Technische Universität Dortmund
Fakultät Statistik
Sommersemester 2024

Bachelorarbeit

Strukturbruchanalyse von Zusammenhangsmaßen von Stromdaten

Betreuerin:
Prof. Dr. Christine Müller

Verfasser:
Kevin Unrau

30.07.2024

Inhaltsverzeichnis

1 Einleitung

Der Klimawandel erfordert ein Umdenken in vielen Lebensbereichen. Dazu gehört insbesondere auch die Wärmeproduktion privater Haushalte in Deutschland, denn eine Studie des Bundesverbandes der Energie- und Wasserwirtschaft (BDEW) aus dem Jahr 2019 zeigt, dass etwa 74% mit den fossilen Energieträgern Erdgas und Öl heizen, wodurch Treibhausgasemissionen entstehen (So heizen die Deutschen, 2019). In diesem Zusammenhang hat der Bundestag 2023 im Rahmen des Gebäudeenergiegesetzes (GEG) beschlossen, dass ab 2024 „möglichst jede neu eingebaute Heizung zu 65 Prozent mit erneuerbaren Energien betrieben werden soll“ (Mit Wärmepumpen Tempo machen für die Klimawende, 2022). Vor allem Wärmepumpen, welche Umweltwärme, also zum Beispiel Energie aus der Luft oder Wasser, nutzen, sollen die klassischen Heizsysteme ablösen. Da diese jedoch strombetrieben sind, ist diese Umstellung mit einer Mehrbelastung des Stromnetzes verbunden. Ziel des Projekts ist es zu untersuchen wie der Stromverbrauch vieler Wärmepumpen an einem Ort zusammenhängt und inwiefern sich dieser Zusammenhang über das Jahr verändert. Insbesondere soll auch ermittelt werden, ob Strukturbrüche vorliegen. Die Grundlage des Projekts bilden Datensätze von 33 Einfamilienhäusern, die den Stromverbrauch der zugehörigen Wärmepumpen enthalten. Die Messungen stammen aus dem Jahr 2019 und wurden während des Forschungsprojekts Wind-Solar-Heat Pump District (WPuQ) in der Nähe von Hameln in Niedersachsen erhoben (Ohrdes et al., 2021). Diese Arbeit ist jedoch nicht Teil vom WPuQ-Projekt und wurde unabhängig davon erstellt.

Im nächsten Abschnitt wird zunächst die Ausgangssituation in Hameln dargelegt, indem das in den Häusern verbaute Heizsystem und der Datenerhebungsprozess kurz erläutert werden. Weiterhin wird der Datensatz beschrieben, hierbei werden die Datenaufbereitung und die für die Analyse genutzten Variablen vorgestellt. Auch die Zielsetzung des Projekts wird genau erklärt. Dabei wird unter anderem genannt mit welchen statistischen Methoden diese Ziele umgesetzt werden sollen. Die Definition dieser Methoden erfolgt in Abschnitt 3, bevor in Abschnitt 4 die Resultate der statistischen Auswertung präsentiert werden. Zum Abschluss werden die Ergebnisse zusammengefasst und kritisch hinterfragt. Außerdem werden weiterführende Forschungsfragen aufgezeigt.

2 Problemstellung

2.1 Ausgangssituation

Die betrachteten Häuser sind Teil einer aus 71 Einfamilienhäusern bestehenden Siedlung, die in Abbildung 1 dargestellt wird und sich am Ohrberg bei Hameln in Niedersachsen befindet (Ohrdes et al., 2021). Alle Häuser erfüllen den Niedrigenergiestandard und wurden Ende der 90er und Anfang der 2000er Jahre erbaut (Schlemminger et al., 2022). Die 33 betrachteten Häuser besitzen eine Wasser-Wasser-Wärmepumpe, welche „an ein kaltes Nahwärmnetz angeschlossen ist“, das die Pumpen ganzjährig mit 10 bis 12°C warmen Was-

ser versorgt (Ohrdes et al., 2021). Nebenbei besitzen die Wärmepumpen einen Heizstab, der als Backup fungiert und bei nicht ausreichender Wärmeproduktion unterstützt (Schlemminger et al., 2022). Die erzeugte Wärme wird über eine Fußbodenheizung im Haus verteilt. Weiterhin erfolgt die Warmwasseraufbereitung insbesondere im Sommer über eine Solaranlage, wobei die Wärmepumpe übernimmt, wenn nicht genug Solarenergie verfügbar ist (Schlemminger et al., 2022). Jedoch besitzt keiner der 33 betrachteten Haushalte eine Photovoltaikanlage.

Abbildung 1: Wind-Solar-Heat Pump District am Ohrberg in der Nähe von Hameln in Niedersachsen (ISFH EnEff:Stadt Verbundvorhaben: Wind-Solar Wärmepumpen-Quartier, o.D.).

Die Messung des Stromverbrauchs wurde im Zeitraum von Mitte 2018 bis Ende 2020 durchgeführt. Dafür wurden in den Häusern Messsysteme installiert, die in Intervallen von 10 Sekunden den momentanen Verbrauch der Wärmepumpe als auch den Verbrauch des Haushalts in Watt erhoben haben (Schlemminger et al., 2022). Diese Daten wurden lokal zwischengespeichert und alle 1-5 Minuten über die Internetverbindung des Hauses an einen zentralen Datenbankserver weitergeleitet. Beim Datenerhebungsprozess gab es teilweise technische Probleme, da die Messsysteme oder die Internetverbindung einzelner Häuser über einen längeren Zeitraum ausgefallen sind, sodass Daten der betroffenen Häuser für diese Zeiträume fehlen (Schlemminger et al., 2022). Da die erforderliche Heizleistung direkt vom Wetter abhängig ist, wurde zusätzlich im gleichen Zeitraum alle 5 Minuten die Temperatur für Hameln in Grad Celsius über eine HTTP REST API des Dienstes WetterOnline angefragt und ebenfalls auf dem zentralen Datenbankserver gespeichert (Schlemminger et al., 2022). Die Daten wurden zusätzlich basierend auf unterschiedlichen Zeitintervallen (1 Minute, 15 Minuten und 60 Minuten) aggregiert, indem das arithmetische Mittel des Stromverbrauchs gebildet wurde (Schlemminger et al., 2022).

Die für dieses Projekt vorliegenden Datensätze, welche im Unterabschnitt 2.3 vorgestellt werden, enthalten jeweils die aggregierten Daten eines Hauses in 15-Minuten-Intervallen für das Jahr 2019.

2.2 Ziele des Projekts

Zu Beginn wird repräsentativ für alle Haushalte eine deskriptive Analyse des ersten Haushalts durchgeführt, um einen Überblick über den Stromverbrauch einer Wärmepumpe im Verlaufe des Jahres zu geben. Für die Bestimmung des Zusammenhangs zwischen den Haushalten werden die täglichen paarweisen Kovarianzen zwischen dem Stromverbrauch der Wärmepumpen der einzelnen Haushalte bestimmt und in Kovarianzmatrizen eingetragen. Kovarianzen sind jedoch nicht vergleichbar, da sie stark von der Varianz der einzelnen Variablen abhängen. Zudem ist es nicht einfach möglich die täglichen Kovarianzmatrizen über die Zeit darzustellen. Aufgrund dessen wird für jeden Tag eine auf der Kovarianzmatrix basierende Hauptkomponentenanalyse durchgeführt. Es kann ermittelt werden welchen Anteil der gemeinsamen Varianz jede Hauptkomponente erklärt. Dieser Anteil der erklärten Varianz der täglichen Hauptkomponenten wird über die Zeit betrachtet. Die zugrundeliegende Idee ist, dass je weniger Hauptkomponenten benötigt werden, um einen Großteil der gemeinsamen Varianz zu erklären, desto stärkere Zusammenhänge sind zwischen den Variablen vorhanden. Ansonsten wäre es nicht möglich die Varianz dieser Variablen über wenige Linearkombinationen zusammenzufassen. Insbesondere der erste Eigenwert sollte sehr stark mit der Korrelation verknüpft sein, da der zugehörige Eigenvektor die Form der Punktewolke bestmöglich mittels einer geraden Linie darstellt. Alternativ hätten auch die Korrelationsmatrizen für die Hauptkomponentenanalyse verwendet werden können, jedoch sind hier alle Variablen in derselben Einheit (Watt) erhoben worden. In solch einem Fall wird empfohlen keine Standardisierung vorzunehmen (Johnson & Wichern, 2007). Um zu entscheiden wie viele Hauptkomponenten in die Analyse mit einbezogen werden sollten, wird ein Scree-Plot betrachtet, der darstellt welchen maximalen Anteil der Varianz die jeweiligen Hauptkomponenten eines Tages über das Jahr erklären. Was mit maximaler Anteil gemeint ist wird in Abschnitt 3.4 erklärt. Anschließend wird der summierte Anteil der erklärten Varianz der ausgewählten Hauptkomponenten in einer Zeitreihe dargestellt und in Bezug auf Strukturbrüche analysiert. Nebenbei wird auch der Einfluss des Wetters untersucht.

Vor der Präsentation des Datenmaterials wird noch eine alternative Idee für die Zusammenhangsanalyse vorgestellt, die jedoch zu Gunsten der vorab eingeführten Idee verworfen wurde. Auch die Gründe dafür sollen kurz dargelegt werden, da dies zum Einen verdeutlichen kann warum die Eigenwerte genutzt werden und zum Anderen weshalb einzelne Tage betrachtet werden und keine gleitenden Fenster von zum Beispiel 10 Tagen. Ursprünglich sollten für die Analyse der Zusammenhänge gleitende Fenster von 10 Tagen verwendet werden, indem für jedes Fenster die Korrelationsmatrix bestimmt wird. Genau wie im letzten Abschnitt beschrieben, können auch in diesem Fall die Matrizen der gleitenden Fenster nicht über die Zeit dargestellt werden. Die erste Lösung für dieses Problem war die Berechnung der Determinante der Korrelationsmatrizen der gleitenden Fenster, denn diese ist nur dann 1, wenn alle paarweisen Korrelationen Null sind. In diesem Fall handelt es sich nämlich um die Einheitsmatrix. Dagegen strebt sie gegen Null, wenn das Korrelationsvolumen ansteigt. Bei

perfekter Multikollinearität wäre die Determinante also 0 (Shrestha, 2021). Somit ist eine Determinante ein gutes Aggregationsmaß, um eine Korrelationsmatrix zusammenzufassen. Das Problem in diesem Fall ist jedoch, dass die Korrelationsmatrizen eine Dimension von 27x27 haben und die Determinante dadurch bereits bei einem nicht so hohen Korrelationsvolumen sehr schnell einen Wert nahe Null erreicht. Dies liegt daran, dass viele der kleineren Eigenwerte bei einer so hohen Dimension im Falle vorhandener paarweiser Korrelationen schnell in der Nähe von Null liegen und die Determinante das Produkt der Eigenwerte ist. Dadurch sind Unterschiede über die Zeit, insbesondere im höheren Korrelationsbereich, sehr schlecht zu erkennen. Die nächste Idee war es dann direkt die größten Eigenwerte zu betrachten, um wie im letzten Abschnitt beschrieben den Anteil der erklärten Varianz zu untersuchen, welcher ein Indikator für den vorhandenen Zusammenhang sein kann. Zu diesem Zeitpunkt wurde noch mit der Korrelationsmatrix gearbeitet, wobei sich später, wie vorhin bereits beschrieben, aufgrund der gleichen Einheit der Variablen, für die Kovarianzmatrix entschieden wurde. Das nächste Problem ist bei der Strukturbruchanalyse aufgetreten, denn dadurch, dass gleitende Fenster genutzt wurden, konnte auf gar keinen Fall Unabhängigkeit zwischen den Daten angenommen werden, welche, wie in 3.6 beschrieben, eine zentrale Annahme ist. Aufgrund dessen wurde sich für die Betrachtung einzelner Tage entschieden, was deutlich unkritischer ist, jedoch auch keine Unabhängigkeit garantiert. Dieser Aspekt wird jedoch in 4.3 bei der Prüfung der Annahmen erneut aufgegriffen.

2.3 Datenmaterial

Die 33 Datensätze repräsentieren jeweils ein Einfamilienhaus (SFH) und sind mit einer Nummer gekennzeichnet. Für dieses Projekt liegen die Nummern 3-12, 14, 16-23, 25, 27-32 und 34-40 vor. Der strukturelle Aufbau der Datensätze ist identisch und wird in Tabelle 1 dargestellt. Ursprünglich erfolgte die Messung der Stromverbrauchsdaten alle 10 Sekunden und die Anfrage der Wetterdaten alle 5 Minuten. Die für das Projekt vorliegenden Datensätze wurden jedoch bereits vorab aggregiert, indem das arithmetische Mittel der Daten in 15-Minuten-Intervallen bestimmt wurde.

Tabelle 1: Variablen der Datensätze

Variable	Beschreibung
index	15-Minuten-Zeitintervalle im Format yyyy-mm-dd hh:mm:ss
PUMPE_TOT	Arithmetisches Mittel des Stromverbrauchs in Watt der Wärmepumpe im zugehörigen Zeitintervall
HAUSHALT_TOT	Arithmetisches Mittel des Stromverbrauchs in Watt des Haushalts im zugehörigen Zeitintervall
TEMPERATURE:TOTAL	Arithmetisches Mittel der Außentemperatur in Grad Celsius in Hameln im zugehörigen Zeitintervall

Diese Intervalle werden über die Variable „index“ abgebildet. Die Spalte „TEMPERATURE:TOTAL“ enthält für alle 33 Datensätze identische Einträge, da sich alle Häuser in Hameln befinden und die Temperatur, wie in Abschnitt 2.1 beschrieben, über eine API des Wet-

terdienstes WetterOnline für den Ort Hameln angefragt wurde. Der Fokus der Untersuchung liegt auf der Variable „PUMPE_TOT“, wohingegen die Variable „HAUSHALT_TOT“ nur zur Vollständigkeit aufgeführt wurde und im Rahmen dieses Projekts nicht weiter betrachtet wird. Da es für „index“ am 31.03.2019, dem Tag der Zeitumstellung, Einträge für die Uhrzeit zwischen 02:00 Uhr und 03:00 gibt, wird angenommen, dass die Zeitumstellung im Zuge der Erhebung ignoriert wurde.

Da die Datensätze das gesamte Jahr 2019 abbilden, sollte es pro Datensatz $4 \cdot 24 \cdot 365 = 35040$ Beobachtungen geben. Jedoch sind die Datensätze nicht vollständig. Für 27 der 33 Haushalte liegen 34807 Beobachtungen vor, weil vereinzelt über das Jahr verteilt Intervalle fehlen. Da für alle 27 Datensätze dieselben Intervalle fehlen und diese sich auf das ganze Jahr aufteilen, sind die Auswirkungen auf die Analyse des Zusammenhangs zwischen den Wärmepumpen der Haushalte vernachlässigbar. Bei den 6 Datensätzen SFH6, SFH17, SFH25, SFH31, SFH37 und SFH40 fehlt dagegen jeweils mindestens ein Monat am Stück. Ursache für das Fehlen ist, wie in 2.1 beschrieben, meistens ein technisches Problem wie der Ausfall der Messsysteme oder der Internetverbindung. Aufgrund dessen werden diese Haushalte nicht für die Analyse berücksichtigt, denn das würde bedeuten, dass die Berechnung der Hauptkomponenten, der in 2.2 beschriebenen Kovarianzmatrizen, je nach Verfügbarkeit der Daten in den Zeitfenstern auf unterschiedlich großen Matrizen basiert. Zum Beispiel hätte die Kovarianzmatrix im Januar eine Dimension von 33x33, da Daten aller 33 Häuser verfügbar sind und im Dezember nur eine von 30x30, weil die Daten von SFH6, SFH17 und SFH25 fehlen. Nicht nur die Dimension würde sich unterscheiden, sondern auch die Variablen auf denen die Matrizen beruhen. Dadurch würde die Hauptkomponentenanalyse nicht zu jedem Zeitpunkt auf denselben Variablen basieren und Aussagen über die Änderung des Zusammenhangs über die Zeit wären nicht mehr möglich. Somit baut die Analyse auf den 27 weitgehend vollständigen Datensätzen auf. Diese werden in der ursprünglichen Reihenfolge, die auf der Nummerierung basiert, eingelesen und fortan mit H1-H27 bezeichnet. Um die Analyse zu vereinfachen, werden die relevanten Daten in einem großen Datensatz vereint, dessen Aufbau in Tabelle 2 dargestellt wird.

Tabelle 2: Aufbau des aufbereiteten Datensatzes

Variable	Beschreibung
index	15-Minuten-Zeitintervalle im Format yyyy-mm-dd hh:mm:ss
day	Tag im Format yyyy-mm-dd
temperature	Arithmetisches Mittel der Außentemperatur in Grad Celsius in Hameln im zugehörigen Zeitintervall
P_H1	Arithmetisches Mittel des Stromverbrauchs in Watt der Wärmepumpe des 1. Haushalts im zugehörigen Zeitintervall
⋮	⋮
P_H27	Arithmetisches Mittel des Stromverbrauchs in Watt der Wärmepumpe des 27. Haushalts im zugehörigen Zeitintervall

3 Statistische Methoden

Im Rahmen dieser Arbeit werden Variablen mit einem Großbuchstaben, Ausprägungen mit dem zugehörigen Kleinbuchstaben und Matrizen mit einem fettgedruckten Großbuchstaben gekennzeichnet. Zudem werden transponierte Vektoren mit einem hochgestellten T markiert. Schätzer werden außerdem mit einem Dach über dem Buchstaben gekennzeichnet.

3.1 Kovarianz und Korrelation

Bekannte Maße, um den linearen Zusammenhang zwischen zwei metrischen Variablen zu quantifizieren, sind die Kovarianz und der darauf basierende Korrelationskoeffizient nach Bravais-Pearson. Seien X und Y zwei metrische Merkmale mit den Ausprägungen $x = (x_1, ..., x_n) \in \mathbb{R}^n$ und $y = (y_1, ..., y_n) \in \mathbb{R}^n$. Sei weiterhin $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ der Mittelwert und $s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ die Standardabweichung von x. Analog dazu seien auch $\bar{y}$ und s_y gegeben. Dann ist die Kovarianz zwischen x und y definiert als

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Dabei handelt es sich um ein Maß für die gemeinsame Varianz der Ausprägungen der Merkmale (Fahrmeir et al., 2016). Die Kovarianz $s_{xy} \in \mathbb{R}$ ist nicht normiert und stark von den Varianzen s_x^2 und s_y^2 abhängig. Um die Kovarianz zu normieren und eine Vergleichbarkeit zu ermöglichen, wird durch das Produkt der Standardabweichungen der Ausprägungen der beiden Variablen geteilt. Daraus resultiert der Korrelationskoeffizient

$$r_{xy} = \frac{s_{xy}}{s_x s_y}.$$

Die Normierung hat den Effekt, dass r_{xy} nur Werte zwischen -1 und 1 annehmen kann (Fahrmeir et al., 2016). Umso näher der Betrag $|r_{xy}|$ an 1 ist, desto näher liegen die Ausprägungen von X und Y in einem Streudiagramm auf einer Geraden. Ein positiver Korrelationskoeffizient suggeriert einen positiven linearen Zusammenhang. Das bedeutet wenn x steigt, dann auch y. Umgekehrt ist ein negatives r_{xy} so zu interpretieren, dass wenn x wächst, dann sinkt y. Da Zusammenhänge nicht zwingend linear sein müssen, bedeutet $r_{xy} = 0$ nur, dass kein linearer Zusammenhang vorliegt.

Falls tatsächlich ein hoher linearer Zusammenhang besteht, liegt es nahe diesen mit einer Geraden zu modellieren. Da r_{xy} jedoch meistens nicht 1 ist, kann in der Regel keine Gerade gefunden werden, die exakt durch alle Punkte im Streudiagramm verläuft. Aufgrund dessen wird das Ziel verfolgt eine lineare Funktion zu finden, die möglichst nah an den Beobachtungen liegt, um den größtmöglichen Anteil der Variabilität in den Daten zu erklären. Daraus ergibt sich das Modell

$$Y = f(X) + \varepsilon,$$

wobei $f(X) = \alpha + \beta X$ mit $\alpha, \beta \in \mathbb{R}$ die lineare Beziehung darstellt und $\varepsilon \in \mathbb{R}^n$ ein zufälliger additiver Fehlerterm ist, da nicht angenommen werden kann, dass der funktionale Zusammenhang exakt zutrifft (Fahrmeir et al., 2016). Die resultierende empirische Beziehung ist somit

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

für $i = 1, ..., n$ (Fahrmeir et al., 2016). Das bekannteste Verfahren, um den Achsenabschnitt α und die Steigung β zu schätzen, ist die Methode der kleinsten Quadrate (Fahrmeir et al., 2016). Hierbei wird die durchschnittliche quadratische Differenz zwischen den tatsächlichen Beobachtungen y_i und den vorhergesagten Werten des Modells $\hat{y}_i$ für $i = 1, ..., n$ minimiert. Die Minimierung der Funktion

$$q(\alpha, \beta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

über die partiellen Ableitungen nach α und β liefert die Schätzer

$$\begin{aligned} \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x}\,, \\ \hat{\beta} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \end{aligned}$$

Somit ergibt sich die Ausgleichsgerade

$$\hat{y} = \hat{\alpha} + \hat{\beta} x.$$

3.2 Korrelations- und Kovarianzmatrix

Wird der Zusammenhang von mehr als zwei Variablen betrachtet, wie auch in diesem Projekt, ist die sogenannte Korrelationsmatrix eine geeignete Methode, um die paarweisen Korrelationen nach Bravais-Pearson darzustellen. Seien $X_1, ..., X_d \in \mathbb{R}^n$ die Variablen und $x_i \in \mathbb{R}^n$ die zugehörigen Ausprägungen für $i = 1, ..., d$ zwischen denen Korrelationen bestimmt werden sollen. Die paarweisen Korrelationen zwischen x_i und x_j für $i, j = 1, ..., d$ werden mit $r_{x_i x_j}$ bezeichnet. Dann ist die Korrelationsmatrix definiert als

$$\mathbf{R} = \begin{bmatrix} r_{x_1x_1} & r_{x_1x_2} & \dots & r_{x_1x_d} \\ r_{x_2x_1} & r_{x_2x_2} & \dots & r_{x_2x_d} \\ \vdots & \vdots & \ddots & \vdots \\ r_{x_dx_1} & r_{x_dx_2} & \dots & r_{x_dx_d} \end{bmatrix} = \begin{bmatrix} 1 & r_{x_1x_2} & \dots & r_{x_1x_d} \\ r_{x_2x_1} & 1 & \dots & r_{x_2x_d} \\ \vdots & \vdots & \ddots & \vdots \\ r_{x_dx_1} & r_{x_dx_2} & \dots & 1 \end{bmatrix} \in [-1, 1]^{dxd}.$$

Der Korrelationskoeffizient nach Bravais-Pearson ist zwischen den Ausprägungen derselben Variablen immer $r_{x_i x_i} = 1$ für i=1,...,d, weshalb auf der Diagonalen ausschließlich Einsen stehen. Weiterhin ist die quadratische Matrix symmetrisch, da $r_{x_i x_j} = r_{x_j x_i}$. Somit enthält nur eine Dreieckshälfte über beziehungsweise unter der Diagonalen relevante Informationen.

Die Matrix der Kovarianzen ist analog aufgebaut. Seien $s_{x_i x_j}$ die paarweisen Kovarianzen für $i, j = 1, ..., d$. Dann ist die Kovarianzmatrix definiert als

$$\mathbf{S} = \begin{bmatrix} s_{x_1x_1} & s_{x_1x_2} & \cdots & s_{x_1x_d} \\ s_{x_2x_1} & s_{x_2x_2} & \cdots & s_{x_2x_d} \\ \vdots & \vdots & \ddots & \vdots \\ s_{x_dx_1} & s_{x_dx_2} & \cdots & s_{x_dx_d} \end{bmatrix} = \begin{bmatrix} s^2_{x_1} & s_{x_1x_2} & \cdots & s_{x_1x_d} \\ s_{x_2x_1} & s^2_{x_2} & \cdots & s_{x_2x_d} \\ \vdots & \vdots & \ddots & \vdots \\ s_{x_dx_1} & s_{x_dx_2} & \cdots & s^2_{x_d} \end{bmatrix} \in \mathbb{R}^{dxd}.$$

Die Kovarianz zwischen den Ausprägungen derselben Variablen entspricht immer der Varianz $s^2_{x_i}$ für i=1,...,d, weshalb auf der Diagonalen die Varianzen stehen. Auch diese Matrix ist symmetrisch. Die im Folgenden vorgestellte Hauptkomponentenanalyse basiert vor allem auf der Kovarianzmatrix.

3.3 Hauptkomponentenanalyse

Ziel der Hauptkomponentenanalyse ist es den Großteil der gemeinsamen Varianz der Variablen eines hochdimensionalen Datensatzes mittels weniger Linearkombinationen der ursprünglichen Variablen zu erklären. Es handelt sich also um ein Verfahren, welches darauf abzielt die Dimension zu reduzieren. Es soll in diesem Projekt eingesetzt werden, um die Dimension der Kovarianzmatrizen zu reduzieren, damit dargestellt werden kann wie sich die Struktur der gemeinsamen Varianz des Stromverbrauchs der Wärmepumpen über die Zeit verändert.

3.3.1 Grundlagen

Da Eigenwerte und Eigenvektoren die Grundlage für dieses Verfahren bilden, werden diese vorab definiert. Davor wird zudem auch noch die Determinante definiert, da diese genutzt wird, um die Eigenwerte einer Matrix zu bestimmen. Sei $\mathbf{A} \in \mathbb{R}^{nxn}$ eine quadratische Matrix mit den Einträgen $a_{ij} \in \mathbb{R}$ für $i, j = 1, ..., n$, wobei i die Zeile repräsentiert und j die Spalte. Sei weiterhin $\mathbf{I}^{nxn}$ die Einheitsmatrix. Dann ist die Determinante $|\mathbf{A}| \in \mathbb{R}$ rekursiv definiert als

$$|\mathbf{A}| = a_{11}, \text{ falls n=1}$$
$$|\mathbf{A}| = \sum_{j=1}^{n} a_{1j}|\mathbf{A}_{1j}|(-1)^{1+j}, \text{ falls n > 1.}$$

Die Matrix $\mathbf{A}_{1j} \in \mathbb{R}^{(n-1)x(n-1)}$ entsteht, indem die erste Zeile und die Spalte j von $\mathbf{A}$ entfernt werden (Johnson & Wichern, 2007). Ein Anwendungsfall der Determinante ist die Bestim-

mung von Eigenwerten, die nun erklärt wird. Ein Nicht-Nullvektor $e \in \mathbb{R}^n$ heißt Eigenvektor von $\mathbf{A}$, falls ein Skalar $\lambda \in \mathbb{R}$ existiert, sodass gilt

$$\mathbf{A}e = \lambda e.$$

Das heißt ein Eigenvektor e verändert sich nur um ein Skalar λ, wenn die lineare Transformation $\mathbf{A}$ auf diesen angewandt wird (Kwak & Hong, 2004). Dieses λ ist ein Eigenwert von $\mathbf{A}$ und e gehört zu diesem λ. Sowohl λ als auch e sind unbekannt. Zunächst werden die Eigenwerte bestimmt, indem ausgenutzt wird, dass die Gleichung

$$(\lambda \mathbf{I}^{nxn} - \mathbf{A})e = 0$$

nur dann Lösungen ungleich dem Nullvektor hat, wenn für λ gilt

$$|(\lambda \mathbf{I}^{nxn} - \mathbf{A})| = 0.$$

Hierbei handelt es sich um ein Polynom von Grad n, sodass maximal n reelle Lösungen für λ existieren, was bedeutet, dass n reelle Eigenwerte existieren können (Kwak & Hong, 2004). Die zugehörigen Eigenvektoren e können bestimmt werden, indem für jeden Eigenwert λ das homogene Gleichungssytem

$$(\lambda \mathbf{I}^{nxn} - \mathbf{A})e = 0$$

gelöst wird (Kwak & Hong, 2004). Wenn n reelle Eigenwerte existieren, gilt, dass die Spur von $\mathbf{A}$, also die Summe der Hauptdiagonalelemente, der Summe aller Eigenwerte von $\mathbf{A}$ entspricht (Kwak & Hong, 2004). Falls die Matrix symmetrisch ist, wie zum Beispiel die Kovarianzmatrix, dann existieren immer n reelle Eigenwerte (Kwak & Hong, 2004). Die Eigenvektoren und Eigenwerte ermöglichen zum Beispiel die Hauptkomponentenanalyse, die nun präsentiert wird.

3.3.2 Hauptkomponenten

Sei $\mathbf{X} \in \mathbb{R}^{nxd}$ eine Matrix, die einen Datensatz mit n Beobachtungen und d Variablen $X_1, ..., X_d$ repräsentiert. Die Variablen $X_1, ..., X_d \in \mathbb{R}^n$ für $i = 1, ..., d$ sind die Spalten der Matrix $\mathbf{X}$. Seien weiterhin $a_1, ..., a_d \in \mathbb{R}^n$ für $i = 1, ..., d$ Vektoren der Dimension n. Basierend auf der Kovarianzmatrix $\mathbf{S} \in R^{dxd}$ von $\mathbf{X}$ können die Linearkombinationen für $\mathbf{X}$ bestimmt werden, die den größten Anteil der gemeinsamen Varianz erklären. Da die Matrix symmetrisch und positiv-semidefinit ist, existieren genau d nicht-negative Eigenwerte $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_d$ mit zugehörigen zueinander orthogonalen Eigenvektoren $e_1, ..., e_d$ (Kwak & Hong, 2004). Positiv-semidefinit bedeutet, dass für $\mathbf{S}$ gilt, dass

$$v^T \mathbf{S} v \geq 0 \quad \forall v \in \mathbb{R}^d.$$

Diese der Größe nach geordneten Eigenwerte und die zugehörigen Eigenvektoren ergeben die zueinander orthogonalen und damit unkorrelierten Hauptkomponenten

$$Z_i = \sum_{j=1}^{d} e_{ij} X_j$$

für $i = 1, ..., d$, wobei e_{ij} der j-te Eintrag des Vektors e_i ist (Johnson & Wichern, 2007). Die erste Hauptkomponente ist die Linearkombination von allen möglichen Linearkombinationen der ursprünglichen Variablen, die den größten Anteil der Varianz erklärt. Die zweite Hauptkomponente ist die Linearkombination der ursprünglichen Variablen, die orthogonal zu Z_1 ist und den größten Anteil der Varianz erklärt. Diese Definition setzt sich für die weiteren Hauptkomponenten fort. Somit maximiert die Linearkombination Z_1 die Varianz

$$max_{a_1} Var(a_1^T \mathbf{X}) = max_{a_1} a_1^T \mathbf{S} a_1 = Var(Z_1) = \lambda_1,$$

wobei $a_1^T a_1 = 1$ (Johnson & Wichern, 2007). Die i-te Hauptkomponente maximiert für $i = 2, ..., d$

$$max_{a_i} Var(a_i^T \mathbf{X}) = max_{a_i} a_i^T \mathbf{S} a_i = Var(Z_i) = \lambda_i,$$

wobei $a_i^T a_i = 1$ und $Cov(a_i^T \mathbf{X}, a_k^T \mathbf{X}) = 0$ für $k < i$, wodurch die Orthogonalität sichergestellt wird (Johnson & Wichern, 2007).

Sowohl die Summe der Varianz der Variablen $X_1, ..., X_d$ als auch die Summe der Eigenwerte $\lambda_1, ..., \lambda_d$ entspricht der Spur der Kovarianzmatrix

$$Spur(\mathbf{S}) = \sum_{i=1}^{d} Var(X_i) = \sum_{i=1}^{d} \lambda_i.$$

Das bedeutet, dass der Anteil der erklärten Varianz der i-ten Hauptkomponente gegeben ist als

$$\text{Anteil erklärter Varianz} = \frac{\lambda_i}{\sum_{j=1}^{d} \lambda_j}$$

für $i = 1, ..., d$ (Johnson & Wichern, 2007). Dieser Anteil der erklärten Varianz der Hauptkomponenten soll in dieser Arbeit über die Zeit dargestellt und im Hinblick auf Strukturbrüche untersucht werden. Um zu entscheiden wie viele der Hauptkomponenten für diese

Analyse betrachtet werden sollen, wird ein Scree-Plot betrachtet, der im Folgenden erklärt wird.

3.4 Scree-Diagramm

Sei $\mathbf{S} \in \mathbb{R}^{dxd}$ weiterhin die Kovarianzmatrix mit den d Eigenwerten $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_d$. Das Scree-Diagramm stellt den Anteil der erklärten Varianz der Hauptkomponenten auf der vertikalen Achse gegen den Index dieser Hauptkomponenten auf der horizontalen Achse dar (Johnson & Wichern, 2007). Dabei werden die Komponenten in Abhängigkeit der Eigenwerte in absteigender Reihenfolge, also vom größten zum kleinsten Eigenwert, sortiert. Dies ermöglicht einen schnellen Überblick über die Relevanz der einzelnen Hauptkomponenten (Johnson & Wichern, 2007). Da in diesem Projekt jedoch für jeden Tag des Jahres 2019 Hauptkomponenten bestimmt werden und dies in 365 Scree-Diagrammen resultieren würde, wird stattdessen ein Scree-Diagramm erstellt, das den maximalen Anteil der erklärten Varianz der jeweils i-ten Komponente für $i = 1, ..., d$ darstellt. Alternativ hätte auch der Durchschnitt betrachtet werden können, jedoch wird durch das Maximum sichergestellt, dass Komponenten, die mindestens an einem Tag besonders relevant waren, miteinbezogen werden. In der Regel ist ein plötzlicher Knick in der Abbildung zu erkennen, der dem sogenannten Ellenbogen-Kriterium seinen Namen gibt. Nach diesem Kriterium wird sich für alle Komponenten entschieden, die sich links von der Knickstelle befinden (Johnson & Wichern, 2007). Nachdem die Anzahl feststeht, kann die Summe der erklärten Varianz dieser Hauptkomponenten in einer Zeitreihe dargestellt werden.

3.5 Zeitreihe

Eine Zeitreihe kann als eine Realisierung $y_{1:n} = (y_1, ..., y_n)$ eines stochastischen Prozess $Y_{1:n} = (Y_1, ..., Y_n)$ definiert werden (Kirchgässner et al., 2013). Angenommen die Variablen des stochastischen Prozesses seien alle stetig. Dann sind die Erwartungswerte für $i = 1, ..., n$ definiert als

$$E[Y_i] = \int_{-\infty}^{\infty} y_i \, f_{Y_i}(y_i) \, dy_i,$$

wobei f_{Y_i} die Dichte der Zufallsvariable Y_i ist (Fahrmeir et al., 2016). Weiterhin sind die Varianz und Autokovarianz für $i, j = 1, ..., n$ gegeben als

$$\begin{aligned} Var[Y_i] &= E[(Y_i - E[Y_i])^2], \\ Cov[Y_i, Y_j] &= E[(Y_i - E[Y_i])(Y_j - E[Y_j])], \end{aligned}$$

wobei $i < j$ (Kirchgässner et al., 2013). Der Name Autokovarianz kommt daher, dass es sich um die Kovarianz zwischen zwei Variablen aus demselben stochastischen Prozess handelt. Sollte der stochastische Prozess multivariat normalverteilt sein, reichen die ersten beiden

Momente aus, um die vollständige Verteilung anzugeben (Kirchgässner et al., 2013). Eine wichtige Annahme in der Zeitreihenanalyse ist die Ergodizität, die besagt, dass die empirischen Momente für $n \to \infty$ gegen die theoretischen Momente konvergieren (Kirchgässner et al., 2013). Dies macht aber nur dann Sinn, wenn für alle $i = 1, ..., n$ gilt, dass der Erwartungswert $E[Y_i] = \mu$ und die Varianz $Var(Y_i) = \sigma_Y^2$ konstant, auch stationär genannt, sind (Kirchgässner et al., 2013). Die Ergodizität bezüglich des Mittelwerts und der Varianz sind also definiert als

$$\lim_{n\to\infty} E\left[\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - \mu\right)^2\right] = 0,$$
$$\lim_{n\to\infty} E\left[\left(\frac{1}{n}\sum_{i=1}^{n} (Y_i - \mu)^2 - \sigma_{Y_i}^2\right)^2\right] = 0.$$

Damit ein stochastischer Prozess als stationär bezeichnet werden kann, muss neben der Stationarität des Erwartungswerts und der Varianz auch die Autokovarianz stationär sein. Die Autokovarianz-Stationarität ist definiert als

$$Cov[Y_i, Y_j] = E[(Y_i - E[Y_i])(Y_j - E[Y_j])] = \gamma(|j - i|).$$

Das heißt die Autokovarianz muss als Funktion darstellbar sein, die nur noch von dem Zeitabstand zwischen i und j abhängt (Kirchgässner et al., 2013). Analog zu 3.1 ist auch die Autokovarianz nicht normiert, sodass Aussagen über die Stärke der linearen Abhängigkeit erst nach einer Normierung möglich sind. Die normierte Version wird als Autokorrelation bezeichnet und ist für stationäre Prozesse definiert als

$$\rho(d) = \frac{E[(Y_i - \mu)(Y_{i+d} - \mu)]}{E[(Y_i - \mu)^2]} = \frac{\gamma(d)}{\gamma(0)} \in [-1, 1]$$

für $d = ..., -1, 0, 1, ...$ (Kirchgässner et al., 2013). Für $d = 0$ ist $\rho(0) = 1$ und weiterhin gilt für alle d die Symmetrieeigenschaft $\rho(d) = \rho(-d)$ (Kirchgässner et al., 2013). Der Schätzer

$$\hat{\rho}(d) = \frac{\sum_{i=1}^{n-d} (y_i - \hat{\mu})(y_{i+d} - \hat{\mu})}{\sum_{i=1}^{n} (y_i - \hat{\mu})^2}$$

kann mittels $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$ bestimmt werden (Kirchgässner et al., 2013). Falls zusätzlich gilt, dass $\mu = 0$ und $Cov[Y_i, Y_j] = 0 \quad \forall i \neq j$, dann handelt es sich um einen sogenannten White-Noise-Prozess (Kirchgässner et al., 2013). In diesem speziellen Fall ist $\rho(d)$ asymptotisch normalverteilt mit Erwartungswert 0 und geschätzter Varianz von $\frac{1}{n}$ (Kirchgässner et al., 2013). Aufgrund dessen kann das approximative 95%-Konfidenzintervall von $\hat{\rho}(d)$ angegeben werden als

$$\pm\frac{1.96}{\sqrt{n}}.$$

Die Stationarität und Autokorrelation sind wichtige Eigenschaften für die Strukturbruchanalyse, die nun vorgestellt wird.

3.6 Strukturbruchanalyse

Ziel der Strukturbruchanalyse ist es eine Zeitreihe, die plötzliche Änderungen aufweist, in homogene Segmente zu zerlegen. Sei eine Zeitreihe $y_{1:n} = (y_1, ..., y_n) \in \mathbb{R}^n$ gegeben. Sei weiterhin $m \in \mathbb{N}$ die Anzahl der Strukturbrüche und $\tau_{1:m} = (\tau_1, ..., \tau_m) \in \{1, ..., n-1\}^m$ die zugehörige Position des Strukturbruchs, wobei $\tau_0 = 0$ und $\tau_{m+1} = n$. Die Positionen sind geordnet, sodass gilt $\tau_i < \tau_j$, genau dann wenn $i < j$. Für die resultierenden $m+1$-Segmente müssen die zugehörigen Verteilungskennzahlen θ_i für $i = 1, ..., m+1$ geschätzt werden. Dafür wird die Likelihood-Funktion

$$L(m, \tau_{1:m}, \theta_{1:m+1}) = p(y_{1:n}|m, \tau_{1:m}, \theta_{1:m+1})$$

genutzt (Eckley et al., 2011). Weiterhin wird angenommen, dass die Daten über die Segmente hinweg bedingt unabhängig sind, sodass für die bedingte Dichte gilt

$$p(y_{1:n}|m, \tau_{1:m}, \theta_{1:m+1}) = \prod_{i=1}^{m+1} p(y_{(\tau_{i-1}+1):\tau_i}|\theta_i).$$

Der resultierende Maximum-Likelihood-Schätzer $\hat{\theta}_i$ für das Segment i ist dann definiert als

$$max_{\theta_i} p(y_{(\tau_{i-1}+1):\tau_i}|\theta_i) = p(y_{(\tau_{i-1}+1):\tau_i}|\hat{\theta}_i).$$

Ein bekanntes Verfahren, um die Zeitpunkte der Strukturbrüche zu ermitteln, ist die Minimierung von

$$\sum_{i=1}^{m+1} [C(y_{(\tau_{i-1}+1):\tau_i})] + \beta f(m),$$

wobei C als Kostenfunktion eines Segments fungiert und die monoton steigende Funktion $\beta f(m)$ eine Strafe ist, um Overfitting zu vermeiden (Killick & Eckley, 2014). Umso höher die Strafe, desto weniger Strukturbrüche werden gefunden. In dem verwendeten R-Paket Changepoint wird das zweifache der negativen Log-Likelihood als Kostenfunktion verwendet (Killick, et al., 2022; Eckley et al., 2011). Weiterhin wird $\beta \approx 3 \cdot m \cdot \frac{log(n)}{2}$ genutzt, was als mBIC bezeichnet wird (Killick et al., 2012; Zhang & Siegmund, 2007).

Binary Segmentation ist ein Algorithmus, der versucht das Minimierungsproblem approximativ für $f(m) = m$ zu lösen (Killick et al., 2012). Zu Beginn wird ermittelt, ob ein τ existiert, welches die Ungleichung

$$C(y_{1:\tau}) + C(y_{(\tau+1):n}) + \beta < C(y_{1:n})$$

erfüllt (Killick et al., 2012). Für die erste Iteration wird $m = 1$ gewählt. Falls kein τ existiert, das dies erfüllt, stoppt der Algorithmus und liefert, dass es keine Strukturbrüche gibt. Wenn genau ein τ gefunden wird, dann werden die Daten in zwei Segmente zerlegt, welche durch das ermittelte τ getrennt werden. Sollten mehrere τ gefunden werden, die die Ungleichung erfüllen, dann wird das τ gewählt, welches den linken Teil der Ungleichung minimiert. Das gleiche Verfahren wird auf die beiden neuen Abschnitte angewandt, wobei nun $m = 2$ ist. Das heißt für jede Iteration werden die bereits ermittelten Strukturbrüche in den Strafterm einberechnet. Dies wird so lange fortgeführt bis kein weiterer Strukturbruch gefunden wird (Eckley et al., 2011).

In diesem Projekt wird eine Strukturbruchanalyse des Mittelwerts durchgeführt, sodass θ in diesem Fall den Erwartungswert definiert. Für das Likelihood-Verfahren zur Ermittlung von Strukturbrüchen bezüglich des Mittelwerts wird im für die Analyse verwendeten Changepoint-Paket angenommen, dass für die Verteilung in den jeweiligen Segmenten gilt

$$\begin{aligned}
&Y_t|\theta_1 \sim N(\theta_1, 1), && 1 \leq t \leq \tau_1 \\
&Y_t|\theta_2 \sim N(\theta_2, 1), && \tau_1 < t \leq \tau_2 \\
&\vdots && \vdots \\
&Y_t|\theta_{m+1} \sim N(\theta_{m+1}, 1), && \tau_m < t \leq \tau_{m+1} = n.
\end{aligned}$$

Insbesondere bedeutet das, dass die Zeitreihe in den jeweiligen Segmenten stationär ist. Die Varianz darf sogar über die gesamte Zeit keine Veränderung aufweisen. Sollte keine Varianz von 1 vorliegen, müssen die Daten vorab standardisiert werden, indem der Mittelwert von jedem Vektoreintrag subtrahiert wird und durch die Standardabweichung der Daten geteilt wird (Killick, 2021). Dabei wird das Ziel verfolgt in den jeweiligen Segmenten eine approximative Varianz von 1 zu erhalten. Vor allem wenn tatsächlich Strukturbrüche vorliegen, kommt es im Rahmen dieser Normalisierung zu Schätzfehlern. Diese sind jedoch deutlich unkritischer als keine Normalisierung vorzunehmen. Das resultierende Modell hat also folgende Form

$$Y_t = \begin{cases}
\theta_1 + \varepsilon_t, & 1 \leq t \leq \tau_1 \\
\theta_2 + \varepsilon_t, & \tau_1 < t \leq \tau_2 \\
\vdots & \vdots \\
\theta_{m+1} + \varepsilon_t, & \tau_m < t \leq \tau_{m+1} = n,
\end{cases}$$

wobei $\varepsilon_t \sim N(0,1)$ für $t = 1, ..., n$ (Killick, 2021).

Eine Überprüfung der Normalverteilung sowie der Unabhängigkeit der Daten ist vor der Analyse nicht möglich. Denn wenn ein Strukturbruch vorliegt, dann verändert sich der Verteilungsparameter θ über die Zeit. Auch die Autokorrelation, die im Falle einer Unabhängigkeit der Daten 0 sein sollte, wird von einem Wechsel des Erwartungswerts ansteigen, da auch diese vom Erwartungswert abhängt. Die Prüfung der Annahmen kann also erst erfolgen, wenn das Modell aufgestellt wurde, indem die Residuen betrachtet werden (Killick, 2021). Dafür werden die Residuen der ursprünglichen Daten von den geschätzten Erwartungswerten der Segmente bestimmt. Das bedeutet die geschätzten Residuen sind definiert als

$$\hat{\varepsilon}_t = \begin{cases} y_t - \hat{\theta}_1, & 1 \leq t \leq \tau_1 \\ y_t - \hat{\theta}_2, & \tau_1 < t \leq \tau_2 \\ \vdots & \vdots \\ y_t - \hat{\theta}_{m+1}, & \tau_m < t \leq \tau_{m+1} = n. \end{cases}$$

Die Residuen sollten alle standardnormalverteilt sein und keine Autokorrelation aufweisen, wenn die Annahmen gestimmt haben (Killick, 2021). Aufgrund dessen werden sie in einem Vektor zusammengefasst und auf Standardnormalverteilung sowie Autokorrelation geprüft (Killick, 2021). Die Prüfung der Standardnormalverteilung kann mittels eines Normal-Quantil-Diagramms erfolgen, welches im folgenden Unterabschnitt vorgestellt wird. Die Autokorrelation wird mittels eines ACF-Diagramms geprüft, das in Abschnitt 3.8 präsentiert wird.

3.7 Normal-Quantil-Diagramm

Das Normal-Quantil-Diagramm vergleicht die theoretischen Quantile der Standardnormalverteilung mit den empirischen Quantilen. Seien $x_{(1)},...,x_{(n)} \in \mathbb{R}$ für $i = 1,...,n$ der Größe nach geordnete Beobachtungen. Diese $x_{(i)}$ werden als die empirischen $\frac{1}{n}$-Quantile aufgefasst und werden nun gegen Quantile der Standardnormalverteilung aufgetragen. Jedoch erfolgt vorab eine Stetigkeitskorrektur, indem statt der $\frac{i}{n}$-Quantile die $\frac{i-0.5}{n}$-Quantile der Standardnormalverteilung verwendet werden, denn dadurch „wird die Approximation der empirischen Verteilung durch eine Normalverteilung verbessert“ (Fahrmeir et al., 2016). Diese angepassten Quantile werden mit $z_{(i)}$ bezeichnet. Das resultierende Diagramm besteht aus den Punkten

$$(z_{(1)}, x_{(1)}), ..., (z_{(n)}, x_{(n)}),$$

wobei z die horizontale Achse ist und x die vertikale Achse (Fahrmeir et al., 2016). Sollte die empirische Verteilung tatsächlich approximativ standardnormalverteilt sein, dann sollten diese Punkte auf beziehungsweise nah an der Winkelhalbierenden liegen.

3.8 Autokorrelationsfunktion-Diagramm

Sei eine Zeitreihe $y_{1:n}$ gegeben. Die Autokorrelation bestimmt den Zusammenhang der Zeitreihe mit einer um d-Zeiteinheiten verzögerten Version der Zeitreihe. Der sogenannte ACF-Plot stellt diese Autokorrelation für aufsteigende $d = 1, ..., n-1$ in einem Stabdiagramm dar (Hyndman & Athanasopoulos, 2013). Es reicht ausschließlich positive d zu betrachten, da die Autokorrelation, wie in 3.5 beschrieben, symmetrisch ist. Weiterhin ist es üblich das 95%-Konfidenzintervall mit zwei horizontalen Linien zu markieren. Falls mehr als 5% der Autokorrelationen dieses Intervall überschreiten, kann nicht davon ausgegangen werden, dass die Autokorrelation Null beträgt.

4 Statistische Auswertung

Die Auswertung erfolgte mittels der Software R und auch alle folgenden Abbildungen wurden mit R erstellt (R Core Team, 2023).

4.1 Deskriptive Analyse des Haushalts H1

Zunächst wird, wie in 2.2 beschrieben, eine deskriptive Analyse des ersten Haushalts durchgeführt. Dafür wurde die Zeitreihe in Abbildung 2 erstellt, die den Stromverbrauch der Wärmepumpe im Verlaufe des Jahres 2019 darstellt.

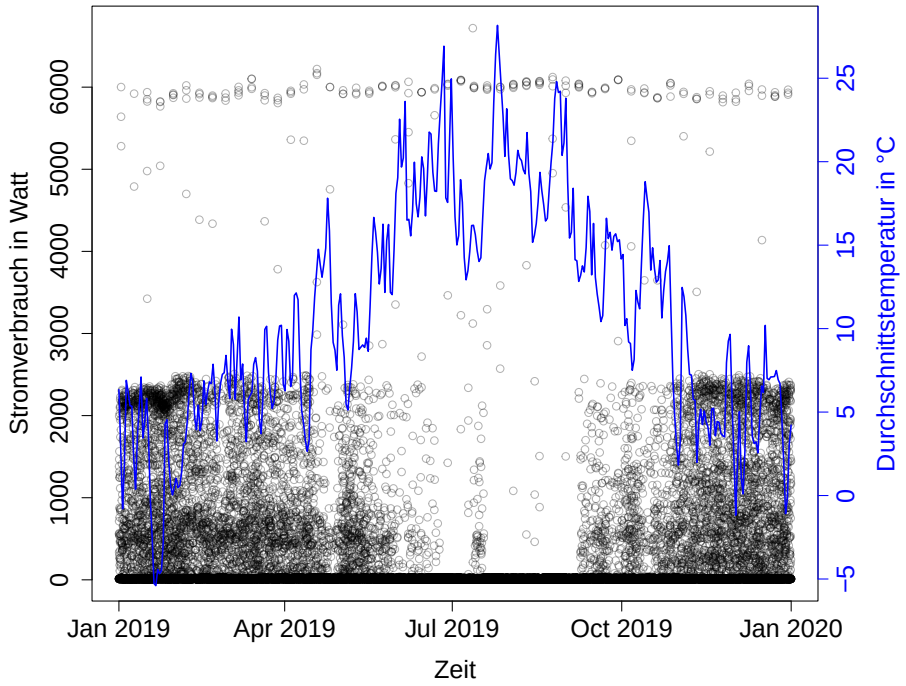

Abbildung 2: Stromverbrauchsdaten der Wärmepumpe des Haushalts H1 und Tagesdurchschnittstemperatur des Ortes Hameln im Verlaufe des Jahres 2019.

Es wurde sich für ein Streudiagramm mit leicht transparenten Punkten entschieden, da durch den sehr wechselhaften Stromverbrauch der Pumpen in einem Liniendiagramm nicht viel

erkennbar wäre. Es ist deutlich zu sehen, dass der Verbrauch der Pumpen im Winter am höchsten ist, denn in den Monaten Januar, Febraur und Dezember sind dunkle Punktewolken um den Bereich von 2000 Watt auszumachen. Diese sind im Frühjahr sowie im Herbst deutlich heller, wobei es im Sommer kaum Punkte in diesem Bereich gibt. Dieses Muster ist auch bei den anderen Haushalten wahrnehmbar. Mitte Juli gibt es zwar wieder einen erhöhten Verbrauch, jedoch ist in diesem Zeitraum auch ein Einbruch der Temperatur wahrzunehmen. Weiterhin sind über das gesamte Jahr verteilt vereinzelt sehr plötzliche extreme Verbräuche von etwa 6000 Watt identifizierbar. Ähnliches ist auch bei anderen Haushalten wahrnehmbar, jedoch nicht bei allen. Eine mögliche Ursache könnte ein Desinfektionsverfahren der Pumpen sein, welches automatisiert in regelmäßigen Abständen die Wärmepumpe durchheizt, um mögliche Erreger abzutöten (Legionellen keine Chance geben, o.D.). Dafür spricht, dass dies auch im Sommer passiert, der Verbrauch jeweils fast identisch ist und regelmäßige Abstände erkennbar sind. Eine weitere Auffälligkeit ist die dunkle durchgezogene Linie nahe 0. Ein wesentlicher Grund dafür ist die gewählte Skala auf der vertikalen Achse zwischen 0 und 6800 Watt, sodass Werte nahe 0 schlecht unterscheidbar sind. Aufgrund dessen werden in Abbildung 3 ausschließlich Beobachtungen dargestellt, die zwischen 0 und 100 Watt liegen.

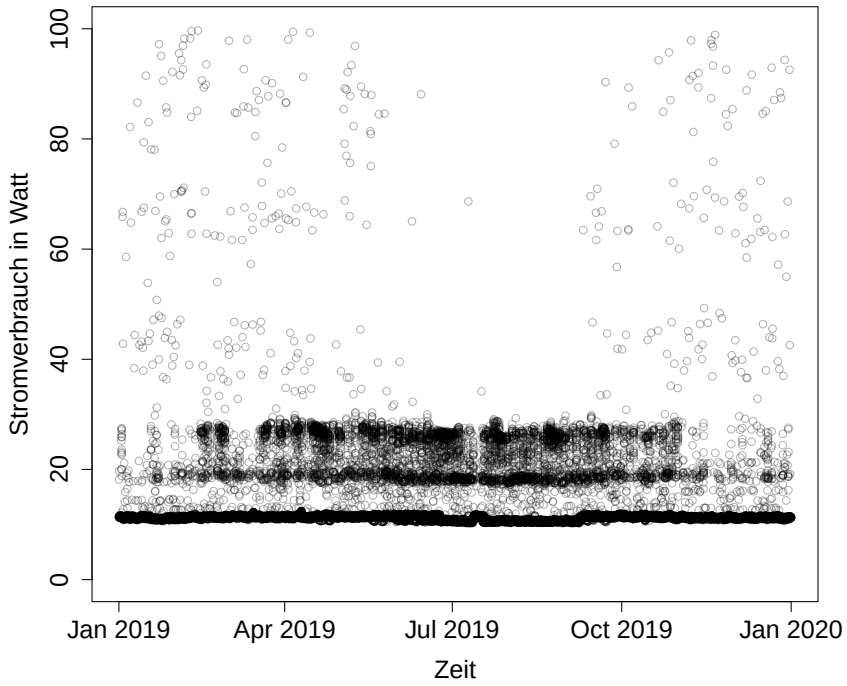

Abbildung 3: Stromverbrauchsdaten der Wärmepumpe des Haushalts H1 zwischen 0 und 100 Watt im Verlaufe des Jahres 2019.

Dadurch ist erkennbar, dass es vor allem im Frühling, Sommer und Herbst sehr viele Beobachtungen zwischen 10 und 30 Watt gibt. Die durchgezogene Linie um 11 Watt resultiert aus dem Standby-Betrieb der Wärmepumpe, welche die Nicht-Heiz-Phasen kennzeichnet.

4.2 Hauptkomponentenanalyse

Für die Hauptkomponentenanalyse werden für jeden Tag basierend auf den Stromverbrauchsdaten der Wärmepumpen Kovarianzmatrizen bestimmt. Diese Kovarianzmatrizen basieren auf jeweils $4 \cdot 24 = 96$ Beobachtungen für jeden der 27 Haushalte, insofern für diesen Tag keine Beobachtungen fehlen. Anschließend werden die zugehörigen der Größe nach geordneten Eigenwerte, die die täglichen Hauptkomponenten repräsentieren, für jede dieser Matrizen bestimmt. Um zu ermitteln welchen Anteil der Varianz jede diese Hauptkomponenten erklärt, wird jeder Eigenwert durch die Summe der Eigenwerte des jeweiligen Tages geteilt. Dieser Anteil der erklärten gemeinsamen Varianz durch die wichtigsten Hauptkomponenten eines Tages wird in Abschnitt 4.3 in einer Zeitreihe dargestellt und auf Strukturbrüche untersucht. Um zu ermitteln welche der Hauptkomponenten zu den Wichtigsten zählen, wird vorab ein Scree-Diagramm betrachtet. Bevor jedoch mittels eines Scree-Diagramms entschieden wird welche Hauptkomponenten für die weitere Analyse berücksichtigt werden, wird die jeweils erste Hauptkomponente genauer betrachtet. Diese sollte, wie in 2.2 beschrieben, besonders stark mit der Korrelation verknüpft sein. Aufgrund dessen wird bestimmt an welchem Tag die jeweils erste Hauptkomponente in den 365 Tagen die maximale Varianz erklärt. Die erste Hauptkomponente vom 28.08.2019 erklärt, verglichen mit den anderen Tagen, mit 95.29% die maximale gemeinsame Varianz. In Abbildung 4, welche mit dem Paket Corrplot erstellt wurde, wird die zugehörige Korrelationsmatrix des Tages betrachtet (Wei & Simko, 2021).

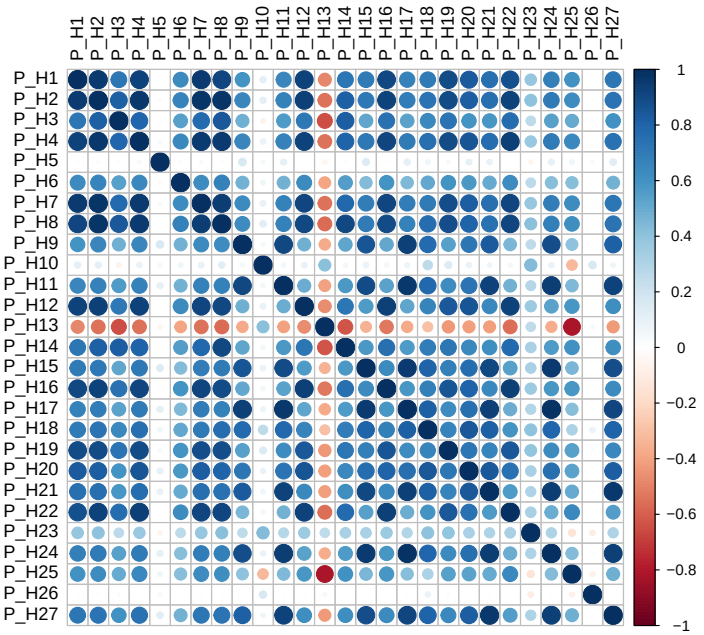

Abbildung 4: Korrelationsmatrix des Stromverbrauchs der Wärmepumpen vom 28.08.2019.

Es ist deutlich zu erkennen, dass die Wärmepumpen an diesem Tag besonders stark positiv korreliert sind. Die höchste Korrelation zwischen dem Stromverbrauch von zwei Wärmepumpen an diesem Tag ist 0.98, also fast maximal. Das zugehörige Streudiagramm mit der Ausgleichsgerade von den Pumpen P_H17 und P_H24 befindet sich in Abbildung 11 im

Anhang. Die hohe positive Korrelation kommt vor allem dadurch zustande, dass die Pumpen in den gleichen Zeitintervallen im Standby-Betrieb sind oder arbeiten. Auffällig ist, dass die Wärmepumpen der Haushalte H5, H10 und H26 kaum positive Korrelationen mit den anderen Haushalten aufweisen. Die Pumpe des Haushalts H13 ist sogar negativ mit den anderen Haushalten korreliert. Wird andererseits der 29.01.2019 betrachtet, bei dem die erste Hauptkomponente mit 15.3%, im Vergleich zu den anderen Tagen, den geringsten Anteil der gemeinsamen Varianz erklärt, fällt in Abbildung 5 auf, dass kein linearer Zusammenhang zwischen den Pumpen der Haushalte vorliegt.

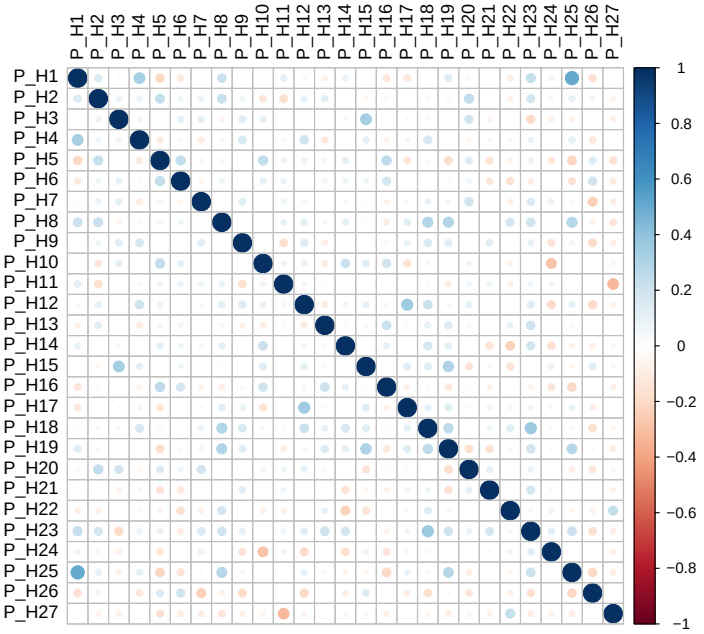

Abbildung 5: Korrelationsmatrix des Stromverbrauchs der Wärmepumpen vom 29.01.2019.

Ein wesentlicher Aspekt, ist, dass der Tag mit der besonders hohen Korrelation zwischen den Haushalten im Sommer liegt und der mit der besonders geringen Korrelation im Winter. Um zu entscheiden welche der Hauptkomponenten für die weitere Analyse berücksichtigt werden, wird der Scree-Plot in Abbildung 6 betrachtet.

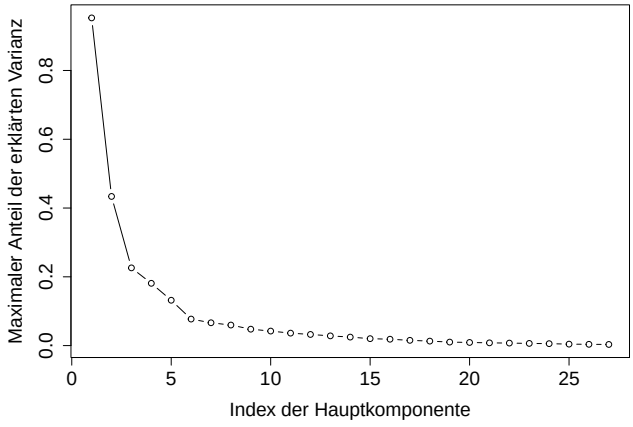

Abbildung 6: Scree-Diagramm des maximalen Anteils der erklärten Varianz der jeweiligen Hauptkomponente für das Jahr 2019.

Die Abbildung stellt für die jeweils i-te Komponente für $i = 1, ..., 27$ den maximalen Anteil der erklärten Varianz dar. Der in 3.4 beschriebene auffällige Knick ist nach der 3. Hauptkomponente zu erkennen, sodass sich für die ersten 3 entschieden wird.

4.3 Strukturbruchanalyse

In Abbildung 7 wird die resultierende Zeitreihe der Summe des Anteils der erklärten Varianz der jeweils drei ersten Hauptkomponenten eines Tages im Verlaufe des Jahres 2019 dargestellt. Auch die Durchschnittstemperatur des Tages wird aufgetragen, da diese, wie es in Abbildung 2 bereits deutlich geworden ist, einen starken Einfluss auf den Stromverbrauch der Wärmepumpen und somit vermutlich auch auf die Struktur des Zusammenhangs zwischen den Pumpen hat. Nebenbei werden die Zeitreihen für den erklärten Anteil der Varianz der drei ersten Hauptkomponenten im Anhang in den Abbildungen 12-14 zur Vollständigkeit auch nochmal einzeln dargelegt.

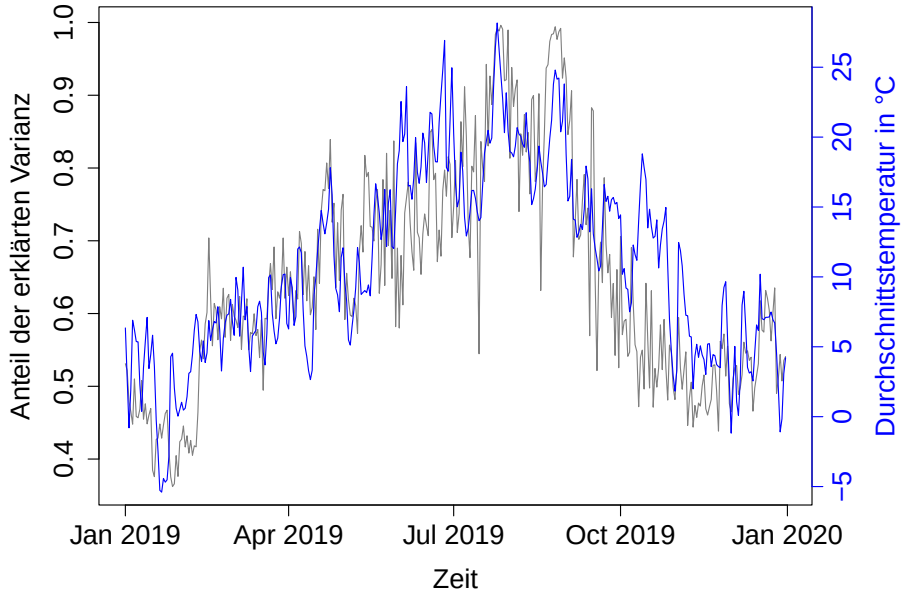

Abbildung 7: Zeitreihe des Anteils der erklärten Varianz der ersten drei Hauptkomponenten und der täglichen Durchschnittstemperatur.

Es ist deutlich zu erkennen, dass die Temperatur und der Anteil der erklärten Varianz sehr stark positiv zusammenhängen. Wenn die Temperatur steigt, dann steigt auch der Anteil der erklärten Varianz. Somit ist die erklärte Varianz im Sommer am höchsten und im Winter am niedrigsten. Insbesondere Ende Januar und Anfang Februar ist der Anteil der erklärten Varianz sehr gering, da die Temperaturen in diesem Zeitraum am niedrigsten waren. Dagegen werden Ende Juli und Ende August die Höchstwerte von bis zu 0.996 erreicht. Das heißt 99.6% der gesamten Varianz des 27.07.2019 kann über nur 3 Hauptkomponenten erklärt werden. Um nun eine Strukturbruchanalyse bezüglich des Mittelwerts auf dieser Zeitreihe durchzuführen, müssen die Daten zunächst standardisiert werden, da das genutzte Paket Changepoint für das Likelihood-Verfahren eine Varianz von 1 erwartet. Diese ist in

diesem Fall nicht gegeben, da die Werte nur zwischen 0.362 am 27.02.2019 und 0.996 am 27.07.2019 liegen. Somit erfolgt eine Standardisierung, indem der Mittelwert subtrahiert wird und durch die Standardabweichung geteilt wird. Von den standardisierten Werten wird angenommen, dass diese normalverteilt sind mit einem konstanten Erwartungswert in den Segmenten, die bestimmt werden sollen, und einer konstanten Varianz über die gesamte Zeit von 1. Weiterhin wird eine Unabhängigkeit der Daten angenommen. Diese Annahmen können, wie in 3.6 beschrieben, erst nach der Bestimmung der Strukturbrüche überprüft werden. Insbesondere die Unabhängigkeit der Daten ist sehr kritisch zu betrachten, da die Temperatur auf jeden Fall positiv autokorreliert ist und der Stromverbrauch der Pumpen sehr stark von der Außentemperatur abhängig ist und somit vermutlich auch positiv autokorreliert ist. Es wird der Binary-Segmentation-Algorithmus und der mBIC-Strafterm verwendet. Wie in Abbildung 8 erkennbar, wurden unter den gemachten Annahmen 3 Strukturbrüche ermittelt.

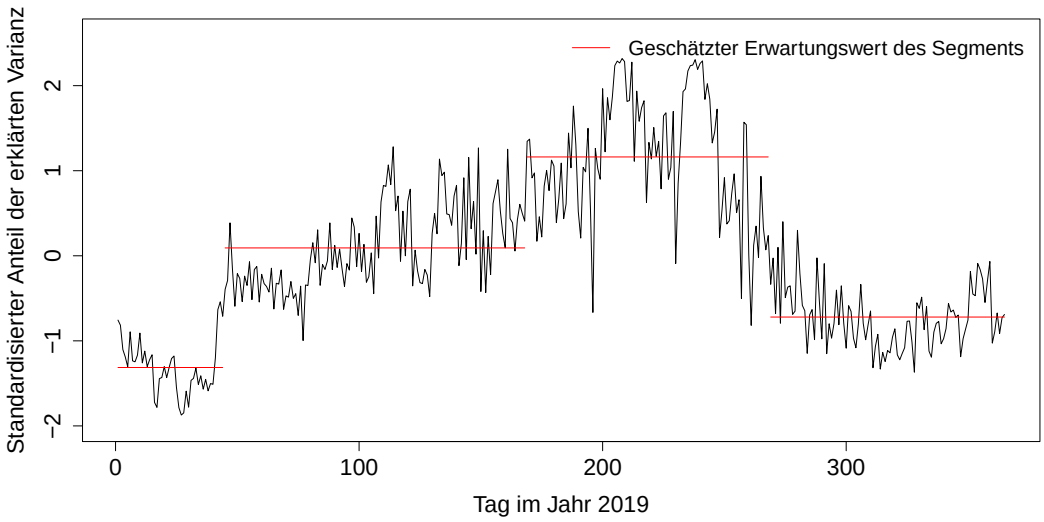

Abbildung 8: Strukturbruchanalyse des Mittelwerts.

Der erste Strukturbruch findet am 13.02.2019 statt, der Zweite am 17.06.2019 und der Letzte am 25.09.2019. Somit wird die Zeitreihe in vier Segmente zerlegt. Der zugehörige geschätzte Erwartungswert für diese Segmente ist ebenfalls mit einer roten Linie gekennzeichnet. Durch die Standardisierung ist die vertikale Achse natürlich nicht mehr interpretierbar. Da der geschätzte Erwartungswert des letzten Segments genau zwischen den geschätzten Erwartungswerten der beiden ersten Segmente liegt, besteht die Überlegung, ob eventuell nur 2 Strukturbrüche notwendig wären, die Anfang und Ende des Sommers kennzeichnen. Auf der anderen Seite macht der Strukturbruch Mitte Februar auch sehr viel Sinn, da der Anteil der erklärten Varianz Ende Januar und Anfang Februar vermutlich durch die extrem kalten Temperaturen sehr gering war.

Um die Annahmen des Modells zu prüfen, werden nun, wie in 3.6 beschrieben, die Residuen der Beobachtungen vom geschätzten Erwartungswert des jeweiligen Segments berechnet. Anschließend wird die Standardnormalverteilung dieser Residuen mittels des in Abbildung 9 dargestellten Normal-Quantil-Diagramms geprüft.

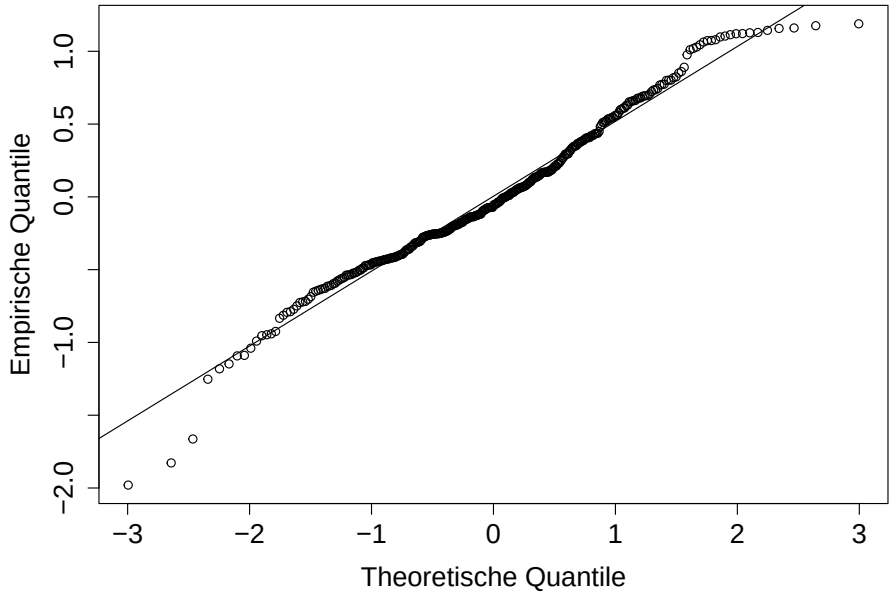

Abbildung 9: Normal-QQ-Plot der Residuen.

An den Verteilungsenden sind Abweichungen zu erkennen, wobei dies bei echten Daten sehr häufig auftritt. Da die Punkte jedoch zum größten Teil nahe oder auf der eingezeichneten Winkelhalbierenden liegen, wird die Standardnormalverteilungsannahme der Residuen nicht verworfen. Dagegen wird die Unabhängigkeitsannahme verworfen, denn im ACF-Plot in Abbildung 10 sind deutlich positive Autokorrelationen zu erkennen, die für $d = 1,...,9$ auch das blau-markierte 95%-Konfidenzintervall überschreiten.

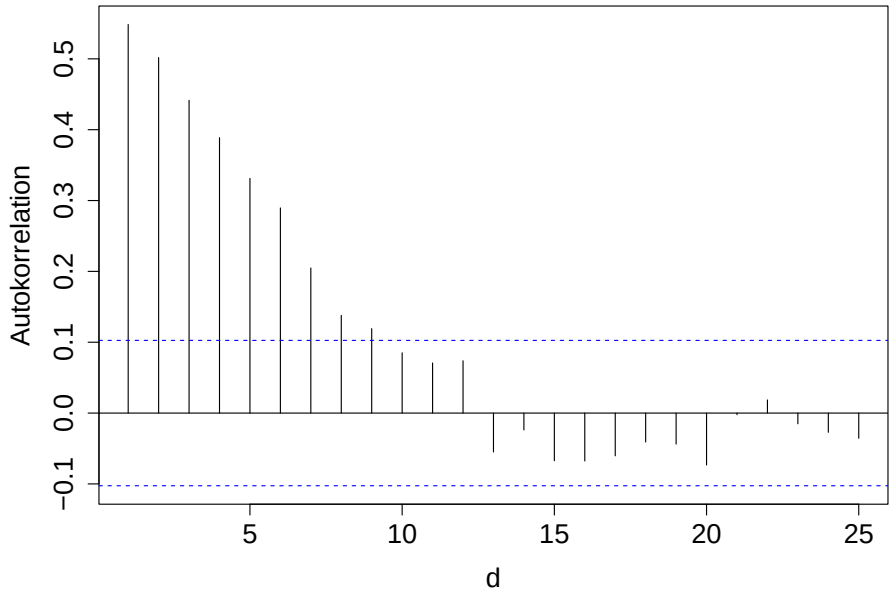

Abbildung 10: ACF-Plot der Residuen.

Dies hängt, wie vorhin bereits erwähnt, vermutlich mit der Autokorrelation des Wetters zusammen. Für die Resultate der Strukturbruchanalyse bedeutet dies, dass sie mit Vorsicht interpretiert werden sollten. Die Autokorrelation tritt sehr häufig bei echten Daten auf. Eine

Möglichkeit dem entgegenzuwirken ist es den Strafterm zu erhöhen oder vorab die Autokorrelation aus den Daten zu entfernen. Dies wird jedoch in dieser Arbeit nicht mehr getan, da die ermittelten Strukturbrüche sinnig waren.

Abschließend wird nun noch der große Unterschied zwischen Winter und Sommer interpretiert. Ein möglicher Grund für die extremen Unterschiede bezüglich des Zusammenhangs der Wärmepumpen im Winter und Sommer könnte sein, dass es durch die niedrige Temperatur im Winter viel mehr Faktoren gibt, die das Heizverhalten der Wärmepumpen beeinflussen. Im Sommer ist die Zieltemperatur im Haus aufgrund der Außentemperaturen sowieso meist ohne die Arbeit der Wärmepumpen erreicht, sodass sie hauptsächlich im Standby-Betrieb laufen können. Und selbst wenn sie mal heizen müssen, kann auch dies deutlich geregelter ablaufen, da nicht so ein großer Temperaturunterschied zwischen Ist- und Zieltemperatur ausgeglichen werden muss. Auch die Warmwasseraufbereitung kann im Sommer hauptsächlich durch die Solaranlage erfolgen. Dadurch sind die Voraussetzungen für ein ähnliches Arbeitsverhalten deutlich besser als im Winter, da hier viele Faktoren einen Einfluss haben, die nun vorgestellt werden. Zum Einen muss die Wärmepumpe die Warmwasseraufbereitung übernehmen und da nicht alle Menschen gleichzeitig duschen, stört dies bereits einen möglichen Zusammenhang zwischen den Pumpen. Weiterhin besitzen die Häuser nicht die gleichen objektiven Voraussetzungen, denn die Häuser unterscheiden sich hinsichtlich der Größe, der Dämmung und der Anzahl der Personen, die dort leben. Darüber hinaus kann nicht davon ausgegangen werden, dass in jedem Haus die gleiche Zieltemperatur eingestellt ist.

5 Zusammenfassung

Aufgrund des Klimawandels werden klassische Heizsysteme, die auf fossilen Brennstoffen basieren, in Deutschland insbesondere von Wärmepumpen abgelöst. Diese nutzen Umweltwärme, also thermische Energie aus der Luft, dem Boden oder dem Wasser. Für die Wärmeproduktion benötigen sie jedoch auch Strom, sodass diese Umstellung mit einer Mehrbelastung des Stromnetzes verbunden ist. In dieser Arbeit wurde der Zusammenhang des Stromverbrauchs vieler Wärmepumpen an einem Ort untersucht und analysiert wie sich dieser über das Jahr ändert. In diesem Kontext wurden auch Strukturbrüche ermittelt. Basis der Untersuchung ist ein Datensatz aus dem WPuQ-Projekt, der Stromverbrauchsdaten der Wasser-Wasser-Wärmepumpen von 33 Haushalten in der Nähe von Hameln enthält. Die Daten stammen aus dem Jahr 2019 und liegen in 15-Minuten-Intervallen vor, wobei die ursprüngliche Erhebung in 10-Sekunden-Intervallen erfolgte. Die Daten der 10-Sekunden-Intervalle wurden mithilfe des arithmetischen Mittels aggregiert. Auch die Außentemperatur des Ortes Hameln ist für jedes Intervall vorhanden. Da für 6 der 33 Haushalte jeweils mindestens ein Monat der Daten fehlt, wurden diese für die Analyse nicht weiter berücksichtigt. Die anderen 27 Haushalte sind weitestgehend vollständig, da nur vereinzelt 15-Minuten-Intervalle fehlen und für alle 27 dieselben betroffen sind.

Als Zusammenhangsmaß zwischen den Wärmepumpen fungiert die paarweise Kovarianz, die statt der Korrelation verwendet wurde, da alle Variablen in derselben Einheit Watt vorliegen. Um die Entwicklung des Zusammenhangs über die Zeit zu untersuchen, wurden diese paarweisen Kovarianzen für jeden Tag bestimmt und in Kovarianzmatrizen festgehalten. Da die Kovarianzmatrizen aufgrund der 27 betrachteten Haushalte eine Dimension von 27x27 besitzen, ist es nicht einfach möglich gewesen die enthaltenen Informationen in einer Zeitreihe darzustellen. Deshalb wurde für jeden Tag eine auf den Kovarianzmatrizen basierende Hauptkomponentenanalyse durchgeführt. Der resultierende Anteil der erklärten Varianz der drei größten täglichen Hauptkomponenten wurde über die Zeit dargestellt und im Hinblick auf Strukturbrüche untersucht. Die Entscheidung für drei Hauptkomponenten basiert auf der Betrachtung eines Scree-Plots. Die Idee dieser Analyse ist es, dass je größer der Anteil der erklärten Varianz der täglich größten Hauptkomponenten ist, desto größer ist auch der Zusammenhang zwischen dem Stromverbrauch der Wärmepumpen. Ansonsten sollte es nicht möglich sein einen Großteil der gemeinsamen Varianz über drei Linearkombinationen der ursprünglichen Variablen zu erklären. Auch die jeweils erste Hauptkomponente der Tage war von besonderem Interesse, da der zugehörige Eigenvektor die Punktewolke bestmöglich mittels einer geraden Linie abbildet.

Vor der eigentlichen Zusammenhangsanalyse wurde der Stromverbrauch der Wärmepumpe eines Haushalts betrachtet, um den Stromverbrauch beziehungsweise die Arbeitsweise von Wärmepumpen über das Jahr besser zu verstehen. Dabei ist aufgefallen, dass Wärmepumpen einen sehr großen Teil der Zeit im Standby-Betrieb verbringen, der sich durch einen stabilen niedrigen Verbrauch kennzeichnet. Sobald die Pumpe anfängt zu arbeiten, ist ein sehr großer Sprung im Stromverbrauch zu beobachten. Weiterhin ist der größte Stromverbrauch im Winter zu erkennen, da in diesem Zeitraum die Außentemperaturen am niedrigsten sind. Im Sommer ist der Verbrauch dagegen sehr gering, da die Zieltemperatur im Haus aufgrund der hohen Außentemperaturen meistens schon ohne viel Arbeit der Wärmepumpen erreicht wird. Eine weitere Auffälligkeit, die jedoch nicht bei allen 27 Pumpen zu beobachten ist, sind regelmäßige, aber sehr vereinzelt auftretende, extrem hohe Stromverbräuche über das gesamte Jahr. Eine mögliche Ursache könnte ein Desinfektionsverfahren sein bei dem die Wärmepumpe einmal durchheizt, um potenzielle Erreger im Heizsystem abzutöten.

Im Rahmen der Hauptkomponentenanalyse wurde die täglich erste Hauptkomponente nochmal gesondert betrachtet. Insbesondere wurde bestimmt an welchem Tag der Anteil der erklärten Varianz am höchsten und am niedrigsten ist. Die erste Hauptkomponente vom 28.08.2019 erklärt mit 95.29%, verglichen mit der ersten Hauptkomponente der anderen Tage, den größten Anteil der Varianz. Da der zugehörige Eigenvektor die Punktewolke bestmöglich mit einer geraden Linie beschreibt, wurde für diesen Tag die Korrelationsmatrix der 27 Haushalte betrachtet. Dabei sind größtenteils hohe positive Korrelationen bis zu 0.98 zwischen den Haushalten zu erkennen. Dagegen erklärt die erste Hauptkomponente vom 29.01.2019 mit 15.3%, verglichen mit der ersten Hauptkomponente der anderen Tage, den geringsten Anteil der Varianz. Dementsprechend sind in der Korrelationsmatrix dieses Ta-

ges auch keine Korrelationen zu erkennen. Von besonderer Relevanz ist die Tatsache, dass die hohen Zusammenhänge an einem Tag im Sommer festgestellt wurden und die nicht vorhandenen linearen Zusammenhänge an einem Tag im Winter.

Wird die Zeitreihe, die den Anteil der erklärten Varianz der täglichen ersten drei Hauptkomponenten darstellt, betrachtet, fällt auf, dass der erklärte Anteil sehr stark mit der Außentemperatur zusammenhängt. Wenn die Temperatur ansteigt, steigt auch der Anteil der erklärten Varianz der drei Hauptkomponenten an. Anhand dieser Zeitreihe wurde eine Strukturbruchanalyse des Mittelwerts durchgeführt. Dafür wurde das R-Paket Changepoint verwendet. Da das verwendete Likelihood-Verfahren in dem Paket annimmt, dass die Varianz über die gesamte Zeit 1 beträgt, wurde die Zeitreihe vorab normalisiert, indem der Mittelwert subtrahiert und durch die Standardabweichung geteilt wurde. Die weiteren Annahmen, welche besagen, dass die Daten über die Segmente hinweg bedingt unabhängig und innerhalb der Segmente normalverteilt, mit einem für das jeweilige Segment geltenden Erwartungswert, sein müssen, können erst nach Ermittlung der Strukturbrüche mittels der Residuen überprüft werden. Aus dieser Analyse resultierten die drei Strukturbrüche des Mittelwerts am 13.02.2019, 17.06.2019 und am 25.09.2019. Da der geschätzte Erwartungswert des vierten Segments zwischen den geschätzten Erwartungswerten der ersten beiden Segmente liegt, hätten eventuell 2 Strukturbrüche, die Anfang und Ende des Sommers markieren, mehr Sinn gemacht. Ursache für den Strukturbruch am 13.02.2019 könnte der besonders kalte Zeitraum zwischen Mitte Januar und Mitte Februar sein. In der anschließenden Prüfung der Annahmen wird die Standardnormalverteilung der Residuen nicht verworfen, jedoch wird die Unabhängigkeit der Daten verworfen, da der ACF-Plot der Residuen auf Abhängigkeiten hinweist, die vermutlich aus der positiven Autokorrelation der Temperatur resultieren, die sehr stark mit dem Anteil der erklärten Varianz der drei ersten Hauptkomponenten zusammenhängt. Da dadurch vorab gemachte Annahmen verletzt werden, sollten die Ergebnisse mit Vorsicht interpretiert werden.

Zusammenfassend lässt sich sagen, dass der Zusammenhang des Stromverbrauchs der Wärmepumpen sehr stark von der Außentemperatur abhängt. Insbesondere Winter und Sommer unterscheiden sich deshalb sehr stark. Ein möglicher Grund dafür ist, dass die objektiven Voraussetzungen der Häuser im Winter einen deutlich größeren Einfluss als im Sommer haben. Im Sommer muss aufgrund der hohen Außentemperatur, wenn überhaupt, nur ein geringer Temperaturunterschied zwischen der Ist- und Zieltemperatur ausgeglichen werden. Das heißt die Pumpen können deutlich geregelter heizen und sind meistens im Standby-Betrieb. Im Winter ist das Heizverhalten dagegen stark von der Dämmung, der Größe des Hauses und der eingestellten Zieltemperatur abhängig. Da die Häuser hinsichtlich dieser Aspekte Unterschiede aufweisen, unterscheidet sich das Heizverhalten im Winter auch stärker als im Sommer. Da die Daten ursprünglich in 10 Sekunden-Intervallen erhoben wurden, wäre es nachfolgend interessant zu untersuchen, ob ähnliche Ergebnisse mit kleineren Aggregationsintervallen von zum Beispiel 5 Minuten, 1 Minute oder sogar 10 Sekunden festgestellt werden können.

Literaturverzeichnis

- Eckley, I. A., Fearnhead, P. & Killick, R. (2011). Analysis of changepoint models. In Cambridge University Press eBooks, 205–224.
- Fahrmeir, L., Heumann, C., Künstler, R., Pigeot, I. & Tutz, G. (2016). Statistik: Der Weg zur Datenanalyse. Springer-Verlag.
- Hyndman, R. J. & Athanasopoulos, G. (2013). Forecasting: principles and practice. In OTexts.
- ISFH EnEff:Stadt Verbundvorhaben: Wind-Solar Wärmepumpen-Quartier – Erneuerbar betriebene Wärmepumpen zur Minimierung des Primärenergiebedarfs (o.D.). `https://isfh.de/wind-solar-waermepumpen-quartier/` (zuletzt abgerufen am 15.07.2024).
- Johnson, R. A. & Wichern, D. W. (2007). Applied Multivariate Statistical Analysis. In Prentice Hall.
- Killick, R., & Eckley, I. A. (2014). changepoint: An R package for changepoint analysis. Journal of Statistical Software, 58(3), 1–19.
- Killick R, Haynes K, Eckley IA (2022). _changepoint: An R package for changepoint analysis_. R package version 2.2.4, `https://CRAN.R-project.org/package=changepoint` (zuletzt abgerufen am 28.07.2024).
- Killick, R., Fearnhead, P. & Eckley, I. (2012). Optimal detection of changepoints with a linear computational cost. Journal of the American Statistical Association, 107(500), 1590–1598.
- Killick, R. (2021). NHS-R Workshop: Introduction to changepoint analysis with R- November 2021. `https://www.youtube.com/watch?v=UfGrLJ7S3sc&t=8523s` (zuletzt abgerufen am 15.07.2024).
- Kirchgässner, G., Wolters, J. & Hassler, U. (2013). Introduction to Modern Time Series Analysis. In Springer texts in business and economics.
- Kwak, J. H. & Hong, S. (2004). Linear Algebra. In Birkhäuser Boston eBooks.
- Mit Wärmepumpen Tempo machen für die Klimawende (2022). `https://www.bundesregierung.de/breg-de/aktuelles/kanzler-viessmann-2070096` (zuletzt abgerufen am 15.07.2024).
- Ohrdes, T., Schneider, E., Kastner, P., Knoop, M., Bast, O., Hagemeier, J., Darnauer, N., Klaas, A., Spielmann, V., Wehrmann, E., Beck, H. (2021) Abschlussbericht EnEff:Stadt Verbundvorhaben: Wind-Solar-Wärmepumpen-Quartier - Erneuerbar betriebene Wärmepumpen zur Minimierung des Primärenergiebedarfs (WPuQ). Institut für Solarenergieforschung GmbH.

- R Core Team (2023). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ (zuletzt abgerufen am 28.07.2024).
- Schlemminger, M., Ohrdes, T., Schneider, E. & Knoop, M. (2022). Dataset on electrical single-family house and heat pump load profiles in Germany. Scientific Data, 9, 56.
- So heizen die Deutschen (2019). https://www.bmwk-energiewende.de/EWD/Redaktion/Newsletter/2019/10/Meldung/direkt-erfasst_infografik.html (zuletzt abgerufen am 22.07.2024).
- Shrestha, N. (2021). Factor Analysis as a Tool for Survey Analysis. American Journal of Applied Mathematics and Statistics, 9(1), 4-11.
- Wärmepumpe: Legionellen keine Chance geben. (o. D.). https://www.wolf.eu/de-de/ratgeber/waermepumpe-legionellen#:~:text=Thermische%20Desinfektion%20durch%20Legionellenschaltung%3A%20W%C3%A4rmepumpen,automatisch%20laufen%20und%20Erreger%20abt%C3%B6ten (zuletzt abgerufen am 23.07.2024).
- Wei, T., & Simko, V. (2021). R package 'corrplot': Visualization of a Correlation Matrix (Version 0.92).
- Zhang, N. R. & Siegmund, D. O. (2007). A Modified Bayes Information Criterion with Applications to the Analysis of Comparative Genomic Hybridization Data. Biometrics, 63(1), 22-32.

Anhang

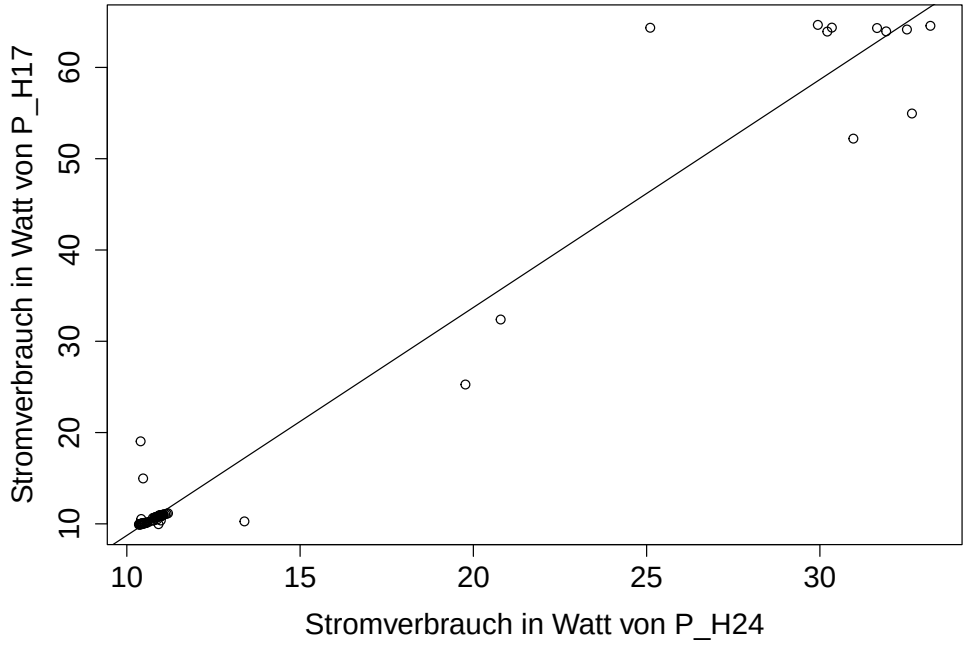

Abbildung 11: Streudiagramm und Ausgleichsgerade zwischen dem Stromverbrauch der Wärmepumpen der Haushalte 24 und 17 vom 28.08.2019.

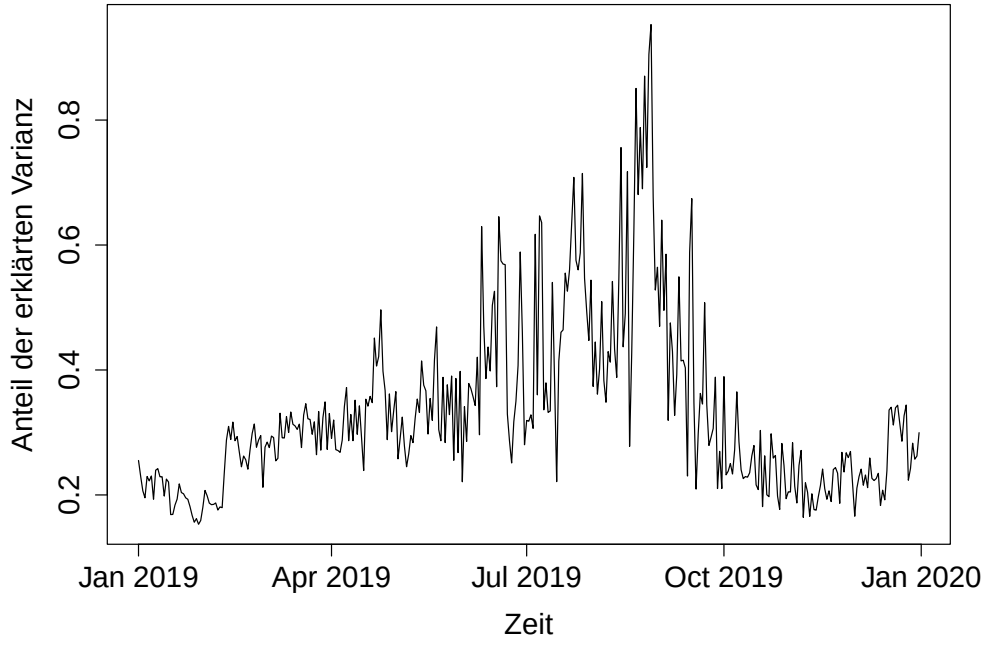

Abbildung 12: Zeitreihe des Anteils der erklärten Varianz der ersten Hauptkomponente.

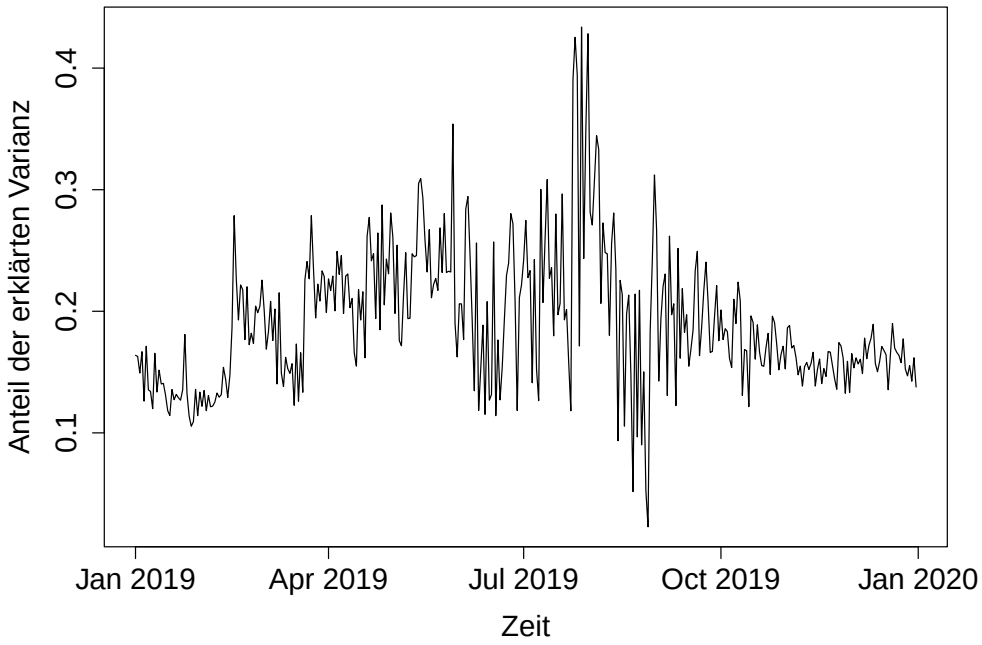

Abbildung 13: Zeitreihe des Anteils der erklärten Varianz der zweiten Hauptkomponente.

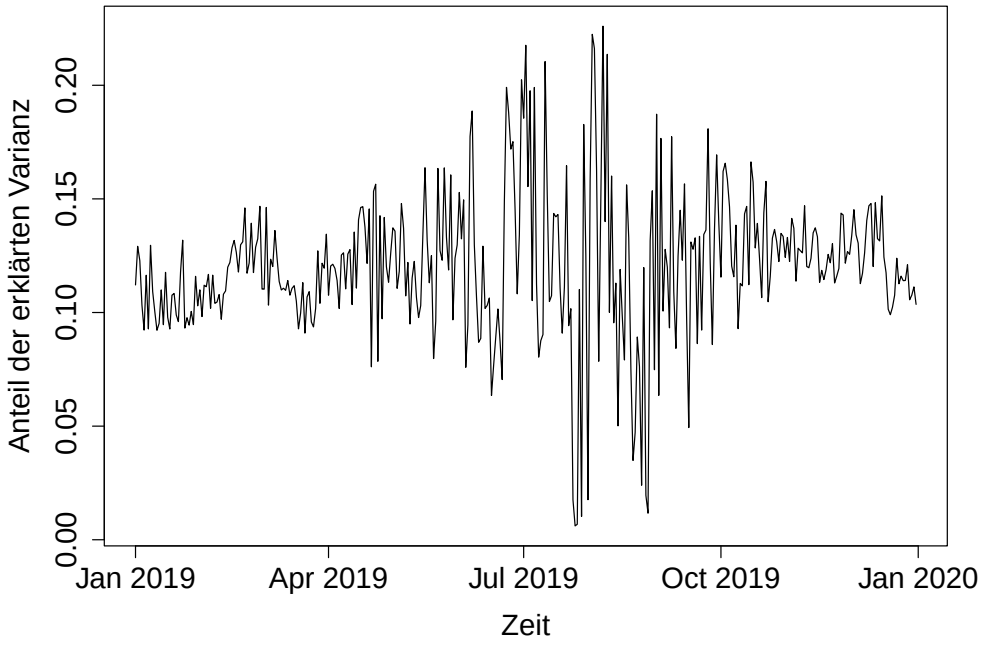

Abbildung 14: Zeitreihe des Anteils der erklärten Varianz der dritten Hauptkomponente.

Eidesstattliche Versicherung

(Affidavit)

Unrau, Kevin

Name, Vorname
(surname, first name)

210206

Matrikelnummer
(student ID number)

☒ Bachelorarbeit
(Bachelor's thesis)

☐ Masterarbeit
(Master's thesis)

Titel
(Title)

Strukturbruchanalyse von Zusammenhangsmaßen von Stromdaten

Ich versichere hiermit an Eides statt, dass ich die vorliegende Abschlussarbeit mit dem oben genannten Titel selbstständig und ohne unzulässige fremde Hilfe erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

I declare in lieu of oath that I have completed the present thesis with the above-mentioned title independently and without any unauthorized assistance. I have not used any other sources or aids than the ones listed and have documented quotations and paraphrases as such. The thesis in its current or similar version has not been submitted to an auditing institution before.

Bielefeld, 29.07.2024

Ort, Datum
(place, date)

Unrau

Unterschrift
(signature)

Belehrung:
Wer vorsätzlich gegen eine die Täuschung über Prüfungsleistungen betreffende Regelung einer Hochschulprüfungsordnung verstößt, handelt ordnungswidrig. Die Ordnungswidrigkeit kann mit einer Geldbuße von bis zu 50.000,00 € geahndet werden. Zuständige Verwaltungsbehörde für die Verfolgung und Ahndung von Ordnungswidrigkeiten ist der Kanzler/die Kanzlerin der Technischen Universität Dortmund. Im Falle eines mehrfachen oder sonstigen schwerwiegenden Täuschungsversuches kann der Prüfling zudem exmatrikuliert werden. (§ 63 Abs. 5 Hochschulgesetz - HG -).

Die Abgabe einer falschen Versicherung an Eides statt wird mit Freiheitsstrafe bis zu 3 Jahren oder mit Geldstrafe bestraft.

Die Technische Universität Dortmund wird ggf. elektronische Vergleichswerkzeuge (wie z.B. die Software „turnitin") zur Überprüfung von Ordnungswidrigkeiten in Prüfungsverfahren nutzen.

Die oben stehende Belehrung habe ich zur Kenntnis genommen:

Official notification:
Any person who intentionally breaches any regulation of university examination regulations relating to deception in examination performance is acting improperly. This offense can be punished with a fine of up to EUR 50,000.00. The competent administrative authority for the pursuit and prosecution of offenses of this type is the Chancellor of TU Dortmund University. In the case of multiple or other serious attempts at deception, the examinee can also be unenrolled, Section 63 (5) North Rhine-Westphalia Higher Education Act (*Hochschulgesetz, HG*).

The submission of a false affidavit will be punished with a prison sentence of up to three years or a fine.

As may be necessary, TU Dortmund University will make use of electronic plagiarism-prevention tools (e.g. the "turnitin" service) in order to monitor violations during the examination procedures.

I have taken note of the above official notification:*

Bielefeld, 29.07.2024

Ort, Datum
(place, date)

Unrau

Unterschrift
(signature)

***Please be aware that solely the German version of the affidavit ("Eidesstattliche Versicherung") for the Bachelor's/ Master's thesis is the official and legally binding version.**