liebevollen Art auch in nicht ganz so leichten Zeiten den Glauben an das Fertigstellen dieser Arbeit nicht verloren hast und mich immer weiter unterstützt und motiviert hast.

Vielen Dank auch an das gesamte Institut für Mathematik und das Rechenzentrum in Landau. Lachen konnte man mit euch immer, was ungemein geholfen hat.

Auch vielen Dank an das gesammte LLinES Team, v.a. Alex, Roberto und Edgar, das es mir ermöglicht hat im Rahmen des Projekts auch andere Winkel der Welt und Lebenseinstellungen kennenzulernen.

Den allergrössten Dank habe ich meiner Familie und insbesondere meinen Eltern auszusprechen. Neben all der Unterstützung während des Studiums und der Promotion habt ihr mir ein Fundament mitgegeben, durch den diese Arbeit erst ermöglicht wurde: "Immer weiter und niemals aufgeben". Vielen Dank Mama, Papa, Jan, Silvie und Laura !

Widmen möchte ich diese Arbeit meinen Eltern und insbesondere meiner Mutter, welche die Fertigstellung dieser Arbeit leider nicht mehr miterleben konnte. Vielen Dank für Alles !

Lebenslauf

Schulausbildung

2000	Allgemeine Hochschulreife
1991-2000	Friedrich Abel Gymnasium, Vaihingen an der Enz
1987-1991	Grundschule Vaihingen an der Enz

Studium

Seit 2013	**Doktorand**, Rheinland-Pfälzischen Technischen Universität Kaiserslautern-Landau bzw Universität Koblenz-Landau (bis 2023)
2012	**Diplom in Umweltwissenschaften**, Diplomarbeit: "Abschätzung der global eingesetzten Insektizidmengen mithilfe eines geografischen Informationssystems"
2005-2012	**Umweltwissenschaften, Diplomstudiengang**, Universität Koblenz-Landau
2001-2004	**Elektrotechnik, Diplomstudiengang**, Universität Ulm

Landau, 08.09.2025: Jörg Rapp