ZVersWiss (2021) 110:21–48
https://doi.org/10.1007/s12297-021-00493-1

ABHANDLUNG

Neuronale Netze und Zeitreihenansätze zur Vorhersage des auslösenden Faktors in der privaten Krankenversicherung

Fatlinda Avdullai · Ralf Korn · Hans-Martin Hoben

Angenommen: 3. Februar 2021 / Online publiziert: 1. März 2021

Zusammenfassung Die Möglichkeit einer Prämienanpassung in der deutschen PKV ist vom Wert des sogenannten auslösenden Faktors abhängig, der mittels einer linearen Extrapolation der Schadenquotienten der vergangenen drei Jahre berechnet wird. Seine frühzeitige, verlässliche Vorhersage ist aus Sicht des Risikomanagements von großer Bedeutung. Wir untersuchen deshalb vielfältige Vorhersageansätze, die von klassischen Zeitreihenansätzen und Regression über neuronale Netze bis hin zu hybriden Modellen reichen. Während bei den klassischen Methoden Regression mit ARIMA-Fehlern am besten abschneidet, zeigt ein neuronales Netz, das mit Zeitreihenvorhersage kombiniert oder auf desaisonalisierten und trendbereinigten Daten trainiert wurde, das insgesamt beste Verhalten.

Schlüsselwörter Krankenversicherungstarife · Auslösender Faktor · Zeitreihenvorhersagemethoden · Neuronale Netze · Regressionsmodelle

F. Avdullai · R. Korn (✉)
Fachbereich Mathematik, TU Kaiserslautern, 67653 Kaiserslautern, Deutschland
E-Mail: korn@mathematik.uni-kl.de

F. Avdullai
E-Mail: avdullai@mathematik.uni-kl.de

H.-M. Hoben
Debeka Krankenversicherungsverein a. G., 56058 Koblenz, Deutschland

R. Korn
Abteilung Finanzmathematik, Fraunhofer ITWM, 67653 Kaiserslautern, Deutschland

Neural networks and time series approaches to predict the triggering factor in private health insurance

Abstract The possibility to adapt a premium in German private health insurance depends on the so-called triggering factor. Its value is determined via a linear extrapolation of the loss ratios of the three preceding years. To predict this value early and in a reliable way is of great importance for risk management purposes. We therefore examine the performance of various prediction methods that range from classical time series methods and regression to neural networks and hybrid methods. While regression with ARIMA errors performs best among the classical methods, the overall best method is the combination of time series predictions with a neural network that is trained on deseasonalized and detrended data.

Keywords Private health insurance tariffs · Triggering factor · Time series prediction methods · Neural networks · Regression models

1 Einleitung und Hintergrund

Private Krankenversicherung nach Art der Lebensversicherung, insbesondere als substitutive Krankenversicherung, bietet in der Regel lebenslangen Versicherungsschutz (siehe z. B. Pitacco 2014). Der Versicherer hat kein Kündigungsrecht und kann auch nicht den Leistungsumfang reduzieren. Um einen solchen lebenslangen Versicherungsschutz zu ermöglichen, kann der Versicherer unter gewissen Bedingungen seine Rechnungsgrundlagen anpassen, was z. B. im Fall einer allgemeinen Kostensteigerung seiner Leistungen mit einer Prämiensteigerung einhergehen kann. Eine solche Kostensteigerung kann das Resultat des allgemeinen Anstiegs von Behandlungskosten sein, was auch als *medizinische Inflation* bezeichnet wird (siehe Becker 2017).

Gemäß ğ 155 Absätze 3,4 des Versicherungsaufsichtsgesetzes (VAG) muss der Versicherer (mindestens) jährlich die rechnungsmäßigen und die tatsächlichen Versicherungsleistungen (sowie die Sterberaten) vergleichen. Im Fall einer Abweichung von mehr als 10 % (5 %), sofern in den Allgemeinen Versicherungsbedingungen (AVB) kein niedrigerer Prozentsatz vereinbart ist, hat der Versicherer alle Prämien der betreffenden Tarifs zu überprüfen. Werden die Abweichungen als dauerhaft angesehen, so sind die Prämien unter Zustimmung eines Treuhänders anzupassen (siehe Milbrodt und Röhrs 2016). Die Basis für diesen Prozess ist die jährliche Berechnung des sogenannten *auslösenden Faktors*.

Der auslösende Faktor ergibt sich aus einer linearen Extrapolation der Gesamtschadenquotienten der letzten drei Jahre inklusive des aktuellen Jahrs, was insbesondere bedeutet, dass er erst berechenbar ist, wenn die vollständigen Daten des aktuellen Jahres vorliegen. Hingegen ist es aus der Sicht des Risikomanagements sehr wünschenswert, möglichst frühzeitig eine verlässliche Prognose für seinen Wert zu erhalten. Folglich ist es von großer Wichtigkeit, eine entsprechende Vorhersagemethodik zu entwickeln und zu implementieren.

Dabei stellen Vorhersageverfahren auf der Basis von Zeitreihenmodellen einen naheliegenden klassischen Ansatz dar, der in der Wirtschaft häufig erfolgreich eingesetzt wurde. Besonders populär sind sogenannte ARIMA-Modelle (AutoRegressive Integrated Moving Average), also lineare Zeitreihen vergangener Beobachtungen und Zufallsfehler. Einfache Vorhersagemethoden wie Regressionsmodelle sagen Zeitreihen unter der Annahme eines linearen Zusammenhangs mit anderen Zeitreihen voraus. Um deren Vorhersagegenauigkeit zu erhöhen, lassen sich Regressionsmodelle mit abhängigen Fehlertermen betrachten, die als ARIMA-Zeitreihe modelliert werden. Hierdurch wird dem ARIMA-Konzept ermöglicht, zusätzliche Informationen zu berücksichtigen.

(Künstliche) Neuronale Netze (ANN) werden gegenwärtig in vielen Gebieten der Wirtschaft, Industrie und Forschung erfolgreich als Vorhersagemethode verwendet. Da sie rein datenbasiert und selbstanpassend sind, benötigen sie nur schwache Modellannahmen, was sie für die praktische Anwendung attraktiv erscheinen lässt. Diese Attraktivität wird durch weitere Eigenschaften (wie z. B. die Fähigkeit zu verallgemeinern oder die universelle funktionale Approximations-eigenschaft) unterstützt. Schließlich sind neuronale Netze nichtlineare Regressions- und Klassifikationsverfahren und somit auch in der Lage, nichtlineare Zusammenhänge adäquat zu behandeln (siehe z. B. Zhang et al 1998). Allerdings zeigen empirische Studien der Vorhersage saisonaler Daten mit neuronalen Netzen unterschiedliche Resultate. Während teilweise berichtet wird, dass neuronale Netze mit der Saisonalität ohne weitere Vorverarbeitung der Daten umgehen können, wird auch teilweise das direkte Gegenteil berichtet. So berichtet Zhang (Zhang und Min 2005), dass Modelle, die Saisonalität und Trendmuster ignorieren, hohe Varianz und schlechte Vorhersagegenauigkeit aufweisen und sich dies nur durch eine vorangeschaltete Desaisonalisierung und Trendbereinigung der Daten beheben lässt.

Viele Studien zur Vorhersage mittels Zeitreihen zeigen auf, dass Vorhersagen auf der Basis kombinierter Modelle denjenigen von individuellen Modellen überlegen sind. Wir haben deshalb auch eine hybride Modellierung untersucht, bei der sowohl ARIMA-Modelle als auch Regression mit neuronalen Netzen kombiniert werden.

Im Rahmen unserer Arbeit werden wir Zeitreihenvorhersagemodelle und neuronale Netze anwenden, um den auslösenden Faktor für einen gegebenen Tarif in der PKV vorherzusagen. Dabei liegt unser Fokus darauf, die beste Vorhersagemethodik aus Zeitreihenansätzen, neuronalen Netzen, hybriden Methoden und neuronalen Netzen, die mit desaisonalisierten und trendbereinigten Daten trainiert wurden, zu finden. Hierzu dienen uns monatliche Daten der Debeka Krankenversicherung als Basis unserer Fallstudie. Zur Beurteilung der Vorhersagegüte der einzelnen Verfahren werden die Wurzel der mittleren Fehlerquadratsumme (RMSE), der mittlere absolute prozentuale Fehler (MAPE) und der mittlere absolute Fehler (MAE) als Qualitätsmaße verwendet.

Dabei ist unsere Arbeit folgendermaßen aufgebaut: Wir werden zunächst die grundlegenden Vorhersagemodelle samt der oben erwähnten Aspekte im nächsten Abschnitt vorstellen und danach die Daten und die Aufgabenstellung beschreiben. Die Präsentation der numerischen Resultate ist Gegenstand des vierten Kapitels, während das fünfte Kapitel die abschließende Diskussion unserer Ergebnisse beinhaltet.

2 Modellansätze zur Vorhersage von Zeitreihen

Daten, die in regelmäßiger Form im Zeitablauf gemessen werden, werden als *Zeitreihe* bezeichnet. Viele Zeitreihen in der praktischen Anwendung besitzen einen *Trend* und/oder *saisonales Verhalten* (z. B. Wetterdaten oder Stromverbrauchsdaten). Dabei verseteht man unter einem Trend einen länger währenden An- oder Abstieg, der nicht notwendigerweise linear sein muss. Saisonales Verhalten liegt typischerweise bei Daten vor, die von Jahreszeiten oder Wochentagen abhängen. Dabei ist die Saisonalität immer durch eine feste und bekannte Frequenz gekennzeichnet (siehe Andrew und Paul 2009).

Um zufällige und deterministische Einflüsse zu trennen, versucht man Zeitreihen in die entsprechenden Komponenten zu zerlegen. Frühe Quellen für solche Zerlegungsansätze sind bereits in den 1920er Jahren zu finden (siehe z. B. Persons 1919). In der Praxis bedeutsame Zerlegungen mit saisonaler Anpassung sind die X-11 Methode, die vom US-amerikanischen Bureau of the Census in den 1950er and 1960er Jahren entwickelt wurde (siehe Shiskin et al 1967). Eine Weiterentwicklung stellt das X-13ARIMA-SEATS-Programm dar (vgl. Monsell 2007). Als Alternative zu diesen Ad-hoc-Methoden wurden auch modellbasierte saisonale Anpassungsverfahren entwickelt wie z. B. das bekannte Box-Jenkins-Verfahren (siehe Box und Jenkins 1976), das auf dem saisonalen ARIMA-Modell basiert und sich bei der Modellwahl an der Autokorrelation der Daten orientiert. In der neueren Zeit werden neuronale Netze als populäre Alternative zur traditionellen Zeitreihenvorhersage angesehen (vgl. Zhang et al 1998). Wir werden in diesem Abschnitt einen Überblick über die klassischen Methoden der Zerlegung und der saisonalen Anpassung geben, uns danach mit Regressionsmodellen und dem Box-Jenkins-Ansatz beschäftigen sowie neuronale Netze und hybride Methoden vorstellen.

2.1 Zerlegungsmethoden mit saisonaler Anpassung

Man unterscheidet zunächst zwischen additiver und multiplikativer Zerlegung. Dabei wird die erste Methode angewendet, wenn die Größenordnung der saisonalen Schwankungen (oder die Variation um einen Trend herum) nicht vom Niveau der Zeitreihe abhängt, während die zweite dann eingesetzt wird, wenn der saisonale Effekt sich mit wachsendem Trend verstärkt (oder abschwächt). Frühe Methoden der Vorhersage mit Zeitreihen sind stark von der Saisonbereinigung der Daten abhängig (siehe Zhang und Min 2005). Bei saisonal adjustierten Daten wird die saisonale Komponente aus den Originaldaten entfernt. So zerlegt man z. B. im additiven Modell die Zeitreihe Y_t in eine saisonale Komponente S_t und eine nicht-saisonale Komponente NS_t:

$$Y_t = S_t + NS_t \tag{1}$$

Gründe für eine saisonale Anpassung liegen in der verbesserten Erkennbarkeit eines Trends oder in der verbesserten Kurzfristvorhersagemöglichkeit auf der Basis der angepassten Daten. Des weiteren lassen sich das Verhältnis zu anderen Zeitreihen,

die ebenfalls saisonale Daten beschreiben, besser erkennen und die Daten einfacher von Monat zu Monat vergleichen (siehe Bell und Hillmer 1984).

Eine populäre Methode, um einen Trend und eine Saisonalität der Zeitreihe Y_t zu schätzen, ist die *Zentrierung durch ein gleitendes Mittel*. Dabei wird zunächst der Trend mittels eines gleitenden Mittels geschätzt und dieses von den Daten abgezogen. Anschließend wird die saisonale Komponente durch Mittelbildung über die so modifizierten Daten einer Saison geschätzt. Diese einfache Methode wird vereinzelt noch in der Praxis eingesetzt.

Die in Shiskin et al (1967) entwickelte *Census Method II-X11* ist eine weitere populäre Methode für die Schätzung der wesentlichen Komponenten einer Zeitreihe und basiert auf glättenden linearen Filtern (gleitenden Mitteln), wobei angenommen wird, dass die Zeitreihe der Daten eine additive oder multiplikative Form besitzt. Für monatliche Daten werden die wesentlichen Schritte von X11, um desaisonalisierte Daten zu erhalten und die Trendkomponente zu schätzen, in Dagum und Bianconcini (2016) beschrieben. In Dagum (1978) wird als Modifikation von Census X11 die *X11ARIMA*-Methode eingeführt, die eine bessere saisonale Anpassung bei Daten mit einer schnell stochastisch wechselnden Saisonalität erreichen soll. Die in Findley et al (1998) entwickelte *X12ARIMA*-Methode verwendet zusätzlich Regressionsmodelle mit ARIMA-Fehlern (regARIMA). Dabei kann die Zeitreihe auch zunächst im Hinblick auf Ausreißer und Kalendereffekte vorverarbeitet werden, bevor die saisonale Anpassung durchgeführt wird.

Eine auf ARIMA-Modellen basierende Zerlegung einer Zeitreihe ist die *SEATS*-Methode („Signal Extraction in ARIMA Time Series"), die in Burman (1980) vorgestellt wird. Die spezielle, vom US Census Bureau entworfene *X-13ARIMA-SEATS*-Methode (siehe Monsell 2007) erlaubt die Schätzung der saisonalen Komponente und eines Trends über lineare Filter wie in X12-ARIMA oder basierend auf einer ARIMA-Modellzerlegung.

2.2 Zeitreihen-Regressionsmodelle

Regressionsmodelle nehmen eine lineare Beziehung der Zeitreihe y zu einer weiteren Zeitreihe x an, wobei y die abhängige Variable und die Vorhersagevariable x die unabhängige Variable ist. Ein saisonales Modell mit einer Saison von s Zeiteinheiten (z. B. $s = 12$ Monate in einem Jahr) und einem Trend μ_t ist dann gegeben durch

$$y_t = \mu_t + s_t + \epsilon_t \tag{2}$$

wobei $s_t = \alpha_i$ gilt, wenn t in die i. Zeiteinheit der Saison fällt ($t = 1, ..., n; i = 1, ..., s$), und ϵ_t die Zeitreihe der Residuen darstellt, die autokorreliert sein kann. So ist z. B. für eine monatliche Zeitreihe Y_t, die mit $t = 1$ im Januar startet, ein saisonales Indikatormodell mit einem linearen Trend durch

$$y_t = \beta_1 t + s_t + \epsilon_t = \begin{cases} \beta_1 t + \alpha_1 + \epsilon_t & t = 1, 13, ... \\ \beta_1 t + \alpha_2 + \epsilon_t & t = 2, 14, ... \\ \vdots & \\ \beta_1 t + \alpha_{12} + \epsilon_t & t = 12, 24, ... \end{cases} \tag{3}$$

gegeben. Dabei können die Parameter in Gl. (3) mittels kleinster Quadrate geschätzt werden, indem man die saisonale Komponente s_t als Dummy-Variablen auffasst.

Falls es mehrere Regressionsmodelle gibt, die z. B. unterschiedliche Saisonalitäten beinhalten können (monatlich, täglich, stündlich), bei denen die Stärke der Einflüsse unterschiedlich ist, so braucht es ein Kriterium für die Modellwahl. Populäre Maßzahlen hierfür sind das adjustierte R^2, Akaikes Informationskriterium (AIC), das korrigierte AIC (AIC_c) oder das Schwarzsche Bayesische Informationskriterium (BIC) (siehe z. B. Burnham und Anderson 2004, Nagelkerke et al 1991 oder Sheather 2009 für die zugehörigen Definitionen). Nachdem das Regressionsmodell gewählt und die Parameter an die Daten angepasst worden sind, werden die Resiuen analysiert. Hierbei sprechen unkorrelierte Residuen mit einem Mittelwert von Null für eine gute Modellanpassung. Sind die Residuen sogar zusätzlich normal verteilt mit konstanter Varianz, so liegt eine zufriedenstellende Modellierung vor (siehe Royston 1982).

2.3 Saisonale ARIMA-Modelle

Saisonale ARIMA-Modelle sind lineare Zeitreihen vergangener Beobachtungen und zufälliger Fehler, die durch die Modellgleichung

$$\Phi_P(B^s)\phi_p(B)(1-B^s)^D(1-B)^d y_t = \Theta_Q(B^s)\theta_q(B)\epsilon_t \tag{4}$$

gegeben sind. Dabei ist s die Länge einer Saison, $(1-B)^d$ und $(1-B^s)^D$ sind der nicht-saisonale und der saisonale Differenzenoperator, während $\Phi_P, \phi_p, \Theta_Q,$ und θ_q Polynome der Ordnungen $P, p, Q,$ und q sind. B ist der Rückwärtsoperator, der durch $B y_t = y_{t-1}$ gegeben ist und ϵ_t ist eine Folge weißen Rauschens, d. h. eine Folge unkorrelierter Zufallsvariablen mit Erwartungswert Null und konstanter Varianz. Dabei werden die Parameter d, D so gewählt, dass durch die Anwendung der zugehörigen Differenzenbildungen $(1-B)^d$ auf der einfachen (bei uns monatlichen) Zeitskala und $(1-B^s)^D$ auf der saisonalen Zeitskala stationäre Daten entstehen. Das Modell wird mit $SARIMA(p,d,q)(P,D,Q)_s$ bezeichnet.

In Box und Jenkins (1976) wird eine Mehrschritt-Modellwahl-Strategie für saisonale ARIMA-Modelle eingeführt. Sie besteht aus Modellwahl, Modellanpassung und Modelldiagnose. Bei der Modellwahl werden die Autokorrelationsfunktion (ACF) und die partielle Autokorrelationsfunktion (PACF) der Daten zur Bestimmung der einzelnen Ordnungen des ARIMA-Modells verwendet, indem man die empirischen Schätzer von ACF und PACF mit dem theoretischen Verhalten der ACF und PACF des jeweiligen Modells vergleicht. Hier kann auch eine Datentransformation wie z. B. die Box-Cox-Transformation verwendet werden, um aus den Daten eine stationäre Zeitreihe (also eine, deren Eigenschaften nicht von der Beobachtungszeit abhängen) zu erzeugen.

In der Anpassungsphase werden die bestmöglichen Schätzer für ein gegebenes Modell bestimmt (je nach Möglichkeit durch Maximum-Likelihood- (MLE) oder Kleinste-Quadrate-Methode). In der Modelldiagnosephase werden die entstandenen Residuen mittels Statistiken und Grafiken analysiert, um das am sparsamsten parametrisierte gute Modell zu definieren, das dann für die Vorhersage verwendet wird.

2.4 Regression mit SARIMA-Fehlern

Standard-Regressionsmodelle sind nicht in der Lage, die komplizierte Zeitreihendynamik eines *SARIMA*-Modells zu beschreiben. Wir werden dies in diesem Abschnitt beheben, in dem wir Fehlerterme bei der Regression erlauben, die eine nichtverschwindende Autokorrelation besitzen. Hierzu nehmen wir an, dass wir Beobachtungen $\{y_1, ..., y_n\}$ vorliegen haben, die als lineare Funktion der k (Prädiktor-) Variablen $\{x_{1,t}, ..., x_{k,t}\}$ und einer Fehlerfolge η_t, die einem *SARIMA*-Modell folgt, gegeben sind. Dann ist das Regressionsmodell mit $SARIMA(p, d, q)(P, D, Q)_s$-Fehlern definiert als

$$\begin{aligned} &y_t = \beta_0 + \beta_1 x_{1,t} + ... + \beta_k x_{k,t} + \eta_t, \\ &\Phi_P(B^s)\phi_p(B)(1 - B^s)^D(1 - B)^d \eta_t = \Theta_Q(B^s)\theta_q(B)\epsilon_t, \end{aligned} \tag{5}$$

wobei ϵ_t eine Folge weißen Rauschens ist. Wir schätzen die Modellparameter mittels Minimierung der Summe der Quadrate der ϵ_t-Werte. Hierdurch werden Probleme mit den AIC_c-Werten in diesem Rahmen vermieden (siehe Hyndman und Athanasopoulos 2018).

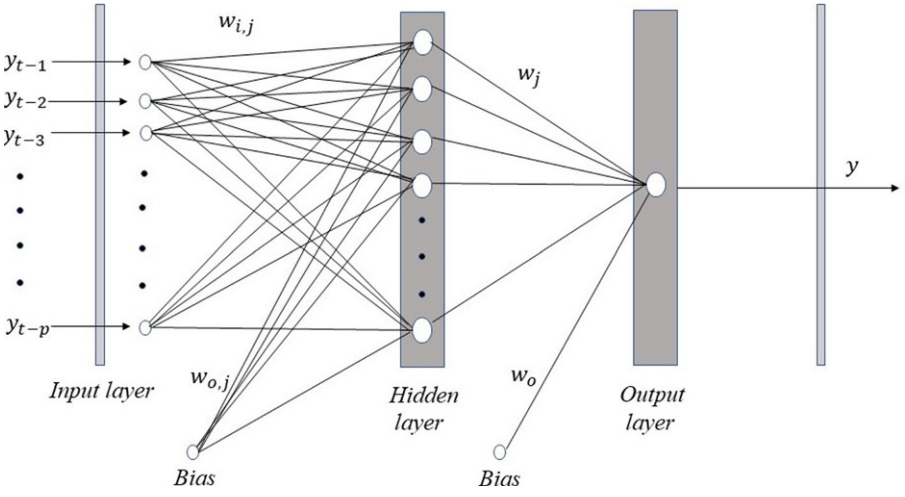

Abb. 1 Struktur eines neuronalen Netzes mit einem Hidden Layer

2.5 Neuronale Netze

(Künstliche) neuronale Netze (ANN) sind daten-basiert, selbst-adaptiv und benötigen nur geringe Modellannahmen, weshalb sie – auch dank vorhandener Open Source Software – mittlerweile allgegenwärtig in Anwendungen in Wirtschaft, Industrie und Forschung zu sein scheinen. Da ein ANN jede Funktion beliebig genau approximieren kann (bei geeignet hoher Neuronenzahl) – man spricht auch von der *universellen Approximationseigenschaft* – gibt es auch mathematische Gründe für ihre Popularität. Nicht zuletzt die Nichtlinearität von ANN kann ihnen einen entscheidenden Vorteil gegenüber traditionellen Vorhersagemethoden wie der Box-Jenkins- oder ARIMA-Methodik verschaffen, wenn der zugrunde liegende Mechanismus der Datenerzeugung nichtlinear ist (siehe Zhang et al 1998).

Aufgrund ihrer Einfachheit – und deshalb auch ihrer Verständlichkeit – werden oft ANN mit nur einem sogenannten *Hidden Layer* (vgl. Abb. 1) für die Modellierung und Vorhersage von Zeitreihen verwendet (siehe Zhang et al 1998). Die Beziehung zwischen Output-Variablen (y_t) und den Input-Variablen ($y_{t-1}, ..., y_{t-p}$) ist dort gegeben gemäß

$$y_t = w_0 + \sum_{j=1}^{q} w_j \sigma \left(w_{0,j} + \sum_{i=1}^{p} w_{i,j} y_{t-i} \right) + \epsilon_t, \tag{6}$$

wo $w_{i,j} (i = 0, 1, 2, ..., p, j = 1, 2, ..., q)$ und $w_j (j = 0, 1, ..., q)$ die Modellparameter darstellen, die sogenannten Gewichte. p ist die Anzahl der Input-Knoten, q die der verborgenen (*hidden*) Knoten (oder *Neuronen*). σ wird als *Aktivierungsfunktion* bezeichnet, während ϵ_t den Fehler darstellt (Abb. 1). Als Aktivierungsfunktion für die Output-Schicht wird in der Regel eine lineare Funktion gewählt. Als Aktivierungsfunktion für die verborgene Schicht (den *Hidden Layer*) werden oft die logistische Funktion, der Tangens Hyberbolicus oder der Positivteil gewählt:

$$\sigma(x) = \frac{1}{1 + \exp(-x)} \tag{7}$$

$$\sigma(x) = \tanh(x) \tag{8}$$

$$\sigma(x) = \max\{0, x\} \tag{9}$$

Die Nichtlinearität der Aktivierungsfunktionen kann dabei helfen, den Input besser an die Werte des Output anzupassen als durch ein lineares Regressionsproblem. Je nach Anwendung können die beiden beschränkten und differenzierbaren Aktivierungsfunktionen vorteilhaft sein, wobei allerdings auch in vielen Anwendungen aus Gründen der Einfachheit der Positivteil verwendet wird. Allerdings scheint es bisher keine abschließende globale theoretische Empfehlung für die Verwendung einer speziellen Aktivierungsfunktion zu geben.

Das ANN zu (6) ist eine nichtlineare Abbildung der Beobachtungen der Vergangenheit in die Zukunft y_t, genauer

$$y_t = f(y_{t-1}, ..., y_{t-p}, w) + \epsilon_t, \tag{10}$$

wobei w der Parametervektor ist und die Funktion f durch die Aktivierungsfunktion und die Netzwerkstruktur gegeben ist. Ein solches neuronales Netzwerk ist also äquivalent zu einem nichtlinearen auto-regressiven Modell. Einer der Gründe, warum solch einfache Strukturen mit lediglich einem Hidden Layer in der Praxis verwendet werden, liegt auch darin, dass tiefe Netze mit mehreren Hidden Layern zwar Daten sehr gut anpassen können, dann aber in der Vorhersage bzw. in der Testphase nicht so gut abschneiden, da sie im wesentlichen die Daten auswendig gelernt und dabei die Fähigkeit verloren haben, das Wesentliche zu lernen. Man spricht dann von *Overfitting* (siehe Khashei und Bijari 2010). Ein überbestimmtes Modell ist insbesondere für die Vorhersage ungeeignet. Die Wahl der Anzahl q der Neuronen des Hidden Layers hängt von den vorliegenden Daten ab und unterliegt keiner systematischen Regel.

Außer der Wahl der geeigneten Anzahl von Neuronen im Hidden Layer des ANN ist es von großer Wichtigkeit, zu wählen, wie viele Beobachtungen der Vergangenheit als Inputvariablen verwendet werden, d. h., die Zahl p muss festgelegt werden. Die Wahlen von p und q sowie die Wahl der Aktivierungsfunktion(en) und der Anzahl der Hidden Layer im ANN werden unter dem Fachbegriff des *Hyperparameter-Tuning* zusammen gefasst. Zwar existieren hierfür Vorschläge in der Literatur, doch sind diese in der Regel sehr komplex, und es existiert auch kein Standardverfahren (siehe Zhang et al 1998). Deshalb besteht die übliche Prozedur darin, eine Anzahl von neuronalen Netzen mit verschiedenen Wahlen der Parameter (p, q) zu betrachten und dann das Paar auszuwählen, das den besten Kompromiss aus guter Anpassung an die Daten in der Trainingsphase (siehe unten) und guter Vorhersageperformance in der Validierungsphase auf einer Testmenge zeigt (siehe Hosseini et al 2006).

Während der sogenannten Traingsphase werden dabei bei jedem gewählten ANN die Gewichtsparameter durch Minimierung des mittleren quadratischen Vorhersagefehlers (MSE) auf den (als Trainingsmenge ausgewählten) Daten bestimmt, d. h., wir minimieren die Summe der quadratischen Differenzen zwischen den beobachteten Werten (y) und den durch das ANN vorhergesagten Werten $(\widehat{y})$:

$$E = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \tag{11}$$

Für diese aufwändige Minimierung wird die sogenannte Backpropagation verwendet (siehe z. B. McClelland et al 1986 oder Khashei und Bijari 2010), die eine effiziente Umsetzung des Gradientenverfahrens zur Bestimmung einer Nullstelle des Gradienten des MSE darstellt. Sie ist in Open Source Software implementiert und somit bequem einsetzbar. Dabei sind in der Regel auch weiterentwickelte Varianten wie z. B. diejenige von Rumelhart und McClelland (siehe McClelland et al 1986) verfügbar.

Schließlich wird die Vorhersagequalität des so ermittelten („trainierten“) Modells auf einer *Testmenge* analysiert, für die Input-Output-Daten vorhanden sind, die aber nicht in die Trainigsphase eingeflossen sind.

Dieses Zusammenspiel von Kalibrierung und Validierung führt dann letztendlich zur Auswahl des für die Anwendung einzusetzenden neuronalen Netzes.

2.6 Die hybride Methodik

Da es in der praktischen Anwendung schwierig ist, zu entscheiden, ob eine Zeitreihe linear oder nicht-linear von einem zugrunde liegenden Prozess abhängt, erscheint auch eine hybride Vorgehensweise der Kombination der oben vorgestellten Ansätze ein guter Kandidat zu sein, unterschiedliche Muster in den Daten zu erkennen und damit auch die Vorhersagequalität zu verbessern.

Zhang (Zhang 2003) schlägt ein hybrides Modell aus einer ARIMA-Zeitreihe und einem neuronalen Netz vor:

$$y_t = L_t + N_t. \tag{12}$$

Dabei bezeichnet y_t die Originalzeitreihe, L_t ihre lineare und N_t ihre nichtlineare Komponente. Die lineare Komponente wird dann durch ein ARIMA-Modell geschätzt und die erhaltenen Residuen

$$e_t = y_t - \widehat{L}_t \tag{13}$$

durch ein ANN modelliert. Mit n Input-Knoten lässt sich das ANN für die Residuen als

$$e_t = f(e_{t-1}, e_{t-2}, ..., e_{t-n}) + \epsilon_t \tag{14}$$

schreiben, wobei die nichtlineare Funktion f durch das ANN bestimmt wird und ϵ_t der zufällige Modellfehler ist. Die Schätzung von e_t mittels (14) stellt dann die Vorhersage der nichtlinearen Komponente N_t der Zeitreihe dar. Wir erhalten somit die Vorhersagewerte der Zeitreihe als

$$\widehat{y_t} = \widehat{L}_t + \widehat{N}_t \tag{15}$$

3 Die gewählte Vorgehensweise

Wir werden Zeitreihenmodellvorhersagemethoden und ANN-Ansätze zur Vorhersage des auslösenden Faktors eines gegebenen Krankenversicherungstarifs anwenden und vergleichen. Zusätzlich werden wir hybride Methoden untersuchen und neuronale Netze, die auf speziell vorverarbeiteten Daten trainiert werden. Dabei wird sich der hybride Ansatz der Kombination von Zeitreihenansätzen und ANN an einem Vorschlag aus Zhang (2003) orientieren. Genauer, wir werden die besten, separat in den Schritten der Modellwahl, -anpassung und -diagnose ermittelten Vorhersagemodelle aus den Klassen der Regressionsmodelle, SARIMA- und Regressionsmodelle mit SARIMA-Fehler für die Modellierung des linearen Modellanteils auswählen sowie die Modellierung des nichtlinearen Teils durch ein ANN betrachten.

Zusätzlich zum Hybridmodell legen wir einen weiteren Fokus auf den Ansatz der Zerlegung einer Zeitreihe in eine saisonale und eine Trendkomponente sowie ihren zufälligen Anteil. Für eine solche Zerlegung wenden wir zunächst das *X-13ARIMA-*

SEATS-Verfahren an, um unsere Daten von Trend und Saisonalität zu bereinigen und trainieren neuronale Netze erst auf den bereinigten Daten. Ein solches Vorgehen wird auch in Zhang und Min (2005) empfohlen, wo auch darauf hingewiesen wird, dass die Anwendung neuronaler Netze auf die Rohdaten in der Regel zu hohen Vorhersagefehlern führt.

Alle Regressions- und SARIMA-Modelle wurden in R mit der Funktion *auto.arima()* und dem *forecast*-Paket implementiert, während für die Erstellung der neuronalen Netze das *Keras*-Paket https://keras.rstudio.com/ Chollet und Allaire (2018) verwendet wurde.

3.1 Die Daten und die Aufgabenstellung

Um aufgrund eines allgemeinen Anstiegs der Kosten der Versicherungsleistungen die Prämie eines Krankenversicherungstarifs anpassen zu dürfen, muss der Versicherer jährlich den sogenannten *auslösenden Faktor* berechnen. Die zugehörige Berechnung der (gewichteten) Versicherungsleistungen geschieht in den folgenden Schritten: Zunächst wird für jedem Monat j der *Schadenquotient* (SQ)

$$\mathrm{SQ}^{(j)} = \frac{S^{(j)}}{\sum_x L_x^{(j)} K_x^{(\mathrm{rech})}/12} \tag{16}$$

berechnet, der sich als Quotient aus tatsächlichen und kalkulatorischen Leistungen ergibt. Dabei sind $L_x^{(j)}$ die Anzahl Versicherter des Alters x im Beobachtungsmonat j und $K_x^{(\mathrm{rech})}$ der rechnungsmäßige Kopfschaden eines Versicherten des Alters x. $S^{(j)}$ ist die Summe der tatsächlichen Leistungen im j. Monat. Der jährliche Schadenquotient – z. B. in 2018 – ergibt sich als

$$\begin{aligned}\mathrm{SQ}^{2018} = \mathrm{SQ}^{1,2018}\frac{S^{1,2018,\mathrm{rech}}}{S^{1,2018,\mathrm{rech}} + \ldots + S^{12,2018,\mathrm{rech}}} + \ldots \\ + \mathrm{SQ}^{12,2018}\frac{S^{12,2018,\mathrm{rech}}}{S^{1,2018,\mathrm{rech}} + \ldots + S^{12,2018,\mathrm{rech}}}\end{aligned} \tag{17}$$

Im zweiten Schritt kann dann der auslösende Faktor der Versicherungsleistungen für das *Berichtsjahr* 2018 berechnet werden, der die Grundlage für eine evtl. Beitragsanpassung in 2020 darstellt. Er ist definiert als:

$$\mathrm{AF}(2018) = \frac{1}{3}\left(\mathrm{SQ}^{(2016)} + \mathrm{SQ}^{(2017)} + \mathrm{SQ}^{(2018)}\right) + \frac{3}{2}\left(\mathrm{SQ}^{(2018)} - \mathrm{SQ}^{(2016)}\right) \tag{18}$$

Liegt der auslösende Faktor für 2018 außerhalb etwa des Intervalls [0,95,1,05], so müssen alle Prämien für den zugehörigen Tarif überprüft und gegebenenfalls in 2020 mit der Zustimmung eines Treuhänders angespasst werden, andernfalls dürfen sie nicht angepasst werden.

Unsere Berechnungen zur Vorhersage des auslösenden Faktors werden mit Daten für einen privaten Krankheitskostenvollversicherungstarif durchgeführt. Alle benötigten Parameter wurden von der Debeka Krankenversicherung zur Verfügung

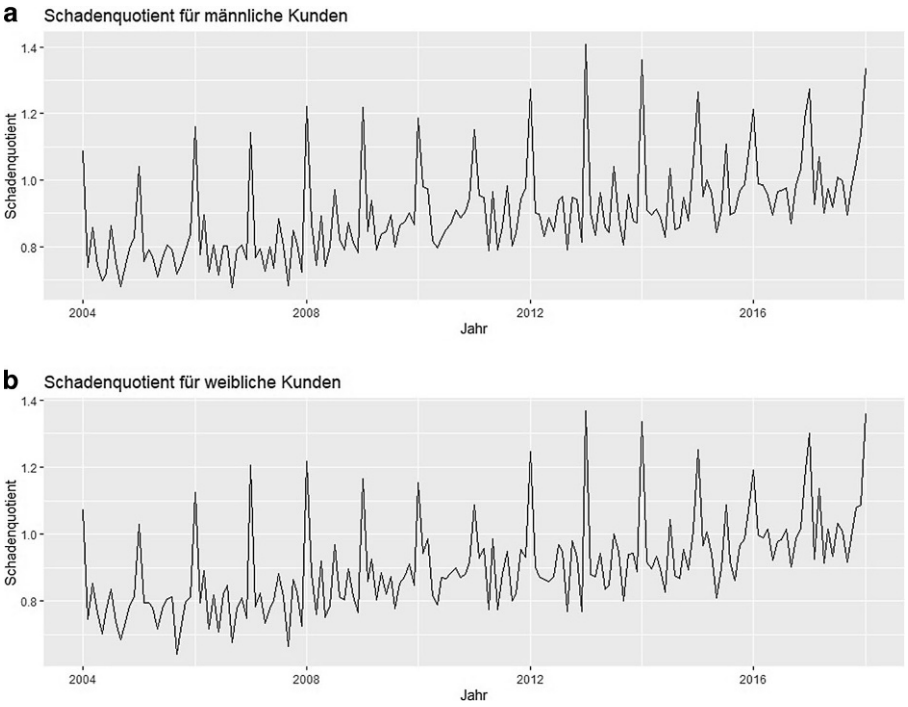

Abb. 2 Schadenquotient: Monatliche Zeitreihe von 2004–2018 für männliche und weibliche Kunden

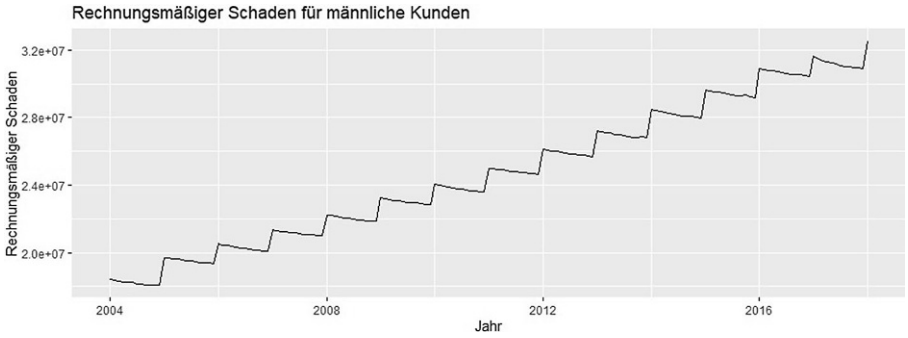

Abb. 3 Rechnungsmäßiger Schaden für männliche Kunden

gestellt. Dabei betrachten wir vier Datensätze der Schadenquotienten und der rechnungsmäßigen Schäden für die Zeit von Januar 2004 bis Dezember 2018, die jeweils 180 monatliche Daten getrennt nach männlicher und weiblicher Population enthalten.

Die Abb. 2 zeigt den Schadenquotient für männliche und weibliche Versicherte. Beide Zeitreihen weisen einen wachsenden Trend und ein saisonales Verhalten mit Höchstwerten im Januar und Tiefstwerten im September auf. Dabei lassen sich die Maxima im Januar mit den dann oft eintreffenden Jahresabrechnungen der Versi-

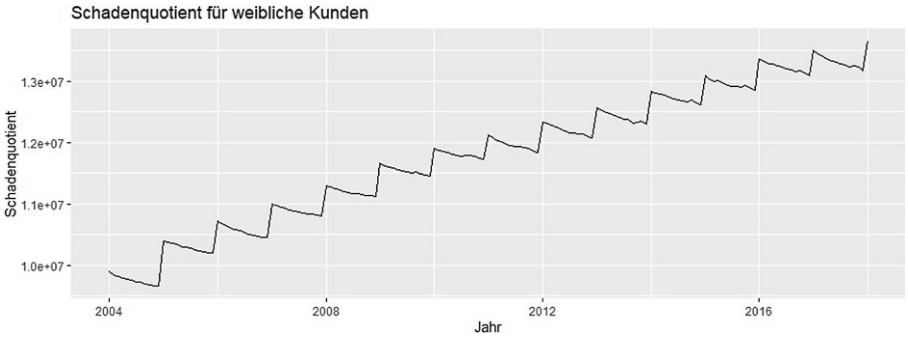

Abb. 4 Rechnungsmäßiger Schaden weibliche Kunden

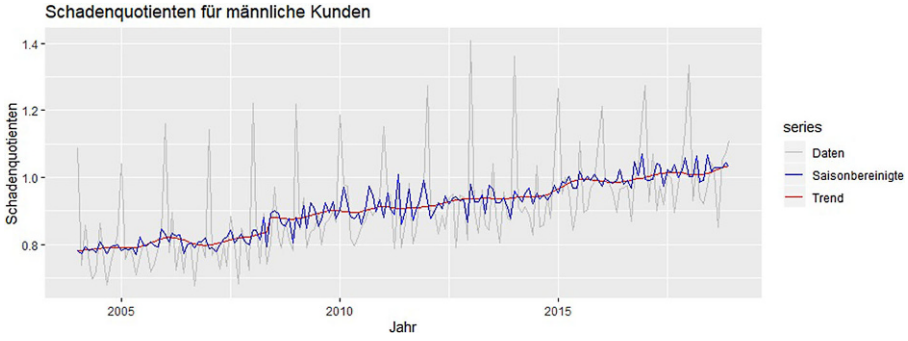

Abb. 5 Schadenquotienten für männliche Kunden: Originaldaten (*grau*), Trendkomponente (*rot*) und saisonbereinigte Daten (*blau*)

cherten erklären. Abb. 3 und 4 zeigt die jeweiligen rechnungsmäßigen Schäden. In beiden Fällen sind Trend und Saisonalität ebenfalls deutlich vorhanden.

3.2 Durchgeführte Untersuchungen

Für die Anwendung neuronaler Netze führen wir zunächst drei Arten der Datenvorverarbeitung durch:

- Trendbereinigung: Mittels der X11-Methode wird ein Trend geschätzt und dann von den Zeitreihen subtrahiert.
- Saisonbereinigung: Die X-13ARIMA-SEATS-Methode wird zur Bereinigung der Zeitreihen von einem saisonalen Einfluss eingesetzt.
- Trend- und Saisonbereinigung: Beide oben genannten Methode werden zur kombinierten Bereinigung von Trend und Saisonalität auf die Originalzeitreihen angewendet.

Es liegen somit vier Typen von Zeitreihendaten vor, die Originaldaten (O), die trendbereinigten (DT), saisonbereinigten (DS) und die trend- und saisonbereinigten Zeitreihen (DSDT), auf die neuronale Netze zur Vorhersage angewendet wer-

den. Abb. 5 zeigt die Originaldaten, die Trendkomponente und die saisonbereinigte Zeitreihe für die monatlichen Schadenquotienten männlicher Kunden.

Im Gegensatz zu den neuronalen Netzen werden die SARIMA-Methode, Regression und die Regression mit SARIMA-Fehlertermen nur auf den Originaldaten untersucht. Mittels des iterativen Prozesses der Modellwahl, -anpassung und -diagnose wird in den drei genannten Ansätzen jeweils das Modell mit der besten Anpassung an die Daten als Vorhersagemodell verwendet. Die ausgewählten Vorhersagemodelle werden zusätzlich im hybriden Ansatz zur Modellierung des linearen Teils der Daten verwendet, während ein neuronales Netz die entstandenen Residuen der lineare Teile modellieren soll.

Zur Bestimmung der besten Struktur eines neuronalen Netzes für jede Zeitreihe wird die Methode der Kreuzvalidierung angewendet. Hierfür werden die vorhandenen Daten in eine Traingsmenge und eine Validierungsmenge aufgeteilt, wobei die Validierungsmenge aus den Daten der letzten 12 Monate besteht und der Rest die Traingsmenge darstellt. Wir variieren die Anzahl der Neuronen im Hidden Layer von 2 bis 14. Die Anzahl der Inputknoten wird mit Hilfe der partiellen Autokorrelationsfunktion (PACF) bestimmt, die den direkten Einfluss einer n Monate zurück liegenden Beobachtung auf die Gegenwart misst.

Für unsere Untersuchungen haben wir uns für ein dreilagiges neuronales Netz mit lediglich einem Hidden Layer entschieden, wie es auch in Abb. 1 zu sehen ist. Es besitzt nur einen Output-Knoten, die Vorhersage des Schadenquotienten bzw. rechnungsmäßigen Schadens. Des Weiteren verwenden wir das Ein-Schritt-Vorhersageverfahren während der Modellanpassung und Modellauswahl mittels Trainings- und Validierungsdaten. Lediglich auf der Testmenge wird ein (iteratives) Mehr-Schritt-Vorhersageverfahren angewendet. Dieses besteht aus wiederholten Ein-Schritt-Verfahren, bei denen im nächsten Schritt den Daten jeweils die Vorhersage aus dem vorangegangen Schritt als zusätzlicher Input hinzugefügt wird. Als Aktivierungsfunktion in den Neuronen der verborgenen Schicht wählen wir die logistische Funktion, während für den Output-Knoten die Identität verwendet wird. Basierend auf der Trainingsmenge wird jedes neuronale Netz dreimal mit verschiedenen, zufällig gewählten Anfangsgewichten trainiert. Das bestangepasste Netz wird dann auch auf der Validierungs- und der Testmenge untersucht, und es werden die zugehörigen Anpassungs- und Vorhersageresultate wiedergeben. Um bei der Minimierung des Fehlerkriteriums die Chance des Erreichens des globalen Minimums zu erhöhen, haben wir den Adam-Optimierer (siehe Kingma und Ba 2014) mit anfänglicher Lernrate von 0,001 und Beschleunigung 0,9 gewählt. Während der Trainingsphase haben wir den Wert der Lernrate bis auf 0,5 gesteigert und dann wieder um einen Faktor 10 verringert, bis wir bessere Werte für den Abstieg erhalten haben. Die maximale Anzahl an Iterationen (*Epochen*) in der Traingsphase wurde auf 1000 festgelegt.

Bei mittels Trend- und Saisonbereinigung vorverarbeiteten Daten wird der aus den neuronalen Netzen erhaltene Output wieder entsprechend zurück transformiert. Die für die Testmenge benötigten Trendwerte werden mit der X11-Methode vorhergesagt, während die saisonalen Werte einfach von denen des letzten Jahres der geschätzten Komponente übernommen werden (siehe Hyndman und Athanasopoulos 2018). Die Anpassung auf den Trainingsdaten sowie die Vorhersageresultate (auf

den Validierungs- und Testdaten) werden für alle Modelle durch die mittlere Fehlerquadratsumme (RMSE), den mittleren absoluten prozentualen Fehler (MAPE) und den mittleren absoluten Fehler (MAE) gemessen.

Genauso wird mit den monatlichen Daten der rechnungsmäßigen Leistung verfahren. Vorverarbeitung der Daten geschieht mit der X11-Methode zur Trendbereinigung und der X-13ARIMA-SEATS Methode für die Saisonbereinigung. Regression mit Trend und saisonalen Vorhersagevariablen und SARIMA-Fehlern als traditionelle Methode sowie hybride Modelle aus einem traditionellen Modell und einem ANN werden ebenfalls betrachtet. Dabei wird das neuronale Netz wie oben bei der Anwendung auf die monatlichen Schadenquotienten erstellt.

4 Numerische Resultate

Tab. 1 listet die Maßzahlen RMSE, MAPE und MAE für Training, Validierung und Test der jeweils besten Modellparametrisierungen der betrachteten Modellklassen zur Vorhersage der monatlichen Schadenquotienten männlicher Versicherter auf.

Dabei erweist sich $SARIMA(2,0,3)(1,1,1)_{12}$ als bestes Zeitreihenmodell. Der Saisonalitätsparameter $s = 12$ ist dabei bei monatlichen Daten naheliegend, die restlichen ARIMA-Parameter zeigen eine recht kurze, aber auch nicht vernachlässigbare Abhängigkeit des vorherzusagenden Schadenquotienten von den direkt vorhergehenden Vorhersagen und Fehlern auf. Basierend auf Gl. (4) kann das Modell als

$$\Phi_1(B^{12})\phi_2(B)(1-B^{12})\log(y_t) = \Theta_1(B^{12})\theta_3(B)\epsilon_t$$

geschrieben werden, wobei es auf die mit dem Logarithmus transformierten Ordiginaldaten angewendet wird und die Darstellungen

$$\begin{aligned}
\Phi_1(B^{12}) &= 1 - \Phi_1 B^{12} \\
\phi_2(B) &= 1 - \phi_1 B - \phi_2 B^2 \\
\Theta_1(B^{12}) &= 1 - \Theta_1 B^{12} \\
\theta_3(B) &= 1 - \theta_1 B - \theta_2 B^2 - \theta_3 B^3
\end{aligned}$$

gelten.

Das beste (klassische) Regressionsmodell (auf Basis der logarithmierten Daten) enthält die Vorhersagevariablen Trend, Saisonalität (vgl. hierfür Abschn. 2.2) sowie die monatlichen Arbeitstage in Rheinland-Pfalz und ab 2016 die monatlichen Arbeitstage und die Anzahl der monatlichen Wochenendtage. Der Grund für diese Variation ab 2016 besteht in der ab dann durch die Debeka Krankenversicherung ermöglichten Einreichung der Kundenrechnungen per Mobile App.

Das beste Regressionsmodell ist die Kombination eines klassischen Regressionsansatzes mit den Arbeitstagen als Prädiktorvariablen und $SARIMA(0,0,0)(0,1,1)_{12}$-Fehlern.

In Tab. 1 fallen einige Resultate ins Auge. Innerhalb der klassischen Modelle erzielen die mit SARIMA-Fehler modellierten Regressionsansätze die besten Re-

Tab. 1 Ergebnisse für den Schadenquotienten männlicher Versicherter

Datentyp	Modell	Training			Validierung			Test		
		RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE
O	ANN	0,0555	4,83	0,0436	0,0499	3,91	0,0395	0,0460	3,78	0,0386
Hybrid	SARIMA mit ANN	0,0440	3,76	0,0339	0,0430	2,69	0,0279	0,0477	3,63	0,0362
Hybrid	Regression mit ANN	0,0340	2,76	0,0255	0,0464	2,95	0,0324	0,0380	3,23	0,0328
Hybrid	Regression mit SARIMA und ANN	0,0379	2,91	0,0270	0,0420	2,83	0,0306	0,0349	2,66	0,0278
DT	ANN	0,0569	4,87	0,0438	0,0587	5,02	0,0509	0,0458	3,85	0,0397
DS	ANN	0,0367	2,75	0,0250	0,0321	2,67	0,0271	0,0383	3,42	0,0340
DSDT	ANN	0,0262	2,11	0,0189	0,0150	1,27	0,0128	0,0335	2,75	0,0275
O	SARIMA	0,0438	3,69	0,0335				0,0468	3,70	0,0359
O	Regression	0,0361	2,89	0,0265				0,0508	4,10	0,0419
O	Regression mit SARIMA	0,0376	2,91	0,0270				0,0350	2,68	0,0280

O = original, *DT* = trendbereinigt, *DS* = desaisonalisiert, *DSDT* = trend- und saisonbereinigt

sultate. Diese Kombination reduziert den Vorhersagefehler und resultiert in einem verbesserten Vorhersagemodell. Auf der anderen Seite schneiden neuronale Netze auf der Basis der Originaldaten sowohl im Hinblick auf Modellierung und Vorhersage schlecht ab. Ihre Performance wird durch eine Vorverarbeitung der Daten mittels Trend- und Saisonbereinigung deutlich verbessert.

Auf den jeweils um Trend und/oder Saisonalität bereinigten Daten sind die Fehler bei den DS- und DSDT-Netzen deutlich geringer als bei den auf den Originaldaten oder den lediglich trendbereinigten Daten trainierten Netzen. Insbesondere hat das DSDT-Netz die kleinsten Fehler bzgl. aller drei verwendeten Fehlermaße. Dies bestätigt auch die Aussage von Zhang Zhang und Min (2005), dass die Trend- und Saisonalitätsbereinigung die wirksamste Datenvorverarbeitung bei der Modellierung und Vorhersage von Zeitreihen mittels neuronaler Netze darstellt.

Tab. 1 beinhaltet auch die Vorhersageresultate der Hybridmodelle. Hier wurden die drei besten Zeitreihenvorhersagen mit (Feedforward) neuronalen Netzen kombiniert. Bei der Kombination des $ARIMA(2,0,3)(1,1,1)_{12}$-Modells mit einem ANN erweist sich eine $2 \times 10 \times 1$-Struktur als optimal, wobei die erste 2 die Anzahl der Inputvariablen, die zweite Zahl die Anzahl der verborgenen Neuronen darstellt und wir lediglich einen Output-Knoten haben. Das zweite Hybridmodell besteht aus der Kombination von Regression mit Trend, Saisonalität, den Werktagen in Rheinland-Pfalz sowie den beiden Zusatzvariablen ab 2016 und einem $4 \times 2 \times 1$ ANN. Schließlich betrachten wir die Kombination aus Regression mit Anzahl der Werktage als Vorhersagevariablen und $SARIMA(0,0,0)(0,1,1)_{12}$-Fehler sowie einem $2 \times 2 \times 1$ ANN.

Die Resultate zeigen, dass ein neuronales Netz allein die Vorhersagegenauigkeit von SARIMA- und Regressionsmodellen verbessern kann, wobei aber Regression mit SARIMA-Fehler den ANN überlegen ist. Das Abschneiden der Hybridmaße zeigt, dass in allen Fällen die Vorhersagefehler gegenüber den Urprungsmodellen deutlich reduziert werden konnten. Wir verweisen hier besonders auf die Kombina-

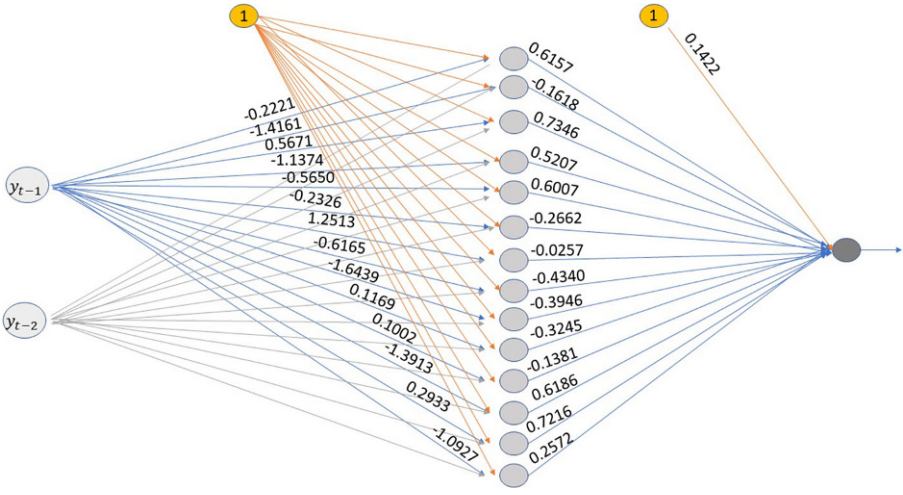

Abb. 6 Neuronale Netzwerkstruktur, die mit trend- und saisonbereinigten Daten trainiert wurde

tion des Regressionsmodells mit ANN, die ein besonders beeindruckende Verbesserung darstellt.

Umgekehrt hat ein auf den trend- und saisonbereinigten Daten trainiertes ANN die beste Vorhersageperformance, was zeigt, dass für den Einsatz von ANN eine geeignete Vorverarbeitung der Daten notwendig ist. Gleichzeitig zeigen diese Resultate auch das Potential der Kombination aktueller Zerlegungsmethoden wie X13-ARIMA-SEATS mit ANN auf.

Die Gleichung und Abb. 6 stellen das auf trend- und saisonbereinigten Daten trainierte Netz dar. Wir veranschaulichen dabei auch einige der 57 trainierten Parameter, wobei wir die übrigen aus Gründen der Übersicht nicht dargestellt haben.

$$y_t = w_0 + \sum_{j=1}^{14} w_j \sigma(w_{0,j} + \sum_{i=1}^{2} w_{i,j} y_{t-i}) + \epsilon_t$$

Die Abb. 7 zeigen die Vergleiche zwischen realisierten und vorhergesagten Werten der besten tradionellen Methoden und solchen, die ANN beinhalten. Hierfür wurden die Daten sukzessive in Trainings- und Vorhersagezeiten aufgeteilt, wobei die jeweils gleiche Struktur der traditionellen Modelle und der Methoden aus Tab. 2 verwendet wurden. Genauer, um z. B. die Werte aus 2013 mit der Mehrschrittvorhersagemethode zu erhalten, wurden die Modelle auf den Daten bis 2012 trainiert. Dabei haben sich die Vorhersagen durch Kombination von ANN und Regression mit SARIMA-Fehlern deutlich verbessert. Die neuronale Netwerkstruktur auf den DSDT-Daten zeigt auch ein sehr gutes Vorhersageverhalten auf den vergangenen Jahren.

Die Struktur aller jeweils optimalen neuronalen Netze für den Schadenquotienten männlicher Versicherter wird in Tab. 2 angegeben. Die Spalte „Lag“ gibt die vergangenen Werte ein, die in das ANN als Input eingehen. So hat z. B. das optimale Netz auf den Originaldaten die folgenden 10 Inputvariablen: $y_{t-1}, y_{t-2}, y_{t-3}, y_{t-6}, y_{t-8}, y_{t-10}, y_{t-12}, y_{t-13}, y_{t-14}, y_{t-15}$.

Tab. 3 zeigt die entsprechenden Modelle für die Schadenquotienten weiblicher Versicherter. Die besten Zeitreihenmodelle sind $SARIMA(2,0,3)(1,1,1)_{12}$, Regression mit den gleichen fünf Variablen wie bei den männlichen Versicherten, sowie Regression mit Arbeitstagen, saisonalen Variablen und $SARIMA(0,0,0)(0,1,1)_{12}$-Fehler. Hier schneidet zunächst das SARIMA-Modell besser ab als die Regression mit SARIMA-Fehler. Allerdings sieht man in Abb. 7, dass dies in den vorangegangenen Jahren nicht der Fall war, so dass insgesamt das Regressionsmodell mit SARIMA-Fehler stabiler als SARIMA allein wirkt. Des Weiteren schneiden hier die ANN nicht besser als die traditionellen Modelle ab.

Kombiniert man neuronale Netze mit den drei Zeitreihenmethoden, zeigt sich in allen Fällen eine Verbesserung bei den Vorhersagefehlern, auch wenn dies im Training nicht zwangsläufig sein muss. Das hybride Modell aus $ARIMA(2,0,3)(1,1,1)_{12}$ und einem neuronalen Netz zeigt im Test insgesamt das beste Verhalten. Es zeigt sich auch bei diesem Datensatz, dass neuronale Netze erst dann richtig gut abschneiden, wenn sie auf den trend- und saisonbereinigten Daten trainiert werden. Dann sind sie mit dem besten Hybridmodell vergleichbar. Die jeweils besten Netzwerkstrukturen werden hier in Tab. 4 wiedergegeben.

Abb. 7 **a** Regression mit SARIMA-Fehler, **b** ANN, **c** Hybrid-Regression mit SARIMA-Fehler und ANN, **d** Trend- und saisonbereinigtes ANN

Tab. 2 Beste neuronale Netzwerkmodelle für den Schadenquotienten männlicher Kunden

Datentyp	Lag	Anzahl der verborgenen Neuronen
O	1–3, 6, 8, 10, 12–15	10
DT	4–12	14
DS	1–3	14
DSDT	1,2	14
Hybrid_1	1,2	10
Hybrid_2	2, 8, 10, 12	2
Hybrid_3	1,2	2

O = Original, DT = trendbereinigte Daten, DS = desaisonalisierte Daten, $DSDT$ = trend- und saisaonbereinigte Daten, $Hybrid_1$ = SARIMA und ANN, $Hybrid_2$ = Regression und ANN, $Hybrid_3$ = Regression mit SARIMA und ANN

Die folgenden Bilder in Abb. 8 zeigen wieder den Vergleich zwischen Vorhersage- und Realwerten der jeweils besten Modelle aus Tab. 3 in den Vorjahren, diesmal für den Schadenquotienten bei weiblichen Kunden.

Auf analoge Art und Weise haben wir die gewählten Modellklassen auch für die Vorhersage der rechnungsmäßigen Schäden angewendet. Dabei erzielen Regression mit Trend und Saisonalität sowie $SARIMA(1,0,0)(2,0,0)_{12}$- bzw. $ARIMA(0,1,0)$-Fehler die besten Ergebnisse bei den traditionellen Modellen für die Vorhersage der rechnungsmäßigen Schäden bei männlichen bzw. weiblichen Versicherten. Diese wurden auch zur Kontrolle zur Vorhersage der vorangegangenen Jahre von 2013 bis 2017 verwendet (vgl. hierzu jeweils das erste und dritte Bild der Abb. 9. Für die Anwendung neuronaler Netze haben wir bei den rechnungsmäßigen Schäden die Daten vor der Modellanpassung nach [0,1] transformiert. Die Modelle, die jeweils am besten bzgl. RMSE, MAPE und MAE abschnitten, waren auf trendbereinigten Daten trainierte $13 \times 6 \times 1$-Netze für die männlichen und $13 \times 2 \times 1$-Netze für die weiblichen Versicherten.

4.1 Vergleich der besten Vorhersagemodelle für den auslösenden Faktor des privaten Krankheitskostenvollversicherungstarifs verschiedener Jahre

Der Zweck unserer Untersuchung bestand in der Ermittlung des besten Vorhersagemodells für den auslösenden Faktor eines Krankheitskostenvollversicherungstarifs. Dabei wurden traditionelle Zeitreihenvorhersagemodelle und neuronale Netze verglichen. Unter den traditionellen Methoden hat sich die Regression mit SARIMA-Fehlern als der beste Ansatz erwiesen. Bei den neuronalen Netzen konnten erst durch Datenvorverarbeitung in Form von Trend- und Saisonbereinigung vergleichbar gute Resultate erzielt werden, die dann sogar teilweise die besten traditionellen Verfahren klar übertrafen. Als eine weitere innovative Komponente haben wir hybride Modelle betrachtet, bei denen die linearen Teile der Vorhersage von traditionellen Verfahren übernommen werden und die Residuen mittels neuronalen Netzen erklärt werden sollen.

Dabei erzielten auf DSDT-Daten trainierte neuronale Netze die besten Resultate für die Schadenquotienten männlicher Versicherter, während die hybride Methode

Tab. 3 Ergebnisse für den Schadenquotienten weiblicher Versicherter

Datentyp	Modell	Training			Validierung			Test		
		RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE
O	ANN	0,0595	5,32	0,0478	0,0703	5,54	0,0582	0,0533	3,61	0,0379
Hybrid	SARIMA mit ANN	0,0451	3,95	0,0356	0,0465	3,42	0,0358	0,0301	2,64	0,0268
Hybrid	Regression mit ANN	0,0347	2,94	0,0268	0,0442	2,98	0,0329	0,0420	3,55	0,0364
Hybrid	Regression mit SARIMA und ANN	0,0369	3,10	0,0281	0,0370	3,07	0,0320	0,0395	3,21	0,0326
DT	ANN	0,0628	5,47	0,0493	0,0795	6,55	0,0679	0,0507	3,47	0,0370
DS	ANN	0,0378	3,19	0,0285	0,0297	2,42	0,0250	0,0372	2,85	0,0294
DSDT	ANN	0,0320	2,62	0,0234	0,0254	2,23	0,0229	0,0396	2,75	0,0285
O	SARIMA	0,0449	3,89	0,0352				0,0305	2,72	0,0275
O	Regression	0,0351	2,97	0,0269				0,0455	3,87	0,0400
O	Regression mit SARIMA	0,0371	3,05	0,0276				0,0399	3,28	0,0337

O = original, *DT* = trendbereinigt, *DS* = desaisonalisiert, *DSDT* = trend- und saisonbereinigt

Tab. 4 Beste neuronale Netzwerkmodelle für den Schadenquotienten weiblicher Kunden

Datentyp	Lag	Anzahl der verborgenen Neuronen
O	1–3, 6, 8, 10, 12–15	4
DT	4–13	14
DS	1–5	12
DSDT	1,2	14
Hybrid_1	1,2	8
Hybrid_2	2,12	4
Hybrid_3	2,4	10

O = Original, *DT* = trendbereinigte Daten, *DS* = desaisonalisierte Daten, *DSDT* = trend- und saisonbereinigte Daten, $Hybrid_1$ = SARIMA und ANN, $Hybrid_2$ = Regression und ANN, $Hybrid_3$ = Regression mit SARIMA und ANN

aus SARIMA und neuronalem Netz für den Schadenquotienten weiblicher Versicherter am besten abschnitt. Bzgl. der Vorhersage der rechnungsmäßigen Schäden zeigten auf trendbereinigten Daten trainierte neuronale Netze sowohl für männliche als auch weibliche Kunden das beste Verhalten.

Die Tab. 5 und 6 beinhalten die *absoluten Fehler* zwischen den wahren und den vorhergesagten Werten (siehe auch die Abb. 7, 8 und 9) des auslösenden Faktors in den verschiedenen Jahren, wobei mit *Modell 1* Regression mit SARIMA-Fehler als beste traditionelle Vorhersagemethode bezeichnet wird. Dabei werden die Schadenquotienten für männliche und weibliche Versicherte durch Regression auf Arbeitstag und $SARIMA(0,0,0)(0,1,1)_{12}$-Fehler vorhergesagt, während die rechnungsmäßigen Schäden mittels Regression mit Trend- und Saisonalitätsvariablen und $SARIMA(1,0,0)(2,0,0)_{12}$-Fehler (für männliche Versicherte) bzw. $ARIMA(0,1,0)$-Fehler (für weibliche Versicherte) vorhergesagt werden.

Auf der anderen Seite waren jeweils neuronale Netze, die mit trend- und saisonbereinigten Daten trainiert wurden, die Modelle, die die beste Vorhersagequalität für den Schadenquotienten männlicher und weiblicher Versicherter lieferten. Für diese von uns mit *Modell 2* bezeichneten Verfahren wurde bei männlichen Versicherten eine $2 \times 14 \times 1$-Netzwerkstruktur und ein hybrides Modell aus einer $2 \times 8 \times 1$-Netzwerkstruktur kombiniert mit einem $SARIMA(2,0,3)(1,1,1)_{12}$-Modell für weibliche Versicherte verwendet. Die rechnungsmäßigen Schäden wurden mit auf trendbereinigten Daten trainierten neuronalen Netzen vorhergesagt, wobei sich bei den männlichen Versicherten eine $13 \times 6 \times 1$-Struktur und bei weiblichen Versicherten eine $13 \times 2 \times 1$-Struktur als optimal erwiesen.

Auch wenn Modell 1 insgesamt geringfügig kleinere absolute Fehler als Modell 2 produziert, soll festgehalten werden, dass beide Modelle jeweils im Hinblick auf eine Überprüfung des Krankenversicherungstarifs zu denselben Entscheidungen führen. Dabei sind in sieben von acht Fällen die Vorhersagen richtig, ob eine Überprüfung durchzuführen ist, d. h., ob der auslösende Faktor nicht mehr in [0,95,1,05] liegen wird. Diese Übereinstimmung bei der Vorhersage, dass eine Überprüfung kommen wird, ist für die Krankenversicherung sehr wichtig, denn sie gibt dem Versicherer eine hohe Planungssicherheit, dass Überprüfungen durchzuführen sein werden.

Abb. 8 **a** Regression mit SARIMA-Fehler, **b** ANN, **c** Hybrid-SARIMA mit ANN, **d** Trend- und saisonbereinigtes ANN

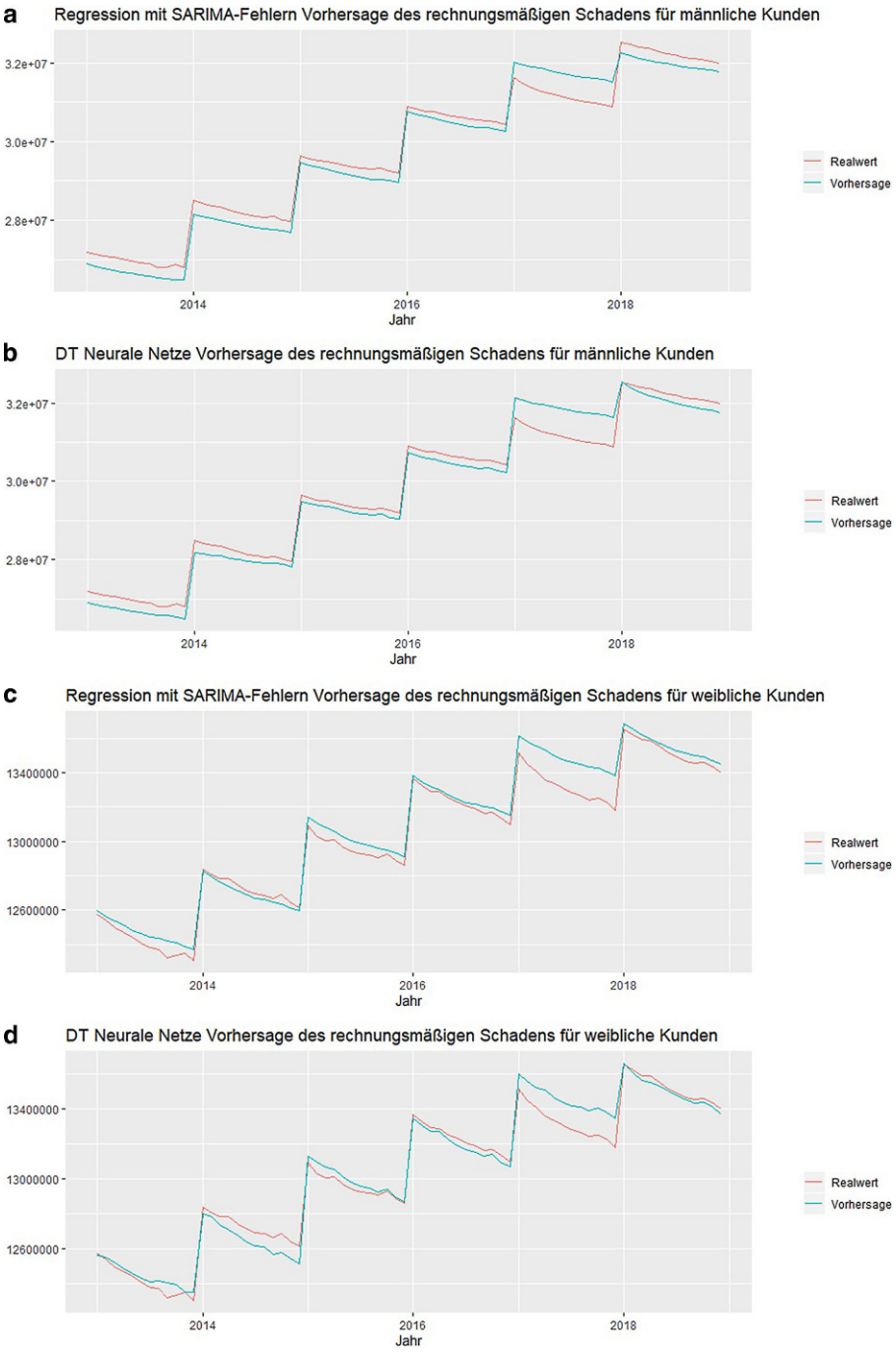

Abb. 9 Vorhersagen der rechnungsmäßigen Schäden mittels Regression mit SARIMA-Fehler und neuronalem Netz auf der Basis trendbereinigter Daten (männliche und weibliche Versicherte)

Tab. 5 Vergleich der vorhergesagten auslösenden Faktoren männlicher Kunden

Jahr	wahrer Wert	Modell 1	abs. Fehler	Modell 2	abs. Fehler
2015	1,0406	1,0262	0,0145	1,0107	0,0299
2016	1,0563	1,0546	0,0016	1,0737	0,0174
2017	1,0318	1,0572	0,0255	1,0638	0,0320
2018	1,0462	1,0515	0,0053	1,0503	0,0041

Tab. 6 Vergleich der vorhergesagten auslösenden Faktoren weiblicher Kunden

Jahr	wahrer Wert	Modell 1	abs. Fehler	Modell 2	abs. Fehler
2015	1,0216	1,0364	0,0149	0,9886	0,0330
2016	1,0716	1,0520	0,0196	1,0564	0,0152
2017	1,0795	1,0591	0,0204	1,0740	0,0055
2018	1,0628	1,0793	0,0165	1,0926	0,0299

Zwar ist die annähernd gleiche, gute Performance der beiden Modelle sehr erfreulich, doch haben auch nicht alle Krankenversicherungstarife die gleichen Datenmuster wie der untersuchte Tarif. So kann beispielsweise keine erfolgreiche Desaisonalisierung als Vorverarbeitung möglich sein oder aber eine hohe Nichtlinearität vorliegen, die für die klassischen Verfahren nicht abbildbar ist. Und natürlich sind auch die persönlichen Vorkenntnisse und Vorlieben der jeweiligen Modellanwender ein nicht zu vernachlässigendes Argument für die Wahl des anzuwendenden Modellrahmens.

Es soll aber auch festgehalten werden, dass wir mit der Klasse der neuronalen Netze und der Hybrid-Verfahren zwei neue Ansätze eingeführt haben, die eine veritable Alternative zur lang etablierten Box-Jenkins-Methodik darstellen und diese in unseren Anwendungen übertreffen.

5 Zusammenfassung und Schlussfolgerungen

Für die Vorhersage des auslösenden Faktors eines Krankenversicherungstarifs haben wir verschiedenste Zeitreihenvorhersagemodelle (Regression, SARIMA-Modell und Regression mit SARIMA-Fehler) mit neuronalen Netzen und Hybrid-Modellen, bei denen neuronale Netze mit klassischen Zeitreihenmethoden kombiniert wurden, verglichen. Dabei schneiden neuronale Netze dann sehr gut ab, wenn die Trainingsdaten mittels Trend- und Saisonbereinigung vorverarbeitet werden.

Als Fallstudie für unsere Untersuchung verwendeten wir monatliche Daten der Debeka Krankenversicherung, die nach männlichen und weiblichen Versicherten getrennt waren. Dabei waren jeweils sowohl die Schadenquotienten als auch die rechnungsmäßigen Schäden vorherzusagen. Dabei verwendeten wir als Maße zur Beurteilung der Vorhersagequalität die Wurzel der mittleren Fehlerquadratsumme (RMSE), den mittleren absoluten prozentualen Fehler (MAPE) und den mittleren absoluten Fehler (MAE).

Insgesamt lässt sich aufgrund unserer Studie feststellen, dass

- moderne Ansätze aus dem Bereich der neuronalen Netze sehr attraktive Methoden für die Vorhersage des auslösenden Faktors sind, wenn sie auf trend- und saisonbereinigten Daten trainiert und validiert werden,
- Hybridansätze, die neuronale Netze mit klassischen linearen Vorhersageverfahren (lineare Regression, ARIMA-Modelle) verknüpfen, die Vorhersageresultate der jeweiligen klassischen Verfahren übertreffen,
- die lineare Regression mit SARIMA-Fehlern, die eher selten untersucht wird, oft populären klassischen Zeitreihenmodellen überlegen ist.

Eine weitere Erkenntnis unserer Untersuchungen besteht darin, dass

- eine naive Anwendung neuronaler Netze auf Daten, die einen (leichten) Trend und eine ausgeprägte Saisonalität aufweisen, in der Regel zu schlechteren Resultaten als eine Anwendung einfacher Zeitreihenmodelle führt.

Gerade diese letzte Aussage zeigt, dass es nicht ratsam ist, vollkommen dem datenbasierten Lernen zu vertrauen, wenn man bereits klare Zusatzinformation über das Verhalten der Zeitreihe besitzt. Eine Datenvorverarbeitung, die offensichtliche Effekte eliminiert, kann als ein Analogon zur Verwendung von Varianzreduktionsverfahren bei der Monte Carlo-Methode angesehen werden. Sie führt in der Regel ebenfalls zur Varianzreduktion der Vorhersageverfahren (vgl. auch Zhang und Min 2005).

Eine weitere wichtige Botschaft unserer Studie ist aber auch, dass es kein gleichmäßig bestes Vorhersageverfahren für den auslösenden Faktor über die Jahre hinweg zu geben scheint. Zwar geht die Tendenz hin zu den oben getroffenen Aussagen, aber zum einen sind die Unterschiede zwischen den guten Verfahren nicht wirklich groß und zum anderen lässt das dem Anwender auch Freiraum, um sich aus den von uns empfohlenen Methode die herauszusuchen, die der persönlichen Vorkenntnis am nächsten kommt oder für die die geeignete Sofwareausstattung zur Implementation vorhanden ist. Gleichzeitig erscheint es aber auch aufgrund dieser Situation eine gute Wahl zu sein, mehr als ein Vorhersageverfahren zu verwenden, um die Stabilität der Vorhersage zu erhöhen bzw. festzustellen, dass es keine eindeutige Tendenz unter den Vorhersagen gibt und somit vorsichtig zu agieren ist.

Unsere Arbeit speziell im Bereich der neuronalen Netze kann als erster Ausgangspunkt ihrer Anwendung zur Vorhersage des auslösenden Faktors (und weiterer Vorhersageprobleme der privaten Krankenversicherung) angesehen werden. Die Eignung weiterer spezieller Architekturen neuronaler Netze wie z. B. rekurrenter Netze oder die Suche nach optimalen Hyperparamertern (wie z. B. die Anzahl weiterer Hidden Layer) können interessante Aspekte zukünftiger Forschung zur Vorhersage des auslösenden Faktors sein. Gleiches gilt auch für die von uns verwendeten hybriden Modellansätze.

Funding Open Access funding enabled and organized by Projekt DEAL.

Literatur

Andrew, V.M., Paul, S.C.: Introductory Time Series with R. Springer, Berlin, Heidelberg, New York (2009)
Becker, T.: Mathematik der privaten Krankenversicherung. Springer, Berlin, Heidelberg, New York (2017)
Bell, W.R., Hillmer, S.C.: Issues involved with the seasonal adjustment of economic time series. J. Bus. Econ. Stat. **2**(4), 291–320 (1984)
Box, G.E., Jenkins, G.M.: Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco (1976)
Burman, J.P.: Seasonal adjustment by signal extraction. J R Stat Soc Ser A Stat Soc **143**(3), 321–337 (1980)
Burnham, K.P., Anderson, D.R.: Multimodel inference: understanding aic and bic in model selection. Sociol. Methods Res. **33**(2), 261–304 (2004)
Chollet, F., Allaire, J.J.: Deep Learning mit R and Keras: Das Praxis-Handbuch von den Entwicklern von Keras and RStudio. MITP, Bonn (2018)
Dagum, E.B.: Modelling, forecasting and seasonally adjusting economic time series with the x-11 arima method. J R Stat Soc SerD **27**(3/4), 203–216 (1978)
Dagum, E.B., Bianconcini, S.: Seasonal Adjustment Methods and Real Time Trend-Cycle Estimation. Springer, Cham (2016)
Findley, D.F., Monsell, B.C., Bell, W.R., Otto, M.C., Chen, B.C.: New capabilities and methods of the x-12-arima seasonal-adjustment program. J. Bus. Econ. Stat. **16**(2), 127–152 (1998)
Hosseini, H.G., Luo, D., Reynolds, K.J.: The comparison of different feed forward neural network architectures for ecg signal diagnosis. Med. Eng. Phys. **28**(4), 372–378 (2006)
Hyndman, R.J., Athanasopoulos, G.: Forecasting: Principles and Practice. OTexts, (2018)
Khashei, M., Bijari, M.: An artificial neural network (p, d, q) model for timeseries forecasting. Expert Syst Appl **37**(1), 479–489 (2010)
Kingma, D.P., Ba, J.: Adam: A method forx stochastic optimization. Proceedings of the 3rd International Conference on Learning Representations (ICLR). Preprint at https://arxiv.org/abs/1412.6980 (2014)
McClelland, J.L., Rumelhart, D.E., et al.: Parallel distributed processing. Explor. Microstruct. Cogn. **2**, 216–271 (1986)
Milbrodt, H., Röhrs, V.: Aktuarielle Methoden der deutschen Privaten Krankenversicherung Bd. 34. VVW GmbH, (2016)
Monsell, B.: The x-13a-s seasonal adjustment program. In: Proceedings of the 2007 Federal Committee On Statistical Methodology Research Conference. http://www.fcsm.gov/07papers/Monsell.II-B.pdf (2007). Zugegriffen: 20. Jan. 2020
Nagelkerke, N.J., et al.: A note on a general definition of the coefficient of determination. Biometrika **78**(3), 691–692 (1991)
Persons, W.M.: Indices of business conditions. Rev. Econ. Stat. **1**, 5–107 (1919)
Pitacco, E.: Health Insurance. Basic Actuarial Models. Springer, Cham (2014)
Royston, J.P.: An extension of shapiro and wilk's w test for normality to large samples. J R Stat Soc Ser C Appl Stat **31**(2), 115–124 (1982)
Sheather, S.: A Modern Approach to Regression with R. Springer, Berlin, Heidelberg, New York (2009)
Shiskin, J., Young, A.H., Musgrave, J.C.: The x-11 variant of the census ii seasonal adjustment program. Technical report, Bd. 15. US Department of Commerce, Bureau of Economic Analysis, Washington (1967)

Zhang, G., Patuwo, B.E., Hu, M.Y.: Forecasting with artificial neural networks: The state of the art. Int J Forecast **14**(1), 35–62 (1998)
Zhang, G.P.: Time series forecasting using a hybrid arima and neural network model. Neurocomputing **50**, 159–175 (2003)
Zhang, G.P., Min, Q.: Neural network forecasting for seasonal and trend time series. Eur J Oper Res **160**(2), 501–514 (2005)