FernUniversität in Hagen

Fakultät für Wirtschaftswissenschaft

Seminararbeit zur Erlangung des Grades eines Bachelor of Science

über das Thema:

Imputationsverfahren für fehlende Daten

Eingereicht bei:	Univ.-Prof. Dr. Robinson Kruse-Becher
Betreuer:	Univ.-Prof. Dr. Robinson Kruse-Becher
von cand.rer.oec.:	Frank Müller
Matr. - Nr.:	9053395
Anschrift:	Calle Castillo 21a 29680 Estepona Spanien
E-Mail:	frank.e.mueller@googlemail.com
Abgabedatum:	2. Januar 2023

Inhaltsverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis:

ACF Autocorrelation function

KDD Knowledge Discovery in Databases

LOCF Last observed value carried forward

MAR Missing At Random

MCAR Missing Completely At Random

MNAR Missing Not At Random

MSE Mean Squared Error

NOCB Next Observation Carried Backward

PACF Partial autocorrelation function

1. Einleitung: Knowledge Discovery in Databases

In den letzten Jahrzehnten stieg mit der Verbreitung des Internets und der Digitalisierung die Verfügbarkeit von Daten und Informationen stark an. Um aus Daten wichtige Informationen zu extrahieren und in Wissen umzuwandeln gibt es viele Methoden. Eine dieser Methoden ist Knowledge Discovery in Databases (KDD):

„*Knowledge Discovery in Databases* is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data." (Fayyad, Piatetsky-Shapiro, & Smyth, 1996)

Nach Fayyad finden sich Anwendungsbereiche für KDD im Geschäftsleben: im Marketing, im Investmentbereich, für Fraud Detection, in der Industrie, der Telekommunikation und der Datenaufbereitung. KDD wurde erstmalig geprägt durch Piatetsky Shapiro in einem Workshop 1989.

Im Folgenden wird der KDD-Prozess, bestehend aus neun Schritten, näher beschrieben (nach Fayyad et al., 1996, S. 6):

Beim ersten Schritt wird das Problem definiert. Man versucht die Domaine zu verstehen, sich das relevante Vorwissen zu erarbeiten und identifizieren des Ziels von dem KDD-Prozess aus Kundensicht.

Anschließend wird im zweiten Schritt das Dataset zu erzeugt. Dazu gehört die Auswahl der relevanten Daten aus den verfügbaren Daten.

Nun müssen im dritten Schritt die Daten bereinigt und aufbereitet werden. Grundlegende Operationen enthalten die Entfernung von Rauschen, das Sammeln von Informationen, um Rauschen zu modellieren, das Festlegen der Strategien für den Umgang mit fehlenden Daten unter Berücksichtigung von Zeitreiheninformation und bekannten Veränderungen. Die Seminararbeit widmet sich hier den Imputationsverfahren für fehlende Daten in Zeitreihen.

Anschließend versucht man gegebenenfalls die Daten zu reduzieren und zu projizieren, um nützliche Merkmale bezüglich der Aufgabenstellung zu entdecken. Hier bietet es sich auch an, die effektive Anzahl der Variablen durch Dimensionsreduktion oder Transformation zu reduzieren.

Im fünften Schritt wird die Data Mining Methode ausgewählt. Die Fragestellung aus Schritt eins und die Zielsetzung wird in Übereinstimmung gebracht.

Im sechsten Schritt geht es um die explorative Analyse und der Auswahl der Hypothese und des Models. Hier werden auch konkrete Data Mining Algorithmen und Methoden ausgewählt, um nach Mustern in den Daten zu suchen.

Das Data Mining im siebten Schritt sucht gezielt nach interessanten Mustern, die im achten Schritt interpretiert werden. Gegebenenfalls werden die letzten beiden Schritte wiederholt, um das Ergebnis zu verbessern.

Im neunten Schritt wird das entdeckte Wissen verwendet. Entweder es wird direkt verwendet, dokumentiert oder weitergegeben.

Im dritten Schritt 3 geht es um die Bereinigung und die Aufbereitung von Daten. Das Ziel der Datenbereinigung ist die Steigerung der Datenqualität. Kurzfristig schafft sie temporäre Abhilfe von Datenqualitätsproblemen. Für eine langfristige Sicherung der Datenqualität braucht man ein systematisches Datenqualitätsmanagement.

Bei der Bereinigung gibt es drei Fehlerklassen nach (Müller & Freytag, 2003):

- Bei semantischen Fehlern stimmt die Semantik der Daten nicht. Beispiele sind Verschmutzungen, Noisy Data und Redundanzen.
- Bei Coverage Fehlern ist der durch die Daten beschriebene Realitätsausschnitt kleiner als der ursprünglich angenommene. Dazu gehören fehlende Werte, Nullwerte und unvollständige Werte.
- Syntaktische Fehler sind Fehler in Form der Daten, die es unmöglich machen, die Daten korrekt zu interpretieren. Unregelmäßigkeiten und unzulässige Werte sind Beispiele für syntaktische Fehler.

Treten nun fehlende Werte auf so hat man nach (Han & Kamber, 2006) S.88 folgende Möglichkeiten:

- Man kann den betroffenen Datensatz ignorieren. Dadurch wird bei den meisten verwendeten Programmen der Datensatz gelöscht und steht damit der weiteren Verwendung nicht mehr zur Verfügung.
- Die andere Möglichkeit wäre eine Imputation. Das Ziel hier ist das Erzeugen eines Ersatzdatensatzes, der dem vollständigen Datensatzes möglichst ähnlich ist. Das Ziel hier sind möglichst gute Ergebnisse hinsichtlich des KDD-Prozesses. Das imputierte Data Set sollte möglichst unverzerrt sein.
- Die fehlenden Werte können manuell eingetragen (imputiert) werden. Je nach Aufgabenstellung und dem Fachwissen des Bearbeiters kann dies zu guten Ergebnissen führen. Diese Methode ist aber nicht automatisierbar und stößt auf Probleme bei großen und komplexen Datenbeständen.
- Des Weiteren kann man eine globale Konstante oder einen statistischen Wert für das Attribut wie einen Median oder Mittelewert nutzen. Diese beiden Methoden führen aber womöglich zu einer starken Verzerrung der statistischen Eigenschaften des Datensatzes.
- In vielen Fällen bietet es sich an, den wahrscheinlichsten Wert mittels Regression, Bayes-Inferenz oder Entscheidungsbäumen zu ermitteln.

2 Literaturübersicht / Theoretischer Hintergrund

2.1 Grundlegend statistische Begrifflichkeiten

2.1.1 Einige Begrifflichkeiten aus der Querschnitts-Statistik

Eine Stichprobe setzte sich in der Querschnitts- Statistik (cross-sectional) aus unabhängigen und identisch verteilten statistischen Einheiten zusammen, so dass sich die Streuung von Schätzern unter der Ausnutzung dieser Unabhängigkeit umgekehrt proportional zur Stichprobengröße verringert. Wenn die Veränderung einer Variablen im Zeitverlauf beobachtet wird, spricht man von einer Zeitreihe. Aus der Kombination von Querschnittsdaten und Zeitreihen entstehen Paneldaten. Wenn die Beobachtungen nur ein Merkmal enthalten, bezeichnet man die Daten als univariat. Bei multivariaten Daten kann man mehrere Merkmale gleichzeitig betrachten. Dadurch kann man das gleichzeitige Auftreten von Merkmalskombinationen untersuchen.

2.1.2 Einige statistische Besonderheiten von Zeitreihen

Nach Mazzoni ist eine Zeitreihe „eine Manifestation eines darunterliegenden (stochastischen) Prozesses. […] Die Abhängigkeit der Zeitreihenwerte wird mit Hilfe ihrer Kovarianz bzw. ihrer Korrelation charakterisiert."

X und Y seien zwei Zufallsvariablen, die nicht voneinander abhängen, dann ergibt sich die Kovarianz zwischen ihnen aus

$$Cov[X,Y] = E[(X - E[X])(Y - E[Y])]$$

Und die Korrelation ρ als eine Normierung der Kovarianz im Bereich [-1,1]

$$\rho[X,Y] = \frac{Cov[X,Y]}{\sqrt{Var[X]}\sqrt{Var[Y]}}$$

Zeitreihen enthalten oft Trends (langfristige Einflüsse) und Saisonalität (periodisch wiederkehrende Einflüsse), die entsprechend modelliert werden müssen.

2.2 Ausfallmechanismen nach Rubin

Wenn in einem Dataset fehlende Werte auftreten, unterscheidet Rubin drei verschiedene Ausfallmechanismen: Missing Completely At Random (MCAR), Missing At Random (MAR) und Missing Not At Random (MNAR). (Little & Rubin, 2020), (Rubin, 1987)

2.2.1 Missing Completely At Random (MCAR)

„Ein vollständig zufälliger Datenausfall (MCAR) liegt dann vor, wenn die Wahrscheinlichkeit fehlender Werte in einer bestimmten Variable weder einen Zusammenhang mit den Ausprägungen dieser Variable selbst noch mit den Ausprägungen irgendeiner anderen in der Umfrage erhobenen Variable aufweist. Sind diese beiden Bedingungen für jede Variable erfüllt, können die Befragten mit vollständigen Beobachtungen als eine einfache Zufallsstichprobe aus den erhobenen Umfragedaten betrachtet werden.“ (Tausendpfund, 2020) In unserem Fall liegt aber keine Umfrage vor, sondern die fehlenden Werte werden manuell erzeugt.

2.2.2 Missing At Random (MAR)

„Von einem zufälligen Datenausfall (MAR) spricht man, wenn die Wahrscheinlichkeit fehlender Werte zwar mit den Ausprägungen anderer beobachteter Variablen zusammenhängt, aber nicht von der Ausprägung der Variable selbst beeinflusst ist. In diesem Fall kann das Fehlen von Werten durch weitere beobachtete Variablen im Datensatz erklärt werden.“ (Tausendpfund, 2020)

2.2.3 Missing Not At Random (MNAR)

„Ein nicht zufälliger Datenausfall (MNAR) liegt schließlich vor, wenn“ die Ausfallwahrscheinlichkeit „von den fehlenden Daten selbst abhängt.“ (Tausendpfund, 2020)

2.3 Behandlung fehlender Werte in Querschnittsdaten

2.3.1 Taxonomie von „Missing Data Methods“ nach Rubin und Little

(Little & Rubin, 2020) beschreiben Methoden für die Behandlung von fehlenden Daten, die sie aus der Fachliteratur abgeleitet haben:

- Verfahren, die auf nur komplett vorhandenen Datensätzen beruhen: hier werden die nicht verfügbaren Datensätze werden ignoriert (complete-case analysis). Diese sind einfach zu implementieren und funktionieren gut bei kleinen Datenmengen. Trotzdem kann es zu ernsthaften Verzerrungen kommen.
- Gewichtete Verfahren: die gesampelten Daten werden anhand ihrer Auftretenswahrscheinlichkeit gewichtet.

- Imputationsverfahren: die fehlenden Werte werden eingefügt und die vervollständigten Daten werden anschließend mit den Standardverfahren analysiert. Die meisten in dieser Arbeit verwendeten Algorithmen fallen in diese Kategorie.
- „Model-based methods“: es wird ein Model für die vollständigen Daten und die zugrunde gelegte Interferenz der Wahrscheinlichkeitesverteilung festgelegt. Die Parameter werden z.B. durch Maximum Likelihood geschätzt. Der Algorithmus Amelia fällt in diese Kategorie.

2.3.2 Ausschluss (Listwise deletion)

Bei der Ausschlussmethode, auch Complete Case analysis oder listwise deletion, werden alle Datensätze mit fehlenden Werten gelöscht. Die MCAR-Standardfehler und Signifikanzgrenzen bleiben korrekt, sind aber oft größer relativ zu den gesamten Daten.

Besonders wenn es viele Variablen gibt, entsteht ein großer Datenverlust. Sollte der Ausfallmechanismus nicht MCAR sondern MAR oder MNAR sein, kann der Ausschluss zu einer starken Verzerrung der der Mittelwerte, Regressionskoeffizienten und Korrelationen führen.

Trotzdem kann die Ausschlussmethoden in bestimmten Anwendungsfällen zu den besten Ergebnissen führen. Dazu benötigt man eine große Datenmenge, MCAR- Fehler, wenige fehlende Daten und wenige Ausreißer. (Buuren, 2018)

2.3.3 Ersetzen durch den Mittelwert/ Median

Die fehlenden Daten werden durch den Mittelwert oder den Median ersetzt.

$$\overline{x} = \sum_{n=1}^{m} x_n$$

Da der Mittelwert/ Median gleichbleibt, hat dieses Verfahren einen guten MSE.

Mit einem starken Peak im Mittelwert / Median ändert sich die Häufigkeitsverteilung der beobachteten Daten, die Standardabweichung wird kleiner und auch die Korrelationen zwischen den Variablen ändern sich, was alles zu einer Verzerrung der Daten führt.

2.3.4 Ersetzen durch Regression

Nach (Little & Rubin, 2020) kann man bei multivariaten Datasets auch die anderen Variablen miteinbeziehen, um bessere Imputationsergebnisse zu bekommen. Dabei kann man verschiedene Regressionsmodelle verwendet wie lineare Regression, Random Forests.

Unter MCAR bekommt man so ein möglichst unverzerrte Schätzung. Falls die Faktoren, die den Datenverlust beeinflussen, Teil des Modells sind, dann sind die

Regressionsparameter auch unverzerrt unter MAR. Bei einer hohen Modellqualität können die imputierten Werte die tatsächlichen Werte fast perfekt schätzen.

Ein Problem stellt hier aber immer noch die schlechte Schätzung der Variabilität. Diese hängt von der erklärten Varianz und dem Anteil der fehlenden Werten ab (Little and Rubin, 2002). Es können außerdem Beziehungen in den Daten künstlich verstärkt werden und somit die Daten verzerren.

2.3.5 Ersetzen durch statistische Regression

Nach (Buuren, 2018, S. 15) kann man die Ergebnisse der Regression verbessern, indem man der Schätzung Rauschen hinzufügt. Dadurch wird teilweise die Verzerrung durch die Korrelationen berücksichtigt. Die Regressionsparameter wie auch die Korrelationen zwischen den Variablen bleiben erhalten.

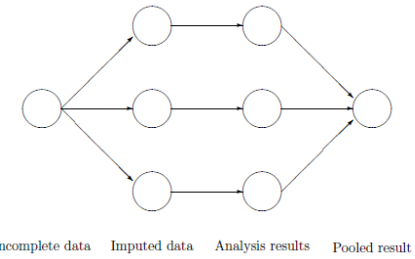


Abb. 1 Vorgehensweise bei multipler Imputation nach Buuren

2.3.6 Multiple Imputation

Nach (Buuren, 2018) werden bei multipler Imputation m > 1 Datensätze aus dem Ursprungsdatensatz erzeugt. Jeder dieser Datensätze wird bearbeitet. Die m Ergebnisse werden in einer letzten Schätzung inklusive des Standardfehlers zusammengefasst (pooling) unter Verwendung von „Rubin´s rules“.

Im ersten Schritt werden die fehlenden Daten in den verschieden Datensätzen durch plausible Werte ersetzt, die aus einer Verteilung gezogen wurden, die für jeden fehlenden Wert modelliert wird. Daraus entstehen m imputierte Datensätze, in denen die beobachteten Werte gleich sind, die sich aber bei den imputierten Werten unterscheiden. Die Unterschiede bei den imputierten Werten gibt Aufschluss über die Unsicherheit über den imputierten Wert.

Nun muss man noch im zweiten Schritt den zu schätzenden Parameter für jeden imputierten Datensatz mit Hilfe einer Analysemethode bestimmen. Die geschätzten Ergebnisse unterscheiden sich.

Die geschätzten Parameter der verschiedenen Schätzungen werden im dritten Schritt, dem Pooling, in einem Ergebnis zusammengefasst. Die Gesamtvarianz enthält die Stichprobenvarianz (within- imputation variance) und die zusätzliche Varianz, die durch die fehlenden Werte erzeugt wurde (between-imputation variance). Die zusammengefassten Schätzungen sollten unverzerrt sein und die korrekten statistischen Eigenschaften haben. (Buuren, 2018)

Bei singulärer Imputation ist zwar der MSE gut, aber andere statistische Eigenschaften des bearbeiteten Datensatzes ändern sich. Bei einer guten multiplen Imputation bleiben die Varianz, die Häufigkeitsverteilung und die Korrelationsmatrix gleich.

2.4 Ausgewählte Verfahren für Imputation von fehlenden Werten in Zeitreihen

Da Zeitreihendaten Querschnittsdaten ähneln, aber auch einige Besonderheiten aufweisen, auf die in 2.1.2 hingewiesen wurde, kann man nun viele Imputationsverfahren übertragen, andere weisen aber deutlich schlechtere Ergebnisse auf.

Zudem muss noch die Lückengröße (Gapsize) als weiterer wichtiger Parameter berücksichtigt werden. In dem Analyseteil werden verschiedene Algorithmen anhand ihrer Performance auf verschiedenen Datasets mit unterschiedlichen Lückengrößen und unterschiedlichen Anteilen von fehlenden Werten miteinander verglichen.

2.4.1 Ausschluss

In einer Zeitreihe führt ein Ausschluss zu einer Unterbrechung dieser. Da viele Zeitreihen durch Autokorrelationen und Saisonalität geprägt sind, kann ein Ausschluss eines Datensatzes zu signifikanten Verzerrungen führen. Bei multivariaten Zeitreihen verstärkt sich dieser Effekt noch. Daher ist ein Ausschluss bei Zeitreihen nicht zu empfehlen, nur in begründeten Ausnahmenfällen kann davon abgesehen werden.

2.4.2 Nachbarpunkt

Der fehlende Wert wird durch den zeitlich nächsten Wert ersetzt. Bei dem Last Observation Carried Forward Verfahren (LOCF) verwendet man den früheren Wert. Beim dem Next Observation Carried Backward (NOCB) wird der nachfolgende Wert benutzt.

2.4.3 Ersetzen durch Lageparameter

Analoge zu der Vorgehensweise bei der fehlenden Werten in Querschnittsdaten werden hier durch den bestimmte Lageparameter der Zeitreihe ersetzt. Diese können das Arithmetisches Mittel, der Median oder der Modus sein. Sollte eine Zeitreihe nicht zetriert sein, führt dieses Verfahren zu extremen Verzerrungen, wie im Analyseteil veranschaulicht wird.

2.4.4 Lineare Interpolation

Bei einer linearen Interpolation werden zwei Werte mit einer Strecke verbunden, bei der $y_1,\ y_2$ die Werte zu den Zeitpunkten $t_1,\ t_2$ darstellen. Für die lineare Interpolation eines Wertes zwischen den beiden Zeitpunkten gilt:

$$y_{int} = y_1 + (t_{int} - t_1)\frac{y_2 - y_1}{x_2 - x_1}$$

Daneben gibt es vielfältige Variationen wie die Spline-Interpolation.

2.4.5 Gleitender Mittelwert/ Moving Average

„Der gleitende Mittelwert ist eine Methode zur Glättung von Messdaten. Anders als der herkömmliche Mittelwert (arithmetisches Mittel), der über die Gesamtheit der vorliegenden Daten gebildet wird, wird der gleitende Mittelwert im Intervall über verschiedene, gleich große Untermengen eines Datensatzes gebildet.“ (Statista, 2022)

Der einfache gleitende Mittelwert (Simple Moving Average, SMA) beinhaltet die Werte $y_1,\ y_2,\ldots,\ y_n$ an den Zeitpunkten $t_1,\ t_2,\ldots,t_n$ mit n als die Fenstergröße.

$$y_{ma} = \frac{y_1 + y_2 + \cdots + y_n}{n}$$

Variationen sind z.B. der gewichtete gleitender Durchschnitt, der exponentiell geglätteter Durchschnitt und der exponentiell gewichtete geglättete Durchschnitt.

2.4.6 Structural Model mit Kalman- Filter

ImputeTS bietet auch einen Kalman Filter an, der zwei Inputs akzeptiert: ein Strukturmodel, das durch einen Maximum Likelihood Schätzer angepasst wird und ein Arima-Model basierend auf dem Paket „auto.arima“. In den Simulationen wird das Strukturmodel „StructTS“ verwendet. Weiter Informationen zu dem Strukturmodel und dem Kalman-Filter findet man bei (Durbin & Koopman, 2012) und bei (Grewal & Andrews, 2014).

2.4.7 Multiple Imputation mit Amelia II

Das Imputationsmodel von Amelia geht nach davon aus, dass die vollständigen Daten multivariat normalverteilt vorliegen. Die zugrundeliegenden Ausfallmechanismen müssen entweder MAR oder MCAR sein (Honacker, King, & Blackwell, Amelia II: A Program for Missing Data, 2011).

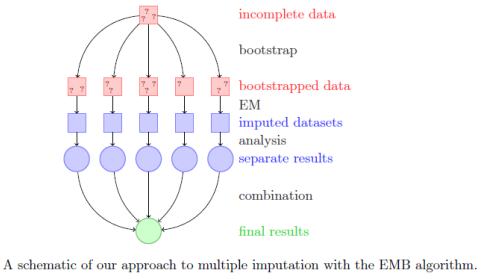


Abb. 2 Multiple Imputation mit Amelia II nach (Honacker, King, Amelia II)

Die Parameter des originalen Datasets sein $\theta = (\mu, \Sigma)$. Die beobachten Daten sein D^{obs} und *M* sei die „missingness matrix“. So ist der Likelihood unserer beobachteten Daten $p(D^{obs}, M|\theta)$.

Unter der MAR- Annahmen können wir dies so schreiben:

$$p(D^{obs}, M|\theta) = p(M|D^{obs})p(D^{obs}|\theta).$$

Da wir uns nur für die Inferenz der vollständigen Datenparameter interessieren, beschreiben wir die Likelihood als: $L(\theta\,|D^{obs}) \propto p(\theta\,|D^{obs})$.

Daraus folgt eine A-posteriori-Wahrscheinlichkeit:

$$p(\theta|D^{obs}) \propto p(D^{obs}|\theta) = \int p(D|\theta)\,dD^{mis}$$

Nach (Honacker, King, & Blackwell, Amelia II: A Program for Missing Data, 2011) S.3-4

Der EM- Algorithmus (Dempster, Laird, & Rubin, 1977) ist eine einfache Methode den Modus der A-posteriori-Wahrscheinlichkeit zu finden. Amelia II kombiniert den klassischen EM-Algorithmus mit einem Bootstrap- Verfahren um aus dieser A-posteriori-Wahrscheinlichkeit Rückschlüsse zu ziehen. Daraus kann man über die fehlenden Wert

Rückschlüsse treffen wie deren Punktschätzung oder auch die Varianz der Punktschätzung.

Die multiple Imputation wird in 2.3.6 genauer erklärt.

3 Pakete für Zeitreihen und Imputation in R

R Paket	Zweck	Monatliche Downloads (20.12.2023)
ImputeTestbench (3.0.3)	Vergleich von Imputationsmethoden	1.655
zoo (1.8-11)	Methoden für die Arbeit mit irregulären Zeitreihen	858.448
forecast (8.19)	Methoden für die Analyse von univariaten Zeitreihen	287.395
ImputeTS (3.3)	Methoden für die singuläre Imputation von univariaten Zeitreihen	35.193
Hmisc (4.7-2)	Sammlung von Methoden für Datenanalyse etc.	864.480
Mice (3.15.0)	Multiple Imputation mit FCS (multivariat, Querschnittsdaten)	91.720
Mi (1.1)	Imputation durch Approximative Bayessche Berechnung	29.203
Amelia II (1.8.1)	Multiple Imputation mit EM- Algorithmus und Bootstrap (multivariat, Zeitreihen)	11.778

R Bibliotheken für Zeitreihen/Imputation

4 Methodik

4.1 Fragestellung

Im praktischen Arbeiten mit Daten treten häufig fehlende Werte auf. Doch man kann oft schlecht beurteilen, welche Imputationsalgorithmen geeignet sind. Im Folgenden werden für die zwei Szenarien „fehlende Werte bei Aktienkursen“ und „fehlende Werte bei Renditezeitreihen“ auf einem historischen Datenmaterial einige der gängigsten Algorithmen getestet und deren Imputationsvorschläge miteinander verglichen. Besonderer Wert wird auf den Anteil der fehlenden Werte und auf die Größe der Lücken (Gapsize) gelegt.

4.2 Dataset

Das Dataset besteht aus 28 Aktienkursen von Unternehmen, die im Dow Jones Industrial Average zwischen dem 1.1.2018 und 31.12.2019 gelistet wurden. Damit der Datensatz vollständig ist, wurden alle Unternehmen entfernt, die in dem Zeitraum aus dem Dow Jones entfernt oder hinzugefügt wurden. Zudem wurde das Unternehmen Visa Inc entfernt, da aufgrund der hohen Korrelation mit Microsoft Corp der Algorithmus Amelia II nicht mehr funktionierte.

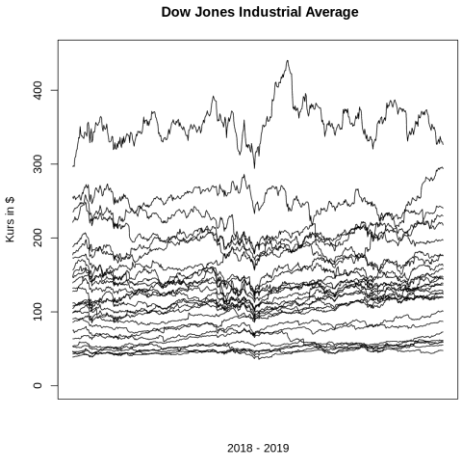


Abb. 3 Kursentwicklung der Dow Jones Aktien zwischen 2018 und 2019

Bei der Wahl des Zeitraums wurde darauf geachtet, dass es keine Wirtschaftskrisen oder größere Strukturbrüche gab. Die Jahre 2018 und 2019 waren in den Vereinigten Staat geprägt von wirtschaftlicher Stabilität unter der Führung des republikanischen Präsidenten Donald J. Trump, in denen der Dow Jones von 26149,39 am 01.01.2018 auf 28256,03 am 01.01.2020 leicht anstieg.

Von den jeweiligen Aktienkursen wird ein R- Dataframe erstellt, der nur die Schlusskurse enthält.

In der Graphik wird beispielhaft der ACF und der PACF der Apple- Aktie dargestellt. Der ACF ist abklingend, der PACF hat nur einen signifikanten Ausschlag bei 1. Dies deutet auf einen AR (1) – Prozess hin. Wahrscheinlich liegt aber keine Stationarität vor.

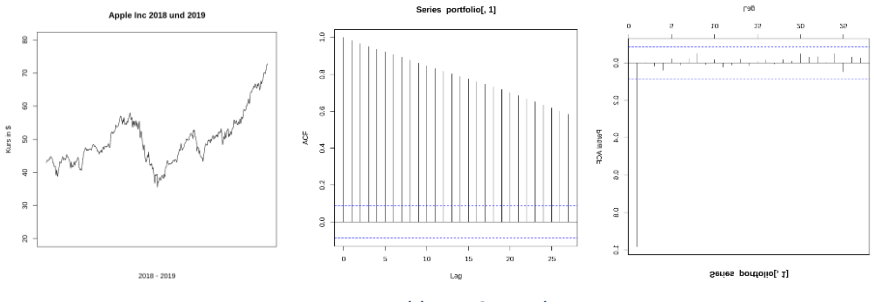


Abb. 5 Apple: Kurs, ACF, PACF

In der Korrelationsmatrix für die Kurse kann man erkennen, dass die Kursentwicklung einiger Aktien mit anderen Aktien stark positiv und negativ korreliert.

Für die Darstellung der Rendite wird die Stetige Rendite ausgewählt, mit dem Kurs $S_0\ zum\ Zeitpunkt\ t_0$, dem Kurs $S_1\ zum\ Zeitpunkt\ t_1$ und der Stetigen Rendite $r_1^s\ zum\ zum\ Zeitpunkt\ t_1$:

$$r_1^s = \ln(S_1) - \ln\,(S_0)$$

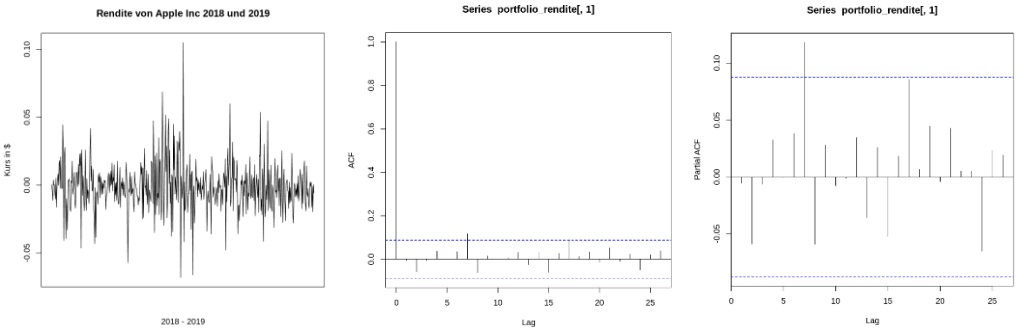


Abb. 6 Apple: stetige Rendite, ACF, PACF

Die Rendite weist deutlich andere statistische Eigenschaften wie der zugehörige Aktienkurs auf. Der ACF zeigt nur einen Ausschlag bei Lag_0, der PACF hat nur einen signifikanten Ausschlag bei Lag_7. Stationarität ist wahrscheinlich nicht gegeben, da im Renditechart Volatilitätscluster zu erkennen sind.

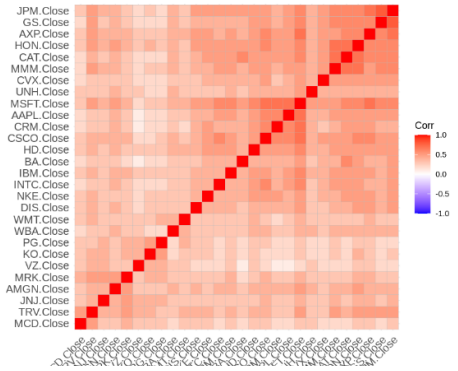


Abb. 7 Korrelationsmatrix Renditen

Die Korrelationsmatrix der Renditen weist eine interessante Besonderheit auf. Keine der Renditen der Aktienkurse korrelieren stark negativ miteinander.

4.3 Erzeugung von amputierten Datasets

Aus dem vollständigen Dataset wird nun ein fehlerhaftes Dataset erstellt. Es werden zufällig Werte gelöscht, um eine MCAR- Fehlerstruktur zu erzeugen. Als Eingabeparameter gibt es die Häufigkeit von Lücken (Gaps) und die Größe der Lücken (Gapsize). Die Lücken können sich auch überlagern. Durch die Überlagerung muss der prozentuale Anteil der NA- Wert nicht dem Produkt von Lückengröße und Lückenhäufigkeit entsprechen.

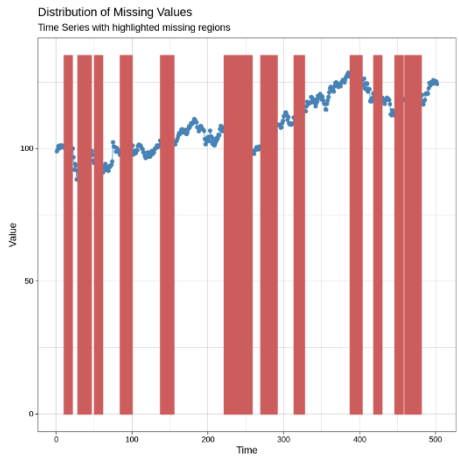


Abb. 8 Ein fehlerhaftes Dataset mit Lücken

Der ursprüngliche Wert wird mit einem „NA“- Eintrag ersetzt. Im Diagramm kann man beispielhaft das Ergebnis einer Amputation sehen. Der Ursprungsgraph ist in blau dargestellt. Die nicht bekannten Werte als Lücken in rot.

4.4 Testen der verschieden Imputationsmethoden

Aus dem amputierten Dataset werden die Werte nach ausgewählten Imputationsmethoden ersetzt, die dem R Paket Amelia und dem R Paket imputeTS entnommen

wurden. ImputeTS steuert folgende Algorithmen bei: Locf, Mean, Interpolation, Moving Average, Kalman Filter.

Locf ist die erste verwendete Imputationsmethode (Last Observation Carried Forward). Hier wird der NA Wert durch seinen zeitlichen Vorgänger imputiert. Bei größeren Lücken wird im zweiten Schritt der letzte NA- Wert in der Lücke durch seinen Nachfolger imputiert, bis die Lücke verschwindet. Diese Paramatereinstellung „na_remaining“ = „rev“ der Funktion Locf ist default.

Die zweite verwendete ist die Imputation durch den arithmetischen Mittelwert (mean) der Zeitreihe.

Die dritte Methode interpoliert über die fehlenden Werte (na_interpolate). Zur Auswahl stehen lineare Interpolation, Spline Interpolation und Stineman Interpolation. Es wird hier die Default- Einstellung lineare Interpolation verwendet.

Der gleitende Durchschnitt (moving average, na_ma) ist die vierte Methode. Hier stehen die Parameter „k“ und „weighting“ zur Verfügung. k definiert die Größe des Betrachtungsfensters. Mit „weighting“ stehen verschiedene Gewichtungsfunktionen zur Verfügung: „simple“ – Simple Moving Average, „linear“ – Linear Weighted Moving Average und „exponentional“ – Exponentional Weighted Moving Average (EWMA) als Default.

$$EWMA_t = \alpha * r_t + (1 - \alpha) * EWMA_{t-1}$$

Als fünfte Methode wird der Kalman Filter mit dem Strukturmodel „StructTS“ verwendet (siehe 2.4.6).

Aus dem Amelia II Paket stammt die sechste Methode Amelia (siehe 2.3.7). Hier werden mehrere in unserem Fall fünf imputierte Datasets erzeugt, die dann mit Hilfe der Rubins Rule zu einem Dataset vereint werden (siehe 2.4.7).

„First estimate some Quantity of interest, Q, such as a univariate mean, regression coefficient, predicted probability, or first difference in each data set j (j= 1,…,m). The overall point estimate $\bar{q}$ of Q is the average of the m separate estimates, q_i:”

Rubin’s rule: $\bar{q} = \frac{1}{m}\sum_{j=1}^{m} q_i$ (Honacker, King, Joseph, & Scheve, 2001)

Da die echten Werte bekannt sind, wird aus dem Durchschnitt der quadrierten Differenz zwischen den geschätzten und dem ursprünglichen Dataset die mittlere quadratische Abweichung (Mean Squared Error/ MSE) berechnet, um die Qualität der verschiedenen Schätzverfahren miteinander vergleichen zu können. Diese Fehlermetrik gibt an, wie stark die Schätzung von der Realität abweicht. n sind die Anzahl der Beobachtungen $\hat{Y}_i$ ist der geschätzte Wert und Y_i der tatsächliche.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Es werden jeweils fünf Simulationsdurchgänge durchgeführt, in denen verschiedene amputierte Datasets erzeugt werden, indem man den Seed- Wert in der Zufallszahlfunktion ändert. Aus den Ergebnissen der Simulationsdurchgänge werden nun die Mittelwerte gebildet.

Es wurde auf die Berechnung anderer Maßzahlen verzichtet, da der Umfang der Simulationen schon ausreichend ist. Man könnte aber argumentieren, dass sich die Aktienkurse um den Faktor 8 unterscheiden können, und hier Fehlschätzungen auf höheren Aktienkursen ein stärkeres Gewicht haben. Abhilfe könnte der mean absolute percentage error (MAPE) bringen.

Zudem wird bei manchen Testläufen auch noch die Ausführungszeit der Algorithmen gemessen, um auch diese Werte miteinander zu vergleichen. Diese Testergebnisse hängen auch stark von dem jeweiligen Testrechner ab und eignen sich nur, um die Geschwindigkeit der verschiedenen Algorithmen miteinander zu vergleichen. Es ist außerdem in der Programmierung noch viel Optimierungspotenzial, daher wird es durchaus möglich sein, die Ausführungsgeschwindigkeit der einzelnen Algorithmen deutlich zu beschleunigen und deren Ergebnisqualität zu verbessern.

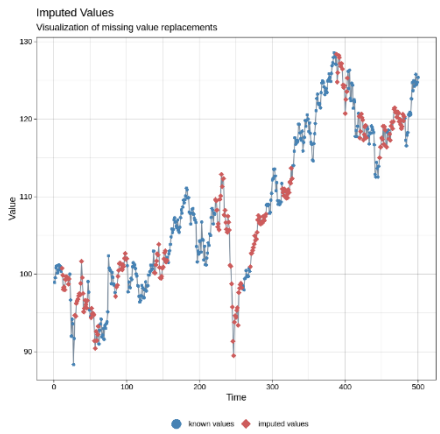


Abb. 9 Imputiertes Dataset

Die jeweiligen Methoden verwenden entweder das imputeTS Paket oder das Amelia- Paket. Sie wurden im automatischen Modus ausgeführt und könnten manuell wahrscheinlich noch stärker optimiert werden.

Im Diagramm sieht man ein Beispiel für ein imputiertes Dataset. Die imputierten Werte sind in rot eingetragen, die nicht amputierten Werte in blau.

5 Ergebnisse

5.1 Analyse der Aktienkurszeitreihen

5.1.1 Auswirkung eines Anstieges von fehlenden Werten

In Abb. 10 kann man den MSE der verschiedenen Imputationsalgorithmen ablesen, nachdem sie aus dem amputierten Dataset die fehlenden Werte ersetzt haben. Bei diesen Aktienkurszeitreihen schneidet der Mean (arithmetischer Mittelwert) erwartbar ziemlich schlecht ab, da diese Zeitreihen meist nicht zentriert sind. In den folgenden

Graphiken für Aktienkurse wird der Mean für eine bessere visuelle Darstellungsqualität herausgenommen.

NA-Anteil in %	Locf	Mean	Interpolation	Moving Average	Kalman	Amelia
1.982072	8.109201	200.0901	3.478061	4.038686	3.480642	16.76638
3.941377	7.026949	194.4962	3.553161	4.334039	3.556783	18.58928
5.886454	8.289582	183.1662	3.325109	3.996924	3.330624	19.47642
7.800228	8.664235	197.0254	4.125283	4.881574	4.123655	19.46788
9.709732	8.982262	193.4574	4.036384	4.752387	4.038415	18.32950
11.617814	8.073234	193.8078	3.577787	4.355358	3.577803	20.19788
13.453330	9.052858	186.0354	3.912874	4.650827	3.908662	20.64345
15.331531	9.061282	190.9525	3.847323	4.568845	3.848688	20.27498
17.078828	8.990552	186.1389	3.842932	4.766195	3.842343	20.50583
18.912920	9.497148	186.2436	4.039089	4.913635	4.051741	21.79967

Abb. 10 Tabelle mit MSE von verschiedenen Imputationsalgorithmen für Aktienkurse bei einer Gapsize = 1

Die Interpolation und der Kalman- Filter weisen den niedrigsten MSE und damit die beste Performance auf, beide Werte liegen so eng beisammen, dass sie sich beide Graphen im Abb. 11 überlagern. Das Moving Average (Gleitender Durchschnitt) ist etwas abgeschlagen auf den dritten Platz. Locf ist deutlich schlechter und wird nur noch von Amelia übertroffen. Der steigende Anteil der fehlenden Werte verschlechtert die Performanz der Algorithmen nur leicht, die Rangfolge der Algorithmen bleibt aber erhalten.

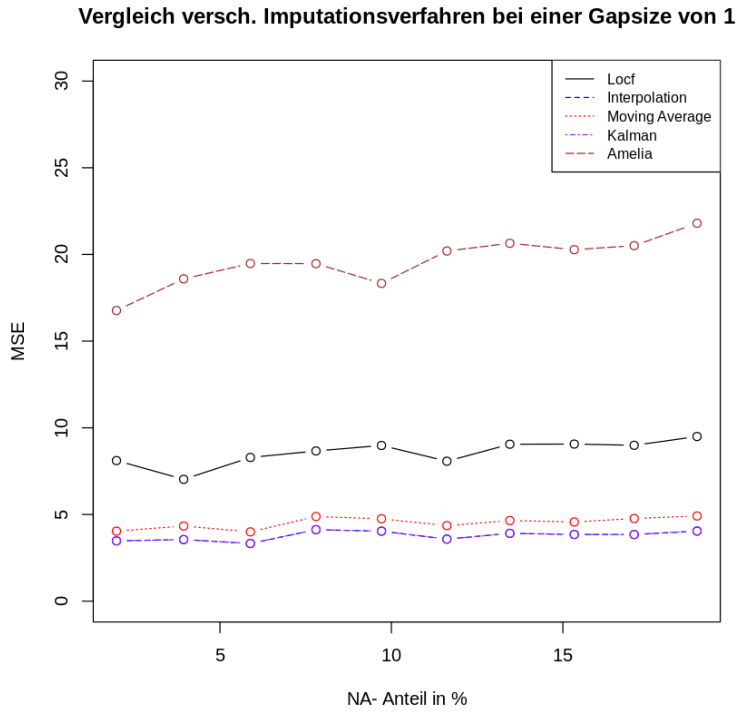


Abb. 11 Vergleich verschiedener Imputationsverfahren für Aktien bei einer Gapsize = 1

In Abb. 12 wird die Ausführungszeit der verschiedenen Algorithmen miteinander verglichen. Die einfachen Algorithmen Locf, Mean, Interpolation und Moving Average weisen eine schnelle Ausführungszeit zwischen 2 ms und 12 ms bei niedrigen und bei hohen Anteilen von fehlenden Werten auf. Der Kalman Filter ist deutlich langsamer mit durchschnittlich 268 ms. Amelia ist nochmals deutlich schlechter mit dem besten Wert von 1591 ms und verlangsamt sich mit steigender Fehlerzahl weiter.

NA-Anteil in %	Locf	Mean	Interpolation	Moving Average	Kalman	Amelia
1.984917	0.002024889	0.002027512	0.006212473	0.003098488	0.2539659	1.590930
3.948492	0.002010107	0.002051830	0.005839348	0.004146576	0.2759976	1.655831
5.890723	0.002091408	0.002057314	0.005897284	0.005141735	0.2725024	1.790688
7.797382	0.002422571	0.002104521	0.006305218	0.006088734	0.2806880	1.971865
9.704041	0.002663612	0.002124786	0.006010532	0.007461548	0.2639463	2.194636
11.646272	0.002183914	0.002087355	0.006077766	0.008800745	0.2675097	2.128295
13.474673	0.002094984	0.002060890	0.005776405	0.009612560	0.2685571	2.203307
15.374217	0.002339125	0.002114296	0.006286144	0.011128664	0.2640531	2.373112
17.124360	0.002183914	0.002417564	0.006146193	0.011269331	0.2797031	2.410002
18.888731	0.002473593	0.002113581	0.005863667	0.012212038	0.2628539	2.509513

Abb. 12 Tabelle mit Ausführungszeit in Sekunden von verschiedenen Imputationsalgorithmen für Aktienkurse bei einer Gapsize = 1

Im Anhang 1 und Anhang 2 kann man den gleichen Untersuchungsgegenstand für eine Gapsize von fünf erkennen. Die Rangfolge der Algorithmen bleibt erhalten, nur der Wert für den MSE verschlechtern sich.

5.1.2 Auswirkung eines Anstieges der Lückengröße

Wie schon im vorherigen Abschnitt angedeutet, führt eine Vergrößerung der Länge der Lücken (Gapsize) zu einer Verschlechterung der Performance der Algorithmen.

NA-Anteil in %	Gapsize	Locf	Mean	Interpolation	Moving Average	Kalman	Amelia
2.768924	1	7.196481	198.3834	3.687023	4.594396	3.688029	18.58746
2.349175	2	11.995250	214.0680	4.998570	5.925325	4.995197	18.47310
2.380478	3	13.026453	171.5468	5.975421	7.021266	5.990758	18.86823
1.980649	4	18.564807	198.8776	6.531539	8.696812	6.549803	15.55574
2.370518	5	20.690210	187.0505	7.904584	10.933521	7.933297	18.54258
2.747581	6	23.883263	219.2992	8.644572	11.927618	8.654932	21.24388
1.576551	7	26.004247	159.8769	9.830204	13.258795	9.847369	25.05423
1.780023	8	27.477848	205.7449	11.214519	13.162087	11.200688	23.57236
1.977803	9	31.182314	170.8174	11.346438	17.457610	11.350560	32.14982
2.178429	10	33.609273	206.8417	12.900465	18.098689	12.898950	33.47083

Abb. 13 Tabelle mit MSE von verschiedenen Imputationsalgorithmen für Aktienkurse bei steigender Länge der Lücken

Es ist in Abb. 14 und Abb.15 zu erkennen, dass alle Algorithmen eine deutlich schlechtere Performance mit einer größeren Gapsize aufweisen. Nur der arithmetische Mittelwert bleibt gleichbleibend schlecht. Diese Entwicklung lässt sich bei einem niedrigen Anteil fehlender Werte (ungefähr 2%) und bei einem hohen Anteil fehlender Werte (ungefähr 18%) feststellen.

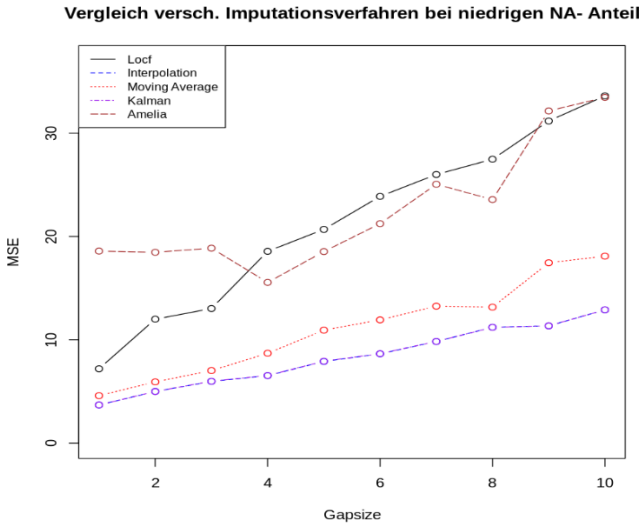


Abb. 14 Vergleich verschiedener Imputationsverfahren für Aktien

bei einem niedrigen Anteil von fehlenden Werten

5.2 Analyse der stetigen Renditezeitreihen

Bei den analysierten Renditezeitreihen präsentiert sich nun ein anderes Bild. Bei der Simulation bei einer Gapsize von 1 weist der Amelia- Algorithmus die mit Abstand beste Performance auf (siehe Abb. 15 und Abb. 16). Gefolgt wird er von Mean und Kalman. Das Moving Average und das Interpolationsverfahren ist etwas schlechter und wird nur noch von Locf negativ übertroffen.

NA-Anteil in %	Locf	Mean	Interpolation	Moving Average	Kalman	Amelia
1.988879	0.0004122474	0.0002050169	0.0003488832	0.0002653279	0.0002087822	0.0001416118
3.949244	0.0004369966	0.0002142593	0.0003572202	0.0002710016	0.0002188883	0.0001473725
5.906758	0.0004371881	0.0002185361	0.0003665650	0.0002782762	0.0002214425	0.0001387276
7.808668	0.0004296269	0.0002166976	0.0003415458	0.0002673703	0.0002199029	0.0001465199
9.710579	0.0004575810	0.0002237089	0.0003699896	0.0002850542	0.0002278417	0.0001531500
11.595381	0.0004845773	0.0002283243	0.0003705190	0.0002924498	0.0002323885	0.0001505811
13.498717	0.0004485720	0.0002150222	0.0003533722	0.0002795179	0.0002188444	0.0001444447
15.305104	0.0004420210	0.0002217923	0.0003525379	0.0002801567	0.0002261276	0.0001479997
17.211292	0.0004526626	0.0002162132	0.0003560512	0.0002805922	0.0002201840	0.0001511867
18.940690	0.0004373017	0.0002128801	0.0003407660	0.0002730416	0.0002167973	0.0001467515

Abb. 15 Tabelle mit MSE von verschiedenen Imputationsalgorithmen für Log-Renditen bei einer Gapsize = 1

Der Anstieg des NA- Anteils hat keine große Auswirkung auf die Performance der verschiedenen Algorithmen.

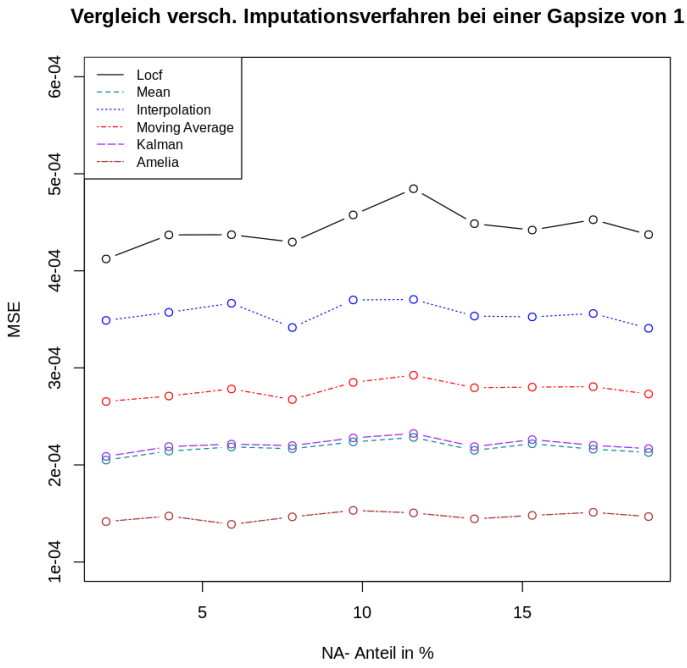


Abb. 16 Vergleich verschiedener Imputationsverfahren für Log- Renditen bei einer Gapsize von 1

Im Anhang 5 und Anhang 6 befinden sich die Untersuchungen für eine größere Lückengröße. Auch hier ändert sich die Reihenfolge der Algorithmen nicht.

6 Diskussion

Die Aktienkurszeitreihen und die Renditezeitreihen unterscheiden sich grundlegen ihrer statistischen Eigenschaften (Stationarität, Varianz, ACF, PACF, usw.). Dies hat auch eine starke Auswirkung auf die gewählten Imputationsalgorithmen. Während bei den analysierten Aktienkursen die Algorithmen Kalman/ Interpolation, Moving Average am besten waren, performten in den Renditezeitreihen Amelia, Kalman/ Mean am besten.

Da es noch viele verschiedene andere Arten von Zeitreihen gibt, kann man nicht einen Algorithmus für alle Fälle empfehlen, sondern die Algorithmen müssen angepasst an die Anforderungsdefinition gezielt selektiert werden.

Es gibt einige Kritikpunkte gegenüber der Methodik, die nicht verschwiegen werden sollten:

Im Anwendungsfall steht nur ein unvollständiges Dataset zur Verfügung. Daher kann man nicht wie mit dem hier genutzten Verfahren die jeweiligen Algorithmen evaluieren. Wenn man trotzdem dieses Evaluierungsverfahren nutzen möchte, könnte man ähnliche, aber vollständige Datensätze verwenden oder man nimmt die restlichen korrekten Daten, um die Algorithmen daran zu testen.

In dieser Analyse wurden auch nur MCAR- Fehler untersucht, die auf eine bestimmte Weise implementiert wurden. Bei anderen Fehlerklassen und anderen Implementierungen werden die Simulationen vielleicht andere Ergebnisse erzeugen.

Auch die Daten wurden speziell ausgesucht. Bei anderen Datengrundlagen, andere Beobachtungszeiträume und auch größeren Datenmengen könnte es zu einer Abweichung der Ergebnisse kommen.

Die Algorithmen wurden nur mittels MSE evaluiert. Auch andere Streuungsmaße und andere Evaluationskriterien wären möglich. Man könnte auch messen, inwieweit die imputierten Werten ähnliche statistische Eigenschaften wie die nicht imputierten Werte haben (z.B. hinsichtlich der Verteilungsfunktion). Man könnte auch ein Data Mining Verfahren auswählen und untersuchen, ob sich die Prognosequalität durch die Imputationen der Imputationsalgorithmen verschlechtert oder verbessert.

Die Imputationsalgorithmen wurden selektiert und stellen nur eine kleine Auswahl der möglichen dar. Es gibt wahrscheinlich andere Algorithmen, die noch deutlich bessere Ergebnisse liefern. Auch die verwendeten Algorithmen können womöglich noch durch Optimierung der Parameter und des Algorithmus noch weiter verbessert werden.

Zudem muss man die Imputationsverfahren im Gesamtkontext sehen. Wie in Kapitel 1 erläutert wurde, ist die Datenbereinigung nur ein Schritt im KDD- Prozess. Das Ziel muss sein, die bestmöglichen Ergebnisse mit einem möglichst geringen Aufwand zu erzielen unter den gegebenen Anforderungen. Daher muss das Imputationsverfahren mit dem Data Mining Verfahren und den Prozessanforderungen abgestimmt werden.

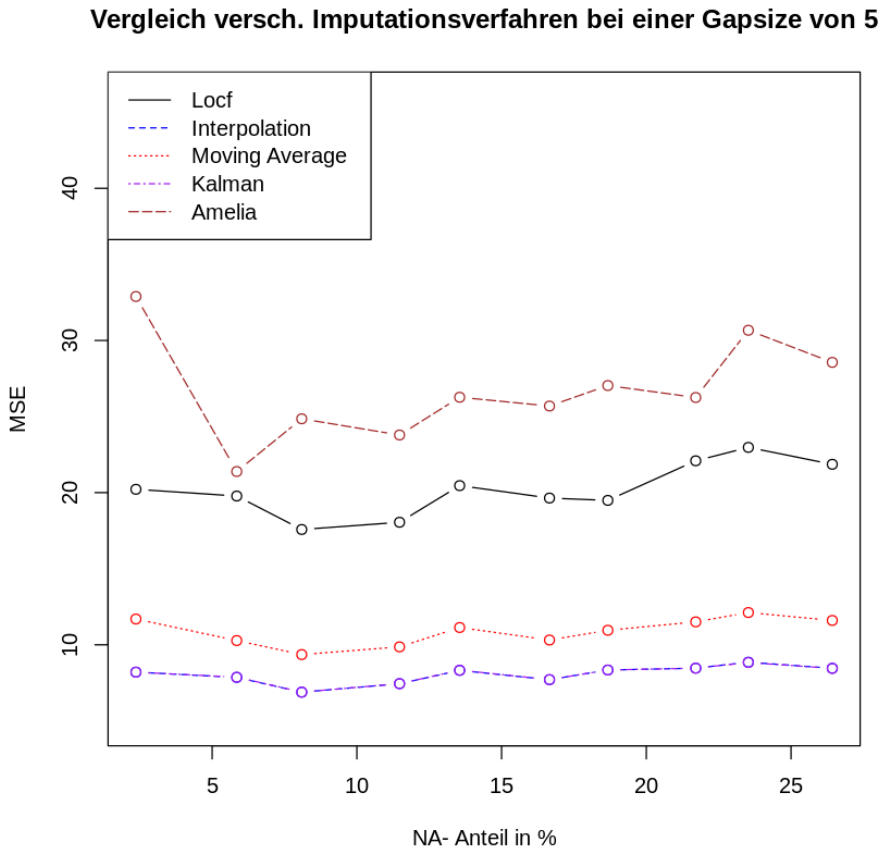
7 Fazit

Abschließend muss ich eingestehen, dass diese Seminararbeit bestenfalls als eine Einführung in die umfangreiche Welt der Behandlung fehlender Daten zu verstehen ist. Hier wurden nur Algorithmen verglichen, die Einbettung in ein umfassenderes Management wurde ignoriert. Da es anscheinend keine universal gut performenden Algorithmen gibt, bedarf es einer maßgeschneiderten Lösung.

8 Anhang

NA-Anteil in %	Locf	Mean	Interpolation	Moving Average	Kalman	Amelia
2.366249	20.21758	220.3633	8.202883	11.693140	8.201926	32.88968
5.849459	19.77670	199.3344	7.857674	10.277959	7.871534	21.38761
8.094764	17.57784	182.0343	6.883757	9.357372	6.891372	24.85234
11.474104	18.05226	198.4652	7.442927	9.862208	7.444673	23.78673
13.551508	20.45991	180.5356	8.315961	11.135493	8.332427	26.26839
16.654809	19.64246	197.5442	7.708683	10.313069	7.711383	25.69145
18.671030	19.49067	195.3000	8.341277	10.954133	8.343387	27.03886
21.707456	22.09120	186.9631	8.462864	11.504209	8.465138	26.25243
23.525896	22.97361	194.7267	8.856954	12.119454	8.833306	30.67093
26.425726	21.86057	187.4357	8.455240	11.600308	8.446192	28.56077

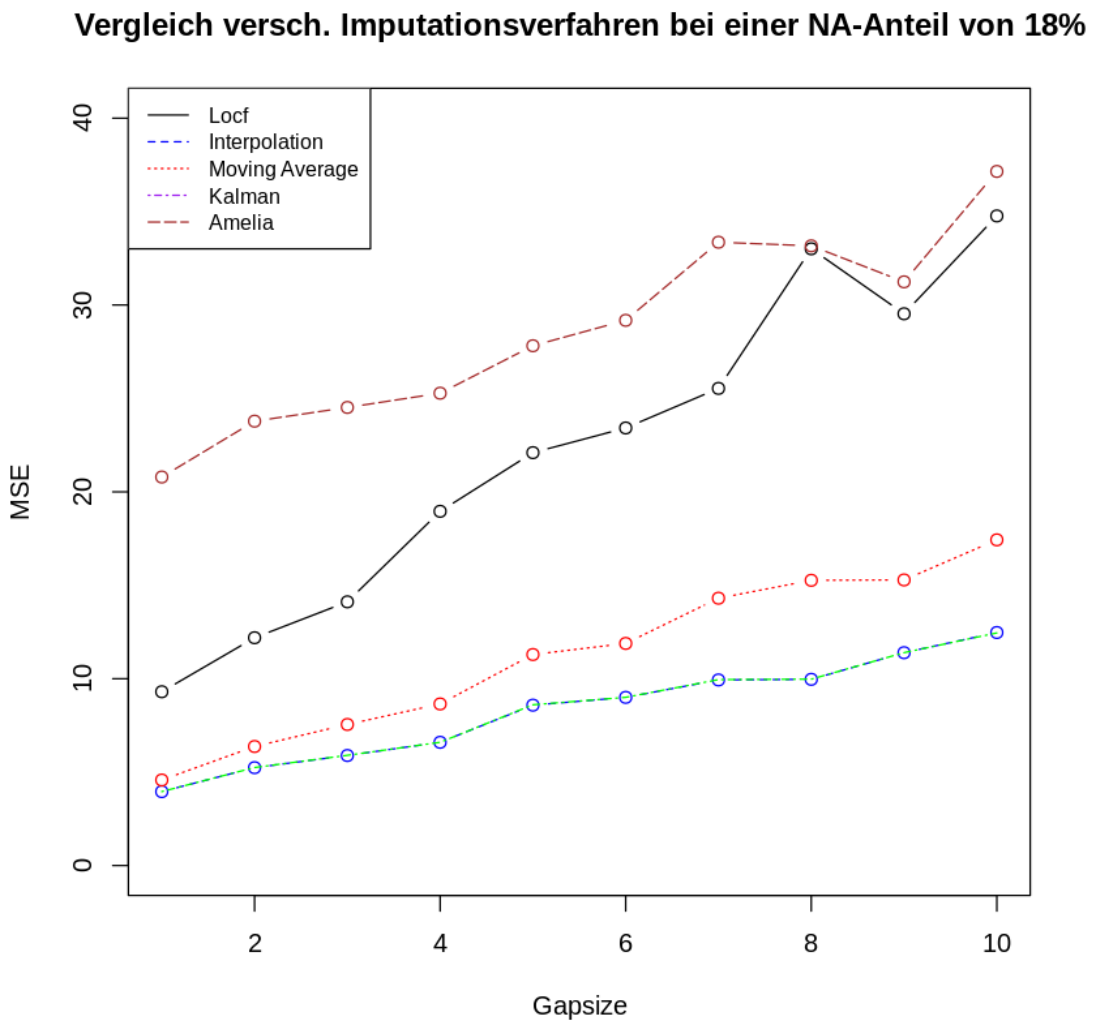
Anhang 1: Tabelle mit MSE von verschiedenen Imputationsalgorithmen für Aktienkurse bei einer Gapsize = 5



Anhang 2: Vergleich verschiedener Imputationsverfahren für Aktien bei einer Gapsize von 5

NA-Anteil in %	Gapsize	Locf	Mean	Interpolation	Moving Average	Kalman	Amelia
18.93426	1	9.302194	190.8430	3.957039	4.577509	3.957405	20.79392
18.86739	2	12.192470	194.8580	5.238627	6.371259	5.246033	23.78285
18.56716	3	14.114034	192.9440	5.889396	7.550010	5.899561	24.51810
18.50740	4	18.958219	193.6824	6.595932	8.648681	6.596606	25.27963
18.55435	5	22.098076	201.9211	8.583706	11.294130	8.604220	27.81902
18.62550	6	23.419460	193.5303	9.000756	11.887530	9.002815	29.18026
18.80763	7	25.536592	210.5549	9.933750	14.308071	9.950350	33.36040
18.19721	8	33.008329	188.0537	9.966121	15.265227	9.974587	33.16705
18.24132	9	29.530736	187.7810	11.391951	15.285534	11.393694	31.23990
18.04639	10	34.767739	201.1043	12.471978	17.431332	12.449017	37.14893

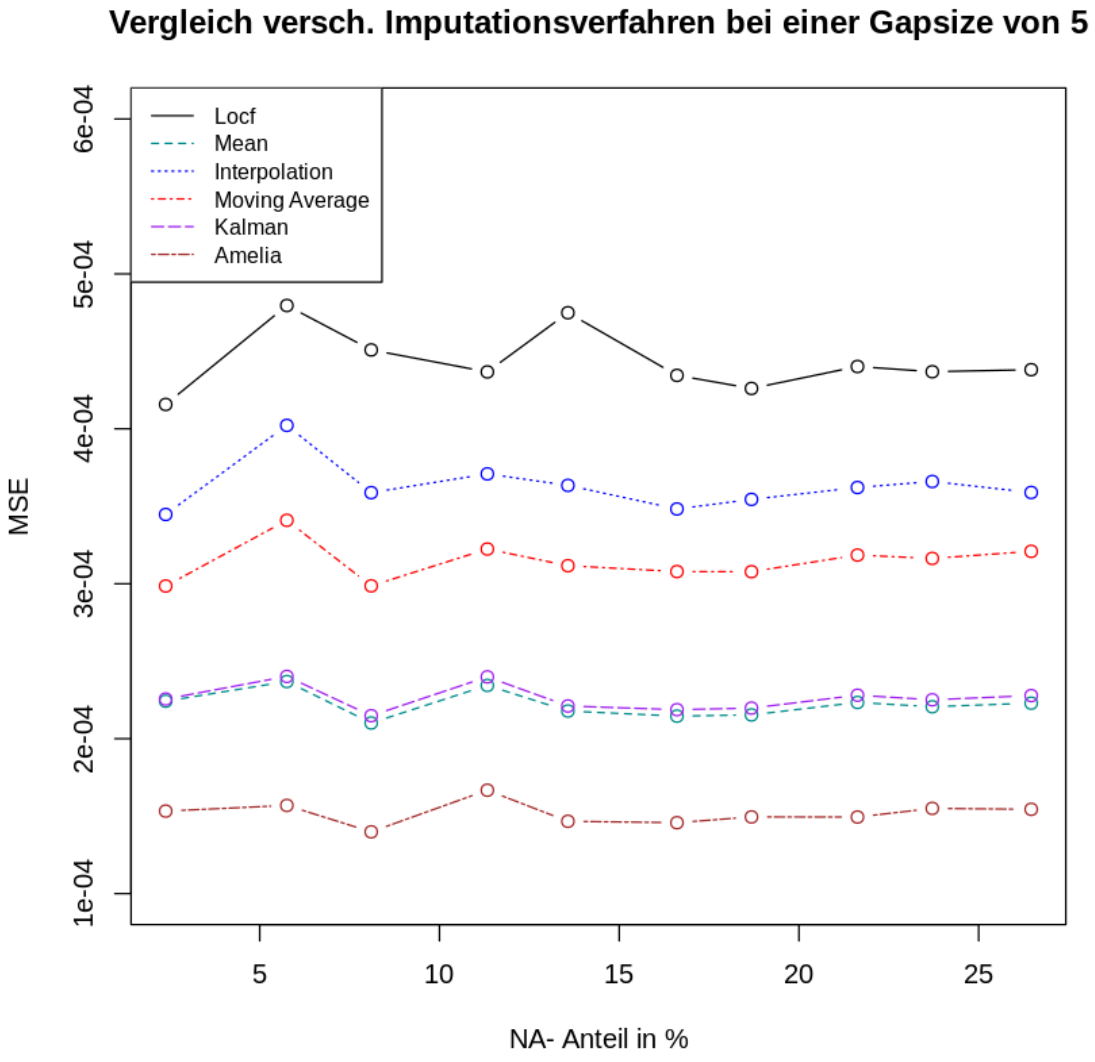
Anhang 3: Tabelle mit MSE von verschiedenen Imputationsalgorithmen für Aktienkurse bei hohem Anteil an fehlenden Werten



Anhang 4: Vergleich verschiedener Imputationsverfahren für Aktien bei hohem Anteil an fehlenden Werten

NA-Anteil in %	Locf	Mean	Interpolation	Moving Average	Kalman	Amelia
2.383804	0.0004157246	0.0002242325	0.0003447297	0.0002985344	0.0002256042	0.0001532698
5.754206	0.0004795865	0.0002368915	0.0004022278	0.0003409524	0.0002401874	0.0001569290
8.100941	0.0004509129	0.0002101416	0.0003588070	0.0002986402	0.0002148253	0.0001398839
11.331622	0.0004367290	0.0002343892	0.0003709060	0.0003224031	0.0002398923	0.0001667528
13.572854	0.0004749166	0.0002178520	0.0003634834	0.0003116432	0.0002210956	0.0001466889
16.606786	0.0004344867	0.0002145732	0.0003482616	0.0003078635	0.0002187331	0.0001457888
18.679783	0.0004260449	0.0002153620	0.0003544503	0.0003077794	0.0002197328	0.0001495013
21.628172	0.0004402144	0.0002234019	0.0003621196	0.0003185101	0.0002280647	0.0001494371
23.709723	0.0004368491	0.0002206202	0.0003660016	0.0003163363	0.0002251109	0.0001550124
26.461363	0.0004381586	0.0002229396	0.0003589089	0.0003209183	0.0002278280	0.0001544311

Anhang 5: Tabelle mit MSE von verschiedenen Imputationsalgorithmen für Log-Renditen bei einer Gapsize = 5



Anhang 6: Vergleich verschiedener Imputationsverfahren für Renditen bei einer Gapsize von 5

9 Literaturverzeichnis

Backhaus, K., Erichson, B., Gensler, S., Weiber, R., & Weiber, T. (2018). *Multivariate Analysemethoden, 16. Auflage.* Berlin Heidelberg: Springer.

Buuren, S. (2018). *Flexible Imputation of Missing Data 2nd Edition.* Boca Raton: CRC Press.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 1-38.

Durbin, J., & Koopman, S. (2012). *Time Series Analysis by State Space Methods (2nd edition).* Oxford: Oxford University Press.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine Volume 17 Number 3*, S. 37-54.

Frawley, W., Piatetsky-Shapiro, G., & Matheus, C. (1992). Knowledge Discovery in Databases: An Overview. *AI Magazine Volume 13 Number 3*, S. 57-70.

Grewal, M., & Andrews, A. (2014). *Kalman Filtering: Theory and Practice with MATLAB, 4th Edition.* Hoboken: Wiley-IEEE Press.

Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques (2. Aufl.).* San Francisco: Morgan Kaufmann.

He, Y., Zhang, G., & Hsu, C.-H. (2022). *Multiple Imputation of Missing Data in Practice.* Boca Raton: CRC Press.

Honacker, J., & King, G. (April 2010). What to Do about Missing Values in Time-Series Cross-Section Data. *American Journal of Political Science, Vol. 54, No. 2*, S. 561-581.

Honacker, J., King, G., & Blackwell, M. (December 2011). Amelia II: A Program for Missing Data. *Journal of Statistical Software, Volume 45, Issue 7*.

Honacker, J., King, G., Joseph, & Scheve. (March 2001). Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. *American Political Science Review, Vol. 95, No. 1*, S. 49-69.

Horton, N., & Kleinman, K. (February 2007). Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models. *The American Statistician,Vol. 61, No. 1 79*, S. 79-90.

Little, R., & Rubin, D. (2020). *Statistical Analysis with Missing Data 3rd Edition.* New York: JohnWiley & Sons, Inc.

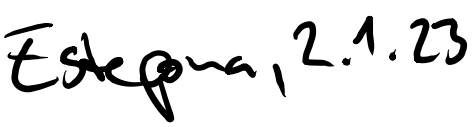
Müller, & Freytag, J.-C. (2003). *Problems, methods, and challenges in comprehensive data cleansing (Arbeitsbericht Nr. HUB-IB-164).* Berlin: Humboldt Universität.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons, Inc.

Tausendpfund, M. (2020). *Fortgeschrittene Analyseverfahren in den Sozialwissenschaften.* Wiesbaden: Springer VS.

10 Eidesstattliche Erklärung

Ich erkläre, dass ich die Seminararbeit selbstständig und ohne unzulässige Inanspruchnahme Dritter verfasst habe. Ich habe dabei nur die angegebenen Quellen und Hilfsmittel verwendet und die aus diesen wörtlich, inhaltlich oder sinngemäß entnommenen Stellen als solche den wissenschaftlichen Anforderungen entsprechend kenntlich gemacht. Die Versicherung selbstständiger Arbeit gilt auch für Zeichnungen, Skizzen oder graphische Darstellungen. Die Arbeit wurde bisher in gleicher oder ähnlicher Form weder derselben noch einer anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht. Mit der Abgabe der elektronischen Fassung der endgültigen Version der Arbeit nehme ich zur Kenntnis, dass diese mit Hilfe eines Plagiatserkennungsdienstes auf enthaltene Plagiate überprüft und ausschließlich für Prüfungszwecke gespeichert wird.



Ort, Datum

Unterschrift