Diese Arbeit wurde vorgelegt am
Institut für Elektrische Anlagen und Netze, Digitalisierung und Energiewirtschaft

Methoden zum Vergleich multipler Smart Meter Messreihen

Methods for the comparison of multiple smart meter measurement series

Bachelorarbeit

von

Herrn Chijun Zhang

1. Prüfer:	Univ.-Prof. Dr.-Ing. Albert Moser
2. Prüfer:	Univ.-Prof. Dr.-Ing. Dr. h. c. dr hab. Kay Hameyer
Betreuer:	M. Sc., M. Sc. Eng. Dominik Mildt
	Aachen, 27. Juli 2020

Kurzfassung

Um den von Menschen erzeugten Klimawandel zu verlangsamen werden die fossilen Energieträger mehr durch erneuerbare Energiequellen ersetzt. Dies führt zur Änderung des auf den fossilen Energiequellen basierenden konventionellen Stromversorgungssystems, dass die stromverbrauchsbasierende Versorgung durch die stromnachfragebasierende Versorgung ersetzt wird, weshalb sich die traditionelle Methode des Stromzählens ändern muss. Dafür ist der Einsatz der Smart Meter von großer Bedeutung.

Ein Smart Meter ist ein Stromzähler, der den Stromverbrauch in Echtzeit zu verschiedenen Zeitauflösungen misst. Die von den Smart Metern gemessenen Daten sind heterogen und addieren sich zu riesigen Datenmengen, weshalb der Umgang mit ihnen in den Bereich der Big Data Anwendungen fällt. Nach der Messung werden die Smart Meter Daten an die Stromversorger übermittelt. Die Stromversorger analysieren diese Daten weiter zur Beschreibung, Prognose und Entscheidung für ihre Energiebetriebe.

Trotz riesiger Datenmengen gibt es aber bestimmte Regelmäßigkeiten in den Smart Meter Messreihen. Ein wichtiger Aspekt bei deren Untersuchung ist dabei die Betrachtung der Ähnlichkeiten mehrerer Messreihen. Durch die Ähnlichkeitssuche in den mehreren Messreihen kann das Stromverbrauchsverhalten besser verstanden werden, und die Prognosen können verbessert werden. Damit können die Energievertriebe ihre Portfolios optimieren.

In Praxis und Forschung werden verschiedene Methoden und Algorithmen der statistischen Analyse und des Maschinellen Lernens eingesetzt. Wie die Methoden und Algorithmen richtig eingesetzt werden können, um die Ähnlichkeiten der Messreihen zu erarbeiten, steht im Fokus dieser Arbeit.

Abstract

In order to slow down the human-induced climate change, the fossil fuels are being replaced more by renewable energy sources. This leads to a change in the conventional power supply system based on fossil energy sources, that the power consumption based supply is replaced by the power demand based supply, therefore the traditional method of electricity metering has to change. The use of smart meters is of great importance for this change.

A smart meter is an electricity meter that measures electricity consumption in real time at various time resolutions. The data measured by smart meters are heterogeneous and add up to enormous amounts of data, which is why handling them falls within the scope of Big Data applications. After measurement the smart meter data are transmitted to the electricity suppliers. The electricity suppliers analyze these data further to describe, forecast and decide on their energy operations.

Despite huge amounts of data, there are certain regularities in the smart meter measurement series. An important aspect of their investigation is to observe the similarities of several measurement series. By searching for similarities in the several measurement series, the power consumption behavior can be better understood, and the forecast accuracy can be improved. This enables the energy distributors to optimize their portfolios.

In practice and research different methods and algorithms of statistical analysis and machine learning are applied. The focus of this thesis is how the methods and algorithms can be used correctly to work out the similarities of the measurement series.

Inhaltsverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis

DLC	Distribution Line Communication
EEG	Erneuerbare Energien Gesetz
GTO	Gate Turn-Off Thyristor
H	Hochhäuser
HH	Hochspannung Hochleistung
IFHT	Institut für Hochspannungstechnik
IGBT	Insulated-Gate Bipolar Transistor
kW	Kilowatt
MS	Mittelspannung
MW	Megawatt
NS	Niederspannung
OLTC	On Load Tap Changer
ONS	Ortsnetzstation
RWTH	Rheinisch-Westfälische Technische Hochschule
VDEW	Verband der Elektrizitätswirtschaft
VDN	Verband der Netzbetreiber
AMSS	Angular Metric for Shape Similarity
DTW	Dynamic Time Warping

1 Einleitung

1.1 Hintergrund und Motivation

Die Energiewende ist heutzutage ein Schlüsselthema für die nachhaltige Entwicklung sowohl der Technologie als auch der Gesellschaft. Ziel dieser Energiewende ist es, den Ausstoß des Treibhausgases CO_2 zu reduzieren. Schlüsselreaktion ist hierbei das Verbrennen der fossilen Energieträger. Hierbei entsteht durch die Reaktion von Kohlenstoff mit Sauerstoff das Treibhausgas CO_2. Dieses Gas sammelt sich in der Atmosphäre und beeinflusst das Klima. Zukunftsorientiert sind die weltweiten Reserven der fossilen Energieträger begrenzt. Um die der Erde schadenden Konsequenzen der Nutzung der fossilen Energien zu vermeiden, und die fossilen Energien mit einem alternativen und nachhaltigen Weg zu ersetzen, ist die Energiewende bzw. der Ausbau des Stromversorgungssystems aus den erneuerbaren Energien nötig.

Jedoch fallen hohe Kosten beim Erwerb der erneuerbaren Energien an [Ala16]. Darüber hinaus hängt die Stromversorgung aus den erneuerbaren Energien wie Solarenergie und Windenergie stark von den Wetterbedingungen ab, wobei sich das Versorgungsmodell „Versorgung folgt Verbrauch“ zu „Nachfrage an Versorgung anpassen“ ändern muss [Deu14]. Diese Änderung beim Ausbau der erneuerbaren Energien erfordert mehr Flexibilität und Stabilität im Versorgungssystem [Brü18]. Für diese Änderung muss der Netzbetrieb neu organisiert werden, damit die wetterbedingte Stromerzeugung aus erneuerbaren Energiequellen mit dem lokalen spezifischen Stromverbrauch harmonisiert werden kann [Deu14]. Die Digitalisierung trägt dazu bei, bzw. ist notwendige Vorrausetzung. Allerdings erfüllt sie sicher nicht allein alle Anforderungen.

Der Einsatz der Smart Meter kann dieser Änderung nicht fehlen, weil die Analyse der von den Smart Metern echtzeitig gemessenen Daten z.B. die auf den gemessenen Spannungen und Leistungsflüssen basierenden Versorgungszustandsschätzung den Stromversorgern ermöglicht, die Versorgung dementsprechend zu regulieren und zu optimieren, damit die Kundennachfragen und die Versorgung koordiniert werden können [Asg17].

1.2 Ziel und Aufbau der Arbeit

Das Ziel der vorliegenden Arbeit ist, mehrere Smart Meter Messreihen zu vergleichen und die Ähnlichkeiten zu identifizieren und zu analysieren.

Im zweiten Kapitel dieser Arbeit werden die Eigenschaften der Smart Meter Daten erklärt, die Quellen der Daten aufgezeigt und die allgemeinen Analysemethoden der Smart Meter Daten vorgestellt. Mit Zeitreihenanalyse und Clusteranalyse werden zuletzt in diesem Kapitel die Methoden der Ähnlichkeitssuche erklärt.

Im dritten Kapitel wird die exemplarische Untersuchung zur Ähnlichkeitssuche modelliert, wobei die Vorgehensweisen der ausgewählten Methoden jeweils aus Zeitreihenanalyse und Clusteranalyse für die Ähnlichkeitssuche in einem realen Smart Meter Datensatz erklärt werden.

Zum Schluss werden die Ergebnisse der exemplarischen Untersuchung im vierten Kapitel interpretiert, und im fünften Kapitel wird diese Arbeit zusammengefasst.

2 Theoretische Grundlagen

Dieses Kapitel beinhaltet den Überblick über die Daten, die Smart Meter bereitstellen, und Ansätze um diese zu analysieren. Anschließend werden die Zeitreihenanalyse und die Clusteranalyse im Detail betrachtet.

2.1 Smart Meter Data

Im Smart Metering werden die konventionellen Ferraris-Stromzähler durch die Smart Meter ersetzt. Bei einem Ferraris-Zähler leiten zwei Spulen den von einem Haushalt verbrauchten Strom, wodurch ein magnetisches Drehfeld aufgebaut wird. Dieses lässt eine Aluminiumscheibe rotieren. Die Aluminiumscheibe bringt einen mechanischen Zähler ins Rollen. Je mehr elektrische Energie dieser Haushalt verbraucht, umso stärker wird das magnetische Feld, wobei sich die Aluminiumscheibe und der Zähler schneller drehen. Der Wert auf dem Zähler steigt deshalb. Der Stromverbrauch in Kilowattstunden wird anhand der auf dem Zähler angezeigten Werte manuell abgelesen, indem die Differenz der Zählerwerte zweier bestimmten Zeitpunkte berechnet wird.

Ein Smart Meter basiert im Gegensatz nicht auf der elektromechanischen Funktionalität, sondern stellt ein elektrisches, mit einer Kommunikationseinheit verbundenes Messsystem dar. Ein Smart Meter misst den Stromverbrauch echtzeitig in einer Zeitauflösung von z.B. einer Minute, 15 Minuten, einer Stunde usw. Die Messdaten werden durch die Advanced Metering Infrastructure, die aus Smart Metern, Datenverwaltungssystem und bidirektionalem Kommunikationsnetzwerk zwischen den Verbrauchern und den Versorgern besteht [Dud18], [Sul19], weiter an die Stromversorger periodisch übermittelt [Mar13], damit die Messdaten sowohl von den Stromverbrauchern als auch von den Stromversorgern jederzeit zugreifbar sind.

Mit den Smart Meter Messdaten können nicht nur die Stromverbraucher über ihren Stromverbrauch besser informiert werden [Bar15], sondern auch die Stromerzeuger können dadurch die Netzanlagen besser verwalten, indem sie die Messdaten zur Analyse bringen. Anhand der Analyse dieser Messdaten können die Stromversorger z.B. die exakten weiteren Versorgungsplanungen erstellen, die Ausfälle schnell

erkennen, die Ursachen für die Ausfälle feststellen und die potenziellen Ausfälle verringern [Bar15].

2.1.1 Eigenschaften der Smart Meter Daten

Eine Datenmenge wird als Big Data bezeichnet, wenn sie über die vier Hauptmerkmale verfügt: hohes Volumen, hohe Schnelligkeit, hohe Vielfalt und geringe Informationsdichte [Ala16]. Die von Smart Metern erfassten Messdaten beziehen sich eindeutig auf diese vier Merkmale, weshalb die Smart Meter Daten für Big Data gehalten werden.

Im Folgenden werden diese vier Hauptmerkmale der Smart Meter Daten erklärt:

Volumen

Die Menge aller gemessenen Daten bestimmt das Volumen der Datenmenge, wobei die Smart Meter Daten leicht ein enormes Datenvolumen liefern [Mar19]. Nach dem Wechsel von den Ferraris-Zählern mit den monatlichen manuellen Ablesungen zu den Smart Metern mit den intelligenten Ablesungen in einer 15-minütigen oder 1-stündigen Zeitauflösung wird ein massives Datenvolumen erzeugt, welches den Stromversorgern Schwierigkeiten über die Verwaltung bietet [Paw17]. Das Datenvolumen der Smart Meter ist wegen der hohen Anzahl der Stromverbraucher riesig [Li16]. Z.B. generiert Smart Meter mit der Abtastrate 1 Datenprobe pro Sekunde und 128 Bytes pro Datenprobe in einem Gebiet mit 500.000 Stromverbrauchern 92 Terabytes Messdaten jeden Tag [Mar13]. Dies stellt die Stromversorger durch die Speicherung und die Verarbeitung solcher riesigen Datenmenge vor große Herausforderungen.

Für die enormen Mengen der Smart Meter Daten können verschiedene Speicherplätze zur Verfügung gestellt werden, die durch Software miteinander verbunden sind, sodass auf die Daten in einem Speicher leicht von einem anderen zugegriffen werden kann [Mar19]. In [Xie14], [Zho16] bietet sich die verteilte und skalierbare Rechnerarchitektur als eine Lösung für die Verwaltung der großen Datenmenge der Daten, die die Synchrophasoren bereitstellen, an [Bha19]. Die Reduktion der Datendimensionalität wird in [Zom12] ausführlich betrachtet [Bha19]. Diese Methode verringert zwar die Datenkomplexität und das Datenvolumen, bewahrt aber die Datenintegrität der originalen Daten auf [Bha19].

Schnelligkeit

Die Schnelligkeit ist der wesentlichste Unterschied zwischen Big Data und konventioneller Datengewinnung [Liu17b]. Einerseits bezieht sie sich darauf, wie schnell und rechtzeitig die Big Data generiert, erfasst, verarbeitet und genutzt werden [Ala16]. Andererseits ist sie zu verstehen, dass die Anforderungen der Big Data an die echtzeitigen Messungen viel schneller als die Anforderungen der traditionellen Datengewinnung an die echtzeitigen Messungen erfüllt werden [Liu17b], wobei die jährliche Zunahme der echtzeitig und kontinuierlich erzeugten Daten mehr als 60% beträgt [Yan14]. Beispielweise werden am Jahresende 2920 Terabytes Messdaten von einer Million Smart Meter generiert, wenn über das ganze Jahr jeder Smart Meter alle 15 Minuten nur einen Messwert zum Stromversorger sendet, wobei die Speicherung und die Übermittlung einer solchen Datenmenge ohne weiteren fortschrittlichen Technologien problematisch wären [Mar19]. Obwohl es schon analytische Algorithmen für die riesigen Datenmengen gibt, sind viele dieser Algorithmen nicht in der Lage, solche Analysen echtzeitig oder in einer ausreichend kurzen Zeitspanne durchzuführen [Ala16], [Paw17]. Beispielweise reicht eine einmalige Berechnung über Nacht für die echtzeitigen Analysenaufgaben wie die Zuverlässigkeitsüberwachung der Stromversorgungskomponenten, die Verhinderung der Stromausfälle oder die Sicherheitsüberwachung nicht aus [Ala16]. Wenngleich sich schon mehrere Techniken zur echtzeitigen Datenanalyse in Forschung und Entwicklung befinden, muss daran noch gearbeitet werden, bis diese Techniken für kommerzielle Nutzung angewendet werden können [Ala16].

Vielfalt

Vielfalt bedeutet im Kontext des Big Data zunehmend diverse Datentypen und große Auswahl an Datenquellen [Liu17b], [Mar19]. In der konventionellen Datengewinnung sind die Daten meistens strukturiert, hingegen können die Big Data, die die Smart Meter bereitstellen, strukturiert, semistrukturiert oder unstrukturiert sein [Yan14]. Die für die Datenanalyse im Smart Metering relevanten Daten stammen nicht nur aus den klassischen Stromdatenquellen wie Stromerzeugung und -verbrauch, die direkt von den Smart Metern gemessen werden, sondern auch aus Sicherheitskameras, Wettervorhersagesystemen, geographischen Karten, Bildern sowie Internet [Ala16]. Im Folgenden dieser Arbeit werden all diese Daten „Smart Meter Daten“ genannt. Ein Energieversorger kann beispielweise die auf seine Energiedienstleistungen bezogenen

sozialen Medien und die Dialoge in seinem Callcenter analysieren, und dann die Informationen daraus als ein Teil seiner Entscheidungs- und Planungsverfahren in die Smart Meter Daten integrieren [Ala13]. Wegen der Multidatentypen der Smart Meter Daten sind die Anforderungen an die Verarbeitungsfähigkeit der Smart Meter Daten viel höher als die Verarbeitungsfähigkeit der Daten aus traditioneller Datengewinnung [Liu17b].

Informationsdichte

Die Informationsdichte einer Datenmenge bezeichnet das Verhältnis von nützlichen Informationen zum Datenvolumen. Die Big Data bzw. die Smart Meter Daten verfügen über eine niedrige Informationsdichte. Mithilfe der breiten Nutzung des Internet of Things in Kommunikationstechnik vom Smart Metering sind die Smart Meter Daten überall zu empfangen [Liu17b], wobei die Größe des Datensatzes zunimmt, aber die Menge der für die Datenanalyse nützlichen Daten und Informationen abnimmt [Mar19]. Z.B. enthält ein Video aus einem Sicherheitskamera nur 1 bis 2 Sekunden nützliches Material für eine ununterbrochene Überwachung [Yan14]. Um die nützlichen Informationen vom ganzen Datensatz zu differenzieren, können erstmal die Maschinellen Algorithmen zur Extraktion nützlicher Daten eingesetzt werden [Liu17b]. Im Anschluss besteht der Schritt darin, zu verstehen, wofür diese extrahierten Daten nützlich und relevant sind [Mar19]. Außerdem können die Messgenauigkeiten der Smart Meter und der anderen auf die Smart Meter Daten bezogenen intelligenten Messgeräte durch die Aufrüstung der Messtechniken verbessert werden, um den Anteil der relevanten Informationen zu erhöhen [Mar19].

Nach der Perspektive der klassischen Merkmale vom Big Data werden die Eigenschaften der Smart Meter Daten im Folgenden aus der Sicht der in der Smart Meter Datengewinnung auftretenden Probleme und Herausforderungen erklärt. Sie werden in [Bha19] in die folgenden vier Kategorien unterteilt.

Ungewissheit

Die Ungewissheit der Smart Meter Daten ist auf die instabile Datenqualität zurückzuführen, die von Genauigkeit, Vollständigkeit und Konsistenz der erfassten Daten abhängt [Bha19]. Die Hauptgründe für falsche Daten sind Messrauschen,

Cyberattacken und Kommunikationslatenzzeiten [Bha19]. Darüber hinaus gibt es noch andere Ursachen für die Ungewissheit der Smart Meter Daten. Ein Beispiel in [Bha19] erläutert, wenn die Sensoren von einem Smart Meter altern, oder von Hackern aus böswilligen Zwecken angegriffen werden, gehen diese Messdaten der Sensoren während der Datenerfassung teilweise verloren. Aus der Sicht der Big Data bieten sich beispielweise zwei Bewertungsmethoden der Datenqualität an. In [Sub15] wird ein Framework auf Basis des Entscheidungsbaums und des mehrdimensionalen Modells zur Auswertung der Datenqualität des Big Data entworfen, das sich mit den Dimensionen des Big Data befasst [Liu17a]. Ein skalierbarer Bewertungsansatz der Datenqualität wird in [Klä16] detailliert vorgestellt, wobei der erste Prototyp zur Untersuchung der Skalierbarkeit in einer Testumgebung mit Multiknoten bereits realisiert wird [Liu17a]. Nach der Auswertung der Datenqualität können die Maßnahmen der Reduzierung der Datenungewissheit ergriffen werden. In [Tsa09] werden die probabilistische Datenanalyse und -gewinnung, bei denen die Datenungewissheit als stochastisches Verfahren modelliert wird, zur Verringerung der Datenungewissheiten eingesetzt [Bha19]. Die Datenvorverarbeitung z.B. Datenbereinigung wird in [Wag12] angewendet, um die Ausreißer zu erkennen, die Inkonsistenzen zu korrigieren, und das Messrauschen zu identifizieren und zu glätten [Bha19].

Sicherheit

Die Smart Meter Daten enthalten die sicherheitsrelevanten und vom Datenschutze betroffenen Informationen der Stromversorger und -verbraucher, weshalb die Datensicherheit ein nicht vernachlässigbares Thema ist. Die Datensicherheit beschäftigt sich dabei hauptsächlich mit drei Unterpunkten, und zwar Datenschutz, Datenintegrität und Datenauthentifizierung [Bha19].

Der Datenschutz wird heutzutage in der Verwendung vom Big Data bedroht, weil z.B. die Vorlieben der Nutzer von den IT-Firmen mithilfe der Analyse der riesigen Datenmengen über das Nutzerverhalten genau bestimmt werden können [Liu17b]. Analog können die Stromverbrauchsdaten, die durch die Smart Meter erfasst werden, die Einsichten in die privaten Informationen der einzelnen Stromverbraucher bzw. Haushalte ermöglichen, wobei die Belegungssituationen und der wirtschaftliche Status dieser Haushalte verraten werden können [Bha19], [Asg17]. Zum Umgang mit solchen Problemen ist die Datenaggregation einer der üblichen Ansätze, wobei die verteilte

Datenaggregation und die differenzielle Datenaggregation aktuell weiterentwickelt werden, um die Datenschutzprobleme zu beseitigen [Bha19].

Die Datenintegrität wird benutzt, um unbefugte Änderungen an den Daten zu verhindern, die auf die physikalischen Angriffe an die Hardware der Smart Meter oder die Cyberattacken in das Datensystem zurückzuführen sind [Bha19]. Beispielweise können die Stromverbraucher nach einer erfolgreichen Cyberattacke ihre persönlichen Daten an dritte verlieren, und riesige Rechnungen erhalten, wobei ihre Datenintegrität verletzt wird [Mar19]. Das Modell der privatsphärenbewahrenden Datenaggregation in [He17] kann die Datenintegrität durch eine digitale Unterschrift oder einen Authentifizierungscode mit Nachrichten sicherstellen [Bha19].

Die Datenauthentifizierung, die sowohl für den Datenschutz als auch für die Datenintegrität notwendig ist, wird von den Smart Meter Daten für die Unterscheidung der rechtmäßigen und unrechtmäßigen Datenzugreifer gebraucht, wofür die Verschlüsselung, das Vertrauensmanagement und die Einbruchserkennung die wichtigen Sicherheitsmechanismen sind [Bha19]. In [Qi16] werden die Datenverschlüsselung und die Unterschriftserstellung zur Datenauthentifizierung im Detail betrachtet [Bha19].

Zeitsynchronisation

Die Zeitsynchronisation der Smart Meter Daten spielt eine entscheidende Rolle bei der echtzeitigen Datenmessung und -übermittlung. Die zeitsynchronisierten Datengenerierung, -kommunikation und -analyse können die Datenanalysten bzw. die Stromversorger in die Lage versetzen, die aussagekräftigen Zusammenhänge zwischen den Ereignissen hinter den Verbrauchs- und Erzeugungsdaten zu identifizieren, was beim echtzeitigen Informieren der gegenwärtigen Situationen und beim Entscheiden über zukünftige Betriebsplanungen bei der forensischen Analyse der vergangenen Ereignisse hilft [Bha19], [Sci12]. Dass die Datenübermittlung, -speicherung und -analyse der Smart Meter Datenströme im Smart Metering aktuell jedoch noch nicht ganz zeitsynchronisiert sind, bringt ein potenzielles Risiko der irreführenden Entscheidungen der Stromversorger, weshalb die Datenströme im Smart Metering zeitsynchronisiert werden sollen [Bha19].

Datenindexierung

Die Datenindexierung ist von Bedeutung für die ordentliche Darstellung der Smart Meter Daten mit riesigem Datenvolumen und die Anfragebearbeitung der bestimmten Datenmengen im ganzen Datenvolumen zur weiteren Datenanalyse. Für die Anfragebearbeitung werden heutzutage z.B. die SQL-Server (Structured Query Language Server) und die gegen die Anfragebearbeitungen in Datensätzen gerichteten Softwareprodukte vom SAP (Systeme, Anwendungen und Produkte in der Datenverarbeitung) eingesetzt [Bha19]. Diese reichen für die echtzeitigen Datenströme in Big Data bzw. Smart Meter Daten aber noch nicht aus [Bha19]. In [Tay98], [Kot02], [Kam93] werden die modernen Datenindexierungstechniken einschließlich R-Bäume, B-Bäume, und Quadbäume, die die Smart Meter Daten effizient indexieren, ausführlich erklärt [Bha19].

2.1.2 Smart Meter Datentypen

Wie im Absatz „Vielfalt“ des Kapitels 2.1.1 Eigenschaften der Smart Meter Daten angekündigt steht in dieser Arbeit der Begriff „Smart Meter Daten“ nicht nur für die gemessenen Daten, sondern auch für die Daten, die aus den Smart Metering bezogenen Informationen stammen. Im Folgenden werden die üblichen Smart Meter Datentypen charakterisiert.

Messdaten

Die Messdaten, die in regelmäßigen Zeitintervallen erfasst und übertragen werden, beziehen sich auf den tatsächlichen Stromverbrauch, der in Kilowattstunden kWh gemessen wird [Ala13]. Neben dem Stromverbrauch können die Spannung, der Strom, die Frequenz, die Wirkleistung, die Blindleistung und der Ein/Aus-Status des elektrischen Betriebs erfasst werden [Ras18]. Abhängig von dem einzelnen Einsatz der Smart Meter und der Region kann das Zeitintervall der Messung von 1 Minute bis 1 Stunde variieren, wobei z.B. eine Leistungsmessung als 15-Minuten-Mittelwert für die Abrechnung übertragen werden kann, und die 1-minütigen Leistungsmesswerte für eine Echtzeit-Zustandserfassung verwendet werden können.

Die Hauptnutzung der Messdaten geht um die genaue Stromverbrauchsabrechnung bei dynamischer Preisgestaltung [Asg17]. Die Messwerte zu Beginn und am Ende eines Abrechnungsintervalls können den Stromverbrauch dieses Zeitintervalls

berechnen, um den diesem Verbrauch entsprechenden Abrechnungspreis zu bestimmen, wobei ein typisches Abrechnungsintervall ein Monat ist [Ala16]. Außerdem können die Messdaten z.B. in einer 15-minütigen Zeitauflösung auf die Daten in einer monatlichen Zeitauflösung statistisch aggregiert werden, um einen Überblick über den monatlichen Stromverbrauch zu erschaffen. Dies kann die Vergleichsanalyse aller Haushalte in einer Gemeinde und die Vergleichsanalyse mit der Nutzungsgeschichte unterstützen [Ala16].

Die Messdaten verfügen über hohes Datenvolumen und hohe Datenschnelligkeit wie die typischen Merkmale vom Big Data, aber wegen der hauptsächlich auf den Stromverbrauch bezogenen Messung sind die Messdaten homogener als Big Data, d.h. die Datenvielfalt der Messdaten ist relativ gering [Ala13].

Eventdaten

Die Eventdaten sind die Informationen über die in Stromerzeugung, -versorgung, -verbrauch, und -messung erfassten Ereignisse, die z.B. den echtzeitigen Status der Smart Meter und der elektrischen Geräten, die Stromausfälle, die Stromqualitätsinformationen usw. umfassen [Paw17]. Diese Ereignisse können sich aus den Attributen wie Quelle und Stellvertreter, Schweregrad und Kategorie zusammensetzen, wobei die Quelle der Sprung dieser Ereignisse ist, und der Stellvertreter diese Ereignisse erfasst und an die berechtigten Parteien übermittelt [Ala16]. Unter den typischen Ereignissen werden die Stromqualitätsinformationen z.B. Leistungsfaktor und Spannungsstabilität allgemein bei der Fehleranalyse, die sowohl die Vorfehler als auch die Nachfehler analysiert, benutzt, um die Zuverlässigkeit des Verteilnetzes zu verbessern [Ala16]. Darüber hinaus ist der Nutzungsgrad der erneuerbaren Energiequellen auch ein wichtiges Ereignis, durch den die Effektivität der Stromerzeugung aus erneuerbaren Energien identifiziert werden kann [Ala16].

Wie Big Data weisen die Eventdaten eine hohe Datenschnelligkeit auf, und im Vergleich zu den Messdaten können die Eventdaten aus diversen Quellen stammen, bzw. aus diversen Stellvertretern empfangen werden, weshalb die Eventdaten über eine hohe Datenvielfalt verfügen [Ala13].

Hilfsinformationen

Die Hilfsinformationen beziehen sich auf die Informationen z.B. die Wetterdaten, die geographischen Daten, die Energiemarktdaten, die Kundendaten usw [Ala13]. In [Che16] wird exemplarisch untersucht, wie die Standorte der Stromverbraucher durch die Smart Meter Daten, die die Details der Zeitstempel der Solarstromerzeugung enthalten, bestimmt werden [Sul19]. In [Sod16] werden die Kundenumfragedaten, die Wetterdaten, die sozialdemographischen Daten und die geographischen Daten der Energiekunden der Arbon Energie AG in der Schweiz im Detail demonstriert und betrachtet. Die Hilfsinformationen können mit den Messdaten und den Eventdaten der Smart Meter zur Datenanalyse integriert werden, um beispielweise die Abhängigkeiten des Stromverbrauchs von Wetter, Orten usw. zu untersuchen.

2.1.3 Smart Meter Datenanalyse

Von der Smart Meter Datenanalyse profitieren die Verteilnetzbetreiber, Stromversorger, und Kunden. Die Verteilnetzbetreiber können durch die Smart Meter Datenanalyse ihren Verteilnetzzustand besser bewerten, um das Verteilnetz besser zu verwalten [Ala16]. Die Stromversorger können mithilfe der Analyse die versteckten und potenziell nützlichen Informationen in der Stromlast innerhalb von riesigen Datensätzen identifizieren, die weiter der Versorgungsplanung und dem Versorgungsmanagement dienen [Bha19]. Die Kunden profitieren besonders von der nach der Smart Meter Datenanalyse erstellten Preisintelligenz bzw. den dynamischen Tarifstrukturen, wobei die Kunden z.B. nach den Tarifen zu den Spitzenlastzeiten und Grundlastzeiten ihre Stromverbrauchsplanungen für die niedrigeren Stromgebühren anpassen können [Paw17].

Bei der Smart Meter Datenanalyse werden verschiedene Algorithmen und Verfahren verwendet, wobei Maschinelles Lernen, Korrelation, Klassifizierung, Regression usw. häufig eingesetzt werden. Je nach dem Analysenziel und dem Anwendungsfall wird die Smart Meter Datenanalyse in die folgenden Analysenmethoden gegliedert: deskriptive, prädiktive und präskriptive Analyse [Bha19], [Dud18].

Deskriptive Analyse

Die deskriptive Analyse wird zur Beschreibung des Stromverbrauchsverhaltens bzw. der Lastprofile verwendet. Die deskriptive Analyse findet die spezifischen

Eigenschaften der Lastprofile mehrerer Haushalte raus, wobei z.B. die Eigenschaften, die den sozioökonomischen Status der einzelnen Haushalte, die Wohnimmobilien und den Bestand der elektrischen Geräte in einem Haushalt erfassen, identifiziert werden können [Dud18]. Einer der üblichen Zwecke der deskriptiven Analyse ist die Beschreibung des täglichen Stromverbrauchs, die in den Lastprofilkurven des täglichen Stromverbrauchs erfolgt. Die Lastprofilkurve des Stromverbrauchs von einem Tag gibt die Informationen über die Grundlast, die Mittellast und die Spitzenlast an diesem Tag [Paw17], [Yu16]. Werden die Lastprofilkurven des Stromverbrauchs von vielen Tagen eines Haushaltes in einem Graphen geplottet, ist es zu bemerken, in welchem Zeitraum der Strom am meisten und am geringsten verbraucht wird, wovon der Stromverbrauch dieses Haushaltes zur Stromgebührenverringerung oder zur Balancierung des täglichen Stromverbrauchs zeitlich angepasst werden kann [Paw17].

Die Stromversorger können ihre Kunden beispielweise durch das Clustern oder die Dimensionsreduktion der Tageslastprofile ihrer Kunden klassifizieren. Dadurch können die Stromversorger mit den Kunden eines gleichen Stromverbrauchsmusters spezifische Verträge abschließen, was zur Marktsegmentierung führt [Mar19], [Wan20]. Die Folge sind bessere Marketing-Strategien im Energiemarkt [Mar19].

Prädiktive Analyse

Die prädiktive Analyse beschäftigt sich mit den Vorhersagen der Betriebszustände und der zukünftigen Entscheidungen [Bha19]. Der Schwertpunkt der prädiktiven Analyse liegt daran, eine riesige Anzahl der historischen Smart Meter Daten zu sammeln, und darauf basierend ein wissenschaftliches und effektives Prognosemodell zu erstellen, wofür durch die effektiven Algorithmen des Deep Learnings eine große Anzahl der exemplarischen Untersuchungen auf den historischen Daten basierend durchgeführt wird [Yan14]. Die daraus resultierenden Erfahrungen werden zusammengefasst und ständig korrigiert. Dieses so erstellte Prognosemodell kann die zukünftigen Lastzustände bzw. die zukünftigen Lastschwankungen gut widerspiegeln [Yan14].

Die erfolgreichen und genauen Ergebnisse der Prognosen des zukünftigen Stromverbrauchs können sich auf den Energiemarkt positiv auswirken, weshalb die genauen Prognosen eine wichtige Rolle für die Stromversorger bei ihren Entscheidungen über den kurzfristigen Betrieb und die mittelfristige Planung spielen [Mar19], [Ala13]. Für eine langfristige Planung reicht die prädiktive Analyse nicht aus [Mar19], [Ala13].

Bei der kurzfristigen prädiktiven Analyse tritt ein Problem auf, dass das Verständnis des Gesamtbildes des Energiebetriebs häufig verloren geht [Ala13]. Für die kurzfristige Planung der Stromversorgung wird die richtige Prognose durch die prädiktive Analyse leicht getroffen. Daraus wird das Ziel der kurzfristigen Planung erfüllt. Ob ergriffene Maßnahmen durch die kurzfristige Prognose rentabel für die langfristige Planung der Stromversorgung sind, kann nur mit der Zeit überprüft werden.

Präskriptive Analyse

Die präskriptive Analyse befasst sich mit der Frage, wie die Smart Meter Datenanalyse zum Lastmanagement beiträgt. Durch die präskriptive Analyse werden den Stromversorgern die Erkenntnisse über die strategische Betriebsverwaltung und die Investitionsplanung langfristig geliefert [Bha19]. In [Wan20] wird diese Analyse in drei Aspekten erläutert. Im ersten Aspekt werden durch die präskriptive Analyse der Smart Meter Daten die soziodemographischen Informationen der Kunden erworben [Wan20]. Diese Informationen werden für das bessere Verständnis der Kunden weiter bearbeitet, womit den Kunden die besseren und die personalisierten Dienstleistungen angeboten werden können [Wan20]. Im zweiten Aspekt dient die präskriptive Analyse dazu, die potenziellen Kunden mithilfe des Marketings mit Demand Response Program ausfindig zu machen und anzusprechen [Wan20]. Im dritten Aspekt werden die auf die Implementierung des Demand Response Program bezogenen Themen erläutert, unter denen die Preisgestaltung des preisbasierenden Demand Response Program erklärt wird [Wan20].

2.2 Zeitreihenanalyse

Als ein Teilbereich der Statistik wird die Zeitreihenanalyse in verschiedensten Gebieten inklusive der Analyse der Smart Meter Messreihen angewendet [Lei98]. Eine Zeitreihe ist eine zeitliche Abfolge, welcher Messpunkte zugeordnet werden [Dei18]. In Smart Meter werden der Strom, die Spannung, die Frequenz, die Wirkleistung und unter bestimmten Bedingungen die Blindleistung elektrischer Geräte zu regelmäßigen Zeitpunkten gemessen. In Kombination mit diesen Zeitpunkten kann eine zeitlich indexierte Abfolge der Messdaten gebildet werden.

Um sinnvolle Aussagen über die zeitabhängigen Charakteristika der Zeitreihe treffen zu können, beschäftigt sich die Zeitreihenanalyse in diesem Kapitel mit der statistischen Darstellung der Zeitreihendaten.

2.2.1 Kenngrößen

Die statistischen Kenngrößen bilden die grundlegende Beschreibung einer Zeitreihe $(x_t)_{t\,=\,1,\ldots,N}$, wobei x_t den Wert zum Zeitpunkt t darstellt, und es insgesamt N Zeitpunkte in dieser Zeitreihe gibt. In diesem Kapitel werden grundlegende statistische Kenngrößen zur Darstellung der Zeitreihe $(x_t)_{t\,=\,1,\ldots,N}$ beschrieben und mathematisch eingeführt.

Arithmetisches Mittel

Durch das arithmetische Mittel

$$\bar{x} = \frac{1}{N}\sum_{t=1}^{N} x_t \tag{1}$$

wird die zentrale Lage der Werte dieser Zeitreihe beschrieben, um die die Datenpunkte dieser Zeitreihe schwanken [Sch01].

Varianz

Die Varianz

$$s^2 = \frac{1}{N}\sum_{t=1}^{N} (\,x_t - \bar{x}\,)^2 \tag{2}$$

beschreibt, wie groß die Datenpunkte dieser Zeitreihe um den mittleren Wert schwanken [Sch01].

Standardabweichung

Die Standardabweichung

$$s = \sqrt{s^2} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(x_t - \bar{x})^2} \quad (3)$$

ist die Quadratwurzel der Varianz. Sie berechnet die Streubreite der Werte dieser Zeitreihe rund um das arithmetische Mittel.

Die bisher erwähnten drei Kenngrößen stellen die statistischen Charakteristika einer einzelnen Zeitreihe dar. Nun wird eine andere Zeitreihe $(y_t)_{t=1,\dots,N}$ eingeführt, die in demselben Zeitraum zu derselben Messungsfrequenz wie $(x_t)_{t=1,\dots,N}$ gemessen wird. Die folgenden Kenngrößen beschreiben den linearen Zusammenhang dieser zwei Zeitreihen.

Kovarianz

Wie stark die Zeitreihe $(x_t)_{t=1,\dots,N}$ und die Zeitreihe $(y_t)_{t=1,\dots,N}$ voneinander linear abhängig sind, beschreibt die Kovarianz

$$c = \frac{1}{N}\sum_{t=1}^{N}(x_t - \bar{x})(y_t - \bar{y}) \quad (4)$$

dieser zwei Zeitreihen, wobei $\bar{y}$ für das arithmetische Mittel der Zeitreihe $(y_t)_{t=1,\dots,N}$ steht. Das Vorzeichen der Kovarianz entscheidet, ob es einen positiven oder negativen linearen Zusammenhang zwischen den beiden Zeitreihen geben kann. Je größer der Absolutbetrag der Kovarianz ist, desto stärker sind die Zeitreihe $(x_t)_{t=1,\dots,N}$ und die Zeitreihe $(y_t)_{t=1,\dots,N}$ voneinander linear abhängig. Umgekehrt besteht bei betragsmäßig kleiner Kovarianz kaum ein linearer Zusammenhang der beiden Zeitreihen.

Korrelation

Die Korrelation beschreibt ebenfalls, wie stark ein linearer Zusammenhang zwischen der Zeitreihe $(x_t)_{t=1,\dots,N}$ und der Zeitreihe $(y_t)_{t=1,\dots,N}$ sein kann. Die Korrelation ist eine normierte Form der Kovarianz. Die Korrelation unterscheidet sich von der Kovarianz darin, dass die Korrelation einen bestimmten Wertebereich [-1,1] hat. Mit dem begrenzten Wertebereich ist die Korrelation außerdem vergleichbar zwischen

verschiedenen Anwendungen. Je näher der Absolutbetrag der Korrelation am Wert 1 ist, desto stärker ist der lineare Zusammenhang der beiden Zeitreihen. Umgekehrt besteht am Wert 0 kaum ein linearer Zusammenhang der beiden Zeitreihen. Das Vorzeichen der Korrelation entscheidet ebenfalls, ob diese beiden Zeitreihen voneinander positiv oder negativ linear abhängig sein können.

Die Korrelation der beiden Zeitreihen wird durch den Korrelationskoeffizienten von Bravais-Pearson

$$r = \frac{\frac{1}{N}\sum_{t=1}^{N}(x_t - \bar{x})\,(y_t - \bar{y})}{\sqrt{\frac{1}{N}\sum_{t=1}^{N}(x_t - \bar{x})^2} \cdot \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_t - \bar{y})^2}} \tag{5}$$

berechnet [Sch01].

Die Abbildung 1 stellt drei unterschiedliche zweidimensionale Verteilungen dar. Bei 1) ist zu erkennen, dass es ein negativer linearer Zusammenhang aller Punkte besteht, weshalb hier sowohl die Kovarianz als auch die Korrelation negativ sind und ihre Beträge relativ groß sind bzw. die Korrelation nahe -1 ist. Bei 3) zeigen die Punkte einen positiven linearen Zusammenhang, weshalb die Kovarianz und die Korrelation beide positiv und relativ groß sind bzw. die Korrelation nahe 1 ist. Bei 2) ist eine andere Verteilung. Alle Punkte sind durcheinander verteilt, wobei kaum ein linearer Zusammenhang dieser Punkte zu sehen ist. Deswegen sind sowohl die Kovarianz als auch die Korrelation nahe 0.

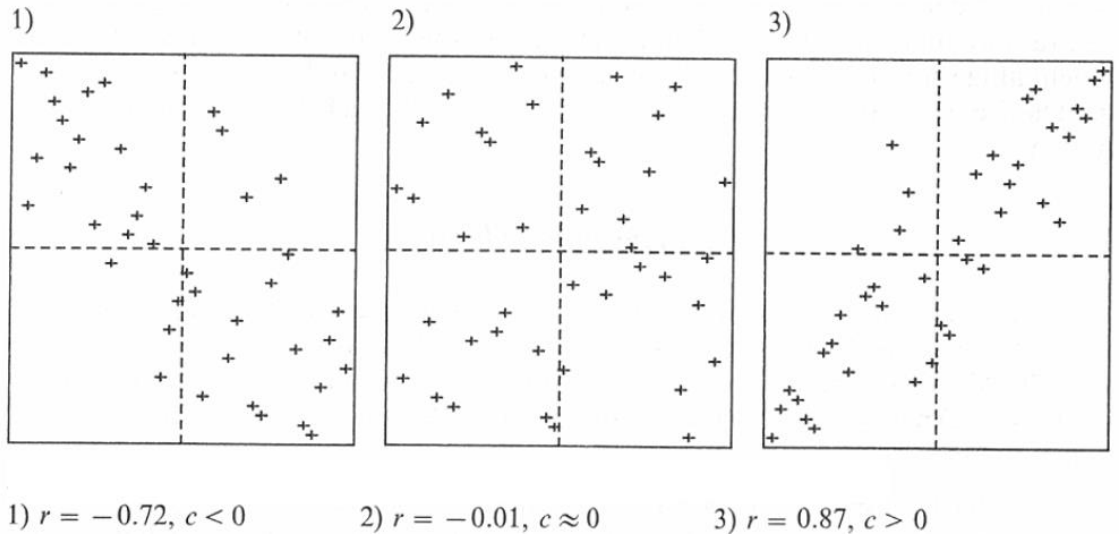

Abbildung 1: Drei zweidimensionale Verteilungen [Bic08]

Die Korrelation bietet einen gesamten Überblick über die lineare Ähnlichkeit der beiden Zeitreihen an. Wenn die beiden Zeitreihen $(x_t)_{t = 1,\dots,N}$ und $(y_t)_{t = 1,\dots,N}$ jeweils als die diskreten Signale $x(t)_{t = 1,\dots,N}$ und $y(t)_{t = 1,\dots,N}$ betrachtet werden, kann aus der Sicht der Signalverarbeitung die Ähnlichkeit der beiden Signale bzw. Zeitreihen durch Kreuzkorrelation gemessen werden.

Kreuzkorrelation

Die Kreuzkorrelation misst die Ähnlichkeit zweier Signale bei unterschiedlichen Zeitverschiebungen der beiden Signale.

Die Kreuzkorrelation der beiden diskreten Signale $x(t)_{t = 1,\dots,N}$ und $y(t)_{t = 1,\dots,N}$ wird durch

$$r_{xy}(\tau) = \sum_{t=-\infty}^{\infty} x^*(t)\, y(t+\tau) \tag{6}$$

bestimmt, wobei $x^*(t)$ für die konjungiert komplexe Funktion des Signals $x(t)_{t = 1,\dots,N}$ steht, und τ für die Zeitverschiebung steht.

Wenn die beiden Signale nun kontinuierlich sind, kann die Kreuzkorrelation durch

$$r_{xy}(\tau) = \int_{-\infty}^{\infty} x^*(t)\, y(t+\tau) dt \tag{7}$$

bestimmt werden.

Ausgehend davon, dass die beiden diskreten Signale $x(t)_{t = 1,\dots,N}$ und $y(t)_{t = 1,\dots,N}$ nur reelle Signalwerte enthalten, ist die konjungiert komplexe Funktion $x^*(t)_{t = 1,\dots,N}$ genau das Signal $x(t)_{t = 1,\dots,N}$ selbst. Die Kreuzkorrelation der beiden diskreten Signale wird dann durch

$$r_{xy}(\tau) = \sum_{t=-\infty}^{\infty} x(t)\, y(t+\tau) \tag{8}$$

bestimmt.

Daraus folgt analog die Formel zur Berechnung der Kreuzkorrelation der beiden Zeitreihen $(x_t)_{t = 1,\dots,N}$ und $(y_t)_{t = 1,\dots,N}$:

$$r_{xy}(\tau) = \sum_{t=-\infty}^{\infty} x_t\, y_{t+\tau} \tag{9}$$

Hier hilft die Zeitverschiebungsgröße τ dabei, die Zeitreihe $(y_t)_{t = 1,\dots,N}$ zu verschieben. Die Zeitverschiebungsgröße τ entscheidet, wie weit die Zeitreihe $(y_t)_{t = 1,\dots,N}$ verschoben wird. Die Positionen der Zeitreihe $(x_t)_{t = 1,\dots,N}$ sind alle festgelegt, zu denen die Positionen der Zeitreihe $(y_t)_{t = 1,\dots,N}$ parallel verschoben werden. Bei jeder Zeitverschiebungsgröße wird die Summe aller Multiplikationsprodukte von den Zeitreihenwerten $(x_t)_{t = 1,\dots,N}$ und den Zeitreihenwerten $(y_t)_{t = 1,\dots,N}$ in den überlappenden Positionen berechnet. Je größer die Kreuzkorrelationswerte der beiden Zeitreihen sind, umso ähnlicher sind die beiden Zeitreihen.

Im folgenden Kapitel wird die Zeitreihe $(x_t)_{t = 1,\dots,N}$ weiter als das Erklärungsobjekt für das Komponentenmodell der Zeitreihenanalyse verwendet.

2.2.2 Komponentenmodell

Die Zeitreihe $(x_t)_{t=1,\dots,N}$ lässt sich aus der Sicht der klassischen Zeitreihenanalyse in die folgenden vier Komponenten zerlegen [Sch10]:

(1) Trend $(g_t)_{t = 1,\dots,N}$

Ein Trend stellt die langfristige Entwicklung einer Zeitreihe dar. Aus der statistischen Sicht ist ein Trend an sich eine Funktion, die entweder monoton seigend oder monoton fallend ist [Sch10].

(2) Konjunkturkomponente $(k_t)_{t = 1,\dots,N}$

Die Konjunkturkomponente beschreibt eine Schwankung mit einer langzeitigen Periode, die mehrere Jahre dauern kann [Sch10].

(3) Saison $(s_t)_{t = 1,\dots,N}$

Eine Saison ist auch eine Schwankungskomponente, deren Periode aber kalenderbedingt ist [Sch10]. Die Periode einer Saison kann einen Tag, einen Monat, eine Jahreszeit usw. dauern. Jedes Jahr wiederholt sich die Saison regelmäßig und relativ unverändert [Sch01].

(4) Restkomponente $(u_t)_{t = 1,\dots,N}$

Die Restkomponente fasst die nicht zu erklärenden Einflüsse oder Störungen zusammen [Sch01]. Im Gesamtbild ändert sie sich unsystematisch und unregelmäßig.

Der Trend und die Konjunkturkomponente können zur glatten Komponente zusammengefasst werden, weil die beiden Komponenten einen langfristigen Einfluss auf die Zeitreihen darstellen [Sch10]. Die Konjunkturkomponente und die Saison sind beide zyklische Komponente, deshalb können sie zur zyklischen Komponente zusammengefasst werden [Sch10]. Hier wird die Konjunkturkomponente in beiden Zusammenfassungen mit aufgenommen.

Für die Verknüpfung dieser vier Komponenten gibt es drei Grundmodellen.

(1) Additives Modell

Im additiven Modell verknüpfen sich die einzelnen Komponenten additiv. Eine Zeitreihe wird durch

$$x_t = g_t + k_t + s_t + u_t \tag{10}$$

beschrieben, wobei die vier Komponenten voneinander unabhängig sind [Lip].

(2) Multiplikatives Modell

Im multiplikativen Modell verknüpfen sich die einzelnen Komponenten multiplikativ. Eine Zeitreihe wird durch

$$x_t = g_t \cdot k_t \cdot s_t \cdot u_t \tag{11}$$

beschrieben, wobei die einzelnen Komponenten einander verstärken [Lei98].

Durch das Logarithmieren kann das multiplikative Modell als ein additives Modell

$$\log x_t = \log(g_t \cdot k_t \cdot s_t \cdot u_t) = \log g_t + \log k_t + \log s_t + \log u_t \tag{12}$$

dargestellt werden.

(3) Gemischtes Modell

Eine Zeitreihe kann auch auf der Kombination des additiven Modells und des multiplikativen Modells basieren. Z.B. sieht sie in diesem Modell wie

$$x_t = g_t + k_t \cdot s_t \cdot u_t \tag{13}$$

oder

$$x_t = g_t \cdot k_t + s_t + u_t \tag{14}$$

aus.

2.2.3 Ähnlichkeitsmessung zweier Zeitreihen

In der Zeitreihenanalyse ist es einer der Schwerpunkte, nach den Ähnlichkeiten der Zeitreihen zu suchen. Es gibt zwei Arten Ähnlichkeiten in Zeitreihen, die formbasierende Ähnlichkeit und die strukturbasierende Ähnlichkeit [Lin09]. Die formbasierende Ähnlichkeit bestimmt die Ähnlichkeit zweier Zeitreihen durch den Vergleich ihrer lokalen Muster, und die strukturbasierende Ähnlichkeit bestimmt die Ähnlichkeit zweier Zeitreihen durch den Vergleich ihrer globalen Strukturen [Lin09]. In diesem Kapitel wird ein Überblick über die Ansätze zur Bestimmung der beiden Ähnlichkeiten angeboten.

2.2.3.1 Formbasierende Ähnlichkeit: Angular Metric for Shape Similarity

Dieser Ansatz baut auf [Nak12] auf. In diesem Kapitel steht die Abkürzung AMSS für Angular Metric for Shape Similarity.

In der AMSS werden die Zeitreihen als die Vektorreihen angesehen, wobei die Ähnlichkeitsberechnung zweier Vektorreihen auf den Richtungen dieser beiden Vektorreihen basiert, sondern nicht auf den Zahlwerten der Datenpunkte in den Zeitreihen [Nak12]. Die Zeitreihen $(x_t)_{t=1,\dots,N}$ und $(y_t)_{t=1,\dots,N}$ aus dem Kapitel 2.2.1 werden hier weiter als die Erklärungsobjekte verwendet. Wie in der Abbildung 2 demonstriert werden die Zeitreihen $(x_t)_{t=1,\dots,N}$ und $(y_t)_{t=1,\dots,N}$ jeweils als die Vektorreihen $(X_t)_{t=1,\dots,N}$ und $(Y_t)_{t=1,\dots,N}$ betrachtet.

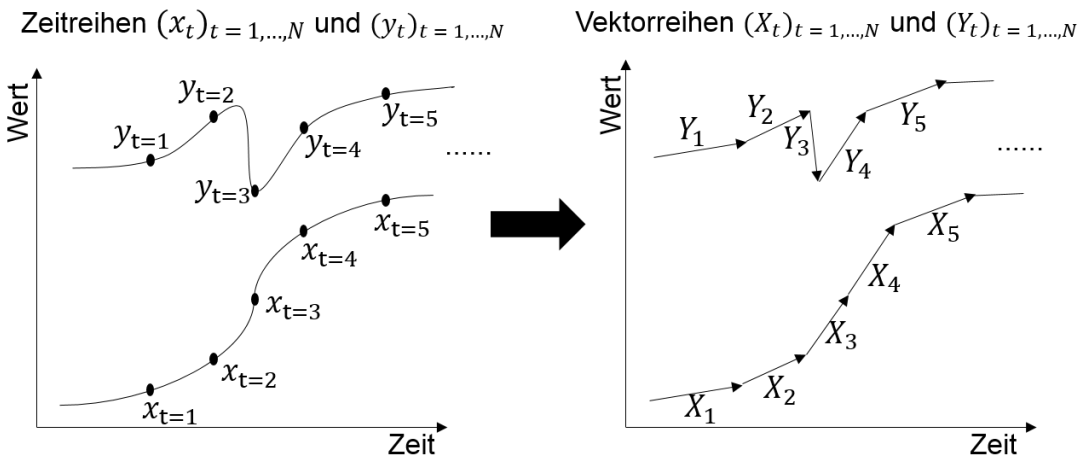

Abbildung 2: Zeitreihen werden als Vektorreihen betrachtet

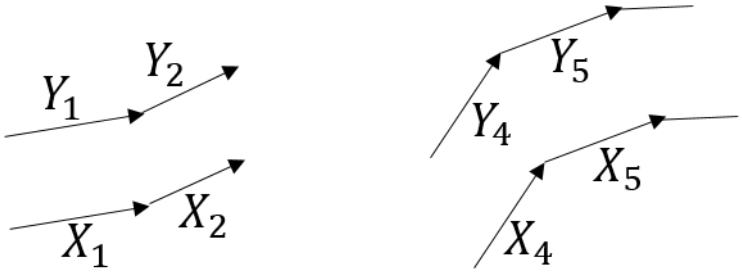

Abbildung 3: Formähnliche Teilvektoren

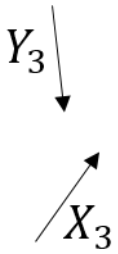

Abbildung 4: Formunähnliche Teilvektoren

In der Abbildung 2 werden 5 Vektoren jeweils aus den beiden Vektorreihen angezeigt. In der Abbildung 3 ist zu erkennen, dass die Teilvektoren X_1, X_2, X_4, X_5 und die Teilvektoren Y_1, Y_2, Y_4, Y_5 in Form zueinander hoch ähnlich sind. Jedoch haben die Vektoren X_3 und Y_3 kaum eine Ähnlichkeit in der Richtung wie in der Abbildung 4 dargestellt.

Die beiden Vektorreihen $(X_t)_{t = 1,\dots,N}$ und $(Y_t)_{t = 1,\dots,N}$ sind außer an den Stellen X_3 und Y_3 zueinander ausgerichtet, d.h. sie weisen beide theoretisch eine hohe globale förmige Ähnlichkeit auf. Trotzdem könnte der lokale Unterschied der Richtungen von X_3 und Y_3 zu einer geringen globalen Ähnlichkeit bei der formbasierenden Ähnlichkeitsberechnung führen [Nak12]. Die AMSS berechnet die globale formbasierende Ähnlichkeit zweier Zeitreihen bzw. Vektorreihen, indem die Teilvektoren mit ähnlichen Richtungen gepaart werden [Nak12].

Die Richtungen jedes Teilvektors von der Vektorreihe $(X_t)_{t = 1,\dots,N}$ und jedes Teilvektors von der Vektorreihe $(Y_t)_{t = 1,\dots,N}$ werden verglichen, wobei der Kosinus des Winkels θ, der wie in der Abbildung 5 illustriert aus den Richtungen des Teilvektors X_t und des Teilvektors Y_t besteht, für den Vergleich eingesetzt wird.

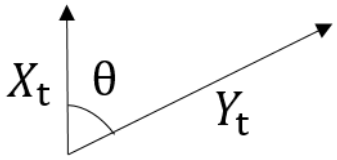

Abbildung 5: Ein Winkel, der aus den Richtungen von X_t und Y_t besteht

Laut [Nak12] weisen X_t und Y_t sehr unterschiedliche Richtungen auf, wenn der Winkel θ größer als 90 Grad ist. In diesem Fall haben X_t und Y_t gar keine förmige Ähnlichkeit. Umgekehrt teilen X_t und Y_t eine förmige Ähnlichkeit, wenn der Winkel θ kleiner als 90 Grad ist.

Im Weiteren wird die rekursive Berechnung für das endliche AMSS-Ergebnis der globalen formbasierenden Ähnlichkeit von den Zeitreihen $(x_t)_{t = 1,\dots,N}$ und $(y_t)_{t = 1,\dots,N}$ bzw. den Vektorreihen $(X_t)_{t = 1,\dots,N}$ und $(Y_t)_{t = 1,\dots,N}$ verwendet, indem alle Teilvektoren der beiden Vektorreihen miteinander verglichen werden, und die formähnlichsten Teilvektoren gepaart und weiter verglichen werden [Nak12].

Das AMSS-Endergebnis der rekursiven Berechnung stellt die globale formbasierende Ähnlichkeit der beiden Zeitreihen dar. Je größer das AMSS-Endergebnis ist, desto stärker sind die beiden Zeitreihen voneinander global förmig abhängig. Umgekehrt bei einem kleinen AMSS-Endergebnis bzw. einem dem Wert 0 nahen AMSS-Endergebnis weisen die beiden Zeitreihen kaum eine formbasierende Ähnlichkeit auf.

2.2.3.2 Strukturbasierende Ähnlichkeit: Distanzmessung

Zum strukturbasierenden Vergleich zweier oder mehrerer Zeitreihen ist die Distanz ein wichtiges Maß. Wie ähnlich zwei Zeitreihen sind, kann durch die Distanz dieser zwei Zeitreihen definiert werden, wobei die Ähnlichkeit umgekehrt proportional zu der Distanz der zwei Zeitreihen ist [Gel15]. Bei einer großen Distanz weisen die beiden Zeitreihen geringe Ähnlichkeit auf, umgekehrt bei einer kleinen bzw. dem Wert 0 nahen Distanz weisen die beiden Zeitreihen große Ähnlichkeit auf [Spi15].

Zur Vereinfachung der Erklärung werden die im Kapitel 2.2.1 erwähnten Zeitreihen $(x_t)_{t = 1,\dots,N}$ und $(y_t)_{t = 1,\dots,N}$ weiter als die Erklärungsobjekte verwendet.

Eine der in der Datenanalyse bzw. der Zeitreihenanalyse am meisten benutzten Distanzmessungen ist die Euklidische Distanz. Die Euklidische Distanz zwischen den Zeitreihen $(x_t)_{t = 1,\dots,N}$ und $(y_t)_{t = 1,\dots,N}$ ist

$$d(x_t, y_t) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_N - y_N)^2} \tag{15}$$

Außerdem wird die Manhattan Distanz auch für die Distanzmessung in der Zeitreihenanalyse verwendet. Die Manhattan Distanz zwischen $(x_t)_{t = 1,\dots,N}$ und $(y_t)_{t = 1,\dots,N}$ wird durch

$$d(x_t, y_t) = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_N - y_N| \tag{16}$$

bestimmt.

Sowohl die Euklidische Distanz als auch die Manhattan Distanz sind die Spezialfälle der Minkowski Distanz, deren Formel ist

$$d(x_t, y_t) = \sqrt[p]{\sum_{t=1}^{n} |x_t - y_t|^p} \tag{17}$$

wobei p eine positive reelle Zahl ist. Bei $p = 1$ ist die Minkowski Distanz die Manhattan Distanz, und bei $p = 2$ ist die Minkowski Distanz die Euklidische Distanz.

Der Vorteil der Distanzmessung mit der Minkowski Distanz und ihren Spezialfällen ist ihr einfaches Verständnis und ihre einfache Implementierung mit großen Datensätzen [Gel15]. Jedoch ist die Nutzung dieser Distanzmessung eingeschränkt, da die zwei Zeitreihen, deren Distanz durch die Minkowski Distanz gemessen wird, die gleiche Anzahl der Datenpunkte haben müssen, und diese Datenpunkte jeweils aus den beiden Zeitreihen paarweise vertikal zur x-Achse bzw. zur Zeitachse ausgerichtet sein müssen

[Gel15], [Lin09]. Um diese Messungsbeschränkung zu überwinden, bietet sich eine andere Methode Dynamic Time Warping zur Distanzmessung an, die im Folgenden DTW genannt wird.

Unterschiedlich zur strengen vertikalen Ausrichtung zweier Zeitreihen können die Zeitreihen im DTW anders ausgerichtet werden. In der Abbildung 6 werden die Ausrichtungsweisen zweier Zeitreihen bei Minkowski Distanzmessung und bei DTW Distanzmessung illustriert.

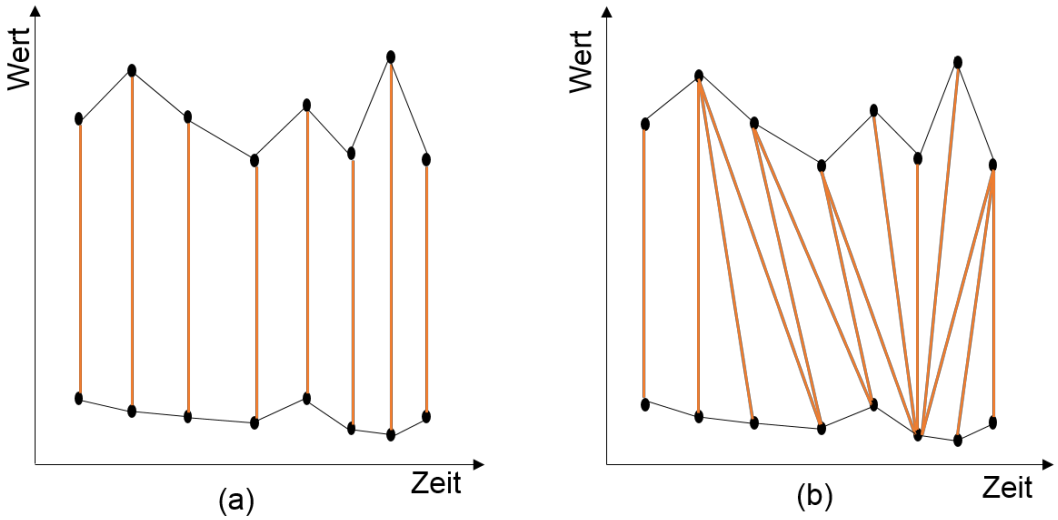

Abbildung 6: a) Zeitreihenausrichtung bei Minkowski Distanzmessung

b) Zeitreihenausrichtung bei DTW Distanzmessung

Mithilfe der dynamischen Programmiertechniken sucht das DTW nach der besten Ausrichtungsweise, die die optimale Distanz der zwei Zeitreihen bestimmt, wobei die Distanz zweier einzelnen Datenpunkte durch z.B. die Euklidische Distanzmessung berechnet werden kann [Lin09]. In [Ber94], [Keo04] wird die Vorgehensweise des DTW im Detail betrachtet und erklärt.

2.3 Clusteranalyse

Das Clustern ist ein Vorgang der Gruppierung einer Menge Objekte in die Klassen ähnlicher Objekte, wobei eine Klasse ein Cluster ist [Ban15]. Um die Ähnlichkeiten in zwei oder mehreren Smart Meter Messreihen zu berechnen, kann das Clustern angewendet werden. An sich ist das Clustern eine Methode des Maschinellen Lernens bzw. des Unüberwachten Lernens. Das Maschinelle Lernen ist eine Anwendung der

künstlichen Intelligenz. Das Ziel des Maschinellen Lernens ist es, dass der Computer Kenntnisse erwirbt ohne die menschliche Assistenz und das explizite Programmieren [Sar18].

Das Maschinelle Lernen kann in die folgende drei Hauptarten gegliedert werden: Überwachtes, Unüberwachtes und Semiüberwachtes Lernen. Beim Überwachten Lernen sind eine Menge historische Eingangsdaten und Ausgangsdaten gegeben, und der Berechnungsvorgang von den Eingangsdaten zu den bekannten Ausgangsdaten wird viel trainiert, um die Fähigkeit des genauen Prognostizierens zu verbessern [Sar18]. Das Unüberwachte Lernen beschäftigt sich mit der Klassifizierung der Eingangsdaten in die verschiedenen Gruppen [Kov99]. Die Eingangsdaten mit ähnlichen Eigenschaften werden in dieselbe Gruppe klassifiziert. Innerhalb einer Gruppe sind die Daten zueinander ähnlicher als zu den Daten der anderen Gruppen [Kov99]. Das Semiüberwachtes Lernen ist eine Kombination des Überwachten Lernens und des Unüberwachten Lernens, wobei der Datentrainingsvorgang im Überwachten Lernen und die Datenklassifizierung im Unüberwachten Lernen kombiniert angewendet werden [Sar18].

Im Folgenden werden die Vorgehensweisen von zwei üblichen Clustermethoden erklärt, und zwar K-Means Clustern und Fuzzy C-Means Clustern.

2.3.1 K-Means Clustern

Im K-Means Clustern werden erstens beliebig K Zentren in einer Datenmenge definiert, auf die die Cluster zentriert sind [Deh10]. Dann werden die Distanzen jedes Datenpunktes und jedes Clusterzentrums berechnet. Jeder Datenpunkt wird dem Cluster, zu dessen Zentrum dieser Datenpunkt nächstgelegen ist, zugeordnet [Kim11]. Für die Distanzberechnung wird die Euklidische Distanzmessung üblich verwendet. Danach werden auf der Zwischendatenverteilung nach der ersten Zuordnung basierend die Clusterzentren aktualisiert bzw. durch die Algorithmen erneut berechnet [Kim11]. Die Datenpunkte werden dann den neuen Clustern, zu deren Zentren diese Datenpunkte nächstgelegen sind, zugeordnet [Kim11]. Daraus entsteht eine Schleife für die Aktualisierung der Clusterzentren und die Zuordnung der Datenpunkte nach der letzten Datenzuordnung. Diese Schleife hört erst auf, wenn die Clusterzentren gemäß den Algorithmen nicht mehr geändert werden können bzw. die Clusterzentren schon in den optimalen Positionen sind [Deh10].

Für die Bestimmung des optimalen Wertes von K können die Silhouette-Koeffizienten eingesetzt werden. Ein Silhouette-Koeffizient eines Datenpunktes in einem Cluster entscheidet, ob dieser Datenpunkt diesem Cluster zugeordnet werden soll oder nicht. Er wird durch

$$S = \frac{b - a}{\max\{a, b\}} \tag{18}$$

berechnet, wobei a für die durchschnittliche Distanz dieses Datenpunktes zu allen anderen Datenpunkten innerhalb seines Clusters steht, und b für die durchschnittliche Distanz dieses Datenpunktes zu allen Datenpunkten des ihm nächstgelegenen Clusters steht.

Der Wertebereich des Silhouette-Koeffizienten ist [-1,1]. Ein 1 naher Silhouette-Koeffizient eines Datenpunktes in einem Cluster bedeutet, dass dieser Datenpunkt wegen seiner hohen Ähnlichkeiten zu den Datenpunkten dieses Clusters diesem Cluster zugeordnet werden soll. Ein -1 naher Silhouette-Koeffizient bedeutet, dass der Datenpunkt eher zu den Datenpunkten des ihm nächstgelegenen Clusters hoch ähnlich ist, und er diesem Cluster nicht zugeordnet werden soll. Ein 0 naher Silhouette-Koeffizient bedeutet, dass dieser Datenpunkt an der Grenze zweier Cluster liegt.

Zur Bestimmung von K bzw. davon, wie viele Clusterzentren beim K-Means Clustern definiert werden, können die durchschnittlichen Silhouette-Koeffizienten aller Datenpunkte in einer Datenreihe bei K = 2, 3, …, 30 oder größer berechnet werden, wobei das K mit dem größten durchschnittlichen Silhouette-Koeffizienten als die Anzahl der Clusterzentren bzw. Cluster festgelegt wird. In diesem Fall sind die Datenpunkte in dieser Zeitreihe allgemein hoch ähnlich zu ihren eigenen Clustern [Vio18]. Darüber hinaus gibt es noch andere Methoden, die das K initialisieren.

Das K-Means Clustern ist zwar nicht kompliziert zu implementieren, aber seine Genauigkeit hängt stark von den Genauigkeiten der Algorithmen für die Berechnung bzw. Aktualisierung der Clusterzentren ab [Est13].

2.3.2 Fuzzy C-Means Clustern

Die Vorgehensweise des Fuzzy C-Means Clusterns ist ähnlich zu der des K-Means Clusterns, aber im Fuzzy C-Means Clustern wird der Begriff Mitgliedschaftsgrad eingeführt. Hierbei gehört ein Datenpunkt nicht zu einem einzelnen Cluster, sondern gehört zu jedem Cluster mit einem Mitgliedschaftsgrad. Der Mitgliedschaftsgrad stellt

dar, wie stark zugehörig ein Datenpunkt zu einem Cluster ist [Aza14]. Im Fuzzy C-Means Clustern werden erstens ebenfalls beliebig Clusterzentren erstellt. Dann werden durch die Algorithmen die Mitgliedschaftsgrade von jedem Datenpunkt zu jedem Clusterzentrum bzw. Cluster berechnet. Auf der Zwischenverteilung der Clusterzentren und den schon berechneten Mitgliedschaftsgraden basierend werden die Clusterzentren aktualisiert bzw. durch die Algorithmen erneut berechnet. Die Mitgliedschaftsgrade aller Datenpunkte werden deswegen für die neuen Cluster nochmal berechnet [Kim11]. Daraus entsteht eine Schleife für die Aktualisierung der Clusterzentren und der Mitgliedschaftsgrade. Diese Schleife hört erst auf, wenn die Mitgliedschaftsgrade gemäß den implementierten Algorithmen optimal verteilt sind.

3 Modellierung

Trotz vieler bisher erwähnten auf die Analyse der Smart Meter Daten bezogenen Methoden liegt der Fokus dieser Arbeit auf dem Vergleich mehrerer Smart Meter Messreihen. Dieses Kapitel befasst sich damit, wie ein Vergleich der Smart Meter Messreihen aus realen Datensätzen modelliert wird, und setzt dazu exemplarisch einige der im Kapitel 2 vorgestellten Methoden ein.

3.1 Datensatz

Für die Modellierung wird hier der Datensatz von [Tja] verwendet. Die Daten stammen aus den Messdaten des IZES-Datensatzes von 497 Haushalten [Tja]. Davon wurden 74 Haushaltlastprofile für das Jahr 2010 ausgewählt, wobei sie alle von einem einzigen Netzbetreiber stammen [Tja]. Die Messlücken des ganzen Jahres betragen weniger als einen Tag [Tja]. Wenn eine Messlücke auftritt, wird sie mit den Messdaten desselben Zeitraums des vorbeigegangenen gleichen Wochentags beseitigt [Tja].

Die Zeitreihen liegen in 1-sekündiger Auflösung vor. Da eine derart hohe Auflösung aktuell allerdings noch kein Standard ist und um die Rechenzeit zu beschränken, wird im Folgenden mit den ebenfalls verfügbaren 1-minütigen Mittelwerten gearbeitet.

Die Zeitreihen haben die Wirkleistungen und Blindleistungen in allen drei Phasen aller 74 Haushalte erfasst. Um die Ähnlichkeitsuntersuchung des Stromverbrauchs durchzuführen, wird die Summe der Wirkleistungen in allen drei Phasen zu jedem Zeitpunkt berechnet. Blindleistungen werden an dieser Stelle nicht weiter betrachtet.

Die Abbildung 7 zeigt die Lastprofile der Haushalte 1 und 2 mit 1-minütiger Zeitauflösung im ganzen Jahre 2010 an. Es ist zu bemerken, dass die Lastprofile stochastisch verteilt sind. Im ersten Blick kann so eine Aussage über den Vergleich dieser beiden Messreihen getroffen werden, dass in beiden Messreihen die auffälligen Peaks im Mai, September und Dezember auftreten. Diese Aussage beschreibt eine beispielhafte Ähnlichkeit der beiden Messreihen. Das Ziel der Modellierung ist es, die Ähnlichkeiten mehrerer Messreihen bzw. des Haushaltsstromverbrauchsverhaltens unter den stochastischen Verteilungen der Lastprofile zu identifizieren.

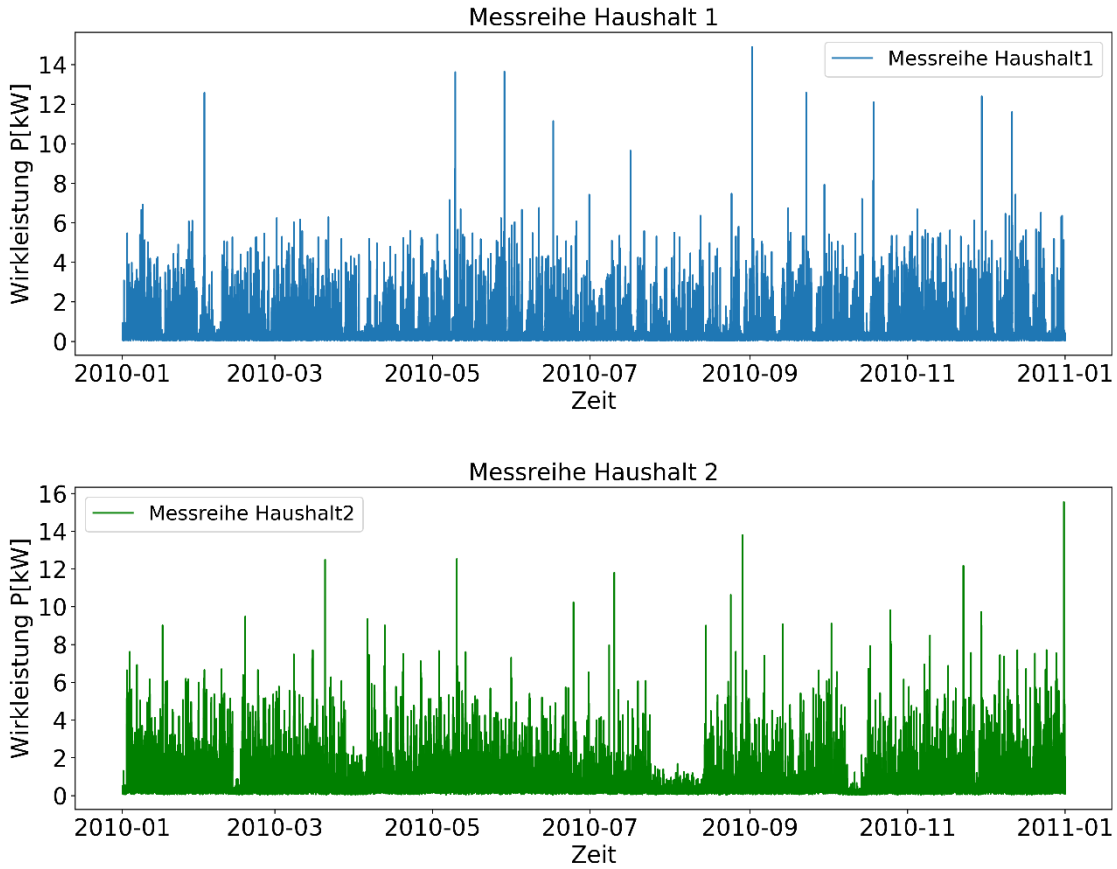

Abbildung 7: Lastprofile der Haushalte 1 und 2 im Jahre 2010

3.2 Korrelationsanalyse

Eine Untersuchung der Korrelation der Zeitreihen bei unterschiedlichen Auflösungen dient dazu, die statistische Beziehung zwischen linearer Ähnlichkeit der Zeitreihen und Messauflösung zu identifizieren.

Nach dem Indexieren der Rohdaten ins Format „Zeitstempel + Wirkleistung“ werden alle 74 Haushaltsmessreihen in Python eingelesen. Mit der Startauflösung 1 Minute fängt die Korrelationsanalyse an, dass die Korrelation jeder Haushaltmessreihe und jeder anderen Haushaltmessreihe berechnet wird. Hieraus ergeben sich insgesamt $\frac{74*73}{2} = 2701$ Korrelationskombinationen. Danach kommt eine Korrelationswertmenge als das Ergebnis für die Auflösung 1 Minute. Analog wird die Korrelationswertmenge bei jeder Auflösung von 2 Minuten bis 120 Minute berechnet, wobei die Verringerung der Zeitauflösung durch Resampling auf die Mittelwerte realisiert wird kann. Am Ende werden die Korrelationswertmengen zu allen 120 verschiedenen Zeitauflösungen erzeugt. Zur deskriptiven Analyse werden die maximalen, durchschnittlichen und

minimalen Werte in allen Korrelationswertmengen zu jeder Zeitauflösung bestimmt, um einen repräsentativen Verlauf der Korrelation zur Messauflösung darzustellen. Dieser Verlauf demonstriert, wie die linearen Ähnlichkeiten aller Haushaltmessreihen in den 120 verschiedenen Zeitauflösungen verteilt sind.

3.3 Analyse der Manhattan Distanzen

Die Distanzmessung zweier Zeitreihen misst die strukturbasierende Ähnlichkeit dieser zwei Zeitreihen wie im Kapitel 2.2.3.2 erläutert. Wegen der einfachen Realisierbarkeit der Manhattan Distanzmessung in Python werden hier die Manhattan Distanzen der Zeitreihen untersucht.

Analog zur Vorgehensweise der Korrelationsanalyse wird die Manhattan Distanz jedes Datenpunktes einer Haushaltmessreihe und jedes Datenpunktes einer anderen Haushaltmessreihe des gleichen Datenpunktes zu den Zeitauflösungen von 1 Minuten bis 120 Minuten berechnet, wobei es für 74 Haushaltmessreihen zu jeder Zeitauflösung ebenfalls insgesamt $\frac{74*73}{2} = 2701$ Vergleichskombinationen gibt. Weil die Manhattan-Distanzen aller Datenpunkte zweier Messreihen eine Distanzzeitreihe bilden, kann diese Manhattan Distanzzeitreihe auf ihren arithmetischen Mittelwert zentriert werden, damit ein Verlauf der zentralen Lage der Punktdistanzen zur Zeitauflösung dargestellt werden kann. Dabei werden die maximalen, durchschnittlichen und minimalen Werte aller Mittelwerte der Manhattan Distanzen der 2701 Vergleichskombinationen zu jeder Auflösung bestimmt. Diese maximalen, durchschnittlichen und minimalen Mittelwerte werden zusammen in einem selben Graphen demonstriert, um einen Überblick des Verlaufs der Manhattan Distanz zur Messauflösung zu verschaffen.

Außerdem kann die Standardabweichung der Manhattan Distanzzeitreihe darstellen, wie konzentriert alle einzelnen Punktdistanzen innerhalb einer Manhattan-Distanzzeitreihe zum Mittelwert verteilt sind. Deswegen wird der Verlauf der Standardabweichung der Manhattan Distanzzeitreihe zur Zeitauflösung untersucht. Die Vorgehensweise davon ist gleich wie die der Analyse der Mittelwerte der Manhattan Distanzreihen.

Mit Mittelwerten und Standardabweichungen der Manhattan-Distanzen zu den verschiedenen Zeitauflösungen stellt sich die Verteilung der allgemeinen strukturbasierenden Ähnlichkeiten aller Haushaltmessreihen dar.

3.4 Kreuzkorrelationsanalyse

Die Kreuzkorrelation wird oft in der Signalverarbeitung für die Ähnlichkeitssuche zweier Signale verwendet. Hier werden die Ähnlichkeiten zwischen den Messreihen von Haushalten 1, 3 und 4 mit Kreuzkorrelation untersucht.

In der Kreuzkorrelationsanalyse werden die Kreuzkorrelationen der Messreihen von Haushalten 1 und 3, 1 und 4, 3 und 4 zu den Zeitverschiebungen von 1 Minute bis 1000 Minuten jeweils berechnet. Hier werden zur Vereinfachung der Speicherung der Berechnungen nur 1000 Minuten als Zeitverschiebungen ausgewählt.

In dieser Analyse ist die Kreuzkorrelation selbst eine Zahlreihe bzw. Zeitreihe zu 1000 Minuten. Die drei Kreuzkorrelationsreihen werden in demselben Graphen dargestellt, um einen Überblick über die Ähnlichkeiten von Haushalten 1, 3 und 4 zu verschaffen.

3.5 K-Means Clusteranalyse

Die Clusteranalyse geht um die Klassifizierung der Lastprofile aller Haushalte, wobei die ähnlichen Lastprofile zu einem gleichen Cluster klassifiziert werden können, um die Ähnlichkeitsverteilung in den Lastprofilen darzustellen. Für diese Clusteranalyse wird das K-Means Clustern angewendet, weil das K-Means Clustern aus seinem integrierten Algorithmus in Python relativ einfach zu implementieren ist. Zur Reduzierung der riesigen Berechnungsdauer werden alle Haushaltmessreihen hier auf die Mittelwerte der Wirkleistungen zur 1-tägigen Messauflösung resampelt.

Vor dem Einsatz des K-Means Clusterns wird erstmal entschieden, wie viele Cluster für die Klassifizierung der Messreihen benutzt werden sollen. Dabei wird der Silhouette-Koeffizient verwendet. Es wird vermutet, dass es für die Klassifizierung 2 bis 30 Cluster geben kann. Die durchschnittlichen Silhouette-Koeffizienten aller Lastpunkte werden für die Clusteranzahl von 2 bis 30 berechnet. Die Clusteranzahl, bei der der durchschnittliche Silhouette-Koeffizient aller Lastpunkte am größten ist, wird für das K-Means Clustern verwendet.

Aber es gibt eine Ausnahme, wenn der größte durchschnittliche Silhouette-Koeffizient bei Clusteranzahl 2 auftritt. Mit 2 Clustern ist die Lastpunkklassifizierung relativ monoton. Um die Lastpunktverteilung in Clustern deutlicher zu illustrieren, wird in diesem Fall die Clusteranzahl verwendet, bei der der durchschnittliche Silhouette-Koeffizient am zweitgrößten ist.

Nach der Bestimmung der Clusteranzahl werden die in Python integrierten K-Means Algorithmen für das K-Means Clustern verwendet, um die optimalen Clusterpositionen zu berechnen und die optimale Zuordnung der Lastpunkte in die Cluster durchzuführen.

4 Ergebnisse und Auswertung

In diesem Kapitel werden die Ergebnisse der in der exemplarischen Modellierung eingesetzten Analysenmethoden ausgewertet.

4.1 Korrelationsverteilung aller Messreihen

In der Abbildung 8 sind die Kurven der maximalen, durchschnittlichen und minimalen Korrelationswerte aller 2701 Korrelationskombinationen mit Auflösungen von 1 Minute bis 120 Minuten dargestellt.

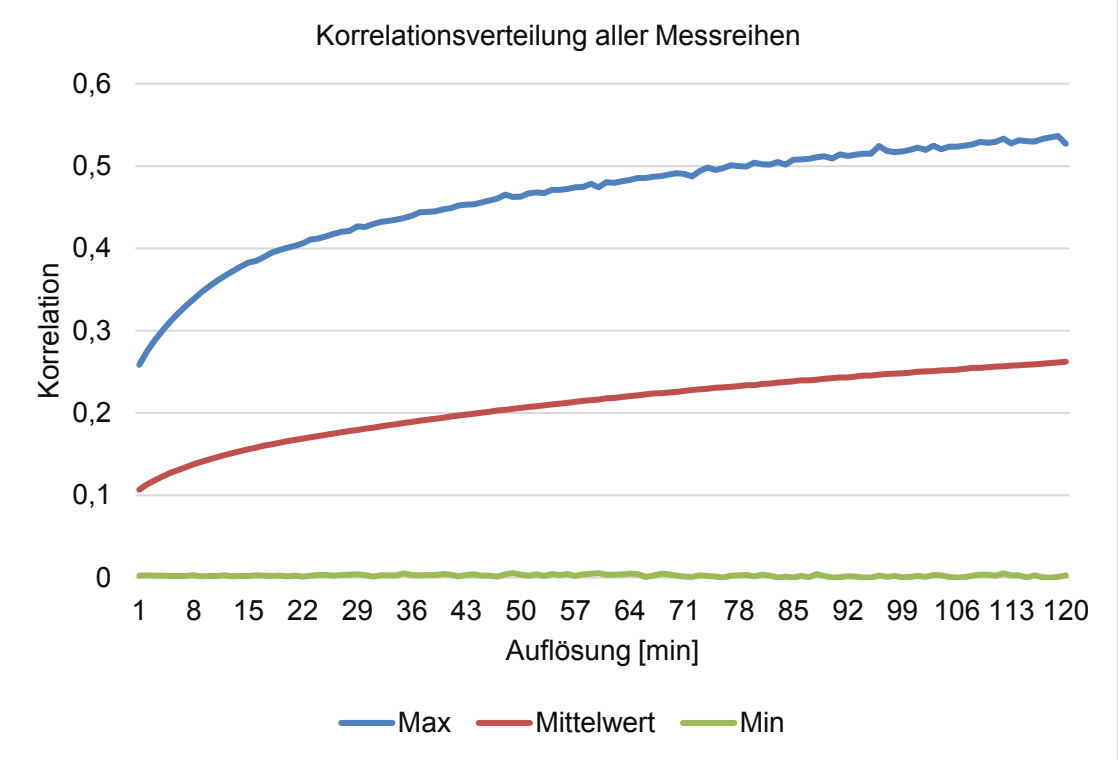

Abbildung 8: Korrelationen aller Messreihen

Die minimalen Werte der Korrelationen bleiben auf einem konstanten Niveau, das dem Wert 0 sehr nah liegt. Der Mittelwert der Korrelationen steigt langsam beim Verringern der Zeitauflösungen. Der maximale Wert steigt am Anfang relativ schnell, aber ab der 30-minütiger Zeitauflösung steigt er fast so langsam wie der Mittelwert. In der Abbildung 8 weisen alle Messreihen mehr linearen Zusammenhang auf beim Verringern der

Mesauflösung. Bei den niedrigen Auflösungen weisen einige Messreihen die Korrelationen über 0,5 auf, wobei sie an diesen Stellen zueinander hoch linear ähnlich sind.

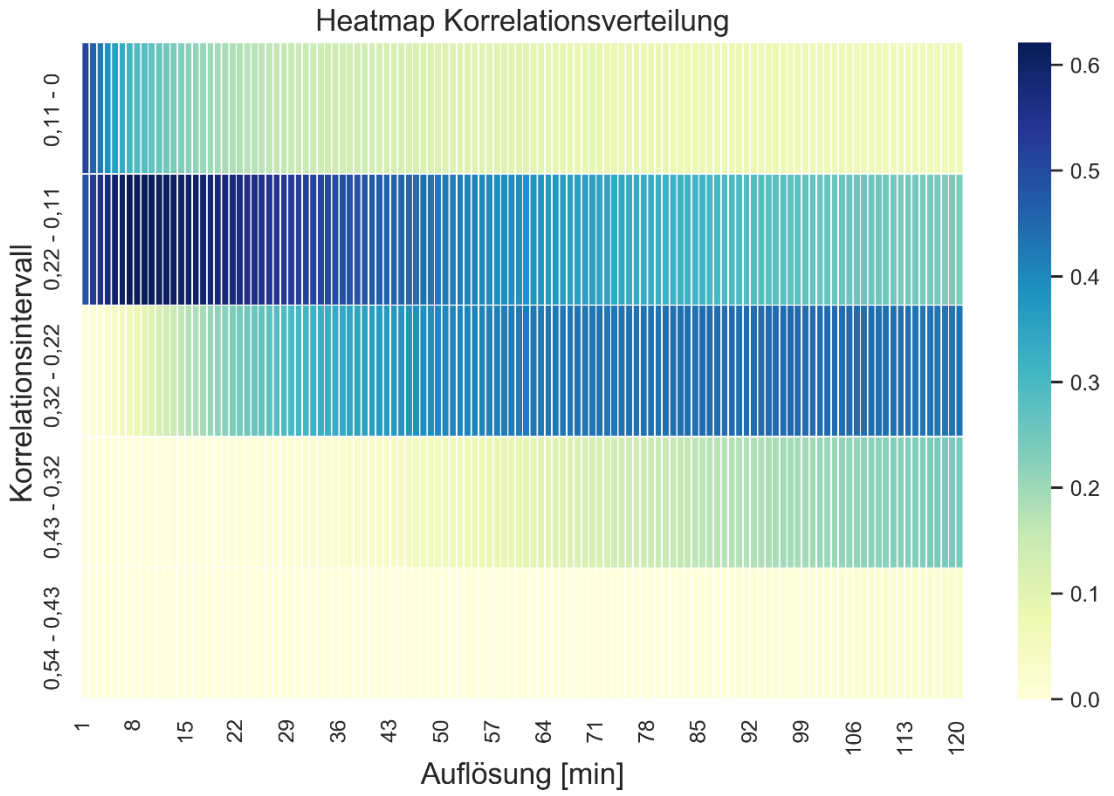

Abbildung 9: Korrelationsverteilung im Heatmap

Die Abbildung 9 illustriert das Heatmap der Korrelationsverteilung in allen Zeitauflösungen. Die Korrelationen werden in fünf gleichlange Intervalle aufgeteilt. Mit 120 Auflösungen werden insgesamt 120 × 5 Pixeln gebildet. Die Farbenintensität jedes Pixels stellt den Anteil der Korrelationen dieses Pixels an allen Korrelationen in der Auflösung dieses Pixels dar. Je dunkler die Farbe eines Intervalls bei einer Auflösung ist, desto mehr Korrelationen bei dieser Auflösung sind in diesem Intervall verteilt. Umgekehrt je heller die Farbe ist, desto weniger Korrelationen bei dieser Auflösung sind in diesem Intervall verteilt. Damit die Unterschiede der Farbenintensitäten im Heatmap deutlich zu erkennen sind, werden deshalb die Korrelationen in fünf gleichlange Intervalle geteilt. Bei einer großen Anzahl der Intervalle werden die Korrelationen fast gleichmäßig in allen Pixeln aufgeteilt, wobei sich die Farbenintensitäten voneinander nicht viel unterscheiden können.

In der Abbildung 9 ist es zu bemerken, je reduzierter die Auflösung ist, desto weniger Korrelationen gibt es bei [0, 0,11] und [0,11, 0,22]. Aber in den übrigen drei Intervallen, die die höheren Korrelationswerte repräsentieren, gibt es mehr Korrelationen beim Reduzieren der Auflösung. Allgemein sind bei jeder Messauflösung die meisten Korrelationen bei [0,11, 0,22], [0,22, 0,32] und [0,32, 0,43] verteilt. Diese drei Intervalle stellen den Durchschnitt aller Korrelationen dar.

4.2 Manhattan Distanzverteilung aller Messreihen

In der Abbildung 10 sind die Kurven der maximalen, durchschnittlichen und minimalen Mittelwerte der Manhattan Distanzen der Datenpunkte aller beider Haushaltmessreihen in den Auflösungen von 1 Minute bis 120 Minuten dargestellt.

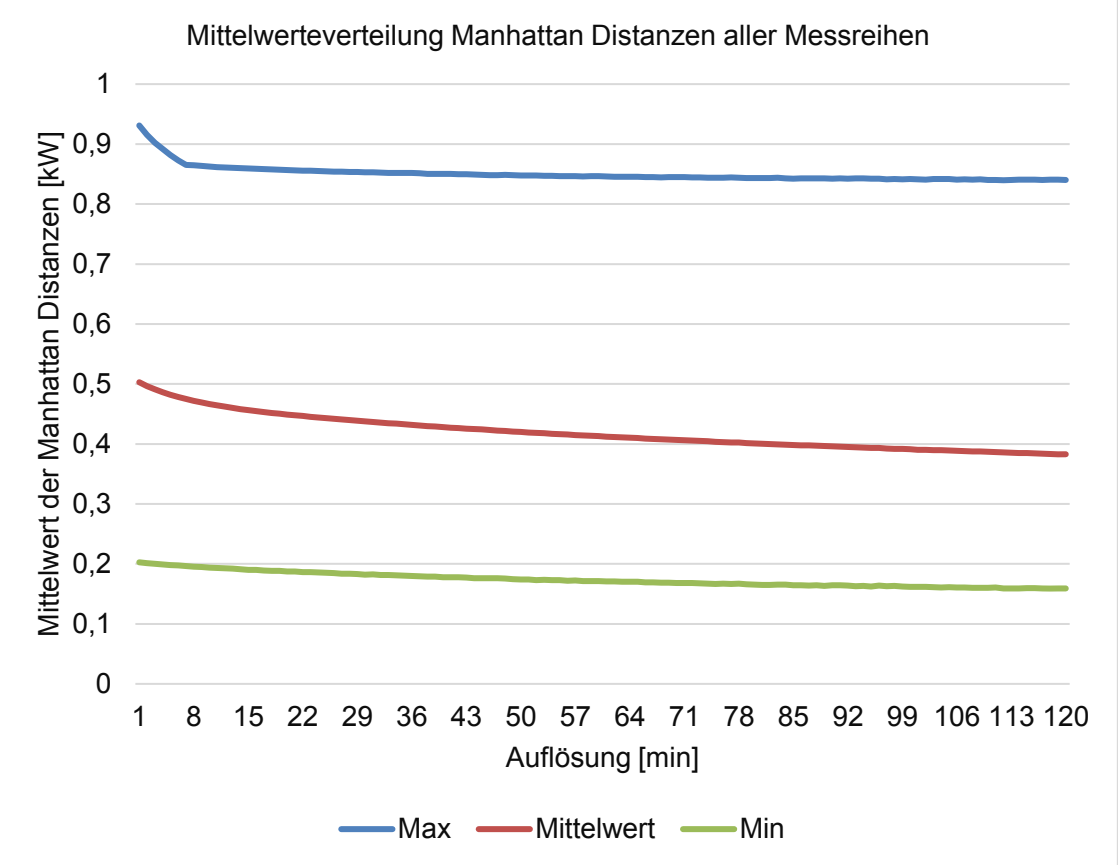

Abbildung 10: Mittelwerte der Manhattan Distanzen aller Messreihen

In der Abbildung 11 sind die Kurven der maximalen, durchschnittlichen und minimalen Standardabweichungen der Manhattan Distanzen der Datenpunkte aller beider Haushaltmessreihen in den Auflösungen von 1 Minute bis 120 Minuten aufgezeigt.

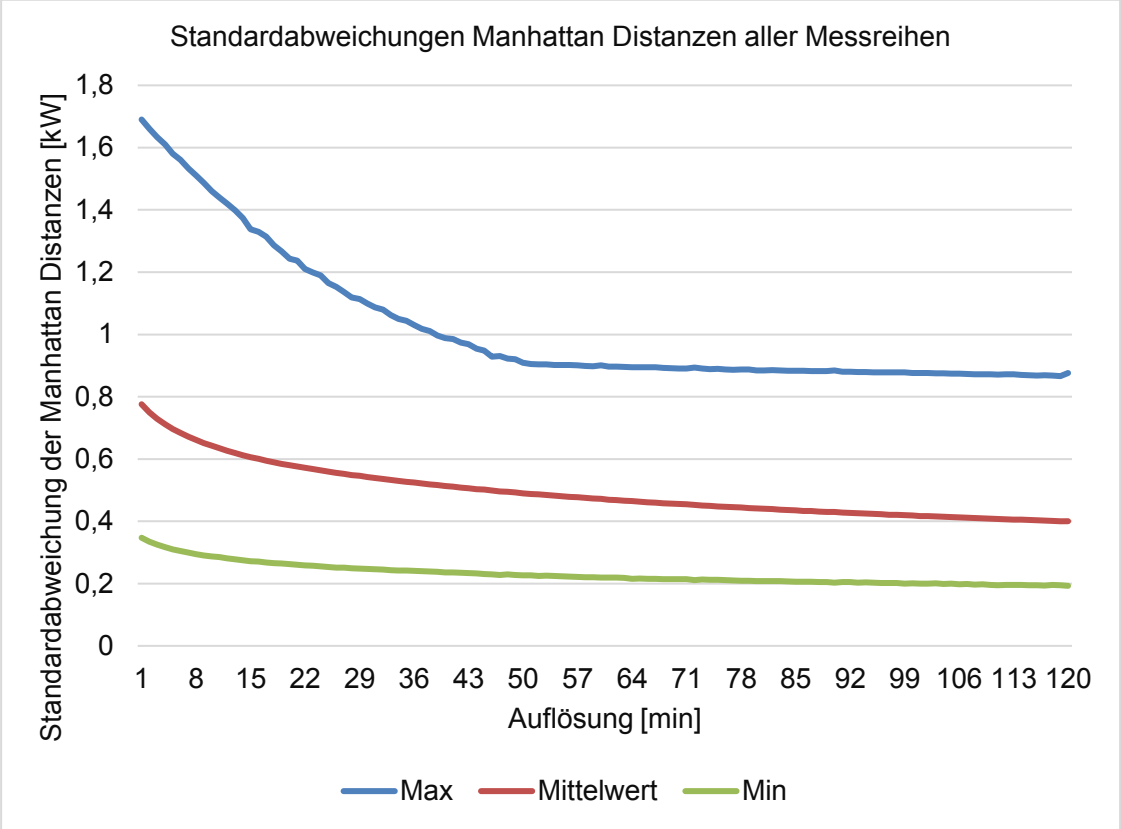

Abbildung 11: Standardabweichungen der Manhattan Distanzen aller Messreihen

Sowohl die Mittelwerte als auch die Standardabweichungen der Manhattan Distanzen der Datenpunkte aller beider Haushaltmessreihen weisen den fallenden Trend auf beim Verringern der Messauflösung. D.h. die zentrale Lage der einzelnen Datenpunktdistanzen fällt, und die einzelnen Punktdistanzen in den Manhattan Distanzreihen befinden sich allgemein näher an der zentralen Lage, wobei die Distanzen einzelner Datenpunkte in den Messreihen mehr konvergieren bei niedrigeren Auflösungen als bei höheren Auflösungen. Zwei Zeitreihen haben dieselbe Form, wenn die Manhattan Distanzen aller Datenpunkte dieser zwei Zeitreihen gleich sind. Daraus sind alle Haushaltmessreihen nicht nur räumlich zueinander näher bzw. in ihrer Struktur, sondern ähneln sich in der Form beim Verringern der Messauflösung.

4.3 Kreuzkorrelationsverteilung

In der Abbildung 12 sind die Kreuzkorrelationskurven der Messreihen von Haushalten 1&3, Haushalten 1&4 und Haushalten 3&4 zu den Zeitverschiebungen von 1 Minute bis 1000 Minuten demonstriert.

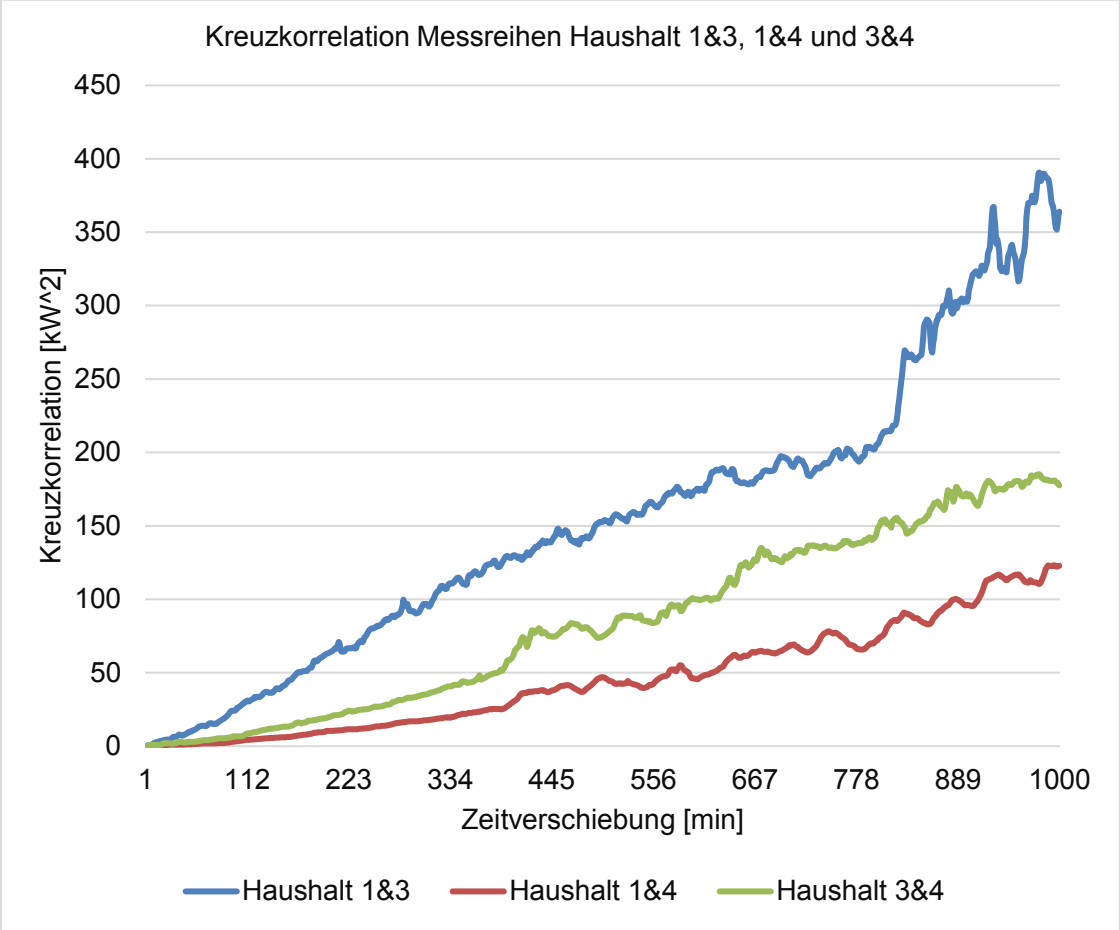

Abbildung 12: Kreuzkorrelation von Haushalten 1&3, 1&4 und 3&4

Beim Zunehmen der Zeitverschiebung weisen die drei Kurven einen steigenden Trend auf. Die Kreuzkorrelationskurve von Haushalten 1&3 dominiert die Kreuzkorrelationskurve von Haushalten 3&4, und die Kreuzkorrelationskurve von Haushalten 3&4 dominiert die Kreuzkorrelationskurve von Haushalt 1&4. D.h. die Zeitreihen von Haushalten 1 und 3 sind am ähnlichsten in allen drei Haushalten, dann folgen die Zeitreihen von Haushalt 3 und 4, die am zweitähnlichsten sind. Die Ähnlichkeit der Zeitreihen von Haushalten 1 und 4 ist am niedrigsten.

Die Ähnlichkeiten der Messreihen von Haushalten 1, 3 und 4 werden mithilfe der Kreuzkorrelation qualitativ beschrieben, weil eine tiefergehende bzw. quantitative Analyse der Anwendungen mit Kreuzkorrelation aus der Signalverarbeitung noch nicht untersucht wurde. Für sinnvolle Aussagen sollten verschiedene Formen der Normierung der Ähnlichkeitskriterien ausprobiert werden.

4.4 K-Means Clustern aller Messreihen

Nach der Bestimmung der durchschnittlichen Silhouette-Koeffizienten der Lastpunkte aller Haushaltzeitreihen bei Clusteranzahlen von 2 bis 30 ist das Ergebnis in der Abbildung 13 aufgezeigt. Der größte durchschnittliche Silhouette-Koeffizient ist bei Clusteranzahl 2, und der zweitgrößte ist bei Clusteranzahl 3. Wie im Kapitel 3.5 erläutert wird 3 als die Clusteranzahl für das K-Means Clustern definiert.

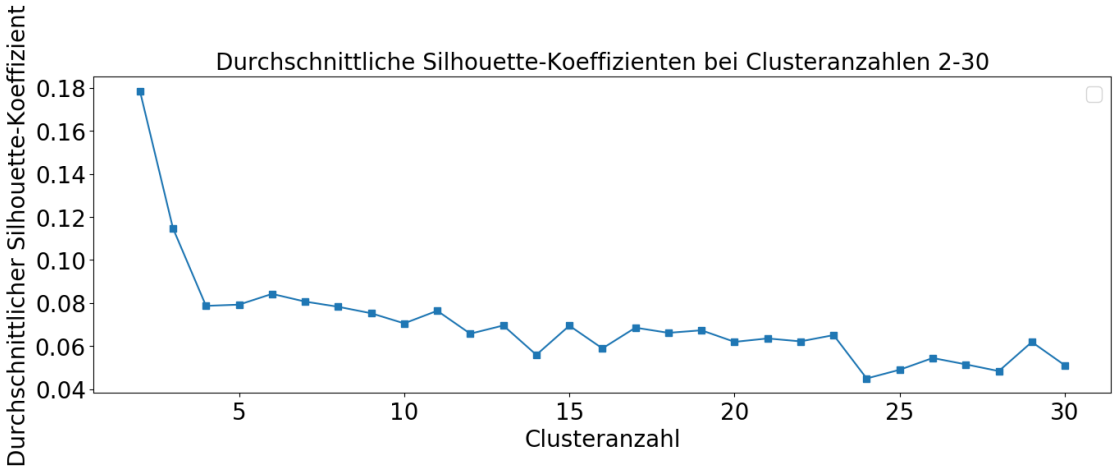

Abbildung 13: Durchschnittliche Silhouette-Koeffizienten

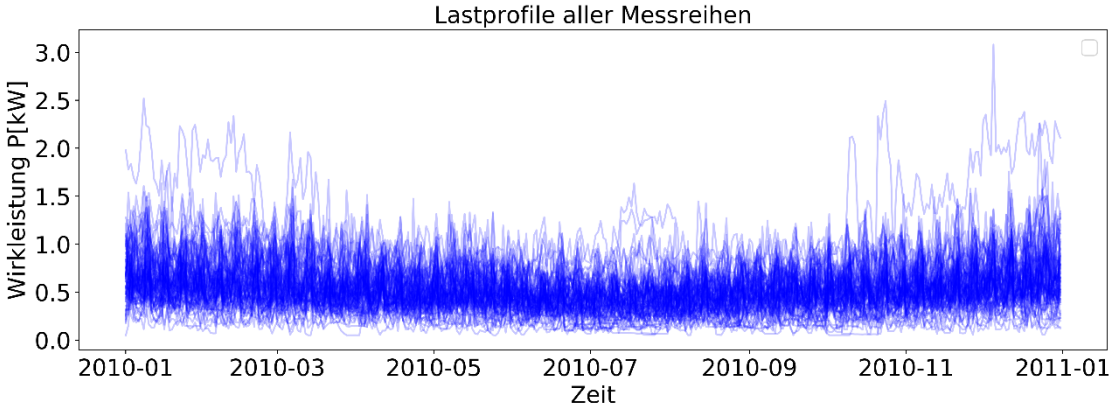

Abbildung 14: Lastprofile aller 74 Haushalte im Jahre 2010

Die Lastprofile aller 74 Haushalte stehen in der Abbildung 14. Um diese zu klassifizieren, wird der K-Means Algorithmus implementiert. Mithilfe der Klassifizierung können die Ähnlichkeitsmuster beim Stromverbrauch der Kunden und die Saisons mit Grund- und Spitzenlast bestimmt werden [Amr16].

Nach der Implementierung sind alle Lastpunkte in drei Clustern verteilt wie in der Abbildung 15 dargestellt. Die Punktfarben Blau, Rot und Grün stehen jeweils für ein Cluster.

Im Grünen Cluster sind die meisten Lastpunkte an den Sommertagen. Die Lastpunkte an den Sommertagen bilden die Grundlast, in der keine hohe Spitzenlast auftritt. Die besonders hohe Spitzenlast tritt im Blauen und Roten Cluster auf. Die meisten Spitzenlastpunkte des Blauen Clusters befinden sich am Oktoberende, Novemberende und Dezemberanfang, und die meisten Spitzenlastpunkte des Roten Clusters befinden sich am Januaranfang und Dezemberende. Im Februar und März treten die Spitzenlastpunkte vom Blauen und Roten Cluster gemischt auf.

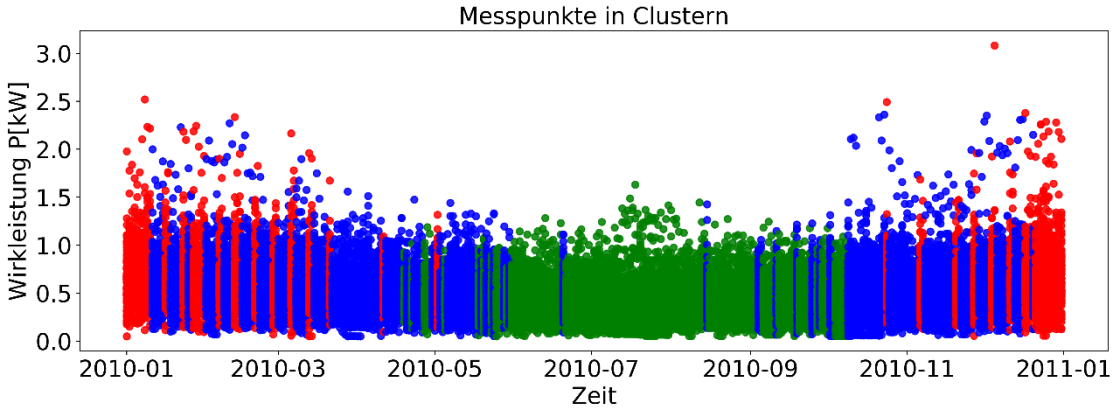

Abbildung 15: Lastprofile aller Haushalte in drei Clustern

5 Zusammenfassung und Ausblick

Beim Ausbau der erneuerbaren Energien ändert sich das Stromversorgungsmodell im Vergleich zur Stromversorgung aus den konventionellen Energiequellen. Im neuen Stromversorgungsmodell muss die Stromerzeugung nicht dem Stromverbrauch folgen, sondern die Stromerzeugung und der Stromverbrauch müssen angesichts der lokalen Stromverbrauchssituation harmonisiert werden. Der Einsatz der Smart Meter bietet eine Plattform um entsprechende Anwendungen umzusetzen. Die Smart Meter messen den Stromverbrauch echtzeitig zu verschiedenen Auflösungen. Diese gemessenen Daten werden zu den Stromversorgern übermittelt für weitere Analysen.

Die Smart Meter Daten verfügen über die Merkmale von Big Data. Dies ist charakteristisch durch die riesige Menge an Daten, hohe Erzeugungs- und Übermittlungsgeschwindigkeit, diverse Datenquellen und niedrige Informationsdichte definiert. Die Smart Meter Daten können für Beschreibungen, Prognosen und Entscheidungen analysiert werden.

Die Smart Meter Messreihen sind Zeitreihen. Die Ähnlichkeit in den Zeitreihen kann strukturbasierend oder formbasierend sein. Für die Suche nach den strukturbasierenden Ähnlichkeiten können die Korrelation, Kreuzkorrelation und Distanzmessung angewendet werden. Für die Suche nach den formbasierenden Ähnlichkeiten können die einzelnen Teilformen der Zeitreihen miteinander verglichen werden. Außerdem können die ähnlichen Datenpunkte in allen Messreihen gruppiert werden, um einen Überblick über die Ähnlichkeitsverteilung in den Messreihen zu erstellen.

Die strukturbasierende Ähnlichkeit der Messreihen hängt stark von der Messauflösung ab. Je niedriger die Messauflösung ist, desto höhere strukturbasierende Ähnlichkeit weisen die Messreihen auf. Die Lastpunkte können nach Grundlast und Spitzenlast in verschiedene Gruppe klassifiziert werden.

Dennoch sind die in dieser Arbeit durchgeführten exemplarischen Untersuchungen nicht tiefergehend. Zur vertiefenden Untersuchung können bei der Kreuzkorrelationsanalyse mithilfe der expliziten Anwendungsfälle der Kreuzkorrelation in der Signalverarbeitung die Ähnlichkeiten der Smart Meter Messreihen quantitativ dargestellt werden. Bei der Distanzmessung kann die dynamische Distanzmessung z.B. Dynamic Time Warping die klassische statistische Distanzmessung ersetzen, um

die genaueren und flexibleren strukturbasierenden Ähnlichkeiten der Messreihen darzustellen. Auch bei der Clusteranalyse kann die Untersuchung erweitert werden. Die Clusteranalyse gibt einen groben Überblick über die Klassifizierung der Lastprofile aller Haushalte über das ganze Jahr 2010 wieder. Es wäre übersichtlicher, wenn alle Lastprofile an einem Tag mit 24 Stunden dargestellt werden, damit ein typischer Tag mit der Grund-, Mittel- und Spitzenlast illustriert werden kann. Die Clusteranalyse gehört zum Unüberwachten Lernen des Maschinellen Lernens. Aus dem Überwachten Lernen können Methoden wie z.B. die Neural Network Classification und die Support Vector Machine für die Klassifizierung der Lastprofile effizient eingesetzt werden.

Literaturverzeichnis

[Ala13] Alahakoon, D.; Yu, X.: Advanced Analytics for Harnessing the Power of Smart Meter Big Data. In 2013 IEEE International Workshop on Inteligent Energy Systems (IWIES), 2013; S. 40–45.

[Ala16] Alahakoon, D.; Yu, X.: Smart Electricity Meter Data Intelligence for Future Energy Systems: A Survey. In IEEE Transactions on Industrial Informatics, 2016, 12; S. 425–436.

[Amr16] Amri, Y.; Fadhilah, A. L.; Fatmawati et al.: Analysis Clustering of Electricity Usage Profile Using K-Means Algorithm. In IOP Conference Series: Materials Science and Engineering 105 (2016) 012020, 2016.

[Asg17] Asghar, M. R.; Dán, G.; Miorandi, D.; Chlamtac, I.: Smart Meter Data Privacy: A Survey. In IEEE Communications Surveys & Tutorials, 2017, 19; S. 2820–2835.

[Aza14] Azad, S. A.; Ali, A. B. M. S.; Wolfs, P.: Identification of Typical Load Profiles using K-means Clustering Algorithm. In Asia-Pacific World Congress on Computer Science and Engineering, 2014.

[Ban15] Banerjee, S.; Choudhary, A.; Pal, S.: Empirical Evaluation of K-Means, Bisecting K-Means, Fuzzy C-Means and Genetic K-Means Clustering Algorithms. In 2015 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), 2015; S. 168–172.

[Bar15] Barai, G. R.; Krishnan, S.; Venkatesh, B.: Smart Metering and Functionalities of Smart Meters in Smart Grid - A Review. In 2015 IEEE Electrical Power and Energy Conference (EPEC), 2015; S. 138–145.

[Ber94] Berndt, D. J.; Clifford, J.: Using Dynamic Time Warping to Find Patterns in Time Series. In Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, 1994; S. 359–370.

[Bha19] Bhattarai, B. P.; Paudyal, S.; Luo, Y. et al.: Big Data Analytics in Smart Grids: State-of-the-Art, Challenges, Opportunities, and Future Directions. In IET Smart Grid, 2019, 2; S. 141–154.

[Bic08] Bickel, S.: Zeitreihenanalyse. https://www.cs.uni-potsdam.de/ml/teaching/ws08/zeitreihenanalyse.

[Brü18] Bründlinger, T.; König, J. E.; Frank, O. et al.: dena-Leitstudie Integrierte Energiewende. Impulse für die Gestaltung des Energiesystems bis 2050, Berlin/Köln, 2018.

[Che16] Chen, D.; Iyengar, S.; Irwin, D.; Shenoy, P.: SunSpot: Exposing the Location of Anonymous Solar-powered Homes. In Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments, 2016.

[Deh10] Dehariya, V. K.; Shrivastava, S. K.; Jain, R. C.: Clustering of Image Data Set Using K-Means and Fuzzy K-Means Algorithms. In 2010 International Conference on Computational Intelligence and Communication Networks 2010 International Conference on Computational Intelligence and Communication Systems, 2010; S. 386–391.

[Dei18] Deistler, M.; Scherrer, W.: Modelle der Zeitreihenanalyse. Springer International Publishing, Cham, 2018.

[Deu14] Deutsche Energie-Agentur GmbH (dena) Hrsg.: Einführung von Smart Meter in Deutschland. Analyse von Rolloutszenarien und ihrer regulatorischen Implikationen. (kurz: dena-Smart-Meter-Studie), Berlin, 2014.

[Dud18] Dudek, G.; Gawlak, A.; Kornatka, M.; Szkutnik, J.: Analysis of Smart Meter Data for Electricity Consumers. In 2018 15th International Conference on the European Energy Market (EEM), 2018.

[Est13] Esteves, R. M.; Hacker, T.; Rong, C.: Competitive K-Means. A new accurate and disributed k-means algorithm for large datasets. In 2013 IEEE International Conference on Cloud Computing Technology and Science, 2013; S. 17–24.

[Gel15] Geler, Z.: Role of Similarity Measures in Time Series Analysis. Doctoral Dissertation, Novi Sad, 2015.

[He17] He, D.; Kumar, N.; Zeadally, S. et al.: Efficient and privacypreserving data aggregation scheme for smart grid against internal adversaries. In IEEE Transactions on Smart Grid, 2017, 8; S. 2411–2419.

[Kam93] Kamel, I.; Faloutsos, C.: Hilbert R-tree: An improved R-tree using fractals. Technical Research Report, 1993.

[Keo04] Keogh, E.; Ratanamahatana, C. A.: Exact indexing of dynamic time warping. In Knowledge and Information Systems, 2004.

[Kim11] Kim, Y.-I.; Kang, S.-J.; Ko, J.-M.; Choi, S.-H.: A study for clustering method to generate Typical Load Profiles for Smart Grid. In 8th International Conference on Power Electronics - ECCE Asia, 2011; S. 1102–1109.

[Klä16] Kläs, M.; Putz, W.; Lutz, T.: Quality evaluation for big data. a scalable assessment approach and first evaluation results. In 2016 Joint Conference of the International Workshop on Software Measurement and the International Conference on Software Process and Product Measurement (IWSM-MENSURA), 2016; S. 115–124.

[Kot02] Kothuri, R. K. V.; Ravada, S.; Abugov, D.: Quadtree and r-tree indexes in oracle spatial: a comparison using GIS data. In Proceedings of the 2002 ACM SIGMOD international conference on Management of data, 2002; S. 546–557.

[Kov99] Kovács, L.; Terstyánszky, G. Z.: Diagnosising faults by supervised and unsupervised learning. In 1999 European Control Conference (ECC), 1999; S. 4238–4242.

[Lei98] Leiner, B.: Grundlagen der Zeitreihenanalyse. R. Oldenbourg Verlag, München, Wien, 1998.

[Li16] Li, R.; Li, F.; Smith, N. D.: Multi-Resolution Load Profile Clustering for Smart Metering Data. In IEEE Transactions on power systems, 2016, 31; S. 4473–4482.

[Lin09] Lin, J.; Li, Y.: Finding Structural Similarity in Time Series Data Using Bag-of-Patterns Representation. In International Conference on Scientific and Statistical Database Management: SSDBM 2009, 2009; S. 461–477.

[Lip] Lippe, P. von der: Einführung in die Zeitreihenanalyse: Deskriptive Statistik; S. 393–436.

[Liu17a] Liu, H.; Huang, F.; Li, H. et al.: A Big Data Framework for Electric Power Data Quality Assessment. In 2017 14th Web Information Systems and Applications Conference, 2017a; S. 289–292.

[Liu17b] Liu, Q.; Zhu, B.; Li, Q.: Impact of Big Data on Electric-power Industry. In 2017 IEEE 2nd Internatinal Conference on Big Data Analysis, 2017b; S. 460–463.

[Mar13] Martins, A. D.; Gurjao, E. C.: Processing of smart meters data based on random projections. In 2013 IEEE PES Conference on Innovative Smart Grid Technologies (ISGT Latin America), 2013.

[Mar19] Marlen, A.; Maxim, A.; Ukaegbu, I. A.; Kumar Nunna, H. S. V. S.: Application of Big Data in Smart Grids: Energy Analytics. In International Conference on Advanced Communications Technology(ICACT), 2019; S. 402–407.

[Nak12] Nakamura, T.; Taki, K.; Nomiya, H. et al.: A Shape-based Similarity Measure for Time Series Data with Ensemble Learning. In Pattern Analysis and Applications, 2012, 16.

[Paw17] Pawar, S.; Momin, B. F.: Smart Electricity Meter Data Analytics: A Brief Review. In 2017 IEEE Region 10 Symposium (TENSYMP), 2017.

[Qi16] Qi, J.; Hahn, A.; Lu, X. et al.: Cybersecurity for distributed energy resources and smart inverters. In IET Cyber-Physical Systems: Theory & Applications, 2016, 1; S. 28–39.

[Ras18] Rashid, M. H.: AMI Smart Meter Big Data Analytics for Time Series of Electricity Consumption. In 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering, 2018; S. 1771–1776.

[Sar18] Saravanan, R.; Sujatha, P.: A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning Approaches in Data Classification. In the Second InternationalConference on Intelligent Computing and Control Systems (ICICCS 2018), 2018; S. 945–949.

[Sch01] Schlittgen, R.; Streitberg, B. H.J.: Zeitreihenanalyse. R. Oldenbourg Verlag, München, Wien, 2001.

[Sch10] Schneider, M.; Mentemeier, S.: Zeitreihenanalyse mit R. https://www.uni-muenster.de/Stochastik/lehre/SS10/BlockprakZeit/Zeitreihenanalyse.pdf.

[Sci12] Sciacca, S.: ‘Big data’ and the need for improved time synchronization standards. Without time synchronization, vast streams of data have limited use. https://www.csemag.com/articles/big-data-and-the-need-for-improved-time-synchronization-standards/.

[Sod16] Sodenkamp, M.; Hopf, K.; Kozlovskiy, I.; Staake, T.: Smart-Meter-Datenanalyse für automatisierte Energieberatungen. ("Smart Grid Data Analytics"), Bamberg, 2016.

[Spi15] Spiegel, S.: Time Series Distance Measures. Segmentation, Classification, and Clustering of Temporal Data. Doktorarbeit, Berlin, 2015.

[Sub15] Subramanian, D. V.; Pradheepkumar, K.; Dhinakaran, K.; Duraimurugan: Catur approach to assess the quality of big data using decision tree and multidimensional model. In Australian Journal of Basic and Applied Sciences, 2015, 9; S. 503–508.

[Sul19] Sulaiman, S. M.; Jeyanthy, P. A.; Devaraj, D.: Smart Meter Data Analysis Issues: A Data Analytics Perspective. In 2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), 2019.

[Tay98] Tayeb, J.; Ulusoy, Ö.; Wolfson, O.: A quadtree-based dynamic attribute indexing method. In The Computer Journal, 1998, 41; S. 185–200.

[Tja] Tjaden, T.; Bergner, J.; Weniger, J.; Quaschning, V.: Repräsentative elektrische Lastprofile für Wohngebäude in Deutschland auf 1-sekündiger Datenbasis, Berlin.

[Tsa09] Tsang, S.; Kao, B.; Yip, K. Y. et al.: Decision trees for uncertain data. In 2009 IEEE 25th International Conference on Data Engineering, 2009; S. 441–444.

[Vio18] Viola, L. G.: Clustering electricity usage profiles with K-means. https://towardsdatascience.com/clustering-electricity-profiles-with-k-means-42d6d0644d00.

[Wag12] Wagstaff, K. L.: Machine Learning that Matters, California, 2012.

[Wan20] Wang, Y.; Chen, Q.; Kang, C.: Smart Meter Data Analytics. Electricity Consumer Behavior Modeling, Aggregation and Forecasting. Science Press and Springer Nature Singapore Pte Ltd. 2020, 2020.

[Xie14] Xie L.; Chen, Y.; Kumar, P. R.: Dimensionality Reduction of Synchrophasor Data for Early Event Detection: Linearized Analysis. In IEEE Transactions on power systems, 2014, 29; S. 2784–2794.

[Yan14] Yang, Y.; Bi, Z.: Advances and Future Challenges in Electric Power Big Data. In 2014 Second International Conference on Advanced Cloud and Big Data, 2014; S. 213–219.

[Yu16] Yu, W.-S.; Fang, Y.-J.: Data analysis of the smart meters and its applications in Tatung University. In 2016 International Conference on Fuzzy Theory and Its Applications (iFuzzy), 2016.

[Zho16] Zhou, D.; Guo, J.; Zhang, Y. et al.: Distributed Data Analytics Platform for Wide-Area Synchrophasor Measurement Systems. In IEEE Transactions on Smart Grid, 2016, 7; S. 2397–2405.

[Zom12] Zomaya, A. Y.; Lee, Y. C.: Energy efficient distributed computing systems. John Wiley & Sons, Inc., Hoboken, 2012.