Rheinland-Pfälzische Technische Universität Kaiserslautern Landau

Vom Fachbereich Mathematik der Rheinland-Pfälzischen Technischen Universität Kaiserslautern-Landau zur Verleihung des akademischen Grades Doktor der Naturwissenschaften (Doctor rerum naturalium, Dr. rer. nat.) genehmigte Dissertation

Forecasting Index Insurance Payouts for Agricultural Risk Using Singular Spectrum Analysis

Rana Amani Desenaldo

Gutachter:
Prof. Dr. Jörn Saß
Prof. Dr. Peter Ruckdeschel

Datum der Disputation: 12. März 2026

DE-386

Abstract

The READI Actuaries Science Applied Research Program in Indonesia is looking for potential research related to risk estimation using climate information along with actuarial assumptions and methods. The monthly accumulated precipitation value is one of the climate information recorded by the Meteorology, Climatology, and Geophysical Agency of Indonesia (BMKG in Indonesian language). This weather factor can be used to calculate the financial risks of paddy crops in all provinces of Indonesia.

The first stage of this calculation is forecasting the precipitation values in 2016-2017 from the data up to 2015 using Singular Spectrum Analysis (SSA) both Univariate and Multivariate accordingly, considering the three areas of calculations: Province, Region, and Country. The second step is to convert these rainfall values to calculate payouts based on several linear index insurance models, with 6 Million IDR per hectare per planting season as the maximum indemnity. The procedure is based on additional analyses, such as comparisons between areas of calculations, between method combinations of the SSA itself, and between various missing data handling methods, are also included in this research. A simulation study based on different rainfall distributions is also included.

The major findings of this research are the better performance of some SSA method combinations and the best linear index insurance model to reproduce the payouts. There are many factors to be considered when discussing good policies and the benefits of such insurance against agricultural risk.

Zusammenfassung

Das READI Actuaries Science Applied Research Program in Indonesien sucht nach potenziellen Forschungsprojekten im Zusammenhang mit der Risikoschätzung unter Verwendung von Klimainformationen sowie versicherungsmathematischen Annahmen und Methoden. Der monatlich akkumulierte Niederschlagswert ist eine der Klimainformationen, die von der indonesischen Behörde für Meteorologie, Klimatologie und Geophysik (BMKG in indonesische Sprache) erfasst werden. Dieser Wetterfaktor kann zur Berechnung der finanziellen Risiken von Reisanbauflächen in allen Provinzen Indonesiens herangezogen werden.

Der erste Schritt dieser Berechnung besteht darin, die Niederschlagswerte für 2016–2017 anhand der Daten bis 2015 unter Verwendung der univariaten und multivariaten Singular-Spektrum-Analyse (SSA) zu prognostizieren, wobei drei Berechnungsbereiche berücksichtigt werden: Provinz, Region und Land. Der zweite Schritt besteht darin, diese Niederschlagswerte umzurechnen, um die Auszahlungen auf der Grundlage mehrerer linearer Indexversicherungsmodelle zu berechnen, wobei 6 Millionen IDR pro Hektar und Anbausaison als maximale Entschädigung gelten. Das Vergehen basiert auf zusätzlichen Analysen, wie z. B. Vergleiche zwischen Berechnungsbereichen, zwischen Methodenkombinationen der SSA selbst und zwischen verschiedenen Methoden zum Umgang mit fehlenden Daten. Auch eine Simulationsstudie basierend auf verschiedenen Niederschlagsverteilungen wird durchgeführt.

Die wichtigsten Ergebnisse dieser Studie sind die bessere Vorhersagekraft einiger SSA-Methoden und die Wahl des besten linearen Indexversicherungsmodells zur Reproduktion der Auszahlungen. Bei der Diskussion über gute Policen und die Vorteile einer solchen Versicherung gegen landwirtschaftliche Risiken sind viele Faktoren zu berücksichtigen.

Acknowledgements

First of all, I would like to thank Prof. Dr. Jörn Sass for his continuous guidance during my time as one of his PhD students. Our regular meetings always gave me new insights to improve my work, and the way that he believes in me really helped me in focusing on finishing my work. It was a fun experience, doing this research with Jörn. He provided me with meaningful feedbacks and valuable ideas, helping me to better this work in each and every way. Jörn is also a caring, professional, and dedicated person in the field, and I am honored to be able to work with him for the past three years.

Secondly, I would also like to thank Prof. Dr. Peter Ruckdeschel from the University of Oldenburg who has given his time and dedication to review my work. His thorough and speedy feedback had really helped with the process. Our interactions throughout the process had been short but meaningful, and I am honored to have shared this experience with him.

Thirdly, I would also like to thank my current and former colleagues in RPTU Kaiserslautern-Landau and Fraunhofer ITWM who have been giving me valuable inputs for this research. They have been the ones witnessing the progress of this research from the very beginning. Special thanks to Indira, Kezang, and Ajla for their strong support, tons of help, and warm friendships during this journey.

Lastly, many thanks (more like infinite) to my family: my sister Rania, my mother Tanti, and my father Erith, who have been very supportive with every aspect of my life. They gave me time, advice, comfort, and happy moments whenever I need them. They have helped me go through so many things. This thesis, the joy and pride that came with it, I also share with them.

And a special shoutout to all my online friends who have constantly cheering me on, especially for this journey. It is nice to always have someone to talk to, no matter the time and location.

Contents

1 Introduction

Motivation and Research Goals

Rice is one of the leading food consumed in Indonesia. According to [Lou10], almost 97% people in Indonesia eat rice as their main staple food, which can be up to three times per day. Hence, paddy fields are significant in sustaining this need. In 2008, a concept called *Periode Musim Tanam Padi* (PMTP) or Paddy Rice Planting Season Period was introduced in Indonesia. According to [Sum06], using this concept, the government can figure out everything related to paddy rice production per province, even per city, easier and faster.

This concept divides paddy rice planting seasons into three periods based on the weather conditions. Of course, paddy rice can be planted at any time and normally can be harvested around four months after being planted. Depending on the weather conditions, the amount of paddy rice being harvested, however, can differ. Farmers do plant paddy rice all year long, but they adjust the amount based on water availability. Hence, this PMTP concept divides the planting season into three, which are:

a) Primary season has the most amount of harvested paddy. This season occurs from November to March, and the harvest period is from February until June.
b) Gadu Season has good quality paddy, but less than the amount harvested in Primary Season. This season occurs from April to July, and the harvest period is from July until October.
c) Draught season is the season with the least amount of harvested paddy. This season occurs from August to October, and the harvest period is from November until January. However, due to the unpromising quantity, farmers usually plant something else during this time.

The water availability that determines the planting seasons is observed from rainfall. In [MRY19], the authors discuss how rainfall values affect paddy rice productivity in a city in Indonesia. However, according to this paper, rainfall doesn't affect paddy rice that much—only by around 4.429%. The remaining 95.571% is affected by other factors. This research is done by seeing annual data from 2009-2013, which is a short amount of time. The effect might not be huge for the productivity, but rainfall is still an important weather factor to be taken into account when it comes to paddy rice crops because they need around 200 mm of rainfall per month according to [Pra+20] and [MRY19]. More than that, paddy crops will most likely be damaged. However, less rainfall is also not good because the paddy rice will not be able to grow properly without enough water.

According to [Ind23], agricultural insurance is given to the farmers in order to protect them from unforeseen risks. This agricultural insurance covers all farming activities, including crops, horticulture, and plantation, and covers all possible losses for crop yields due to natural disasters, pests, climate changes, and possibly others. With the global climate change that also affects Indonesia, there is a possibility that losses may increase. At the moment, the only insurance company that has this agricultural insurance product is PT Asuransi Jasa Indonesia, or simply known as Jasindo. This company is owned by the government and was established in 1973.

The agricultural insurance product owned by Jasindo is called Asuransi Usaha Tani Padi (AUTP), which is translated to Paddy Farming Business Insurance. Based on [Ind18], the sum insured for this product is 6 Millions IDR per hectare (around 340 EUR), and the premium is 180,000 IDR (around 10 EUR), with 80% of the value being covered by the government, so the farmers only need to pay 36,000 IDR (around 2 EUR). However, only maximum two hectares of paddy field per farmer can be covered using this insurance. The farmers can file for claims when their crops fulfill the following conditions:

a) The crops have been planted for at least 10 days,
b) The crops have to be older than 30 days old (direct-seeded paddy rice),
c) The damage intensity of the crops is more than 75%, and
d) More than 75% area of each field plot area is damaged.

The current statistics that they have published in [Keu23] is that in 2015-2018, they managed to cover 2.5 Millions hectare of paddy field out of 3.5 Millions hectare of paddy field that became their target, which makes it 72.50% of the target. After that, until July 2019, only 375,278.28 hectare out of 1 Million hectare were covered, which is 37.53% of the target. After July 2019, the insurance company paid around 10.9 Billions IDR (around 620,000 EUR) to cover 1,824.49 hectare of paddy field that were claimed to be damaged. Furthermore, in [Ard23], it is explained that unfortunately, most paddy rice fields are still not covered by AUTP. Only 3-6% of paddy rice fields are already insured by this program. In 2023, Jasindo has finished claims in the amounts of 695 Millions IDR (around 40,800 EUR) to cover 115.7 hectare land. They admitted to still have around 4,861 hectare of claims to finish.

How the payouts for AUTP are calculated is unknown, but in theory, payouts *can* be calculated using weather factors. To begin with, by setting limits and thresholds on the monthly accumulated precipitation value according to the needs of paddy crops, we can set the number of payouts for each rainfall value. These monthly rainfall values can also be used to forecast not only the upcoming rainfall values, but also the payout values.

As a start, it is important to observe the rain situation in Indonesia. According to former researches, especially in [APD23], climate changes are correlated to time series analysis. One of the methods that are often used is Singular Spectrum Analysis, also known as SSA. This method is a nonparametric spectral estimation method that combines time series analysis, multivariate statistics, multivariate geometry, dynamical systems and signal processing. With a lot of various mathematical methods taken into account, the provided results will help in analyzing, forecasting, and also in decision making. In addition, according to [REA], the READI Actuaris Science Applied Research Program in Indonesia is looking for potential research related to risk estimation using climate information with actuarial assumptions and methods. This statement strengthens our motivation to forecast the rainfall values in Indonesia, and then use the results to illustrate the upcoming payout values to be prepared by the insurance company.

Related Literature

In [APD23], the authors calculated the payout forecast of agricultural insurance for the upcoming two years. The calculation was done based on coffee crops using Univariate Singular Spectrum Analysis (USSA) for a Linear Index Insurance Model. However, only one linear index insurance model is included in this research. The SSA method also only includes forecasting by the recurrent method with row-formed trajectory matrix with $L = N/2$. In order to decide which weather stations to use, they used Principal Component Analysis (PCA). This paper also uses USSA with Linear Index Insurance Model, but with more varieties.

More SSA calculations were done in [MRY19]. They compare the results of USSA and MSSA (Multivariate Singular Spectrum Analysis). Here, they use four combinations of MSSA: HMSSA-R, HMSSA-V, VMSSA-R, and VMSSA-V. The first letter means Horizontal or Vertical, which is the trajectory matrix stack form. The last letter shows the forecasting method, Recurrent or Vector. The results show that MSSA outperforms SSA, by an accuracy comparison using MSE. It seems like the highest accuracy is shown by HMSSA-R (Horizontal/Row MSSA Recurrent). In addition, they also use 1, 5, and 10 steps ahead forecast throughout the simulations.

A comparison between SSA using vector forecast and recurrent forecast is done in [Gho+17]. The comparison was done through a lot of tests such as RMSE, MAE, impact of data distribu-

tions, stationarity and non-stationarity, frequency of data, series length, DC metric, and Henon series simulation. The calculation was done not only through simulations, but also through real data. The results show that SSA-V (Vector) is on average slightly better than SSA-R (Recurrent).

Data Source

The rainfall dataset for our calculations is obtained from Meteorology, Climatology, and Geophysical Agency of Indonesia (*Badan Meteorologi, Klimatologi, dan Geofisika Indonesia* – BMKG). One meteorological station from each province is chosen randomly, so we have 34 stations in total from 34 provinces as per 2021. The obtained data are daily, but for the calculation, they are accumulated to be monthly. The duration of the data is from 2000-2015 (16 year, 192 months, 6575 days) with an additional dataset from 2016-2017 (2 years, 24 months, 731 days) for forecast analysis. There are a lot of missing values in the database, which will be discussed further in Chapter 2.

Outline

The structure of this thesis is as follows. Firstly, in Chapter 2, we discuss about how the missing data values are handled. There are several imputation techniques to be considered, both single-value imputation and multiple imputation, along with their advantages and disadvantages. The imputed datasets are compared and explored before the calculation is started.

In Chapter 3, we discuss about Data Grouping, specifically using Weighted Values and Clustering. Since the calculation will be done in various levels of areas (Province, Region, Cluster, and Country), obtaining the weights for each province is necessary to group the data together so that one series can be formed. We only use one method to obtain the weights, but there are several calculations needed to be done due to the various imputation methods we use in the previous chapter. Clustering is also necessary to group the provinces for the Cluster calculation. We can also see the series for each level of areas after the grouping is conducted.

Chapter 4 has the most time-series-related analysis. We start by introducing Singular Spectrum Analysis (SSA), both Univariate and Multivariate, and then we proceed to discuss the various error tests and compare all the possible method combinations. Several combinations of the SSA methods, combined with the former missing data handling methods, produce different values for the error measures that need to be compared and analyzed. The most effective method(s), the error measure comparisons, and the rainfall forecast results can also be seen in that chapter.

It is finally time to discuss about the actuarial aspect in Chapter 5, where the discussion is about the payouts obtained from the rainfall values, both from the data and from the forecast. Several linear index insurance models are used to obtain different payout scenarios, and further analysis is done to determine which scenario with which methods of missing data handling and SSA provides the best insurance design.

After all the calculations have been done, Chapter 6 discusses SSA and Payout Conversion using simulated rainfall datasets. The simulation considers various distributions to generate the data and these distributions will help with the comparison between method combinations. The results for every distribution are discussed thoroughly.

Lastly, all the key points of the calculations and the analysis will be put together in our conclusion in Chapter 7.

2 Missing Data

According to [Kan13], missing data, or missing values, is the data that is not stored in the observation. This problem is very common in almost all research and unfortunately can have a significant effect to the results. Firstly, the missing data can reduce statistical power that might end up creating a difference when it comes to hypothesis testing. Secondly, the estimation of parameters can also be biased because of the missing data. Thirdly, the missing data can cause the data samples to be less representative, which is a huge problem when the sample size itself is already small. Lastly, the analysis can get more complicated with the absence of data.

When it comes to the types of missing data, there are three types: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Normally, one would not know the type of their missing data, however, the most realistic assumption would be MAR. Data are considered to be missing at random when the data are not related to the specific missing values that are expected to be obtained, but the probability depends on the set of observed responses. This is unlike MCAR where the probability is not dependent to the observed responses.

2.1 Data Exploration

In Table 1, we have a summary of each province's daily rainfall values without excluding the missing values (raw data can be seen in Appendix A). Most provinces have a lot of zero values, which results in them having minimum, first quartile, and median equal to 0. This is, however, not the case for all provinces because there are some median values that are larger than 0. We can see a huge gap between the third quartile and the maximum values, showing a lot of outliers as observed in Figure 2a. All the maximum values show extreme rain, showing how rainy it can be even if it is only for a day. We can also see the number of missing values for each province in the NA's column.

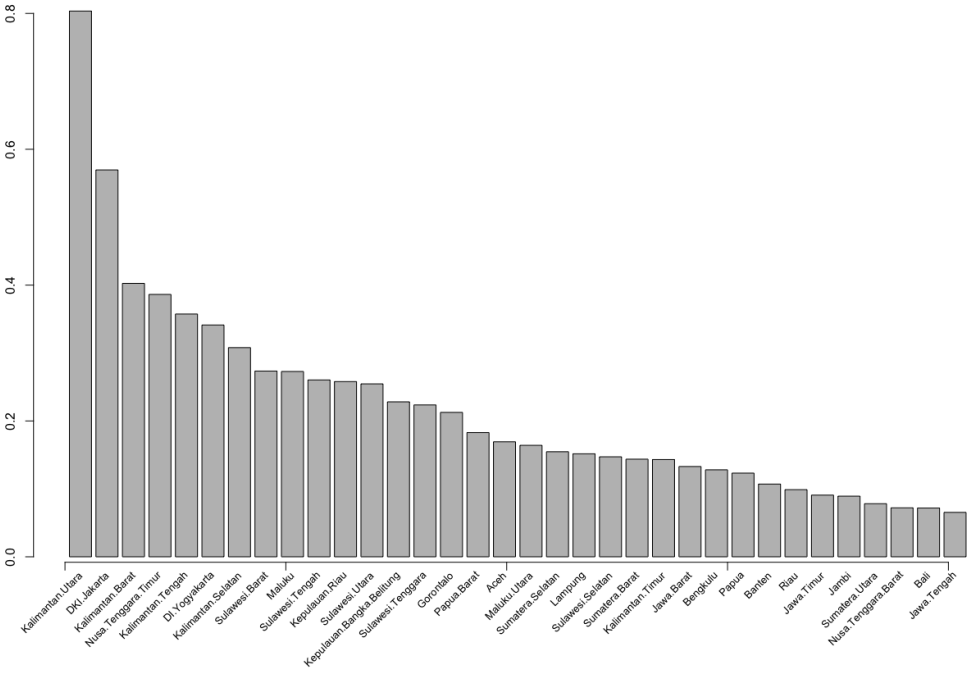


Figure 1: Empty Data Percentage for Each Province

The obtained datasets unfortunately but unsurprisingly have a lot of missing values. In total, we have 6575 days of data, which means 6575 values for each province. Figure 1 shows the percentage of missing values from the entire datasets for each province, ordered from the highest percentage to the lowest percentage. The province with the highest number of missing data is Kalimantan Utara, with 80.35% missing values (5283 missing days). However, this is due to Kalimantan Utara being a new province, barely established in 25 October 2012. Before becoming its own province, Kalimantan Utara used to be a part of Kalimantan Timur, which only has 14.31% missing values (941 missing days).

Table 1: Daily Rainfall Data Summary for Each Province (with missing values)

Province	**Min**	**Q1**	**Median**	**Mean**	**Q3**	**Max**	**NA's**
Aceh	0	0.0	0.0	4.928	3.3	188.8	1113
Sumatera Utara	0	0.0	0.2	7.411	6.8	170.8	514
Sumatera Barat	0	0.0	1.7	13.346	14.4	470.0	945
Riau	0	0.0	0.4	7.611	7.4	201.5	650
Jambi	0	0.0	0.2	6.697	5.9	140.4	587
Sumatera Selatan	0	0.0	0.6	8.306	9.0	214.1	1017
Bengkulu	0	0.0	1.0	10.578	11.0	236.0	842
Lampung	0	0.0	0.0	5.841	5.0	204.9	997
Kepulauan Bangka Belitung	0	0.0	2.3	10.056	13.3	196.1	1500
Kepulauan Riau	0	0.0	0.5	7.732	6.9	279.5	1696
DKI Jakarta	0	0.0	0.2	8.425	8.7	305.0	3745
Jawa Barat	0	0.0	0.8	7.122	8.3	122.9	874
Jawa Tengah	0	0.0	0.0	6.212	5.0	170.4	430
DI Yogyakarta	0	0.0	0.0	6.497	5.0	364.1	2244
Jawa Timur	0	0.0	0.0	5.006	3.0	142.5	597
Banten	0	0.0	0.0	4.973	3.0	316.3	704
Bali	0	0.0	0.0	5.423	2.9	161.1	472
Nusa Tenggara Barat	0	0.0	0.0	4.205	1.0	218.0	474
Nusa Tenggara Timur	0	0.0	0.0	3.664	1.2	157.9	2540
Kalimantan Barat	0	0.0	1.0	9.288	9.7	194.0	2647
Kalimantan Tengah	0	0.6	4.5	13.131	17.3	164.6	2350
Kalimantan Selatan	0	0.0	2.3	9.676	12.2	136.1	2025
Kalimantan Timur	0	0.0	0.7	7.089	7.5	153.5	941
Kalimantan Utara	0	0.0	1.0	7.895	8.8	157.2	772
Sulawesi Utara	0	0.0	2.0	8.155	9.0	206.0	1674
Sulawesi Tengah	0	0.0	1.0	7.611	7.0	214.9	1712
Sulawesi Selatan	0	0.0	0.0	9.336	9.0	270.0	968
Sulawesi Tenggara	0	0.0	1.0	6.970	7.0	163.0	1470
Gorontalo	0	0.0	0.0	4.888	4.0	142.0	1398
Sulawesi Barat	0	0.0	0.0	5.131	3.0	505.0	1798
Maluku	0	0.0	1.0	7.630	8.0	272.0	1794
Maluku Utara	0	0.0	0.0	6.215	5.0	188.0	1079
Papua Barat	0	0.0	2.0	9.709	11.0	248.8	1202
Papua	0	0.0	1.0	7.487	7.6	179.4	810

Next to Kalimantan Utara, we have DKI Jakarta (now DK Jakarta) with 56.96% missing values (3745 missing days). Unlike Kalimantan Utara, there is no justification to this, along

with the rest of the provinces. Fortunately, the next highest percentage is less than 50%, with Kalimantan Barat having 40.26% missing values (2647 missing days). After Kalimantan Barat, the provinces with around 30-40% missing values (1973-2630 missing days) are Nusa Tenggara Timur, Kalimantan Tengah, DI Yogyakarta, and Kalimantan Selatan. Next, with 20-30% missing values (1315-1973 missing days), we have Sulawesi Barat, Maluku, Sulawesi Tengah, Kepulauan Riau, Sulawesi Utara, Kepulauan Bangka Belitung, and Gorontalo. With 10-20% (658-1315 missing days) missing values, there are Papua Barat, Aceh, Maluku Utara, Sumatera Selatan, Lampung, Sulawesi Selatan, Sumatera Barat, Kalimantan Timur, Jawa Barat, Bengkulu, Papua, and Banten. Lastly, with less than 10% missing values (less than 658 missing days), the provinces are Riau, Jawa Timur, Jambi, Sumatera Utara, Nusa Tenggara Barat, Bali, Jawa Tengah.

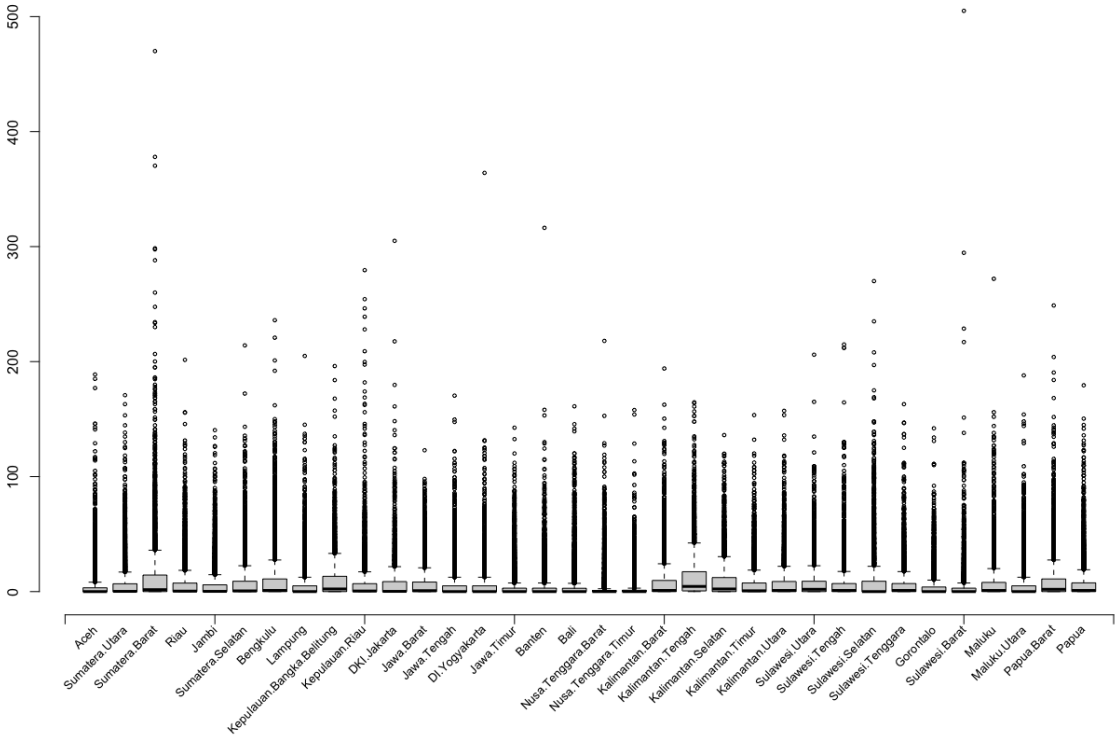
In Figure 2, we have the boxplot of the real data. According to Figure 2b, the rainfall values are mostly 0-40 mm/day, which is considered no rain or cloudy (0 mm/day), light rain (0.5-20 mm/day), or medium rain (20-50 mm/day) by [Geo]. However, in Figure 2a, we can see that we have a lot of outliers that reach high values up to 500 mm/day. Those are considered extreme rain (above 150 mm/day). All provinces show a similar behavior when it comes to outliers, but when it comes to no outliers, there are some provinces with higher rainfall values than the others, and there are also some provinces with less rain. Aceh, Jawa Timur, Banten, Bali, Nusa Tenggara Barat, and Nusa Tenggara Timur, for example, have less rain. Meanwhile, Sumatera Barat, Kalimantan Tengah, and Kepulauan Bangka Belitung seem to have a lot of rain.

2.2 Missing Data Handling Methods

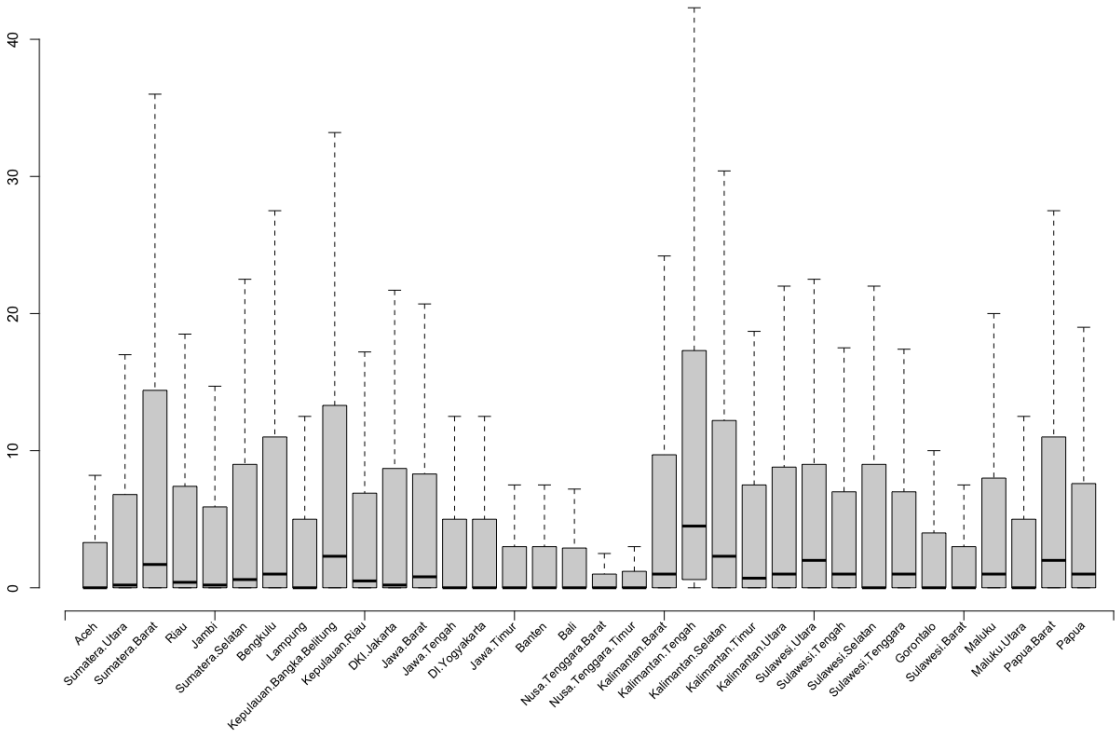
Certainly, the best way to handle missing data is to plan good data collection procedures. Unfortunately, this is not possible when we have secondary data. Hence, a technique to handle missing data is needed. The first method is listwise or case deletion, where the missing data are completely ignored for the rest of the calculation. This cannot be used for time series data, specifically this data, because we want to take daily values into consideration. Skipping a day might affect the calculation for the entire month or even the entire year. The second technique is pairwise deletion. This is similar to case deletion, but the data are only deleted when specific data-points are needed to test a particular assumption. The third technique is imputation or substitution. There are a lot of imputation methods to be taken into account, both single-value and multiple imputations. The single-value imputation methods that are going to be discussed in this thesis are Mean Substitution, Median Substitution, Last Observation Carried Forward, Linear Interpolation, Null Substitution, Bootstrap, and Distribution Fill. Meanwhile, the multiple imputation method we look at in this thesis is Multiple Imputation by Chained Equations.

2.2.1 Single-Value Imputation

Single-Value Imputation works by substituting only a single value into each missing value. This method works perfectly when we don't have a lot of missing values because it will most likely fill in the data based on the already existing patterns, and it shouldn't alter the data too significantly. Single-Value Imputation also works faster and is easier to compute than Multiple Imputation. If there are a lot of missing data, however, this method might change the original property and distribution of the data, causing errors in the calculation process leading up to the analysis. As mentioned before, there are seven single-value imputation methods to be considered in this research: Mean Substitution, Median Substitution, Last Observation Carried Forward, Interpolation, Null Substitution, Bootstrap, and Distribution Fill.



(a) with outliers



(b) without outliers

Figure 2: Boxplot of Daily Rainfall Values for Each Province

Mean Substitution
Mean substitution is a method of filling in the missing data values using the mean values of the other observations of the same variable. Unfortunately, due to the tendency of being inconsistently biased and underestimating errors, mean substitution is not a preferred option. However, it is still important to see how different the results can be using this method. This method would e.g. be good for a dataset that is normally distributed. So, we substitute the i-th missing data value by:

$$\hat{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_j, \tag{2.1}$$

where x_j is the j-th existing data value from the same variable, and n is the number of existing data values in the variable. After all the $\hat{x}_i$ values are filled, the new dataset with these mean values can be processed. For the data in this research, we calculate the mean of each month of each province. These mean values will then replace all the missing values in that month. For months with completely missing values, the mean values will be obtained from a neighboring province with the assumption of similar weather condition.

Median Substitution
Unfortunately, the mean substitution technique may lead to inconsistent bias and underestimation of errors. After all, the mean is known to be biased when the data are not normally distributed. Hence, mean substitution should not be the only technique used. Using the same idea, median substitution will also be used. Median is known to be more robust, to show the center of a dataset more equally than mean, especially when the data are skewed positively or negatively. The missing data values will be filled with median values of the other observations of the same variable. Median substitution might not have as many errors as mean substitution considering that it performs better for a dataset that is not normally distributed. So, we substitute the i-th missing data value by:

$$\hat{x}_i = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{\left(\frac{n+1}{2}\right)} + x_{\left(\frac{n}{2}\right)}\right) & \text{if } n \text{ is even,} \end{cases} \tag{2.2}$$

where $x_{(...)}$ is the existing data value from the same variable ordered from the lowest value to the highest value, and n is the number of existing data values in the variable. The index marks the location of the median value. After all the $\hat{x}_i$ values are filled, the new dataset with these median values can be processed. Just like mean substitution, we calculate the median of each month of each province. These median values will then replace all the missing values in that month. For months with complete missing values, the median values will be obtained from a neighboring province under the assumption of similar weather condition.

Last Observation Carried Forward (LOCF)
The fourth method that is used to handle missing data is last observation carried forward (LOCF). This method takes the last existing observation and uses it to fill in the gaps between the last existing observation until before the next existing observation. There are two types of LOCF: forward and backward. For this research, we will start with the forward method. So, we substitute the i-th missing data value by:

$$\hat{x}_i = \begin{cases} x_{i-1} & \text{if } x_{i-1} \text{ exists} \\ \hat{x}_{i-1} & \text{if } x_{i-1} \text{ does not exist,} \end{cases} \tag{2.3}$$

where x_i is the i-th existing data value from the same variable. Unfortunately, this forward method will not be able to fill in the missing values that are located at the start of the series. So, to fill in the remaining missing values, we will use the backward method with the imputed data from after the forward method is applied. The backward method's i-th missing data value is substituted by:

$$\hat{x}_i = \begin{cases} x_{i+1} & \text{if } x_{i+1} \text{ exists} \\ \hat{x}_{i+1} & \text{if } x_{i+1} \text{ does not exist,} \end{cases} \tag{2.4}$$

again, where x_i is the i-th existing data value from the same variable. By combining the two methods, all the missing gaps will be filled. Once all the $\hat{x}_i$ values are filled, the new dataset with the imputed values can be processed. The disadvantage of this method is that there will be a lot of similar values if the missing values are grouped together often. This can cause the data to be either extremely huge if the last observation before the gaps have huge values, or extremely small if the last observation before the gaps have low values, e.g. 0. In the case of rainfall values, this issue can mean either the area rained heavily (possibly causing floods) or the area was extremely dry because there was no rain. This method is highly dependent on the number of missing values.

Linear Interpolation

The fifth method that is used to handle missing data is interpolation. This method creates a pattern between the last existing observation and the next known observation from the same variable, e.g. in the form of a straight line, and fits the imputed data points into that pattern. This method is typically used in time series data or any kinds of datasets that have specific patterns to follow. For this research, we will use a straight line as the pattern, or simply linear interpolation. So, we substitute the i-th missing data value by:

$$\hat{x}_i = x_{t_1} + \frac{x_{t_2} - x_{t_1}}{t_2 - t_1}(i - t_1), \tag{2.5}$$

where x_{t_1} is the value of the last existing observation, x_{t_2} is the value of the next known observation, i is the time between t_1 and t_2 that contains the missing value in question, t_1 is the time of the last existing observation, and t_2 is the time of the next known observation. Once all the $\hat{x}_i$ values are filled, the new dataset with the imputed values can be processed. The disadvantage of this method, however, is that it makes all the data points somewhat connected to each other, while in reality, this is probably not the case. It might yield a bias as well, especially when there are a lot of missing values.

Null Substitution

In this part, we would also like to see how the calculations would be if all the missing values are replaced by 0. We simply change all the missing values with zero. So, the i-th missing data value is formulated as:

$$\hat{x}_i = 0. \tag{2.6}$$

After all the $\hat{x}_i$ values are filled, the new dataset can be processed. Of course, this method will certainly lead to a misleading analysis, but the use of Null Substitution here is to be compared with the other methods. In reality, this case is not possible because this means there was no rain in the area at all for a lot of days. This can lead to a long draught period.

Bootstrap
The Bootstrap technique is also used to fill the missing data gaps. Instead of getting new values from some calculations, the missing data gaps are filled with values chosen randomly from the existing observations. However, Bootstrap is not a deterministic method. In order to make it deterministic, the random seed in R needs to be set. By iterating this process, the random sampling goes on until bootstrap can no longer fill in the gaps. The formula that can be used to picture this technique to substitute the i-th missing data value is:

$$\hat{x}_i \in \{x_j \mid j \in \mathcal{O}\} \tag{2.7}$$

where x_j is the j-th existing observed value where j is chosen randomly, and $\mathcal{O}$ is the set of existing observed value. The advantages of this method include simplicity, speed, and data distribution preservation. Since the method only takes existing values to fill in the empty gaps, the distribution of the data is less likely to change. However, this method also has some disadvantages, including possible bias, lack of variability, and lack of pattern recognition. Bootstrap does not check for patterns of the data, so the pattern might change with this method. While it preserves data distribution, the imputed data might have less variability due to having similar values appearing at different points. These disadvantages also apply to the other single-imputation methods mentioned before, for some more extremely.

Distribution Fitting (Fill)
Finding suitable distributions for each province is not really easy. First of all, rainfall values are always equal or greater than zero, and they are interval data. So, we need distributions for continuous data that allow zero values. As a start, we will try to fit the data into three distributions: Exponential, Normal, and Uniform. Exponential distribution allows for positive values, but it cannot be zero. Normal distribution allows for both positive and negative values, but it allows a huge range of data, so it also works for rainfall values that are rather unpredictable. Uniform distribution is unlikely to happen, but some of the histograms show balanced probabilities. Using the obtained distribution, we will attempt on filling in the missing values. We will generate random values using the rate of the Exponential distribution of each province, and then insert the values to substitute the missing ones. This method will be called "Fill" going forward for simpler writing. The i-th missing data value is substituted by:

$$\hat{x}_i = y_i, \quad Y_i \sim \text{Exponential}(\lambda) \tag{2.8}$$

where y_i is the i-th generated data value from Y_i that follows an exponential distribution and is independent of all $Y_j, j \neq i$, and λ is the parameter of the exponential distribution obtained through the fitting of the existing data values.

2.2.2 Multiple Imputation

A Multiple Imputation method fills in missing values by generating numbers based on the distributions and relationships of the observed data. Unlike Single-Value Imputation that only fills in missing values once, Multiple Imputation goes through a trial and error process until the best value to fill in the empty spot is found. According to [LSA15], this imputation method has two stages. Firstly, generate the replacement values for many times based on the statistical characteristics of the data, e.g. correlations and distributions. This first step will produce numerous datasets with replaced missing values. Secondly, the datasets will be analyzed and combined until the best results are obtained. Any analysis can be conducted as long as it fits the objective of the research. Here, the Multiple Imputation method that we are going to use is only Multiple Imputation by Chained Equations.

Multiple Imputation by Chained Equations (MICE)
Multiple Imputation by Chained Equations (MICE) is a robust method that fills in missing values by iterating predictive models using other variables in the dataset. The iterations should be run until convergence has been met. MICE has a lot of methods. The ones that are going to be used for this research are:

Predictive Mean Matching (PMM)
This method will create regression models based on existing variables and values, find the predicted values of the missing data, and then find the nearest existing values to those predicted values. PMM does not impute data with new values, but instead, fills in the values with the closest value with the predicted outcome of the regression models. This method will also use all existing values that are given without excluding any of it. All of the variables being put in for this process will be taken into consideration, which means all the predictors of the regression models will be there no matter how much they affect the models. The advantage of this method is that it uses existing values to fill in the missing data, so the distribution will stay similar. This method also doesn't capture the pattern of the missing data, which is perfect for data that are missing randomly. However, this method works best for linear data, which is not the case here, and it always needs at least one predictor, or at least two variables for the input.

Classification and Regression Trees (CART)
This is one of the few methods in MICE that allow non-linear relationships between the variables which fit rainfall data values really well. CART mixes regression models for continuous and categorical data, but this method can also work for non-categorical data. When the data is non-categorical, CART will separate the data into several categories to fit the decision tree. The advantage of this method is that it works for any kind of data with non-linear relationships. This method also works better for data that are missing at random. However, CART also requires more than one variable to be put in to create the decision tree. It also doesn't use existing values for the imputation, so it is possible for CART to change the data distribution.

Lasso Linear Regression (LASSO)
The only similarity of LASSO and PMM is that they both use linear regression models. However, LASSO doesn't impute the missing values like PMM does. As a start, LASSO will create regression models using existing values. However, after that, LASSO will analyze the significance of each predictor to the response. It involves shrinking coefficients (making the coefficients 0) to determine the best results. Once the predicted results are obtained, unlike PMM, LASSO will immediately use these predicted values to impute the missing values. The advantage of this method is that it makes sure that the imputed values are the best predicted results because it has gone through coefficient shrinkage analysis. This method is also good for the case of MCAR. However, this method requires at least two predictors, which means at least three variables, for the input. This method also only works well for linear relationships.

Random Indicator for Nonignorable Data (RI)
RI is the only method among the other MICE methods here that captures the missingness pattern, which means that it is not good to impute data that are missing completely at random. This is actually the only MICE method that checks the pattern of the missing data. As a start, RI changes all the missing values for that variable to binary patterns (1 for observed and 0 for missing) and creates logistic regression models using the other variables as predictors. This will determine why the value is missing based on other variables. As for the imputation itself, we need a linear regression model with the binary component as one of the predictors. There is

one extra predictor in the linear regression models compared to the logistic regression models. Even though the final model is linear regression, the logistic regression model helps dealing with non-linear relationships. This method only requires at least one predictor, which means two variables for the method to work, but it works better with more predictors.

Random Sample from Observed Values (SAMPLE)
This method can be said to be a mix of PMM and Bootstrap. First, SAMPLE will use regression models to get the predicted values of the missing values. However, instead of imputing the missing values using the predicted values or using the closest similar values like PMM, SAMPLE will draw a random value from the distribution of existing values and predicted values within that range. This is where it is similar to Bootstrap, but instead of drawing from the entire existing data, SAMPLE only draws from the values within a specific range based on the predicted values, residuals, and distributions. This method can be used with only one predictor, which means only two variables are required. This method can also work with non-linear relationships. Because this method isn't limited to linear regressions, it can also work with logistic regressions for binary data and even for decision trees for data with more categories.

2.3 Data After Imputation

In this section, we discuss the data after all the missing values are imputed. After imputation, the daily rainfall values will be accumulated to produce monthly rainfall values. This is done to adjust the unit with paddy rice needs (200 mm/month) and monthly payout values based on [APD23]. Unfortunately, obtaining monthly data is not possible from the real data due to the incompleteness. Hence, all the monthly rainfall values being discussed moving forward are all from the imputed daily data. The monthly rainfall values can be seen in Appendix B.

2.3.1 Mean Substitution

Here we summarize the values after Mean Substitution is conducted. In Table 2, the substitution has managed to increase some values in the first quartile and median. Previously, in Table 1, all values in the first quartile are 0 except for 1 province, but now there are 4 provinces with larger than 0 first quartile. These provinces are Kepulauan Bangka Belitung, Kalimantan Barat, Kalimantan Tengah, and Kalimantan Selatan. The remaining values are more or less similar. Surprisingly, some of them are even less than the statistics of the real data. There are some significant changes in mean and third quartile. The maximum and minimum values also shows that Mean Substitution does not change the range of the daily data.

Next, in Figure 4, we have the boxplot of accumulated imputed daily rainfall values that form monthly rainfall values. In Figure 4b, we can see that most monthly rainfall values are around 0-1000 mm. However, if we see Figure 4a, there are again, some outliers. However, this time, the outliers are not that many. The worst outlier is the one in DKI Jakarta with almost 2000 mm/month of rainfall. Aside from that, the rainfall values in these provinces seem normal. Sulawesi Selatan and Sumatera Barat seem to have a wide range of rainfall values, and they even reach 1000 mm/month. The others are mostly centered below 400 mm/month.

Next, we have the histogram of the monthly imputed data in Figure 3 and the summary statistics in Table 3. The data is highly positively skewed. The maximum value, as seen in Figure 4a, almost reaches 2000. which is 1946.6 mm/month. The minimum value is of course, 0, meaning the months in question had no rain at all. From the total of 216 months and 34 provinces, resulting in 7344 values, there are 330 months with zero rainfall (around 4.49%).

Table 2: Daily Rainfall Data Summary for Each Province after Mean Substitution

Province	Min	Q1	Median	Mean	Q3	Max
Aceh	0	0	0.32	5.36	6.00	188.80
Sumatera Utara	0	0	0.68	7.42	7.60	170.80
Sumatera Barat	0	0	3.80	14.06	18.10	470.00
Riau	0	0	1.00	7.84	9.50	201.50
Jambi	0	0	0.60	6.74	7.47	140.40
Sumatera Selatan	0	0	1.70	8.33	10.50	214.10
Bengkulu	0	0	2.70	10.73	13.00	236.00
Lampung	0	0	0.70	6.38	8.50	204.90
Kepulauan Bangka Belitung	0	0.03	5.50	10.48	14.69	196.10
Kepulauan Riau	0	0	3.50	8.53	11.01	279.50
DKI Jakarta	0	0	3.10	8.30	11.29	305.00
Jawa Barat	0	0	1.50	7.17	9.19	122.90
Jawa Tengah	0	0	0	6.29	6.20	170.40
DI Yogyakarta	0	0	1.00	6.30	7.68	364.10
Jawa Timur	0	0	0	5.21	4.70	142.50
Banten	0	0	0	5.02	4.54	316.30
Bali	0	0	0	5.45	3.60	161.10
Nusa Tenggara Barat	0	0	0	4.35	2.00	218.00
Nusa Tenggara Timur	0	0	0	3.93	4.17	157.90
Kalimantan Barat	0	0.10	7.90	11.16	17.00	194.00
Kalimantan Tengah	0	2.00	10.04	13.44	18.04	164.60
Kalimantan Selatan	0	0.20	5.58	9.77	12.80	136.10
Kalimantan Timur	0	0	1.80	7.49	9.57	153.50
Kalimantan Utara	0	0	2.00	8.26	10.60	157.20
Sulawesi Utara	0	0	4.00	8.15	10.01	206.00
Sulawesi Tengah	0	0	2.00	7.35	8.25	214.90
Sulawesi Selatan	0	0	1.00	9.25	9.99	270.00
Sulawesi Tenggara	0	0	2.00	7.54	9.89	163.00
Gorontalo	0	0	1.00	5.18	6.04	142.00
Sulawesi Barat	0	0	0.80	5.49	5.41	505.00
Maluku	0	0	2.70	7.53	9.00	272.00
Maluku Utara	0	0	1.00	6.35	7.13	188.00
Papua Barat	0	0	3.50	9.77	12.17	248.80
Papua	0	0	2.00	7.81	8.90	179.40

Statistic	Value
n	7344
Min	0.0
Q1	100.75
Median	208.10
Mean	234.95
Q3	332.71
Max	1946.56
0's	330

Table 3: Summary of Monthly Rainfall Values after Mean Substitution

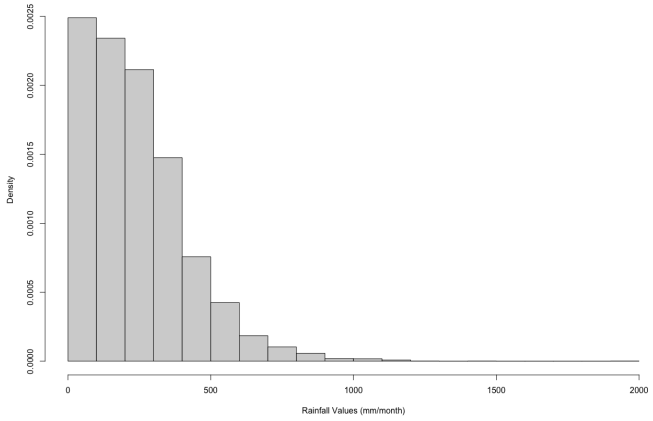
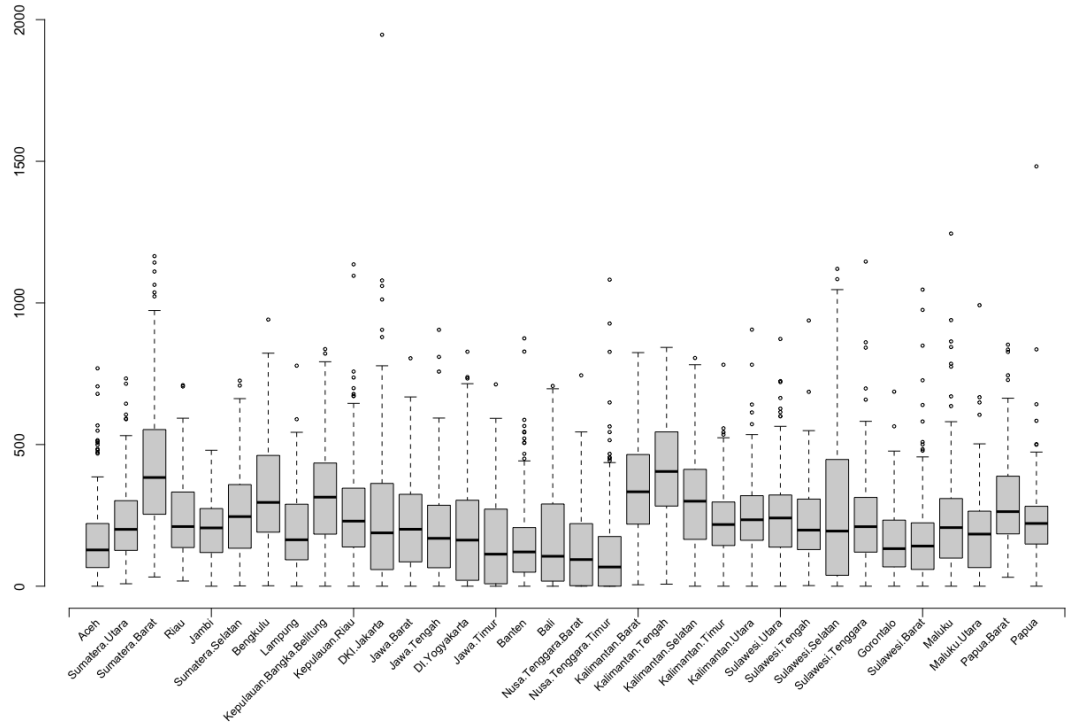
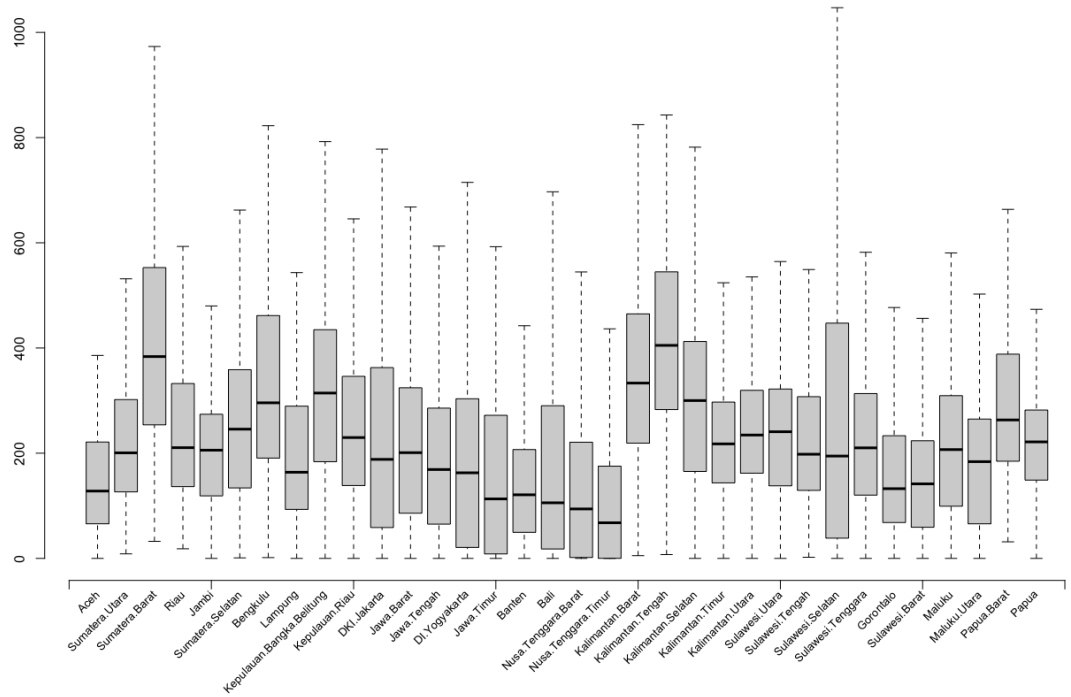


Figure 3: Histogram of Monthly Rainfall Values after Mean Substitution



(a) with outliers



(b) without outliers

Figure 4: Boxplot of Daily Rainfall Values for Each Province after Mean Substitution

2.3.2 Median Substitution

Here we summarize the values after Median Substitution is conducted. In Table 5, we can see that there are not many changes to the first quartile. Kalimantan Tengah, the only province with first quartile that wasn't zero, stays in the same condition, only with different value (from 0.6 to 1.5). The increment in median values is not that significant as well. Just like Mean Substitution, the mean and third quartile values have some changes by some being higher and some being lower, but again, nothing significant. The maximum values remain the same as well, which means that Median Substitution also doesn't change the range of data. Seeing that the other values are mostly more similar compared to Mean Substitution, we can also say that Median Substitution is more robust.

In Figure 6, the boxplot of the accumulated imputed daily rainfall values is shown. Figure 6a shows a lot of outliers, but not as high as the one we had in Figure 4a when the data was imputed through Mean Substitution. In this case, this shows that Median Substitution provides more low values than Mean Substitution, indicating that the median values of each province are mostly lower than the mean values (positively skewed data). Figure 6b also shows less rainfall values interval which is between 0-800 mm except for two provinces: Sulawesi Selatan and Sumatera Barat. DKI Jakarta seems to have normal range in here, but with outliers, it is quite similar to the other two provinces mentioned before. With Mean Substitution, none of the mean values of each province reaches 400 mm/month.

The histogram of the monthly imputed data in Figure 5 shows that this data is also positively skewed. Interestingly, however, the histogram keeps going lower with the rainfall values getting higher. The maximum value in Table 4 is only 1166. which is far below what Mean Substitution has. The first quartile, median, mean, and third quartile are also lower than Mean Subsitution. However, Median Substitution has 451 zero values (around 6.14%), which is more than what Mean Substitution has. This also shows that some of the median values are zero, and it makes the entire data much lower than Mean Substitution.

Statistic	Value
n	7344
Min	0.00
Q1	80.97
Median	179.78
Mean	203.57
Q3	290.00
Max	1166.00
0's	451

Table 4: Summary of Monthly Rainfall Values after Median Substitution

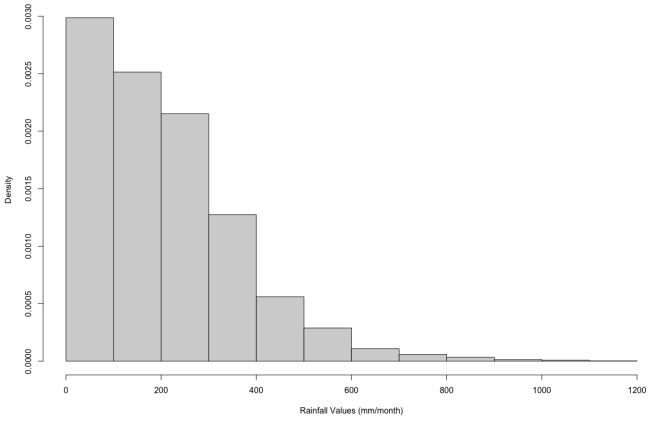
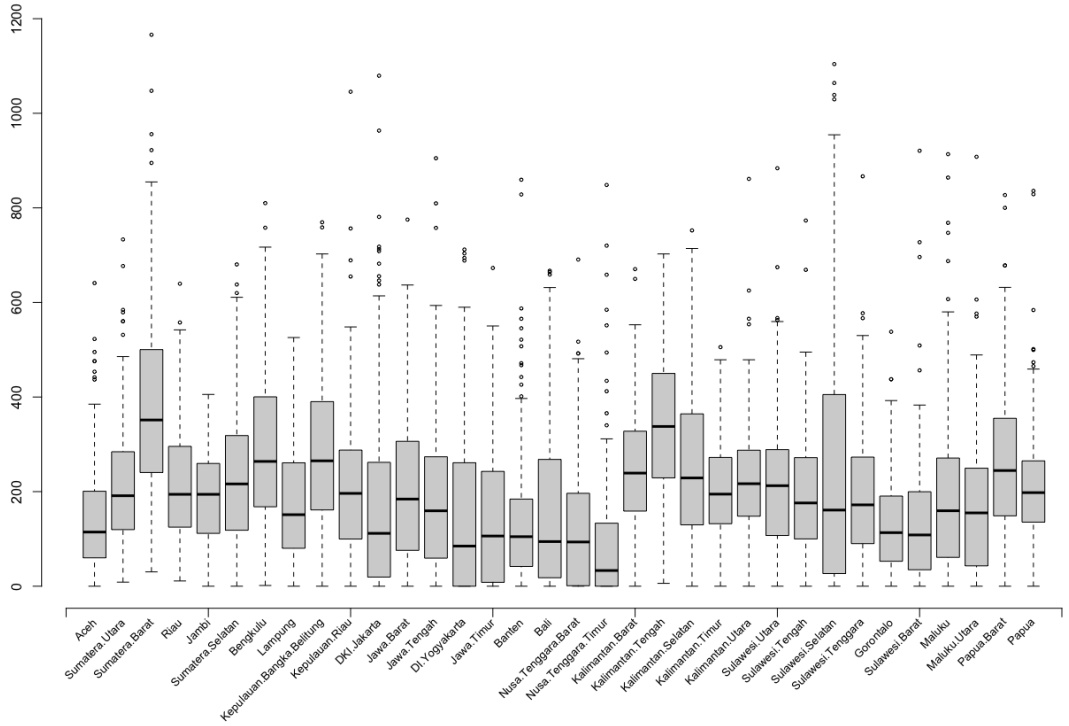


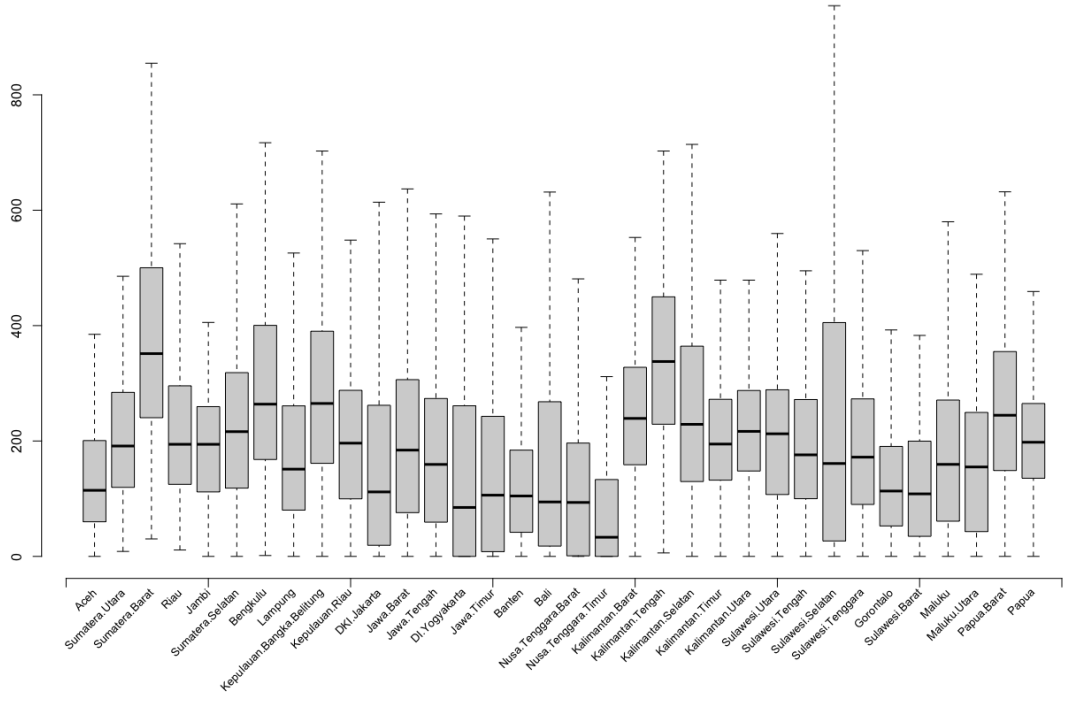
Figure 5: Histogram of Monthly Rainfall Values after Median Substitution

Table 5: Daily Rainfall Data Summary for Each Province after Median Substitution

Province	**Min**	**Q1**	**Median**	**Mean**	**Q3**	**Max**
Aceh	0	0	0	4.62	4.20	188.80
Sumatera Utara	0	0	0.40	7.00	6.00	170.80
Sumatera Barat	0	0	2.80	12.77	14.35	470.00
Riau	0	0	0.50	7.17	6.50	201.50
Jambi	0	0	0.30	6.27	5.20	140.40
Sumatera Selatan	0	0	1.00	7.41	7.20	214.10
Bengkulu	0	0	1.05	9.65	9.10	236.00
Lampung	0	0	0.20	5.66	6.00	204.90
Kepulauan Bangka Belitung	0	0	3.60	9.21	11.35	196.10
Kepulauan Riau	0	0	1.50	6.91	6.70	279.50
DKI Jakarta	0	0	0.25	5.99	6.90	305.00
Jawa Barat	0	0	1.00	6.63	7.50	122.90
Jawa Tengah	0	0	0	5.92	4.50	170.40
DI Yogyakarta	0	0	0	4.83	3.30	364.10
Jawa Timur	0	0	0	4.74	3.00	142.50
Banten	0	0	0	4.61	2.90	316.30
Bali	0	0	0	5.13	2.60	161.10
Nusa Tenggara Barat	0	0	0	4.04	1.50	218.00
Nusa Tenggara Timur	0	0	0	2.99	1.90	157.90
Kalimantan Barat	0	0	4.30	8.26	9.30	194.00
Kalimantan Tengah	0	1.50	6.10	11.04	12.40	164.60
Kalimantan Selatan	0	0	3.20	8.20	9.60	136.10
Kalimantan Timur	0	0	1.20	6.61	7.00	153.50
Kalimantan Utara	0	0	1.50	7.48	7.82	157.20
Sulawesi Utara	0	0	2.00	7.04	8.00	206.00
Sulawesi Tengah	0	0	1.00	6.26	5.00	214.90
Sulawesi Selatan	0	0	0	8.40	7.00	270.00
Sulawesi Tenggara	0	0	1.00	6.24	6.80	163.00
Gorontalo	0	0	0	4.27	3.00	142.00
Sulawesi Barat	0	0	0	4.32	2.70	505.00
Maluku	0	0	1.00	6.32	6.00	272.00
Maluku Utara	0	0	0	5.56	4.00	188.00
Papua Barat	0	0	2.60	8.68	8.85	248.80
Papua	0	0	1.50	7.11	7.20	179.40



(a) with outliers



(b) without outliers

Figure 6: Boxplot of Daily Rainfall Values for Each Province after Median Substitution

2.3.3 LOCF Imputation

Here we summarize the values after LOCF is imputed. Just like the other two methods, as shown in Table 7, LOCF Imputation also did not change the range of the data, because the maximum and minimum values remain the same. The changes are not significant. As usual, some values are increased, some are decreased.

We have the boxplot of LOCF Imputation data in Figure 8 both with and without outliers. Figure 8a shows that the outliers for LOCF are very high, reaching more than 2500 mm/month in Riau and Nusa Tenggara Timur. There are quite a few other provinces with high outlier values, but not as high as the mentioned two. In Figure 8b, it shows that the range of rainfall values is between 0-1000 mm/month without outliers, which is more than Median Substitution but similar to Mean Substitution. Aside from Sulawesi Selatan and Sumatera Barat that reach 1000 mm/month in range, the other provinces are mostly below 800 mm/month, except for Bengkulu.

Table 6 shows that the summary is almost similar to Mean Substitution, and obviously above Median Substitution. Interestingly, all the statistics in LOCF Imputation have lower values than Mean Substitution except for the minimum and maximum value. Their minimum values are both equal to 0, meanwhile the maximum value for LOCF Imputation is much higher than Mean Substitution. This indicates some high values being the last known observations. As for the number of zero values, LOCF Imputation apparently has 411 zero values (around 5.59%), which is closer to Median Substitution, also indicating a lot of zero values as the last known observations. As for the histogram in Figure 7, again it shows a positively skewed distribution. Like Median Substitution, the density also gets lower with more rainfall values, but the decrement is bigger than Median Substitution.

Statistic	Value
n	7344
Min	0.00
Q1	95.65
Median	197.75
Mean	233.66
Q3	321.20
Max	2650.50
0's	411

Table 6: Summary of Monthly Rainfall Values after LOCF Imputation

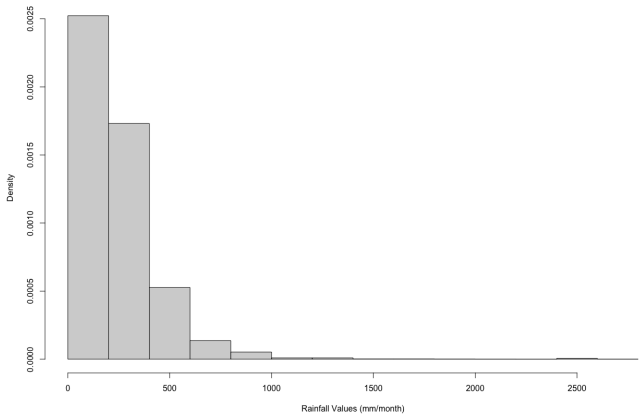
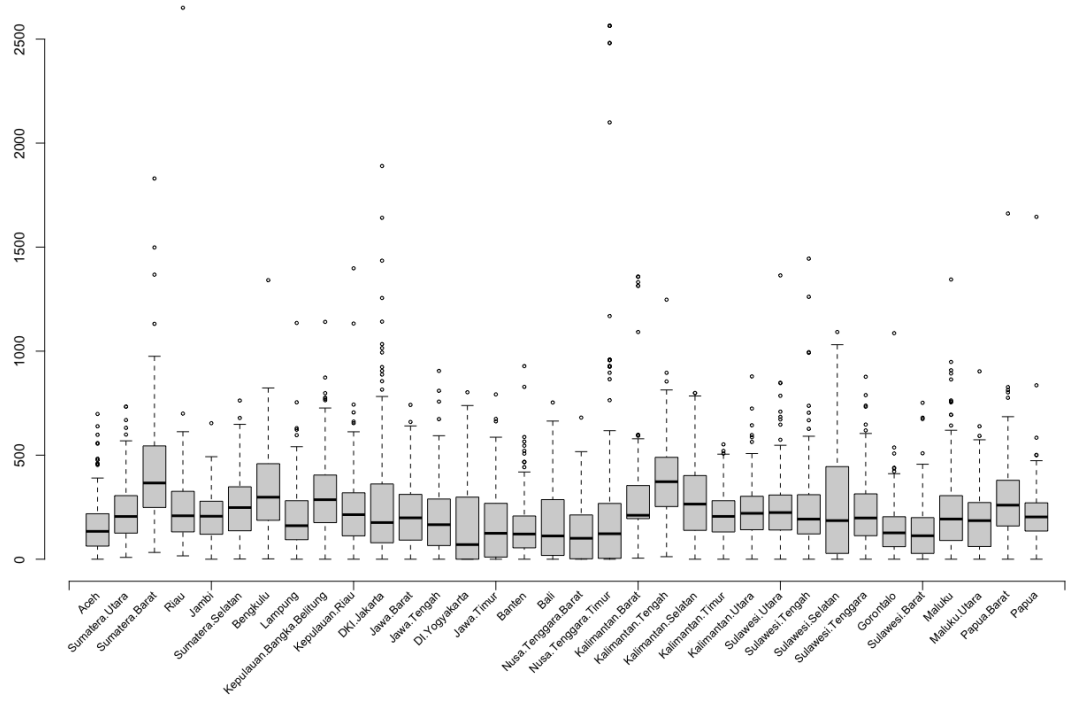


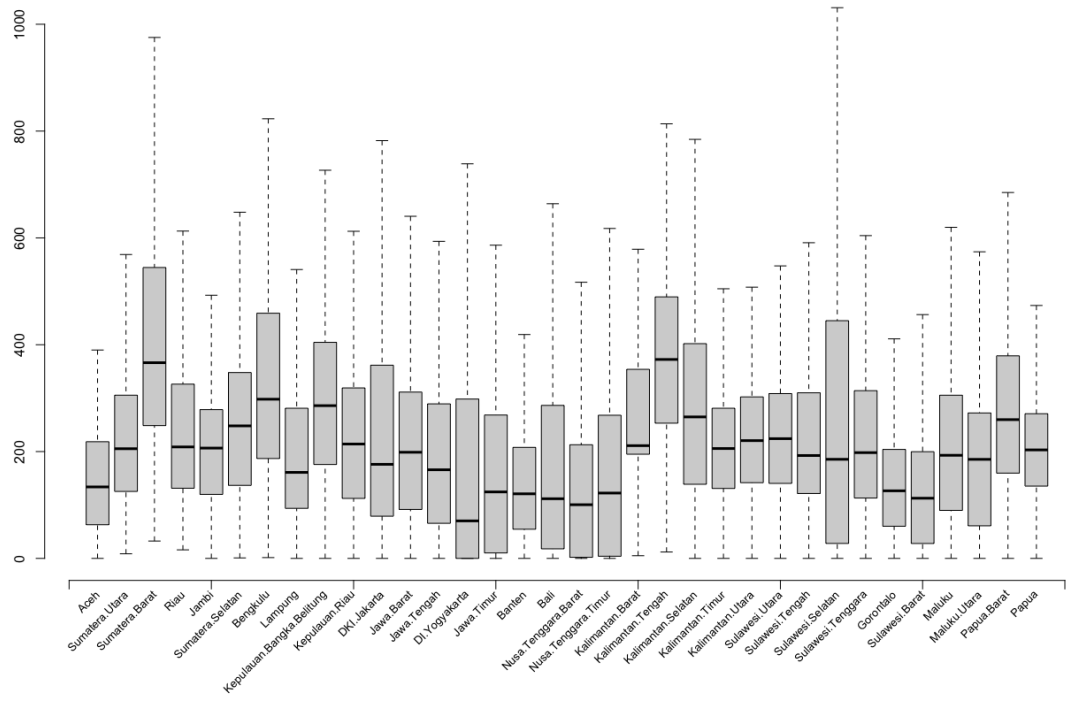
Figure 7: Histogram of Monthly Rainfall Values after LOCF Imputation

Table 7: Daily Rainfall Data Summary for Each Province after LOCF Imputation

Province	**Min**	**Q1**	**Median**	**Mean**	**Q3**	**Max**
Aceh	0	0	0	5.22	4.50	188.80
Sumatera Utara	0	0	0.30	7.45	6.80	170.80
Sumatera Barat	0	0	2.00	13.91	15.10	470.00
Riau	0	0	0.50	8.09	7.85	201.50
Jambi	0	0	0.20	6.77	6.20	140.40
Sumatera Selatan	0	0	0.90	8.37	9.00	214.10
Bengkulu	0	0	1.00	10.92	11.00	236.00
Lampung	0	0	0.20	6.41	6.30	204.90
Kepulauan Bangka Belitung	0	0	2.80	10.15	13.30	196.10
Kepulauan Riau	0	0	1.00	7.88	8.00	279.50
DKI Jakarta	0	0	0.50	8.80	10.50	305.00
Jawa Barat	0	0	0.80	7.06	8.30	122.90
Jawa Tengah	0	0	0	6.28	5.00	170.40
DI Yogyakarta	0	0	0	5.14	2.10	364.10
Jawa Timur	0	0	0	5.25	3.40	142.50
Banten	0	0	0	5.00	3.70	316.30
Bali	0	0	0	5.47	3.00	161.10
Nusa Tenggara Barat	0	0	0	4.32	2.00	218.00
Nusa Tenggara Timur	0	0	0	10.14	9.70	157.90
Kalimantan Barat	0	0	6.50	9.58	6.50	194.00
Kalimantan Tengah	0	0.80	4.60	12.63	16.20	164.60
Kalimantan Selatan	0	0	2.40	9.20	11.40	136.10
Kalimantan Timur	0	0	0.50	6.80	7.00	153.50
Kalimantan Utara	0	0	1.00	7.64	8.20	157.20
Sulawesi Utara	0	0	2.00	8.12	9.00	206.00
Sulawesi Tengah	0	0	1.00	7.90	7.60	214.90
Sulawesi Selatan	0	0	0	9.18	9.85	270.00
Sulawesi Tenggara	0	0	1.00	7.28	8.40	163.00
Gorontalo	0	0	0	4.81	3.50	142.00
Sulawesi Barat	0	0	0	4.40	2.00	505.00
Maluku	0	0	2.00	7.63	7.00	272.00
Maluku Utara	0	0	0	6.32	6.00	188.00
Papua Barat	0	0	2.30	9.56	10.50	248.80
Papua	0	0	1.40	7.32	7.40	179.40



(a) with outliers



(b) without outliers

Figure 8: Boxplot of Daily Rainfall Values for Each Province after LOCF Imputation

2.3.4 Linear Interpolation

Here we summarize the values after Linear Interpolation is conducted. In Table 9, the range of the data remains the same. The changes made are not very significant as well.

The boxplots for Linear Interpolation monthly data are shown in Figure 10. It seems like Nusa Tenggara Timur has some really high daily values seeing that the outlier reaches over 2000 mm/month in Figure 10a. The other provinces also have some outliers, but not as huge as Nusa Tenggara Timur. The second province with the highest outlier values would be Papua Barat, but even the outlier is only around 1600 mm/month. Seeing the rainfall values without outliers like in Figure 10b, the range is around 0-1000 mm/month again. For this, the conditions are more or less similar to the other imputation methods. Sulawesi Selatan and Sumatera Barat are once again the highest, followed by Kalimantan Tengah and Bengkulu.

The summary of the monthly data for all provinces combined can be seen in Table 8. Judging from the values, Linear Interpolation is similar to Mean Substitution and LOCF Imputation. However, the first quartile and median values for Linear Interpolation are the lowest amongst the three. Meanwhile, the median, third quartile, and maximum values are higher than LOCF Imputation, but lower than Mean Substitution. According to the statistics, it is more similar to LOCF Imputation. As for the zero values, it has 360 zero values (around 4.90%) that indicates that most linear patterns in the data were formed not between zero and another zero. Just like the other substitution methods, the histogram in Figure 9 also shows that this aggregated data is positively skewed.

Statistic	Value
n	7344
Min	0.00
Q1	94.63
Median	198.80
Mean	229.38
Q3	323.85
Max	2343.70
0's	360

Table 8: Summary of Monthly Rainfall Values after Linear Interpolation

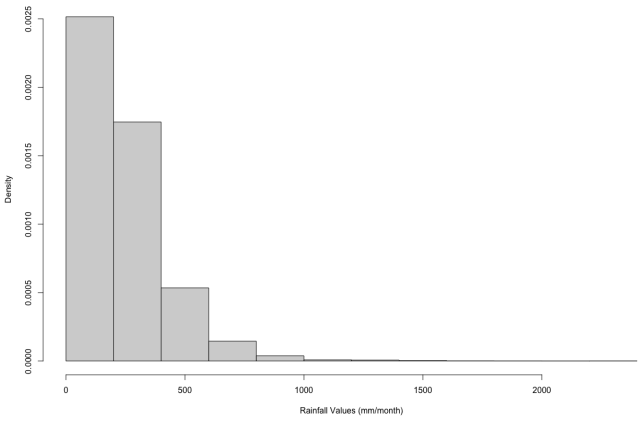
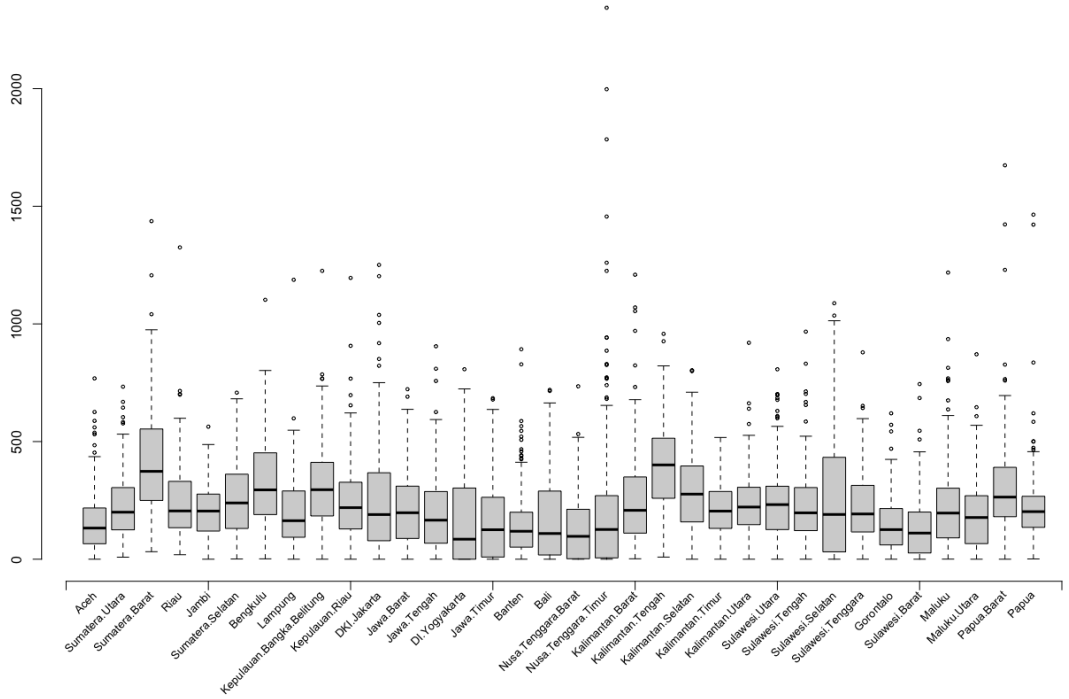


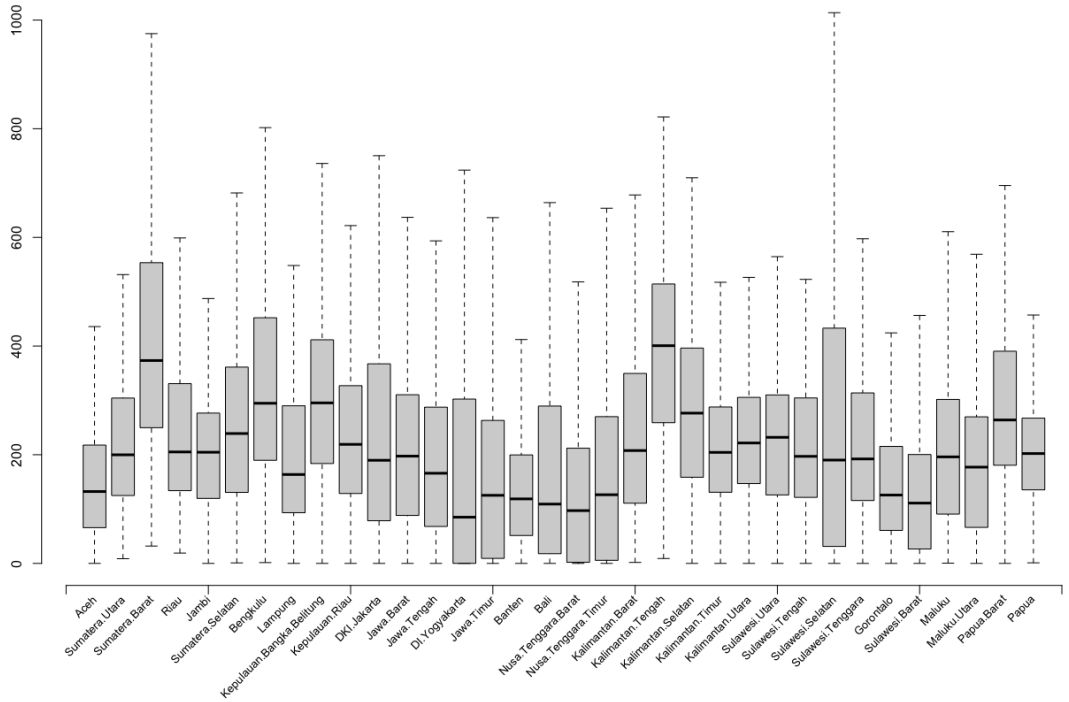
Figure 9: Histogram of Monthly Rainfall Values after Linear Interpolation

Table 9: Daily Rainfall Data Summary for Each Province after Linear Interpolation

Province	**Min**	**Q1**	**Median**	**Mean**	**Q3**	**Max**
Aceh	0	0	0	5.25	4.94	188.80
Sumatera Utara	0	0	0.40	7.38	7.10	170.80
Sumatera Barat	0	0	2.50	13.78	16.00	470.00
Riau	0	0	0.60	8.00	8.35	201.50
Jambi	0	0	0.40	6.76	6.50	140.40
Sumatera Selatan	0	0	1.10	8.31	9.50	214.10
Bengkulu	0	0	2.00	10.71	12.00	236.00
Lampung	0	0	0.40	6.42	6.70	204.90
Kepulauan Bangka Belitung	0	0	3.60	10.33	14.00	196.10
Kepulauan Riau	0	0	1.61	8.04	9.00	279.50
DKI Jakarta	0	0	1.63	8.29	10.20	305.00
Jawa Barat	0	0	1.00	7.00	8.30	122.90
Jawa Tengah	0	0	0	6.28	5.20	170.40
DI Yogyakarta	0	0	0	5.27	3.58	364.10
Jawa Timur	0	0	0	5.23	3.80	142.50
Banten	0	0	0	4.98	3.80	316.30
Bali	0	0	0	5.45	3.20	161.10
Nusa Tenggara Barat	0	0	0	4.31	2.00	218.00
Nusa Tenggara Timur	0	0	0.40	7.35	8.94	157.90
Kalimantan Barat	0	0	2.60	8.30	6.68	194.00
Kalimantan Tengah	0	1.35	6.20	13.02	17.60	164.60
Kalimantan Selatan	0	0	3.00	9.32	11.78	136.10
Kalimantan Timur	0	0	1.00	6.90	7.50	153.50
Kalimantan Utara	0	0	1.10	7.70	8.60	157.20
Sulawesi Utara	0	0	2.40	7.91	9.00	206.00
Sulawesi Tengah	0	0	1.00	7.51	7.50	214.90
Sulawesi Selatan	0	0	0.14	8.98	9.00	270.00
Sulawesi Tenggara	0	0	1.00	7.14	8.68	163.00
Gorontalo	0	0	0	4.81	4.00	142.00
Sulawesi Barat	0	0	0.02	4.33	2.40	505.00
Maluku	0	0	1.90	7.29	7.42	272.00
Maluku Utara	0	0	0.20	6.22	5.50	188.00
Papua Barat	0	0	2.90	10.17	11.90	248.80
Papua	0	0	1.31	7.48	7.60	179.40



(a) with outliers



(b) without outliers

Figure 10: Boxplot of Daily Rainfall Values for Each Province after Linear Interpolation

2.3.5 Null Substitution

Here we summarize the values after Null Substitution is conducted. In Table 11, obviously the maximum values would not change because all the missing values are changed into zero. The first quartile for Kalimantan Tengah, however, changes to 0 because we have more zero values in the data. Overall, this imputation method lowers all the statistics in each province, which is to be expected.

With the daily values mostly being 0, the aggregated monthly values decrease as well. In Figure 12, we can see that the values are overall smaller. Even in Figure 12a, the highest outlier value is only a little over 1000 mm/month. There are not as many outliers as well, seeing that the monthly values are mostly decreased and now have smaller values compared to the other imputation methods. The ones without outlier in Figure 12b, however, seems similar to Median Substitution because it ranges from 0-800 mm/month except for Sulawesi Selatan. Even Sumatera Barat has a huge difference with Sulawesi Selatan here.

The histogram in Figure 11 is surprisingly pretty balanced. It was expected that zero would have a high density, but instead, the density decreases nicely. And again, the data forms a positively skewed distribution, which can also be seen from Table 10 with the median being lower than the mean. Overall, the statistic values are also the lowest amongst other methods. All these values have around 20-30 points of difference with Median Substitution statistics, which is the second lowest so far. The maximum value is even lower by 80 points. But of course, this happens because this method has a lot of zero values which results in a lot of months without rains as well. In total, there are 668 zero values (around 9.09%) in this imputed data.

Statistic	Value
n	7344
Min	0.00
Q1	62.10
Median	158.85
Mean	180.58
Q3	262.72
Max	1084.00
0's	668

Table 10: Summary of Monthly Rainfall Values after Null Substitution

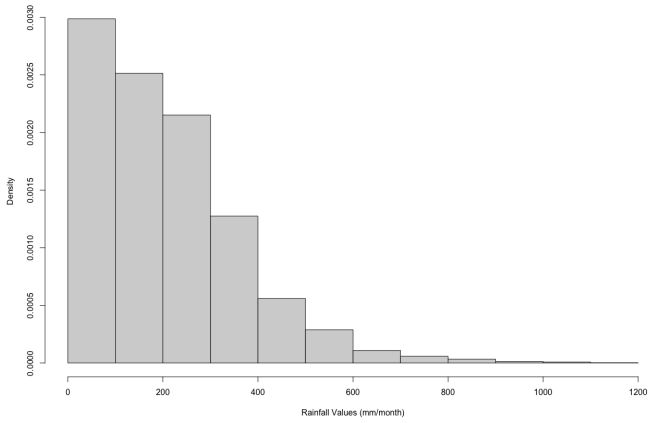
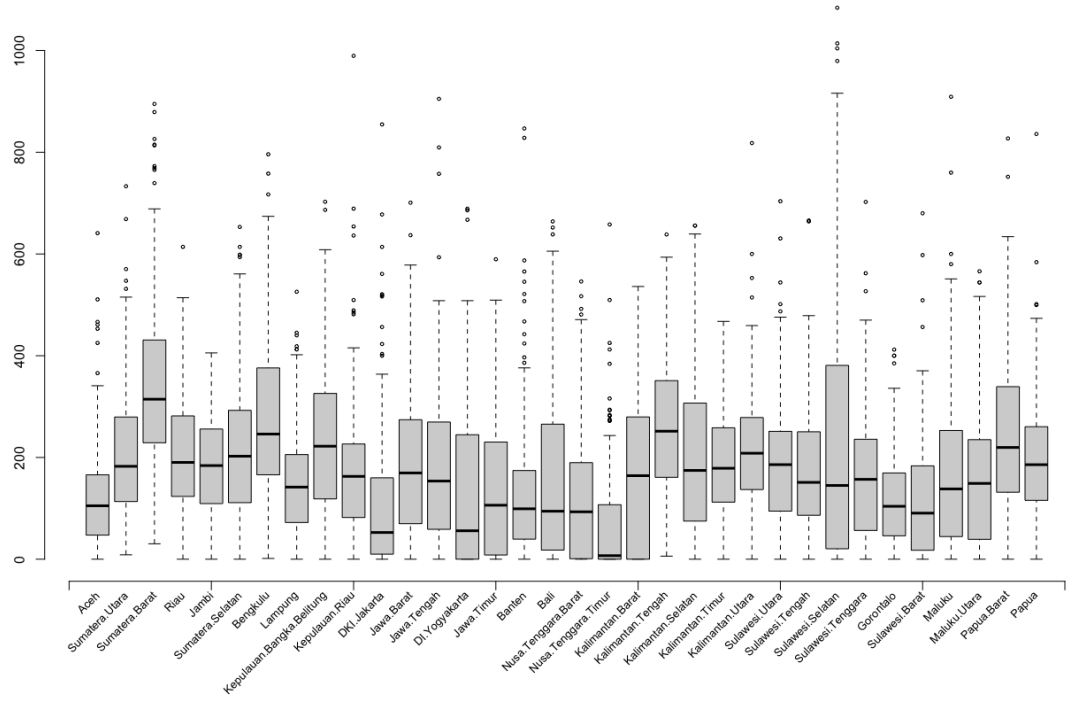


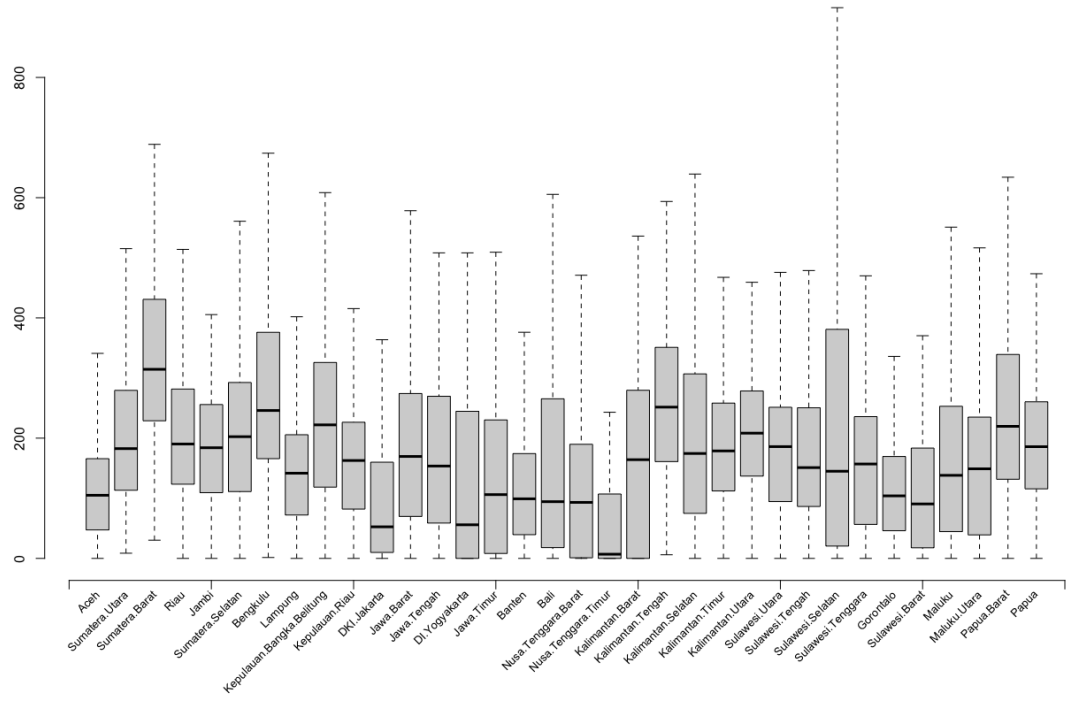
Figure 11: Histogram of Monthly Rainfall Values after Null Substitution

Table 11: Daily Rainfall Data Summary for Each Province after Null Substitution

Province	**Min**	**Q1**	**Median**	**Mean**	**Q3**	**Max**
Aceh	0	0	0	4.09	2.00	188.80
Sumatera Utara	0	0	0	6.83	5.60	170.80
Sumatera Barat	0	0	0.20	11.43	10.40	470.00
Riau	0	0	0.20	6.86	5.90	201.50
Jambi	0	0	0	6.10	4.70	140.40
Sumatera Selatan	0	0	0	7.02	6.40	214.10
Bengkulu	0	0	0.10	9.22	8.00	236.00
Lampung	0	0	0	4.96	3.00	204.90
Kepulauan Bangka Belitung	0	0	0.40	7.76	8.50	196.10
Kepulauan Riau	0	0	0	5.74	3.70	279.50
DKI Jakarta	0	0	0	3.63	0	305.00
Jawa Barat	0	0	0	6.18	6.00	122.90
Jawa Tengah	0	0	0	5.81	4.00	170.40
DI Yogyakarta	0	0	0	4.28	0.80	364.10
Jawa Timur	0	0	0	4.55	2.00	142.50
Banten	0	0	0	4.44	2.00	316.30
Bali	0	0	0	5.03	2.20	161.10
Nusa Tenggara Barat	0	0	0	3.90	1.00	218.00
Nusa Tenggara Timur	0	0	0	2.25	0	157.90
Kalimantan Barat	0	0	0	5.55	2.40	194.00
Kalimantan Tengah	0	0	0.40	8.44	8.65	164.60
Kalimantan Selatan	0	0	0	6.70	6.10	136.10
Kalimantan Timur	0	0	0	6.07	5.60	153.50
Kalimantan Utara	0	0	0.40	6.97	7.00	157.20
Sulawesi Utara	0	0	0	6.08	6.00	206.00
Sulawesi Tengah	0	0	0	5.63	4.00	214.90
Sulawesi Selatan	0	0	0	7.96	5.60	270.00
Sulawesi Tenggara	0	0	0	5.41	4.00	163.00
Gorontalo	0	0	0	3.85	2.00	142.00
Sulawesi Barat	0	0	0	3.73	1.00	505.00
Maluku	0	0	0	5.55	4.00	272.00
Maluku Utara	0	0	0	5.20	3.00	188.00
Papua Barat	0	0	0.60	7.93	8.00	248.80
Papua	0	0	0.20	6.56	6.10	179.40



(a) with outliers



(b) without outliers

Figure 12: Boxplot of Daily Rainfall Values for Each Province after Null Substitution

2.3.6 Bootstrap

Here we summarize the values after Bootstrap is implemented. In Table 13, the maximum values are more or less the same as for the other methods, which means the range of the data stays the same. Overall, the values are quite similar to the daily data before imputation. There is only one province with non-zero first quartile, which is Kalimantan Tengah, the same as the real data statistics shown in Table 1. Median, mean, and third quartile values are also very similar. This is surprising, considering that Bootstrap was supposed to be very random, and yet it does not change the statistics much. Although, as expected, Bootstrap has higher values than Null Substitution. Compared to Mean Substitution, Linear Interpolation, and LOCF Imputation, however, Bootstrap has smaller values. Meanwhile, compared to Median Substitution, the values vary. Median Substitution has some higher values, but Bootstrap also has some higher values.

Figure 14 has the boxplot of monthly accumulated imputed rainfall data values from Bootstrap imputation. Figure 14a shows that there are a lot of outliers, but they are not too extreme. The range of the data is 0-1200 mm/month with outliers and 0-1000 mm/month without outliers as shown in Figure 14b. Actually, without outliers, the range can be even lower (0-800 mm/month) if we exclude two provinces: Sumatera Barat and Sulawesi Selatan.

The histogram of the monthly imputed data in Figure 13 shows that most of the rainfall values are around 200 mm/month. There are also quite a lot of zero values, but not as much as the 100-200 mm/month and 200-300 mm/month intervals. This is unlike the other methods. All of the previous methods always have 0 with the highest probability of showing, but not with Bootstrap. Just like the previous methods, the data is also positively skewed. So far, Bootstrap has the least amount of zero values compared to all the previous methods as can also be seen from Table 12, with only 227 zero values (around 3.09%). The range of the data is similar to Null Substitution, the median value is similar to Mean Substitution, but the mean value is similar to Linear Interpolation.

Statistic	Value
n	7344
Min	0.00
Q1	116.57
Median	209.40
Mean	228.33
Q3	324.70
Max	1085.00
0's	227

Table 12: Summary of Monthly Rainfall Values after Bootstrap

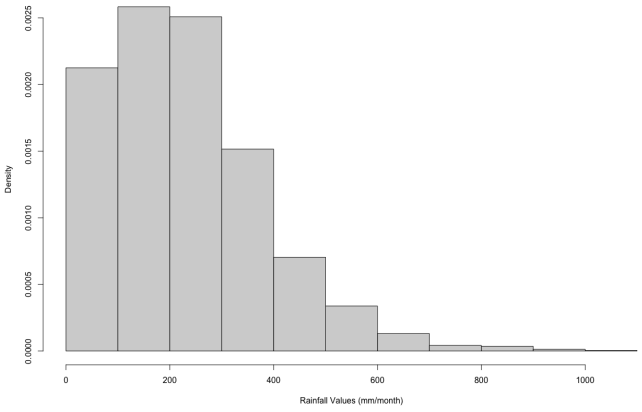
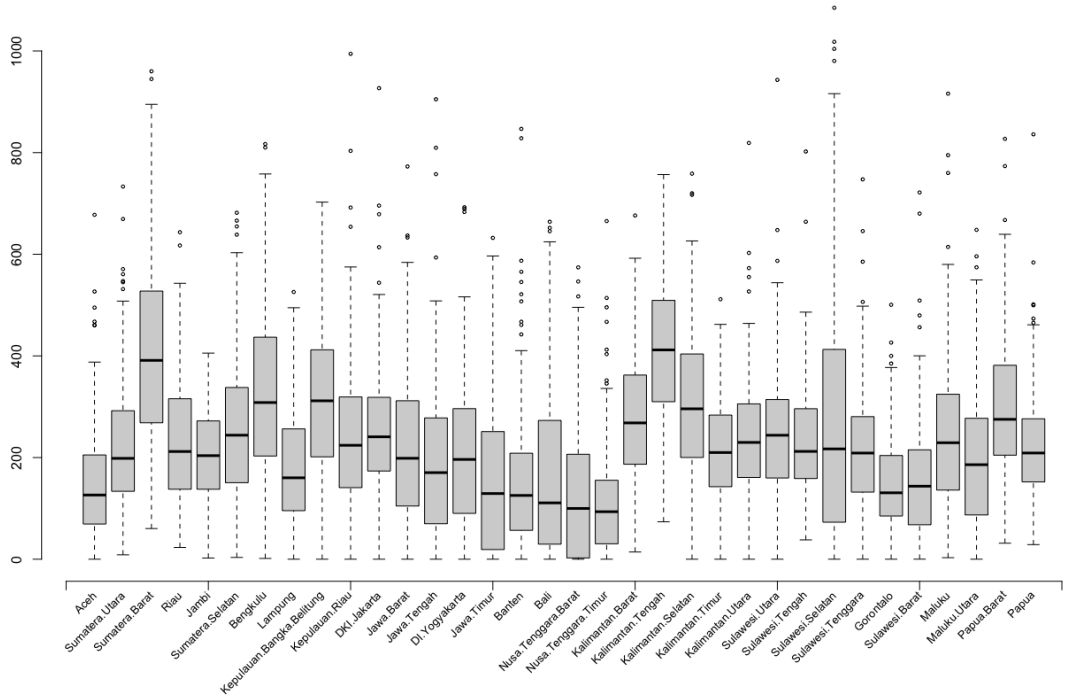


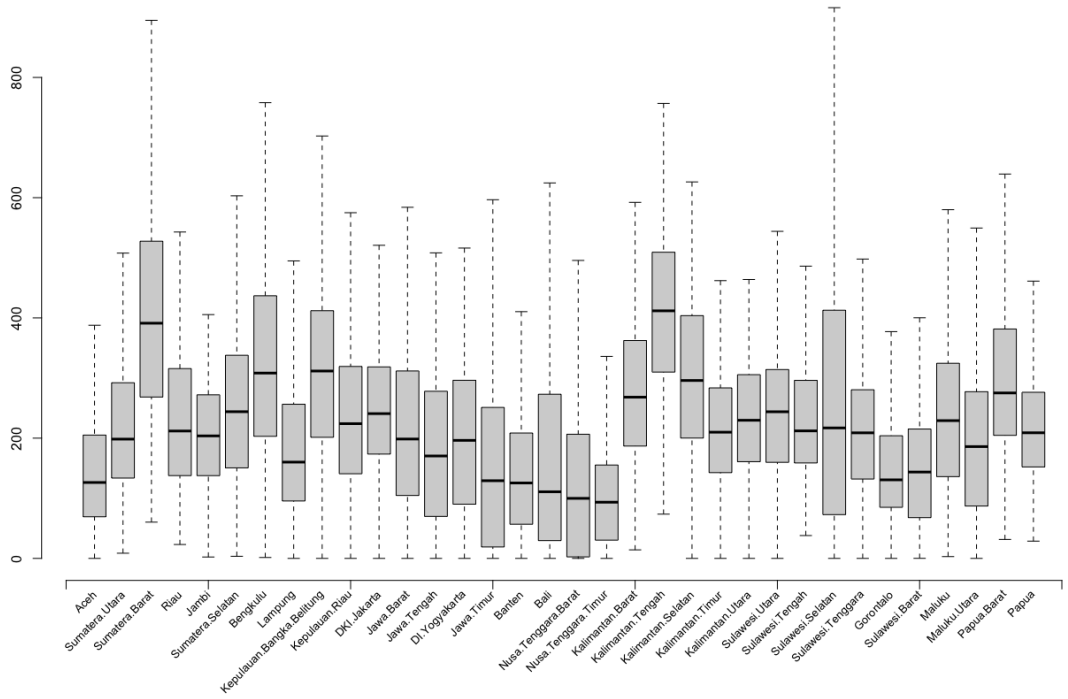
Figure 13: Histogram of Monthly Rainfall Values after Bootstrap

Table 13: Daily Rainfall Data Summary for Each Province after Bootstrap

Province	**Min**	**Q1**	**Median**	**Mean**	**Q3**	**Max**
Aceh	0	0	0	4.86	3.00	188.80
Sumatera Utara	0	0	0.20	7.40	6.70	170.80
Sumatera Barat	0	0	1.70	13.50	14.30	470.00
Riau	0	0	0.30	7.59	7.20	201.50
Jambi	0	0	0.20	6.70	6.00	140.40
Sumatera Selatan	0	0	0.60	8.35	9.00	214.10
Bengkulu	0	0	1.00	10.64	11.00	236.00
Lampung	0	0	0	5.83	4.95	204.90
Kepulauan Bangka Belitung	0	0	2.30	10.23	13.50	196.10
Kepulauan Riau	0	0	0.50	7.74	7.00	279.50
DKI Jakarta	0	0	0.30	8.33	8.60	305.00
Jawa Barat	0	0	0.80	7.18	8.40	122.90
Jawa Tengah	0	0	0	6.23	5.00	170.40
DI Yogyakarta	0	0	0	6.54	5.00	364.10
Jawa Timur	0	0	0	5.02	3.00	142.50
Banten	0	0	0	5.04	3.00	316.30
Bali	0	0	0	5.32	2.90	161.10
Nusa Tenggara Barat	0	0	0	4.24	1.00	218.00
Nusa Tenggara Timur	0	0	0	3.68	1.30	157.90
Kalimantan Barat	0	0	1.00	9.22	9.70	194.00
Kalimantan Tengah	0	0.60	4.70	13.46	18.00	164.60
Kalimantan Selatan	0	0	2.40	9.74	12.40	136.10
Kalimantan Timur	0	0	0.70	7.04	7.50	153.50
Kalimantan Utara	0	0	1.00	7.85	8.70	157.20
Sulawesi Utara	0	0	2.00	8.16	9.00	206.00
Sulawesi Tengah	0	0	1.00	7.71	7.00	214.90
Sulawesi Selatan	0	0	0	9.19	8.00	270.00
Sulawesi Tenggara	0	0	1.00	6.95	7.00	163.00
Gorontalo	0	0	0	4.89	4.00	142.00
Sulawesi Barat	0	0	0	5.04	3.00	505.00
Maluku	0	0	1.00	7.93	8.00	272.00
Maluku Utara	0	0	0	6.20	5.00	188.00
Papua Barat	0	0	2.00	9.77	11.10	248.80
Papua	0	0	1.00	7.44	7.60	179.40



(a) with outliers



(b) without outliers

Figure 14: Boxplot of Daily Rainfall Values for Each Province after Bootstrap

2.3.7 Distribution Fill

Here we summarize the values after Distribution Fill is implemented. In Table 15, we can see that even though the values themselves are not necessarily higher than Bootstrap, especially for mean and third quartile, Distribution Fill has a lot of values that are larger than zero. The median and third quartile values are mostly larger than Bootstrap, but the mean values are similar. As for maximum values, they are the same, which means that the data range is also the same.

Figure 16 shows the boxplot of the monthly imputed rainfall values with and without outliers. With outlier, as seen from Figure 16a, the data range can exceed 1000 mm/month, although that only happens in two provinces: Kepulauan Riau and Sulawesi Selatan. The other provinces are still in the 0-1000 mm/month range. However, without outliers as shows in Figure 16b, the range can decrease to 0-800 mm/month if we do not include Sumatera Barat and Sulawesi Selatan. With these two provinces, the range reaches around 900 mm/month.

Based on Table 14, we can see that Distribution Fill provides the least zero values compared to the previous methods, only 201 zero values (around 2.74%). This method also provides the second smallest data range (after Null Substitution). As for the other statistics, no methods match it perfectly, but Distribution Fill does provide higher values compared to Median Substitution and Null Substitution. As for the histogram in Figure 15, just like Bootstrap, the zero values are lower in frequency compared to 100-200 and 200-300 mm/month intervals. There is a slight difference to Bootstrap in the frequency for the second two bars in the histogram. In Figure 13, Bootstrap has 100-200 mm/month interval with the highest frequency, meanwhile Distribution Fill has 200-300 mm/month interval as the highest frequency.

Statistic	Value
n	7344
Min	0.00
Q1	122.21
Median	213.13
Mean	227.93
Q3	309.37
Max	1101.38
0's	201

Table 14: Summary of Monthly Rainfall Values after Distribution Fill

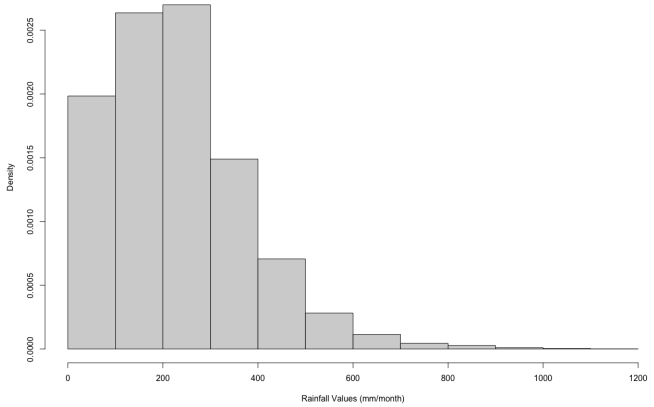
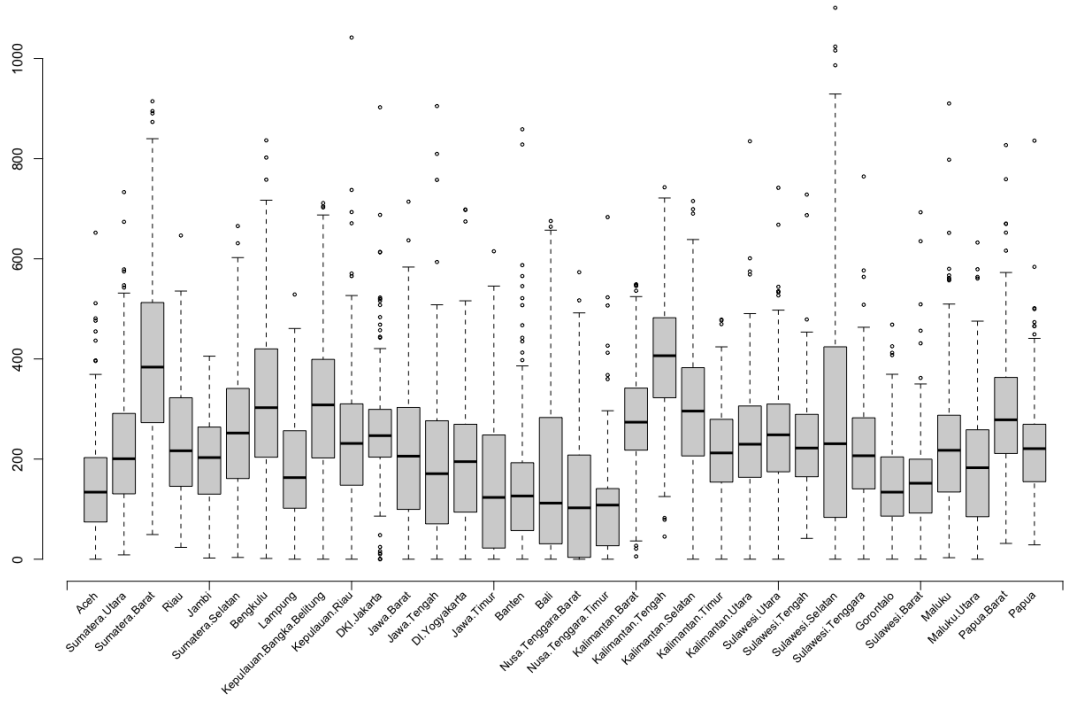


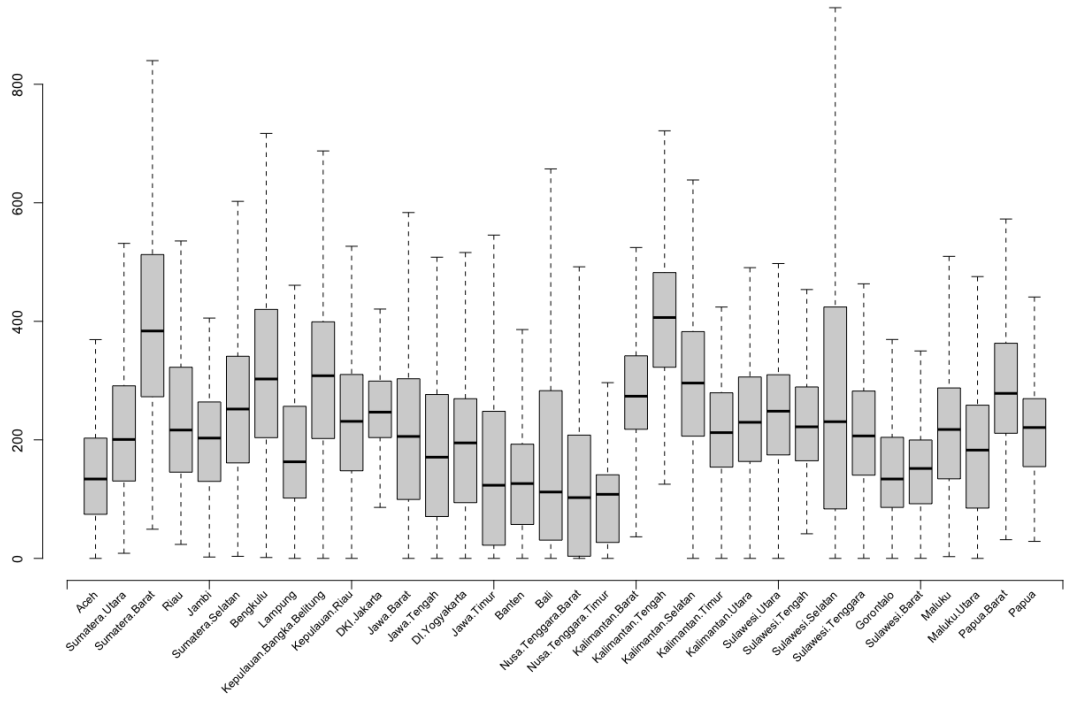
Figure 15: Histogram of Monthly Rainfall Values after Distribution Fill

Table 15: Daily Rainfall Data Summary for Each Province after Distribution Fill

Province	**Min**	**Q1**	**Median**	**Mean**	**Q3**	**Max**
Aceh	0	0	0.11	4.93	4.69	188.80
Sumatera Utara	0	0	0.60	7.38	7.50	170.80
Sumatera Barat	0	0	2.80	13.40	15.58	470.00
Riau	0	0	0.88	7.61	8.02	201.50
Jambi	0	0	0.50	6.68	6.60	140.40
Sumatera Selatan	0	0	1.56	8.36	10.11	214.10
Bengkulu	0	0	2.00	10.54	12.00	236.00
Lampung	0	0	0.50	5.85	6.00	204.90
Kepulauan Bangka Belitung	0	0.13	3.72	10.04	13.50	196.10
Kepulauan Riau	0	0	2.00	7.79	8.81	279.50
DKI Jakarta	0	0.45	4.02	8.46	11.11	305.00
Jawa Barat	0	0	1.40	7.14	8.68	122.90
Jawa Tengah	0	0	0	6.21	5.50	170.40
DI Yogyakarta	0	0	1.50	6.55	7.69	364.10
Jawa Timur	0	0	0	5.00	3.99	142.50
Banten	0	0	0	4.97	4.00	316.30
Bali	0	0	0	5.44	3.77	161.10
Nusa Tenggara Barat	0	0	0	4.22	2.00	218.00
Nusa Tenggara Timur	0	0	0.60	3.68	3.77	157.90
Kalimantan Barat	0	0.10	3.50	9.22	11.60	194.00
Kalimantan Tengah	0	1.40	6.48	13.10	17.70	164.60
Kalimantan Selatan	0	0.20	3.93	9.66	12.80	136.10
Kalimantan Timur	0	0	1.40	7.14	8.30	153.50
Kalimantan Utara	0	0	1.60	7.96	9.30	157.20
Sulawesi Utara	0	0	3.00	8.18	10.00	206.00
Sulawesi Tengah	0	0	2.00	7.62	8.50	214.90
Sulawesi Selatan	0	0	1.00	9.39	10.00	270.00
Sulawesi Tenggara	0	0	1.82	7.00	8.30	163.00
Gorontalo	0	0	0.80	4.91	5.00	142.00
Sulawesi Barat	0	0	0.90	5.14	5.00	505.00
Maluku	0	0	2.18	7.61	9.00	272.00
Maluku Utara	0	0	0.93	6.16	6.00	188.00
Papua Barat	0	0	3.00	9.68	11.63	248.80
Papua	0	0	1.60	7.54	8.48	179.40



(a) with outliers



(b) without outliers

Figure 16: Boxplot of Daily Rainfall Values for Each Province after Distribution Fill

2.3.8 MICE-PMM

Here we summarize the values after MICE-PMM is applied. Table 17 shows that the data still have the same range as the other methods. In general, MICE-PMM has higher median values compared to the real daily data. The mean values also increase for most, but there are also some cases where they are lower in MICE-PMM than in the real daily data. The third quartile values, however, are increased for all except for three provinces: Kepulauan Bangka Belitung, Kalimantan Selatan, and Kalimantan Utara. Overall, MICE-PMM's results are not similar to any of the previous methods.

According to Figure 18, the imputed data values from MICE-PMM also have outliers, but they are not too extreme. The data range between the one with outliers and the one without look almost similar. In Figure 18a, the highest outlier value is owned by Sulawesi Selatan, which does not exceed 1200 mm/month. The next highest outlier value would be around 1000 mm/month. Meanwhile, in Figure 18b, Sulawesi Selatan still has the biggest range compared to the other provinces. The boxplot actually looks pretty similar to Bootstrap in Figure 14.

This histogram of MICE-PMM's imputed data in Figure 17 looks very similar to the one that Bootstrap has in Figure 13. They have more values in the 100-200 mm/month and 200-300 mm/month intervals compared to the ones below 100 mm/month. Not just the histogram, but the statistics in Table 16 are also very similar to Bootstrap. The range of the data is only different by 0.4 point, and all the other values are close to one another. The number of 0 values is, however, not as close to Bootstrap. It is actually between Bootstrap and Distribution Fill, with only 211 values (around 2.87%).

Statistic	Value
n	7344
Min	0.00
Q1	115.13
Median	210.97
Mean	226.50
Q3	311.52
Max	1084.60
0's	211

Table 16: Summary of Monthly Rainfall Values after MICE-PMM

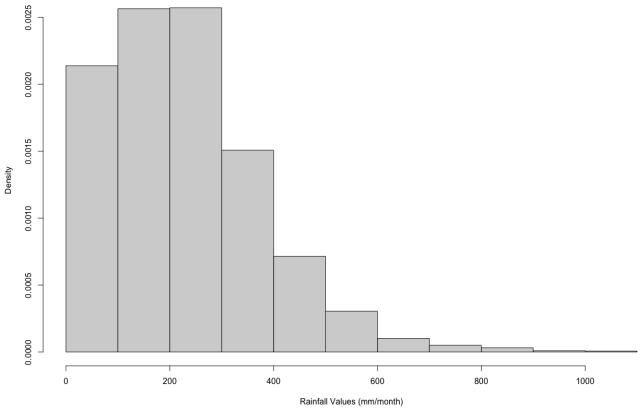
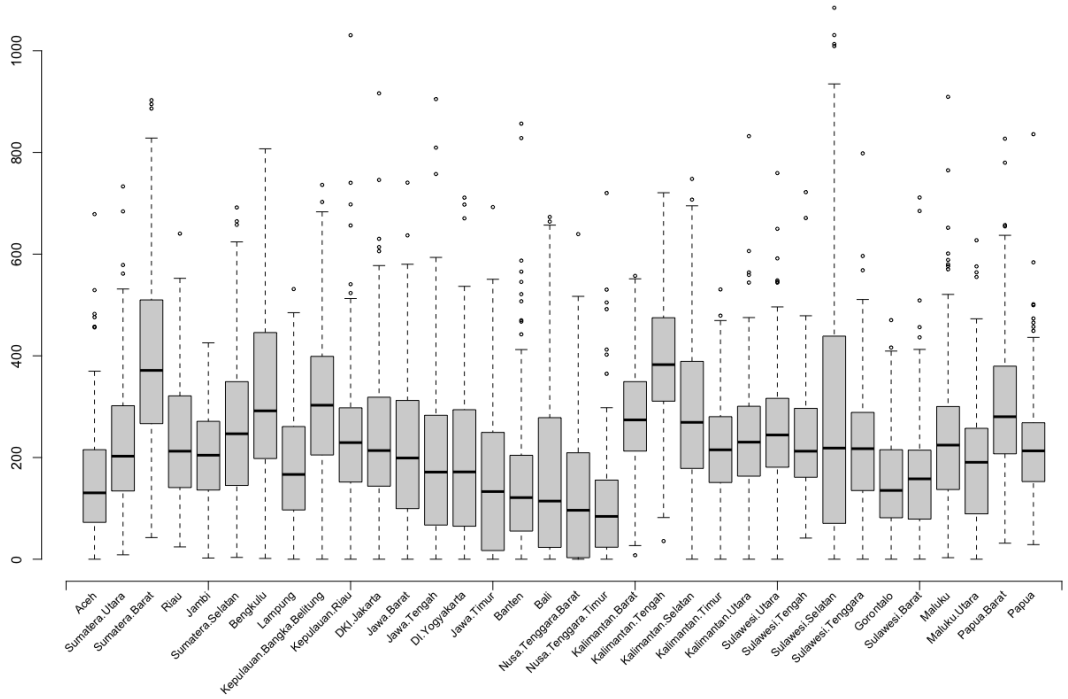


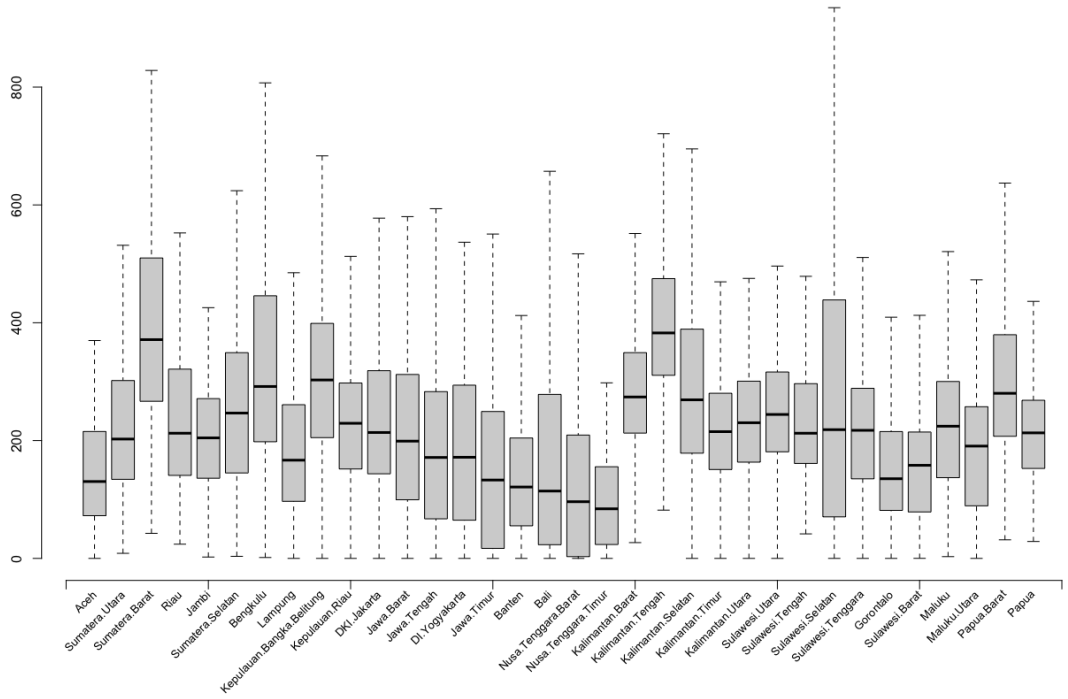
Figure 17: Histogram of Monthly Rainfall Values after MICE-PMM

Table 17: Daily Rainfall Data Summary for Each Province after MICE-PMM

Province	**Min**	**Q1**	**Median**	**Mean**	**Q3**	**Max**
Aceh	0	0	0	5.08	5.00	188.80
Sumatera Utara	0	0	0.60	7.42	7.52	170.80
Sumatera Barat	0	0	2.70	13.17	15.10	470.00
Riau	0	0	0.80	7.62	8.05	201.50
Jambi	0	0	0.50	6.72	6.80	140.40
Sumatera Selatan	0	0	1.46	8.35	10.00	214.10
Bengkulu	0	0	2.00	10.57	12.00	236.00
Lampung	0	0	0.50	5.90	6.00	204.90
Kepulauan Bangka Belitung	0	0.10	4.10	9.90	13.10	196.10
Kepulauan Riau	0	0	1.80	7.60	8.16	279.50
DKI Jakarta	0	0	3.16	7.98	9.90	305.00
Jawa Barat	0	0	1.40	7.07	8.70	122.90
Jawa Tengah	0	0	0	6.28	5.80	170.40
DI Yogyakarta	0	0	0.70	6.40	7.33	364.10
Jawa Timur	0	0	0	5.17	4.00	142.50
Banten	0	0	0	4.94	3.80	316.30
Bali	0	0	0	5.42	3.40	161.10
Nusa Tenggara Barat	0	0	0	4.21	2.00	218.00
Nusa Tenggara Timur	0	0	0	3.57	3.26	157.90
Kalimantan Barat	0	0	3.60	9.26	11.80	194.00
Kalimantan Tengah	0	1.76	7.10	12.64	16.50	164.60
Kalimantan Selatan	0	0.20	4.00	9.22	11.90	136.10
Kalimantan Timur	0	0	1.10	7.13	8.20	153.50
Kalimantan Utara	0	0	1.50	7.83	9.26	157.20
Sulawesi Utara	0	0	3.10	8.25	10.40	206.00
Sulawesi Tengah	0	0	2.00	7.55	8.20	214.90
Sulawesi Selatan	0	0	0.80	9.43	10.00	270.00
Sulawesi Tenggara	0	0	1.72	7.03	8.39	163.00
Gorontalo	0	0	0.50	4.95	5.00	142.00
Sulawesi Barat	0	0	0.42	5.26	4.96	505.00
Maluku	0	0	2.36	7.68	9.00	272.00
Maluku Utara	0	0	0.60	6.19	6.32	188.00
Papua Barat	0	0	3.20	9.73	11.80	248.80
Papua	0	0	1.70	7.48	8.37	179.40



(a) with outliers



(b) without outliers

Figure 18: Boxplot of Daily Rainfall Values for Each Province after MICE-PMM

2.3.9 MICE-CART

Here we summarize the values after MICE-CART is applied. In Table 19, with both being MICE, the values that MICE-CART has are pretty similar to the ones in MICE-PMM. Some of the values are lower in MICE-CART like the third quartile of Kepulauan Bangka Belitung, but some values are also higher like the third quartile of Sumatera Utara. Even so, the differences are not significant. The data range is again not changed.

In Figure 20, we can see a similar case happening to MICE-PMM. MICE-CART also doesn't change the range of the data even after being imputed and aggregated. With outliers, as shown in Figure 20a the data still ranges up to 1200 mm/month. The outlier values are slightly bigger than for MICE-PMM, but it does not change the range much. Without outliers, as shown in Figure 20b, the data still ranges 0-1000 mm/month. Sulawesi Selatan, as always has the biggest range. With outliers, the highest outliers are owned by Sulawesi Selatan and Kepulauan Riau.

The statistics from MICE-CART as shown in Table 18 have the closest values to Bootstrap and MICE-PMM. When MICE-CART's statistics are relatively high, they are closer to Bootstrap, but when the values are lower, they are closer to MICE-PMM. It does not have too many zero values, only 219 values (around 2.98%), which is between Bootstrap and MICE-PMM again. The histogram in Figure 19 also has the same situation: it looks similar to Bootstrap and MICE-PMM.

Statistic	Value
n	7344
Min	0.00
Q1	113.09
Median	210.74
Mean	227.50
Q3	316.60
Max	1097.80
0's	219

Table 18: Summary of Monthly Rainfall Values after MICE-CART

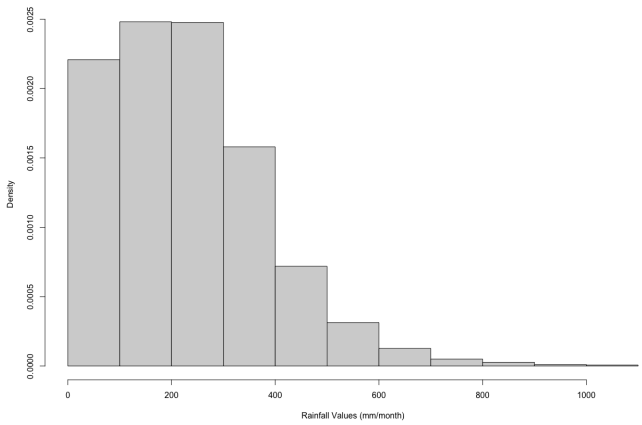
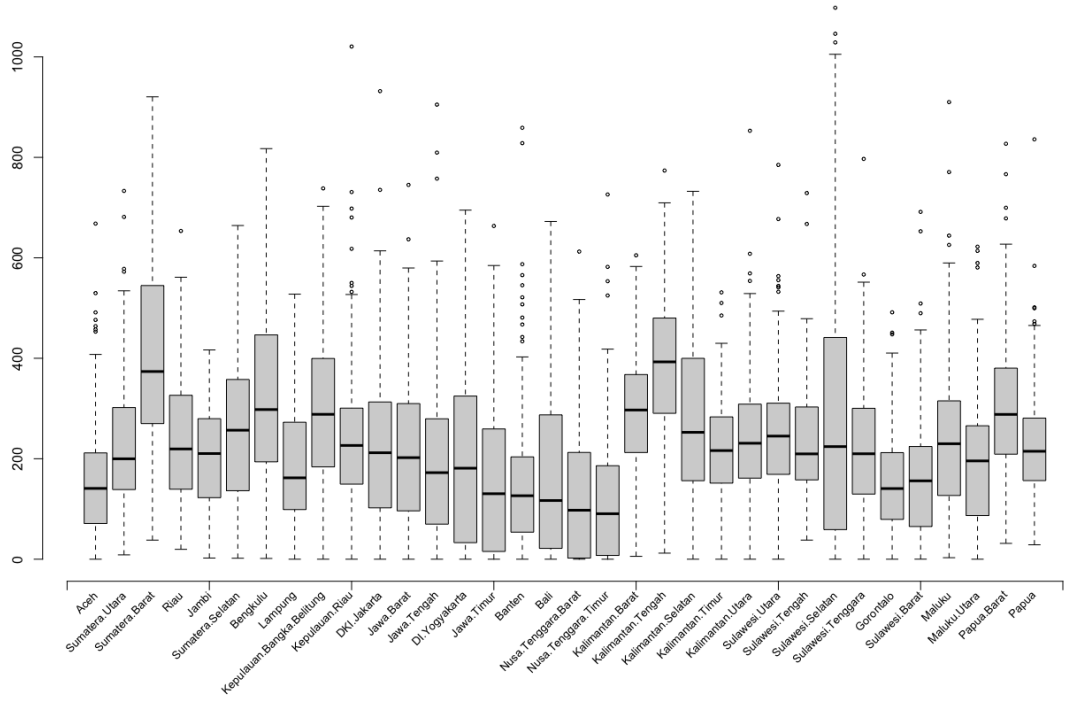


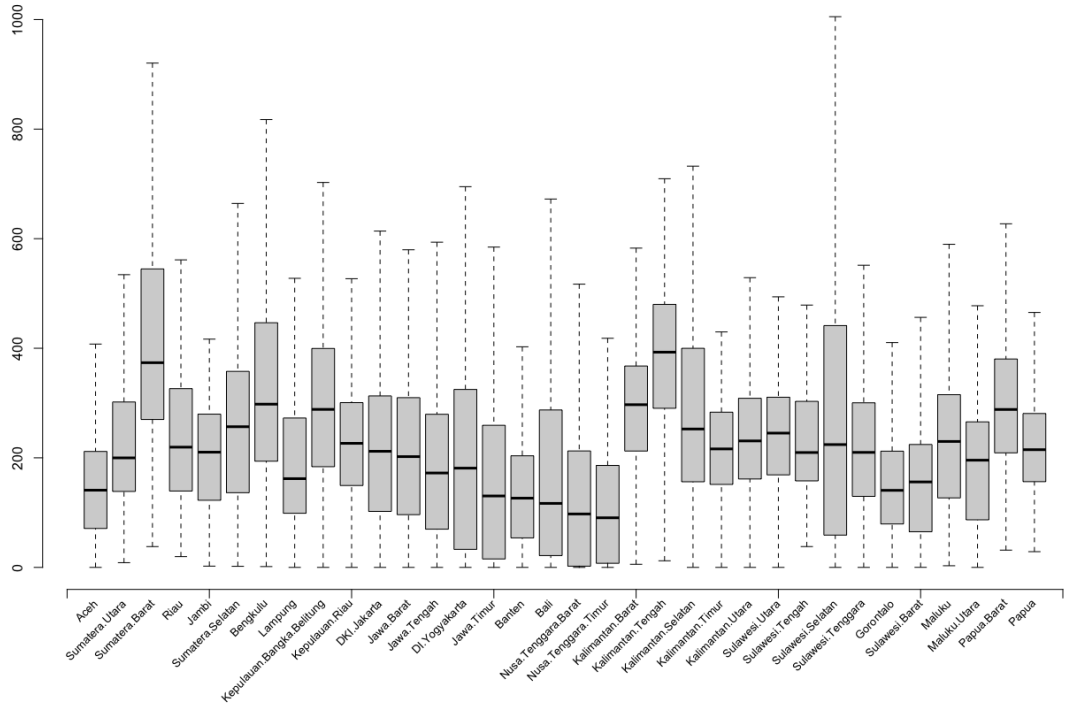
Figure 19: Histogram of Monthly Rainfall Values after MICE-CART

Table 19: Daily Rainfall Data Summary for Each Province after MICE-CART

Province	**Min**	**Q1**	**Median**	**Mean**	**Q3**	**Max**
Aceh	0	0	0	5.13	5.00	188.80
Sumatera Utara	0	0	0.60	7.44	7.57	170.80
Sumatera Barat	0	0	2.80	13.33	15.54	470.00
Riau	0	0	0.80	7.71	8.50	201.50
Jambi	0	0	0.50	6.79	7.00	140.40
Sumatera Selatan	0	0	1.40	8.43	10.19	214.10
Bengkulu	0	0	2.00	10.58	12.00	236.00
Lampung	0	0	0.40	5.98	6.20	204.90
Kepulauan Bangka Belitung	0	0	3.52	9.69	13.00	196.10
Kepulauan Riau	0	0	1.70	7.71	8.48	279.50
DKI Jakarta	0	0	2.20	7.47	9.07	305.00
Jawa Barat	0	0	1.30	7.14	8.90	122.90
Jawa Tengah	0	0	0	6.30	5.85	170.40
DI Yogyakarta	0	0	0.40	6.52	7.46	364.10
Jawa Timur	0	0	0	5.18	4.10	142.50
Banten	0	0	0	4.99	3.94	316.30
Bali	0	0	0	5.46	3.40	161.10
Nusa Tenggara Barat	0	0	0	4.21	2.00	218.00
Nusa Tenggara Timur	0	0	0	3.79	3.28	157.90
Kalimantan Barat	0	0	3.80	9.59	12.41	194.00
Kalimantan Tengah	0	1.60	6.74	12.58	16.55	164.60
Kalimantan Selatan	0	0.14	3.40	8.97	11.40	136.10
Kalimantan Timur	0	0	1.30	7.18	8.50	153.50
Kalimantan Utara	0	0	1.52	7.90	9.40	157.20
Sulawesi Utara	0	0	3.00	8.23	10.27	206.00
Sulawesi Tengah	0	0	2.00	7.71	9.00	214.90
Sulawesi Selatan	0	0	0.60	9.47	10.00	270.00
Sulawesi Tenggara	0	0	1.60	7.07	8.60	163.00
Gorontalo	0	0	0.40	4.98	5.00	142.00
Sulawesi Barat	0	0	0.22	5.21	5.00	505.00
Maluku	0	0	2.00	7.75	9.00	272.00
Maluku Utara	0	0	0.60	6.34	6.99	188.00
Papua Barat	0	0	3.10	9.78	11.89	248.80
Papua	0	0	1.70	7.52	8.50	179.40



(a) with outliers



(b) without outliers

Figure 20: Boxplot of Daily Rainfall Values for Each Province after MICE-CART

2.3.10 MICE-LASSO

Here we summarize the values after MICE-LASSO is applied. In Table 21, the values are similar to both MICE-PMM and MICE-CART. However, MICE-LASSO has, by far, the highest third quartile values among the previous MICE methods. The maximum values are also the same as the original data, showing that the data range does not change.

According to Figure 22, the range of the monthly aggregated rainfall values is pretty similar. However, Figure 22a shows that there are more outliers compared to MICE-CART, even though the difference is not significant. There are 4 data points that exceed 1000 mm/month. Without outliers, as shown in Figure 22b, the range of the imputed data stays in 0-1000 mm/month, just like the other two MICE methods.

The histogram of MICE-LASSO in Figure 21 looks similar to the previous MICE methods. However, it also looks similar to Distribution Fill, especially the 100-200 and 200-300 mm/month intervals. According to the data summary in Table 20, the third quartile value and maximum value are between MICE-PMM and MICE-CART, but not with the other statistics. The first quartile, median value, and mean value are more similar to Distribution Fill. This method also only has 210 zero values (around 2.86%), which is close to the lowest (Distribution Fill with only 201 zero values).

Statistic	Value
n	7344
Min	0.00
Q1	122.10
Median	214.88
Mean	230.20
Q3	314.98
Max	1092.79
0's	210

Table 20: Summary of Monthly Rainfall Values after MICE-LASSO

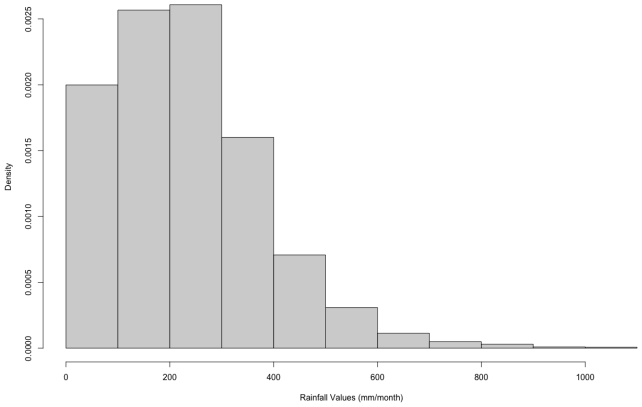
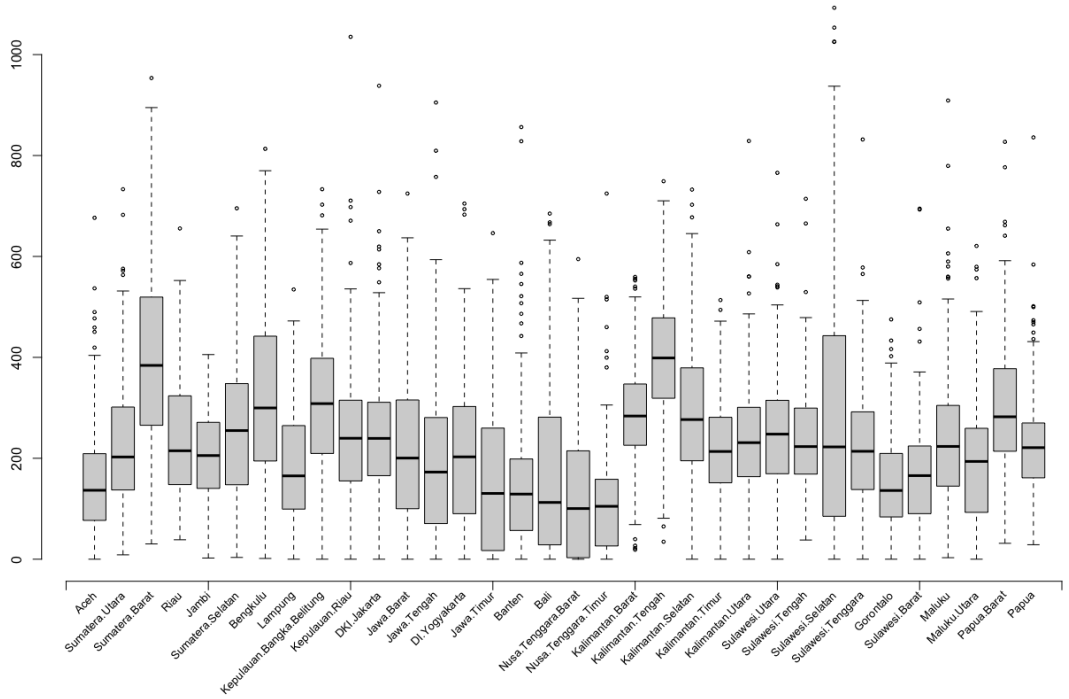


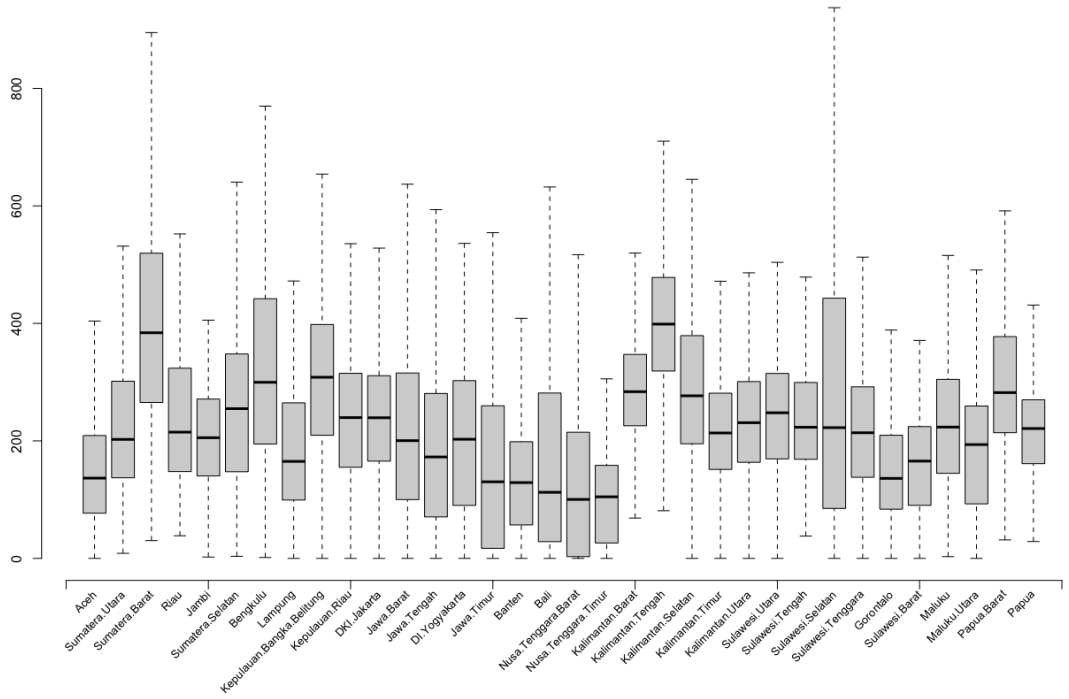
Figure 21: Histogram of Monthly Rainfall Values after MICE-LASSO

Table 21: Daily Rainfall Data Summary for Each Province after MICE-LASSO

Province	**Min**	**Q1**	**Median**	**Mean**	**Q3**	**Max**
Aceh	0	0	0	5.10	5.39	188.80
Sumatera Utara	0	0	0.50	7.44	7.80	170.80
Sumatera Barat	0	0	2.60	13.36	16.40	470.00
Riau	0	0	0.70	7.68	8.70	201.50
Jambi	0	0	0.40	6.72	7.39	140.40
Sumatera Selatan	0	0	1.50	8.39	10.60	214.10
Bengkulu	0	0	2.00	10.66	13.00	236.00
Lampung	0	0	0.20	5.97	6.76	204.90
Kepulauan Bangka Belitung	0	0	4.30	10.00	14.08	196.10
Kepulauan Riau	0	0	1.80	7.92	10.22	279.50
DKI Jakarta	0	0	4.10	8.45	12.20	305.00
Jawa Barat	0	0	1.40	7.12	9.05	122.90
Jawa Tengah	0	0	0	6.27	6.00	170.40
DI Yogyakarta	0	0	0.99	6.64	8.75	364.10
Jawa Timur	0	0	0	5.22	4.60	142.50
Banten	0	0	0	5.05	4.50	316.30
Bali	0	0	0	5.49	4.00	161.10
Nusa Tenggara Barat	0	0	0	4.25	2.00	218.00
Nusa Tenggara Timur	0	0	0	3.85	4.80	157.90
Kalimantan Barat	0	0	4.31	9.44	13.29	194.00
Kalimantan Tengah	0	1.28	7.71	12.85	17.90	164.60
Kalimantan Selatan	0	0	4.40	9.45	12.89	136.10
Kalimantan Timur	0	0	1.30	7.18	8.60	153.50
Kalimantan Utara	0	0	1.50	7.89	9.60	157.20
Sulawesi Utara	0	0	3.34	8.21	11.00	206.00
Sulawesi Tengah	0	0	2.00	7.76	9.79	214.90
Sulawesi Selatan	0	0	0.80	9.60	11.58	270.00
Sulawesi Tenggara	0	0	1.70	7.15	9.24	163.00
Gorontalo	0	0	0.10	5.02	6.00	142.00
Sulawesi Barat	0	0	0.10	5.51	6.38	505.00
Maluku	0	0	2.42	7.77	10.44	272.00
Maluku Utara	0	0	0.40	6.30	7.00	188.00
Papua Barat	0	0	3.20	9.84	12.64	248.80
Papua	0	0	1.60	7.56	9.04	179.40



(a) with outliers



(b) without outliers

Figure 22: Boxplot of Daily Rainfall Values for Each Province after MICE-LASSO

2.3.11 MICE-RI

Here we summarize the values after MICE-RI is applied. In Table 23, MICE-RI seems to be slightly higher in values compared to the other MICE-methods, especially MICE-PMM and MICE-CART. Compared to MICE-LASSO, there are some higher values, but there are also some lower values for median, mean, and third quartile values. The range of the data is still the same as the original data because the maximum values do not change.

After being aggregated monthly, the boxplot of MICE-RI values are shown in Figure 24. There are not that many outliers as shown in Figure 24a, but the range of the data is the widest compared to the other MICE methods by far. In Figure 24b, the boxplot shows that without outliers, the highest value still exceeds 1000 mm/month. Unlike the other MICE methods, the data range of MICE-RI without outliers is 0-1100 mm/month.

As shown in Table 22, MICE-RI has similar first quartile value to MICE-LASSO, similar mean value to LOCF Imputation, similar third quartile value also to LOCF Imputation, and similar maximum value to MICE-CART. Overall, MICE-RI has the highest statistics values except for third quartile and maximum values. MICE-RI only has 208 zero values (2.83%), which is not that many and almost similar to MICE-LASSO. As for the histogram in Figure 23, it looks very similar to MICE-LASSO. In general, the results of these two MICE methods have a lot of similarities. MICE-RI's histogram also looks like the histogram of Distribution Fill.

Statistic	Value
n	7344
Min	0.00
Q1	124.90
Median	220.20
Mean	235.17
Q3	321.81
Max	1098.95
0's	208

Table 22: Summary of Monthly Rainfall Values after MICE-RI

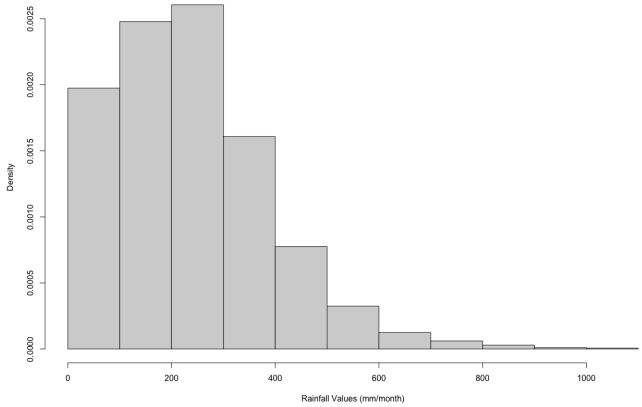
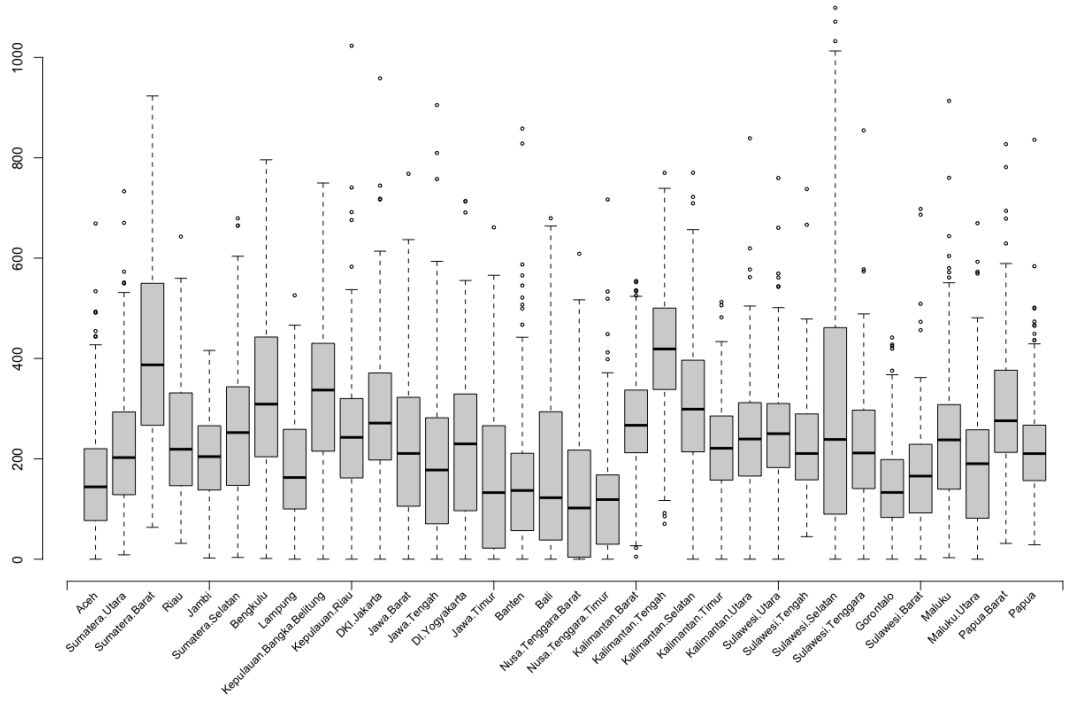


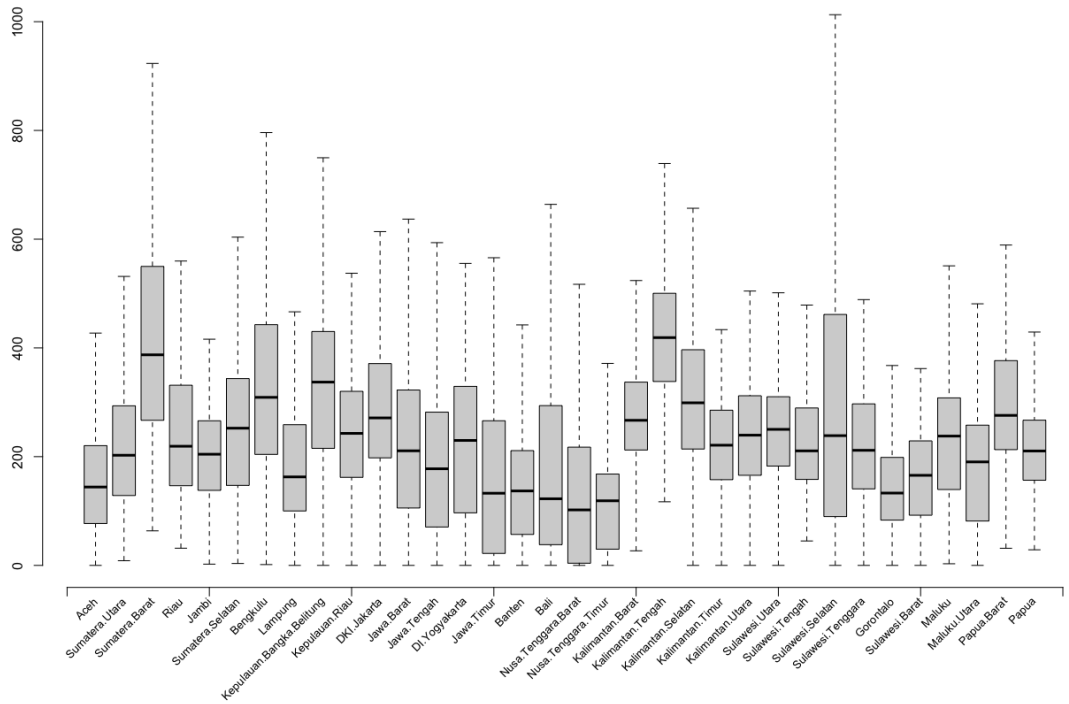
Figure 23: Histogram of Monthly Rainfall Values after MICE-RI

Table 23: Daily Rainfall Data Summary for Each Province after MICE-RI

Province	**Min**	**Q1**	**Median**	**Mean**	**Q3**	**Max**
Aceh	0	0	0	5.29	6.00	188.80
Sumatera Utara	0	0	0.40	7.31	7.50	170.80
Sumatera Barat	0	0	3.00	13.76	17.42	470.00
Riau	0	0	0.80	7.83	9.10	201.50
Jambi	0	0	0.40	6.71	7.26	140.40
Sumatera Selatan	0	0	1.30	8.33	10.60	214.10
Bengkulu	0	0	2.00	10.69	13.00	236.00
Lampung	0	0	0.10	5.83	6.20	204.90
Kepulauan Bangka Belitung	0	0	5.20	10.85	16.20	196.10
Kepulauan Riau	0	0	2.00	8.18	10.60	279.50
DKI Jakarta	0	0	5.88	9.63	14.43	305.00
Jawa Barat	0	0	1.60	7.41	10.00	122.90
Jawa Tengah	0	0	0	6.40	6.60	170.40
DI Yogyakarta	0	0	1.50	7.41	10.53	364.10
Jawa Timur	0	0	0	5.27	4.60	142.50
Banten	0	0	0	5.19	5.00	316.30
Bali	0	0	0	5.77	4.80	161.10
Nusa Tenggara Barat	0	0	0	4.30	2.00	218.00
Nusa Tenggara Timur	0	0	0	4.07	5.30	157.90
Kalimantan Barat	0	0	3.60	9.07	12.32	194.00
Kalimantan Tengah	0	1.50	8.80	13.46	18.90	164.60
Kalimantan Selatan	0	0	4.80	9.84	13.90	136.10
Kalimantan Timur	0	0	1.30	7.27	8.98	153.50
Kalimantan Utara	0	0	1.60	8.09	10.40	157.20
Sulawesi Utara	0	0	3.40	8.27	11.00	206.00
Sulawesi Tengah	0	0	2.00	7.57	9.00	214.90
Sulawesi Selatan	0	0	1.00	9.95	12.64	270.00
Sulawesi Tenggara	0	0	1.71	7.19	9.23	163.00
Gorontalo	0	0	0	4.88	5.52	142.00
Sulawesi Barat	0	0	0	5.50	6.38	505.00
Maluku	0	0	2.23	7.88	10.89	272.00
Maluku Utara	0	0	0.26	6.26	7.00	188.00
Papua Barat	0	0	3.00	9.77	12.50	248.80
Papua	0	0	1.50	7.43	8.61	179.40



(a) with outliers



(b) without outliers

Figure 24: Boxplot of Daily Rainfall Values for Each Province after MICE-RI

2.3.12 MICE-SAMPLE

Here we summarize the values after MICE-SAMPLE is applied. In Table 25, there are a few third quartile values that are lower than the mean values. The median values are also mostly larger than 0. As usual, the maximum values have no differences, which means the data range also stays the same. Overall, the values are the closest to MICE-PMM compared to the other MICE methods. Compared to all other methods, MICE-SAMPLE is the closest to Distribution Fill.

In Figure 26, there are the boxplot of the aggregated imputed data with and without outliers using MICE-SAMPLE. In Figure 26a, there are quite a few outlier values that exceed 1000 mm/month. Without outliers as shown in Figure 26b, the overall range is still 0-1000 mm/month. The only provinces that exceed the 0-800 mm/month range are Sulawesi Selatan and Sumatera Barat. Without these two provinces, the range would be 0-800 mm/month.

MICE-SAMPLE only has 203 zero values (around 2.76%), as shown in Table 24. It is the second least amount of zero values (after Distribution Fill) and is the least amount of zero values amongst MICE methods. The other statistics of MICE SAMPLE, aside from maximum value, also look very similar to Distribution Fill, but not as close to MICE-PMM as one would expect from the daily values. The maximum value is still the closest to Distribution Fill, but not that close since the value itself is between Distribution Fill and Median Substitution. As for the histogram in Figure 25, it is similar to Distribution Fill again with a slight height difference between the 100-200 mm/month and 200-300 mm/month intervals.

Statistic	Value
n	7344
Min	0.00
Q1	122.19
Median	212.36
Mean	227.74
Q3	309.11
Max	1124.60
0's	203

Table 24: Summary of Monthly Rainfall Values after MICE-SAMPLE

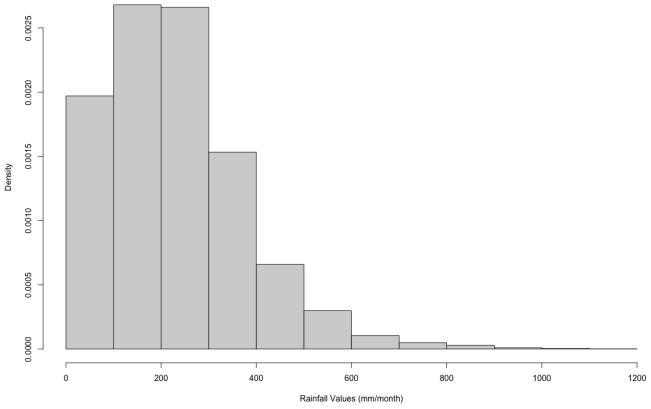
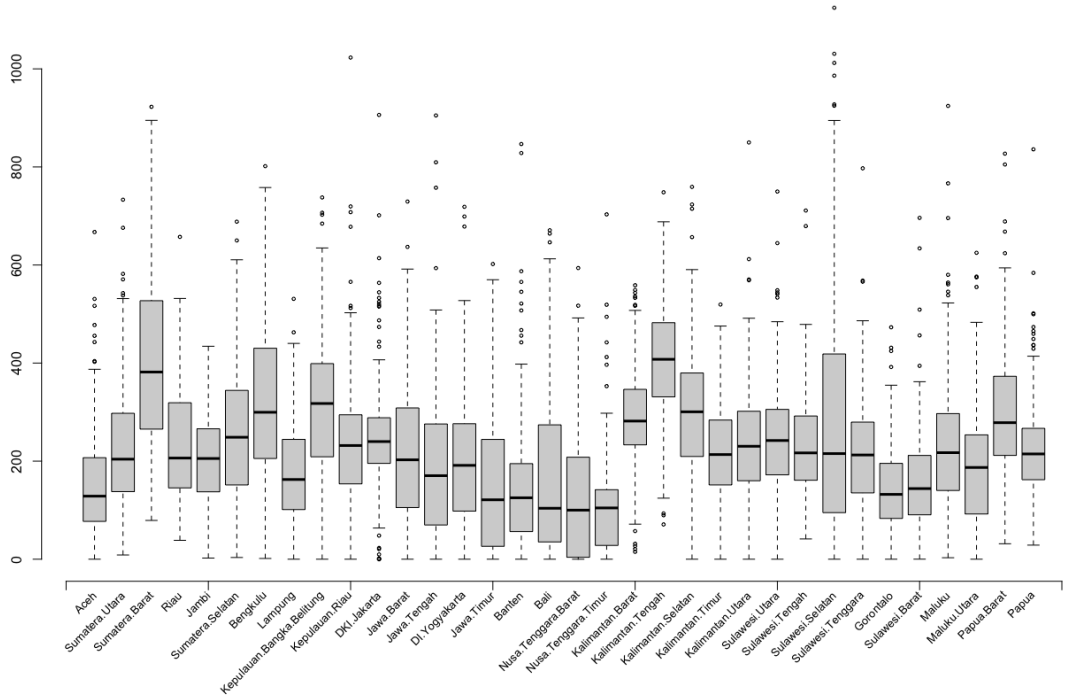


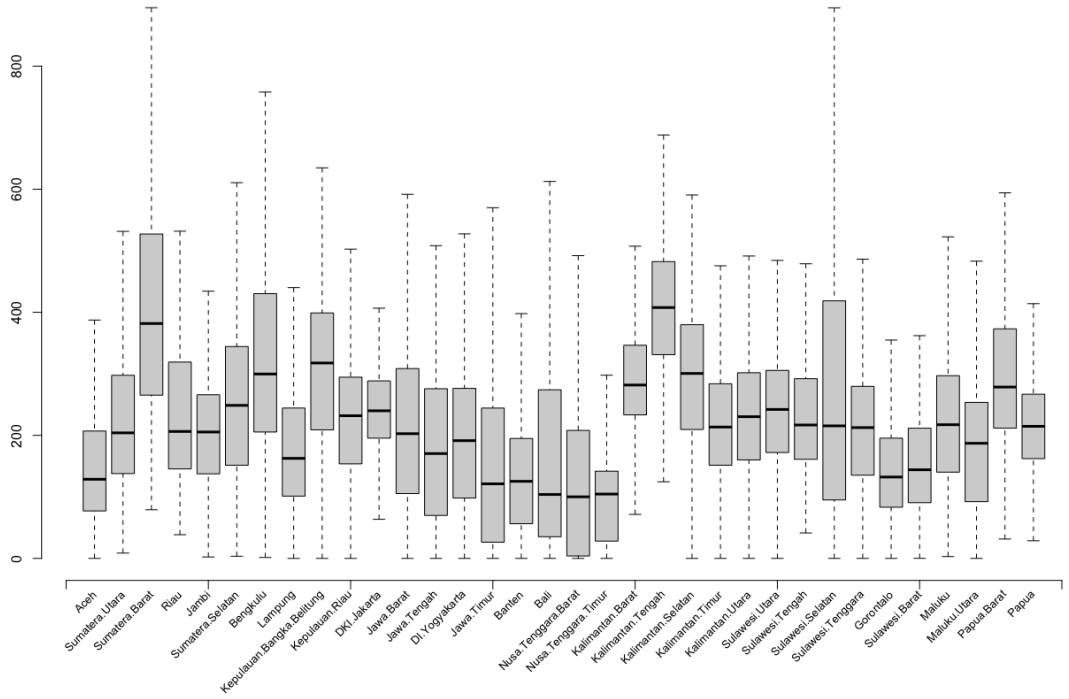
Figure 25: Histogram of Monthly Rainfall Values after MICE-SAMPLE

Table 25: Daily Rainfall Data Summary for Each Province after MICE-SAMPLE

Province	**Min**	**Q1**	**Median**	**Mean**	**Q3**	**Max**
Aceh	0	0	0	4.98	4.72	188.80
Sumatera Utara	0	0	0.60	7.43	7.52	170.80
Sumatera Barat	0	0	3.00	13.42	15.70	470.00
Riau	0	0	0.80	7.63	8.10	201.50
Jambi	0	0	0.50	6.68	6.70	140.40
Sumatera Selatan	0	0	1.60	8.31	10.00	214.10
Bengkulu	0	0	2.00	10.50	12.00	236.00
Lampung	0	0	0.40	5.81	6.00	204.90
Kepulauan Bangka Belitung	0	0.20	4.50	10.11	13.34	196.10
Kepulauan Riau	0	0	2.00	7.73	8.50	279.50
DKI Jakarta	0	0.30	4.10	8.28	10.74	305.00
Jawa Barat	0	0	1.50	7.11	8.81	122.90
Jawa Tengah	0	0	0	6.20	5.55	170.40
DI Yogyakarta	0	0	1.20	6.55	7.80	364.10
Jawa Timur	0	0	0	5.04	3.90	142.50
Banten	0	0	0	4.94	3.90	316.30
Bali	0	0	0	5.42	3.60	161.10
Nusa Tenggara Barat	0	0	0	4.21	2.00	218.00
Nusa Tenggara Timur	0	0	0.18	3.66	3.56	157.90
Kalimantan Barat	0	0.06	4.00	9.40	12.36	194.00
Kalimantan Tengah	0	1.98	8.00	13.14	17.50	164.60
Kalimantan Selatan	0	0.30	4.60	9.67	13.00	136.10
Kalimantan Timur	0	0	1.46	7.10	8.43	153.50
Kalimantan Utara	0	0	1.60	7.88	9.30	157.20
Sulawesi Utara	0	0	3.20	8.11	10.00	206.00
Sulawesi Tengah	0	0	2.00	7.55	8.70	214.90
Sulawesi Selatan	0	0	1.00	9.34	10.00	270.00
Sulawesi Tenggara	0	0	2.00	7.03	8.65	163.00
Gorontalo	0	0	0.64	4.89	5.00	142.00
Sulawesi Barat	0	0	0.56	5.16	5.00	505.00
Maluku	0	0	2.40	7.64	9.00	272.00
Maluku Utara	0	0	0.80	6.20	6.20	188.00
Papua Barat	0	0	3.22	9.73	11.80	248.80
Papua	0	0	1.70	7.49	8.31	179.40



(a) with outliers



(b) without outliers

Figure 26: Boxplot of Daily Rainfall Values for Each Province after MICE-SAMPLE

2.3.13 Summary

Each method provides different results that can picture how, more or less, the data can be in real life. Judging from the maximum values seen in Table 26, if during this duration there were some months with extreme rain, the imputation methods that fit the best would be Mean Substitution, LOCF Imputation, or Linear Interpolation. On the contrary, if there were a few years going on without a lot of months with rain, Median Substitution, Null Substitution, or MICE methods would fit better. Meanwhile, Bootstrap, Distribution Fill, and MICE-SAMPLE would fit better when the rain situation were more balanced without any extreme rainfall values or extreme number of days with no rain.

Table 26: Summary of Monthly Rainfall Values after Each Missing Data Method

Method	**Min**	**Q1**	**Median**	**Mean**	**Q3**	**Max**	**0's**
Mean	0	100.75	208.10	234.95	332.71	1946.56	330
Median	0	80.97	179.78	203.57	290.00	1166.00	451
Boot	0	116.57	209.40	228.33	314.70	1085.00	227
LOCF	0	95.65	197.75	233.66	321.20	2650.50	411
Int	0	94.63	198.80	229.38	323.85	2343.70	360
MICE.pmm	0	115.13	210.97	226.50	311.52	1084.60	211
MICE.cart	0	113.09	210.74	227.50	316.60	1097.80	219
MICE.lasso	0	122.10	214.88	230.20	314.98	1092.79	210
MICE.ri	0	124.90	220.20	235.17	321.81	1098.95	208
MICE.sample	0	122.19	212.36	227.74	309.11	1124.60	203
Fill	0	122.21	213.13	227.93	309.37	1101.38	201
Null	0	62.10	158.85	180.58	262.72	1084.00	668

What is surprising from the methods, seen from Figure 27, is that Mean Substitution, Median Substitution, LOCF Imputation, Linear Interpolation, and Null Substitution have similar histogram shape: spike on zero and consistently going downwards. Meanwhile, the other methods such as Bootstrap, Distribution Fill, and all MICE methods also have similar histogram shape: spike on either the second or third interval, not on the first interval, and consistently going downwards afterwards. This difference shows how methods that focus on creating new values tend to have more zero values compared to the methods that focus on taking values from the data or the distribution of the data. The first category also gives higher outlier values compared to the second category, except for Null Substitution due to obvious reasons. In summary, the first method tends to extend the range and have extreme amount of values (both low and high), while the second category tends to create spikes in the middle of the intervals, not on the first or the last interval. The second category also consists of methods that can all be included in multiple imputation methods, even though Bootstrap and Distribution Fill in this research work as single imputation.

There are also some significant differences between the MICE methods, though not as big as compared to the other methods. Take the maximum value, for instance. With these 5 methods, the maximum values vary from 1084.60 up to 1124.60 mm/month, which means the range itself is around 40 mm/month. This was not expected because they come from the same main method, only with different approaches. According to the statistics, it seems like MICE-PMM and MICE-CART are more similar, meanwhile MICE-LASSO, MICE-RI, and MICE-SAMPLE are more similar to one another. Comparing to the other methods, the first two MICE methods are similar to Bootstrap, and the last three MICE methods are similar to Distribution Fill.

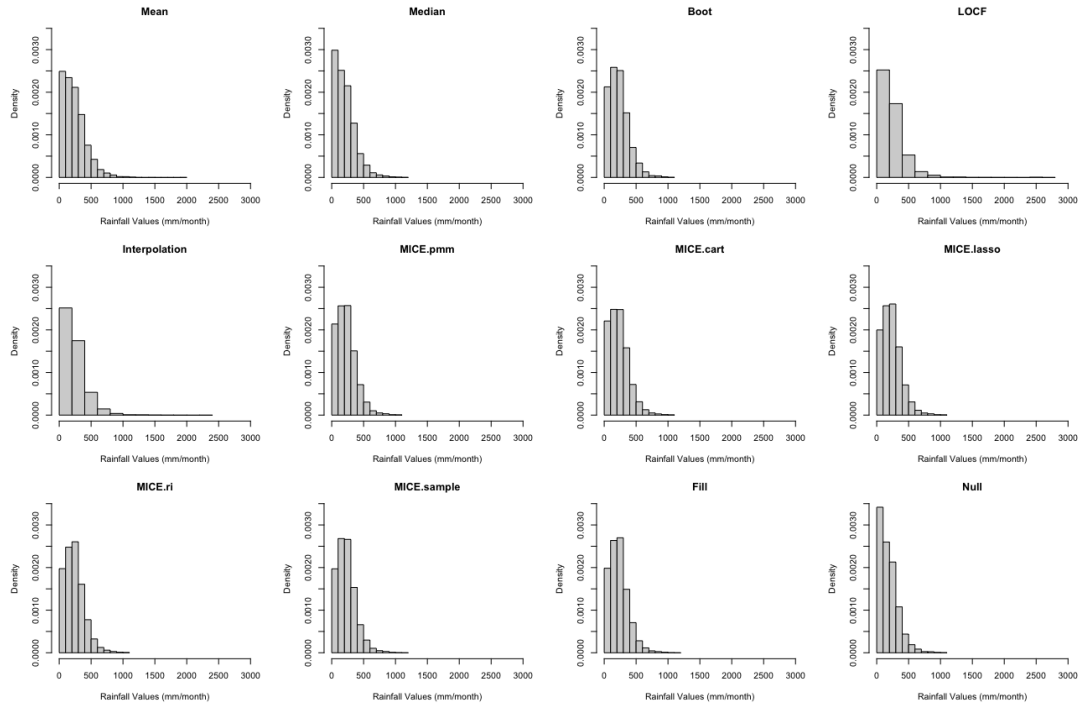


Figure 27: Histogram of Monthly Rainfall Values after Imputation (All Methods)

In general, Null Substitution has the lowest values as expected. With over 600 zero values, all the statistics values in Table 26 are the lowest. Mean Substitution produces high monthly rainfall intensity but at the same time, also a lot of zero values. As expected, this method is not robust and it transforms the imputed data values into much higher values. Also, due to the many zero values, this imputation method also has low first quartile. Comparing Mean to Median Substitution, Median obviously has much lower values, as expected due to all the histogram being positive-skewed, which means the median values obtained for the imputation would be lower than the mean values. Median Substitution also has the second lowest statistics valus after Null Substitution, excluding the maximum value. LOCF Imputation and Linear Interpolation are interesting: both methods start with low first quartile median values, but they suddenly rise up in the mean value, third quartile, and even maximum values, becoming the two methods with the highest maximum values. This imputation shows that the original data might have a lot of missing gaps between high values, causing the LOCF Imputation to repeatedly impute high values and Linear Interpolation to create linear models between two high values (or at least one of them is high).

As for the rest of the methods, Bootstrap, Distribution Fill, MICE-PMM, MICE-CART, MICE-LASSO, MICE-RI, and MICE-SAMPLE, they all produce similar results, more similar compared to the rest of the methods. These methods provide rather high first quartile values, followed by median values that are also high, but they show balance in the third quartile values, followed by rather low maximum values. All of these methods also have rather low zero values, not more than 250 zero values while the rest of the methods have over 300 zero values. These methods show that resampling the data and taking the distribution into account do not change the patterns of the data. Surprisingly, Bootstrap and MICE methods do not produce a lot of zero values. This is also the case for Distribution Fill, but since the imputed values are generated from exponential distribution, the non-existence of zero is justified. The number of zero values that Distribution Fill has is the same as the number of zero values that the original data has.

3 Data Grouping

After all the missing values have been imputed, the next step would be grouping the data. At the moment, the data have 34 variables which represent 34 provinces in Indonesia. However, province is not the only scale that we want to work on. We also want to see how the calculations would be if it is calculated based on regions (several provinces combined into one) and on the entire country (all the provinces combined into one). We also look at clusters. In order to do this, we need one variable containing the rainfall values for each region, the country, and each cluster. Obtaining these variables is possible through creating series by re-expressing the data using Weighted Values. Additionally, for cluster calculation, the clustering process needs to be done before putting in the weights.

3.1 Weighted Values

The purpose of weighted values here is to form one series from several provinces, depending on the area that we calculate. In other words, we re-express the data for region, country, and cluster calculations before processing it for forecasting. Deciding how to calculate the weight for each province is important. In the case of grouping rainfall values, we decided to use the area of each province. Table 27 shows the area and weight of each province for Country calculation. This dataset is obtained from the Indonesian Minister of Home Affairs Decree No. 300.2.2-2138 of 2025. We use the same area values for all calculations (Region and Cluster), but we use different weights depending on the number of provinces in each group.

Formation of a Single Series
After obtaining the weight for each province, we combine the rainfall values into one series with the formula as follows:

$$Y_i = \sum_{j=1}^{34} w_j \cdot x_{ij}, \tag{3.1}$$

where w_j represents the weight of province j that belongs to the group (Region, Country, or Cluster) and x_{ij} represents the rainfall value of province j at time i.

3.2 Clustering

Clustering has two steps: distance calculation and grouping with linkage methods. Distance calculation is to measure how different variables are from one another, meanwhile linkage methods are needed to group the variables based on these distances.

3.2.1 Distance

There are several formulas that can be used to obtain distance. Every distance works differently, depending on what kind of clustering is needed. Here, we use Euclidean, Maximum, Manhattan, and Canberra distance formulas. x_i represents the i-th value of variable x, and y_i represents the i-th value of variable y. We have N data points and p variables.

Table 27: Area of Each Province with Country Weight

Province	**Area (km^2)**
Aceh	56,835.019
Sumatera Utara	72437.755
Sumatera Barat	42,107.674
Riau	89900.78
Jambi	49,023.037
Sumatera Selatan	86771.918
Bengkulu	20,122.21
Lampung	33570.758
Kepulauan Bangka Belitung	16,670.225
Kepulauan Riau	8170.375
DKI Jakarta	660.982
Jawa Barat	37053.331
Jawa Tengah	34,337.489
Daerah Istimewa Yogyakarta	3170.363
Jawa Timur	48,055.876
Banten	9355.763
Bali	5,582.827
Nusa Tenggara Barat	19631.991
Nusa Tenggara Timur	46,378.105
Kalimantan Barat	147018.063
Kalimantan Tengah	153,430.363
Kalimantan Selatan	37125.426
Kalimantan Timur	126,951.758
Kalimantan Utara	69900.886
Sulawesi Utara	14,500.275
Sulawesi Tengah	61496.983
Sulawesi Selatan	45,330.55
Sulawesi Tenggara	36139.303
Gorontalo	12,024.982
Sulawesi Barat	16590.667
Maluku	46,133.832
Maluku Utara	31465.977
Papua Barat	99411.648
Papua	312,820.717
TOTAL	**1,890,177.908**

Euclidean Distance
This distance measures the square root of sum of square of differences between each element of the two variables. It is best used for continuous data.

$$d_{\text{euclidean}}(x, y) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2}$$

Maximum Distance
This distance finds the highest distance value from all the absolute differences between each element of the two variables. It is best used for categorical data.

$$d_{\text{maximum}}(x, y) = \max_i |x_i - y_i|$$

Manhattan Distance
This distance calculates the sum of all the absolute of difference between each element of the two variables. It is best used for city block distance between two vectors, meaning it only measures along axes (x and y axes). Unlike Euclidean distance that can form a straight line whenever, Manhattan distance needs to calculate by grids.

$$d_{\text{manhattan}}(x, y) = \sum_{i=1}^{N} |x_i - y_i|$$

Canberra Distance
This distance divides the absolute of difference between each element of the two variables by the sum of absolute of sum of each element, and then sums up all these fractions together. This method is best used for non-negative values, e.g. counts.

$$d_{\text{canberra}}(x, y) = \sum_{i=1}^{N} \frac{|x_i - y_i|}{|x_i + y_i|}$$

Distance Matrix
After obtaining all the distances, we create a distance matrix (from the same type of distance) that is symbolized with D. We will use D for the grouping later on.

$$D = \begin{pmatrix} d(1,1) & d(1,2) & \dots & d(1,p) \\ d(2,1) & d(2,2) & \dots & d(2,p) \\ \vdots & \vdots & \ddots & \vdots \\ d(p,1) & d(p,2) & \dots & d(p,p) \end{pmatrix}$$

3.2.2 Linkage Methods

After obtaining all the distances of each variable pair and forming D, we move on to the linkage method in order to group the variables together. There are several linkage methods that will be used in this thesis: Single, Complete, Average, McQuitty, Median, Centroid, and Ward's Linkage Method. $d(x, y)$ represents the distance between variable x and variable y, meanwhile the d in each method represents the new distance after we merge the variables into temporary clusters until we reach the number of clusters that we want.

Single Linkage Method
We cluster the variables based on the lowest value in matrix D. This process goes on until the number of clusters reaches the number of clusters that we want. This method usually works best for data with high outliers because this clustering does not consider the outliers.

$$d_{\text{single}} = \min d(x, y)$$

Complete Linkage Method
This method starts with clustering the lowest distance in matrix D. Afterwards, we obtain the new distance between the merged cluster with the other variables by choosing the maximum distance between the current distance values. This process goes on until the number of clusters reaches the number of clusters that we want. This method generally produces more well-separated clusters compared to Single Linkage Method, but it depends on the outliers as well.

$$d_{\text{complete}} = \max d(x, y)$$

Average Linkage Method
This method measures the sum of distances between each observation in each cluster, and then weighted by the number of observations in each cluster. Unlike Single and Complete, the Average Linkage Method requires each distance value. The clustering process starts by merging two clusters with the lowest distance, and then the new distance between the newly-formed cluster and the other variables will be calculated with the formula below. This process goes on until the wanted number of clusters is reached. This method is also good to provide well-separated clusters, like Complete Linkage Method, but it is better at handling the existence of outliers compared to Complete Linkage Method.

$$d_{\text{average}} = \frac{1}{n_X \times n_Y} \sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} d(x_i, y_j)$$

McQuitty Linkage Method
This method also measures the sum of distance between each observation in each cluster, but instead of weighting it by the each cluster's number of observation, it sums the number of observation in both clusters. This method has a similar behavior to Single Linkage Method. Just like the former methods, it starts by merging two variables with the lowest distance, and the new distance will be formed with this formula.

$$d_{\text{mcquitty}} = \frac{1}{n_X + n_Y} \sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} d(x_i, y_j)$$

Median Linkage Method
This method simply calculates the median of all the distances between each observation in each cluster. After that, the lowest median value determines the merged clusters, and the process goes on until the number of cluster wanted is reached.

$$d_{\text{median}} = \text{median}(d(x, y))$$

Centroid Linkage Method
This method starts by merging two variables with the lowest distance. Afterwards, we calculate the new centroids μ_{XY} by using all the data points from Cluster X and Cluster Y. Using the same distance calculations to form matrix D, we recalculate all the distances between the new cluster XY with the other variables, but this time, we use μ_{XY} as the data points for XY. $||.||$ represents the same distance formula being chosen to form matrix D. This method works best to group data with fewer similarities.

$$\mu_{XY} = \left(\frac{x_1 + y_1}{2}, \frac{x_2 + y_2}{2}, \ldots, \frac{x_N + y_N}{2} \right) \tag{3.2}$$
$$d_{\text{centroid}} = ||\mu_X - \mu_Y||$$

Ward's Linkage Method
Just like the other methods, it starts by finding the lowest distance. The distance between the new cluster and the other variables afterwards are calculated by obtaining new centroids μ_{XY} just like Equation (3.2). Using the formulas below, we obtain the new distance. There are two types of Ward's Linkage Method used in this thesis. The first one does uses square root, meanwhile the second one uses the raw squared value. The first one tends to produce clusters with relatively even number of members, meanwhile the second one prioritizes minimizing within-cluster variance without concerning about the number of members in the cluster.

$$d_{\text{ward.D}} = \sqrt{\frac{n_X n_Y}{n_X + n_Y} ||\mu_X - \mu_Y||^2}$$
$$d_{\text{ward.D2}} = \frac{n_X + n_Y}{n_X n_Y} ||\mu_Y - \mu_Y||^2$$

3.3 Country Grouping

We start with transforming the time series for each province to obtain the time series for the entire country. We calculate the weight of each province based on their area. Afterwards, we see the final series of Country after the weights are imputed in the calculation.

3.3.1 Province Weight

By dividing the area of each province in Table 27 with the total of area, we obtain the weight for each province as shown in Table 28. The highest contribution for the creation of Country series is apparently given by Papua province with 16.55%, and the lowest is given by DKI Jakarta which is only 0.03%. These weights are applied to all missing data methods, without exception. Hence, all missing data methods will have the same weights applied to their series.

Aside from Papua, two other provinces that give the highest contribution to the Country series are Kalimantan Tengah and Kalimantan Barat, and both happen to belong to the Kalimantan region, while Papua belongs to the Papua region. As for the other two provinces aside from DKI Jakarta with the lowest weight or contribution are DI Yogyakarta and Bali. DKI Jakarta and DI Yogyakarta both belong to Jawa region, while Bali belongs to Nusa Tenggara region. These regions are based on the region division made by the government written in Table 29.

3.3.2 Final Country Series

The final series for the country is shown in Figure 28. In general, the series between missing methods looks fairly similar. These series range from 100-400 mm/month in general, although there are some missing data methods that have higher ranges. Some months in Mean, LOCF, and Interpolation exceed 400 mm, almost 500 in fact. As for Null, it mostly ranges from 0-300 mm/month with some values exceeding 300 mm/month.

It is almost impossible to spot the differences between these series, but the most obvious differences can be seen from the spikes that are located in these timestamps 1) early 2001 until mid-2001; 2) mid-2002 until mid-2003; and 3) mid-2014 until mid-2015. For the first timestamp, the obvious difference can be observed from Interpolation where the spike is almost 500 mm/month, very high compared to the other methods. As for the second timestamp, there are different kinds of fluctuations there, showing different number of small spikes in this timestamp. Some of them only have one spike, like Mean, Median, and Null, but some of them have two, like Bootstrap and Interpolation (almost). As for the third timestamp, aside from between MICE, they look very different. LOCF and Interpolation look similar, so do Mean and Median.

Overall, it is safe to say that the final series of Country shows mostly similar results between the missing data methods. Due to this, it is expected that the forecast between each of them will not be too different. In addition, the values for the series can be seen in Appendix C.

Table 28: Weight (%) of Each Province for Country based on Land Area

Province	**Weight (%)**
Aceh	3.01
Sumatera Utara	3.83
Sumatera Barat	2.23
Riau	4.76
Jambi	2.59
Sumatera Selatan	4.59
Bengkulu	1.06
Lampung	1.78
Kepulauan Bangka Belitung	0.88
Kepulauan Riau	0.43
DKI Jakarta	0.03
Jawa Barat	1.96
Jawa Tengah	1.82
Daerah Istimewa Yogyakarta	0.17
Jawa Timur	2.54
Banten	0.49
Bali	0.30
Nusa Tenggara Barat	1.04
Nusa Tenggara Timur	2.45
Kalimantan Barat	7.78
Kalimantan Tengah	8.12
Kalimantan Selatan	1.96
Kalimantan Timur	6.72
Kalimantan Utara	3.70
Sulawesi Utara	0.77
Sulawesi Tengah	3.25
Sulawesi Selatan	2.40
Sulawesi Tenggara	1.91
Gorontalo	0.64
Sulawesi Barat	0.88
Maluku	2.44
Maluku Utara	1.66
Papua Barat	5.26
Papua	16.55

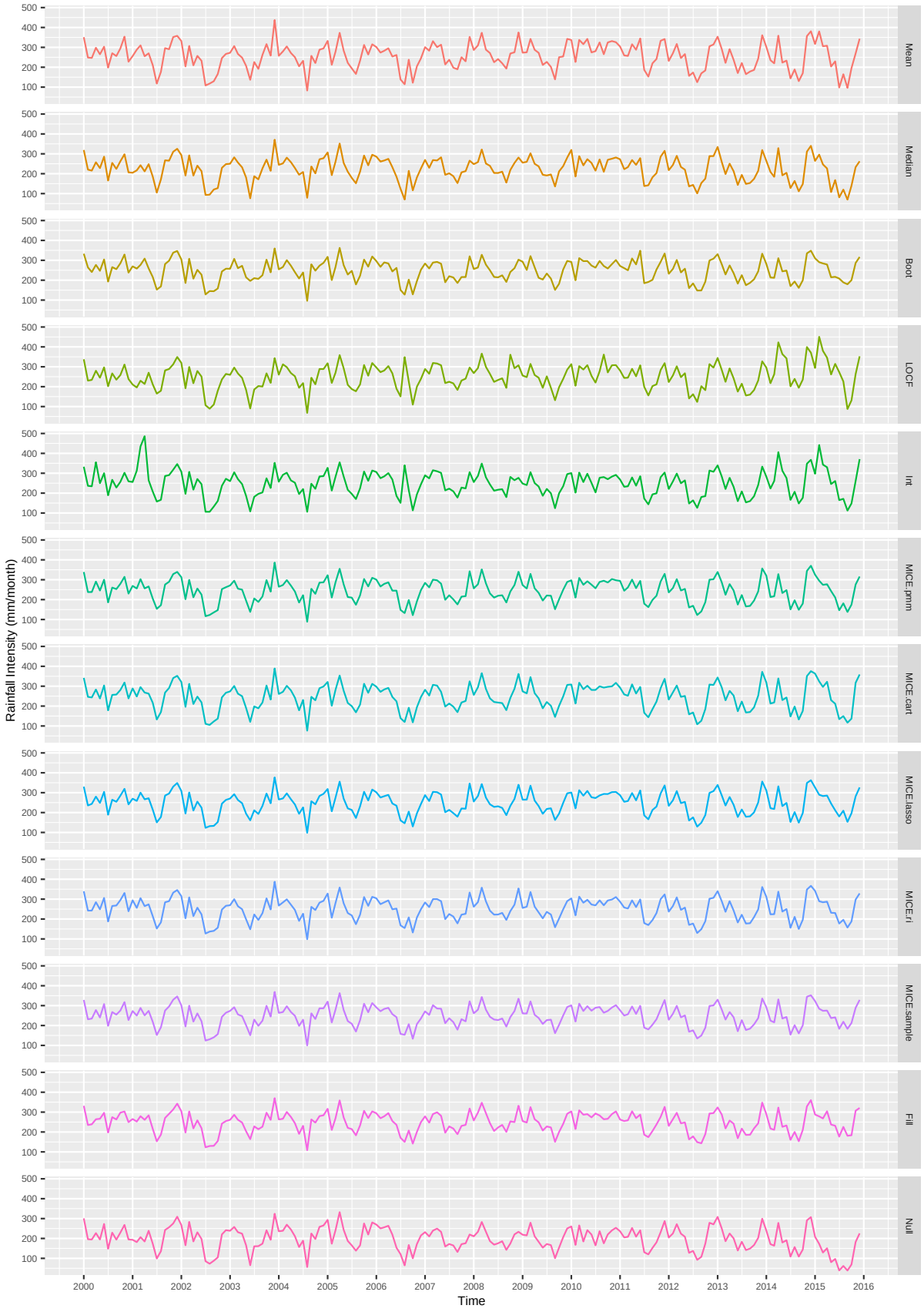


Figure 28: Final Series of Country Data (Weighted)

3.4 Region Grouping

Another area that we are going to calculate is Region. In Indonesia, the provinces are divided into 7 main regions with at least two provinces in each region. This division will determine the grouping which will affect the weights of each province. So, we see the list of provinces for each region, followed by the weight of each province for the region series. In the end of this section, there will be 7 different series for Region calculation.

3.4.1 List of Regions

The list of provinces in each region is as written in Table 29. This division is based on the main island where each province belongs to. There are 5 big islands in Indonesia and 2 main archipelagos of Indonesia, creating 7 main regions for the province regions according to [Sta]. As per 2017 as far as the datasets were obtained, there were 34 provinces in Indonesia (as per 2022, 4 more provinces were added to Papua region due to political reasons).

Table 29: List of Provinces per Region in Indonesia

Region	**Provinces**
Sumatera (10)	Aceh, Sumatera Utara, Sumatra Barat, Riau, Kepulauan Riau, Jambi, Bengkulu, Sumatra Selatan, Kepulauan Bangka Belitung, Lampung
Jawa (6)	Banten, DKI Jakarta, Jawa Barat, Jawa Tengah, DI Yogyakarta, Jawa Timur
Nusa Tenggara (3)	Bali, Nusa Tenggara Barat, Nusa Tenggara Timur
Kalimantan (5)	Kalimantan Barat, Kalimantan Tengah, Kalimantan Utara, Kalimantan Timur, Kalimantan Selatan
Sulawesi (6)	Sulawesi Utara, Gorontalo, Sulawesi Tengah, Sulawesi Barat, Sulawesi Selatan, Sulawesi Tenggara
Maluku (2)	Maluku, Maluku Utara
Papua (2)	Papua, Papua Barat

Sumatera, Jawa, Kalimantan, Sulawesi, and Papua are the 5 big islands of Indonesia. Meanwhile, Nusa Tenggara and Maluku are 2 main archipelagos. The areas with big islands tend to have more provinces compared to those that belong to a group of islands. For a clearer visualization, Figure 29 shows the map of Indonesia with these divisions.

3.4.2 Province per Region Weight

The weight of each province used to obtain one series is written in Table 30 with the region division based on Table 29. The more provinces in a region there are, of course, the less contribution each province can give. Maluku and Papua only have two provinces, but the contribution is not half each. As for Maluku, the land areas are still more or less the same, which makes the contribution almost half, but the provinces of Papua region are not similar in size. Papua province is much bigger, hence contribution 75% to the series. Land size also affects Nusa Tenggara provinces, with Bali only contributing 7% while Nusa Tenggara Timur contributes almost 65%. Every region seems to always have some provinces that dominate the contribution, and at the same time, some provinces that do not contribute much. The biggest contributions for each region are given by Riau, Jawa Timur, Nusa Tenggara Timur, Kalimantan Tengah, Sulawesi Tengah, Maluku, and Papua respectively, while the lowest contributions for each region are given by Kepulauan Riau, DKI Jakarta, Bali, Kalimantan Selatan, Gorontalo,

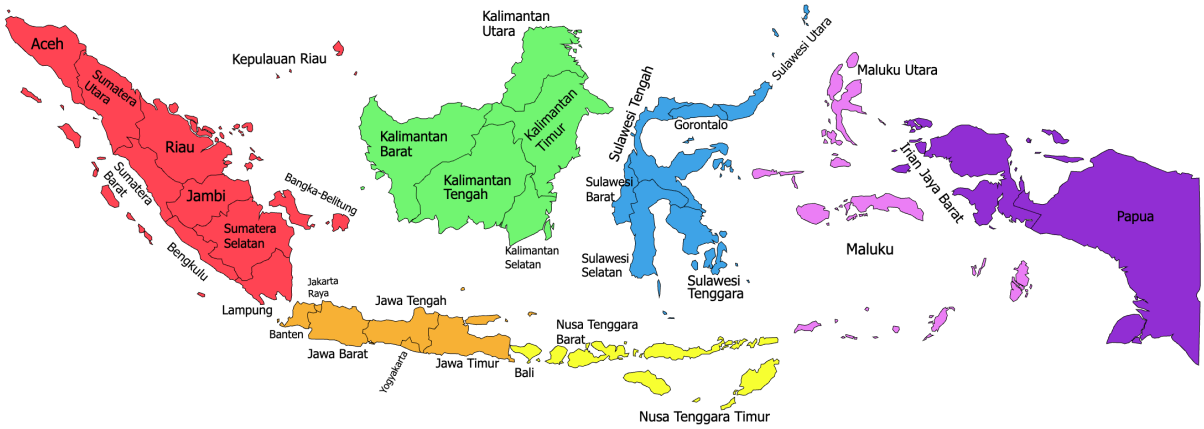


Figure 29: Map of Indonesia Divided by Regions

Maluku Utara, and Papua Barat respectively. Just like Country, we use the same weights for all missing methods.

3.4.3 Final Region Series

Through this part, we discuss how the Region series looks like after weights are imputed. Figure 30 shows the rainfall values of the 10 provinces of Sumatera. We can clearly see the differences between these series and Figure 28. Sumatera seems to have a lot of extreme changes from time to time. The values mostly range from 50-500 mm/month, except for Null that only ranges from 50-400 mm/month. These series also look similar between one another. There are some small differences, like in early 2015 for example, where the values are high for Mean, Median, Bootstrap, LOCF, and Interpolation, but they are low for all MICE methods, Fill, and Null.

Figure 31 shows the series of Jawa, which is extremely different compared to Sumatera, but also different from Country. The rainfall value changes tend to be more extreme. At times, they are really high, reaching the range of 600-800 mm/month, but the next time, they are very low, mostly in the 0-200 mm/month range. In some cases, such as Mean, Median, LOCF, and Interpolation, the series range from 0-800 mm/month. As for the other methods, the series only ranges from 0-600 mm/month, and the changes in these methods are still similar to changes in Country. There are some visible differences in the spikes in early 2002, early 2006, and early 2014.

There are significant differences in Figure 32, the series of Nusa Tenggara. All of the missing methods provide similar results with the range of 0-600 mm/month, except for LOCF and Interpolation. Somehow, LOCF has some extremely high values that reach 1500 mm/month. Interpolation also has some high spikes that reach with one reaching 1250 mm/month. The possible explanation for this would be that there are a lot of missing values between 2007-2009 and 2014-2015 in Bali, Nusa Tenggara Barat, and Nusa Tenggara Timur with high observation values. This causes LOCF to repeat the high values several times and Interpolation to create linear models that produce high values, causing even higher values once the daily imputed values are aggregated.

Kalimantan region has quite interesting results as shown in Figure 33. So far, this region might have the smallest range. Most of the missing data methods range from 100-500 mm/month, such as Interpolation and all the MICE methods. Median is almost in that range, but it has some observations below 100 mm/month. Meanwhile, Bootstrap and Fill have a smaller range around 200-400 mm/month with some observations above and below it, though

Table 30: Weight (%) of Each Province for Region based on Land Area

Province	**Weight (%)**
Aceh	11.95
Sumatera Utara	15.23
Sumatera Barat	8.85
Riau	18.90
Jambi	10.31
Sumatera Selatan	18.24
Bengkulu	4.23
Lampung	7.06
Kepulauan Bangka Belitung	3.51
Kepulauan Riau	1.72
DKI Jakarta	0.50
Jawa Barat	27.94
Jawa Tengah	25.89
Daerah Istimewa Yogyakarta	2.39
Jawa Timur	36.23
Banten	7.05
Bali	7.80
Nusa Tenggara Barat	27.42
Nusa Tenggara Timur	64.78
Kalimantan Barat	27.51
Kalimantan Tengah	28.71
Kalimantan Selatan	6.95
Kalimantan Timur	23.75
Kalimantan Utara	13.08
Sulawesi Utara	7.79
Sulawesi Tengah	33.05
Sulawesi Selatan	24.36
Sulawesi Tenggara	19.42
Gorontalo	6.46
Sulawesi Barat	8.92
Maluku	59.45
Maluku Utara	40.55
Papua	75.88
Papua Barat	24.12

nothing too extreme. LOCF has the range of 0-500 mm/month, meanwhile Null has the range of 0-400 mm/month. Mean is the odd one out, seeing that it has the range of 0-600 mm/month, and it has a high spike at the end of 2013. This spike can be caused by an outlier in the data, causing the average value to be extremely high.

The differences between missing data methods for Figure 34 can be seen in 2012-2016. The spikes in 2013 for Sulawesi have different heights, and 2015 shows very different series between these methods. The range itself is rather similar for all of the methods: 0-500 mm/month, with some exceeding this number. However, the tail in 2015 looks interesting. We can safely say that a lot of missing values were there, and the imputed values make Null has really low values in 2015. It seems like the existing values in 2015 are mostly small because LOCF and Interpolation are also rather low, but seeing that Mean and Median have high values, some existing values might have high values to boost the rainfall intensity. What is also interesting is that MICE methods also show very different results between one another.

Maluku region also has rather similar results as shown in Figure 35. The main differences are the spike in mid-2011 and the second spike in early 2014. The spikes in 2011 vary, but they are still visible. Mean and LOCF's spikes are very high, meanwhile Null seems to have the lowest spikes, showing that there were a lot of missing values during this time. The second spike in 2014 is also interesting because for some methods, the spikes are not even showing. We can see the spikes in Mean, Median, LOCF, and Interpolation, but the spikes are not visible in Bootstrap, all MICE methods, Fill, and Null. This also indicates that there are a lot of missing values in here, and the existing values are pretty high.

The last region is Papua which is shown in Figure 36. The rough series shows that there are a lot of changes going on. There are also some visible spikes that are significantly different with the other methods. Interpolation has two spikes in 2000-2002 which are not owned by any other methods. In 2006, LOCF and Interpolation have a spike that, again, is not owned by any other methods. In 2015, Mean, LOCF, and Interpolation have a spike that only belong to them. The only spike that shows up in all methods is the one in 2013, and these spikes are surprisingly similar. Another significant difference that can be spotted is how the series looks like in 2015-2016. Some can be high, some can be really low. Papua also has a huge range in general, with 0-700 mm/month for Null, which has the lowest values, and 0-1000 mm/month for LOCF and Interpolation that seem to have the highest values.

In general, the series looks similar for the different missing data methods. There are some significant differences, especially when it comes to spikes. There are also some differences especially shown by Mean, LOCF, and Interpolation. It is also shown that when there are too many members of the group (Sumatera) or too less members of the group (Maluku and Papua), the series looks more rough. With too less members, the differences are more significant and more easily spotted as well. In addition, the values for the series can be seen in Appendix C.

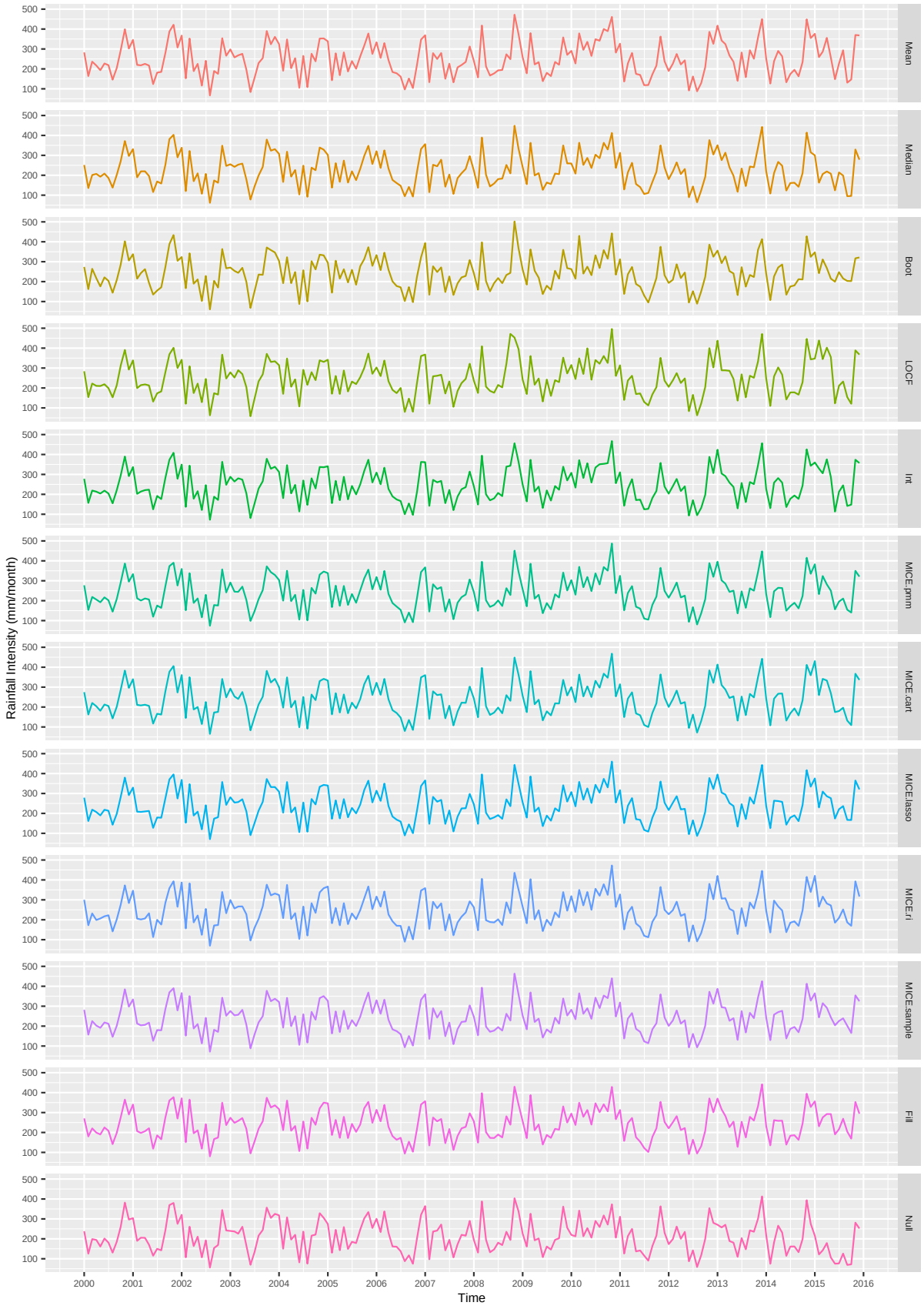


Figure 30: Final Series of Sumatera Region Data (Weighted)

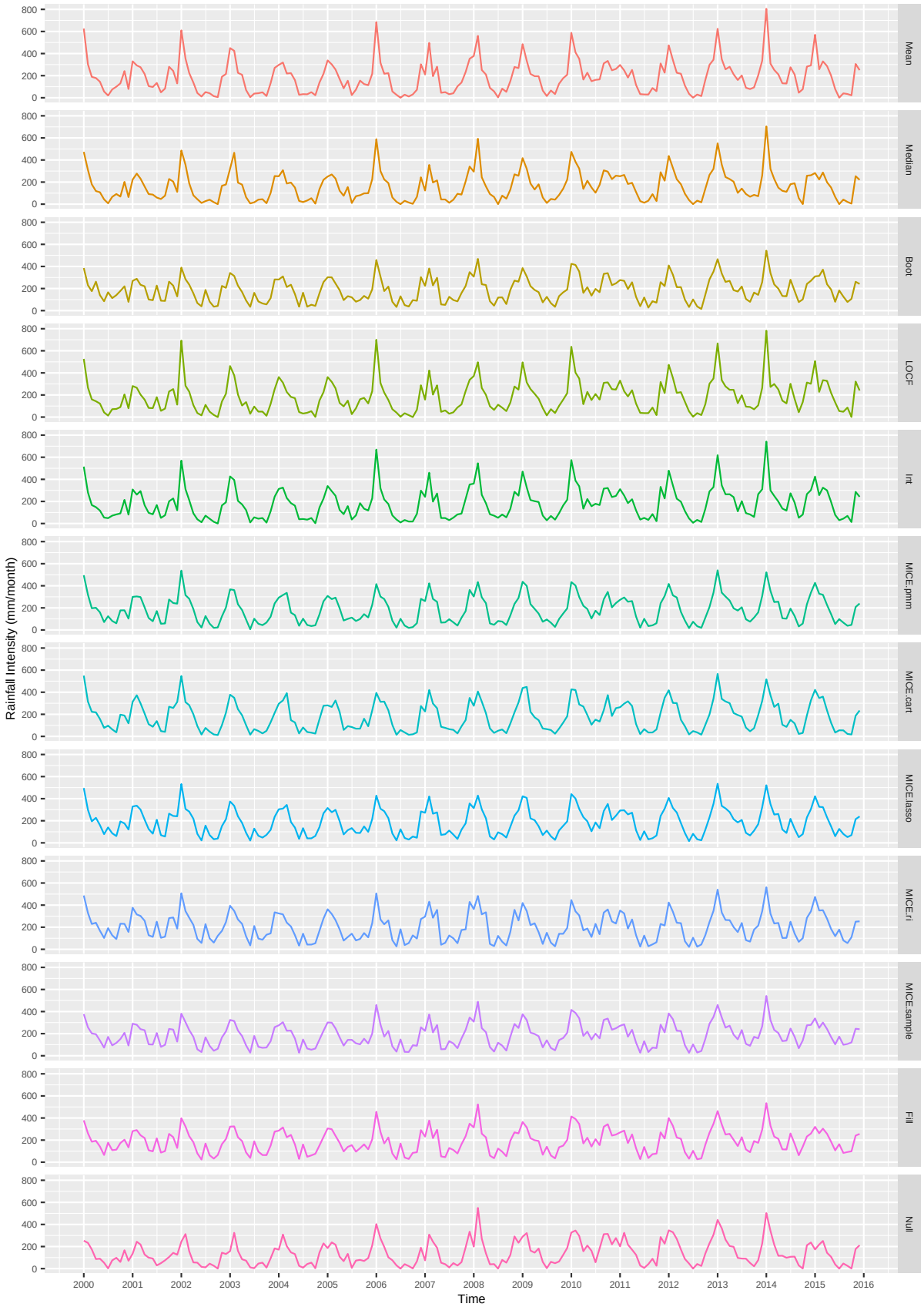


Figure 31: Final Series of Jawa Region Data (Weighted)

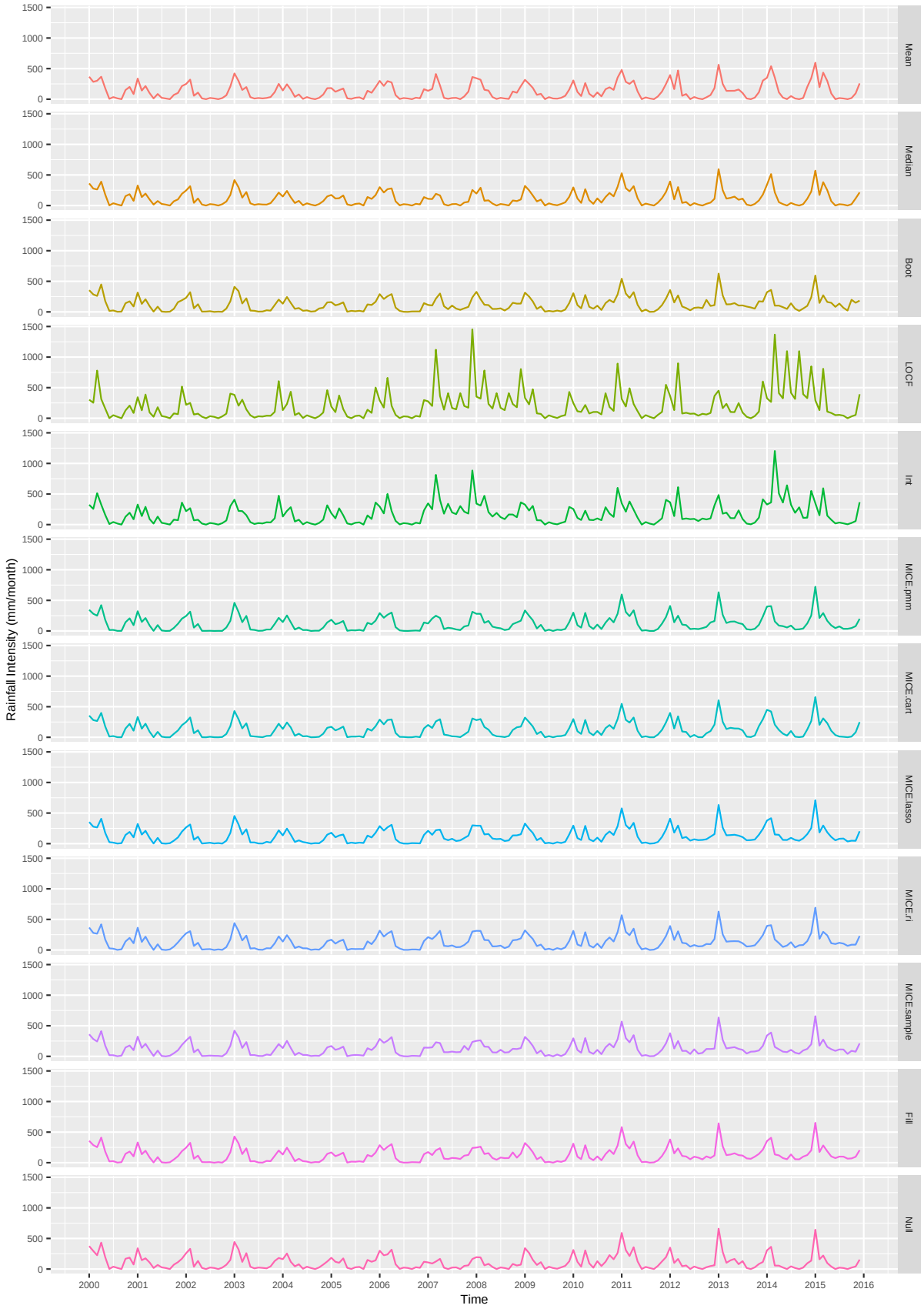


Figure 32: Final Series of Nusa Tenggara Region Data (Weighted)

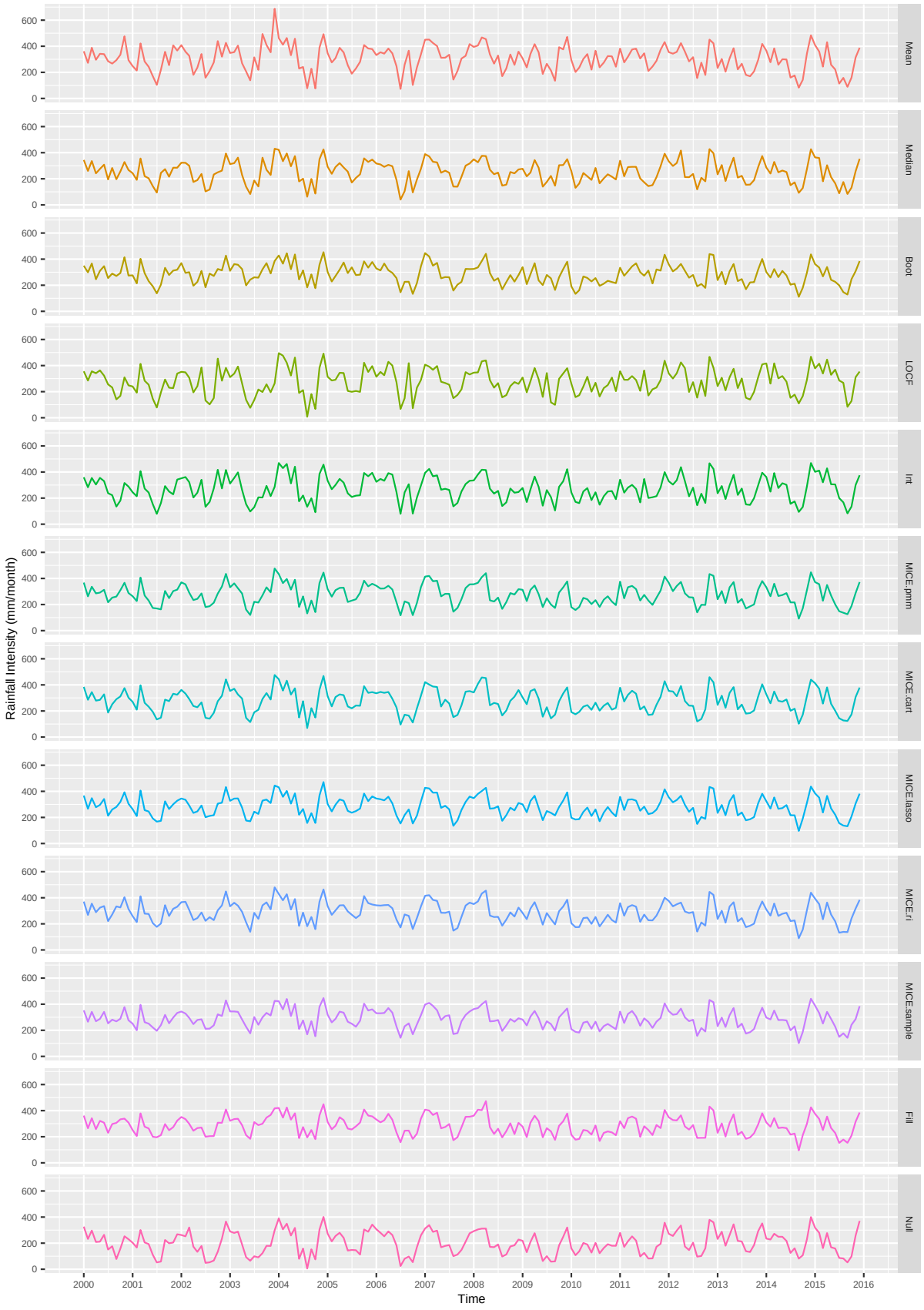


Figure 33: Final Series of Kalimantan Region Data (Weighted)

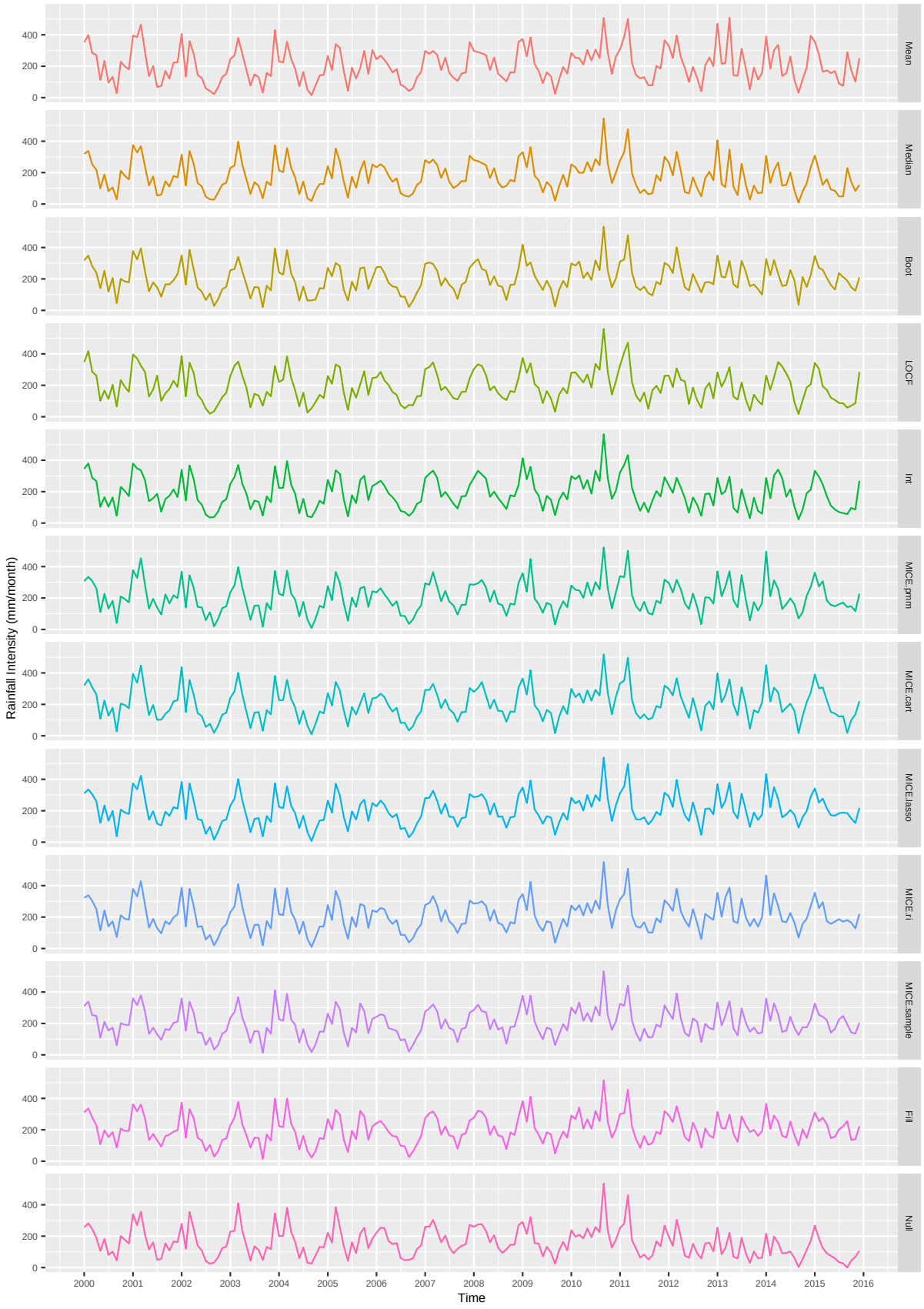


Figure 34: Final Series of Sulawesi Region Data (Weighted)

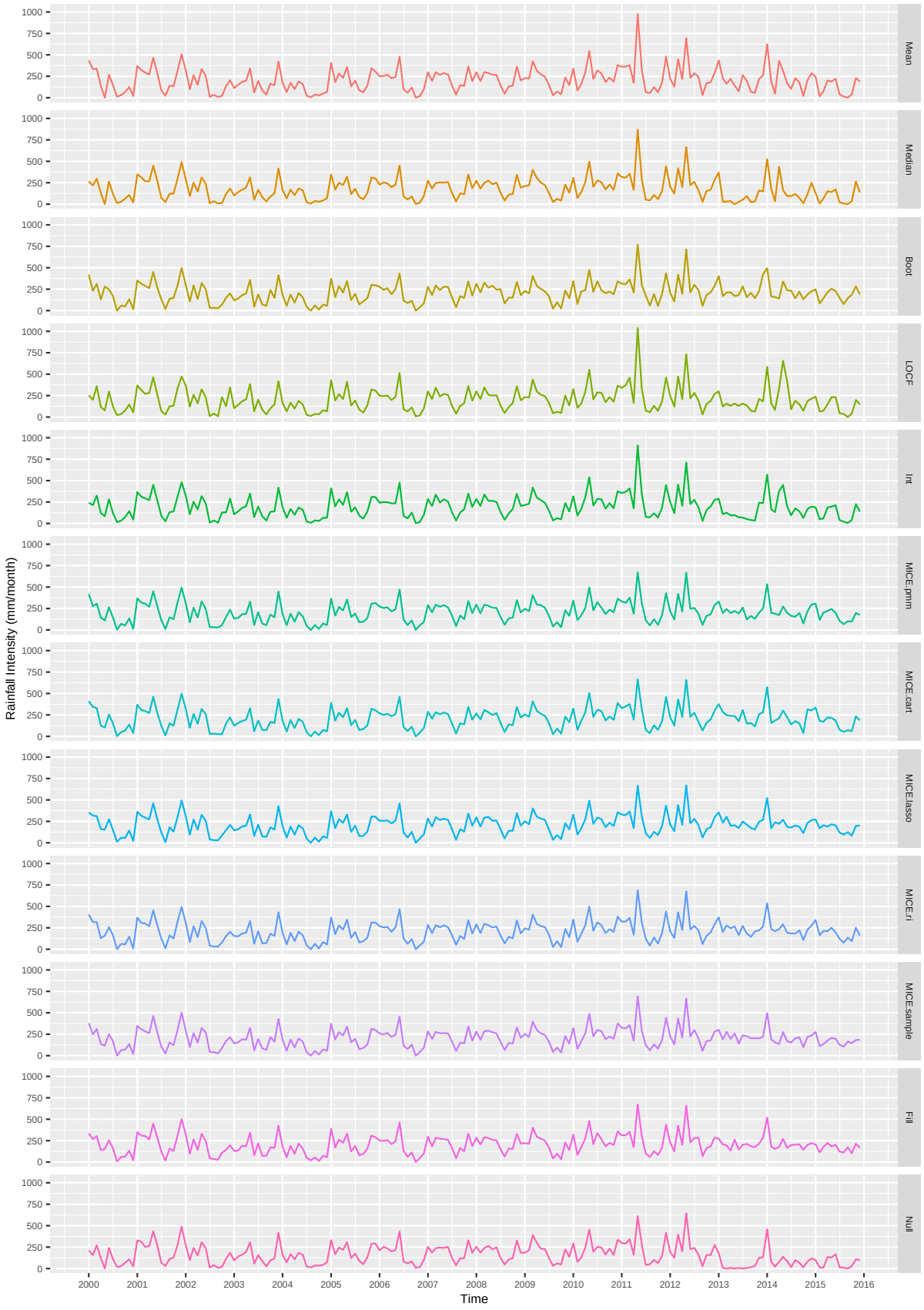


Figure 35: Final Series of Maluku Region Data (Weighted)

Figure 36: Final Series of Papua Region Data (Weighted)

3.5 Cluster Grouping

The next approach that we are going to work on is clustering. After grouping the provinces based on the official region separation, grouping the provinces based on statistically-processed groups is also needed. Clustering is done for each missing data method. With 4 Distance calculations and 8 Linkage methods, there are 32 combinations of clustering that can be tried. Based on the clustering results of these combinations, we attempt to pick the best fitting clusters for the provinces and for the missing data methods. Once the clusters are set, weights are imputed to each cluster for each missing data method.

3.5.1 Choosing Clusters

There are 32 combinations from 4 types of Distance calculations and 8 Linkage Methods that can be used to determine the clusters for each missing data method. By the procedure, the first province of the list, Aceh, will always be in Cluster 1. As shown in Table 31, each province except Aceh has more than one cluster option. Some of them have less than 5 cluster options, but most of them have 7 cluster options. Sumatera Utara has 2 cluster options, meanwhile Sumatera Barat has 3 cluster options. There are 4 provinces with 4 cluster options: Riau, Jambi, Sumatera Selatan, and Jawa Barat. There are 13 provinces each for the group with 6 cluster options and 7 cluster options. For the group with 6 cluster options, the provinces are Lampung, Kepulauan Bangka Belitung, Kepulauan Riau, DKI Jakarta, Jawa Tengah, DI Yogyakarta, Jawa Timur, Banten, Bali, Nusa Tenggara Barat, Nusa Tenggara Timur, Kalimantan Tengah, and Kalimantan Timur. Finally, the last group that has 7 cluster options consist of these provinces: Kalimantan Barat, Kalimantan Selatan, Kalimantan Utara, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, Gorontalo, Sulawesi Barat, Maluku, Maluku Utara, Papua Barat, and Papua.

We consider using the number of clusters that each province is assigned to as its own set of clusters. If we see Table 31 again, it means Cluster 1 will only have Aceh, Cluster 2 will have Sumatera Utara, Cluster 3 will have Sumatera Barat, Cluster 4 will have Riau, Jambi, Sumatera Selatan, and Jawa Barat, and so on. However, the number of clusters that each province is in does not explain anything regarding the data. These numbers only show that using different clustering methods combined with various data imputation methods, the cluster results may vary.

Breaking down the clusters to each missing data method, we see that most of the clustering methods show that most provinces are alike and the calculation put almost all provinces together in Cluster 1. This causes some problems in the clustering process because not all 7 clusters have members. At most, there are only 5 clusters. However, we want to make sure we have 7 clusters so that the number is consistent to the number of Region. Note that clusters are nominal, which means they only function as labels. Clusters do not show that some provinces are better than the other provinces, or even vice versa. Note that the clustering process usually starts from the first province, which is Aceh. In Table 32, we see that Aceh is consistently in Cluster 1. The reason why Aceh is the only province staying in Cluster 1 no matter which method is used is because Aceh is the base of the clustering. Every province is being compared to Aceh. If the province is similar to Aceh, it joins Cluster 1. If not, the province starts a new cluster or joins another already existing cluster.

We also consider using the most chosen cluster per province as its own set of clusters. Using 4 distance formulas and 7 linkage methods in Section 3.2, it means we have 28 different combinations to cluster the provinces. Our original plan is to find a cluster number that shows up the most often for each province in these 28 combinations. However, again, this clusters do not explain anything except the data similarities. For example, Jawa Barat and Jawa Tengah

are the only members of Cluster 2 using 20 combinations and the only members of Cluster 3 using the remaining 8 combinations. This does not prove that Cluster 2 is better than Cluster 3. Since clusters are nominal, they are actually the same, especially since they have the exact same members. Hence, using the most chosen clusters is not a good option.

Table 31: Summary of Cluster Options for Each Province

Province	**Cluster**	**Number of Clusters**
Aceh	1	1
Sumatera Utara	1,2	2
Sumatera Barat	1,2,3	3
Riau	1,2,3,4	4
Jambi	1,2,3,4	4
Sumatera Selatan	1,2,3,4	4
Bengkulu	1,2,3,4,5	5
Lampung	1,2,3,4,5,6	6
Kepulauan Bangka Belitung	1,2,3,4,5,6	6
Kepulauan Riau	1,2,3,4,5,6	6
DKI Jakarta	1,2,3,4,5,6	6
Jawa Barat	1,2,3,4	4
Jawa Tengah	1,2,3,4,5,6	6
DI Yogyakarta	1,2,3,4,5,6	6
Jawa Timur	1,2,3,4,5,6	6
Banten	1,2,3,4,5,6	6
Bali	1,2,3,4,5,6	6
Nusa Tenggara Barat	1,2,3,4,5,6	6
Nusa Tenggara Timur	1,3,4,5,6,7	6
Kalimantan Barat	1,2,3,4,5,6,7	7
Kalimantan Tengah	1,2,3,4,5,6	6
Kalimantan Selatan	1,2,3,4,5,6,7	7
Kalimantan Timur	1,2,3,4,6,7	6
Kalimantan Utara	1,2,3,4,5,6,7	7
Sulawesi Utara	1,2,3,4,5,6,7	7
Sulawesi Tengah	1,2,3,4,5,6,7	7
Sulawesi Selatan	1,2,3,4,5,6,7	7
Sulawesi Tenggara	1,2,3,4,5,6,7	7
Gorontalo	1,2,3,4,5,6,7	7
Sulawesi Barat	1,2,3,4,5,6,7	7
Maluku	1,2,3,4,5,6,7	7
Maluku Utara	1,2,3,4,5,6,7	7
Papua Barat	1,2,3,4,5,6,7	7
Papua	1,2,3,4,5,6,7	7

Table 32: Cluster Options of Each Province and Each Missing Data Method using All Distance and Linkage Method Combinations

Province	**Cluster Options**					
	Mean	**Median**	**Boot**	**LOCF**	**Interpolation**	**PMM**
Aceh	1	1	1	1	1	1
Sumatera Utara	1,2	1,2	1,2	1,2	1,2	1,2
Sumatera Barat	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3
Riau	1,2,3	1,2,3,4	1,2,3,4	1,2,3,4	1,2,3	1,2,3,4
Jambi	1,2,3	1,2,3,4	1,2,3,4	1,2,3,4	1,2	1,2,3,4
Sumatera Selatan	1,2,3	1,2,4	1,2,4	1,2,3,4	1,2	1,2,3,4
Bengkulu	1,2,3,4	1,2,3,4,5	1,2,3,4,5	1,2,3,4,5	1,2,3,4	1,2,3,4,5
Lampung	1,3	1,2,3	1,3,6	1,2,3,4	1,2,3,4	1,3,4,5,6
Kepulauan Bangka Belitung	1,2,3,4,5	1,2,3,4,5,6	1,2,3,4,5,6	1,2,3,4,5	1,2,3,4,5	1,2,3,4,5
Kepulauan Riau	1,2,3,4,5	1,2,3,4	1,2,3,4	1,2,3,4	1,2,3	1,2,3,4
DKI Jakarta	1,2,3,4,5,6	1,2,3,4,5	1,2,4,5	1,2,3,4,5,6	1,2,3,4,5	1,2,3,4,5,6
Jawa Barat	1,2,3	1,2,3,4	1,2,4	1,2,3,4	1,2,4	1,2,3,4
Jawa Tengah	1,3,4,5,6	1,3,5,6	1,4,5,6	1,3	1,2,3,4,5,6	1,3,4,5,6
DI Yogyakarta	1,3,4,6	1,3,4,5,6	1,2,3,4,6	1,2,3,4	1,2,3,4,5,6	1,2,3,4,5,6
Jawa Timur	1,3,4,5,6	1,3,4,5,6	1,2,3,4,6	1,3,4,5	1,3,4,5,6	1,2,3,4,5,6
Banten	1,3,5,6	1,3,5,6	1,3,4,5,6	1,3	1,2,3,4,5,6	1,3,4,5,6
Bali	1,3,4,5,6	1,3,4,5,6	1,3,4,5,6	1,3,4,5	1,4,5,6	1,3,4,5,6
Nusa Tenggara Barat	1,3,5,6	1,3,4,5,6	1,4,5,6	1,5,6	1,4,5,6	1,3,4,5,6
Nusa Tenggara Timur	1,5,6	1,3,5,6,7	1,5,6	4,5,6,7	4,5,6,7	1,3,4,5,6,7
Kalimantan Barat	1,2,3,4,5,6	1,2,3,4,5	1,2	1,2,3,4,5,6	1,2,3,4,5,6	1,2,3,4,5
Kalimantan Tengah	1,2,3,4,5,6	1,2,3,4,5	1,2,3,4,5,6	1,2,3,4,5,6	1,2,3,4,5,6	1,2,3,4,5,6
Kalimantan Selatan	1,2,3,4,7	1,2,3,4,5,6	1,2,3,4,5,6	1,2,3,4,5,6	1,2,3,4	1,2,3,4,5,6
Kalimantan Timur	1,2,3,6,7	1,2,3,4	1,2,3,4	1,2,3,4,7	1,2,6	1,2,3,4,6,7
Kalimantan Utara	1,2,3,6,7	1,2,3,4	1,2,4	1,2,3,4,7	1,2,4,6	1,2,3,4,6,7
Sulawesi Utara	1,2,3,6,7	1,2,3,4,6,7	1,2,4,5	1,2,3,4,6,7	1,2,6,7	1,2,3,4,6
Sulawesi Tengah	1,2,3,7	1,2,3,4,6,7	1,2,4,5	1,2,3,4,6,7	1,2,4,6,7	1,2,3,4,5,6
Sulawesi Selatan	1,3,5,6,7	1,3,5,6,7	4,5,6,7	1,3,4,5,6,7	1,2,4,5,6,7	3,4,5,6,7
Sulawesi Tenggara	1,2,3,6,7	1,2,3,4,6,7	1,2,3,4	1,2,3,4,7	1,2,4,6,7	1,2,3,4,5,6
Gorontalo	1,2,3,6,7	1,2,3,4,6,7	1,3,7	1,3,4,7	1,2,6,7	1,2,3,4,5,6
Sulawesi Barat	1,3,5,6,7	1,3,4,5,7	1,4,5,6,7	1,3,4,7	1,6,7	1,2,3,4,5,6
Maluku	1,2,3,6,7	1,2,3,4,6,7	1,2,4,6,7	1,2,3,4,6,7	1,2,3,4,5,6,7	1,2,3,4,6,7
Maluku Utara	1,2,3,6,7	1,2,3,4,6,7	1,3,4,7	1,2,3,4,7	1,2,6,7	1,2,4,5,6,7
Papua Barat	1,2,3,5	1,2,3,4,7	1,2,5,7	1,2,3,4,6,7	1,2,4,6,7	1,2,3,7
Papua	1,2,3,6,7	1,2,3,4,6	1,2,4	1,2,3,4,7	1,2,3,7	1,2,3,4,6

Table 32: Cluster Options of Each Province and Each Missing Data Method using All Distance and Linkage Method Combinations (cont.)

Province	**Cluster Options**					
	CART	**LASSO**	**RI**	**SAMPLE**	**Fill**	**Null**
Aceh	1	1	1	1	1	1
Sumatera Utara	1,2	1,2	1,2	1,2	1,2	1,2
Sumatera Barat	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3
Riau	1,2,3,4	1,2,3,4	1,2,3,4	1,2,3,4	1,2,3,4	1,2,3,4
Jambi	1,2,3,4	1,2,3,4	1,2,3,4	1,2,3,4	1,2,3,4	1,2,3,4
Sumatera Selatan	1,2,4	1,2,3,4	1,2,4	1,2,3,4	1,2,4	1,2,3,4
Bengkulu	1,2,3,4,5	1,2,3,4,5	1,2,3,4,5	1,2,3,4,5	1,2,3,4,5	1,2,3,4
Lampung	1,2,3,4,5	1,2,3,4,5	1,2,3,4,5	1,3,4	1,2,3,4,6	1,3,4,5
Kepulauan Bangka Belitung	1,2,3,4,5	1,2,3,4,5	1,2,3,4,5,6	1,2,3,4,5,6	1,2,3,4,5,6	1,2,3,4
Kepulauan Riau	1,2,3,4,5	1,2,3,4,5,6	1,2,3,4,5,6	1,2,3,4	1,2,3,4	1,2,3,4
DKI Jakarta	1,2,4,5,6	1,2,3,4,5	1,2,3,4,5,6	1,2,4,5,6	1,2,4,5,6	1,2,3,4,5
Jawa Barat	1,2,4	1,2,3,4	1,2,4	1,2,4	1,2,4	1,2,3,4
Jawa Tengah	1,2,4,5,6	1,3,4,5,6	1,2,4,5,6	1,4,5,6	1,2,4,5,6	1,4,5,6
DI Yogyakarta	1,2,3,4,5,6	1,2,3,4,5	1,2,3,4,5	1,2,3,4,5,6	1,2,3,4,5,6	1,3,4,5
Jawa Timur	1,2,3,4,5,6	1,2,3,4,5,6	1,2,3,4,5,6	1,2,3,5,6	1,2,3,5,6	1,3,4,5
Banten	1,2,4,5,6	1,3,4,5,6	1,2,3,4,5,6	1,3,4,6	1,2,3,4,5,6	1,4,5,6
Bali	1,2,3,4,5,6	1,3,4,5,6	1,2,3,4,5,6	1,3,4,5,6	1,3,4,5,6	1,3,4,5
Nusa Tenggara Barat	1,2,4,5,6	1,3,4,5,6	1,4,5,6	1,3,4,5,6	1,4,5,6	1,3,4,5
Nusa Tenggara Timur	1,5,6	1,4,5,6	1,3,5,6,7	1,3,4,5,6,7	1,3,4,5,6,7	1,3,4,5,6
Kalimantan Barat	1,2,3,4,5	1,2,3,4,5	1,2,3,4	1,2,3,4,5	1,2,4,6	1,2,4,5,6,7
Kalimantan Tengah	1,2,3,4,5	1,2,3,4,5,6	1,2,3,4,5,6	1,2,3,4,5,6	1,2,3,4,5,6	1,2,3,4
Kalimantan Selatan	1,2,3,4,5,6	1,2,3,4,5,6	1,2,3,4,5,6	1,2,3,4,5,6	1,2,3,4,5,6	1,2,3,4,5,6
Kalimantan Timur	1,2,3,4,6,7	1,2,3,4,6	1,2,3,4,6,7	1,2,3,4,6	1,2,3,4,6	1,2,3,4,6
Kalimantan Utara	1,2,3,4,6,7	1,2,3,4,6	1,2,4,6,7	1,2,4,5,6	1,2,4,6	1,2,3,4,6
Sulawesi Utara	1,2,4,6	1,2,3,4,6	1,2,3,4,6	1,2,3,4,6	1,2,4,6	1,2,4,6
Sulawesi Tengah	1,2,3,4,6	1,2,3,4,5,6	1,2,4,5,6	1,2,4,5,6	1,2,4,5,6	1,2,3,4,5
Sulawesi Selatan	5,6,7	4,5,6,7	3,5,6,7	5,6,7	5,6,7	1,4,5,6,7
Sulawesi Tenggara	1,2,3,4,6	1,2,3,4,6	1,2,4,6	1,2,4,5,6	1,2,4,5,6,7	1,2,3,4
Gorontalo	1,2,3,6,7	1,2,3,4,6	1,2,3	1,3	1,2,3	1,2,3,4,6,7
Sulawesi Barat	1,2,3,6,7	1,2,4,5,6	1,2,4,5,6	1,3,6	1,2,5,6	1,2,3,4,6,7
Maluku	1,2,4,6,7	1,2,3,4,6,7	1,2,4,6,7	1,2,4,6,7	1,2,4,6,7	1,2,3,4,6,7
Maluku Utara	1,2,3,4,6,7	1,2,3,4,5,6,7	1,2,4,7	1,2,4,5,6,7	1,2,4,5,6,7	1,2,3,4,6,7
Papua Barat	1,2,4,7	1,2,3,5,6,7	1,2,3,6,7	1,2,7	1,2,7	1,2,7
Papua	1,2,4,6	1,2,3,4,5,6	1,2,4,6	1,2,3,4,6	1,2,4,5,6	1,2,3,4,5

Due to aforementioned reasons, picking the most chosen clusters for each province does not work. If we continue picking the most chosen clusters for each province based on these options, some of the missing data methods will not get 7 clusters, which is not what we want. Aside from this technique, there is also an option to pick some Distance and Linkage Method combinations, preferably one or two with different member proportions. By picking two Distance-Linkage combinations, we have enough clusters that can be compared and analyzed between one another.

The chosen two combinations are Manhattan Distance with Ward's Linkage Method (the first one) and Euclidean Distance with Single Linkage Method. These two combinations are chosen because they have extreme differences in terms of the number of cluster members. Euclidean-Single has only one member for 6 clusters, and the rest of the provinces belong to Cluster 1, as shown in Table 34. There is an exception to Mean missing data method where Cluster 3 has two provinces, which makes Cluster 1 has 27 provinces, and the rest of the clusters only have 1 province. Meanwhile, Manhattan-Ward's is better balanced. As shown in Table 33, there are still clusters that only have 1 province as their only member, but there are only maximum 2 clusters per missing data method, which is still acceptable. The clusters with the most members for each missing data method only have 9-14 provinces as their members, so the other clusters don't only have 1 province as their member. The extreme difference will also help in the comparison later on.

After using all possible combinations of distance and linkage methods, we find out that using this dataset, there are several linkage methods that tend to provide clusters where one cluster has the most members with an extreme number. In this case, these methods cluster 25-28 provinces in one cluster, and then put 1-2 province in the other clusters. These methods include Single, Median, Centroid, and Average Linkage Methods. It makes sense for median, centroid, and average that focus on the center point of the data, but it is surprising for single that focuses on the minimum values. However, most of the monthly rainfall values are only 0 mm/month, and since 0 is the minimum value for rainfall, the single linkage method most likely captures these provinces to be similar due to the number of 0 values they have.

Meanwhile, Ward's Linkage Method shows the opposite: they tend to give more balanced results, in the case of the number of cluster members. We still have one cluster that have an extreme number of members, but this extreme cluster only has 11-14 members, which means the other clusters still have a lot of provinces to divide within them. There are still some clusters with only 1 province as their members, but not as many as the previous methods that focus on the center points.

As for the distance methods, Euclidean Distance seems to be the one that fits the cluster calculation the best. Combined with the center-point linkage methods, Euclidean provides the most matches with the most chosen clusters for each province and each missing data method. The same thing happens when Euclidean is combined with Ward's Linkage Method: Euclidean also provides the most matches. After Euclidean, Manhanttan Distance provides the most matches when being combined to the center-point linkage methods and Maximum and Canberra Distances provide the most matches when being combined to Ward's Linkage Method.

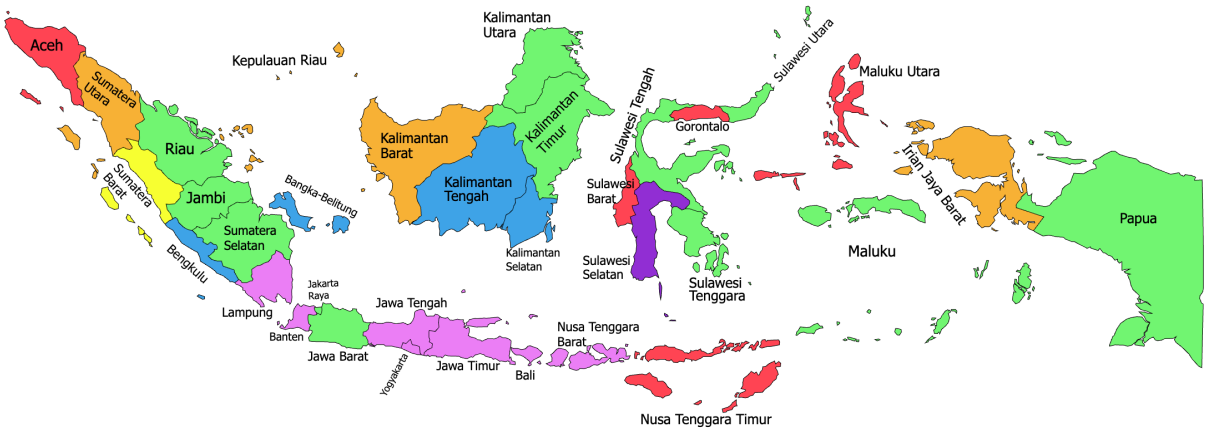
By having two different clusters, we can compare each other. We compare the clusters that are based on Manhattan-Ward's and Euclidean-Single as shown in Table 35. We have some clusters that are similar between these two methods. 15 provinces, which is almost half, of Mean Substitution have matching clusters, followed by Median Substitution and MICE-RI that have 11 matching provinces. MICE-LASSO and Null Substitution also have quite a lot of matches: 10 provinces in total. Aside from these missing data methods, the others have 5-9 provinces with matching clusters within the two clustering methods. This comparison show that even though we have chosen the methods with different outcome conditions, they can still be quite similar to one another.

Table 33: Clusters of Each Province and Each Missing Data Method using Manhattan Distance and Ward's Linkage Method

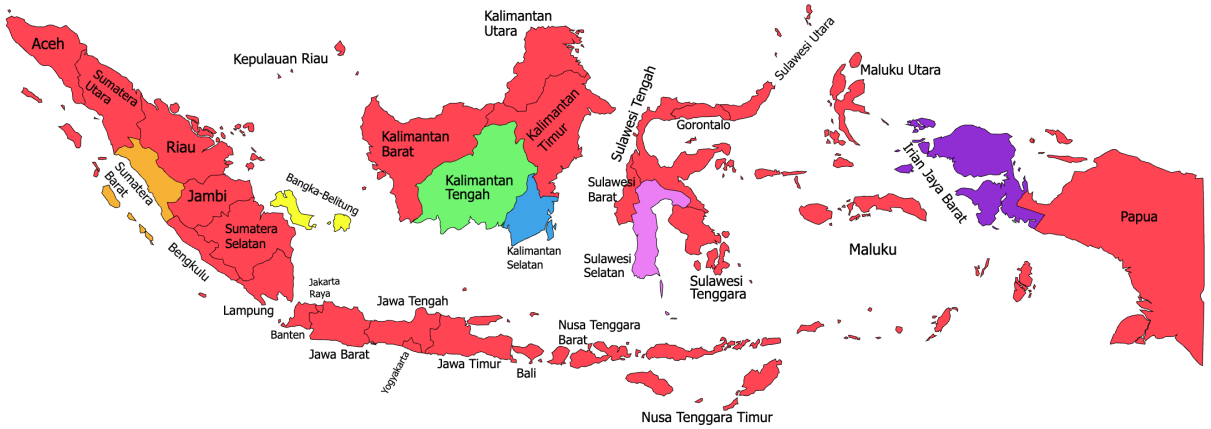
Province	**Clusters**											
	Mean	**Median**	**Boot**	**LOCF**	**Int**	**PMM**	**CART**	**LASSO**	**RI**	**SAMPLE**	**Fill**	**Null**
Aceh	1	1	1	1	1	1	1	1	1	1	1	1
Sumatera Utara	2	2	2	2	2	2	2	1	2	2	2	2
Sumatera Barat	3	3	3	3	3	3	3	2	3	3	3	3
Riau	1	4	4	2	1	4	4	1	2	4	4	4
Jambi	1	4	4	2	1	4	4	1	2	4	4	4
Sumatera Selatan	1	4	4	2	1	4	4	1	4	4	4	4
Bengkulu	4	5	5	2	4	5	5	3	4	5	5	3
Lampung	1	1	6	1	1	6	1	4	1	1	6	5
Kepulauan Bangka Belitung	4	5	5	2	4	2	5	3	5	5	5	4
Kepulauan Riau	2	2	2	2	2	2	2	5	2	2	2	2
DKI Jakarta	5	1	4	4	5	4	6	1	4	4	4	5
Jawa Barat	1	4	4	2	1	4	4	1	4	4	4	4
Jawa Tengah	6	1	6	1	5	6	6	4	1	6	6	5
DI Yogyakarta	6	1	6	1	5	6	6	1	4	6	6	5
Jawa Timur	6	1	6	1	5	6	6	4	1	6	6	5
Banten	6	1	6	1	5	6	6	4	1	6	6	5
Bali	6	1	6	1	5	6	6	4	1	6	6	5
Nusa Tenggara Barat	6	1	6	1	5	6	6	4	1	6	6	5
Nusa Tenggara Timur	6	1	1	5	6	6	6	4	1	6	6	5
Kalimantan Barat	4	4	2	2	2	2	5	3	2	5	4	2
Kalimantan Tengah	4	5	5	6	4	5	5	3	5	5	5	4
Kalimantan Selatan	4	5	5	6	4	5	5	3	4	5	5	4
Kalimantan Timur	7	4	4	7	1	4	4	6	6	4	4	6
Kalimantan Utara	7	4	4	7	1	4	4	6	6	4	4	6
Sulawesi Utara	1	4	4	7	7	2	4	6	6	4	4	1
Sulawesi Tengah	1	6	4	7	7	4	4	6	6	4	4	1
Sulawesi Selatan	5	7	7	4	5	7	7	7	7	7	7	7
Sulawesi Tenggara	1	6	4	7	7	4	4	6	6	4	4	1
Gorontalo	1	6	1	7	7	1	1	6	1	1	1	1
Sulawesi Barat	1	1	1	7	7	1	1	6	1	1	1	1
Maluku	1	6	4	7	7	4	4	6	6	4	4	1
Maluku Utara	1	6	1	7	7	1	1	6	1	1	1	1
Papua Barat	2	2	2	2	2	2	2	5	2	2	2	2
Papua	1	6	4	2	2	4	4	1	2	4	4	1

Table 34: Clusters of Each Province and Each Missing Data Method using Euclidean Distance and Single Linkage Method

Province	**Clusters**											
	Mean	**Median**	**Boot**	**LOCF**	**Int**	**PMM**	**CART**	**LASSO**	**RI**	**SAMPLE**	**Fill**	**Null**
Aceh	1	1	1	1	1	1	1	1	1	1	1	1
Sumatera Utara	1	1	1	1	1	1	1	1	1	1	1	1
Sumatera Barat	2	2	2	2	2	2	2	2	2	2	2	2
Riau	1	1	1	1	1	1	1	1	1	1	1	1
Jambi	1	1	1	1	1	1	1	1	1	1	1	1
Sumatera Selatan	1	1	1	1	1	1	1	1	1	1	1	1
Bengkulu	1	1	1	1	1	3	3	1	1	1	1	3
Lampung	1	1	1	1	1	1	1	1	1	1	1	1
Kepulauan Bangka Belitung	3	3	3	1	1	1	1	1	3	3	1	1
Kepulauan Riau	4	1	1	1	1	1	1	3	1	1	3	1
DKI Jakarta	5	4	1	3	3	1	1	1	1	1	1	1
Jawa Barat	1	1	1	1	1	1	1	1	1	1	1	1
Jawa Tengah	1	1	1	1	1	1	1	1	1	1	1	1
DI Yogyakarta	1	1	1	1	1	1	1	1	1	1	1	1
Jawa Timur	1	1	1	1	1	1	1	1	1	1	1	1
Banten	1	1	1	1	1	1	1	1	1	1	1	1
Bali	1	1	1	1	1	1	1	1	1	1	1	1
Nusa Tenggara Barat	1	1	1	1	1	1	1	1	1	1	1	1
Nusa Tenggara Timur	1	1	1	4	4	1	1	1	1	1	1	1
Kalimantan Barat	3	1	1	5	1	1	1	1	1	1	1	4
Kalimantan Tengah	6	5	4	6	5	4	4	4	4	4	4	1
Kalimantan Selatan	1	1	5	1	1	5	5	5	5	5	5	5
Kalimantan Timur	1	1	1	1	1	1	1	1	1	1	1	1
Kalimantan Utara	1	1	1	1	1	1	1	1	1	1	1	1
Sulawesi Utara	1	1	1	1	1	1	1	1	1	1	1	1
Sulawesi Tengah	1	1	1	1	1	1	1	1	1	1	1	1
Sulawesi Selatan	7	6	6	1	6	6	6	6	6	6	6	6
Sulawesi Tenggara	1	1	1	1	1	1	1	1	1	1	1	1
Gorontalo	1	1	1	1	1	1	1	1	1	1	1	1
Sulawesi Barat	1	1	1	1	1	1	1	1	1	1	1	1
Maluku	1	1	1	1	1	1	1	1	1	1	1	1
Maluku Utara	1	1	1	1	1	1	1	1	1	1	1	1
Papua Barat	1	7	7	7	7	7	7	7	7	7	7	7
Papua	1	1	1	1	1	1	1	1	1	1	1	1



(a) using Manhattan-Ward's



(b) using Euclidean-Single

Figure 37: Map of Indonesia Divided by Clusters (Using Bootstrap Data)

It is also interesting to see the provinces that have matching clusters (see Table 35). Aceh, always being in the same cluster, always matches no matter which missing data method is being used. Some other provinces that also have many matching clusters between these two clustering methods are Sulawesi Barat (9 matches), Gorontalo (8 matches), Maluku Utara (8 matches), and Lampung (7 matches). There are also some provinces that have no matches at all: Kepulauan Bangka Belitung, Kepulauan Riau, Kalimantan Barat, Sulawesi Selatan, and Papua Barat. Meanwhile, the other provinces have matches ranging from 1 to 3, with one province having 5 matching clusters. Most provinces, there are 12 in total, have 3 matches.

We also show the Clusters of Bootstrap Data based on Table 33 and Table 34 in Figure 37, and the differences between Manhattan-Ward's and Euclidean-Single is more obvious through the pictures. With Manhattan Distance and Ward's Linkage Method, we obtain more balanced number of provinces per cluster (see Figure 37a). However, using Euclidean Distance and Single Linkage Method, one cluster has all the provinces except for six provinces which are parts of other clusters (see Figure 37b).

Table 35: Comparing Clusters Between Manhattan-Ward's and Euclidean-Single

Province	**Clusters**											
	Mean	**Median**	**Boot**	**LOCF**	**Int**	**PMM**	**CART**	**LASSO**	**RI**	**SAMPLE**	**Fill**	**Null**
Aceh	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Sumatera Utara								✓				
Sumatera Barat								✓				
Riau	✓				✓			✓				
Jambi	✓				✓			✓				
Sumatera Selatan	✓				✓			✓				
Bengkulu												✓
Lampung	✓	✓		✓	✓		✓		✓	✓		
Kepulauan Bangka Belitung												
Kepulauan Riau												
DKI Jakarta	✓							✓				
Jawa Barat	✓				✓			✓				
Jawa Tengah		✓		✓					✓			
DI Yogyakarta		✓		✓				✓				
Jawa Timur		✓		✓					✓			
Banten		✓		✓					✓			
Bali		✓		✓					✓			
Nusa Tenggara Barat		✓		✓					✓			
Nusa Tenggara Timur		✓	✓						✓			
Kalimantan Barat												
Kalimantan Tengah		✓		✓								
Kalimantan Selatan			✓			✓	✓			✓	✓	
Kalimantan Timur					✓							
Kalimantan Utara					✓							
Sulawesi Utara	✓											✓
Sulawesi Tengah	✓											✓
Sulawesi Selatan												
Sulawesi Tenggara	✓											✓
Gorontalo	✓		✓			✓	✓		✓	✓	✓	✓
Sulawesi Barat	✓	✓	✓			✓	✓		✓	✓	✓	✓
Maluku	✓											✓
Maluku Utara	✓		✓			✓	✓		✓	✓	✓	✓
Papua Barat												
Papua	✓							✓				✓
TOTAL	15	11	6	9	8	5	6	10	11	6	5	10

3.5.2 Province per Cluster Weight

Table 36 shows the weight of each province to obtain one series for each cluster using the combination of Manhattan-Ward's clustering. Just like Region in Table 30, the more provinces in a cluster, the less contribution each province can give. Due to some clusters only having 1 province in them, some provinces will have 100% contribution for these clusters, showing that the series of the cluster is exactly the same as the series of the province. A cluster example for this would be Cluster 3 in Mean missing data method, since this cluster only has Sumatera Barat as its member and this province has 100% weight. Meanwhile, clusters with 2 provinces have equal weight: 50% for both provinces, just like Maluku and Papua in Region calculation. In this case, Cluster 5 in Mean missing data method has only two provinces: DKI Jakarta and Sulawesi Selatan, and both have 50% weight.

Table 37 shows the same thing, but using the combination of Euclidean-Single clustering method. Since this combination provides one cluster with a lot of provinces and several clusters with only 1 or 2 provinces as the members, some of the provinces that belong to the cluster with a lot of members certainly have less weight. Take Aceh, for example, as a province that remains in one cluster for the entire calculation. Aceh's weights in Manhattan-Ward's are higher than in Euclidean-Single, about 2-5 times bigger. This is because in Manhattan-Ward's, Cluster 1 only has at most 14 members, meanwhile Euclidean-Single's Cluster 1 can have up to 28 members in one cluster alone. Meanwhile, we still have Sumatera Barat with 100% weight on both methods.

3.5.3 Final Cluster Series

The final series after the weighted values are implemented is discussed in this part. There are 14 different series to look at, with the combination of 7 clusters and 2 methods of clustering. Some of them look almost similar to one another within the same clustering method. Figure 38 shows the series of Cluster 1. The series from Manhattan-Ward's and Euclidean-Single look similar, ranging from 0-500 mm/month, except for one data point in Mean Substitution for Euclidean-Single that reaches almost 600 mm/month.

Figure 39 shows the series of Cluster 2. The Euclidean-Single series is only for Sumatera Barat. MICE-LASSO in Manhattan-Ward's is also only for Sumatera Barat, which is why the graphs for the two clusters are the same. However, the other missing data methods in Manhattan-Ward's have more than one province in Cluster 2, which makes the other series different from Euclidean-Single. This time, the series looks rather different for the two clustering methods. Manhattan-Ward's has around 0-500 mm/month range, except for MICE-LASSO, which is caused by the cluster only having one province as the member. Meanwhile, Euclidean-Single has big range, especially seen in LOCF Imputation and Interpolation. Again, this is due to the cluster only having one province.

The final series of Cluster 3 is shown in Figure 40. Cluster 3 only consists of Sumatera Barat alone for all missing methods in Manhattan-Ward's except for MICE-LASSO (Cluster 2) and Null Substitution. Cluster 3 of Manhattan-Ward's also consists of 5 provinces. This is, however, not the case for Euclidean-Single, since Sumatera Barat is in Cluster 2. However, that means that we can say that the values of Cluster 3 of Manhattan-Ward's and Cluster 2 of Euclidean-Single are mostly the same. This also shows that Sumatera Barat is most likely to have a very different data pattern that makes it hard to be grouped with other provinces. As for Cluster 3 of Euclidean-Single, aside from Mean Substitution, it always consists of one province that is either Bengkulu, Kepulauan Bangka Belitung, Kepulauan Riau, or DKI Jakarta. The only Cluster 3 of Euclidean-Single that has more than one province is Mean-Substitution. Overall, we have a larger range in this series due to this cluster having mostly one province as its member, with some data points almost reaching 2000 mm/month.

Table 36: Weight (%) of Each Province for Cluster (Manhattan-Ward's)

Province	Proportion for Each Cluster (%)											
	Mean	**Median**	**Boot**	**LOCF**	**Int**	**PMM**	**CART**	**LASSO**	**RI**	**SAMPLE**	**Fill**	**Null**
Aceh	0.06	0.21	0.35	0.27	0.10	0.49	0.38	0.08	0.18	0.38	0.49	0.10
Sumatera Utara	0.40	0.40	0.22	0.08	0.11	0.20	0.40	0.10	0.09	0.40	0.40	0.22
Sumatera Barat	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.68
Riau	0.10	0.14	0.10	0.10	0.16	0.10	0.10	0.13	0.12	0.10	0.08	0.19
Jambi	0.06	0.08	0.05	0.05	0.09	0.05	0.05	0.07	0.06	0.05	0.05	0.10
Sumatera Selatan	0.10	0.14	0.09	0.09	0.16	0.09	0.09	0.12	0.47	0.09	0.08	0.18
Bengkulu	0.05	0.09	0.09	0.02	0.09	0.10	0.05	0.05	0.11	0.05	0.09	0.32
Lampung	0.04	0.12	0.22	0.16	0.06	0.17	0.22	0.17	0.11	0.22	0.17	0.17
Kepulauan Bangka Belitung	0.04	0.07	0.07	0.02	0.07	0.05	0.04	0.04	0.10	0.04	0.07	0.04
Kepulauan Riau	0.05	0.05	0.02	0.01	0.01	0.02	0.05	0.08	0.01	0.05	0.05	0.02
DKI Jakarta	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Jawa Barat	0.04	0.06	0.04	0.04	0.07	0.04	0.04	0.05	0.20	0.04	0.03	0.08
Jawa Tengah	0.21	0.13	0.22	0.16	0.21	0.17	0.21	0.17	0.11	0.21	0.17	0.17
DI Yogyakarta	0.02	0.01	0.02	0.02	0.02	0.02	0.02	0.00	0.02	0.02	0.02	0.02
Jawa Timur	0.29	0.18	0.31	0.23	0.29	0.24	0.29	0.24	0.15	0.29	0.24	0.24
Banten	0.06	0.03	0.06	0.04	0.06	0.05	0.06	0.05	0.03	0.06	0.05	0.05
Bali	0.03	0.02	0.04	0.03	0.03	0.03	0.03	0.03	0.02	0.03	0.03	0.03
Nusa Tenggara Barat	0.12	0.07	0.13	0.09	0.12	0.10	0.12	0.10	0.06	0.12	0.10	0.10
Nusa Tenggara Timur	0.28	0.17	0.28	1.00	1.00	0.23	0.28	0.24	0.15	0.28	0.23	0.23
Kalimantan Barat	0.39	0.24	0.45	0.16	0.23	0.41	0.39	0.39	0.19	0.39	0.14	0.45
Kalimantan Tengah	0.41	0.67	0.67	0.81	0.67	0.73	0.41	0.41	0.90	0.41	0.67	0.33
Kalimantan Selatan	0.10	0.16	0.16	0.19	0.16	0.18	0.10	0.10	0.20	0.10	0.16	0.08
Kalimantan Timur	0.64	0.20	0.14	0.31	0.23	0.14	0.14	0.31	0.36	0.14	0.12	0.64
Kalimantan Utara	0.36	0.11	0.08	0.17	0.13	0.08	0.08	0.17	0.20	0.08	0.06	0.36
Sulawesi Utara	0.02	0.02	0.02	0.03	0.07	0.04	0.02	0.03	0.04	0.02	0.01	0.02
Sulawesi Tengah	0.07	0.12	0.07	0.15	0.28	0.07	0.07	0.15	0.17	0.07	0.06	0.10
Sulawesi Selatan	0.99	1.00	1.00	0.99	0.27	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Sulawesi Tenggara	0.04	0.07	0.04	0.09	0.17	0.04	0.04	0.09	0.10	0.04	0.03	0.06
Gorontalo	0.01	0.02	0.07	0.03	0.06	0.10	0.08	0.03	0.04	0.08	0.10	0.02
Sulawesi Barat	0.02	0.06	0.10	0.04	0.08	0.14	0.11	0.04	0.05	0.11	0.14	0.03
Maluku	0.05	0.09	0.05	0.11	0.21	0.05	0.05	0.11	0.13	0.05	0.04	0.08
Maluku Utara	0.04	0.06	0.19	0.08	0.14	0.27	0.21	0.08	0.10	0.21	0.27	0.05
Papua Barat	0.55	0.55	0.30	0.11	0.16	0.28	0.55	0.92	0.13	0.55	0.55	0.30
Papua	0.35	0.63	0.34	0.33	0.49	0.34	0.34	0.44	0.40	0.34	0.29	0.53

Table 37: Weight (%) of Each Province for Cluster (Euclidean-Single)

Province	Proportion for Each Cluster (%)											
	Mean	**Median**	**Boot**	**LOCF**	**Int**	**PMM**	**CART**	**LASSO**	**RI**	**SAMPLE**	**Fill**	**Null**
Aceh	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
Sumatera Utara	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Sumatera Barat	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Riau	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06
Jambi	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
Sumatera Selatan	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06
Bengkulu	0.01	0.01	0.01	0.01	0.01	1.00	1.00	0.01	0.01	0.01	0.01	1.00
Lampung	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
Kepulauan Bangka Belitung	0.10	1.00	1.00	0.01	0.01	0.01	0.01	0.01	1.00	1.00	0.01	0.01
Kepulauan Riau	1.00	0.01	0.01	0.01	0.01	0.01	0.01	1.00	0.01	0.01	1.00	0.01
DKI Jakarta	1.00	1.00	0.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Jawa Barat	0.03	0.02	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
Jawa Tengah	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
DI Yogyakarta	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Jawa Timur	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
Banten	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Bali	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Nusa Tenggara Barat	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Nusa Tenggara Timur	0.03	0.03	0.03	1.00	1.00	0.03	0.03	0.03	0.03	0.03	0.03	0.03
Kalimantan Barat	0.90	0.10	0.10	1.00	0.10	0.10	0.10	0.10	0.10	0.10	0.10	1.00
Kalimantan Tengah	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.10
Kalimantan Selatan	0.03	0.02	1.00	0.03	0.02	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Kalimantan Timur	0.09	0.08	0.08	0.09	0.08	0.09	0.09	0.08	0.08	0.08	0.08	0.08
Kalimantan Utara	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Sulawesi Utara	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Sulawesi Tengah	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
Sulawesi Selatan	1.00	1.00	1.00	0.03	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Sulawesi Tenggara	0.02	0.02	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
Gorontalo	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Sulawesi Barat	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Maluku	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
Maluku Utara	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
Papua Barat	0.07	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Papua	0.21	0.20	0.21	0.22	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21

Figure 41 shows the series of Cluster 4. Euclidean-Single only has one province for Cluster 4: either Kepulauan Riau, DKI Jakarta, Nusa Tenggara Timur, Kalimantan Barat, or Kalimantan Tengah. Unlike Euclidean-Distance, Manhattan-Ward's that has more than one province as their members. Overall, the range of the final series is from 0-500 mm/month, especially for those with more than one cluster member. However, there are some interesting cases where the maximum value reaches 1000 mm/month, even almost 1500 mm/month. This happens in Manhattan-Ward's LOCF Imputation where the cluster only has two provinces. As for Euclidean-Single LOCF Imputation and Interpolation, it is obvious that there are some big imputed values that affected this series because some of the data points exceed 2000 mm/month.

The final series of Cluster 5 is shown in Figure 42. Again, Euclidean-Single only has one member for Cluster 5 which is either DKI Jakarta, Kalimantan Barat, Kalimantan Tengah, or Kalimantan Selatan, but mostly Kalimantan Selatan. Meanwhile, Manhattan-Ward's has more than one province in this cluster, except for LOCF Imputation which only has Nusa Tenggara Timur as its member. This is the same as Euclidean-Single's Cluster 4, which is why their series is identical. Most of the series still have 0-500 mm/month range, but there are also some that reach 0-1000 mm/month range. Meanwhile, some extreme cases like Manhattan-Ward's and Euclidean-Single in Mean Substitution and LOCF Imputation reach 0-2000 mm/month range. This is most likely due to the less number of members (Manhattan-Ward's in Mean Substitution only consists of two provinces).

Figure 43 shows the series of Cluster 6. This time, there is one missing data method that only has one province with Manhattan-Ward's (Interpolation), and all Cluster 6's that are grouped by Euclidean-Single only have one province. The province in Manhattan-Ward's Interpolation is Nusa Tenggara Timur, which is the same as the Cluster 4 of Euclidean-Single. Again, in the case of clusters with more than 2 provinces, we tend to have 0-500 mm/month range, meanwhile the rest of them tend to exceed this range. Euclidean-Single in this case has 0-1500 mm/month range, meanwhile the highest value of Manhattan-Ward's in Interpolation even exceeds 2000 mm/month.

The final series of Cluster 7 is shown in Figure 44. Manhattan-Ward's has one province in most cluster except for Mean Substitution, LOCF Imputation, and Interpolation, and this province is Sulawesi Selatan. As for Euclidean-Single, it only has 1 province, which is mostly Papua Barat. This graph once again shows that clusters with more than two members have smaller values, ranging from 0-500 mm/month. Manhattan-Ward's for Mean Substitution is still in 0-1000 mm/month range because it only has two provinces in the cluster. Meanwhile, Euclidean-Single definitely exceeds the 0-500 mm/month range. Most of Manhattan-Ward's series looks similar to the Cluster 6 of Euclidean-Single because they both only have Sulawesi Selatan as their member for some missing data methods.

After seeing all the series, it seems like Clustering helps to lessen the extreme values that are obtained from the imputation process. By balancing the number of provinces in each cluster while making sure each province has more than 2 provinces, the rainfall values become less extreme. Less extreme values means they will also become more predictable when it comes to forecasting. In addition, the values for the series can be seen in Appendix C.

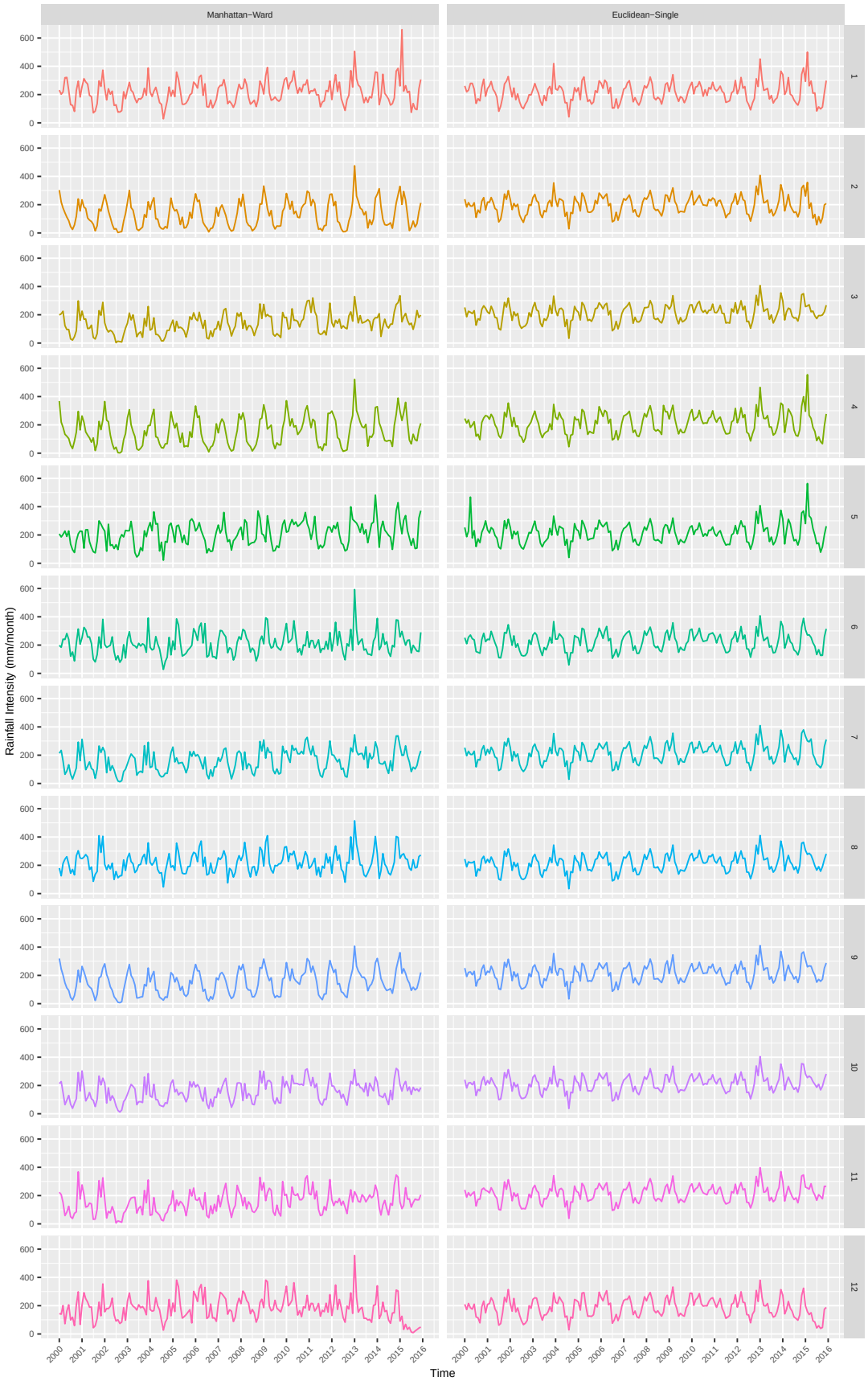


Figure 38: Final Series of Cluster 1 Data (Weighted)

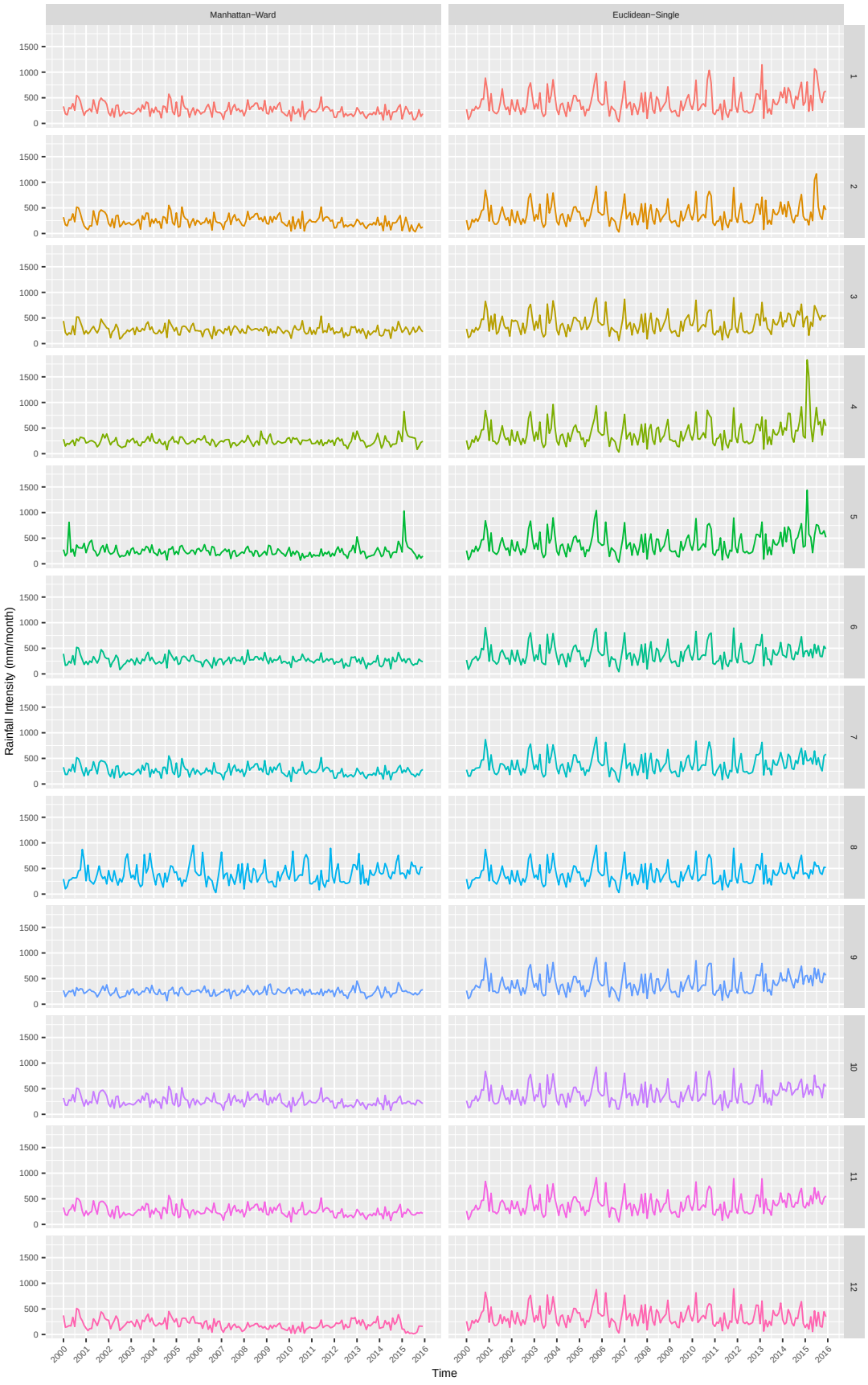


Figure 39: Final Series of Cluster 2 Data (Weighted)

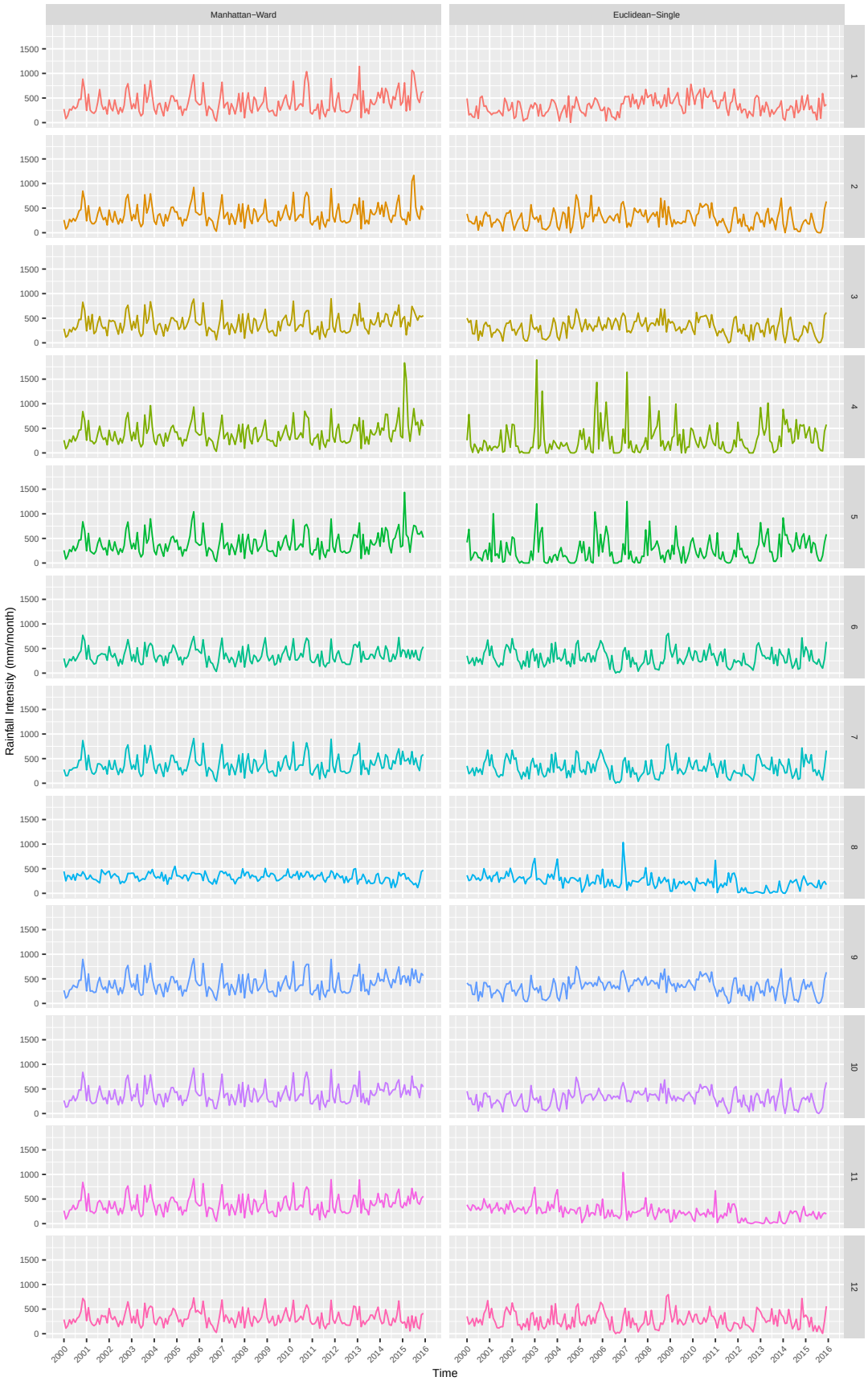


Figure 40: Final Series of Cluster 3 Data (Weighted)

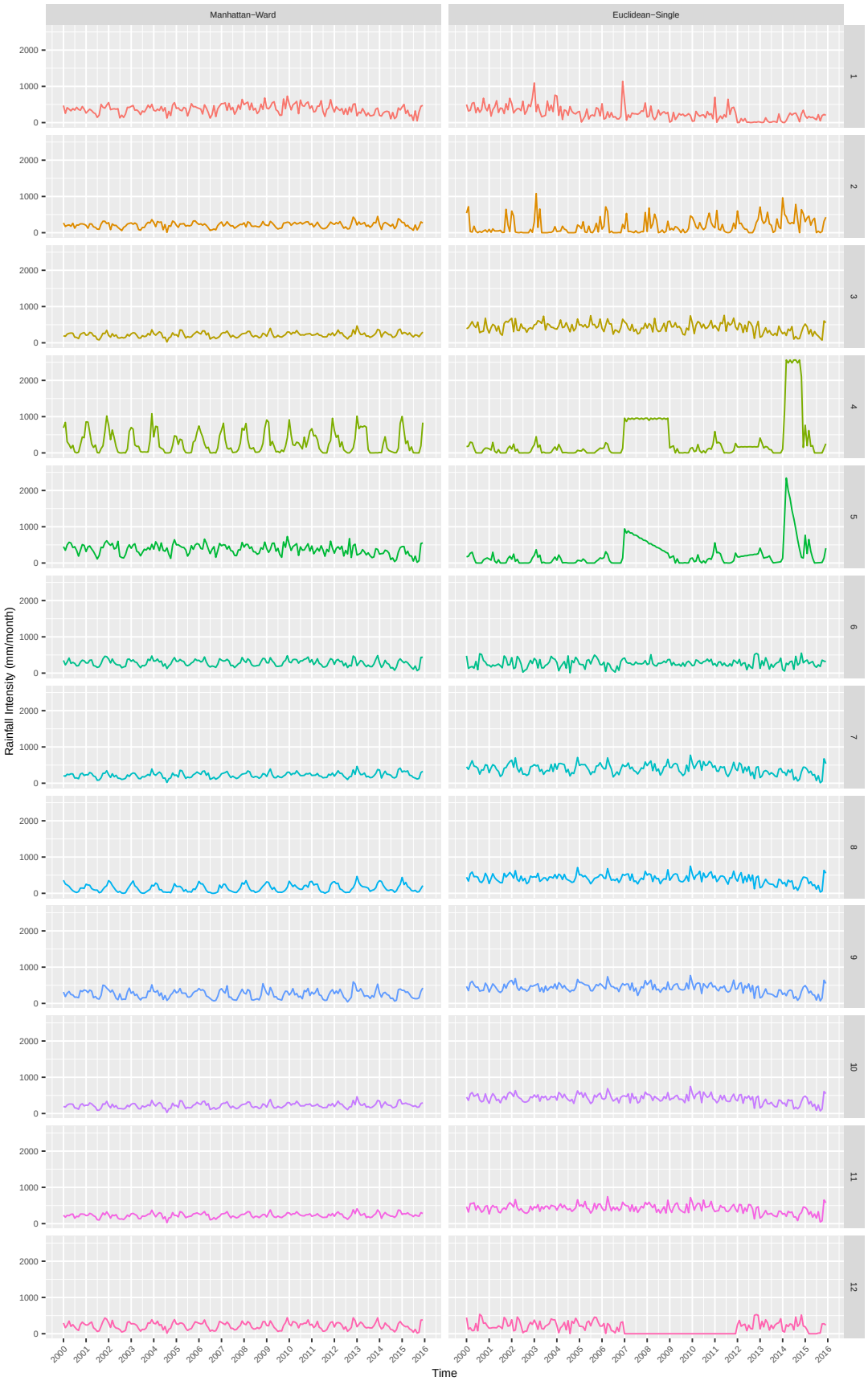


Figure 41: Final Series of Cluster 4 Data (Weighted)

Figure 42: Final Series of Cluster 5 Data (Weighted)

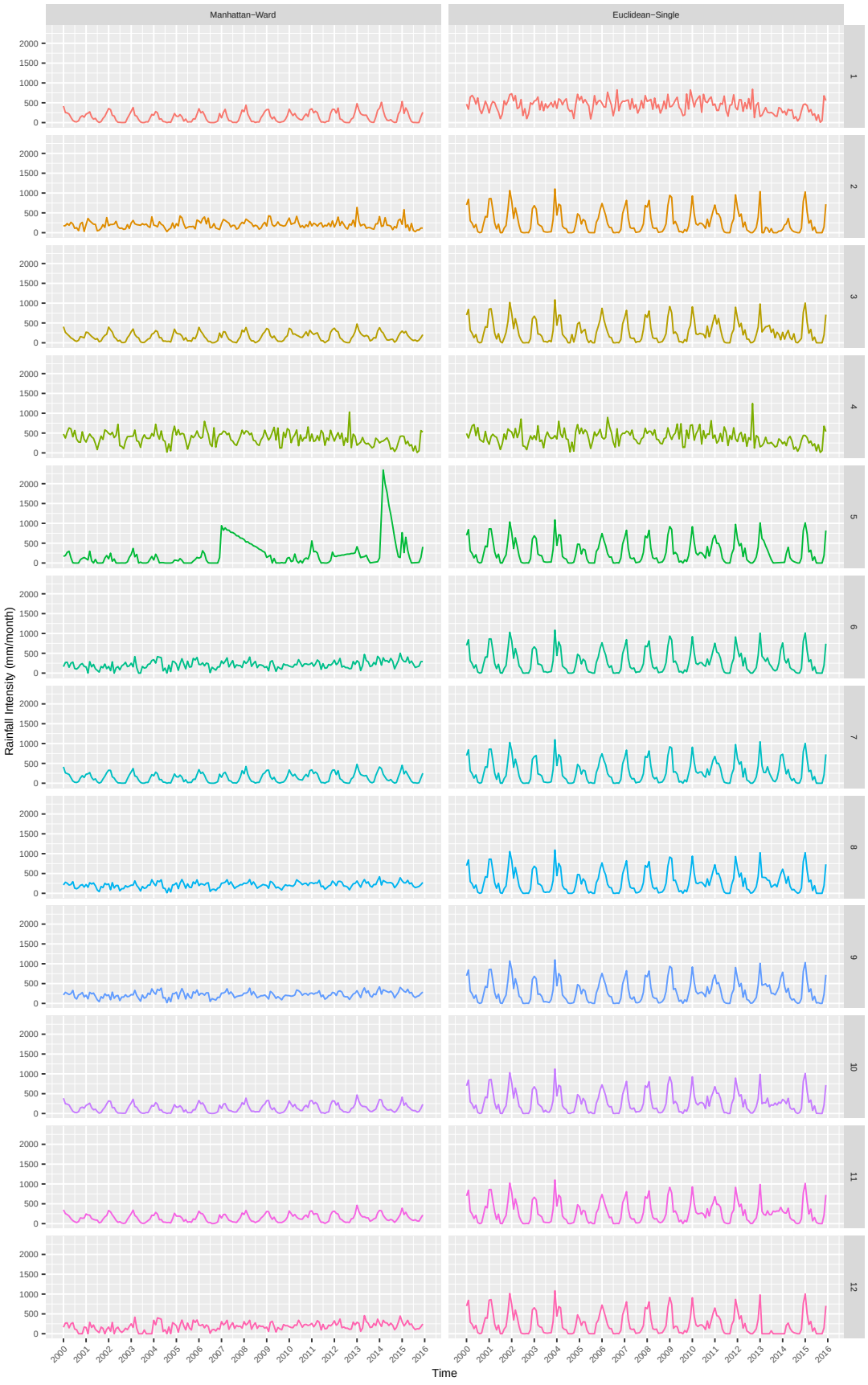


Figure 43: Final Series of Cluster 6 Data (Weighted)

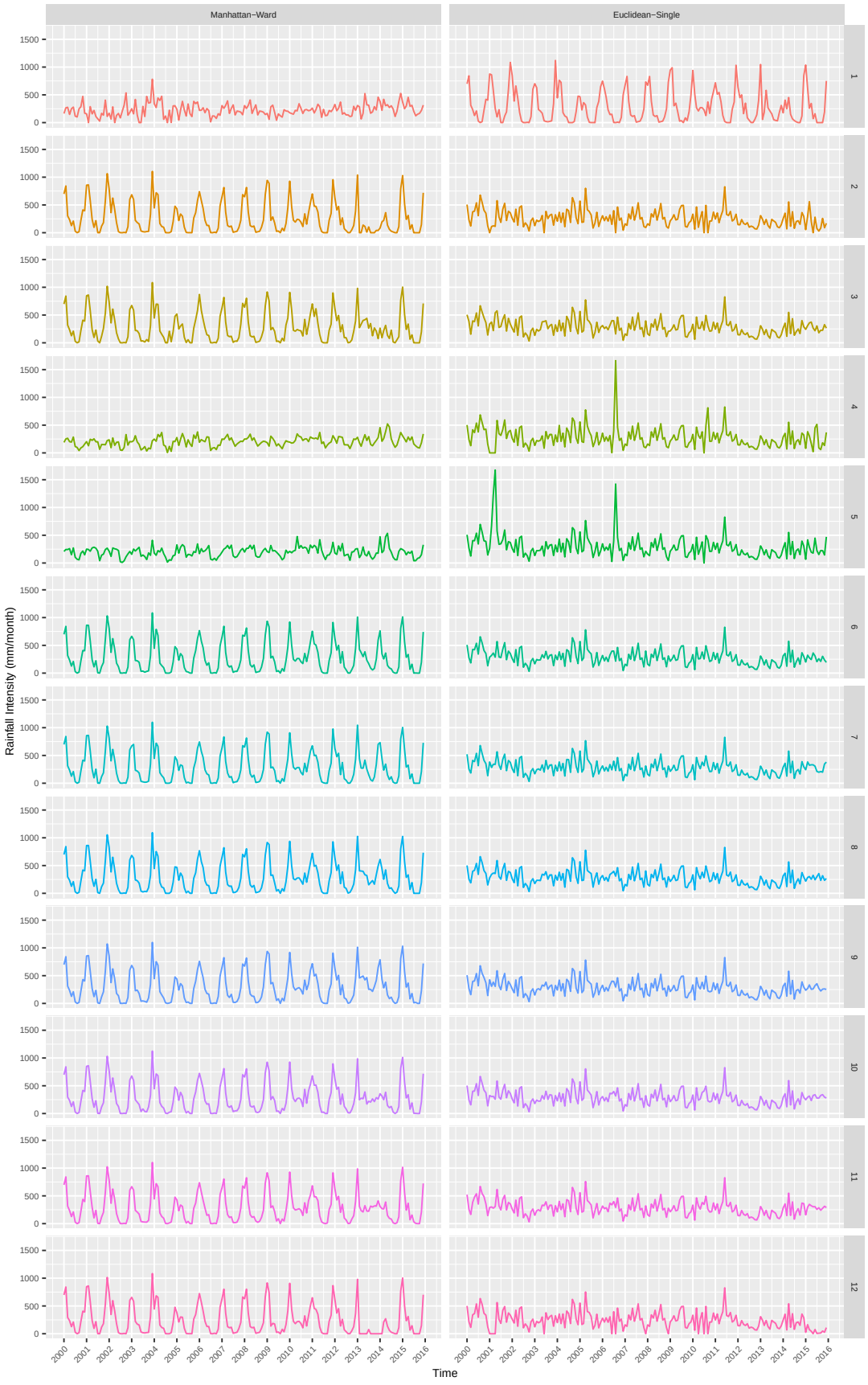


Figure 44: Final Series of Cluster 7 Data (Weighted)

4 Singular Spectrum Analysis

Singular Spectrum Analysis (SSA) is a nonparametric spectral estimation method that combines time series analysis, multivariate statistics, multivariate geometry, dynamical systems and signal processing. Two main applications for SSA are: filtering & smoothing, and forecasting. Both applications are used in the calculation, but the focus of the research is forecasting. We first describe the method for the Univariate SSA in Section 4.1 before we provide some theoretical background in Section 4.2 for further explanations.

4.1 Univariate Singular Spectrum Analysis

a) Stage 1: Decomposition

(i) Embedding
This step can be considered as a mapping process that transforms one-dimensional time series $\{y_t : t = 1, \ldots, N\}$ to a multi-dimensional time series $\{x_t : t = 1, \ldots, N\}$ with values $x_t = (y_t, \ldots, y_{t+L-1}) \in \mathbb{R}^L$ where L is the window length with $2 \leq L \leq N/2$. The multidimensional series $\mathbf{X} = (X_1, \ldots, X_K)$ is called trajectory matrix, or Hankel matrix, needed for the next step. The default method for this transformation in the R package is column-based trajectory matrix, which means the number of columns is equal to the window length. When it is a row-based trajectory matrix, the number of rows is equal to the window length. Using window length L and K, where $K = N - L + 1$, we have a matrix with $L \times K$ dimension as our row-based trajectory matrix:

$$\mathbf{X} = \begin{bmatrix} y_1 & y_2 & y_3 & \cdots & y_K \\ y_2 & y_3 & y_4 & \cdots & y_{K+1} \\ y_3 & y_4 & y_5 & \cdots & y_{K+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_L & y_{L+1} & y_{L+2} & \cdots & y_N \end{bmatrix}$$

And we have a matrix with $K \times L$ dimension as our column-based trajectory matrix, which is the transposed version of the row-based trajectory matrix:

$$\mathbf{X}^T = \begin{bmatrix} y_1 & y_2 & y_3 & \cdots & y_L \\ y_2 & y_3 & y_4 & \cdots & y_{L+1} \\ y_3 & y_4 & y_5 & \cdots & y_{L+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_K & y_{K+1} & y_{K+2} & \cdots & y_N \end{bmatrix}$$

(ii) Singular Value Decomposition (SVD)
The following calculations are presented only for the row-based trajectory matrix. We calculate $\mathbf{X}\mathbf{X}^T$ in order to obtain the eigenvalues that, in decreasing order, are denoted by $\lambda_1 \geq ... \geq \lambda_L \geq 0$. We also obtain the corresponding eigenvectors $U_1, ..., U_L$ of $\mathbf{X}\mathbf{X}^T$. These are left eigenvectors of $\mathbf{X}$. The SVD can be defined as $\mathbf{X} = \mathbf{X}_1 + ... + \mathbf{X}_L$, where $\mathbf{X}_i = \sqrt{\lambda_i} U_i V_i^T$, and $V_i = \mathbf{X}^T U_i / \sqrt{\lambda_i}$. The collection $(\sqrt{\lambda_i}, U_i, V_i)$ is called the i-th eigentriple, formed by the eigenvalues and the left and right eigenvectors.

b) Stage 2: Reconstruction

(i) Eigentriple Grouping
Eigentriple Grouping is the process of choosing sets $I_1, ..., I_B$ to distinguish the signal and the noise. Each set contains integer values with the minimum number of 1 and the maximum number L/B of elements per set with $B \leq L$.

Each set can have a different number of values as long as there are no overlapping values between sets, e.g. For $L = 30$, the sets can be $\{I_1 = \{1, ..., 10\}, I_2 = \{11, ..., 20\}, I_3 = \{21, ..., 30\}\}$ or $\{I_1 = \{1, ..., 30\}\}$. It is also possible to not use all indises up to L, if we only want to use two sets with $\{I_1 = \{1, ..., 15\}, I_2 = \{16, ..., 20\}\}$. It is also possible for each set to have a different number of values, and possible for the values to not be consecutive, e.g. $\{I_1 = \{2, 3, 5, 15\}\}$.

We split the trajectory matrix $\mathbf{X}$ into components arising from several disjoint subsets denoting $\mathbf{X} \equiv \mathbf{X}_{I_1} + ... + \mathbf{X}_{I_B}$. We can freely choose these sets using various sizes, for our case, we decided to only use one set ($B = 1$), which is $I_1 = \{1, ..., d\}$. Hence, we have:

$$\tilde{\mathbf{X}} \equiv \mathbf{X}_{I_1}, i.e. \tag{4.1}$$

$$\tilde{\mathbf{X}} \equiv \sum_{j \in I_1} \mathbf{X}_i = \mathbf{X}_1 + ... + \mathbf{X}_d \tag{4.2}$$

with $d < L$.

(ii) Diagonal Averaging
The purpose of diagonal averaging is to transform each matrix $\mathbf{X}_{I_j}$ to a new series $\tilde{y}_{I_j} = \left(\tilde{y}_{I_j,1}, ..., \tilde{y}_{I_j,N}\right)$. Let each $\mathbf{X}_{I_j}$ be filled with elements $y^*_{I_j,l,k}, 1 \leq l \leq L, 1 \leq k \leq K$. In order to apply diagonal averaging, we use:

$$\tilde{y}_{I_j,t} = \begin{cases} \dfrac{1}{t} \displaystyle\sum_{m=1}^{t} y^*_{I_j,m,t-m+1} & \text{for } 1 \leq t < L \\ \dfrac{1}{L} \displaystyle\sum_{m=1}^{L} y^*_{I_j,m,t-m+1} & \text{for } L \leq t \leq K \\ \dfrac{1}{N-t+1} \displaystyle\sum_{m=t-K+1}^{N-K+1} y^*_{I_j,m,t-m+1} & \text{for } K < t \leq N \end{cases} \tag{4.3}$$

c) Stage 3: Forecasting
There are two forecast methods in SSA: Recurrent and Vector. The main difference is that Recurrent Forecast first obtain approximations of $y_1, \ldots, y_n$ by diagonalizing $\tilde{X}$ and then forecast, while Vector Forecast first forecasts additional columns of $\tilde{X}$ and then diagonalizes.

(i) Recurrent Forecast
We refer to Section 4.2 where we motivate the formulas in detail. For any vector U denote by U^{∇} the vector consisting of the first $L-1$ components of the vector U. Let $\hat{y}_{N+1}, ..., \hat{y}_{N+h}$ be the h terms of the SSA recurrent forecast. These are obtained by the following formula:

$$\hat{y}_i = \begin{cases} \tilde{y}_t & \text{for } t = 1, ..., N \\ \sum_{j=1}^{L-1} \alpha_j \hat{y}_{t-j} & \text{for } t = N+1, ..., N+h \end{cases}$$

where $\tilde{y}_i, i = 1, ..., N$ is the reconstructed series (noise reducted series) from $\tilde{X}$ and vector $A = (\alpha_{L-1}, ..., \alpha_1)$ is computed by:

$$A = \begin{bmatrix} \alpha_{L-1} \\ \vdots \\ \alpha_1 \end{bmatrix} = \frac{1}{1-v^2} \sum_{i-1}^{r} U_{i,L} U_i^{\nabla}$$

(ii) Vector Forecast
Consider the following matrix:

$$\Pi = \mathbf{V}^{\nabla}(\mathbf{V}^{\nabla})^T + (1-v^2)AA^T,$$

where $\mathbf{V}^{\nabla} = [U_1^{\nabla}, ..., U_d^{\nabla}]$. Now consider the linear operator θ where

$$\theta Z = \begin{bmatrix} \Pi \\ A^T \end{bmatrix} Z^{\Delta}$$

where for Z in $\mathbb{R}^L$, Z^{Δ} denotes the vector consisting of the last $L-1$ entries, i.e.

$$Z^{\Delta} = \begin{bmatrix} Z_2 \\ \vdots \\ Z_L \end{bmatrix}.$$

Then

$$\hat{X}_{.k} = \begin{cases} \tilde{X}_{.k} & \text{for k = 1,...,K} \\ \theta \hat{X}_{.k-1} & \text{for k = K+1,...,K+h+L-1} \end{cases}$$

where $\tilde{X}_{.k}$'s are the columns of the approximation of the trajectory matrix (after grouping and eliminating noise components). Now, by constructing matrix $\hat{X}_{.1}, \ldots,$ $\hat{X}_{.K+k+L-1}$ and performing diagonal averaging, we obtain a new series $\tilde{y}_1, ..., \tilde{y}_{N+h+L-1}$, where $\hat{y}_{N+1}, ..., \hat{y}_{N+h}$ form the h terms of the SSA vector forecast.

4.2 Theory for Decomposition and Forecast

In this section we provide some theory which explains the ideas behind the algorithms in the Univariate SSA in Section 4.1. The Multivariate SSA in Section 4.3 will also be based on these results. The proofs we gather in Subsection 4.2.1. Let X be a row-based trajectory matrix, $\lambda_1 \geq \ldots \geq \lambda_L \geq 0$ be the eigenvalues of XX^T, and $U_1, \ldots, U_L$ be the corresponding eigenvectors of XX^T. We have

$$\mathbf{X} = \begin{bmatrix} y_1 & y_2 & y_3 & \cdots & y_K \\ y_2 & y_3 & y_4 & \cdots & y_{K+1} \\ y_3 & y_4 & y_5 & \cdots & y_{K+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_L & y_{L+1} & y_{L+2} & \cdots & y_N \end{bmatrix}, \quad K = N - L + 1.$$

Proposition 4.1. *We have $\mathbf{X} = \mathbf{X}_1 + \ldots + \mathbf{X}_L$, and from the eigenvalue decomposition, we got $X_i = U_i U_i^T X$.*

Remark 4.2. *Note that U_i are the left eigenvectors of $\mathbf{X}$ and for the right eigenvectors of $\mathbf{X}$, $V_i = \frac{X^T U_i}{\sqrt{\lambda_i}}$. We have $X_i = \sqrt{\lambda_i} U_i V_i^T$.*

Lemma 4.3. *(i)* $X_{lk} = X_{l'k'} \iff l + k = l' + k'$ *for* $l, l' \in \{1, \ldots, L\}$, $k, k' \in \{1, \ldots, K\}$.

(ii) $y_n = X_{lk}$, *if* $n = l + k - 1$.

(iii) $\mathbf{X} = \mathbf{X}_1 + \ldots + \mathbf{X}_d$ *if* $d = rank(\mathbf{X})$.

As approximation of $\mathbf{X}$, we may think of $\tilde{\mathbf{X}}$ as signal part and $\mathbf{X}_{d+1} + \ldots + \mathbf{X}_L$ as noise part of $\mathbf{X}$.

Notation:

We denote

$$U_i = \begin{bmatrix} U_{i,1} \\ \vdots \\ U_{i,L} \end{bmatrix}, \quad U_i^{\nabla} = \begin{bmatrix} U_{i,1} \\ \vdots \\ U_{i,L-1} \end{bmatrix}, \quad i = 1, \ldots, L,$$

$$X_{\cdot k} = \begin{bmatrix} X_{1k} \\ \vdots \\ X_{Lk} \end{bmatrix}, \quad X_{\cdot k}^{\nabla} = \begin{bmatrix} X_{1k} \\ \vdots \\ X_{L-1,k} \end{bmatrix}, \quad X_{\cdot k}^{\Delta} = \begin{bmatrix} X_{2k} \\ \vdots \\ X_{Lk} \end{bmatrix}, \quad k = 1, \ldots, K$$

$$v^2 = \sum_{i=1}^{d} U_{i,L}^2$$

$$\alpha_j = \frac{1}{1 - v^2} \sum_{i=1}^{d} U_{i,L} U_{i,L-j}$$

$$A := \begin{bmatrix} \alpha_{L-1} \\ \vdots \\ \alpha_1 \end{bmatrix} = \frac{1}{1 - v^2} \sum_{i=1}^{d} U_{i,L} U_i^{\nabla}$$

Theorem 4.4. *We have*

$$\tilde{X}_{lk} = \sum_{i=1}^{d} U_{il} \sum_{m=1}^{L} U_{im} X_{mk} = v^2 X_{Lk} + (1-v^2) A^T X_{\cdot k}^{\nabla}, \quad k = 1, \ldots, K \tag{4.4}$$

Corollary 4.5. *If $X \approx \tilde{\mathbf{X}}$, then $\mathbf{X}_{Lk} \approx A^T X_{\cdot k}^{\nabla}$, $l = 1, \ldots, L, k = 1, \ldots, K$ and $\tilde{X}_{Lk} \approx A^T \tilde{X}_{\cdot k}^{\nabla}$ for $k = 1, \ldots, K$.*

Corollary 4.6. *We have for $l \leq L-1, k = 1, \ldots, K$:*

$$\tilde{X}_{lk} = \sum_{m=1}^{L-1} \sum_{i=1}^{d} U_{il} U_{im} X_{m+1,k-1} + \sum_{i=1}^{d} U_{iL} U_{iL} X_{Lk}$$

If $X_{Lk} \approx \tilde{X}_{Lk}$, then for $l \leq L-1$

$$\tilde{X}_{lk} = \sum_{m=1}^{L-1} \left(\sum_{i=1}^{d} U_{il} U_{im} \right) X_{m+1,k-1} + \sum_{i=1}^{d} U_{iL} U_{iL} A^T \underbrace{X_{\cdot k}^{\nabla}}_{X_{\cdot,k-1}^{\Delta}}$$

Denoting $V^{\nabla} = \left(U_1^{\nabla}, \ldots, U_d^{\nabla}\right)$ this yields

$$\tilde{X}_{\cdot k}^{\nabla} = \left(V^{\nabla}(V^{\nabla})^T + (1-v^2) A A^T\right) X_{\cdot,k-1}^{\Delta}$$

Further,

$$\tilde{X}_{Lk} = A^T X_{\cdot,k-1}^{\Delta}$$

Remark 4.7. *Recurrent and Vector Forecast*

a) When we first obtain $\tilde{y}$ by diagonal averaging from $\tilde{X}$, then by Corollary 4.5 and using that $y_n = X_{L,n-L+1}$, $y_{n-j} = X_{L,n-L+1-j}$, i.e.

$$X_{\cdot,n-L+1}^{\nabla} \approx \begin{bmatrix} y_{n-L+1} \\ \vdots \\ y_{n-1} \end{bmatrix},$$

we get

$$y_n \approx A^T \begin{bmatrix} y_{n-L+1} \\ \vdots \\ y_{n-1} \end{bmatrix}$$

i.e.

$$y_n = \sum_{j=1}^{L-1} \alpha_j y_{n-j}$$

So we use as estimates

$$\hat{y}_n = \begin{cases} \tilde{y}_n & n \leq N \\ \sum_{j=1}^{L-1} \alpha_j \hat{y}_{n-j} & n > N \end{cases}$$

We do the latter for $n = N+1, \ldots, N+h$

b) When we first forecast, then do diagonal averaging, we use forecast for columns when finding $\tilde{X}$. By Corollary 4.6, for $\tilde{X} \approx X$ we have

$$\tilde{X}_{\cdot k} \approx \begin{bmatrix} \Pi \\ A^T \end{bmatrix} \tilde{X}^{\Delta}_{\cdot,k-1} \text{ for } \Pi = \boldsymbol{V}^{\nabla}(\boldsymbol{V}^{\nabla})^T + (1-v^2)AA^T$$

and then we will use

$$\hat{X}_{\cdot k} = \begin{cases} \tilde{X}_{\cdot k} & k \leq K \\ \begin{bmatrix} \Pi \\ A^T \end{bmatrix} \hat{X}^{\Delta}_{\cdot,k-1} & k > K \end{cases}$$

Doing the forecast for $k = K+1, \ldots, K+h+L-1$ we obtain $\hat{y}_n$ by diagonal averaging for $n = N+1, \ldots, N+h$

4.2.1 Proofs

Proposition 4.1 and Remark 4.2 follow by the theory of eigenvalue decompositions.

Proof of Lemma 4.3

(i) Follows by construction.

(ii) Follows by construction.

(iii) Follows by $\lambda_{d+1} = \ldots = \lambda_L = 0$, $\lambda_d > 0$ if $d = \text{rank}(X)$.

We use for $d < L$ the approximation

$$\tilde{X} = X_1 + \ldots + X_d$$

when d should fulfill $\sum_{i=1}^{d} U_{i,L}^2 < 1$.

Proof of Theorem 4.4

Note that $\tilde{X}_{lk} = \sum_{i=1}^{d} \left(U_i U_i^T X\right)_{lk} = \sum_{i=1}^{d} U_{il} \sum_{m=1}^{L} U_{im} X_{mk}$

For $l = L$ this yields

$$\begin{aligned}
\tilde{X}_{Lk} &= \sum_{i=1}^{d} U_{il} \sum_{m=1}^{L} U_{im} X_{mk} \\
&= \underbrace{\sum_{i=1}^{d} U_{iL}^2}_{v^2} X_{Lk} + \sum_{i=1}^{d} U_{iL} \sum_{m=1}^{L-1} U_{im} X_{mk} \\
&= v^2 X_{Lk} + \sum_{m=1}^{L-1} \underbrace{\sum_{i=1}^{d} U_{iL} U_{im}}_{(1-v^2)\alpha_{L-m}} X_{mk} \\
&= v^2 X_{Lk} + (1-v^2) \sum_{m=1}^{L-1} \alpha_{L-m} X_{mk} \\
&= v^2 X_{Lk} + (1-v^2) A^T X^{\nabla}_{\cdot k}.
\end{aligned}$$

Proof of Corollary 4.5

If $X \approx \tilde{X}$, then by Theorem 4.4, replacing $\tilde{X}_{Lk}$ by X_{Lk}, we get $X_{Lk} = v^2 X_{Lk} + (1-v^2) A^T X_{\cdot k}^{\nabla}$ Solving for X_{Lk} yields

$$X_{Lk} = \frac{1}{1-v^2}(1-v^2) A^T X_{\cdot k}^{\nabla} = A^T X_{\cdot k}^{\nabla}.$$

Replacing X by the approximation $\tilde{X}$, we also get $\tilde{X}_{Lk} \approx A^T X_{\cdot k}^{\nabla}$.

Proof of Corollary 4.6

From Equation (4.4) in Theorem 4.4, we get for $l \leq L-1$

$$\begin{aligned}
\tilde{X}_{lk} &= \sum_{i=1}^{d} U_{il} \sum_{m=1}^{L} U_{im} X_{mk} \\
&= \sum_{m=1}^{L} \left(\sum_{i=1}^{d} U_{il} U_{im} \right) X_{mk} \\
&= \sum_{m=1}^{L-1} \left(\sum_{i=1}^{d} U_{il} U_{im} \right) \underbrace{X_{mk}}_{X_{m+1,k+1}} + \sum_{i=1}^{d} U_{il} U_{iL} X_{Lk}
\end{aligned}$$

Now, if $X_{Lk} \approx \tilde{X}_{Lk}$ by Theorem 4.4. we can approximate X_{Lk} by $A^T X_{\cdot k}^{\nabla}$ and get

$$\begin{aligned}
\tilde{X}_{lk} &\approx \sum_{m=1}^{L-1} \left(\sum_{i=1}^{d} U_{il} U_{im} \right) X_{m+1,k+1} + \underbrace{\sum_{i=1}^{d} U_{il} U_{iL}}_{(1-v^2)\alpha_{L-l}} A^T X_{\cdot k}^{\nabla} \\
&= (V^{\nabla}(V^{\nabla})^T X_{\cdot,k-1}^{\Delta})_l + (1-v^2)\alpha_{L-l} A^T \underbrace{X_{\cdot k}^{\nabla}}_{X_{\cdot,k-1}^{\Delta}},
\end{aligned}$$

since

$$V^{\nabla}(V^{\nabla})^T = \begin{bmatrix} U_{1,1} & \dots & U_{d,1} \\ \vdots & \ddots & \vdots \\ U_{1,L-1} & \dots & U_{r,L-1} \end{bmatrix} \begin{bmatrix} U_{1,1} & \dots & U_{1,L-1} \\ \vdots & \ddots & \vdots \\ U_{d,1} & \dots & U_{r,L-1} \end{bmatrix} = \left(\sum_{i=1}^{d} U_{il} U_{im} \right)_{l,m=1,\dots,L-1},$$

and thus as vector

$$\begin{aligned}
\tilde{X}_{\cdot k}^{\nabla} \approx \begin{bmatrix} X_{1,k} \\ \vdots \\ X_{L-1,k} \end{bmatrix} &= V^{\nabla}(V^{\nabla})^T X_{\cdot,k-1}^{\Delta} + (1-v^2) A A^T X_{\cdot,k-1}^{\Delta} \\
&= \left(V^{\nabla}(V^{\nabla})^T + (1-v^2) A A^T \right) X_{\cdot,k-1}^{\Delta}
\end{aligned}$$

From Theorem 4.4 we further obtain for $\tilde{X}_{Lk} \approx X_{Lk}$

$$\tilde{X}_{Lk} \approx A^T X_{\cdot,k}^{\nabla} = A^T X_{\cdot,k-1}^{\Delta}.$$

4.3 Multivariate Singular Spectrum Analysis

Using more or less the same steps as for the Univariate SSA, multivariate singular spectrum analysis (MSSA) is also applied in this thesis. MSSA is able to generate the forecast of a lot of time series at once by taking other series into consideration. Let $\{y_t : t = 1, \dots, N\}, y_t = \begin{bmatrix} y_t^{(1)} \\ \vdots \\ y_t^{(m)} \end{bmatrix}$, denote a sample of an M-variate time series with length N.

a) Stage 1: Decomposition

(i) Embedding
Just like in Univariate SSA, we transform the time series data into a trajectory matrix $\mathbf{X}$. Denote by $\mathbf{X}^{(m)}, m = 1, ..., M$, the Hankel matrix which is associated to the m^{th} time series, $y_1^{(m)}, ..., y_N^{(m)}$. A row-based trajectory matrix with L and $K = N - L + 1$ for multivariate SSA will look like this:

$$\mathbf{X}^{(m)} = \begin{bmatrix} y_1^{(m)} & y_2^{(m)} & y_3^{(m)} & \dots & y_K^{(m)} \\ y_2^{(m)} & y_3^{(m)} & y_4^{(m)} & \dots & y_{K+1}^{(m)} \\ y_3^{(m)} & y_4^{(m)} & y_5^{(m)} & \dots & y_{K+2}^{(m)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_L^{(m)} & y_{L+1}^{(m)} & y_{L+2}^{(m)} & \dots & y_N^{(m)} \end{bmatrix}$$

A column-based trajectory matrix for multivariate SSA will look like this:

$$\left(\mathbf{X}^{(m)}\right)^T = \begin{bmatrix} y_1^{(m)} & y_2^{(m)} & y_3^{(m)} & \dots & y_L^{(m)} \\ y_2^{(m)} & y_3^{(m)} & y_4^{(m)} & \dots & y_{L+1}^{(m)} \\ y_3^{(m)} & y_4^{(m)} & y_5^{(m)} & \dots & y_{L+2}^{(m)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_K^{(m)} & y_{K+1}^{(m)} & y_{K+2}^{(m)} & \dots & y_N^{(m)} \end{bmatrix}$$

After creating the trajectory matrix of each series, they need to be stacked with each other. There are two options for stacking: Row (Vertical) and Column (Horizontal) stacking, which would look like

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \vdots \\ \mathbf{X}^{(M)} \end{bmatrix} \quad \text{or} \quad \mathbf{X} = \begin{bmatrix} \mathbf{X}^{(1)} & \dots & \mathbf{X}^{(M)} \end{bmatrix},$$

respectively. The same can be done for the signal component $\mathbf{S}$ (the approximation $\tilde{\mathbf{X}}$ of $\mathbf{X}$) and noise component $\mathbf{N}$ (the remaining part).

(ii) Singular Value Decomposition (SVD)
In this step, $\mathbf{X}$ will be decomposed by singular value decomposition as $\mathbf{X} = \mathbf{X}_1 + ... + \mathbf{X}_r$ where $\mathbf{X}_i$'s are unitary matrices and r represents the rank of $\mathbf{X}$. Denoting by $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r > 0$ the eigenvalues of $\mathbf{X}\mathbf{X}^T$ and $U_1, U_2, ..., U_d$, the corresponding eigenvectors, we have first r:

$$\mathbf{X}_j = U_j U_j^T \mathbf{X}, \quad j = 1, 2, ..., r.$$

b) Stage 2: Reconstruction

(i) Grouping
Considering $\mathbf{X}_i$ to be associated to the i-th largest singular value of $\mathbf{X}$, this step intends to separate the signal and noise components as follows:

$$\mathbf{X} = \underbrace{\mathbf{X}_1 + ... + \mathbf{X}_d}_{\hat{\mathbf{S}}=\text{Signal}} + \underbrace{\mathbf{X}_{d+1} + ... + \mathbf{X}_r}_{\hat{\mathbf{N}}=\text{Noise}}$$

(ii) Diagonal Averaging
In this step, using anti-diagonal averaging on each block of $\hat{\mathbf{S}}$, the denoised time series will be reconstructed. We use notation $\tilde{\mathbf{X}} = \hat{\mathbf{S}}$ for the approximation to show the results of this step.

c) Stage 3: Forecasting
Just like Univariate SSA, there are also two forecast methods in Multivariate SSA: Recurrent and Vector. And since there are various types of forms of trajectory matrix and stacking in MSSA, there will also be more formulas to be considered for both Recurrent and Vector forecast methods. With two forms of the trajectory matrix, two forms of matrix stack, and two forms of forecast methods, there will be eight possible results in MSSA as written in Table 38.

Table 38: Eight Possible Combinations of MSSA

Trajectory Form	**Trajectory Stack**	**Forecasting**	**Abbreviaton**
Column	Column	Recurrent	Col-M/Col-S/Rec
Column	Row	Recurrent	Col-M/Rowl-S/Rec
Column	Column	Vector	Col-M/Col-S/Vec
Column	Row	Vector	Col-M/Row-S/Vec
Row	Column	Recurrent	Row-M/Col-S/Rec
Row	Row	Recurrent	Row-M/Row-S/Rec
Row	Column	Vector	Row-M/Col-S/Vec
Row	Row	Vector	Row-M/Row-S/Vec

Column Stack
Let $U_j, j = 1, ..., r$ be the j-th eigenvector of $\mathbf{X}\mathbf{X}^T$. Note that $U_j \in \mathbb{R}^L$. Denote by $\mathbf{V}$ the matrix of its first d eigenvectors, corresponding to the d largest singular values of $\mathbf{X}$ and as for Univariate SSA V^{∇} the first $L-1$ rows of $\mathbf{V}$ and $v^2 = \sum_{i=1}^{d} U_{iL}^2$. We set $\tilde{\mathbf{X}} = \tilde{\mathbf{S}} = \sum_{i=1}^{d} U_i U_i^T \mathbf{X}$. $\tilde{\mathbf{X}}$ is of dimension $L \times KM$,

$$\tilde{\mathbf{X}} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,K} & x_{1,K+1} & \dots & x_{1,2K} & x_{1,2K+1} & \dots & x_{1,MK-1} & x_{1,MK} \\ x_{2,1} & x_{2,2} & \dots & x_{2,K} & x_{2,K+1} & \dots & x_{2,2K} & x_{2,2K+1} & \dots & x_{2,MK-1} & x_{2,MK} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{L,1} & x_{L,2} & \dots & x_{L,K} & x_{L,K+1} & \dots & x_{L,2K} & x_{L,2K+1} & \dots & x_{L,MK-1} & x_{L,MK} \end{bmatrix}$$

The gray color shows the response, while the other rows are the regressors. This motivates the calculation for the Recurrent method, which now works the same as the Univariate SSA. This leads to the analogous estimates as above via

Column Stack + Recurrent Forecast

$$\mathbf{W} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ 0 & A^T \end{bmatrix}, \quad A = \frac{1}{1-v^2} \sum_{i=1}^{d} U_{iL} U_i^{\nabla},$$

where $\mathbf{I}$ is the $(L-1) \times (L-1)$ identity matrix and $\mathbf{0}$ is a column vector with $L-1$ zeros. Then, the h-steps ahead forecasts can be obtained by:

$$\hat{y}_{T+h}^{(m)} = (\mathbf{W}^h)_{L,\cdot} \tilde{X}_{\cdot,mK}, m = 1, \ldots, M, h = 1, 2, \ldots.$$

The coefficients $\mathbf{W}_{L,\cdot}^h$ are generated by the whole system of time series with consideration of correlations among time series. In addition, $\tilde{X}_{\cdot,mK}$ is smoothed again based on the information of all time series. It should be noticed, however, that the forecasts for all individual time series are made by using the same coefficients. This works analogously to the recurrent forecast in the Univariate SSA, we only use here the matrix W for a consistent notation with the forecast for the Row Stack below. Note that we could also diagonalize first, changing the notation slightly.

Column Stack + Vector Forecast
Considering the same notation as before, we define:

$$\mathbf{W} = \begin{bmatrix} \mathbf{0} & \mathbf{\Pi} \\ 0 & A^T \end{bmatrix}, \quad \mathbf{\Pi} = \mathbf{V}^{\nabla} \left(\mathbf{V}^{\nabla}\right)^T + (1-v^2) A A^T$$

where $\mathbf{0}$ is the column vector with $L-1$ zeros. Then, the h-steps ahead forecasts can be obtained by:

$$\hat{y}_{T+h}^{(m)} = \frac{1}{L} \sum_{l=h}^{h+L-1} \left(\mathbf{W}^l \tilde{X}_{\cdot,mK}\right)_{L-l+k}, \quad m = 1, \ldots, M, h = 1, 2, \ldots$$

To better understand how the Recurrent and Vector forecast differ, we compare the equations of $\hat{y}_{T+h}^{(m)}$. For Recurrent, we multiply by the coefficient $\left(\mathbf{W}^h\right)_{L,\cdot}$ to produce the forecast of the next $y^{(m)}$. However, $\tilde{\mathbf{X}}$ in the Vector method is multiplied by the coefficients $\left(\mathbf{W}^l\right)_{L-l+k,\cdot}$, and the forecast is produced by averaging. As for Univariate SSA this is forecasting the columns, but always stacking for $y^{(m)}$ from the last column corresponding to m.

Row Stack
Denote by $\mathbf{U}_1, \ldots, \mathbf{U}_d$ the first d eigenvectors of $\mathbf{X}\mathbf{X}^T$ corresponding to the d largest singular values of $\mathbf{X}$ and again $\tilde{\mathbf{X}} = \sum_{i=1}^{d} U_i U_i^T \mathbf{X}$. This matrix has dimension $ML \times K$,

$$\tilde{\mathbf{X}} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,K} \\ x_{2,1} & x_{2,2} & \dots & x_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{L,1} & x_{L,2} & \dots & x_{L,r} \\ x_{L+1,1} & x_{L+1,2} & \dots & x_{L+1,K} \\ x_{L+2,1} & x_{L+2,2} & \dots & x_{L+2,K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{2L,1} & x_{2L,2} & \dots & x_{2L,K} \\ x_{2L+1,1} & x_{2L+1,2} & \dots & x_{2L+1,K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{(M-1)L+1,1} & x_{(M-1)L+1,2} & \dots & x_{(M-1)L+1,K} \\ x_{(M-1)L+2,1} & x_{(M-1)L+2,2} & \dots & x_{(M-1)L+2,K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{ML-1,1} & x_{ML-1,2} & \dots & x_{ML-1,K} \\ x_{ML,1} & x_{ML,2} & \dots & x_{ML,K} \end{bmatrix}$$

Like for Column Stack, the gray color shows response variables for the recurrent forecast, while the remaining fows are the regressors.

Row Stack + Recurrent Forecast

Note that now the eigenvectors U are of dimension LM. Aside from the dimension, we can use the same notation, i.e. let $V = (U_1, \dots, U_d)$. But now assume that $\mathbf{V}^{\nabla}$ is constructed by removing the rows $L, 2L, ..., ML$ from $\mathbf{V}$, and $\mathbf{V}_{\nabla}$ is the matrix that is constructed by stacking the rows $L, 2L, ..., ML$ of $\mathbf{V}$, we can define:

$$\mathbf{W} = \begin{Bmatrix} \mathbf{0} & \mathbf{I} \\ 0 & \mathcal{A}_{1,\cdot} \\ \mathbf{0} & \mathbf{I} \\ 0 & \mathcal{A}_{2,\cdot} \\ \vdots & \vdots \\ \mathbf{0} & \mathbf{I} \\ 0 & \mathcal{A}_{M,\cdot} \end{Bmatrix}, \quad \mathcal{A} = (I_{M\times M} - \mathbf{V}_{\nabla}\mathbf{V}_{\nabla}^T)^{-1}\mathbf{V}_{\nabla}\mathbf{V}_{\nabla}^T$$

where $\mathbf{I}$ is the $(L-1)\times(L-1)$ identity matrix, $\mathbf{0}$ is a column vector with $L-1$ zeros and $[0, \mathcal{A}_{L,\cdot}]$ is a vector of size ML. Then, the h-steps ahead forecasts can be obtained by:

$$\hat{y}_{T+h}^{(m)} = \left(\mathbf{W}^h\right)_{mL,\cdot} \tilde{\mathbf{X}}_{\cdot,K}, \quad m = 1, \dots, M, h = 1, 2, \dots$$

Row Stack + Vector Forecast
Considering the same notation as before, we can define:

$$\mathbf{W} = \begin{Bmatrix} \mathbf{0} & \mathbf{\Pi}_1 \\ 0 & \mathcal{A}_{1,\cdot} \\ \mathbf{0} & \mathbf{\Pi}_2 \\ 0 & \mathcal{A}_{2,\cdot} \\ \vdots & \vdots \\ \mathbf{0} & \mathbf{\Pi}_M \\ 0 & \mathcal{A}_{M,\cdot} \end{Bmatrix} \quad , \quad \mathbf{\Pi} = \mathbf{V}^{\nabla} \left(\mathbf{V}^{\nabla}\right)^T + \mathcal{A}^T (I_{M\times M} - \mathbf{V}^{\nabla} \left(\mathbf{V}^{\nabla}\right)^T)\mathcal{A},$$

where $\mathbf{0}$ is a column vector with $L-1$ zeros and $\mathbf{\Pi}_j$ represents the rows number $(j-1)(L-1)+1, ..., j(L-1)$ of $\mathbf{\Pi}, j = 1, ..., M$. Then, the h-steps ahead forecasts can be obtained by:

$$\hat{y}_{T+h}^{(m)} = \frac{1}{L} \sum_{l=h}^{h+L-1} \left(\mathbf{W}^l\right)_{mL-l+h} \tilde{\mathbf{X}}_{\cdot,mK}, \quad m = 1, \ldots, M, h = 1, 2, \ldots \quad .$$

4.4 Parameter Selection

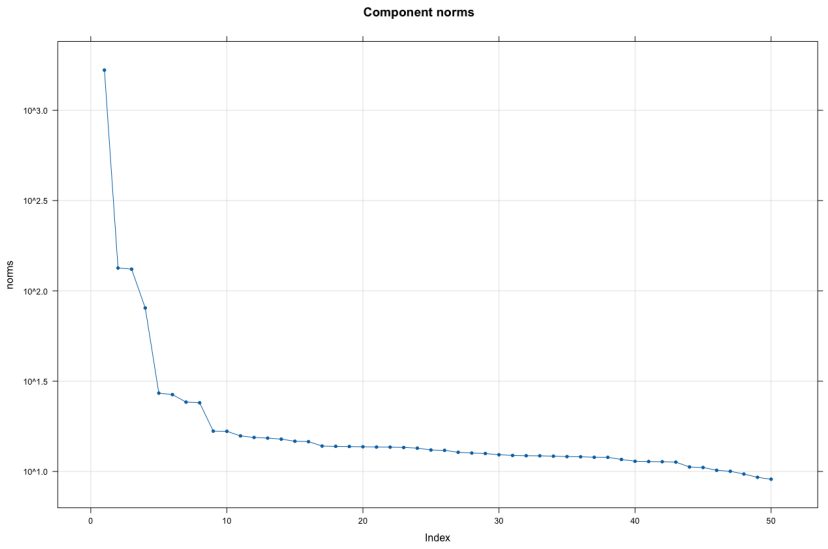
Even though Singular Spectrum Analysis is a nonparametric method, there are some important values that need to be determined in order to provide the best results. These "parameters" are window length L and the number of eigentriples for grouping. For the entire calculation, it is decided to use $L = 24$ and $d = 6$. However, we would also like to see how different the results could be if we use another value for window length and eigentriples. Therefore, we also use $L = 84$ and $d = 48$. We chose two types of L based on the seasonality patterns (2 years and 7 years), and we choose the d accordingly, but also with seasonality patterns (6 months and 4 years).

In the Embedding step, both Univariate and Multivariate SSA need to create a trajectory matrix with a certain window length L. L is the only parameter in the decomposition stage. Selecting the proper window length can be based on the information of the time series itself or by some other means. Theoretically, L should be large enough, but not more than $N/2$. As for the information itself, we can include the seasonal component into account. For instance, if the series has an annual pattern, then we can use $L = 12$. Or, if the pattern is biannually, then we can use $L = 24$. The number of window length will determine the number of eigentriples being used as well. Some supplementary information to determine the number of window lengths are:

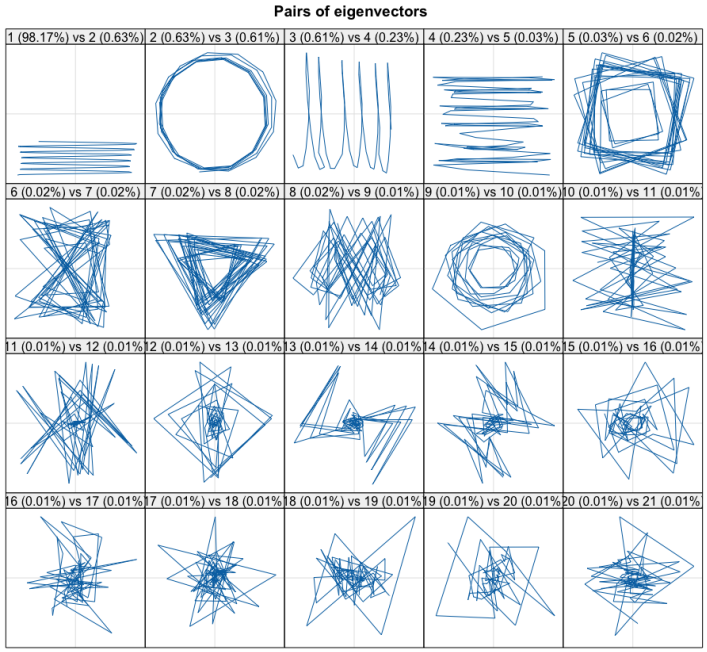
a) Auxiliary Information
This kind of information can always help the process to provide more accurate results. This information can help us select the proper group and choose the proper window length, which will lead to proper forecasting. The seasonality of the time series is important to be used as a base to make assumptions.

b) Singular Values
Checking the breaks in the eigenvalue spectra can prove useful. When the values are similar in height, these pairs are related to specific periods. Although we might not know which period for now, but we know that there are some eigentriple pairs that can prove the seasonal pattern of the time series. As an example, below we have the eigenvalue spectra. Some obvious pairs are 2-3, 5-6, 7-8, and 9-10. However, there might be more if we use

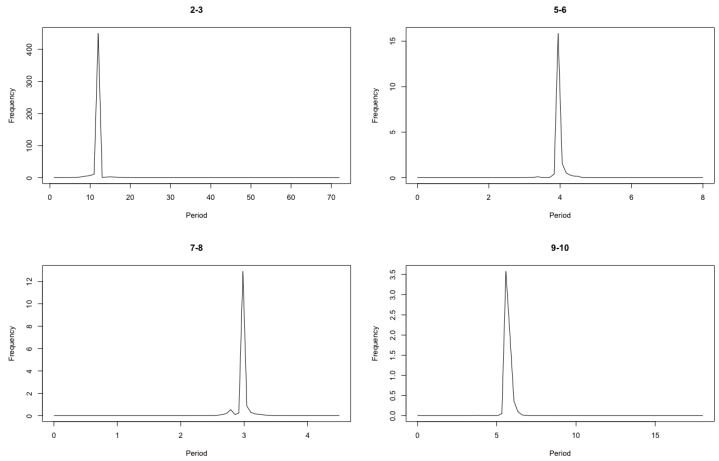
other methods. Anyway this also helps to decide on d, i.e. on the number of eigenvalues used to explain the signal part.



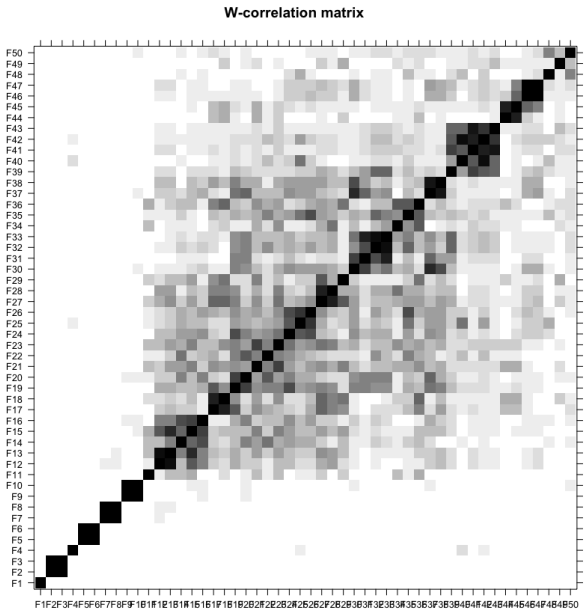
c) Pairwise Scatterplots
This method is the easiest way to determine the eigentriple pairs and the seasonal pattern type. By seeing the patterns formed by the pairs of eigenvectors, we can determine the seasonal pattern of a time series. According to the example below, Pair 2-3 has a dodecagonal form (12 sides), showing that the series has an annual pattern. Pair 5-6 has a square form (4 sides), showing that the series has a pattern for every 4 months. Pair 7-8 has a triangle form, showing a pattern every 3 months. Pair 9-10 shows a semesterly pattern through its hexagonal form. There are two more patterns that can be analyzed, but not in these plots. They are a dodecagran (star with 12 vertices) which resembles a seasonal pattern of 2.4 months and a pentagram (star with 5 vertices) which resembles a seasonal pattern of 2.5 months.



d) Periodogram Analysis
Periodogram can be helpful because it shows us the frequencies that need to be considered. Based on this frequency, we have to look for the eigentriples that belong to this frequency. The results of this method usually align with the results from pairwise scatterplots. Using the four pairs from before, we see that Pair 2-3 align in the period of 12, Pair 5-6 in the period of 4, Pair 7-8 in the period of 3, and Pair 9-10 in the period of 6.



e) Separability
Separability shows how well different components can be separated from each other. We can analyze this through w-correlation. If the absolute value of the w-correlation is small or closer to 0, it means the eigentriples are separable. If the absolute value of the w-correlation is large or closer to 1, then the eigentriples are badly separable. The results from this method usually align with the results from pairwise scatterplots as well. Those who form a specific pattern will have high correlation and therefore, are badly separable. According to the plot below, 2-3, 5-6, 7-8, 9-10, 12-13, 27-28, 32-33, 37-38, 41-42, 44-45, and 46-47 seem to be badly separable. According to the former analysis based on pairwise scatterplots and periodogram, we do have Pair 2-3, 5-6, 7-8, and 9-10.



f) Error Values
Aside from using information obtained through the time series, it is also possible to calculate everything using various window length and eigentriples through iteration. The error values obtained from these forecast will help to determine the best window length to be used, along with the good number of eigentriples to be used as well.

4.4.1 Choosing L as Number of Row or Column

In the SSA approaches, according to [GZ13], L must be relatively small but still sufficiently large ($L \leq N/2$), and K must be very large ($K \to \infty, K = N - L + 1$). The reason why L needs to be sufficiently large is because we want to capture enough pattern of the whole series through L. With a big enough number of L, we might be able to see the important parts and analyze the characteristics of the data. If L is too small, these data chunks might get mixed up and possibly confusing. When L is big enough, the results will usually stay stable and small changes will not mess things up that much. The best possible approach when it comes to choosing L for SSA would be repeating the process several times with different values of L.

In this thesis, we look at both settings, i.e. using L as the number of rows and as the number of columns. When L is the number of rows, the dimension of the trajectory matrix becomes $L \times K$ and when L is the number of columns, the dimension becomes $K \times L$. L as the number of rows is preferred because it preserves the time structure well by overlapping windows of the time series. The matrix shape with L as the number of rows is also good for all the procedures needed in SSA. However, that doesn't necessarily mean that L as the number of column is bad. It can still be processed and it still has meanings. We only need to remember when interpreting that the time structure is put the opposite way. According to [Gol20], trajectory matrices between the two approaches can be considered as transposed.

Even though L as the number of trajectory matrix rows fits better for the entire process of SSA, the position itself actually does not matter as long as the interpretation follows the structure of the matrix as well. If L is the number of rows, U becomes the left singular vector with patterns along the window length (how values change within a small window), while V becomes the right singular vector with patterns along the time direction (how the windows move over time). If L is the number of column, then the meanings change: U patterns the time change while V patterns the value change.

Whether L is used as the number of rows or columns, thus will affect MSSA that includes matrix stacking before calculation. L as the number of rows followed by column/horizontal stacking is preferrable and the most common because it makes interpretation easier. With $L \times K$ matrix and column stack (side-by-side), the rows of the matrix show the past (lagged) values, while the columns show time windows across all series. However, if we use row stack (on top of each other) with $L \times K$ matrix, the rows show the past (lagged) values of each series, while the columns show the time windows. As for $K \times L$ matrix and column stack, the rows show time windows, while the columns show lags across series, but flipped view. Lastly, if we use $K \times L$ matrix with row stack, we will get the time window across series through the rows and the lag positions through the columns.

4.5 SSA Combinations

In summary, there are 2 different forms of the trajectory matrix and two different methods for forecasting, so there will be 4 possible combinations for Univariate SSA. Meanwhile, adding 2 different forms of trajectory matrix stack will result in 8 possible combinations for Multivariate SSA as shown in Table 38 In total, there are 12 combinations and all of them will be used for the calculation and comparison. All these combinations are written in Table 39.

Table 39: All SSA Possible Combinations

SSA Type	Trajectory Form	Stack	Forecast Method	Abbreviation
Univariate	Column	-	Recurrent	USSA-Col-M/Rec
Univariate	Column	-	Vector	USSA-Col-M/Vec
Univariate	Row	-	Recurrent	USSA-Row-M/Rec
Univariate	Row	-	Vector	USSA-Row-M/Vec
Multivariate	Column	Column	Recurrent	MSSA-Col-M/Col-S/Rec
Multivariate	Column	Row	Recurrent	MSSA-Col-M/Row-S/Rec
Multivariate	Column	Column	Vector	MSSA-Col-M/Col-S/Vec
Multivariate	Column	Row	Vector	MSSA-Col-M/Row-S/Vec
Multivariate	Row	Column	Recurrent	MSSA-Row-M/Col-S/Rec
Multivariate	Row	Row	Recurrent	MSSA-Row-M/Row-S/Rec
Multivariate	Row	Column	Vector	MSSA-Row-M/Col-S/Vec
Multivariate	Row	Row	Vector	MSSA-Row-M/Row-S/Vec

4.6 RMSE and MAE

There will be two kinds of error values used to evaluate the goodness of each method combination for the reconstructed data (2000-2015) and the forecast (2016-2017). The first error value is Root of the Mean Square of Errors (RMSE). As per Equation (4.5), RMSE calculates the square of each error value, averages these square values, and calculates the square root of this average.

$$RMSE = \left(\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \right)^{\frac{1}{2}} \tag{4.5}$$

The second error value is Mean of Absolute value of Errors (MAE) which is written in Equation (4.6). MAE also calculates the difference, but instead of averaging the square of the error values, it averages the absolute error values.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |(y_i - \hat{y}_i)| \tag{4.6}$$

4.7 Error Values Comparison

In this section, we compare the error values for each missing data method, each area of calculation, and each SSA combination method in two different situations: $\{L = 24, d = 6\}$ and $\{L = 84, d = 48\}$. Our goal is to find combination methods with the lowest error possible. Since it is ineffective to check through all the error values one by one, we group them into several categories and calculate the average error values. These average values are used for the comparisons.

4.7.1 SSA Methods and Areas Comparison

Figure 45 shows the error values of SSA with $\{L = 24, d = 6\}$. The error values of Reconstructed is much lower than Forecast. Of course, MAE is also lower than RMSE. We can also see that USSA-Col-M/Rec has the lowest value almost in all error categories and compared to other SSA methods. Aside from this method, USSA-Row-M/Rec and USSA-Row-M/Vec also have low error values in Reconstructed, but they have high error values for Province in Forecast.

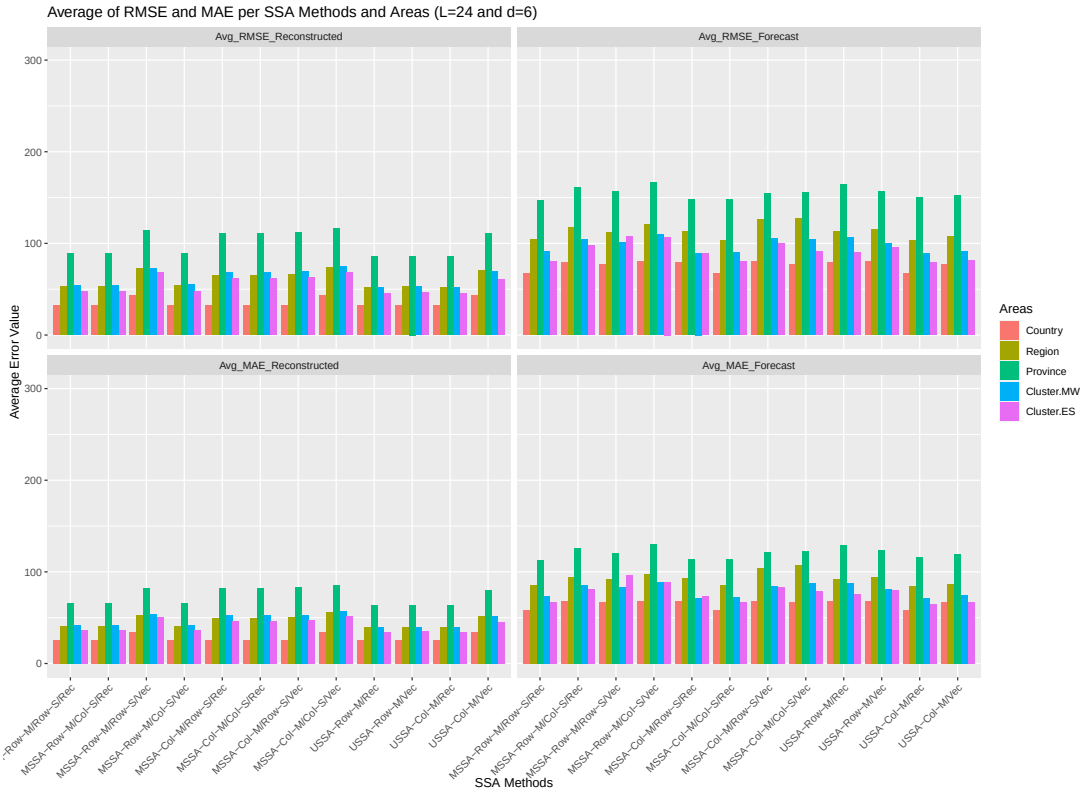


Figure 45: Average RMSE and MAE based on SSA Methods and Areas $\{L = 24, d = 6\}$

USSA-Col-M/Vec is the highest in Reconstructed, but almost the lowest in Forecast. As for MSSA, it looks like MSSA-Row-M/Row-S/Rec has the lowest error values, but almost similar to MSSA-Row-M/Col-S/Rec and MSSA-Row-M/Col-S/Vec as well. The area of calculation with the lowest error value goes to Country, followed by Cluster and Region which is balanced, and lastly, Province. In Forecast, however, Region has higher error values compared to Cluster, while in Reconstructed, they are more or less the same. Overall, Manhattan-Ward's Cluster also has higher error values than Euclidean-Single Cluster. With some SSA methods, the error values for Euclidean-Single Cluster are almost similar to Country.

As for SSA with $\{L = 84, d = 48\}$ as shown in Figure 46, the error values of Reconstructed are also much lower than Forecast. The observations are more or less the same as SSA with $\{L = 24, d = 6\}$ where the MAE values are lower than RMSE values, Country has the lowest error values while Province has the highest, and Region and Clusters seem to perform similarly. However, there are some slight differences in the SSA methods that perform better. Overall, the error values with this window length and number of eigentriples are higher than before. The error values of the SSA method combinations are more different compared to each other, unlike before. This shows that SSA with $\{L = 24, d = 6\}$ has more consistent results no matter which combination is used. With $\{L = 84, d = 48\}$, we have to be more careful upon

choosing the SSA method because some of them can result in really high error values. As for the SSA combination itself, USSA-Col-M/Vec seems to have the lowest error values in Forecast again, but the highest in Reconstructed. Meanwhile, USSA-Row-M/Rec has the lowest error values in Reconstructed, but the highest in Forecast. As for MSSA, the lowest error values in Reconstructed seem to be given by either MSSA-Row-S/Row-M/Rec or MSSA-Row-S/Col-M/Rec, but the lowest error values in Forecast are given by MSSA-Row-S/Row-M/Vec, followed by MSSA-Col-S/Row-M/Vec. Overall, in this comparison, MSSA performs better than USSA.

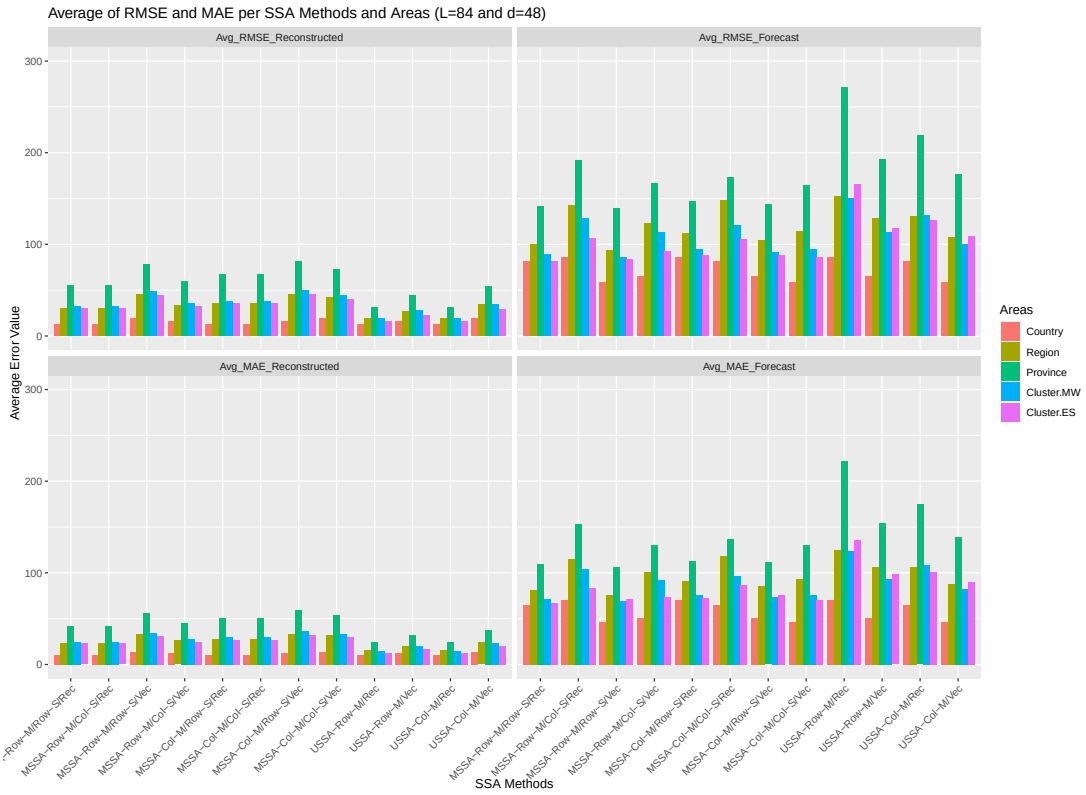


Figure 46: Average RMSE and MAE based on SSA Methods and Areas $\{L = 84, d = 48\}$

4.7.2 Missing Data Methods and Areas Comparison

The second comparison is averaged error values based on missing data methods and areas of calculation. Figure 47 shows this comparison for SSA with $\{L = 24, d = 6\}$. Just like the previous comparisons, Reconstructed error values are lower than Forecast, and MAE is also lower than RMSE. Country still has the lowest error values, while Province has the highest error values. Region and Clusters are mostly similar, although Region has slightly higher error values.

As for the methods, judging from the Reconstructed error values, it seems like all of the missing data methods perform rather similarly. Forecast error values show more extreme differences. The error values for Mean Substitution, LOCF Imputation, and Interpolation are high. Median Substitution also has rather high error values. Meanwhile, all the other methods seem to perform rather similarly. Bootstrap, all MICE methods, Distribution Fill, and Null Substitution have similar error values. Seeing this result, except for Null Substitution, it seems like missing data methods where the missing values are randomly imputed perform better than the ones that are imputed using fixed values e.g. Mean Substitution and Median Substitution.

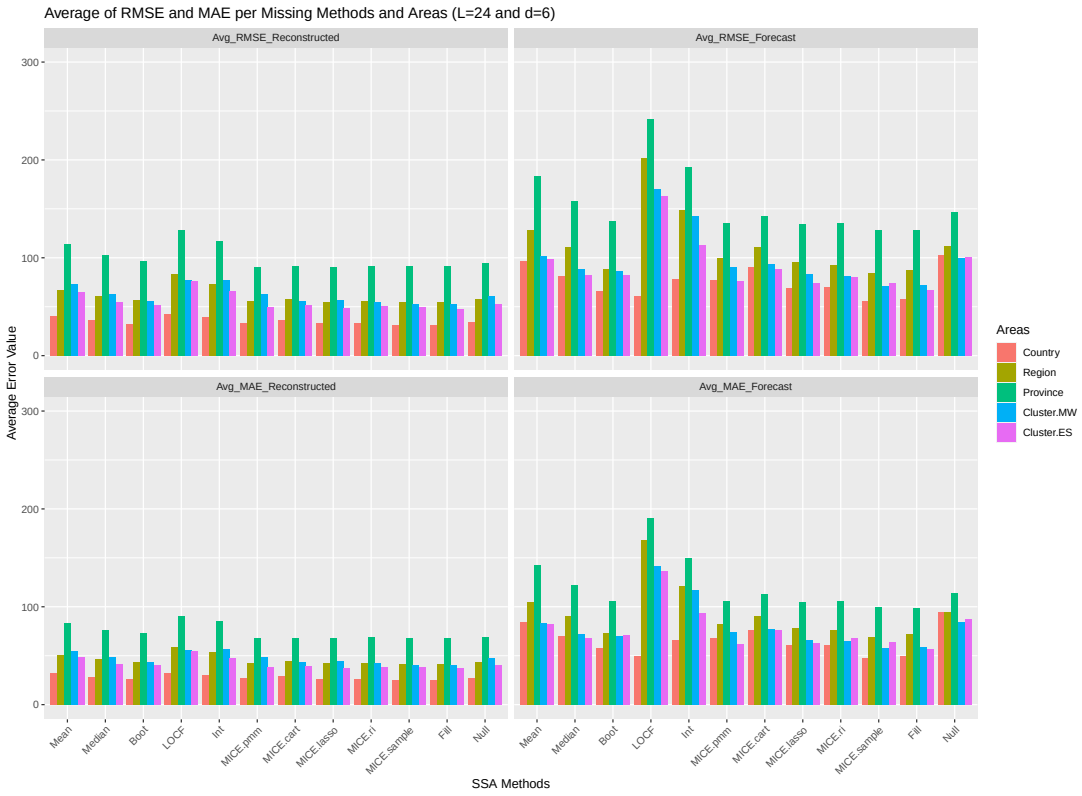


Figure 47: Average RMSE and MAE based on Missing Methods and Areas $\{L = 24, d = 6\}$

In Figure 48 now shows the comparison between missing data methods and areas for SSA with $\{L = 84, d = 48\}$. Again, we have much higher error values compared to SSA with $\{L = 24, d = 6\}$, showing that this pair of window length and eigentriples doesn't perform better in comparison. The error values for Reconstructed look almost similar for all missing data methods, but LOCF Imputation and Interpolation are slightly higher, followed by Mean Substitution. The differences are even clearer in Forecast, also with the same order: LOCF Imputation is the highest, followed by Interpolation, Mean Substitution, and Median Substitution, which is also similar with $\{L = 24, d = 6\}$. Reconstructed error values also show that Region and Clusters perform similarly, Province performs slightly worse, and Country performs the best. It is similar in Forecast. The only difference is that the gap between error values are more obvious. In some cases, Manhattan-Ward's Cluster performs better than Euclidean-Single Cluster, but in other cases, the opposite happens. Overall, aside from the higher error values in general, this comparison produces similar results as SSA with $\{L = 24, d = 6\}$.

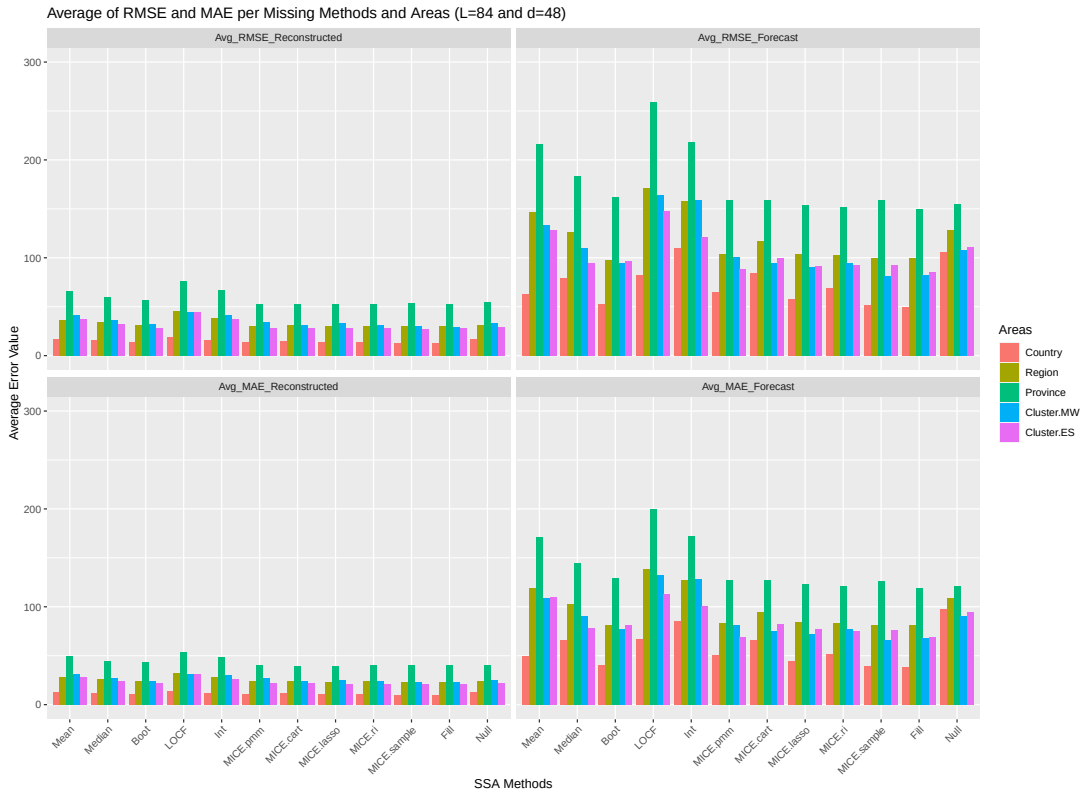


Figure 48: Average RMSE and MAE based on Missing Methods and Areas $\{L = 84, d = 48\}$

4.7.3 SSA Methods and Missing Data Methods Comparison

Figure 49 shows the third comparison: averaged error values based on SSA methods and missing data methods for SSA with $\{L = 24, d = 6\}$. We have consistent results with Reconstructed having lower error values than Forecast and MAE always having lower values than RMSE. All SSA methods seem to be performing similarly, especially in Reconstructed. Some method combinations are higher than the other, but the differences are not significant. On the other hand, Forecast shows different results. The differences are very obvious, especially for LOCF and Interpolation. The highest error for these two missing methods are found in MSSA-Col-M/Row-S/Vec and MSSA-Col-M/Col-S/Vec. There are also some high error values for the two missing data methods in MSSA-Row-M/Col-S/Rec, MSSA-Row-M/Col-S/Vec, and MSSA-Col-M/Row-S/Rec. Meanwhile, the other missing methods seem to be behaving similarly. Mean Substitution, Median Substitution, and Null Substitution has slightly higher error values, but the differences aren't as big as LOCF and Interpolation. This means missing data methods that constantly have low error values are Bootstrap and all MICE methods. As for the SSA methods itself, USSA-Col-M/Vec seems to perform the best in Forecast but the worst in Reconstructed, while MSA-Col-M/Col-S/Rec and MSSA-Row-M/Row-S/Rec seem to perform the best in Forecast. The best MSSA method in Reconstructed is MSSA-Row-M/Col-S/Vec.

We also look at Figure 50 where the third comparison is shown for SSA with $\{L = 84, d = 48\}$. Unlike the former two comparisons, we have lower error values this time. Reconstructed shows the differences very clearly with most error values are below 50. In Forecast, however, the difference isn't as significant, except for LOCF and Interpolation that performs better in general. There are still some high error values from the two missing methods, but they come from mostly USSA, which again doesn't perform better than MSSA. The error values of Mean Substitution, Median Substitution, and Null Substitution are also higher and more obvious in this comparison.

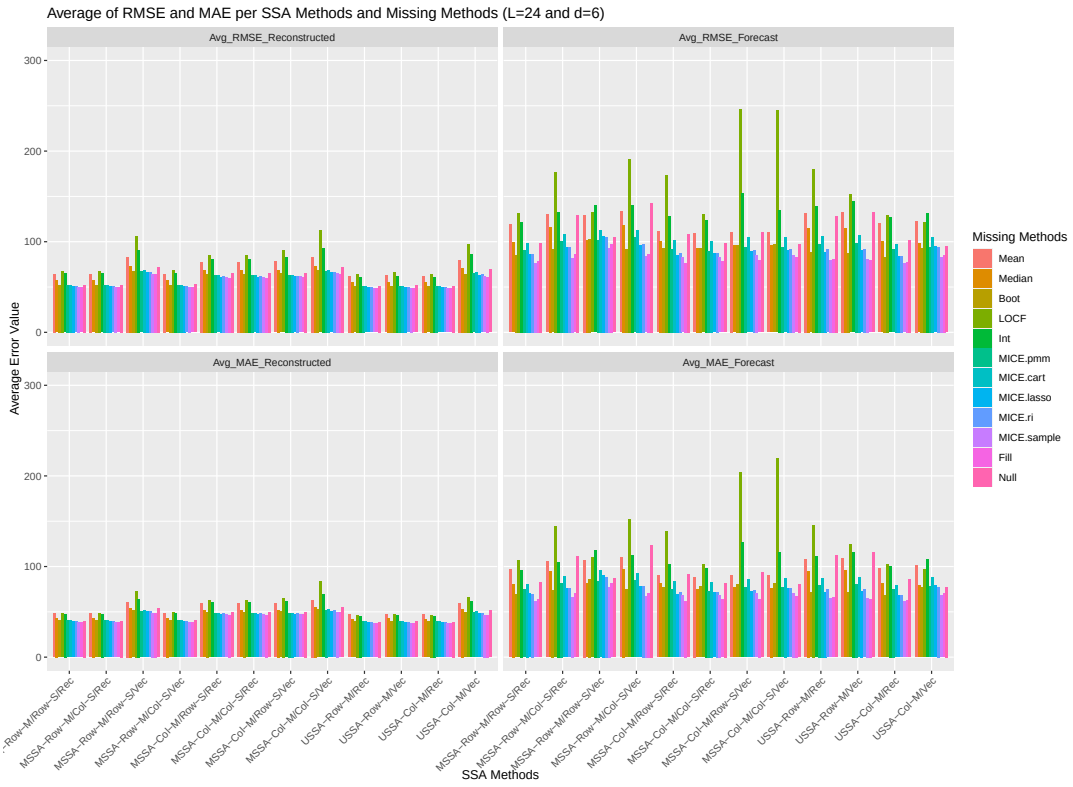


Figure 49: Average RMSE and MAE Values based on SSA and Missing Methods $\{L = 24, d = 6\}$

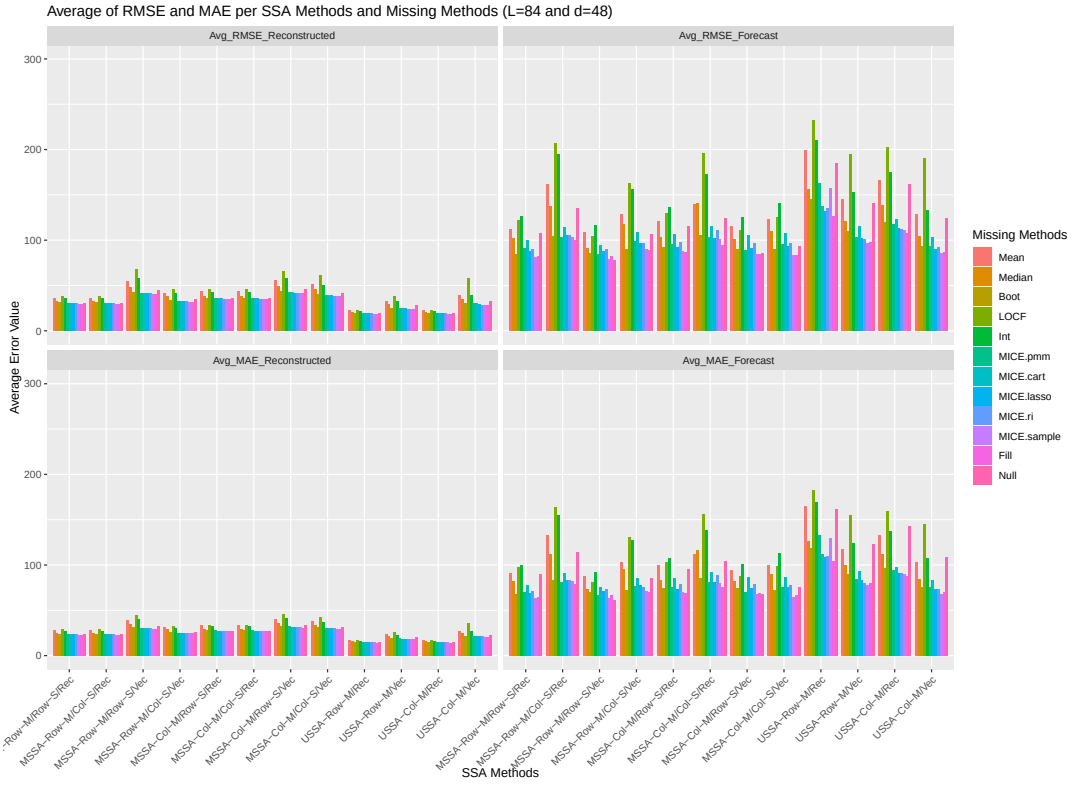


Figure 50: Average RMSE and MAE Values based on SSA and Missing Methods $\{L = 84, d = 48\}$

4.7.4 Summary of Comparisons

There seem to be different conclusions if we want to choose the best method with the lowest error possible, especially if we see Reconstructed and Forecast error values separately. In Forecast, SSA with $\{L = 24, d = 6\}$ has significantly lower error values compared to SSA with $\{L = 84, d = 48\}$. However, it is not the case for Reconstructed because SSA with $\{L = 84, d = 48\}$ have lower error values than SSA with $\{L = 24, d = 6\}$ there. We can say that SSA with $\{L = 84, d = 48\}$ is better for Reconstructed data, while SSA with $\{L = 24, d = 6\}$ is better for Forecast. However, if we take a look again at Figure 49, all missing methods perform better and more similarly under SSA with $\{L = 84, d = 48\}$, which helps if we are not sure which missing method to use. SSA with $\{L = 84, d = 48\}$ guarantees SSA combinations that work better with any missing methods, while SSA with $\{L = 24, d = 6\}$ guarantees SSA combinations that work better with any area of calculation (see Figure 45).The explanation why $\{L = 84, d = 48\}$ does work better for the reconstruction but not in the forecast is that the high d leads to include too much noise in the signal part. In other words, the signal part $\sum_{i=1}^{d} X_i$ should have a lower d than $d = 48$. For $d = 6$, the separation of signal and noise seems to work better, hence the forecast is better.

There are some slightly better SSA methods in Reconstructed with lower error values: USSA-Row-M/Rec, USSA-Row-M/Vec, USSA-Col-M/Rec for USSA and MSSA-Row-M/Row-S/Rec, MSSA-Row-M/Col-S/Rec, and MSSA-Row-M/Col-S/Vec for MSSA. As for Forecast, the slightly better SSA methods are USSA-Col-M/Vec for USSA and MSSA-Row-M/Row-S/Rec and MSSA-Row-M/Row-S/Vec for MSSA. Overall, USSA-Col-M/Rec and MSSA-Row-M/Row-S/Rec are the SSA methods that have the lowest error values. As an addition, in terms of forecasting, Vector forecast seems to provide lower error values compared to Recurrent forecast, especially upon comparing two SSA methods that have the same trajectory matrix form and same stacking method in MSSA.

As for the missing data method, there are several options that perform well: Bootstrap, all MICE methods, and Distribution Fill. All of these missing data methods perform similarly, meanwhile the rest of them: LOCF Imputation, Interpolation, Mean Substitution, Median Substitution, and Null Substitution have high error values. As for the area of calculations, Country is the best pick, especially after seeing how high Province can get. However, if we want various results considering Country only gives one result for all provinces which do not really make sense in reality, Region or Clusters can also be considered. Cluster might be better considering that they have lower forecast error values.

What is interesting from these results is that the best SSA methods do not exactly fit the description in Subsection 4.4.1. It is said that L needs to be large enough to cover the time effect. Hence, we expected that SSA with $\{L = 84, d = 48\}$ had lower error values compared to SSA with $\{L = 24, d = 6\}$. That is the case with Reconstructed, but unfortunately not with Forecast. But as explained above, this result is very positive since it shows that the lower d separates the signal part better than the high d and thus is better for forecasting, while the higher d rather leads to an overfitting, which is better for the reconstruction, but not for forecasting. These effects from the different d seem to dominate the effects from the L. However, we can also argue that bigger L helps with the trend extraction and noise reduction part of SSA, but maybe not so much with the forecasting part. Which window length that we use should depend on the purpose of our calculation. As for the L position in the trajectory matrix, indeed, it does not matter as much as we expected. There are some better forms, but the differences between all SSA methods are not too significant. In addition, all the error values can be seen in Appendix E.

4.8 Rainfall Forecast Results

Based on the previous summary, we will show the forecast results from one of the provinces in all possible categories. We use two SSA methods with $\{L = 24, d = 6\}$ and the lowest Forecast error values: USSA-Col-M/Vec for USSA and MSSA-Row-M/Row-S/Vec for MSSA. We use only one missing data method for easier comparison: Bootstrap. Lastly, the province chosen is Aceh because it keeps having different "group members" in Region, Manhattan-Ward's Cluster, and Euclidean-Single Cluster, and it is never alone in these groups. The comparisons will be done between the five different areas of calculation: Country, Region, Manhattan-Ward's Cluster, Euclidean-Single Cluster, and Province.

Figure 51 shows the comparison between the actual data and the forecast values during the forecast period: 2016-2017 (on the right side of the blue vertical line). The real data in both USSA and MSSA are the same because we use the same database for these two methods. However, the actual series differ for each area due to the weighted values being applied. Province (the lower graph) has the actual data, while the other areas have the weighted values.

Overall, the forecast values seem to capture the pattern of the actual data rather nicely, especially for Country. It is of course normal for Country to have the same results for USSA and MSSA calculations because we only have one series which makes USSA and MSSA work the same way, unlike other areas that have more than one series to be calculated with for MSSA. USSA seems to be fluctuating more extremely in Region and Province compared to MSSA. Also, USSA and MSSA seem to perform rather similarly for the two clusters.

What is interesting from this observation is that there seems to be no methods that can perfectly capture the pattern of Province, especially when it goes up and down very quickly. USSA Province is the closest one to capture the patterns, but the other forecast values don't seem to be accurate. Meanwhile, MSSA Province doesn't even capture the fluctuations. Also, both real data and forecast values stay in the range of 0-400 mm/month. Although, in fairness, the real data values also do not exceed 400 mm/month. Overall, the range seems to stay the same, just the patterns that aren't always captured accurately, but that can of course not be expected due to the variability of the rainfall over the year. The general pattern including the seasonability seems to be captured quite well. In addition, all forecasted rainfall values can be seen in Appendix D.

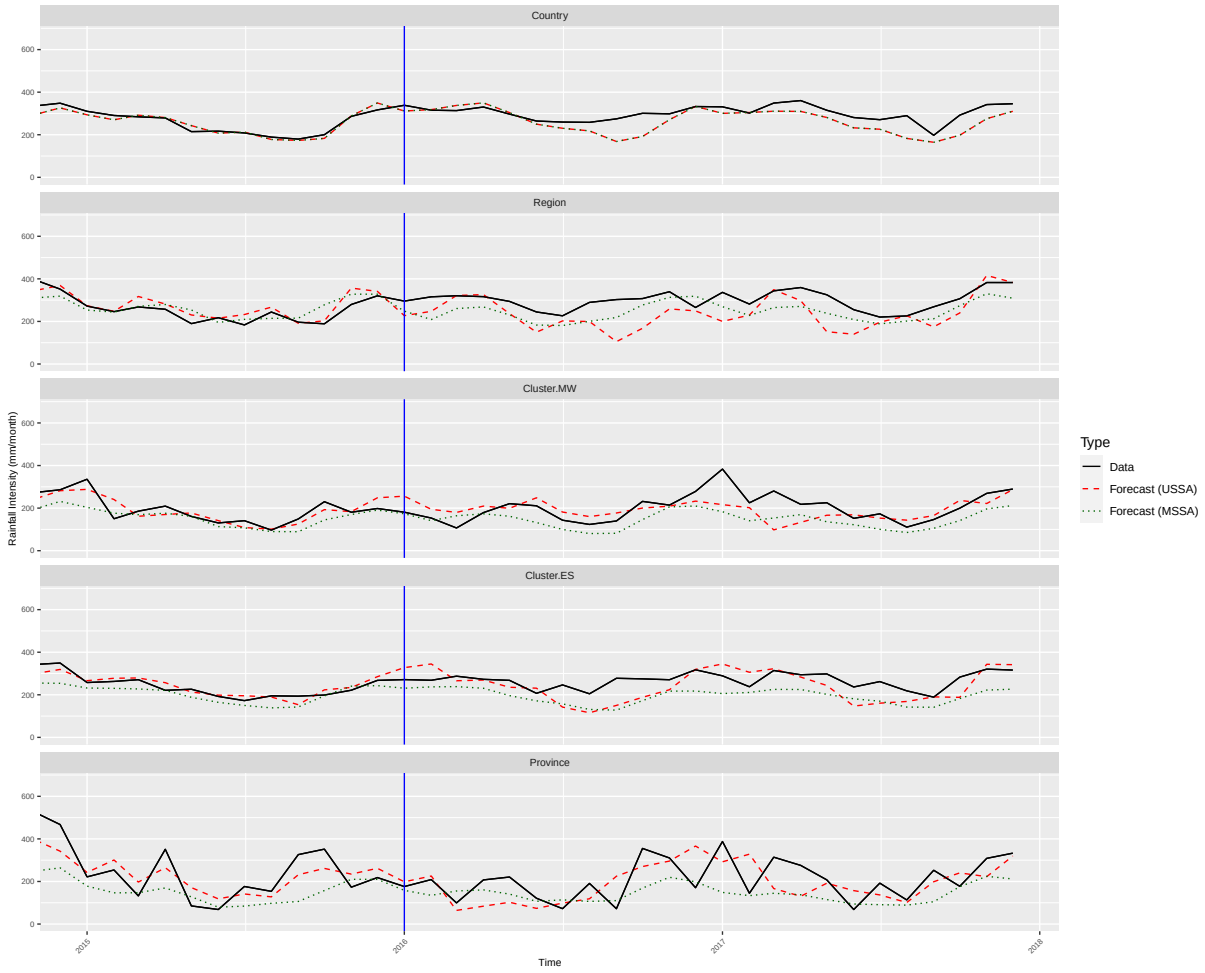


Figure 51: Actual Data and Forecast Values of Rainfall Intensity in Aceh Province

5 Payout Conversion

After forecasting the rainfall intensity and obtaining the results, the next step is converting these rainfall intensity values into payout values. The conversion is calculated based on a linear index insurance model for various scenarios. After converting all the rainfall intensity values, analysis related to these payout values is also conducted to see which scenario fits the best.

5.1 Various Payout Designs

According to the linear index insurance model in [APD23], payouts for coffee crops are determined by the rainfall intensity. The lower and upper limit of the rainfall intensity are set to be 40 and 241 mm/month, and they also set the lower and upper thresholds to be 57 and 135 mm/month. We need to set appropriate limits and thresholds for paddy rice.

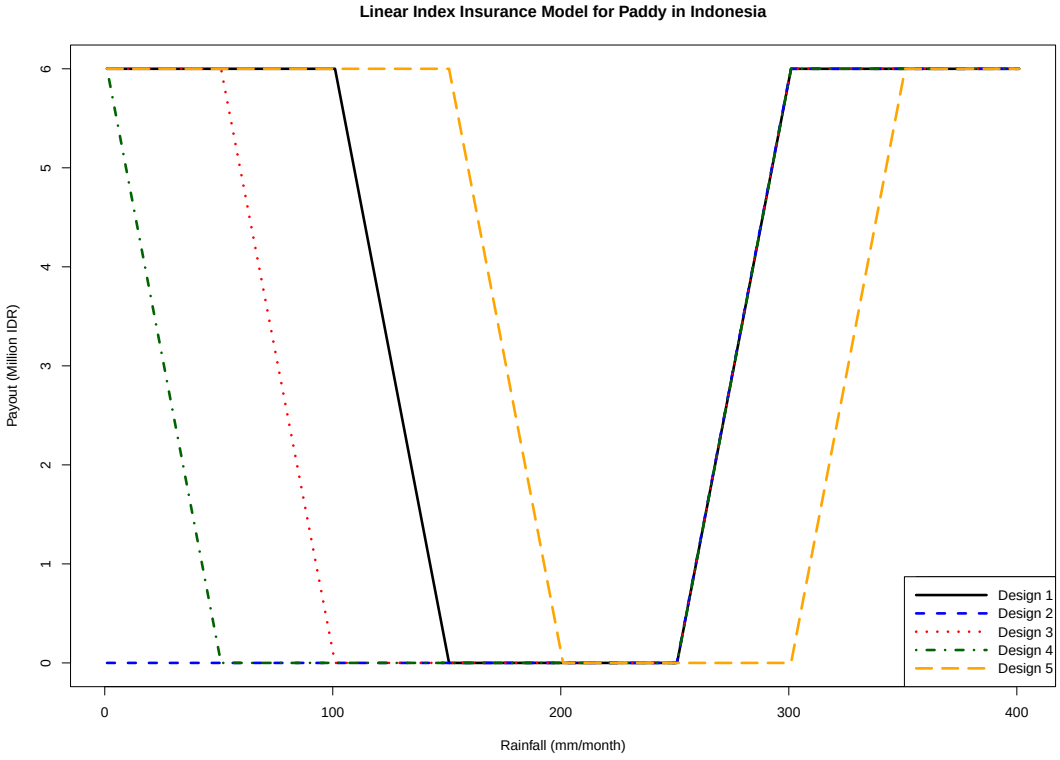


Figure 52: Linear Index Insurance Models for Payout Conversion

According to [Car+18], there are several types of climates in Indonesia that also determine the best period of paddy planting. For some climate types, it is best to begin paddy rice planting when the rainfall intensity is between 100 and 200 mm/month, but in some areas, it is better when the rainfall intensity is above 200 mm/month. However, it is also mentioned that paddy rice crops only need around 150 mm/month of rainfall during the rainy season. Therefore, we decided to use several payout designs to cover all of these options. Currently, there is only one agricultural insurance company in Indonesia. They set the maximum indemnity for paddy rice to be 6 Million IDR per hectare per growing season (around 315 EUR with current rate: 1 EUR $\approx$ 19,000 IDR). We use this value to model the rainfall intensity according to aforementioned situations. The models are shown in Figure 52.

Design 1 (black) uses 200 mm/month as the middle point of the payout model. The lower and upper limits for this model are 100 and 300 mm/month, while the lower and upper thresholds are 150 and 250 mm/month. In this design, we consider 200 mm/month as the average rainfall intensity that the paddy rice crops need. More than 250 mm/month or lower than 150 mm/month, the crops would be proportionally damaged. Above 300 mm/month or lower than 100 mm/month, the crops are fully damaged. However, considering that farmers might use irrigation or do crop rotations, we decided to use Design 2 (blue) as another possibility where the lower limit and threshold are both 0 mm/month, while the upper limit and threshold remain at 250 and 300 mm/month. Here, we assume that the insurance company will not pay any damage caused by no rain.

Now, we discuss the payout designs that are made based on the different climate types in Indonesia that would affect the paddy rice planting. We start with the one that requires 100 until 200 mm/month rainfall intensity which is pictured in Design 3 (red). Here, the lower limit, lower threshold, upper threshold, and upper limit respectively are 50, 100, 250, and 300 mm/month. We extended the upper threshold a bit from 200 mm/month considering the average needs of paddy rice crops, hence we didn't use 200 mm/month as the upper threshold. We use the same reason to motivate Design 4 (green), but instead of using 100-200 interval, we used 150 as the middle point and set the thresholds from there. In this case, the lower limit, lower threshold, upper threshold, and upper limit respectively are 0, 50, 250, and 300 mm/month. Lastly, Design 5 (orange) uses the climate situation where paddy rice crops need over 200 mm/month of rainfall intensity. So, we decided that the lower limit, lower threshold, upper threshold, and upper limit respectively are 150, 200, 300, and 350 mm/month. Of course, we still need to set a high limit because too much rain can also damage the crops.

Using the rainfall forecast values that we obtain from SSA and these linear index models, we obtain the forecast payout values that insurance companies need to prepare. We try seeing which payout model performs the best, followed by seeing the actual values of this model to one of the chosen SSA methods. Following Section 4.8, we will apply the model to USSA-Col-M/Vec and MSSA-Row-M/Row-S/Vec with $\{L = 84, d = 48\}$ using the Aceh province data and the Bootstrap imputation method. All area levels are used (Country, Region, Manhattan-Ward's Cluster, Euclidean-Single Cluster, and Province).

5.2 Conversion Formula

In order to obtain the payout values for each rainfall value per month based on Figure 52, we need a specific formula. According to [APD23], the formula is as follows:

$$PO(x) = \begin{cases} M.I. & x < LL \quad or \quad x \geq UL \\ m_0(x) & LL \leq x < LT \\ 0 & LT \leq x < UT \\ m_1(x) & UT \leq x < UL \end{cases} \tag{5.1}$$

with x as the rainfall value (forecast or non-forecast), M.I. as the maximum indemnity which is 6 Million IDR in this case, $LL = \{100, 0, 50, 0, 150\}$, $LT = \{150, 0, 100, 50, 200\}$, $UT = \{250, 250, 250, 250, 300\}$, and $UL = \{300, 300, 300, 300, 350\}$, following the order of Design 1-5. We also need formulas to obtain m_0 and m_1 due to them not being constant. The formulas are:

$$m_0(x) = \frac{LT - x}{LT - LL} \times M.I. \tag{5.2}$$

$$m_1(x) = \frac{x - UT}{UL - UT} \times M.I. \tag{5.3}$$

The limits and thresholds used to visualize Figure 52 are written in Table 40. These are also the values that we use for Equation (5.1).

Table 40: Limits and Thresholds for Each Payout Design

Design	LL	LT	UT	UL
Design 1	100	150	250	300
Design 2	0	0	250	300
Design 3	50	100	250	300
Design 4	0	50	250	300
Design 5	150	200	300	350

5.3 Model Evaluation Criteria

After obtaining the payout values for each rainfall value, we can start evaluating the models. In order to compare the payout models to one another, we must set some criteria for easier evaluation. There are 6 different indicators that we use to evaluate these models: Mean, Standard Deviation, Value-at-Risk (VaR), Conditional Tail Expectation (CTE), 75th Percentile, and the Difference between Mean and Median (MM).

5.3.1 Mean (Evaluation Criteria)

The first indicator that we use is Mean, or the average of the payout values, or the expected payout values. The higher the Mean, the higher the payout values are, which means the insurance company has to pay more, and vice versa. We want Mean to be as low as possible so that the insurance company does not have to pay a lot, but that it still makes sense for the farmers so that they are not at a disadvantage. The formula is as follows:

$$\overline{PO} = \frac{1}{n}\sum_{i=1}^{n} PO(x_i) \tag{5.4}$$

for the rainfall values x_i.

5.3.2 Standard Deviation

The second indicator is Standard Deviation. This indicator pictures how dispersed the payout values are. If they are more dispersed, it shows extreme changes which will be hard to be predicted in the future. That is why we also want Standard Deviation to be as low as possible. However, having it close to zero is also not a good idea because it means that the payout values remain almost the same for a period of time, and while this is easy to predict, it is also realistically impossible. The formula is as follows:

$$SD(PO) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(PO(x_i) - \overline{PO}\right)^2} \tag{5.5}$$

5.3.3 Value-at-Risk (VaR)

The third indicator is Value-at-Risk (VaR). This indicator measures or quantifies the possibility of large payouts under a certain confidence level. The lower VaR is, the lower the possibility of obtaining high payout values, and vice versa. In this case, we want VaR to be as low as possible,

but not too close to zero because it means most payout values are at 0, which is realistically impossible. The formula is as follows:

$$VaR_{\alpha}(PO) = \inf\{x \in \mathbb{R} : F_{PO}(x) \geq \alpha\}$$

We chose the confidence level to be 95%, which means we adjust the VaR value to be as follows:

$$VaR_{0.95}(PO) = \inf\{x \in \mathbb{R} : F_{PO}(x) \geq 0.95\} \tag{5.6}$$

5.3.4 Conditional Tail Expectation (CTE)

The fourth indicator is Conditional Tail Expectation (CTE), also known as Tail Value-at-Risk (TVaR). This indicator pictures the expected loss in extreme cases beyond VaR. Since we want to be able to control the large losses, we need CTE to be as low as possible. Lower CTE also suggests better performances of the model. The formula to calculate CTE is as follows:

$$CTE(PO) = \frac{1}{N_{CTE}} \sum_{i:PO(x_i)>VaR_{0.95}(PO)}^{N_{CTE}} PO(x_i) \tag{5.7}$$

with N_{CTE} is the number of payout values beyond VaR (tail observations).

5.3.5 75th Percentile

The fifth indicator is 75th Percentile. The formula itself is similar to VaR, showing that it also quantifies the loss beyond a certain confidence level. However, we chose the 75th Percentile because it shows the three quarter of the payout values. Comparing this to VaR also helps us see how extreme the payout values can get. Since our goal is to minimize high payouts, we want this indicator to be as low as possible as well. The formula is as follows:

$$P_{75}(PO) = q_{0.75} = \inf\{x \in \mathbb{R} : F_{PO}(x) \geq 0.75\} \tag{5.8}$$

5.3.6 The Difference between Mean and Median (DMM)

The sixth and last indicator is the Difference between Mean and Median (DMM). This helps us measure the skewness of the payout values. If the difference is zero, then the payout values are symmetrical or well-balanced between high values and low values. If the difference is negative (less than zero; median is greater than mean), that means we have a left-skewed histogram that shows a lot of high payout values. On the contrary, if the difference is positive (greater than zero; mean is greater than median), then we have a right-skewed histogram that shows a lot of low payout values. Since our goal is to have as low payout values as possible, then it is best if we obtain positive difference. The formula is as follows:

$$DMM(PO) = \overline{PO} - med(PO) \tag{5.9}$$

$$med(PO) = \begin{cases} PO_{\left(\frac{n}{2}\right)} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(PO_{\left(\frac{n+1}{2}\right)} + PO_{\left(\frac{n}{2}\right)}\right) & \text{if } n \text{ is even} \end{cases} \tag{5.10}$$

with $PO_{(...)}$ indicating that the values of $PO(x_i)$ are sorted.

5.4 Model Evaluations

5.4.1 Payout Indicators Evaluation

We convert all the forecasted rainfall values from SSA with $\{L = 24, d = 6\}$ into payout values using Equation (5.1). After obtaining all the payout values, we group them based on the missing data methods, SSA methods, areas of calculations, payout models, and whether the rainfall values are reconstructed or forecast. We calculate the average values based on these groups and inspect the differences between payout models per evaluation criteria. The actual payout values can be seen in Appendix F.

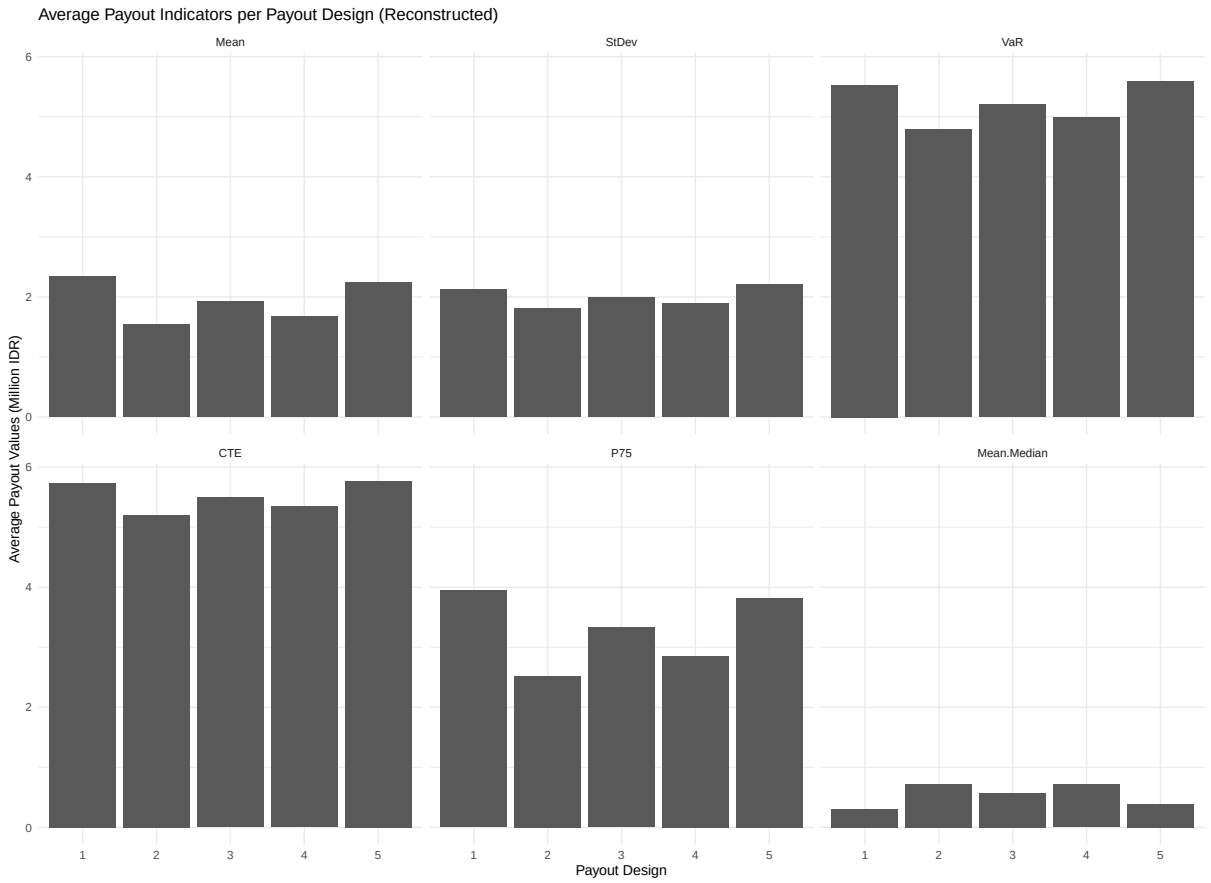


Figure 53: Average Payout Forecast Values for Each Model per Evaluation Criteria

Firstly, we see the comparison between payout models for the forecast values in Figure 53. The mean values for each payout model range from 2 Million IDR to 3.5 Million IDR, with Design 2 having the lowest mean and Design 1 along with Design 5 having the highest mean. Meanwhile, the standard deviation is mostly similar. There are slight differences, which are not significant. VaR is always above 5 Million IDR, with the lowest VaR owned by Design 2 (only slightly below 5) and the highest VaR owned by Design 5. This applies to CTE as well, where the values are all above 5 Million IDR, all similar, and Design 2 is the lowest while Design 5 is the highest. The 75th Percentile shows that Design 2 has the lowest value again, but this time, Design 1 has the highest value. The differences between payout models are more obvious in the 75th Percentile graph. Lastly, the differences between DMM values also look quite significant. All values are surprisingly positive, showing that most mean values are higher than the median values. Design 2 and Design 4 have the biggest DMM values, while Design 1 and Design 5 have

the lowest values.

According to the criteria in Section 5.3, Design 2 shows the best performance because it has the lowest value in Mean, Standard Deviation, VaR, and CTE. However, we need to note that Design 2's model pictures that the insurance company doesn't have to give the payouts when there is no rain (see Figure 52). Due to this, of course the payout values in general would be low, especially if there are a lot of cases where the month has less rain or no rain at all. Besides, Design 2 also has the highest DMM value. Hence, even though Design 2 might be the product that the insurance company would love to have, it might not be the best choice for the entire situation.

For completely opposite reasons, Design 1 and Design 5 are also not the models with the best performance. They consistently have the two highest values in Mean, Standard Deviation, VaR, CTE, and even the 75th Percentile. The only criteria where these two designs are not the highest is DMM. However, because our goal is to pick the lowest in everything and the lowest positive value for DMM, Design 1 and Design 5 are also proven to not have the best performance.

Aside from these two models, Design 4 seems to work well. It has the second lowest values in all criterias, even though it has the highest DMM value. As for Design 3, it is "right in the middle" of every model. It is not the lowest, but also not the highest. Design 3 also has rather low values, shown through the DMM value that is lower than Design 2 and Design 4. According to this evaluation, it looks like it is either Design 3 or Design 4 that works the best. In addition, all the indicator metrics can be seen in Appendix G.

5.4.2 Distribution Evaluation

Aside from evaluating the payout indicators, it is also possible to evaluate the forecast payout values based on their distributions. Figure 54 shows the histogram of the forecast payout values for each payout design. We can see that the first range (0-1 Million IDR) always has the highest probability. Ideally, an insurance company will not prefer a scenario where they have to pay high values more than low values, which means, all designs are actually more preferable for insurance companies than for farmers. Unfortunately, from a farmer's perspective, they would prefer to obtain higher payouts when loss occurs, and these designs do not seem to be in favor of the farmers. That means, it is better to at least pick the designs where the last range (5-6 Million IDR) is high enough so that the design will at least help the farmers a bit. Design 2, Design 3, and Design 4 might not be a good option, seeing how high the first range is. Design 1 and Design 5 look like a good compromise for both the insurance company and the farmers.

All of the forecast payout values follow the pattern of the Beta distribution with $\alpha < 1$ and $\beta < 1$, following a U-shaped Beta distribution. This means the values are heavy on the lower and upper bounds, which is visible from the histogram. These parameters show low density in the middle range. If we refer to the payout design in Figure 52, the shape is quite similar. The middle range is only covered in $m_0(x)$ and $m_1(x)$ that cover smaller range compared to the extreme values of 0 and 6. This scenario is actually close to a Binary situation, where 0 means no payouts are given and 1 means payouts are given.

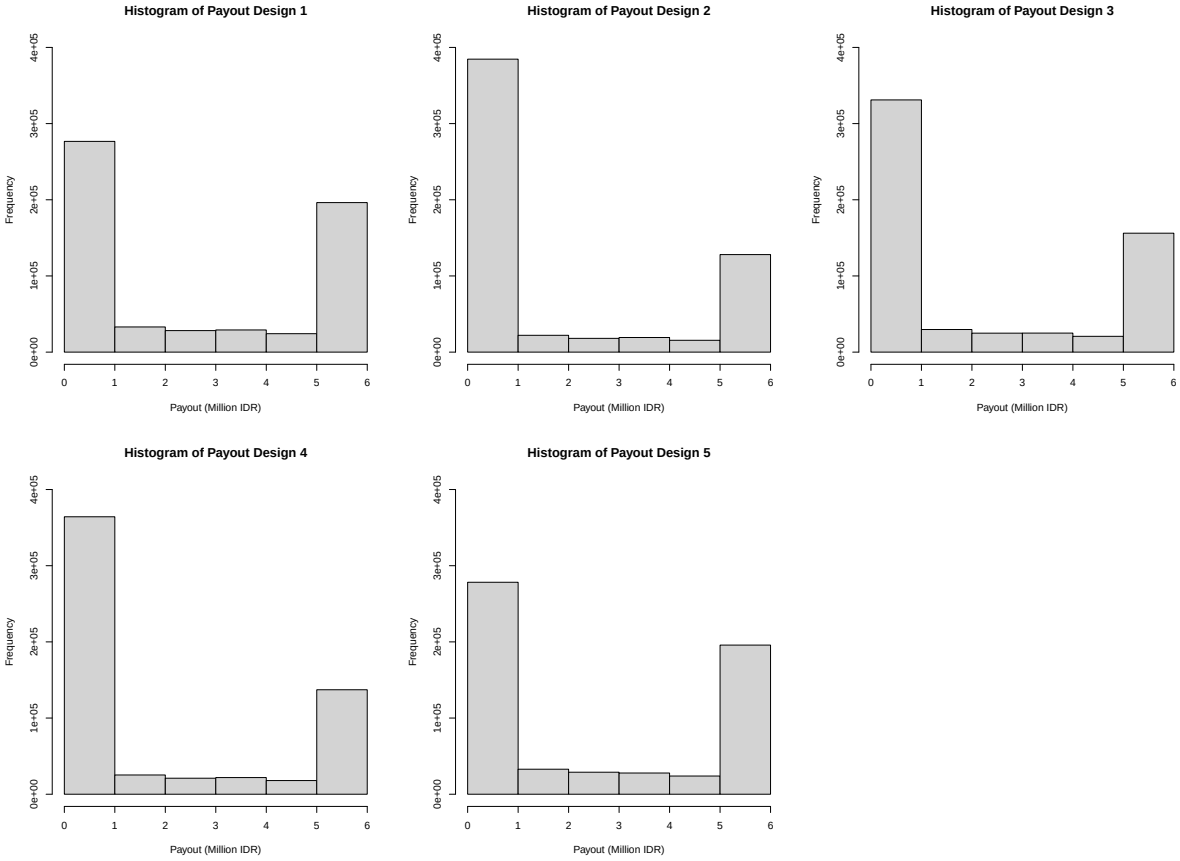


Figure 54: Histogram of Payout Values for Each Model

6 Simulation

Through this part, we are redoing everything starting from daily rainfall values aggregation up to error values and payout values analysis. However, instead of using actual data with imputed values, we use randomly generated data using various distributions. The aim is threefold. First, we want to see how far we can go by just fitting distributions to the observed rainfall data. Second, it is to check the SSA procedure and comparing error values in the forecast. Third, we want to see if we can refine the model to include seasonality aspects. Note that, since we don't simulate the imputation part, the results are not always directly comparable to the results in Chapter 4.

6.1 Simulation Procedures and Results

There are several steps in the simulation, including obtaining parameters, generating randomized data, and applying all SSA combinations into these new generated values. We explain the procedures and summarize the results here.

6.1.1 Obtaining Parameters

First, we need to obtain the parameters for the distributions that we are going to use to generate the data. There are three distributions used to generate the daily rainfall values: Exponential, Normal, and Gamma. We use Exponential distribution because it captures extreme number of observations with low values (although it cannot be zero). We use Gamma for the same reason, but we want to see how the data behave with two parameters in the distribution. Meanwhile, we use Normal distribution because some of the rainfall values are also pretty symmetrical. We use these following formulas to obtain the needed parameters from the real data.

$$\frac{1}{\hat{\lambda}} = \frac{1}{\bar{x}} = \big(\frac{1}{n}\sum_{i=1}^{n} x_i\big)^{-1} \tag{6.1}$$

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{6.2}$$

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{6.3}$$

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{x_i>0\}} \tag{6.4}$$

$$\hat{\alpha} = \frac{\bar{x}^2}{s^2} = \frac{\big(\frac{1}{n}\sum_{i=1}^{n} x_i\big)^2}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{6.5}$$

$$\hat{\beta} = \frac{\bar{x}}{s^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2} \tag{6.6}$$

We use $1/\hat{\lambda}$ as the rate for our Exponential distribution, $\hat{\mu}$ and $\hat{\sigma}$ as the parameters for our Normal distribution, $\hat{p}$ as the proportion for our Binary patterns (rainfall or no rainfall), and $\hat{\alpha}$ and $\hat{\beta}$ as the parameters for our Gamma distribution. Note that since our actual data have a lot of missing values, this parameter estimation only considers the existing values. Hence, the

imputation methods in Chapter 2 take no part in this simulation. The parameters obtained from these formulas are written in Table 41. Since we cannot use zero values to estimate the parameters for Gamma distribution, we change all the zero values to 10^{-10}, ensuring that it is not zero, but still a really small value that should not affect the calculation significantly.

Table 41: Estimated Parameteres for Various Distributions per Province

Province	$\mathbf{1/\hat{\lambda}}$	$\hat{\boldsymbol{\mu}}$	$\hat{\boldsymbol{\sigma}}$	$\hat{\boldsymbol{p}}$	$\hat{\boldsymbol{\alpha}}$	$\hat{\boldsymbol{\beta}}$
Aceh	0.203	4.93	13.03	0.40	0.059	0.0120
Sumatera Utara	0.135	7.41	15.76	0.48	0.069	0.0093
Sumatera Barat	0.075	13.35	27.90	0.57	0.077	0.0058
Riau	0.131	7.61	15.99	0.49	0.077	0.0102
Jambi	0.149	6.70	14.24	0.47	0.069	0.0104
Sumatera Selatan	0.120	8.31	16.73	0.52	0.073	0.0088
Bengkulu	0.095	10.58	21.13	0.55	0.076	0.0072
Lampung	0.171	5.84	13.88	0.43	0.063	0.0108
Kepulauan Bangka Belitung	0.099	10.06	17.13	0.64	0.098	0.0097
Kepulauan Riau	0.129	7.73	18.44	0.51	0.074	0.0096
DKI Jakarta	0.119	8.43	18.69	0.48	0.069	0.0082
Jawa Barat	0.140	7.12	13.35	0.54	0.077	0.0108
Jawa Tengah	0.161	6.21	14.32	0.40	0.060	0.0097
DI Yogyakarta	0.154	6.50	15.83	0.40	0.061	0.0093
Jawa Timur	0.200	5.01	12.70	0.35	0.055	0.0111
Banten	0.201	4.97	13.16	0.36	0.059	0.0118
Bali	0.184	5.42	14.35	0.36	0.056	0.0103
Nusa Tenggara Barat	0.238	4.21	12.11	0.30	0.051	0.0122
Nusa Tenggara Timur	0.273	3.66	10.54	0.29	0.053	0.0143
Kalimantan Barat	0.108	9.29	18.18	0.54	0.079	0.0085
Kalimantan Tengah	0.076	13.13	20.26	0.77	0.157	0.0120
Kalimantan Selatan	0.103	9.68	16.44	0.64	0.094	0.0097
Kalimantan Timur	0.141	7.09	13.93	0.53	0.074	0.0105
Kalimantan Utara	0.127	7.90	15.08	0.56	0.078	0.0099
Sulawesi Utara	0.123	8.15	14.98	0.62	0.088	0.0108
Sulawesi Tengah	0.131	7.61	16.46	0.51	0.069	0.0091
Sulawesi Selatan	0.107	9.34	20.39	0.45	0.062	0.0066
Sulawesi Tenggara	0.143	6.97	14.52	0.51	0.071	0.0101
Gorontalo	0.205	4.89	11.56	0.41	0.061	0.0124
Sulawesi Barat	0.195	5.13	15.76	0.39	0.059	0.0115
Maluku	0.131	7.63	16.37	0.55	0.075	0.0098
Maluku Utara	0.161	6.21	14.12	0.44	0.063	0.0101
Papua Barat	0.103	9.71	18.39	0.63	0.097	0.0100
Papua	0.134	7.49	15.31	0.55	0.078	0.0104

Aside from these parameters, we also need to consider the patterns of the daily rainfall values using Binary patterns, Rain (1) or No Rain (0). Since this Binary pattern only considers the days with rain, it means when the rainfall value is 0 (no rain), the day is not considered to be a part of the sample. Hence, for Binary simulation, we need to use different parameters that do not consider 0 values. Table 42 shows the adjusted parameters that do not consider zero values.

The other consideration for the simulation is the monthly rainfall values. Just like how

we substitute mean and median per month for each province, we use monthly parameters for each province. However, for some months that have completely missing values, we obtain their parameters from the previous month or year (LOCF), depending on where the missing values are. The parameters for this part can be seen in Appendix H.1.

Table 42: Estimated Parameteres for Various Distributions per Province without Zero Values

Province	$\mathbf{1/\hat{\lambda}}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\alpha}$	$\hat{\beta}$
Aceh	0.082	12.14	18.19	0.742	0.0611
Sumatera Utara	0.070	14.34	19.52	0.625	0.0436
Sumatera Barat	0.044	22.75	33.36	0.660	0.0290
Riau	0.077	13.06	19.17	0.534	0.0409
Jambi	0.077	12.94	17.64	0.609	0.0471
Sumatera Selatan	0.066	15.14	20.17	0.680	0.0449
Bengkulu	0.054	18.44	25.16	0.679	0.0368
Lampung	0.078	12.88	18.29	0.699	0.0542
Kepulauan Bangka Belitung	0.068	14.69	18.99	0.683	0.0465
Kepulauan Riau	0.072	13.92	22.94	0.608	0.0437
DKI Jakarta	0.061	16.41	23.44	0.696	0.0424
Jawa Barat	0.080	12.50	15.67	0.710	0.0568
Jawa Tengah	0.069	14.52	18.94	0.684	0.0471
DI Yogyakarta	0.066	15.06	21.26	0.660	0.0438
Jawa Timur	0.072	13.82	17.99	0.731	0.0529
Banten	0.081	12.28	18.39	0.618	0.0503
Bali	0.068	14.69	20.54	0.681	0.0464
Nusa Tenggara Barat	0.071	13.99	18.73	0.751	0.0537
Nusa Tenggara Timur	0.086	11.61	16.12	0.667	0.0574
Kalimantan Barat	0.064	15.66	21.39	0.614	0.0392
Kalimantan Tengah	0.064	15.70	21.23	0.632	0.0403
Kalimantan Selatan	0.069	14.56	18.32	0.734	0.0504
Kalimantan Timur	0.078	12.87	16.66	0.721	0.0560
Kalimantan Utara	0.074	13.56	17.71	0.706	0.0520
Sulawesi Utara	0.078	12.88	17.14	0.795	0.0617
Sulawesi Tengah	0.067	14.85	20.53	0.750	0.0505
Sulawesi Selatan	0.048	20.64	26.18	0.766	0.0371
Sulawesi Tenggara	0.075	13.30	17.84	0.758	0.0570
Gorontalo	0.086	11.57	15.46	0.775	0.0670
Sulawesi Barat	0.080	12.57	22.69	0.650	0.0517
Maluku	0.073	13.76	19.97	0.751	0.0546
Maluku Utara	0.072	13.85	18.40	0.764	0.0552
Papua Barat	0.070	14.19	20.75	0.629	0.0443
Papua	0.077	12.98	18.30	0.711	0.0548

Finally, the last consideration for the simulation is to use not only monthly parameter values, but to combine it with the Binary pattern. Hence, we have a monthly Binary simulation that calculates the parameters from monthly values, but also only using values above zero. Excluding zeros from the parameter calculation, we obviously have more missing values for this parameter generation. Hence, we apply LOCF again to obtain the parameters from the previous month or year. These parameters can also be seen in Appendix H.2.

6.1.2 Generating Randomized Data

After obtaining the parameters for each distribution and each province, we can generate the randomized data. The simulations are done 100 times for $N = 6575$ days.

We randomize the daily values using the parameters obtained above for each distribution (Exponential, Normal, and Gamma), and each province n times with $n = 100$ (see Table 41), change all values below zero to zero (specifically for Normal distribution), and aggregate the values monthly. The daily generated-data summary for each distribution is written in Table 43.

Table 43: Summary of Randomized Daily Data

Statistic	**Exponential**	**Normal**	**Normal** (≥ 0)	**Gamma**
Min	≈0.00	-117.05	0.00	≈0.00
Q1	1.99	-3.26	0.00	≈0.00
Median	4.90	7.12	7.12	0.01
Mean	7.48	7.48	10.79	7.48
Q3	10.10	17.81	17.81	1.14
Max	199.31	149.33	149.33	1462.46
SD	8.12	16.39	12.24	28.28
0's	0	0	7195353	0
0's per Prov & Sim	0	0	2116.28	0

Exponential distribution starts at almost zero value as the minimum value, progressing to 1.99 as first quartile and 7.48 as median rather quickly, unlike Gamma that starts with almost 0 and continues with almost 0 for first quartile and with 0.01 for median. The fastest one is Normal. Before being capped at minimum zero, it starts at -117.05 as the lowest value, proceeding to -3.26 as the first quartile, and 7.12 as median, which is also higher than Exponential's median. The third quartile is even bigger, 17.81, much bigger than Exponential which is only 10.10, and even much bigger than Gamma which is 1.14. Interestingly enough, the jump to the maximum value takes a different turn. Gamma has the highest maximum value, making it have the widest range and also the highest jump from the third quartile. Meanwhile, Normal has the lowest maximum value, and Exponential is more or less in the middle. The mean values of each distribution also don't picture this sudden jump. They are all the same, except for Normal that has been adjusted to zero values. Unsurprisingly, of course, the standard deviation for Gamma is the highest, followed by Normal, and then Exponential. After the negative values are adjusted, Normal's standard deviation decreases to 12.24, showing that by changing the negative values to zero, we're also decreasing the variability. We can also see that the data generation does not produce any zero values, but after Normal-generated data values are adjusted, the number of zero jumps from 0 to 7,195,353 (around 2117 zero values per province per simulation), which shows how many negative values that the actual Normal-generated data values have. Clearly, after setting the negative values to 0 for the Normal distribution, the mean becomes too large. In Section 6.3 we show how to set the parameters such that we end up with the correct mean.

As for Binary, we use the proportions we obtain before as the probability for Bernoulli distributions. We make Binary patterns for the rain days with 1 and 0 for each distribution and each province n times with $n = 100$. Once we have all the patterns, we substitute all 1 values by generating Exponential, Normal, and Gamma distributions with the parameters we have in Table 42 while the 0 values remain as 0. We also substitute all negative values in Normal data generation to zero. Afterwards, we aggregate the values monthly. The daily generated-data summary for each distribution under the binary pattern is written in Table 44.

Table 44: Summary of Randomized Daily Data under Binary Distribution

Statistic	**Exponential**	**Normal**	**Normal (≥ 0)**	**Gamma**
Min	0.00	-149.14	0.00	0.00
Q1	0.00	0.00	0.00	0.00
Median	0.00	0.00	0.00	0.00
Mean	7.09	7.09	8.44	7.09
Q3	9.59	13.67	13.67	7.89
Max	293.19	184.65	184.65	369.45
SD	12.67	16.06	14.46	14.47
0's	11412649	11412649	13975020	11412649
0's per Prov & Sim	3356.66	3356.66	4110.30	3356.66

Just like before, Gamma and Exponential start with 0 as the minimum value. However, unlike in Table 43, this time it is not almost 0, but rather actual 0 due to the Binary pattern before the distribution-based data generation. This goes for Gamma's first quartile and median too, showing that the start for this distribution is really low. The jump for Exponential is also not that high: from 0 to 1.99 (first quartile) and to 4.90 (median). As for Normal, as usual, the jump from the minimum value to first quartile is quite high: -149.14 to 0, but then the median is also 0. The third quartile for Exponential and Normal is almost the same, but Exponential is now lower than Normal. Gamma still has the lowest value with 7.89 as its third quartile. However, the values jump again to the maximum value with Gamma having the biggest jump: to 369.45. Meanwhile, just like before, Exponential and Normal have almost similar maximum values: 199.31 and 184.65. The mean values are almost similar for all of them, even for Normal after the negative values are adjusted to zero. Surprisingly, Normal and Gamma have the same mean, and both are lower than Exponential's mean. The standard deviations, on the other hand, are not the same. Exponential has the lowest, followed by Gamma, and then Normal. However, this time, after the negative values are adjusted, Normal's standard deviation becomes similar to Gamma's standard deviation. As for the zero values, seeing from before that the data generation does not produce any zero values, and seeing that the number of zero values are all the same except for Normal after adjustment, we can safely say that the Binary pattern produces the number of zero values, 11,412,649 of them to be exact (around 3357 zero values per province per simulation). As for Normal after adjustment, the number of zero values increases to 13,975,020 (around 4111 zero values per province per simulation). The difference is 2,562,371 which is the number of negative values obtained from the random data generation.

Compared to Table 43, of course, we have more zero values, which makes more sense because in reality, it is not always raining. Surprisingly, however, even though there are a lot more zero values with the Binary distribution, the statistics haven't changed much. We still have similar minimum values, first quartile, mean, standard deviation, and even third quartile and also maximum values for some distributions.

Next, we also consider the monthly parameters to generate the data for the simulation. The process is similar to the first data generation (see Table 41): we randomize the daily values using obtained parameters for each distribution, each province, and each month n times with $n = 100$. change all values below zero to zero (specifically for Normal distribution), and aggregate the values monthly. The daily generated-data summary using monthly parameters is written in Table 45.

Table 45: Summary of Randomized Daily Data with Monthly Parameters

Statistic	**Exponential**	**Normal**	**Normal (≥ 0)**	**Gamma**
Min	≈0.00	-382.28	0.00	≈0.00
Q1	1.15	-0.59	0.00	≈0.00
Median	3.87	4.97	4.97	0.02
Mean	7.88	7.88	10.12	7.88
Q3	9.91	15.08	15.08	3.03
Max	537.90	421.92	421.92	2432.88
SD	11.55	16.59	13.82	27.79
0's	0	0	6413687	0
0's per Prov & Sim	0	0	1886.38	0

Without the Binary pattern, we have no zero values again, excluding the adjusted values in Normal distribution. This time, the data ranges are wider than before. Gamma and Exponential still start with almost zero values as their minimum values, but the maximum values are very high: 537.90 and 2432.88. As for Normal, the range is also quite wide: with -382.28 as the minimum value and 421.92 as the maximum value. The adjusted negative values managed to bring the range for Normal lower, but it is still the biggest compared to the previous summaries. Aside from the range, the mean for all these distributions are mostly similar, and they are also similar with the previous summaries: always in the range of 6-8. The standard deviations vary between distributions, but Normal's standard deviation is the closest one to its previous summaries compared to Exponential and Gamma. Gamma's standard deviation is, however, pretty close to the first generation (one parameter per province).

Lastly, similar to the previous Binary, we use the proportions as the probability for Bernoulli distribution. We make Binary patterns for the rain days with 1 and 0 for each distribution and each province n times with $n = 100$. Once we have all the patterns, we substitute all 1 values by generating Exponential, Normal, and Gamma distributions with the monthly parameters that exclude zero values, while the 0 values remain as 0. We also substitute all negative values in Normal data generation to zero. Afterwards, we aggregate the values monthly. The daily generated-data summary using monthly parameters under the binary pattern is written in Table 46.

Table 46: Summary of Randomized Daily Data with Monthly Binary Parameters

Statistic	**Exponential**	**Normal**	**Normal (≥ 0)**	**Gamma**
Min	0.00	-690.51	0.00	0.00
Q1	0.00	0.00	0.00	0.00
Median	0.00	0.00	0.00	0.00
Mean	6.707	6.713	7.59	6.699
Q3	7.79	10.39	10.39	6.94
Max	1085.47	894.11	894.11	907.84
SD	13.77	15.47	14.36	15.02
0's	11412649	11412649	13354564	11412649
0's per Prov & Sim	3356.66	3356.66	3927.66	3356.66

Based on the summary, this generation has the biggest range for all distributions, except for Gamma. Seeing the jump from 0 to 1085.47 (minimum to maximum value) for Exponential and

the jump from -690.51 to 894.11 (also minimum to maximum value) for Normal, it is clear that the ranges are so wide. Gamma's range is also wide, seeing that the minimum value is 0 and the maximum value is 894.11. However, this range is not as wide as the one with monthly parameters (see Table 45). Despite the wide range, the other values seem to be similar to one another. The mean and standard deviation are all similar, in the range of 6-8 and 13-16 respectively. The number of adjusted negative values is also similar to the Binary one, showing that there are not too many changes despite the difference in parameters.

6.1.3 Applying SSA

After obtaining the randomized daily values, we aggregate them into monthly values and apply SSA to these monthly values. All method combinations are used for this part, and everything is run 100 times using $\{L = 84, d = 48\}$. From this process, we obtain the error values for each method combination, and then we use only the forecast part of the error values (2016-2017) to create a graph comparing the RMSE and MAE values.

As shown in Figure 55, the lowest error values belong to USSA-Col-M/Vec for univariate and MSSA-Row-M/Row-S/Vec for multivariate. Aside from these two, USSA-Row-M/Vec also seems to perform rather well, so do MSSA-Col-M/Row-S/Vec, MSSA-Col-M/Col-S/Vec, and MSSA-Row-M/Row-S/Rec. Overall, this simulation seems to show that Vector forecast performs better than its Recurrent counterpart. As for the trajectory matrix form (M), Row and Column perform almost similarly. Lastly, for the matrix stack (S) in Multivariate SSA, Row performs slightly better than Column. Overall, MSSA seems to perform better compared to USSA. Of course, MSSA-Row-M/Col-S/Rec and MSSA-Col-M/Col-S/Rec has higher error values than USSA-Row-M/Vec and USSA-Col-M/Vec, but that is a rare case. Compared to the highest error values in USSA, those two methods are still lower. Hence, we can say that MSSA performs better than USSA in this simulation.

As for the distributions for the simulation, seeing the calculations using normal parameters (no months, zero values included, no binary) the lowest error values belong to Exponential, followed by Normal, then Gamma. Gamma also has really high error values, quite far compared to Exponential and Normal. Under Binary distribution, the order is still the same: Exponential, Normal, and Gamma, although now Gamma's error values are closer to Normal and Exponential compared to the normal parameters. Using the monthly parameters, again we have the same order: Exponential, Normal, Gamma, but this time, all distributions show high error values, much higher than the other calculations. Lastly, using monthly parameters under Binary distribution, the order from lowest to highest changes to Exponential, Gamma, and Normal, although in here, the difference is not too significant. Hence, it is safe to say that Exponential proves to simulate the situation better compared to the other distributions. Normal can be a good option, but we have to remember to adjust the negative values. Meanwhile, Gamma might not be a good choice due to its tendency to generate extremely high values.

6.1.4 Converting to Payout

Aside from getting the error values from the SSA calculations, we also obtain the rainfall values, which we convert into five different payout designs just like in Chapter 5. Just like before, we create a graph comparing each design based on several metrics and we also only use the forecast part of the rainfall values (2016-2017).

Figure 56 shows how each distribution behaves with different payout designs. Starting from the entire-series parameters, we see that Normal has high values in Mean, VaR, CTE, and also 75th Percentile. All of its DMM values, however, are negative, which is what we are hoping for. Interestingly, the standard deviation for Normal is rather low, showing that there is less

variability, and seeing how high the other values are, it is possible that Normal has mostly high values. Design 5 is has the lowest values for Normal, but even it is still pretty high. On the contrary, Exponential and Gamma have much lower values, with Exponential being the lowest. Exponential has low VaR values that don't even reach 3 Million IDR, showing how low all the values are. The DMM values are always positive, showing tendency of higher payout values. However, with the low VaR values, it is possible that the aforementioned higher payout values are not as high as Normal. Also, Design 1 and Design 5 seem to behave similary for Exponential, while Design 2, Design 3, and Design 4 are similar. As for Gamma, it is higher than Exponential, but none of the values reach 5 Million IDR. The DMM values are also positive, but with VaR values that are lower than 5 Million IDR, the highest payout value might not be as much as Normal. All payout designs behave slightly differently for Gamma.

Next, we observe the payout designs under Binary pattern. Again, we have the same order: Exponential is the lowest, followed by Gamma, and lastly Normal. This time, however, the payout metrics of Normal are not as high. In fact, none of the values from the three distributions here are high. They are all under 5 Million IDR. However, this time, none of them have negative DMM values. All of them have positive DMM values, but followed by low VaR values and also CTE, the higher payout values must not be that high. They also have low standard deviation values, showing low variabilities.

Moving on to the monthly parameters generation, just like the error values, they also show the highest payout metrics. The mean values might not be as high as the first Normal distribution, but the other indicators clearly show how high the payout values are. What is interesting is that most of the DMM values for the monthly parameters generation are negative. Normal, specifically, has all negative values for the DMM, while Exponential only has it for Design 1 and Design 3, and Gamma has it for all except Design 2. However, it is also important to note that even though most of the payout values are low according to the DMM, there are still some high payout values that are pictured through VaR and CTE. Compared to the other distributions, these three distributions using the monthly parameters also have the highest standard deviation values.

As for the combination of monthly parameters and Binary pattern, the order is the opposite: Gamma is the lowest, followed by Normal, and Exponential is the highest. This only doesn't happen for Design 5 where Normal is the lowest, although the difference with Gamma isn't too big. Exponential has really high values, especially in VaR, CTE, and even 75th Percentile. At least the DMM values are negative for Design 1 and Design 3, but aside from that we can see that Exponential's indicators are too high. Gamma, on the other hand, is really low. The VaR and CTE are only around 3 Million IDR, nothing beyond that. It is almost as low as Exponential from the one-series parameter. The DMM values, however, are all positive, even though the value isn't that big. As for Normal, the DMM values are lower than Gamma, and the other indicators are also not as high as Exponential.

Overall, the payout designs seem to be performing similarly. No payout designs that stand out alone. Although, if we want to avoid obtaining extreme values, Design 5 might be the best choice, especially seen through the Mean values. Design 5 is the only payout design where Normal distribution doesn't have high mean values. The other designs have around 5 Million IDR, while for Design 5, the Mean is less than 4 Million IDR. Aside from that, Monthly Normal also has the lowest mean in Design 5. There are also three negative DMM values in Design 5: Normal, Monthly Normal, and Monthly Gamma. Model-wise, however, Design 5 has the highest payout values (see Figure 52). Even so, based on this simulation, we can safely say that Design 5 gives the best indications for lower payout scenario.

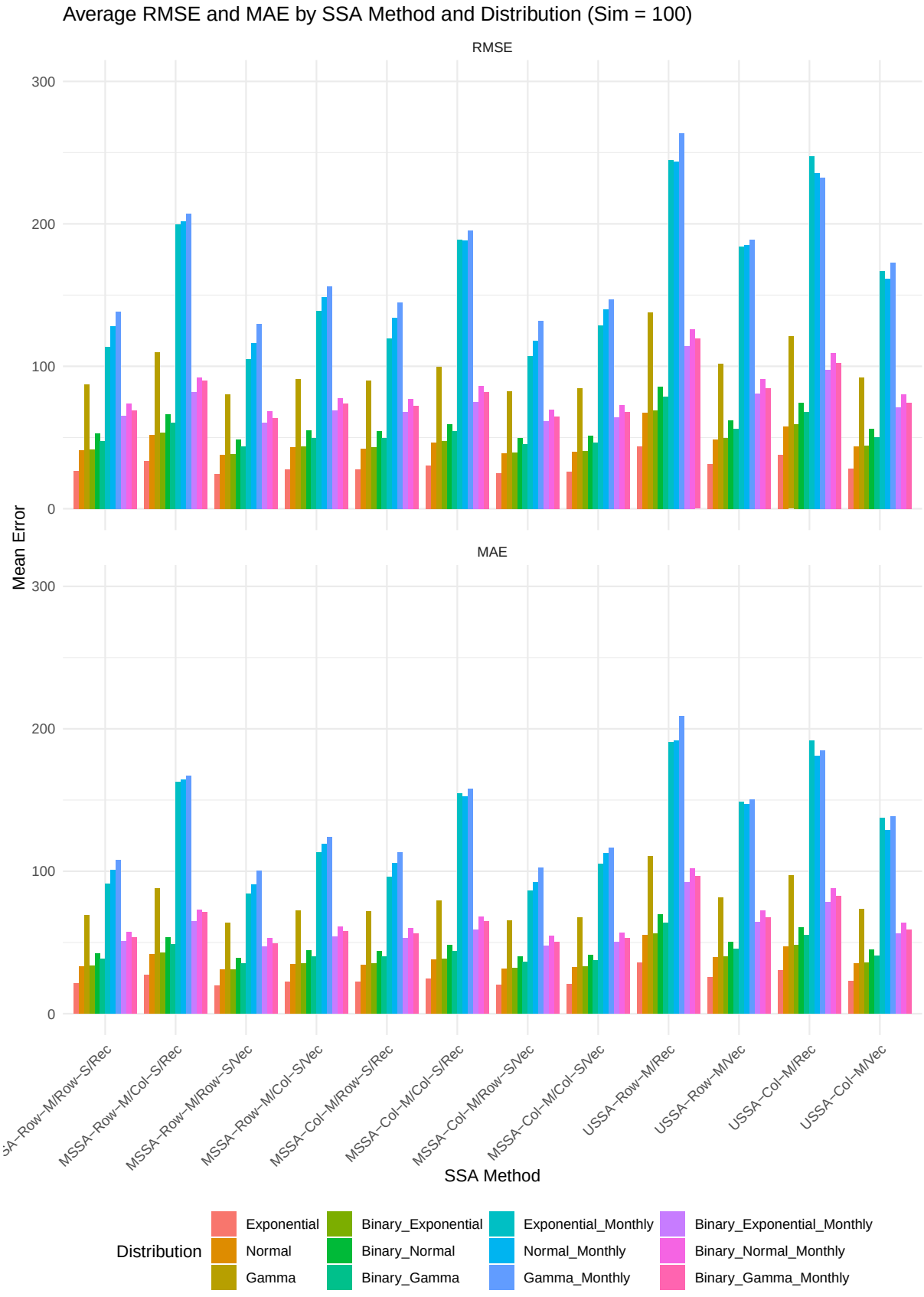


Figure 55: RMSE and MAE Values of Each SSA Method (Sim = 100)

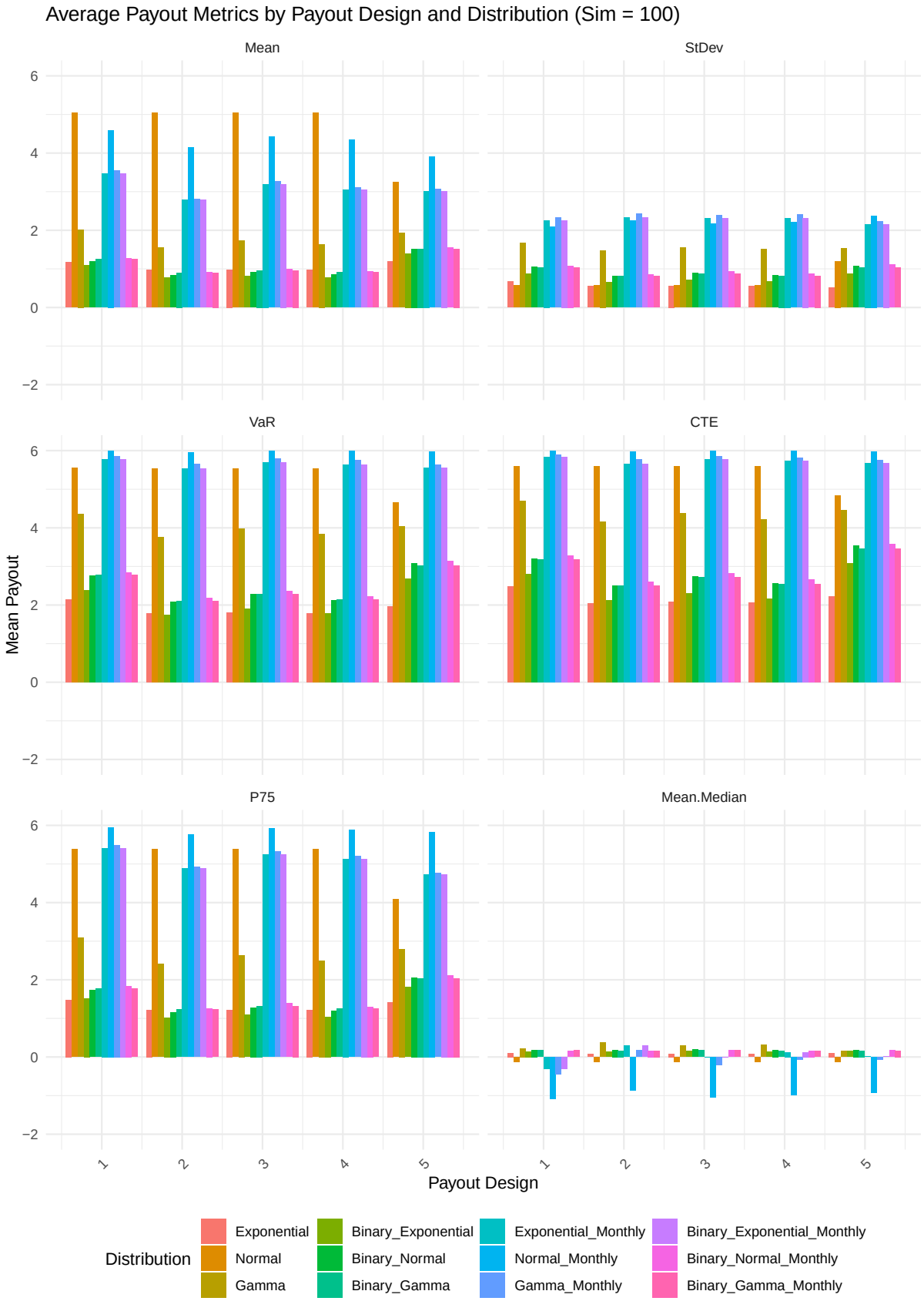


Figure 56: Average Metrics of Each Payout Design (Sim = 100)

6.2 Comparisons with Actual Data (Bootstrap)

The main purpose of a simulation is to portray a real life situation as accurate as possible. Hence, it is necessary to compare our simulation results with our actual calculation results. Through this part, we compare the simulation results with the actual calculations that used Bootstrap data imputation method. The daily imputed data values can be seen in Subsection 2.3.6. Note that we only compare the error values, not the payout designs.

Figure 57 shows the RMSE and MAE values between forecast values from the simulation and actual values from the real data. We use Equation (4.5) and Equation (4.6) to calculate the error values with y_i as the actual data values and $\hat{y}_i$ as the forecast values from the simulation. Once we obtain all RMSE and MAE values from each calculation combination, we calculate the average per SSA Method and per Simulation Distribution. The results are shown in Figure 57.

Overall, if we compare these results to actual values only (see Figure 50), the error values are almost similar to the ones that Bootstrap Forecast has. Some SSA methods have slightly higher error values, but some are similar and even lower. The lower ones are especially visible in MSSA-Col-M/Row-S/Vec, while the higher ones, on the other hand, are visible in the monthly parameter simulations.

Additionally, if we compare these results to the simulation (see Figure 55), the error values in simulation are mostly lower. This shows that simulation has more stable results than actual data. Even though the error values are lower, that doesn't necessarily mean that the results are better. It just means that the forecast values are surprisingly well predictable, which is guaranteed due to the data also being generated through various distributions.

Overall, the simulation is more accurate than the actual calculation. The most mismatch shown through Figure 57 mostly came from actual calculation, not from simulation. However, of course not all distributions have close results to the actual calculation. If we want to pick distributions that have lower error values than Bootstrap calculation, we can pick any distribution from the simulation except for Gamma and all the Monthly distributions. However, if we want to pick the one that has the closest error values, Gamma is the best option.

But we cannot draw too many conclusions here since first the parameters are filled to the imputed values only, i.e. in the simulation we don't consider the imputed values. Second, the simulation more or less only says how much we can predict data from the same distribution, it does not capture any dependence of the data.

6.3 Simulation with Corrected Normal Parameters

We also want to adjust the parameters of Normal distribution in order to fit the data better for a more accurate simulation. The formulas that we have in Equation (6.2) and Equation (6.3) are obtained from MLE assuming that $x \in (-\infty, \infty)$. However, it is impossible for rainfall values to be negative, which is why we needed to adjust Normal-generated rainfall negative values to zero which leads to means which are too high. Hence, we adjust the parameters $\hat{\mu}$ and $\hat{\sigma}$ into the corrected parameters $\hat{\mu}^*$ and $\hat{\sigma}^*$ by taking the positive rainfall values into account. We calculate the value as follows.

$$\begin{aligned} P(X > 0) &= \hat{p} \\ P\left(\frac{X - \hat{\mu}^*}{\hat{\sigma}^*} > \frac{0 - \hat{\mu}^*}{\hat{\sigma}^*}\right) &= \hat{p} \\ P\left(Z > -\frac{\hat{\mu}^*}{\hat{\sigma}^*}\right) &= \hat{p}, \text{with } Z \sim N(0,1) \end{aligned} \tag{6.7}$$

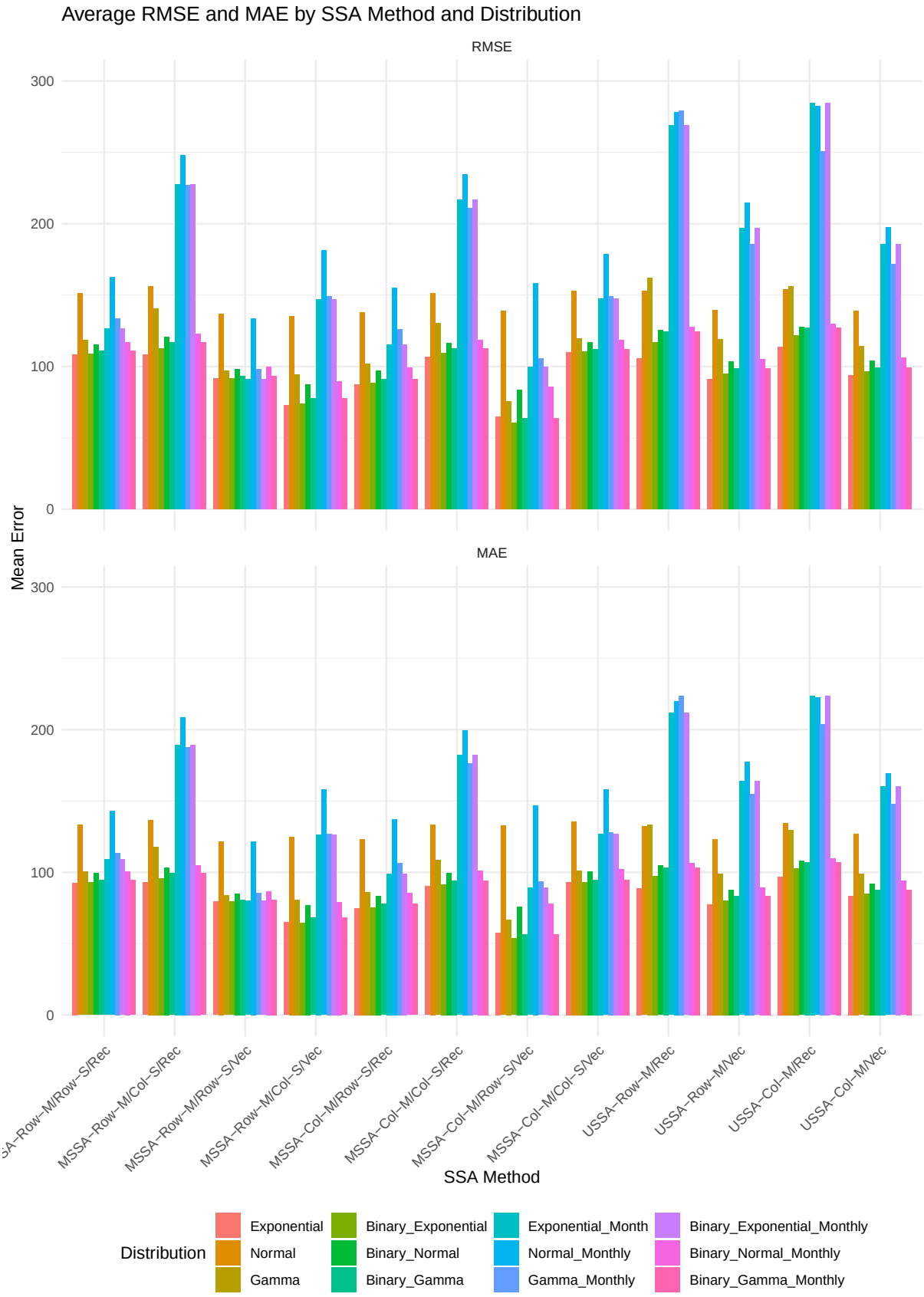


Figure 57: RMSE and MAE Values between Data and Simulation

$$q_{\hat{p}}^{(Z)} = -\frac{\hat{\mu}^*}{\hat{\sigma}^*} \tag{6.8}$$

$$\hat{\sigma}^* = -\frac{\hat{\mu}^*}{q_{\hat{p}}^{(Z)}} \tag{6.9}$$

Using $\hat{p}$ from the Binary pattern calculation in Table 41, we obtain the corrected normal parameters. We insert these parameters into the expectation calculation where $x \in [0, \infty)$, which we use as the lower and upper bounds of the integral.

$$\begin{aligned}
\bar{x} &= E[X\mathbf{1}_{\{X>0\}}] \\
&= \int_0^\infty z \frac{1}{\sqrt{2\pi(\hat{\sigma}^*)^2}} e^{-\frac{1}{2}\left(\frac{z-\hat{\mu}^*}{\hat{\sigma}^*}\right)^2} dz \\
&= \int_0^\infty (z-\hat{\mu}^*) \frac{1}{\sqrt{2\pi(\hat{\sigma}^*)^2}} e^{-\frac{1}{2}\left(\frac{z-\hat{\mu}^*}{\hat{\sigma}^*}\right)^2} dz + \int_0^\infty \hat{\mu}^* \frac{1}{\sqrt{2\pi(\hat{\sigma}^*)^2}} e^{-\frac{1}{2}\left(\frac{z-\hat{\mu}^*}{\hat{\sigma}^*}\right)^2} dz \\
\text{With } \hat{p} &= \int_0^\infty \frac{1}{\sqrt{2\pi(\hat{\sigma}^*)^2}} e^{-\frac{1}{2}\left(\frac{z-\hat{\mu}^*}{\hat{\sigma}^*}\right)^2} dz \text{ we obtain:} \\
\bar{x} &= \frac{1}{\sqrt{2\pi(\hat{\sigma}^*)^2}} \left(-(\hat{\sigma}^*)^2 e^{-\frac{1}{2}\left(\frac{z-\hat{\mu}^*}{\hat{\sigma}^*}\right)^2} \right) \Bigg|_0^\infty + \hat{\mu}^* \hat{p} \\
&= \frac{(\hat{\sigma}^*)^2}{\sqrt{2\pi(\hat{\sigma}^*)^2}} e^{-\frac{1}{2}\left(\frac{\hat{\mu}^*}{\hat{\sigma}^*}\right)^2} + \hat{\mu}^* \hat{p}.
\end{aligned}$$

Using $\hat{\sigma}^*$ in Equation (6.9), we obtain

$$\begin{aligned}
\bar{x} &= \frac{\left(-\frac{\hat{\mu}^*}{q_{\hat{p}}^{(Z)}}\right)^2}{\sqrt{2\pi\left(-\frac{\hat{\mu}^*}{q_{\hat{p}}^{(Z)}}\right)^2}} \exp\left[-\frac{1}{2}\left(\frac{\hat{\mu}^*}{-\frac{\hat{\mu}^*}{q_{\hat{p}}^{(Z)}}}\right)^2\right] + \hat{\mu}^* \hat{p} \\
\bar{x} &= \frac{\hat{\mu}^*}{q_{\hat{p}}^{(Z)}\sqrt{2\pi}} e^{-\frac{1}{2}\left(q_{\hat{p}}^{(Z)}\right)^2} + \hat{\mu}^* \hat{p} \\
\bar{x} &= \hat{\mu}^* \left(\frac{1}{q_{\hat{p}}^{(Z)}\sqrt{2\pi}} e^{-\frac{1}{2}\left(q_{\hat{p}}^{(Z)}\right)^2} + \hat{p} \right) \qquad (6.10) \\
\hat{\mu}^* &= \bar{x} \left(\frac{1}{q_{\hat{p}}^{(Z)}\sqrt{2\pi}} e^{-\frac{1}{2}\left(q_{\hat{p}}^{(Z)}\right)^2} + \hat{p} \right)^{-1} \qquad (6.11)
\end{aligned}$$

Using these equations, we obtain the corrected parameters that we use for a new simulation using Normal distribution. The corrected parameters, the binary proportion, and the $q_{\hat{p}}^{(Z)}$ are shown in Table 47.

Table 47: Corrected Parameteres for Normal Distribution per Province

Province	$\hat{\mu}^*$	$\hat{\sigma}^*$	$\hat{p}$	$q_{\hat{p}}^{(Z)}$
Aceh	-4.55	17.46	0.40	0.26
Sumatera Utara	-1.10	19.93	0.48	0.06
Sumatera Barat	4.70	27.15	0.57	-0.17
Riau	-0.37	19.53	0.49	0.02
Jambi	-1.46	18.56	0.47	0.08
Sumatera Selatan	0.88	19.70	0.52	-0.04
Bengkulu	3.07	22.46	0.55	-0.14
Lampung	-3.10	18.27	0.43	0.17
Kepulauan Bangka Belitung	5.93	16.74	0.64	-0.35
Kepulauan Riau	0.44	18.82	0.51	-0.02
DKI Jakarta	-0.86	22.19	0.48	0.04
Jawa Barat	1.55	15.83	0.54	-0.10
Jawa Tengah	-5.42	21.69	0.40	0.25
DI Yogyakarta	-5.49	22.50	0.40	0.24
Jawa Timur	-8.00	21.07	0.35	0.38
Banten	-6.95	19.98	0.36	0.35
Bali	-7.92	22.12	0.36	0.36
Nusa Tenggara Barat	-12.09	22.53	0.30	0.54
Nusa Tenggara Timur	-11.21	20.20	0.29	0.55
Kalimantan Barat	2.24	20.35	0.54	-0.11
Kalimantan Tengah	11.13	15.02	0.77	-0.74
Kalimantan Selatan	5.65	16.19	0.64	-0.35
Kalimantan Timur	1.12	16.33	0.53	-0.07
Kalimantan Utara	2.34	16.69	0.56	-0.14
Sulawesi Utara	4.43	14.21	0.62	-0.31
Sulawesi Tengah	0.43	18.53	0.51	-0.02
Sulawesi Selatan	-3.72	27.82	0.45	0.13
Sulawesi Tenggara	0.49	16.85	0.51	-0.03
Gorontalo	-3.53	16.29	0.41	0.22
Sulawesi Barat	-5.42	18.88	0.39	0.29
Maluku	1.98	16.52	0.55	-0.12
Maluku Utara	-2.80	18.88	0.44	0.15
Papua Barat	5.53	16.48	0.63	-0.34
Papua	2.01	16.13	0.55	-0.12

We obtain $\hat{p}$ from Table 41. As we can see, half of the provinces has negative mean. Aside from that, the standard deviations are quite big too compared to the data's original standard deviations. Having negative mean will definitely generate negative values for Normal distribution. The same as having low positive mean value with high standard deviations. Seeing that we will definitely have negative values after these data generations, this means changing negative values to zero is still necessary for this simulation. Hence, it is also important for us to look at how different the values will look like if we change all negative values to zero.

Table 48: Summary of Randomized Daily Data for Corrected Normal

Statistic	**C.Normal**	**C.Normal (≥ 0)**
Min	-136.25	0
Q1	-13.86	0
Median	-0.51	0
Mean	-0.88	7.48
Q3	12.47	12.47
Max	137.36	137.36
SD	20.07	11.25
0's	0	11412608
0's per Prov & Sim	0	3356.65

Table 48 shows the summary of the generated values that we obtain using Corrected Normal parameters. We have the original values (C.Normal) and the positive-only values (C.Normal ≥ 0). We can see a shift in the means to be correct means (from -0.88 to 7.48) and standard deviations (from 20.07 to 11.25). Minimum value and the first quartile also change due to them being negative, but aside from that, the other statistics are the same. The huge number of zero values shows how many negative values we generate from the Corrected Normal parameters. Compared to the number of zero values in Table 43, we have double the number of negative values here. Judging from the parameters, however, it makes sense for Corrected Normal to have more negative values than the usual Normal. Note that for a negative mean, the cut off Normal distribution has a decreasing probability density formula which makes it less attractive as alternative to Exponential or Gamma distribution.

Figure 58 shows the error values of Corrected Normal along with Exponential, Normal, and Gamma that are obtained from Figure 55. No changes on the three distributions because we only put them in the same graph as Corrected Normal for comparison. Here, we can clearly see that Corrected Normal has slightly lower error values compared to Normal, which is what we expected. The error values are still not as low as Exponential, but this confirms that adjusting the normal parameters leads to a higher forecast accuracy.

Figure 58: RMSE and MAE Values between Distributions + Corrected Normal

6.4 Simulation with Lognormal Distribution

Aside from Exponential and Gamma, there is also Lognormal distribution that works for positive values. However, since this distribution doesn't work well to fit datasets that have a lot of zero values, we use Lognormal distribution only with the Binary pattern. We use $\hat{\mu}_{LN}$ and $\hat{\sigma}_{LN}$ as our parameters for the Lognormal distributions to distinguish it with the Normal parameters $\hat{\mu}$ and $\hat{\sigma}$. We estimate the two parameters as follows.

$$\hat{\mu}_{LN} = \frac{1}{n}\sum_{i=1}^{n} \ln(x_i) \tag{6.12}$$

$$\hat{\sigma}_{LN} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(\ln(x_i) - \hat{\mu}_{LN}\right)^2} \tag{6.13}$$

Aside from Lognormal Distribution, we also want to compare the results with Corrected Normal Parameters under the Binary assumption. Hence, we need to slightly adjust $\hat{\mu}^*$ in Equation (6.11) using conditional probability in order to obtain $\hat{\mu}^{**}$.

$$\bar{x} = E[X|X>0] = \frac{E[X\mathbf{1}_{\{X>0\}}]}{P(X>0)}.$$

Since we have $E[X\mathbf{1}_{\{X>0\}}]$ from Equation (6.10), and $P(X>0)$ from Equation (6.7), we can substitute them as:

$$\bar{x} = \frac{\hat{\mu}^{**}\left(\frac{1}{q_{\hat{p}}^{(Z)}\sqrt{2\pi}}e^{-\frac{1}{2}\left(q_{\hat{p}}^{(Z)}\right)^2} + \hat{p}\right)}{\hat{p}}$$

$$\hat{\mu}^{**} = \frac{\bar{x}}{\hat{p}}\left(\frac{1}{q_{\hat{p}}^{(Z)}\sqrt{2\pi}}e^{-\frac{1}{2}\left(q_{\hat{p}}^{(Z)}\right)^2} + \hat{p}\right)^{-1}$$

$$\hat{\mu}^{**} = \frac{\mu^*}{\hat{p}} \tag{6.14}$$

$$\hat{\sigma}^{**} = -\frac{\hat{\mu}^{**}}{q_{\hat{p}}^{(Z)}} \tag{6.15}$$

Table 49: Lognormal and Corrected Normal Parameteres per Province (Binary)

Province	$\hat{\mu}_{LN}$	$\hat{\sigma}_{LN}$	$\hat{\mu}^{**}$	$\hat{\sigma}^{**}$
Aceh	1.69	1.33	-11.44	43.95
Sumatera Utara	1.68	1.60	-2.30	41.69
Sumatera Barat	2.20	1.52	8.27	47.74
Riau	1.39	1.80	-0.75	39.66
Jambi	1.55	1.64	-3.12	39.61
Sumatera Selatan	1.83	1.53	1.70	38.04
Bengkulu	2.02	1.49	5.54	40.52
Lampung	1.69	1.45	-7.17	42.23
Kepulauan Bangka Belitung	1.80	1.54	9.29	26.21
Kepulauan Riau	1.62	1.58	0.86	36.96
DKI Jakarta	1.93	1.49	-1.79	45.80
Jawa Barat	1.68	1.47	2.88	29.37
Jawa Tengah	1.79	1.51	-13.51	54.06
DI Yogyakarta	1.79	1.54	-13.61	55.75
Jawa Timur	1.80	1.41	-22.72	59.85
Banten	1.51	1.60	-19.09	54.88
Bali	1.80	1.45	-21.99	61.41
Nusa Tenggara Barat	1.84	1.37	-40.90	76.19
Nusa Tenggara Timur	1.54	1.53	-38.73	69.79
Kalimantan Barat	1.75	1.64	4.12	37.43
Kalimantan Tengah	1.78	1.61	14.44	19.49
Kalimantan Selatan	1.86	1.45	8.88	25.44
Kalimantan Timur	1.72	1.44	2.12	30.97
Kalimantan Utara	1.75	1.46	4.21	30.04
Sulawesi Utara	1.81	1.29	7.11	22.83
Sulawesi Tengah	1.90	1.33	0.85	36.37
Sulawesi Selatan	2.25	1.38	-8.33	62.27
Sulawesi Tenggara	1.80	1.36	0.96	32.93
Gorontalo	1.68	1.31	-8.51	39.32
Sulawesi Barat	1.59	1.47	-14.00	48.77
Maluku	1.82	1.33	3.62	30.16
Maluku Utara	1.85	1.34	-6.35	42.81
Papua Barat	1.68	1.57	8.76	26.10
Papua	1.72	1.42	3.65	29.35

Table 49 shows the parameters that we obtain for Lognormal and Corrected Normal using Binary. We see that the Lognormal parameters are mostly alike, in the range of 1 until 3. As for the Corrected Normal distribution, all parameters are around twice the parameters in Table 47.

Table 50: Summary of Randomized Daily Data for Lognormal and Binary Corrected Normal

Statistic	**C.Normal**	**C.Normal** (≥ 0)	**Lognormal**
Min	-395.70	0	0
Q1	0	0	0
Median	0	0	0
Mean	-0.87	7.48	8.89
Q3	0.09	0.09	5.70
Max	316.46	316.46	42605.10
SD	29.92	17.66	41.04
0's	11412649	16754602	11412649
0's per Prov & Sim	3356.65	4927.82	3356.65

Table 50 shows the summary of generated values from Corrected Normal and Lognormal under the Binary pattern. We also have two versions of Corrected Normal here: the originally generated values and the positive only values. Just like what we see in Table 48, the change of statistics from Corrected Normal to the adjusted values only occurs to minimum, mean, and standard deviation values. The first quartile, median, third quartile, and maximum values stay the same. The change of standard deviation (from 29.92 to 17.66) and the change for mean (from -0.97 to 7.48) are more or less similar, which is around 8. As for Lognormal, it has the same minimum value and first quartile as Adjusted Corrected Normal, but the other statistics are higher. The maximum value is very high, reaching 42000. This does not really make sense for a daily rainfall value, thus the heavy-tailed Lognormal distribution does not seem to be a good option. As for the zero values, both distributions show that the zero values come from the Binary pattern, since Lognormal and Corrected Normal have the same number of zero values.

Figure 59 shows the error values of each distribution under the Binary distribution. The values for Exponential, Normal, and Gamma are obtained from Figure 55. As seen here, Corrected Normal works similarly to Normal, but the error values are slightly higher than Normal. Meanwhile, Lognormal's error values are very high. especially in USSA. Under USSA-Row-M/Vec and USSA-Col-M/Vec the error values exceed 300 by a lot, which shows that Lognormal doesn't work well to simulate rainfall values. As for Corrected Normal, it looks like it works better not under the Binary assumption. Even compared to Corrected Normal error values in Figure 58, the Binary one still has higher error values.

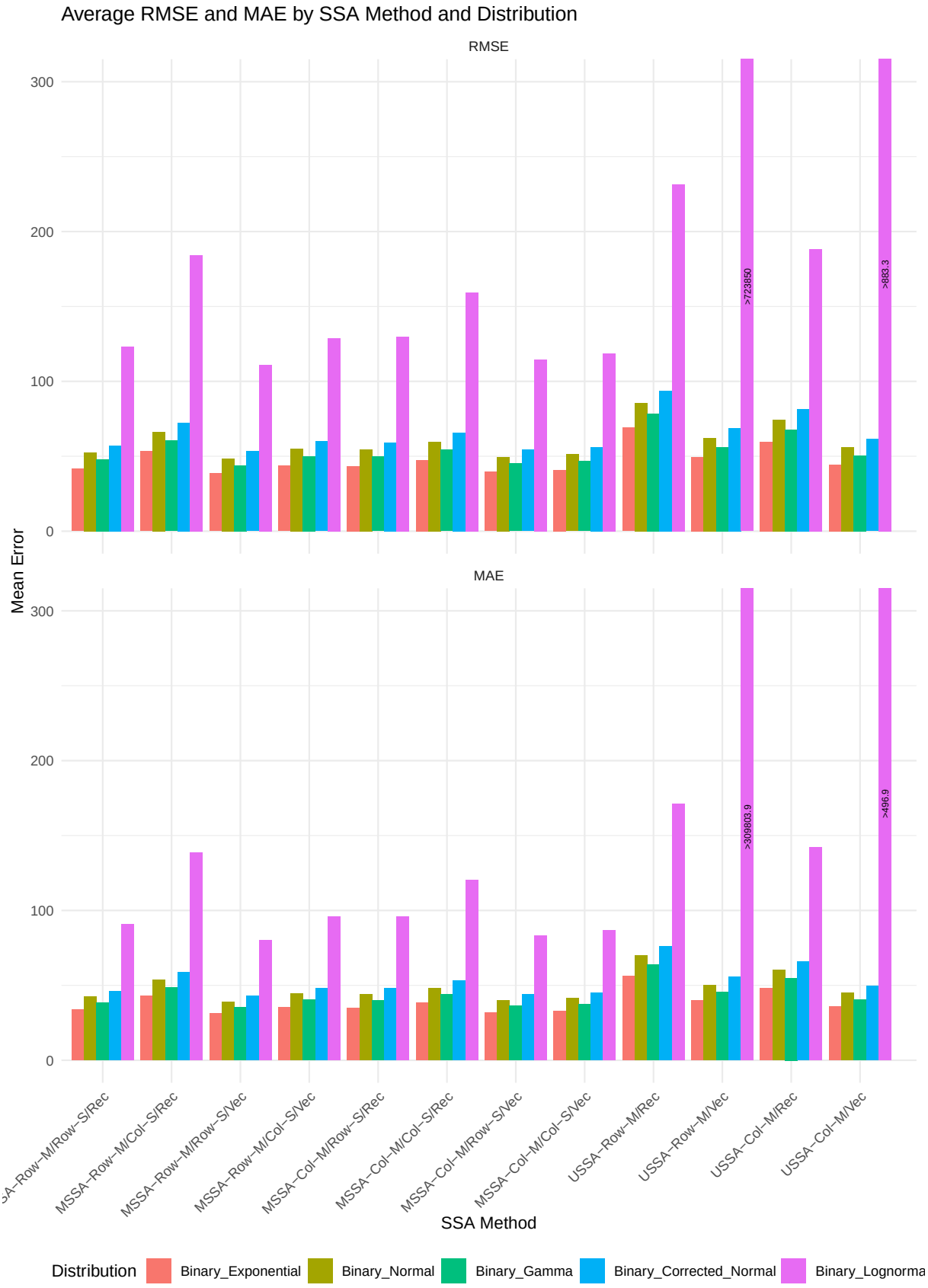


Figure 59: RMSE and MAE Values between Distributions + Corrected Normal and Lognormal (Binary)

6.5 Simulation with Different Distributions Every Year

Each month of each province must have different rainfall patterns. That is what started this simulation. By checking the distribution for each month and each province, we can get more accurate results. We use an R function called `simukde::find_best_fit` to obtain the distribution for each month and each province. We use Bootstrap daily data for the distribution fitting so that there would be no missing values. Unfortunately, checking monthly distribution is proven to be not possible due to some months having all zero values (no rainfall). Hence, we decided to change the distribution fitting from monthly to yearly. This package proposes that some series has Weibull distributions with shape parameter $\hat{\eta}$ and scale parameter $\hat{k}$. Hence, we need the formulas to estimate the parameters as follows.

$$\hat{\eta} = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^{\hat{k}} \right)^{\frac{1}{\hat{k}}} \tag{6.16}$$

We need $\hat{k}$ to obtain $\hat{\eta}$. In order to get $\hat{k}$, we iterate Equation (6.17) until we obtain the $\hat{k}$ that matches the equation.

$$0 = \frac{\sum_{i=1}^{n} x_i^{\hat{k}} \ln x_i}{\sum_{i=1}^{n} x_i^{\hat{k}}} - \frac{1}{\hat{k}} - \frac{1}{n} \sum_{i=1}^{n} \ln x_i \tag{6.17}$$

The obtained parameters for each year and each province can be seen in Appendix. With 34 provinces and 18 years of data, we have 612 different series to be fitted into various distributions. Aside from Weibull (12 series), the R function shows that some province's yearly rainfall series fit Exponential (10 series), Gamma (41 series), and Normal (264 series). There are 285 series that can't be tested by the R function. Hence, we consider the 285 series to fit Gamma. After generating the values using these distributions and the parameters accordingly, we obtain the daily rainfall values.

Table 51: Summary of Randomized Daily Data for Yearly Distribution

Statistic	**Value**
Min	0
Q1	0
Median	0.11
Mean	9.32
Q3	11.52
Max	28258.43
SD	329.81
0's	2885904
0's per Prov & Sim	848.79

The summary of the values obtained from the data generation with different distributions is shown in Table 51. First of all we see that the estimation provided by the R-Package is not based on fitting only the mean exactly as we did for most of the distributions above. But overall, these statistics are almost similar to Lognormal (see Table 50). The maximum value is not as high as Lognormal, but it is still very high considering that this value represents daily rainfall intensity. The number of zero values are also quite a lot, although not as many as Binary pattern.

Figure 60 shows the error values of the Yearly Distribution compared to the Monthly simulations. The summary for the monthly simulation values can be seen in Table 45. Yearly

Distribution has lower error values compared to the monthly calculations, showing that with more proper distributions, the forecasting can be more accurate. It shows that yearly distributions provide more accurate forecast results than monthly distributions.

Aside from this, we would also like to compare the Yearly Distribution with the usual Exponential, Normal, and Gamma simulations (without Binary and without Monthly). The summary for these generated values can be seen in Table 43. We have the comparison between these values and the yearly distributions in Figure 61. Unfortunately, compared to these three, Yearly Distribution has higher error values. This figure shows that using one distribution for the entire series provides more accurate forecast results compared to yearly distributions.

6.5.1 Comparison with Other Distributions Using Yearly Parameters

Comparing Yearly Distribution with Monthly simulations just like what we did in Figure 60 is not an apple-to-apple comparison, which is why in this part, we also simulate the Yearly simulations using Exponential, Normal, and Gamma distributions.

The results are quite interesting (see Figure 62). As we can see, there are different behaviors in USSA and MSSA. In MSSA, one distribution works better than different distribution per year per province. We can see that Exponential has the lowest error values, followed by Normal, Gamma, and then the Yearly Distribution. Although sometimes Gamma has higher error values than Yearly Distribution.

On the other hand, Yearly Distribution work better in USSA. The error values are similar to MSSA, but since the other distributions have much higher error values, Yearly Distribution now has the lowest error values. The order of the lowest to highest is actually the complete opposite: Yearly Distribution, Gamma, Normal, and Exponential. In some cases, Normal and Gamma are quite similar, but Gamma is generally lower.

This comparison shows that if we want the calculation to be simple and assume that all provinces have the same distribution, it is better to use MSSA. However, if we want to assume that every province has different distributions (which is more realistic), it is better to use USSA.

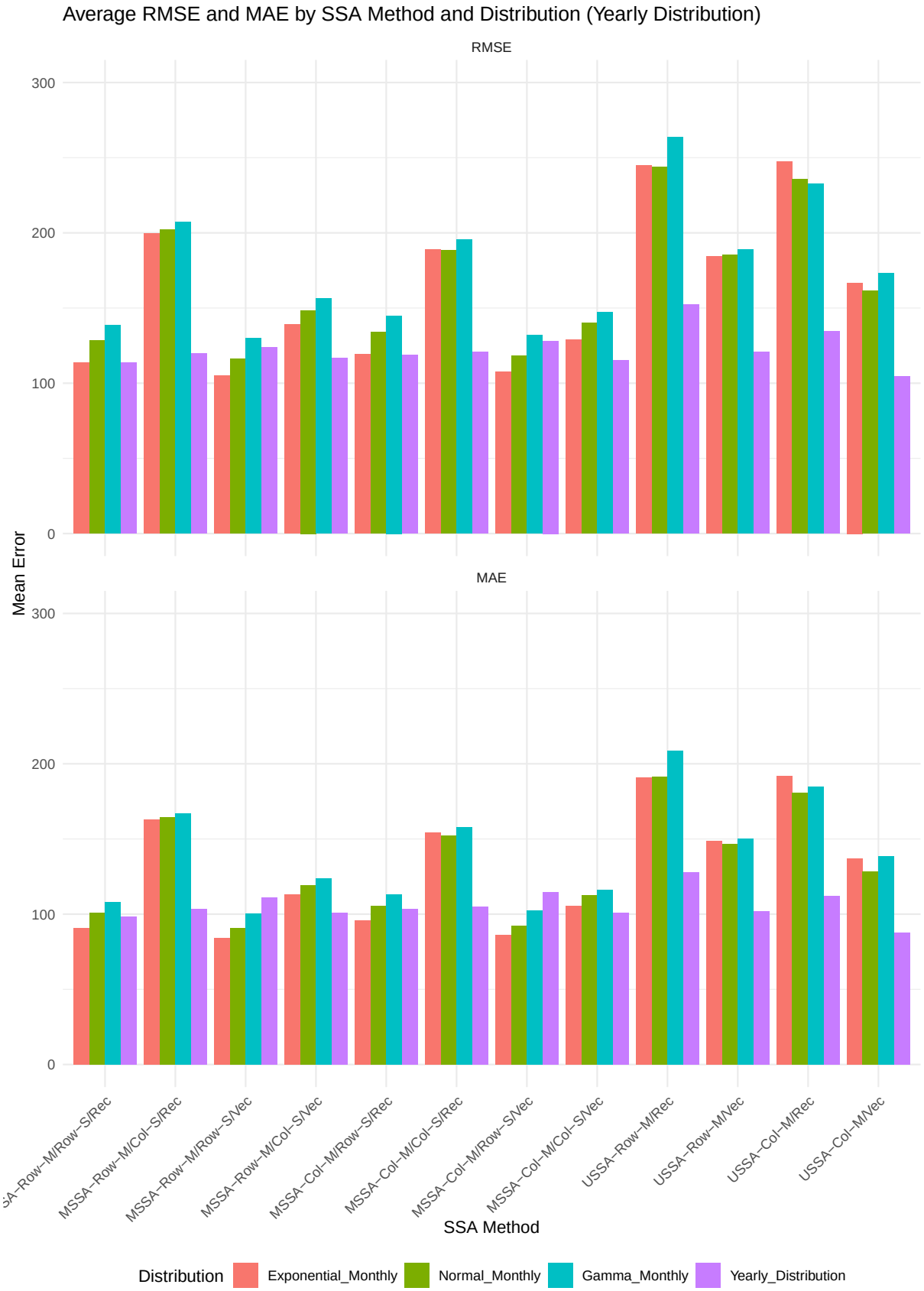


Figure 60: RMSE and MAE Values between Distributions with Monthly and Yearly Parameters

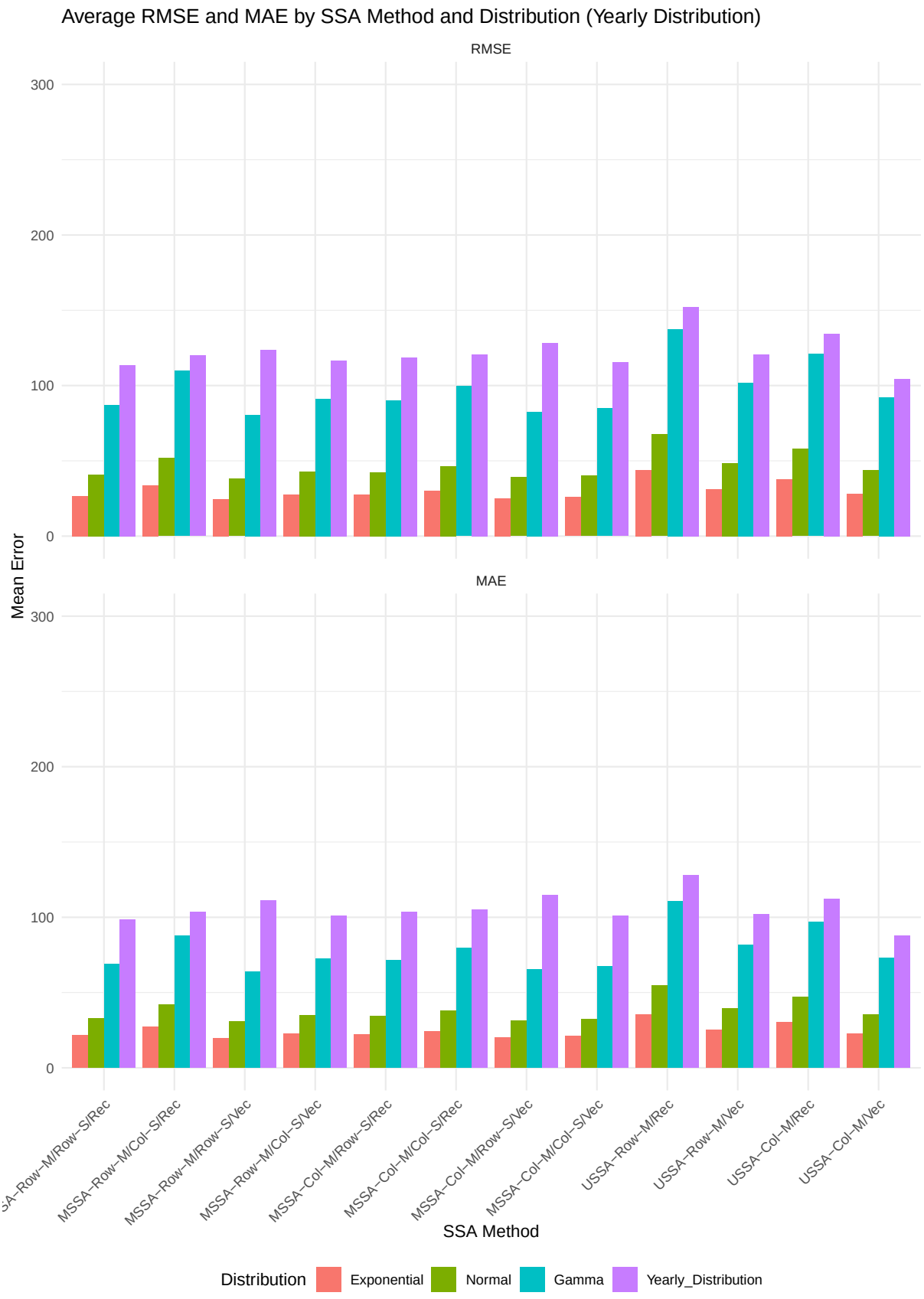


Figure 61: RMSE and MAE Values between Distributions with Entire Series and Yearly Parameters

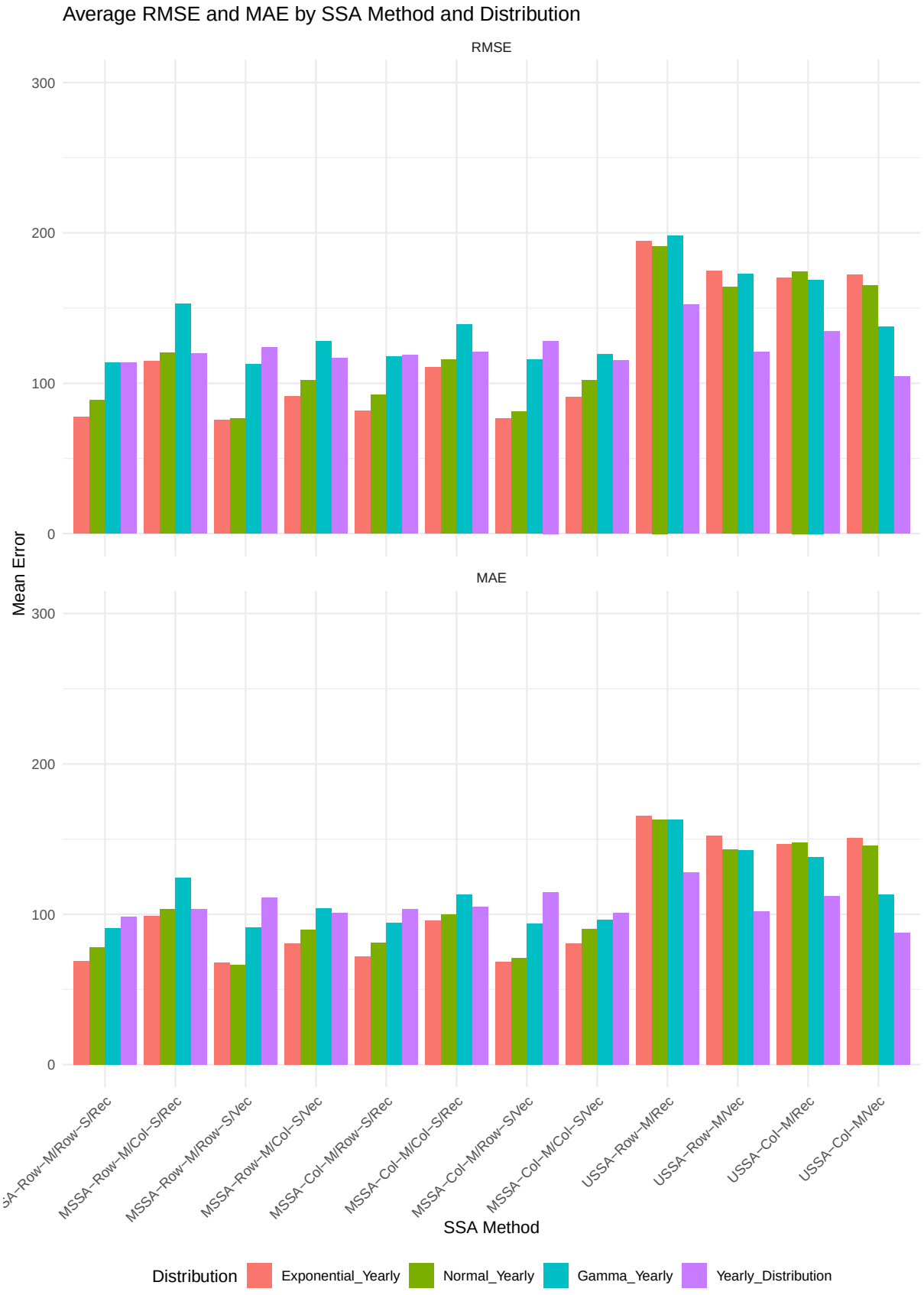


Figure 62: RMSE and MAE Values between Distributions with Yearly Parameters

6.6 Simulation with Scaled Mean (Seasonality)

We attempted another simulation. This time, we do not use randomly generated values and we do not run it for 100 times. Instead, we use the overall mean of each province, each month, each year, and the entire values of Bootstrap-imputed datasets. Using these values, we obtain the mean of each month per year per province, which we just need to aggregate monthly and apply SSA as usual. The formula to obtain the mean is as follows.

$$\bar{X}^{p,m,y} = \bar{X}^{p} \times \frac{\bar{X}^{m}}{\bar{\bar{X}}} \times \frac{\bar{X}^{y}}{\bar{\bar{X}}} \tag{6.18}$$

Using Equation (6.18), we are able to obtain the scaled rainfall mean for every province, every month, and every year which supposedly capture the seasonality of the rainfall values. Afterwards, we multiply these values by 28, 29, 30, or 31 according to the month and the year. We aggregate the values as usual, and we finally have the entire series of daily rainfall values.

Table 52: Summary of Randomized Daily Data for Scaled Mean (Seasonality)

Statistic	**Value**
Min	2.01
Q1	5.32
Median	6.98
Mean	7.51
Q3	9.19
Max	21.96
SD	3.02
0's	0

According to Table 52, Seasonality has the lowest range compared to the other series in the entire simulation. Seasonality has the lowest mean and the only minimum value that is not negative or close to zero. Due to the calculation using scaled mean from all the provinces, months, and years, we also have no zero values in Seasonality. Even so, the series still consists of mostly low values (less than 22).

The error values for Seasonality are compared with the yearly simulations (see Figure 62). It is obvious that the error values are much lower than the other methods. Aside from the possibility that Seasonality provides more accurate forecast results, the small values also play a huge role in the calculation.

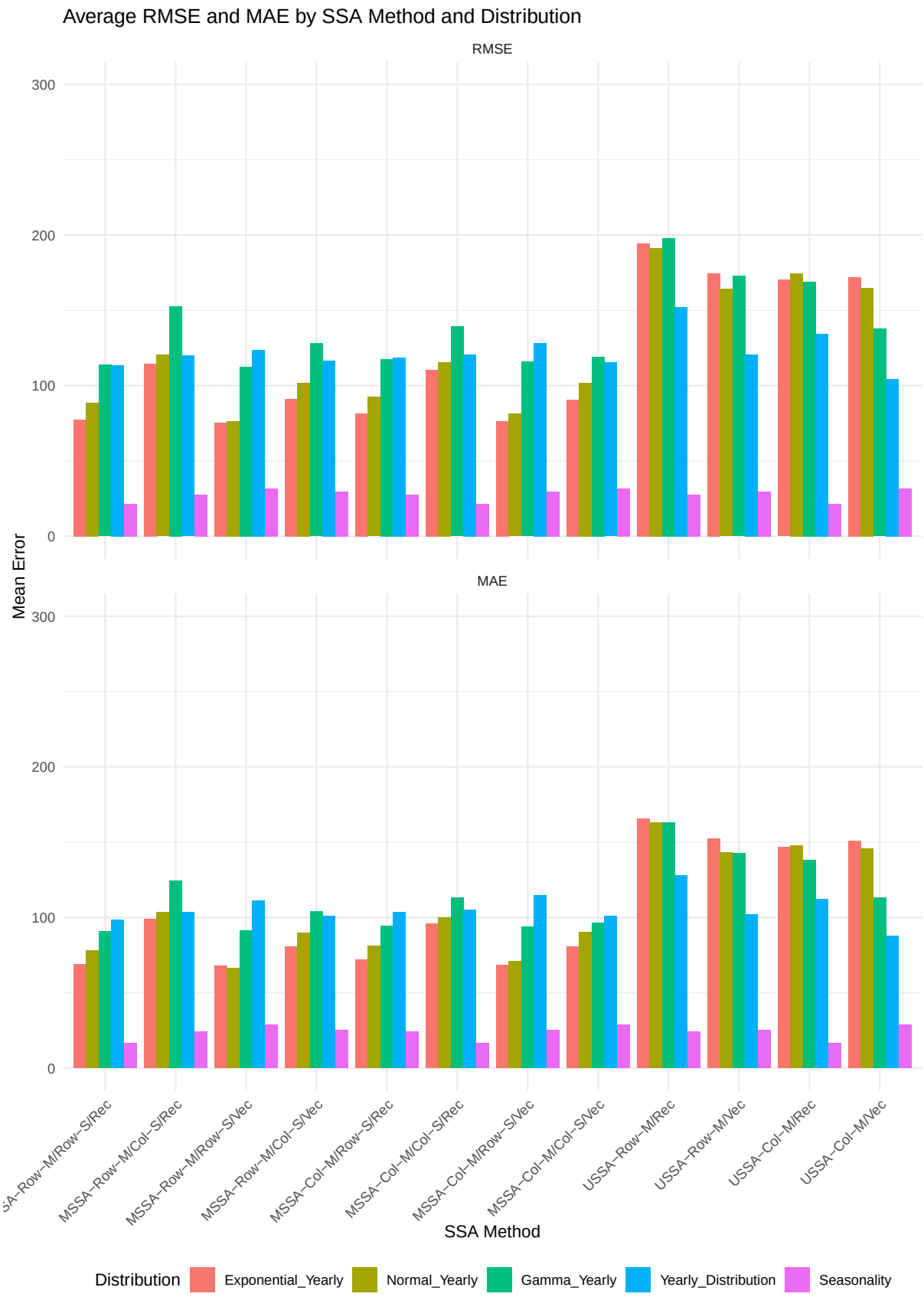


Figure 63: RMSE and MAE Values between Distributions with Seasonality

Aside from the comparison between error values, we also want to see the seasonality of each month based on the mean values. Figure 64 shows the average values per month. The thick black line with dots represents $\bar{X}^m$, which is the mean values per month from the entire series. This figure also shows various lines with different colors which represent the mean values per month per year from the entire series, notated as $\bar{X}^{m,y}$, which was not a part of Equation (6.18). We create intervals by using the minimum and maximum values among these mean values.

$$\bar{X}^m_{upper} = \max\left(\bar{X}^{m,y}\right) \tag{6.19}$$

$$\bar{X}^m_{lower} = \min\left(\bar{X}^{m,y}\right) \tag{6.20}$$

with $m = \{1, 2, ..., 12\}$ and $y = \{2000, 2001, ..., 2015, 2016, 2017\}$.

According to $\bar{X}^m$, the rainfall values are mostly pretty high in January, and then it slowly decreases and reaches the low point in August. Starting September, it goes up again until December, which has the highest mean value. According to the intervals, other years also have that pattern. Most of them have high rainfall values in the beginning of the year, followed by the lowest point in August, and lastly, the rainfall values go up again until the end of the year. This pattern is consistent to the fact that the rainy season in Indonesia usually lasts from October to March. All in all this part shows how one may introduce seasonality in a straightforward way into the simulation. Evidently that improves the simulation result considerably.

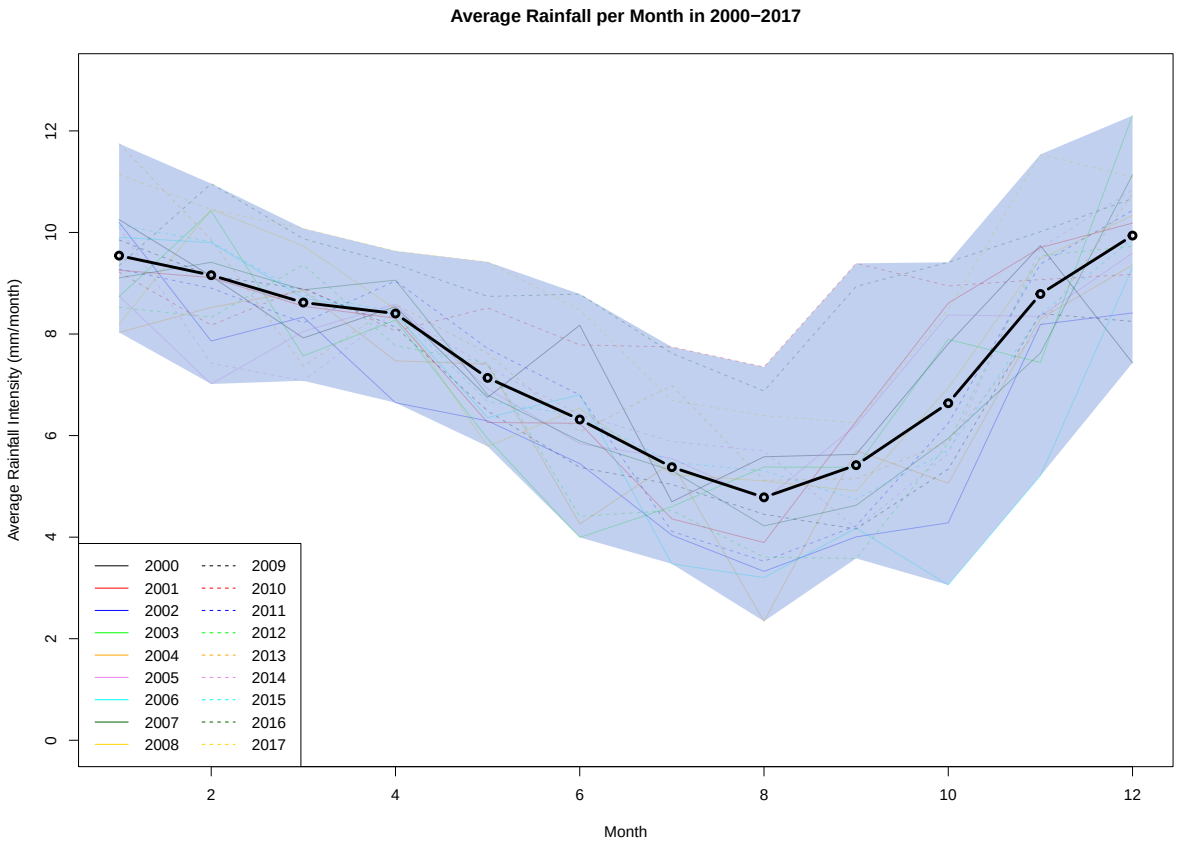


Figure 64: Average Daily Rainfall Values per Month in 2000-2017

7 Conclusion

In this chapter, we summarize all our findings and relate them to our research aims and focus. We also discuss the limitations and the recommendations from this thesis for further research in the future.

Weather components, specifically rainfall intensity, are one of the most important factors that contribute to the growth of paddy rice crops in Indonesia. By forecasting rainfall intensity, we are also able to forecast the possible damages to paddy rice crops and estimate the total payouts that the insurance company has to prepare. This work aims to contributing to more risk-related research in Indonesia using climate information and actuarial assumptions. However, in the process, there are several obstacles that need to be tackled, such as incomplete datasets and a lot of possible scenarios that might affect the outcome.

Hence, the first step is dealing with these obstacles. Our first obstacle are the missing values. We use 12 different imputation methods to fill in the gaps in the data, both single-value imputations and multiple imputation. After imputing all necessary values, we obtain different results that can picture various situations in real life. If there are mostly months with high rainfall intensity, it is best to use Mean Substitution, Last Observation Carried Forward Imputation (LOCF), or Linear Interpolation. On the other hand, if there are mostly months with low rainfall intensity, it is best to use Median Substitution, Null Substitution, or most Multiple Imputation by Chained Equations (MICE). However, if the rainfall intensity is more well-balanced within those months without any extremely high or low values, it is best to use Bootstrap, Distribution Fill, or MICE-SAMPLE (one of the approaches in MICE).

Once all datasets are imputed, the second procedure is to group them into four different area categories: Province, Region, Country, and Cluster. The datasets are set per Province by default, which means there is no need to adjust the data for Province calculation. As for the other categories, we use weighted values according to the area size of each province to form one series to represent each group of the category. We also use two different clustering techniques to obtain two sets of clusters for comparison purposes. With 34 provinces, 7 regions, 7 clusters and two types of clustering, and 1 country, we have 56 different series per imputation method. Through this process, we managed to obtain various rainfall intensity series with some similarities, but also with extreme differences. Interestingly, there are also some grouping cases that help lessen the extreme values obtained from the imputation process.

The third procedure is using these series as input of the Singular Spectrum Analysis (SSA) that also has several variations. There are 2 types of variable inputs (Univariate and Multivariate), 2 types of trajectory matrix forms (Row and Column), 2 types of matrix stacks for the Multivariate version (Row and Column), and 2 types of forecasting methods (Recurrent and Vector). In total, there are 4 different Univariate SSA methods and 8 different Multivariate SSA methods that can be used for all the series obtained from the grouping and the imputation. After going through all the calculations for each series, it turns out that the methods that perform the best for forecasting, seen from their error values, are USSA-Col-M/Vec for univariate, along with MSSA-Row-M/Row-S/Rec and MSSA-Row-M/Row-S/Vec for multivariate. In terms of data reconstruction, most methods perform well, but the differences between method performances are more obvious when we compare the forecasting error values. Overall, Vector forecast seems to work better in comparison to Recurrent forecast. MSSA also performs better than USSA, although with specific parameters, the differences might not be significant. Our results for different L and d show that indeed d must not be chosen too high in order to avoid an overfit and to get better forecast results.

Converting the forecasted values of the rainfall intensity to insurance payout is the fourth

procedure. There are five different linear model that we use for all forecasted values based on the weather condition in different locations in Indonesia. All these values have the same pattern, only with different limits and thresholds. All these models perform similarly, resulting in U-shaped Beta distribution with $\alpha < 1$ and $\beta < 1$. The only differences are in how high the payout values in average according to several evaluation criterias: Mean, Standard Deviation (SD), Value-at-Risk (VaR), Conditional Tail Expecation (CTE), the 75th Percentile, and the Difference between Mean and Median (DMM). Design 3 and Design 4 seem to be the best options according to the evaluation criterias, but Design 1 and Design 5 are better when it comes to the distribution.

After all these processes are done, a simulation study is used to show that the results are robust and to compare with simple forecasts just estimating several rainfall distributions. However, the simulation does not go through the missing values imputation process, only starting from grouping until payout conversion. We generate three datasets using different distributions: Exponential, Normal, and Gamma. Afterwards, we mix and match the datasets accordingly. We have Binary calculations where we only consider the days with rainfall, Monthly calculation where we estimate the parameters of each province every month, and a mix of Binary and Monthly. Aside from these three distributions, we also take some other distributions into account: Lognormal, Corrected Normal, and Weibull. We also do another simulation by assuming different distributions for every province every year and another simulation where we consider the overall mean values for provinces, months, and years to create a new simulated series. Each process is run 100 times and the summary is taken based on the average values. Overall, the results are quite consistent. The best SSA methods are USSA-Col-M/Vec for univariate and MSSA-Row-M/Row-S/Vec for multivariate, which is consistent with the previous SSA calculations. Meanwhile, the best payout design in the simulation part is Design 3, which is consistent with the evaluation criteria, but not with the payout distribution.

We faced several limitations while conducting this research, including data limitations and memory issues. Unfortunately, comparison with real data payouts was not possible due to the inaccessibility of the actual claim data from the insurance company. Simulations were also not running smoothly due to computer memory exhaustion problems. Hence, more simulations were not possible for the research.

There are still rooms to expand this research using available techniques and data. For instance, a machine learning-based imputation can be attempted to fill in the missing values, along with using SSA not only for rainfall values, but also for other factors that are related to climate and paddy rice crops. On the other hand, based on the ideas in the simulation study, one could build a fully parametric model for the rainfall and compare this to the non-parametric SSA approach in detail. Different payout models can also be used to develop different payout scenarios from which both the insurance company and the farmers may benefit. If possible, obtaining the actual payout data from the insurance company can also help the research to be more accurate and more usable for the future.

References

[APD23] Adriana L. Abrego-Perez, Natalia Pacheco-Carvajal, and Maria C. Diaz-Jimenez. "Forecasting Agricultural Financial Weather Risk Using PCA and SSA in an Index Insurance Model in Low-Income Economies". In: *Applied Science* 13.4 (2023), p. 2425. DOI: https://doi.org/10.3390/app13042425.

[Ard23] Prisma Ardianto. *Ramai-Ramai Klaim Asuransi Gagal Panen*. 2023. URL: https://investor.id/finance/343727/ramairamai-klaim-asuransi-gagal-panen.

[Car+18] R E Caraka et al. "Analysis of plant pattern using water balance and cimogram based on oldeman climate type". In: *IOP Conference Series: Earth and Environmental Science* 195.1 (2018), p. 012001. URL: https://doi.org/10.1088/1755-1315/195/1/012001.

[Geo] Badan Metereologi Klimatologi dan Geofisika (BMKG). *Probabilistik Curah Hujan*. URL: https://www.bmkg.go.id/cuaca/probabilistik-curah-hujan.bmkg.

[Gho+17] Mansi Ghodsi et al. "Vector and recurrent singular spectrum analysis: which is better at forecasting?" In: *Journal ofApplied Statistics* (2017), pp. 1872–1899. DOI: http://dx.doi.org/10.1080/02664763.2017.1401050.

[Gol20] Nina Golyandina. "Particularities and commonalities of singular spectrum analysis as a method of time series analysis and signal processing". In: *WIREs Computational Statistics* 12.4 (2020), e1487. ISSN: 1939-0068. DOI: 10.1002/wics.1487. URL: http://dx.doi.org/10.1002/wics.1487.

[GZ13] Nina Golyandina and Anatoly Zhigljavsky. *Singular Spectrum Analysis for Time Series*. Heidelberg: Springer, 2013. ISBN: 978-3-662-62436-4.

[Ind18] Menteri Pertanian Republik Indonesia. *Keputusan Menteri Pertanian Republik Indonesia Nomor 30/Kpts/SR.210/B/12/2018*. 2018. URL: https://psp-pertanian-go-id.webpkgcache.com/doc/-/s/psp.pertanian.go.id/storage/125/Pedoman-Bantuan-Premi-Asuransi-Usahatani-Padi-Tahun-2019.pdf.

[Ind23] Menteri Pertanian Republik Indonesia. *Peraturan Menteri Pertanian Nomor 30 Tahun 2023*. 2023. URL: https://peraturan.bpk.go.id/Download/320056/Permentan%20Nomor%2030%20Tahun%202023.pdf.

[Kan13] Hyun Kang. "The prevention and handling of the missing data". In: *Korean Journal of Anesthesiology* 64.5 (2013), pp. 402–406. URL: https://doi.org/10.4097/kjae.2013.64.5.402.

[Keu23] Otoritas Jasa Keuangan. *Peraturan Menteri Pertanian Nomor 30 Tahun 2023*. 2023. URL: https://sikapiuangmu.ojk.go.id/FrontEnd/CMS/Article/10525.

[Lou10] J. E. Louhenpessy. *Sagu Harapan dan Tantangan*. Jakarta: PT Bumi Aksara, 2010. ISBN: 9789790105188.

[LSA15] Peng Li, Elizabeth A. Stuart, and David B. Allison. "Multiple Imputation: A Flexible Tool for Handling Missing Data". In: *The Journal of the American Medical Association* 314.18 (2015), pp. 1966–1967. DOI: https://doi.org/10.1001/jama.2015.15281.

[MRY19] Rahim Mahmoudvand, Paulo Canas Rodrigues, and Masoud Yarmohammadi. "Forecasting Daily Exchange Rates: A Comparison Between SSA and MSSA". In: *REVSTAT – Statistical Journal* 17.4 (2019), pp. 599–616. URL: https://doi.org/10.57805/revstat.v17i4.282.

[Pra+20] E P A Pratiwi et al. "Precipitation and flood impact on rice paddies: Statistics in Central Java, Indonesia". In: *IOP Conference Series: Earth and Environmental Science* 612.1 (2020), p. 012040. DOI: 10.1088/1755-1315/612/1/012040. URL: https://doi.org/10.1088/1755-1315/612/1/012040.

[REA] READI-Project. *Risk Management, Economic Sustainability, and Actuarial Science Development in Indonesia*. University of Waterloo. URL: https://uwaterloo.ca/risk-management-economic-sustainability-actuarial-science-development-indonesia/sites/ca.risk-management-economic-sustainability-actuarial-science-development-indonesia/files/uploads/files/cc_and_actuaries_in_indonesia_-_research_and_.

[Sta] International Organization for Standardization (ISO). *ID - Indonesia*. URL: https://www.iso.org/obp/ui/#iso:code:3166:ID.

[Sum06] Sumarno. "Periodisasi musim tanam padi sebagai landasan manajemen produksi beras nasional". In: *Sinar Tani* 3136 (2006), pp. 2–3.

Appendices

Appendix A Raw Data and Imputed Daily Rainfall Data

A.1 Raw Data: Daily Rainfall Values for Aceh in Year 2000

Day	Month											
	Jan	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
1	7	0	7	0	0	52	13	0	0	0	0	1
2	4	0	0	-	0	17	1	0	0	24	0	0
3	29	0	0	-	0	3	22	0	0	18	0	0
4	15	0	0	0	0	0	1	0	0	10	0	0
5	21	0	0	0	0	4	0	12	0	0	-	0
6	0	0	0	2	0	0	0	2	0	0	0	0
7	17	3	-	0	0	0	0	1	9	0	0	0
8	0	0	2	-	0	8	1	2	23	0	0	0
9	0	0	0	4	0	-	1	2	2	0	26	11
10	4	0	-	5	0	0	0	0	0	0	6	1
11	24	0	3	2	15	1	0	0	2	0	-	4
12	1	0	0	0	4	12	0	15	0	0	0	116
13	0	0	0	0	0	0	0	0	0	3	0	26
14	0	0	0	-	31	0	0	0	0	-	10	32
15	0	0	0	-	23	1	0	0	0	0	25	1
16	0	6	0	-	21	0	0	0	-	0	-	1
17	0	33	0	0	0	-	0	0	38	-	69	0
18	0	0	1	0	0	-	0	0	-	0	0	5
19	0	1	0	0	0	0	0	2	2	0	22	0
20	6	0	0	0	0	0	0	-	0	2	-	0
21	0	18	0	0	2	0	0	-	5	5	129	9
22	0	12	0	1	15	0	0	0	3	1	141	0
23	0	0	1	0	-	0	0	0	14	-	185	30
24	0	0	-	13	2	0	0	0	-	3	24	0
25	0	0	0	-	0	0	0	0	0	4	4	4
26	2	0	4	1	0	12	-	0	1	-	-	0
27	0	29	26	5	0	1	1	0	0	0	0	0
28	0	0	-	0	0	14	0	0	6	-	0	0
29	0	0	-	0	5	6	-	0	1	0	0	0
30	0	-	2	13	2	5	23	0	0	0	0	0
31	0	-	1	-	2	-	0	0	-	0	-	0

Data for the remaining years, provinces, and imputed values can be accessed through this link: `https://tinyurl.com/desenaldo-appendix-a`

Appendix B Aggregated Monthly Rainfall Values

B.1 Aggregated Monthly Rainfall Values for Aceh using Mean Substitution Imputed Values

Year	Month											
	Jan	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
2000	130.00	102.00	56.04	60.00	126.07	151.11	67.34	38.48	117.78	83.46	769.20	241.00
2001	232.00	66.00	42.00	56.40	71.00	43.00	21.00	21.81	112.50	468.10	114.44	260.00
2002	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2003	181.72	208.00	153.85	184.00	137.29	53.08	66.13	124.00	65.56	332.73	141.11	145.46
2004	52.04	88.04	332.96	90.00	51.67	100.00	25.66	44.00	106.67	127.58	226.80	236.00
2005	241.11	40.00	157.00	68.89	109.61	162.00	37.00	41.00	77.00	206.00	127.00	255.00
2006	177.00	169.12	209.77	102.69	81.93	147.69	62.00	32.03	161.79	130.41	214.00	99.41
2007	173.82	8.96	130.00	288.89	341.00	107.14	79.57	22.00	82.00	128.00	182.00	197.00
2008	162.00	20.00	164.00	0.00	68.41	20.36	54.52	59.79	65.36	73.92	313.45	179.59
2009	306.68	126.56	129.54	219.23	223.20	22.00	8.04	147.56	128.00	44.43	366.92	304.04
2010	186.00	133.00	108.50	227.59	62.00	215.77	126.82	80.82	86.54	115.00	512.22	355.97
2011	169.39	89.38	310.00	153.21	62.00	20.69	57.87	75.29	159.23	54.56	184.44	144.67
2012	102.96	87.00	104.37	87.78	102.30	41.00	32.15	41.69	87.78	135.48	230.77	196.33
2013	548.89	254.05	347.59	265.50	290.29	385.85	288.64	113.85	705.43	159.51	280.88	332.94
2014	163.61	90.84	8.04	124.44	100.75	74.25	38.00	142.71	151.18	498.67	567.44	506.85
2015	171.35	191.80	201.50	678.86	139.50	20.25	135.42	200.09	295.60	473.11	257.40	385.81
2016	231.96	181.48	98.58	288.18	302.25	81.19	82.25	247.82	61.94	366.91	358.00	218.42
2017	515.45	164.97	369.34	344.43	174.54	77.82	85.25	106.43	337.09	273.57	480.95	484.58

Data for the remaining provinces and imputation methods can be accessed through this link:
`https://tinyurl.com/desenaldo-appendix-b`

Appendix C Grouped Monthly Rainfall Values (Weighted Values)

C.1 Grouped Monthly Rainfall Values for Country using Mean Substitution Imputed Values

Year	Month											
	Jan	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
2000	351.06	248.59	246.81	298.08	264.96	303.35	198.13	270.42	256.16	294.55	353.86	227.80
2001	256.12	288.21	310.05	255.14	270.06	210.12	117.89	177.08	296.15	290.36	351.90	357.90
2002	331.76	204.13	307.10	209.74	257.33	232.49	108.39	117.88	129.75	165.97	246.21	267.02
2003	272.57	306.51	266.57	248.29	207.04	136.67	226.20	191.61	262.54	316.31	258.46	437.39
2004	257.49	279.48	304.72	271.56	250.42	204.66	232.40	82.75	257.37	221.09	288.43	294.47
2005	332.49	212.56	278.12	373.02	283.85	221.88	194.09	166.23	232.99	312.15	263.89	315.30
2006	301.62	273.75	282.83	295.33	253.92	261.64	138.68	114.40	238.02	122.02	204.73	244.57
2007	301.16	283.09	330.98	301.03	313.36	214.12	237.94	197.64	189.56	250.41	229.90	352.79
2008	286.48	309.67	373.24	288.22	271.47	225.28	241.03	219.68	192.98	269.13	275.05	374.84
2009	273.18	275.08	342.43	287.66	272.47	211.51	226.85	201.83	138.87	250.51	254.06	342.48
2010	337.03	226.52	337.57	315.76	342.45	274.49	280.66	324.77	265.36	324.66	331.56	325.79
2011	304.21	259.36	256.81	315.73	287.75	344.68	186.84	152.91	219.90	241.36	333.13	341.71
2012	231.57	270.42	317.07	246.02	266.28	156.46	173.48	124.93	169.35	183.88	305.40	314.34
2013	353.75	291.49	221.97	291.05	237.05	170.71	221.40	165.03	178.45	186.54	243.17	360.93
2014	302.33	235.38	219.57	358.68	223.28	233.49	143.88	188.92	130.12	170.05	356.70	380.75
2015	316.76	380.18	305.47	308.24	203.65	229.66	97.91	164.81	96.06	195.59	269.12	343.46
2016	377.88	357.16	345.66	399.25	316.42	318.06	273.61	250.17	302.88	327.27	318.97	309.88
2017	356.79	326.45	370.03	400.08	359.03	293.66	266.64	238.22	203.15	322.13	368.16	368.08

Data for the remaining area levels and imputation methods can be accessed through this link:
`https://tinyurl.com/desenaldo-appendix-c`

Appendix D Forecast Rainfall Values

D.1 Forecast Rainfall Values for Country using Mean Substitution and MSSA-Row-M/Row-S/Rec

Year	Month											
	Jan	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
2000	290.64	278.31	287.24	296.84	271.02	228.28	201.21	219.51	271.95	321.55	333.94	307.62
2001	275.78	266.42	278.51	284.32	260.44	212.16	174.02	181.28	231.87	292.38	326.37	320.57
2002	294.21	272.03	264.71	255.47	228.21	181.75	140.12	132.59	165.31	220.06	266.10	285.64
2003	281.32	269.32	258.09	246.22	226.01	198.50	179.61	185.12	221.27	272.03	310.23	321.42
2004	309.13	296.38	290.57	282.10	255.37	211.61	175.15	171.49	210.78	265.88	301.43	302.57
2005	285.74	278.38	291.87	304.39	286.37	238.18	189.65	177.23	208.61	257.39	289.76	294.30
2006	287.27	286.42	295.02	296.24	270.69	221.73	172.97	153.25	167.19	199.12	233.54	260.43
2007	281.45	298.29	307.83	302.58	279.43	243.18	210.02	194.63	203.44	231.06	265.40	295.10
2008	313.79	323.40	323.11	310.56	285.32	250.42	217.66	201.68	213.21	248.61	286.74	309.55
2009	310.97	303.96	298.60	292.70	277.67	247.95	213.80	194.73	207.26	248.09	290.02	309.51
2010	304.11	293.24	299.69	318.31	324.94	303.33	266.20	245.60	260.90	302.15	334.17	332.43
2011	304.95	281.80	284.50	302.92	304.20	269.38	214.63	185.94	208.54	263.98	310.37	316.63
2012	291.03	265.66	258.72	262.47	253.89	216.42	166.73	139.01	160.52	220.82	284.78	316.83
2013	309.32	282.01	262.73	257.48	247.82	217.04	173.79	143.51	155.86	211.24	279.41	321.13
2014	319.05	290.35	266.62	259.50	249.99	220.66	173.22	140.40	152.91	216.48	297.16	348.83
2015	352.80	327.96	300.11	281.88	260.74	220.92	161.06	120.60	126.58	193.26	283.91	345.98
2016	366.38	348.10	313.82	283.44	255.84	217.15	165.11	121.12	117.46	168.45	254.73	335.17
2017	375.58	369.49	335.63	296.11	257.85	213.70	161.88	117.11	107.19	148.36	230.03	319.15

Data for the remaining area levels, imputation methods, and SSA methods can be accessed through this link:
`https://tinyurl.com/desenaldo-appendix-d`

Appendix E Forecast Error Values

E.1 Forecast Error Values for Country using Mean Substitution

SSA Method	RMSE		MAE	
	Reconstructed	Forecast	Reconstructed	Forecast
MSSA-Row-M/Row-S/Rec	36.94	91.65	29.83	80.94
MSSA-Row-M/Col-S/Rec	36.94	99.72	29.83	86.04
MSSA-Row-M/Row-S/Vec	49.28	92.30	39.01	81.98
MSSA-Row-M/Col-S/Vec	37.04	103.05	29.97	89.81
MSSA-Col-M/Row-S/Rec	36.94	99.72	29.83	86.04
MSSA-Col-M/Col-S/Rec	36.94	91.65	29.83	80.94
MSSA-Col-M/Row-S/Vec	37.04	103.05	29.97	89.81
MSSA-Col-M/Col-S/Vec	49.28	92.30	39.01	81.98
USSA-Row-M/Rec	36.94	99.72	29.83	86.04
USSA-Row-M/Vec	37.04	103.05	29.97	89.81
USSA-Col-M/Rec	36.94	91.65	29.83	80.94
USSA-Col-M/Vec	49.28	92.30	39.01	81.98

Data for the remaining area levels and imputation methods can be accessed through this link: `https://tinyurl.com/desenaldo-appendix-e`

Appendix F Converted Payout Values

F.1 Forecast Error Values for Country using Mean Substitution

Year	Month											
	Jan	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
2000	4.88	3.40	4.47	5.62	2.52	0.00	0.00	0.00	2.63	6.00	6.00	6.00
2001	3.09	1.97	3.42	4.12	1.25	0.00	0.00	0.00	0.00	5.09	6.00	6.00
2002	5.31	2.64	1.77	0.66	0.00	0.00	1.19	2.09	0.00	0.00	1.93	4.28
2003	3.76	2.32	0.97	0.00	0.00	0.00	0.00	0.00	0.00	2.64	6.00	6.00
2004	6.00	5.57	4.87	3.85	0.64	0.00	0.00	0.00	0.00	1.91	6.00	6.00
2005	4.29	3.41	5.02	6.00	4.36	0.00	0.00	0.00	0.00	0.89	4.77	5.32
2006	4.47	4.37	5.40	5.55	2.48	0.00	0.00	0.00	0.00	0.00	0.00	1.25
2007	3.77	5.79	6.00	6.00	3.53	0.00	0.00	0.00	0.00	0.00	1.85	5.41
2008	6.00	6.00	6.00	6.00	4.24	0.05	0.00	0.00	0.00	0.00	4.41	6.00
2009	6.00	6.00	5.83	5.12	3.32	0.00	0.00	0.00	0.00	0.00	4.80	6.00
2010	6.00	5.19	5.96	6.00	6.00	6.00	1.94	0.00	1.31	6.00	6.00	6.00
2011	6.00	3.82	4.14	6.00	6.00	2.33	0.00	0.00	0.00	1.68	6.00	6.00
2012	4.92	1.88	1.05	1.50	0.47	0.00	0.00	1.32	0.00	0.00	4.17	6.00
2013	6.00	3.84	1.53	0.90	0.00	0.00	0.00	0.78	0.00	0.00	3.53	6.00
2014	6.00	4.84	1.99	1.14	0.00	0.00	0.00	1.15	0.00	0.00	5.66	6.00
2015	6.00	6.00	6.00	3.83	1.29	0.00	0.00	3.53	2.81	0.00	4.07	6.00
2016	6.00	6.00	6.00	4.01	0.70	0.00	0.00	3.47	3.90	0.00	0.57	6.00
2017	6.00	6.00	6.00	5.53	0.94	0.00	0.00	3.95	5.14	0.20	0.00	6.00

Data for the remaining area levels, imputation methods, and SSA methods can be accessed through this link:
`https://tinyurl.com/desenaldo-appendix-f`

Appendix G Indicator Metrics of Payout Values

G.1 Indicator Metrics of Payout Values for Country using Mean Substitution and MSSA-Row-M/Row-S/Rec

Metric	Reconstructed					Forecast				
	1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
Mean	2.72	2.66	2.66	2.66	1.02	3.18	2.49	2.49	2.49	3.28
StDev	2.49	2.53	2.53	2.53	1.75	2.66	2.86	2.86	2.86	2.63
VaR	6.00	6.00	6.00	6.00	5.63	6.00	6.00	6.00	6.00	6.00
CTE	6.00	6.00	6.00	6.00	5.96	6.00	6.00	6.00	6.00	6.00
P75	5.55	5.55	5.55	5.55	1.28	6.00	6.00	6.00	6.00	6.00
DMM	0.22	0.72	0.72	0.72	1.02	0.68	0.55	0.55	0.55	3.28

Data for the remaining area levels, imputation methods, and SSA methods can be accessed through this link:
`https://tinyurl.com/desenaldo-appendix-g`

Appendix H Simulation Parameters

H.1 Exponential Monthly Simulation Parameters for Aceh

Year	Month											
	Jan	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
2000	0.24	0.28	0.55	0.50	0.25	0.20	0.46	0.81	0.25	0.37	0.04	0.13
2001	0.13	0.42	0.74	0.53	0.44	0.70	1.48	1.42	0.27	0.07	0.26	0.12
2002	0.13	0.42	0.74	0.53	0.44	0.70	1.48	1.42	0.27	0.07	0.26	0.12
2003	0.17	0.13	0.20	0.16	0.23	0.57	0.47	0.25	0.46	0.09	0.21	0.21
2004	0.60	0.33	0.09	0.33	0.60	0.30	1.21	0.70	0.28	0.24	0.13	0.13
2005	0.13	0.70	0.20	0.44	0.28	0.19	0.84	0.76	0.39	0.15	0.24	0.12
2006	0.18	0.17	0.15	0.29	0.38	0.20	0.50	0.97	0.19	0.24	0.14	0.31
2007	0.18	3.13	0.24	0.10	0.09	0.28	0.39	1.41	0.37	0.24	0.16	0.16
2008	0.19	1.45	0.19	0.19	0.45	1.47	0.57	0.52	0.46	0.42	0.10	0.17
2009	0.10	0.22	0.24	0.14	0.14	1.36	3.86	0.21	0.23	0.70	0.08	0.10
2010	0.17	0.21	0.29	0.13	0.50	0.14	0.24	0.38	0.35	0.27	0.06	0.09
2011	0.18	0.31	0.10	0.20	0.50	1.45	0.54	0.41	0.19	0.57	0.16	0.21
2012	0.30	0.33	0.30	0.34	0.30	0.73	0.96	0.74	0.34	0.23	0.13	0.16
2013	0.06	0.11	0.09	0.11	0.11	0.08	0.11	0.27	0.04	0.19	0.11	0.09
2014	0.19	0.31	3.86	0.24	0.31	0.40	0.82	0.22	0.20	0.06	0.05	0.06
2015	0.18	0.15	0.15	0.04	0.04	0.04	0.23	0.15	0.10	0.07	0.12	0.08
2016	0.13	0.16	0.31	0.10	0.10	0.37	0.38	0.13	0.48	0.08	0.08	0.14
2017	0.06	0.17	0.08	0.09	0.18	0.39	0.36	0.29	0.09	0.11	0.06	0.06

Data for the remaining provinces and distributions can be accessed through this link:
`https://tinyurl.com/desenaldo-appendix-h`

H.2 Binary-Exponential Monthly Simulation Parameters for Aceh

Year	Month											
	Jan	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
2000	0.08	0.07	0.19	0.20	0.09	0.10	0.13	0.19	0.11	0.13	0.02	0.05
2001	0.06	0.06	0.17	0.13	0.14	0.19	0.38	0.32	0.10	0.03	0.13	0.05
2002	0.06	0.06	0.17	0.13	0.14	0.19	0.38	0.32	0.10	0.03	0.13	0.05
2003	0.09	0.04	0.09	0.07	0.09	0.17	0.13	0.07	0.17	0.04	0.06	0.11
2004	0.15	0.11	0.04	0.07	0.16	0.10	0.21	0.18	0.14	0.11	0.08	0.06
2005	0.08	0.08	0.09	0.11	0.12	0.06	0.14	0.15	0.08	0.09	0.10	0.07
2006	0.03	0.07	0.06	0.12	0.12	0.09	0.13	0.16	0.06	0.16	0.07	0.15
2007	0.08	0.25	0.07	0.03	0.04	0.13	0.09	0.23	0.13	0.09	0.08	0.06
2008	0.08	0.20	0.07	0.07	0.17	0.32	0.18	0.20	0.11	0.16	0.07	0.10
2009	0.06	0.04	0.10	0.03	0.05	0.05	0.71	0.10	0.09	0.16	0.06	0.05
2010	0.04	0.04	0.13	0.07	0.13	0.06	0.11	0.12	0.13	0.08	0.05	0.05
2011	0.07	0.13	0.05	0.08	0.17	0.20	0.13	0.10	0.07	0.23	0.05	0.11
2012	0.10	0.17	0.10	0.10	0.15	0.10	0.25	0.15	0.08	0.11	0.06	0.11
2013	0.06	0.11	0.09	0.11	0.11	0.08	0.11	0.27	0.04	0.19	0.11	0.09
2014	0.06	0.06	0.57	0.10	0.14	0.14	0.30	0.07	0.11	0.05	0.03	0.05
2015	0.18	0.15	0.15	0.04	0.04	0.04	0.22	0.14	0.09	0.07	0.12	0.08
2016	0.08	0.08	0.08	0.07	0.08	0.25	0.28	0.10	0.40	0.08	0.08	0.14
2017	0.05	0.11	0.08	0.08	0.17	0.23	0.27	0.23	0.09	0.09	0.05	0.06

Data for the remaining provinces and distributions can be accessed through this link: `https://tinyurl.com/desenaldo-appendix-h`

H.3 Parameters for Different Distribution Every Year Simulation (Year 2000)

Province	Distribution	Par1	Par2
Aceh	Gamma	0.06	0.01
Sumatera Utara	Gamma	0.06	0.01
Sumatera Barat	Normal	11.94	26.13
Riau	Gamma	0.08	0.01
Jambi	Gamma	0.07	0.01
Sumatera Selatan	Normal	7.81	15.43
Bengkulu	Normal	10.45	20.72
Lampung	Gamma	0.06	0.01
Kepulauan Bangka Belitung	Normal	10.18	18.51
Kepulauan Riau	Normal	10.66	17.58
DKI Jakarta	Gamma	0.07	0.01
Jawa Barat	Normal	5.43	11.16
Jawa Tengah	Gamma	0.06	0.01
DI Yogyakarta	Gamma	0.06	0.01
Jawa Timur	Gamma	0.05	0.01
Banten	Gamma	0.06	0.01
Bali	Gamma	0.06	0.01
Nusa Tenggara Barat	Gamma	0.05	0.01
Nusa Tenggara Timur	Gamma	0.05	0.01
Kalimantan Barat	Normal	9.93	20.26
Kalimantan Tengah	Weibull	0.39	8.19
Kalimantan Selatan	Exponential	0.08	-
Kalimantan Timur	Normal	6.16	12
Kalimantan Utara	Normal	6.64	12.66
Sulawesi Utara	Normal	10.53	21.65
Sulawesi Tengah	Normal	7.59	16.69
Sulawesi Selatan	Gamma	0.06	0.01
Sulawesi Tenggara	Gamma	0.06	0.01
Gorontalo	Gamma	0.07	0.01
Sulawesi Barat	Gamma	0.06	0.01
Maluku	Normal	8.33	18.44
Maluku Utara	Gamma	0.04	0.02
Papua Barat	Normal	13.23	20.84
Papua	Normal	7.46	15.08

Data for the remaining years can be accessed through this link:
`https://tinyurl.com/desenaldo-appendix-h`

All appendices can be accessed here:
`https://tinyurl.com/desenaldo-appendices`

Or scan using this QR code (same link):

Scientific and Personal Career

Education

10/2022 - 03/2026	PhD in Mathematics *RPTU Kaiserslautern-Landau, Germany*
08/2019 - 07/2021	Master of Actuarial Science *Institut Teknologi Bandung, Indonesia*
08/2015 - 02/2019	Bachelor of Statistics *Universitas Padjadjaran, Indonesia*

Experience

06/2021 - present	Staff of Technical Division *Setya Widodo Actuarial Consultant, Indonesia*
01/2018 - present	Head of Technical Division *PT Tama Aktuaria Sejahtera, Indonesia*

Wissenschaftlicher und Beruflicher Werdegang

Akademische Ausbildung

10/2022 - 03/2026	Doktorandin der Mathematik *RPTU Kaiserslautern-Landau, Deutschland*
08/2019 - 07/2021	Masterstudium in Versicherungsmathematik *Institut Teknologi Bandung, Indonesien*
08/2015 - 02/2019	Bachelorstudium in Statistik *Universitas Padjadjaran, Indonesien*

Berufserfahrung

06/2021 - heute	Mitarbeiterin, Technische Abteilung *Setya Widodo Actuarial Consultant, Indonesien*
01/2018 - heute	Leiterin, Technische Abteilung *PT Tama Aktuaria Sejahtera, Indonesien*