Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Abschlussarbeit

Sascha Rödel

Erkennung von täglichen Verbrauchsmustern aus Energieverbrauchsdaten häuslicher Fahrzeugladestationen zur Bestimmung der Wirkungsweise dynamischer Preisberechnung

Fakultät Technik und Informatik
Department Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Sascha Rödel

Erkennung von täglichen Verbrauchsmustern aus Energieverbrauchsdaten häuslicher Fahrzeugladestationen zur Bestimmung der Wirkungsweise dynamischer Preisberechnung

Abschlussarbeit eingereicht im Rahmen der Bachelorprüfung

im Studiengang Wirtschaftsinformatik
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuender Prüfer : Prof. Dr. Ulrike Steffens
Zweitgutachter : Prof. Dr.-Ing. Sebastian Rohjans

Sascha Rödel

Thema der Arbeit

Erkennung von täglichen Verbrauchsmustern aus Energieverbrauchsdaten häuslicher Fahrzeugladestationen zur Bestimmung der Wirkungsweise dynamischer Preisberechnung

Stichworte

Energieverbrauch, Elektrische Fahrzeuge, Häusliche Ladestation, Clustering, Data Mining, CRISP-DM, Unüberwachtes Lernen, Klassifizierung, Verbrauchsmuster, Dynamische Preisberechnung

Kurzzusammenfassung

Aktuell beruht der Treibstoff für den Transportsektor größtenteils auf fossilen Brennstoffen und ist deshalb für einen signifikanten Anteil der Treibhausgase verantwortlich. Sowohl rein elektrische - als auch teilelektrische Fahrzeuge (Hybridfahrzeuge) können diese Treibhausgase drastisch reduzieren und werden in Zukunft einen hohen Stellenwert für die Gesellschaft einnehmen. Elektrische Fahrzeuge können im Gegensatz zu konventionellen Fahrzeugen auch in häuslichen Umgebungen durch entsprechende Ladestationen aufgeladen werden. Aus einem hohen Durchdringungsniveau durch elektrische Fahrzeuge in Verbindung mit einem unkontrollierten Konsumverhalten resultiert ein starker Anstieg des Spitzenlastbedarfes. Dieser kann resultierend durch einen Ausbau der Stromversorgung zu höheren Strompreisen oder Ausfällen in der Energieversorgung führen. Ziel dieser Arbeit ist es herauszufinden, ob eine dynamische Preisberechnung einen positiven Einfluss auf das Ladeverhalten der Autofahrer hat. Hierfür werden anfallende Daten gesammelt und anschließend durch Data Mining verarbeitet und analysiert. Durch die Erkennung von Verbrauchsmustern auf Basis täglicher Zeitreihen kann der Einfluss einer dynamischen Preisberechnung auf das Nutzungsverhalten überprüft werden. Diese Arbeit kommt zu dem Schluss, dass Verbraucher, die an einer dynamischen Preisberechnung partizipieren, deutlich geringere relative Häufigkeiten an Verbrauchsmustern zu Spitzenlastzeiten aufweisen. Es ist deshalb von einem positiven Einfluss durch dynamische Preisberechnung auf das Nutzungsverhalten auszugehen.

Sascha Rödel

Title of the paper

Recognition of daily use patterns from energy consumption data of residential vehicle charging stations to determine the effect of dynamic pricing

Keywords

Energy Consumption, Electric Vehicles, Residential Vehicle Charging Station, Clustering, Data Mining, CRISP-DM, Unsupervised Learning, Classification, Usage Pattern, Dynamic Pricing

Abstract

Currently the fuel for the transport sector is largely based on fossil fuels and therefore accounts for a significant part of greenhouse emissions. Both pure electric and partially electric vehicles (hybrid vehicles) can drastically reduce these greenhouse gases and will have a high value for the society in the future. In contrast to conventional vehicles, electric vehicles can also be charged by charging stations in residential environments. A high level of penetration by electric vehicles combined with uncontrolled consumption behavior results in a sharp increase in the peak load demand. This can result in a buildout in the power supply at higher current prices or could create shortages of electric power. The aim of this thesis is to determine, whether dynamic pricing has a positive influence on the charging behavior of drivers. For this purpose, accumulated data are collected and then processed and analysed by data mining. Through the recognition of consumption patterns on the basis of daily time series, the influence of dynamic pricing on the consumption behaviour can be examined. This work concludes that consumers who participate in a dynamic pricing trial have significantly lower relative frequencies of consumption patterns at peak load times. Therefore a positive influence of dynamic pricing on the consumption behaviour must be assumed.

Inhaltsverzeichnis

I. Abbildungsverzeichnis

II. Tabellenverzeichnis

1 Einleitung

1.1 Problemstellung

Der Transportsektor beruht aktuell noch größtenteils auf fossilen Brennstoffen und ist deshalb für einen signifikanten Anteil der Treibhausgase verantwortlich.[1] Personenkraftwagen sind wesentlich für die Nutzung fossiler Brennstoffe verantwortlich und benötigen mehr als 80 Prozent der Transportenergie im Personenverkehr.[2] Unter der Annahme, dass der Anteil erneuerbarer Energien stark steigen wird, stellen sowohl elektrische als auch Hybridfahrzeuge eine der wichtigsten zukünftigen Technologien dar, um Treibhausgase drastisch zu reduzieren.

Zahlreiche Modelle elektrischer Fahrzeuge werden aktuell von unterschiedlichen Herstellern produziert.[3] Die Reichweite dieser Fahrzeuge liegt größtenteils noch unter 65 Kilometern, wobei die Nennleistung zwischen einigen 10 Kilowatt für kleinere Fahrzeuge bis zu mehreren 100 Kilowatt für Performance-Fahrzeuge liegt.

Der Markt für elektrische Fahrzeuge ist aktuell noch sehr limitiert. Insbesondere durch neue Technologien, wie Akkumulatoren mit einer sehr hohen Leistungsdichte, ist mit einem großen Wachstum zu rechnen. Wegen stark fallender Lithium-Ion-Preise rechnet der Informationsdienstleister Bloomberg L. P. mit einer Marktpenetration

[1] Vgl. Putrus G. A. et al.(2009)
[2] Vgl. Umweltbundesamt (2012), S. 30
[3] Vgl. Todd J. (2013)

durch elektrische Fahrzeuge von 35% im Jahr 2040.[4] Dieses Wachstum wird einen deutlichen Einfluss auf das öffentliche Stromnetz haben.[5] Dies ist unter anderem auf die Tatsache zurückzuführen, dass elektrische Fahrzeuge im Gegensatz zu konventionellen Fahrzeugen auch in häuslichen Umgebungen durch entsprechende Ladestationen aufgeladen werden können. Adam Ruder, ein Projektmanager der New York State Energy Research, schreibt hierzu:

> „If PEV charging coincides with peak demand from other sources of peak load, the results may include the need for upgrade to the distribution system, the need for new transmission to relieve congestion, higher wholesale market prices, and a greater reliance on less efficient peaking units.“[6]

Das Nutzerverhalten der Autofahrer elektrischer Fahrzeuge spielt also eine wichtige Schlüsselrolle in Bezug auf die Netzstabilität und Netzinfrastruktur. Erhöhte Investitionen der Energieversorger in die Netzinfrastruktur spiegeln sich direkt in den Strompreisen wider. Aus einem hohen Durchdringungsniveau durch elektrische Fahrzeuge in Verbindung mit einem unkontrollierten Konsumverhalten resultiert ein starker Anstieg des Spitzenlastbedarfes. Dieser kann die Netzstabilität gefährden und zu Ausfällen in der Stromversorgung führen.[7] Es lassen sich also neben den Energieversorgern auch die Verbraucher als Stakeholder in Bezug auf diese Thematik identifizieren.

Ein Ansatz, den Spitzenlastbedarf zu reduzieren und einen Teil des Strombedarfs außerhalb der Spitzenlastzeit zu verlagern, ist die dynamische Preisberechnung. Bei dieser Form der Preisstrategie variiert der Preis für Elektrizität abhängig von der Uhrzeit des Tages, wobei der Strompreis zu Spitzenlastzeiten teurer und außerhalb dieser Zeiten günstiger ist.[8] Eine zeitbasierte, dynamische Preisabrechnung kann hier

[4] Vgl. MacDonald J.(2016)
[5] Vgl. Villafafila-Robles R. et al.(2013)
[6] Vgl. Ruder A. (2015), S. 5
[7] Vgl. Salameh Z. (2015), S. 111
[8] Vgl. Ruder A. (2015), S. 11

einen Anreiz für den Verbraucher darstellen, sein Nutzungsverhalten zu ändern, weil dieser günstigeren Strom beziehen kann.

1.2 Zielsetzung

Ziel dieser Arbeit ist es herauszufinden, ob eine dynamische Preisberechnung einen positiven Einfluss auf das Ladeverhalten der Autofahrer elektrischer Fahrzeuge hat. Hierfür werden anfallende Energiemessdaten durch Smart Meter gesammelt und anschließend zu täglichen Zeitreihen zusammengesetzt. Diese Zeitreihen geben den jeweiligen Messwert einer Ladestation im 15-Minuten-Takt über 24 Stunden wieder.

hour		00:00	00:15	00:30	00:45	01:00	01:15	01:30	01:45
dataid	date								
1642	2013-10-01	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0
	2013-10-02	0.4	0.0	0.0	0.0	0.0	0.0	0.00	0.0
	2013-10-03	0.0	0.0	0.0	0.0	0.0	0.2	0.19	0.0

hour		02:00	02:15	...	21:30	21:45	22:00	22:15	22:30
dataid	date			...					
1642	2013-10-01	0.0	0.0	...	0.00	0.01	0.01	0.00	0.00
	2013-10-02	0.0	0.0	...	0.01	0.01	0.00	0.26	0.17
	2013-10-03	0.0	0.0	...	0.01	0.01	0.00	0.00	0.00

hour		22:45	23:00	23:15	23:30	23:45
dataid	date					
1642	2013-10-01	0.0	0.01	0.0	0.00	0.01
	2013-10-02	0.0	0.00	0.0	0.00	0.00
	2013-10-03	0.0	0.00	0.0	0.46	0.00

Abbildung 1: Exemplarischer Aufbau täglicher Zeitreihen

Die Daten werden von der Entwicklungs- und Forschungsorganisation PecanStreet Inc. mittels einer PostgreSQL-Datenbank zur Verfügung gestellt. Die Zeitreihen können anschließend entsprechend ihrer Distanz zueinander in verschiedene Cluster eingeteilt werden. Gruppierungen ähnlicher Zeitreihen können somit identifiziert und in Form eines Verbrauchsmusters ausgedrückt werden. Das Verbrauchsmuster ist ebenfalls eine Zeitreihe und wird durch den Mittelwert aller Zeitreihen eines Clusters gebildet. Hohe Lastspitzen innerhalb eines Tages können so erkennbar gemacht werden. Die folgende Abbildung zeigt den exemplarischen Aufbau von sechs Verbrauchsmustern:

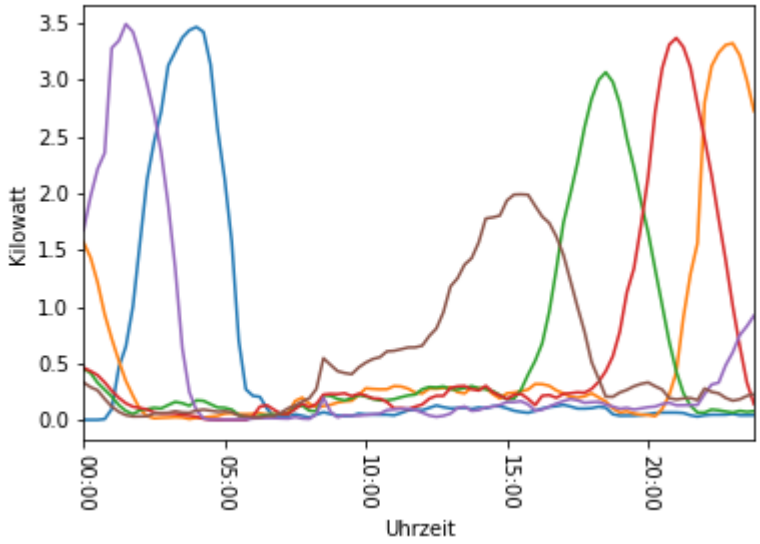


Abbildung 2: Exemplarischer Aufbau von sechs Verbrauchsmustern.

Die Erkennung der Verbrauchsmuster findet sowohl für Verbraucher statt, die durch eine dynamische Preisberechnung einen wirtschaftlichen Anreiz erhalten, ihren Strom außerhalb der Spitzenlastzeit zu beziehen, als auch solche, die keinen wirtschaftlichen Anreiz erhalten. Unterschiede in den Verbrauchsmustern dieser beiden Verbraucherklassen können somit sichtbar und der Einfluss dynamischer Preisberechnung auf die Verbrauchsmuster messbar gemacht werden. Um diesen Einfluss beurteilen zu können, werden die einzelnen Verbrauchsmuster abhängig von ihrer Erscheinungsform in Bezug auf die Netzstabilität als kritisch oder unkritisch klassifiziert. Das geschieht abhängig davon, ob das Verbrauchsmuster einen hohen Energieverbrauch zu Spitzenlastzeiten verzeichnet. In solchen Fällen ist es möglich, den Verbraucher automatisch zu benachrichtigten, wenn ein kritisches Verbrauchsmuster auftritt. Neben der Erkennung von Verbrauchsmustern und der Bestimmung des Einflusses einer dynamischen Preisberechnung auf die Muster, werden in dieser Arbeit die folgenden Fragen beantwortet:

- Wie viele Verbrauchsmuster treten unter dem Einfluss dynamischer Preisberechnung auf? Wie viele ohne diesen Einfluss?
- Wie sind die Muster abhängig vom Einfluss verteilt?
- Wie sind die Muster zu bewerten?

- Wie hoch ist der relative Anteil an kritischen Ladephasen aller Verbrauchsmuster im Verhältnis zur Gesamtzahl der Ladephasen unter dem Einfluss dynamischer Preisberechnung? Wie hoch ist dieser Anteil ohne diesen Einfluss?

1.3 Gang der Untersuchung

Die folgenden Schritte werden für die Realisierung der Zielsetzung ausgeführt. Zunächst wird in Kapitel 1.4 der Bezugsrahmen dieser Arbeit grafisch dargestellt. Dieser gibt an, wie die einzelnen Kapitel dieser Arbeit zueinander in Beziehung stehen. Der Stand der Forschung in Kapitel 2 repräsentiert aktuelle Forschungsarbeiten, welche für diese Arbeit relevant sind. In Kapitel 3 werden die technischen Grundlagen der Clusteranalyse erläutert, welche für die Durchführung der Datenanalysen nötig sind. Anschließend wird in Kapitel 4 das Prozessmodell CRISP-DM beschrieben. Dieser Entwurf beschreibt die einzelnen Schritte in ihrer Reihenfolge, um die Datenanalysen aus Kapitel 5 durchzuführen. In Kapitel 5 werden die Datenanalysen durchgeführt, dokumentiert und die Ergebnisse präsentiert. Schlussendlich enthält Kapitel 6 eine Zusammenfassung der gesamten Arbeit.

1.4 Bezugsrahmen

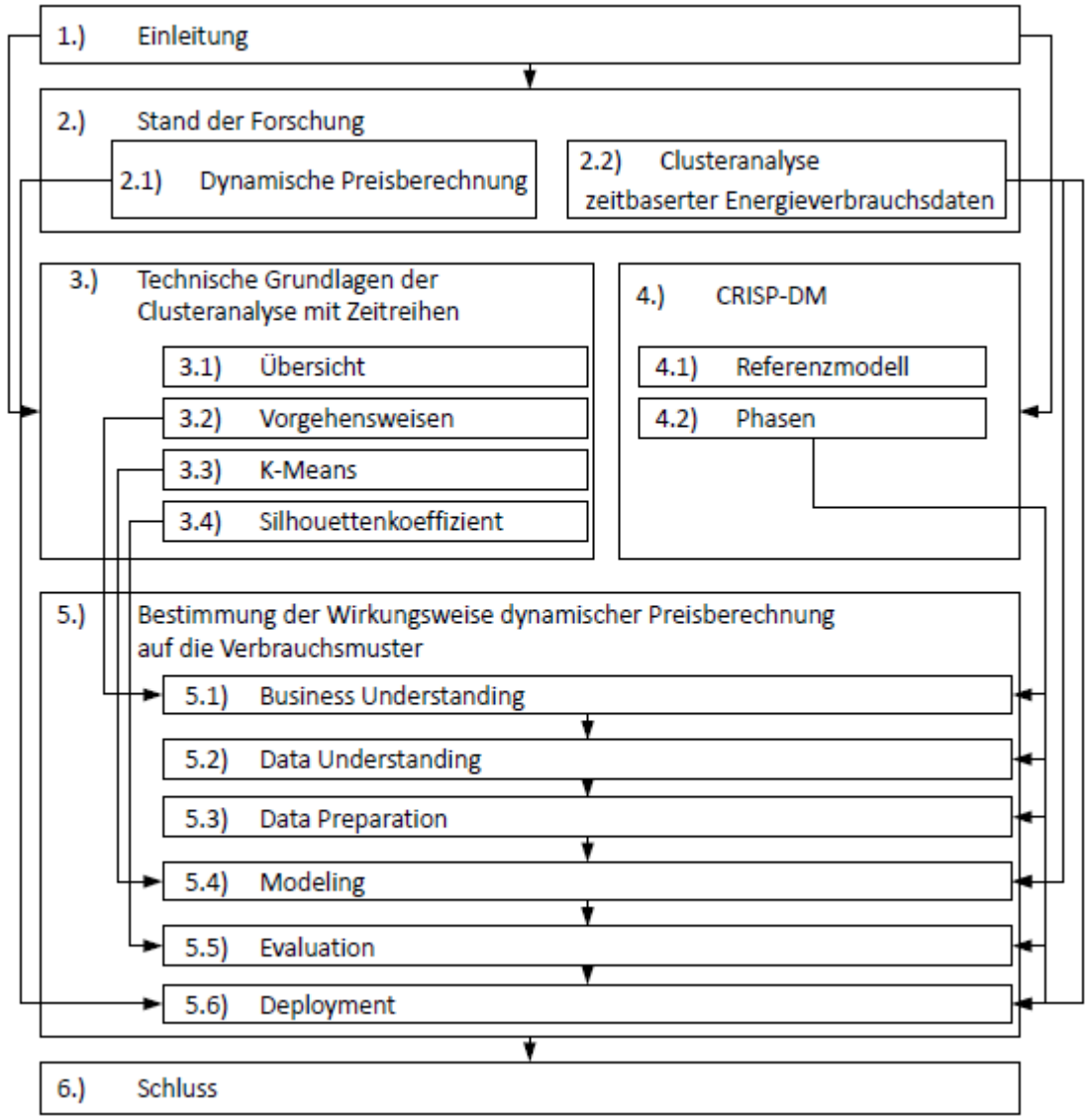


Abbildung 3: Bezugsrahmen

2 Stand der Forschung

2.1 Dynamische Preisberechnung

Elektrizität ist nicht kostensparend speicherbar und ihre Produktion ist abhängig von kurzfristigen Kapazitätsbeschränkungen. Die Nachfrage hingegen ist hoch dynamisch. Der Großhandelspreis, welcher Angebot und Nachfrage reflektiert, variiert. Der Endkundenpreis hingegen ist typischerweise über mehrere Monate konstant. Dieser reflektiert also nicht den sich stündlich ändernden Großhandelspreis und damit auch nicht die schwankende Nachfrage.[9] Diese Mechanik war eine der Ursachen für die kalifornische Energiekrise im Jahr 2000 – 2001. Neben mehreren großflächigen Zusammenbrüchen des Stromnetzes und der Insolvenz eines Energiekonzerns, wurden viele Industrieunternehmen geschädigt, welche von einer zuverlässigen Stromversorgung abhängig waren.[10] Dieses Ereignis war Anlass vieler Versuche, um herauszufinden, inwieweit Verbraucher ihren Bedarf reduzieren, wenn sie an einem variablen Tarif partizipieren. Faruqui et al. (2009) werteten hierzu insgesamt 15 Feldstudien aus und berechneten die Substitutionselastizität. Diese gibt die relative Änderung der Nachfrage im Anschluss an eine Preisänderung an. Sie kamen zu dem Schluss, dass die Bedarfsreduzierung von wenig bis substantiell variiert. Dies war davon abhängig, welches Preismodell eingesetzt wurde, wie groß die Preisschwankungen waren und ob die Verfügbarkeit einer Enabling-Technologie, wie beispielsweise programmierbarer Thermostate oder intelligenter Netzkomponenten, gegeben ist. In jedem Fall bilden Anreizprogramme somit den Schlüssel, um sowohl die Spitzenlast zu verringern[11], als auch Preisschwankungen der Großmarktpreise

[9] Vgl. Borenstein S. et al. (2002), S. 5

[10] Vgl. Sweeney J. (2002), S. 28

[11] Vgl. Aswin Raj. C. (2015), S. 122

abzumildern. Es wurden keine Arbeiten gefunden, welche die Wirkungsweise dynamischer Preisberechnung durch die Erzeugung von Lastprofilen mittels Clusteranalyse erforscht haben.

2.2 Clusteranalyse zeitbasierter Energieverbrauchsdaten

Die erhöhe Verbreitung von Automated Meter Reading – Technologie (AMR) sowie die damit verbundene Entstehung intelligenter Netze, bietet ein großes Volumen an Sensor- und Messdaten.[12] Daten, die mit einem Zeitstempel versehen sind, können zu Zeitreihen beliebiger Länge zusammengesetzt werden. Die Clusteranalyse vollständiger Zeitsequenzen wurde in der Forschung oft zur Erkennung von Lastprofilen verwendet. Figueiredo et al. (2005) erzeugten ein Framework zur Klassifizierung von Kunden basierend auf dem jeweiligen Lastprofil. Dent et al. (2011), Abreu et al. (2012) beschrieben eine Methode zur Definition von Lastprofilen für private Stromverbraucher, um Ähnlichkeiten zwischen den Arten der Verbraucher zu bestimmen. Hernández et al. (2012) erkannten Nutzungsprofile in Industrieparks mittels einer selbstorganisierenden Karte (engl. SOM). Momtazpour et al. (2012) erzeugten Lastprofile zur Standortbestimmung von öffentlichen Ladestationen. Meshram et al. (2013) erzeugten Lastprofile zur Unterstützung der Ressourcenplanung. Sie nutzten hierfür den Algorithmus K-Means und konnten insgesamt fünf tägliche Muster erkennen. Iglesias et al (2013) erzeugten Lastprofile, um verschiedene Distanzfunktionen miteinander zu vergleichen. Die Euklidische Distanz erzielte die besten Ergebnisse. Robinson et al. (2013) identifizierten Lastprofile englischer Autofahrer und überprüften, wie diese Lastprofile bei öffentlichen und privaten Ladestationen verteilt waren. Xydas et al. (2016) erstellten ein Risikomodell für öffentliche Ladestationen elektrischer Fahrzeuge. Als Parameter für die Berechnung der Risikostufe wurden ebenfalls Lastprofile genutzt. Ein zweiter Ansatz in der Erkennung von Lastprofilen besteht darin, den Lernalgorithmus nicht direkt auf die

[12] Vgl. Popeană (2015), S. 14

Zeitreihen anzuwenden.[13] Hierbei werden zunächst neue Variablen aus den Verbrauchsdaten abgeleitet und anschließend als Eingabe für den Lernalgorithmus verwendet. Die Reduzierung der Attribute führt zu einer schnelleren Datenverarbeitung und macht den Lernalgorithmus weniger anfällig für fehlende Werte. Die abgeleiteten Variablen müssen allerdings mit Bedacht ausgewählt werden, um der Komplexität der Verbrauchsdaten gerecht zu werden. Die Schwerpunkte der Cluster bilden außerdem nicht mehr den entsprechenden Mittelwert der Verbrauchsdaten ab und können nicht mehr als Verbrauchsmuster interpretiert werden.

[13] Vgl. Haben et al. (2015), S. 4

3 Technische Grundlagen der Clusteranalyse mit Zeitreihen

3.1 Übersicht

Das Ziel der Clusteranalyse ist es, Strukturen in einem ungekennzeichneten Datensatz zu identifizieren, indem die Daten in homogenen Gruppen organisiert werden. Die Ähnlichkeit der Objekte innerhalb einer Gruppe soll maximiert werden, während die Ähnlichkeit der Objekte zwischen den Gruppen minimiert werden soll.[14] Die Clusteranalyse ist somit immer dann notwendig, wenn die Daten ungekennzeichnet sind, unabhängig davon, welcher Datentyp den Daten zugrunde liegt. Die meisten Daten sind statisch, ihre Werte bleiben mit der Zeit also unverändert oder ändern sich nur selten. Der überwiegende Teil an Clusteranalysen wurde auf statischen Daten ausgeführt. Han et al. (2001) teilten die Clusteralgorithmen in fünf Kategorien ein.

1. Partitionierendes Verfahren: Erzeugt aus n unbeschrifteten Datensätzen k Partitionen, wobei jede Partition ein Cluster repräsentiert, mindestens ein Objekt enthält und $k \leq n$ gilt. Es sind hierbei unterschiedliche Formen der Gruppenbildung möglich. Ein Objekt kann genau einer Partition, mehrerer Partitionen oder allen Partitionen in einem bestimmten Grad angehörig sein. Die gewünschte Art der Gruppenbildung ist von der Datenmenge, als auch der Repräsentation der Cluster abhängig. Der in dieser Arbeit verwendete Algorithmus K-Means zählt zu den harten Methoden, d. h. er ordnet jedem

[14] Vgl. Liao (2005), S.1857

Objekt genau eine Partition zu. Jeder Cluster wird dabei durch seinen Mittelwert repräsentiert.

2. Hierarchisches Verfahren: Erzeugt eine hierarchische Baumstruktur aus Clustern. Die Ähnlichkeit der Cluster wird durch ein ausgewähltes Distanzmaß bestimmt. Je nach Berechnungsvorschrift gibt es zwei wichtige Typen von Verfahren. Bei einem divisiven Verfahren gehören zunächst alle Objekte zu einem Cluster. Anschließend werden die Cluster schrittweise in kleinere Cluster aufgeteilt bis jeder Cluster nur noch aus einem Objekt besteht. Bei einem agglomerativen Verfahren bildet zunächst jedes Objekt einen Cluster. Schrittweise werden die bereits gebildeten Cluster zu größeren Clustern zusammengefasst bis alle Objekte zu einem Cluster gehören. In jedem Fall können einmal gebildete Cluster nicht mehr verändert werden. Die Anzahl der gewünschten Cluster muss für die Durchführung der Analyse nicht bekannt sein.
3. Dichtebasierte Verfahren: Identifizieren zusammenhängende Gebiete im Variablenraum als Cluster. Die Cluster wachsen dabei solange bis ein festgelegtes Dichtekriterium erfüllt ist. Es kann deshalb passieren, dass bestimmte Objekte keinem Cluster zugeordnet werden können. Diese werden als Ausreißer oder Rauschen in den Daten identifiziert.
4. Grid-basierte Verfahren: Erzeugen zunächst eine hierarchische Gitterstruktur, wobei der Variablenraum in eine endliche Zahl an Zellen unterteilt wird. Für jede Zelle werden entsprechende statistische Kennzahlen vorberechnet. Entsprechend einer Query wird das Konfidenzintervall für jede Zelle anhand ihrer statistischen Informationen berechnet. Die jeweilige Zelle wird als relevant oder irrelevant gekennzeichnet. Für relevante Zellen wird dieser Vorgang eine Ebene tiefer wiederholt. Anschließend werden zusammenhängende Räume, bestehend aus den relevanten Zellen, als Cluster zurückgegeben.[15]

[15] Vgl. Wang et al. (1997) S. 4

5. Modell-basierte Verfahren: Basieren auf der Annahme, dass die vorliegenden Daten aus einer Verteilung entstanden sind, welche aus einer Mischung mehrerer Komponenten besteht. Jede Komponente wird durch eine Dichtefunktion beschrieben und hat eine assoziierte Wahrscheinlichkeit oder Gewichtung in der Mischung. Anhand festgelegter Parameter und den Daten wird ein mathematisches Modell angelernt.

Im Gegensatz zu statischen Daten umfassen die Zeitreihen eines Merkmals Werte, die sich mit der Zeit ändern. Bei einer Menge von ungekennzeichneten Zeitreihen ist es das Ziel, Gruppen ähnlicher Zeitreihen zu bestimmen. Dabei wird ebenso wie bei statischen Daten ein Verfahren benötigt, um die Cluster zu bilden. Der Algorithmus wird entsprechend der Form der Daten und dem Anwendungszweck bestimmt. Die Zeitreihen können hierbei diskret oder kontinuierlich, gleichmäßig oder ungleichmäßig abgetastet, univariat oder multivariat, gleicher oder ungleicher Länge sein. Ungleichmäßig abgetastete Daten müssen zunächst in gleichmäßig abgetastete Daten umgewandelt werden, bevor eine Clusteranalyse angewendet werden kann. Hierfür stehen unterschiedliche Verfahren zu Verfügung.[16]

3.2 Vorgehensweisen

Verschiedene Algorithmen wurden entwickelt, um eine Clusteranalyse auf Zeitreihen unterschiedlichster Formen anzuwenden. Blendet man die Unterschiede dieser Algorithmen aus, so versuchen sie im Kern bereits existierende Clusteralgorithmen für statische Daten so anzupassen, dass diese auch für Zeitreihen nutzbar gemacht werden können oder Zeitreihen in statische Daten umgewandelt werden, sodass Clusteralgorithmen für statische Daten angewendet werden können. Die folgende Abbildung zeigt die möglichen Vorgehensweisen, um eine Clusteranalyse auf Zeitreihen anzuwenden.

[16] Vgl. Liao (2005), S.1858

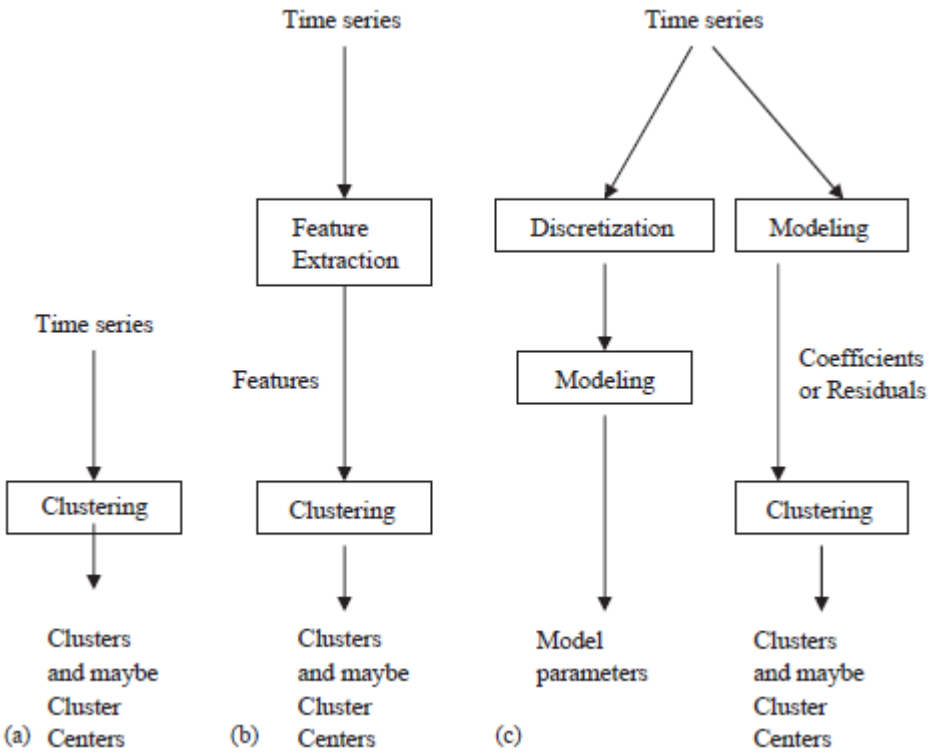


Abbildung 4: Mögliche Vorgehensweisen der Clusteranalyse zeitbasierter Daten

Quelle: Liao (2005), S.1859

Der Rohdaten-Ansatz (a) nutzt die Zeitreihen als direkte Eingabe für den Algorithmus. Lediglich das Distanzmaß muss sinnvoll für die Zeitreihen modifiziert und ausgewählt werden. Der Feature-basierte Ansatz (b) gewinnt statistische Kennzahlen aus den Rohdaten und führt anschließend eine Clusteranalyse auf diesen durch. Es findet also gleichzeitig eine Dimensionsreduktion statt. Hierbei ist es wichtig, möglichst aussagekräftige statistische Kennzahlen zu nutzen, um einen Informationsverlust zu vermeiden. Die Repräsentanten der Cluster können dann allerdings nicht mehr als Mittelwert der Rohdaten genutzt werden. Der Modell-basierte Ansatz (c) erzeugt aus den Rohdaten Koeffizienten oder Modellparameter und wendet anschließend eine Clusteranalyse auf diese an.

3.3 K-Means

Um einen Überblick zu gewährleisten, wurden in der Übersicht die verschiedenen Verfahren kurz erläutert, in die sich die unterschiedlichen Algorithmen einordnen lassen. Dieses Kapitel geht nun detailliert auf den Algorithmus K-Means ein, welcher in dieser Arbeit verwendet wurde und zu den partitionierenden Verfahren gehört. Dieser Algorithmus wurde vor mehr als 30 Jahren entwickelt.[17] Die Hauptidee bestand darin, eine Funktion zu minimieren, welche die totale Distanz zwischen allen Objekten und ihren jeweiligen Clusterzentren wiedergibt. Sein Lösungsansatz basiert auf einem iterativen Schema, welches mit willkürlich initiierten Clusterzentren beginnt.[18] In jeder Iteration werden die Objekte entsprechend der aktuellen Clusterzentren einem Cluster zugeordnet und die Clusterzentren entsprechend ihrer Objekte neuberechnet. Der Algorithmus führt diese beiden Schritte solange aus, bis die Funktion nicht weiter minimiert werden kann. Gegeben seien *n* Objekte$\{x_j \mid j = 1, \dots, n\}$, k-Means bestimmt k Clusterzentren$\{\mu_i | i = 1, \dots, k\}$, indem die Funktion $J = \sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - \mu_i\|^2$ minimiert wird. Das Distanzmaß $\|\cdot\|$ ist hierbei eine Funktion, welche die Ähnlichkeit zwischen zwei Datenpunkten, hier einem Objekt und einem Clusterzentrum, misst. In dieser Arbeit wurde das quadrierte euklidische Distanzmaß verwendet, es können aber auch andere Distanzmaße verwendet werden. Die euklidische Distanz zwischen den beiden P-Dimensionale Vektoren (x_i und v_j) wird dann wie folgt berechnet:

$$d_E = \sqrt{\sum_{k=1}^{P} (x_{ik} - v_{jk})^2}$$

[17] Siehe MacQueen (1967)

[18] Vgl. Liao (2005), S.1860

Der Algorithmus besteht generell aus den folgenden drei Schritten.

1. Initialisierung: Wähle k zufällige Mittelwerte: $m_1, \dots, m_k$ aus dem Datensatz.
2. Zuordnung: Jedes Objekt wird demjenigen Cluster zugeordnet, bei dem die Cluster-Varianz am wenigsten erhöht wird.
   $S_i^{(t)} = \left\{ x_j : \left\| x_j - m_i^{(t)} \right\|^2 \leq \left\| x_j - m_{i^*}^{(t)} \right\|^2, \forall : i^* = 1, \dots, k \right\}$
3. Aktualisieren: Die Mittelpunkte der Cluster werden entsprechend neu berechnet. $m_i^{(t+1)} = \frac{1}{\left| S_i^{(t)} \right|} \sum_{x_j \in S_i^{(t)}} x_j$

Die Schritte 2-3 werden solange wiederholt, bis sich die Zuordnungen nicht mehr ändern. Die Anzahl der Cluster muss im Voraus bekannt sein, kann jedoch experimentell durch ein Evaluierungskriterium (siehe nächster Abschnitt) bestimmt werden. Fehlerhafte Datenobjekte, wie beispielsweise Ausreißer, können die berechneten Clusterzentren erheblich verschieben. Der Algorithmus hat keine Vorkehrungen gegen derartige Effekte, weshalb die Vorverarbeitung der Daten besonders wichtig ist. Das Ergebnis der Analyse hängt stark von gewählten Startpunkten ab, weshalb der Algorithmus mehrmals mit verschiedenen Startpunkten durchgeführt wird.

3.4 Silhouettenkoeffizient

Der Silhouettenkoeffizient ist ein unabhängiges Evaluierungskriterium zur Bestimmung der Qualität einer Clusteranalyse. Er kann außerdem für partitionierende Verfahren genutzt werden, um die optimale Anzahl an Clustern zu bestimmen.[19] Der Silhouettenkoeffizient ist definiert als das arithmetische Mittel aller n_C Silhouetten des Clusters C.

$$s_C = \frac{1}{n_C} \sum_{o \in C} s(o)$$

[19] Vgl. Rousseeuw (1986) S. 53

Die Silhouette eines Objektes o ist, wenn o zum Cluster A gehört, definiert als:[20]

$$S(o) = \begin{cases} 0 & wenn\ dist(A,o)=0 \\ \frac{dist(B,o)-dist(A,o)}{\max\{dist(A,o),dist(B,o)\}} & sonst \end{cases}$$

$dist(A,o)$ ist hierbei die Distanz eines Objektes o zum Cluster A und $dist(B,o)$ die Distanz eines Objektes o zum nächstgelegenen Cluster B. Dabei wird die Differenz des Abstands $dist(B,o) - dist(A,o)$ gewichtet mit der maximalen Distanz. Damit folgt, dass $S(o)$ für ein Objekt o zwischen -1 und 1 liegt. Folgende Erkenntnis kann aus einer Silhouette gewonnen werden:

1. Ist die Silhouette $S(o) < 0$, dann liegen die Objekte des nächstgelegenen Clusters B näher an dem Objekt o als die Objekte des Cluster A zu dem Objekt o gehört. Das Clustering weist eine schlechte Qualität auf.
2. Ist die Silhouette $S(o) \approx 0$ liegt das Objekt zwischen zwei Clustern.
3. Ist die Silhouette nahe 1, so liegt das Objekt in einem Cluster. Das Clustering weist dann eine sehr gute Qualität auf.

Die Distanz $dist(A,o)$ wird berechnet als

$$dist(A,o) = \frac{1}{n_A}\sum_{a\in A} dist(a,o)$$

der Mittelwert der Distanz zwischen allen Objekten im Cluster A und dem Objekt o. n_A ist hierbei die Anzahl der Objekte im Cluster A. Analog wird die Distanz zum nächstgelegenen Cluster B als die minimale durchschnittliche Distanz berechnet:

[20] Vgl. Ester et al. (2000), S.66

$$dist(B,o) = \ min_{c \neq a} dist(C,o) = min_{c \neq a} \left(\frac{1}{n_c} \sum_{c \in C} dist(c,o) \right)$$

Es wird für alle Cluster C, die das Objekt o nicht enthalten, die Distanz $dist(C,o)$ berechnet. Der nächstgelegene Cluster B ist derjenige, der die kleinste Entfernung $dist(C,o)$ aufweist. Der Silhouettenkoeffizient kann grafisch dargestellt werden. Für alle Beobachtungen, die zu einem Cluster gehören, wird der Wert der Silhouette als waagerechte Linie dargestellt. Die Beobachtungen in einem Cluster werden dabei nach der Größe der Silhouette geordnet. Im Folgenden wird ein Silhouttenplot für einen Beispieldatensatz gezeigt. Die rote senkrechte Linie gibt den durchschnittlichen Silhouettenkoeffizient wieder.

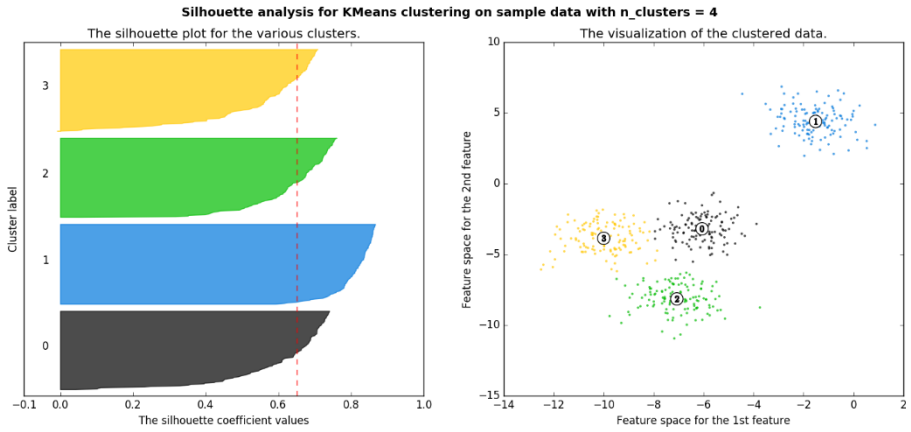


Abbildung 5: Exemplarischer Silhouettenplot zur Bestimmung der Qualität einer Clusteranalyse bei 2-dimensionalen Datenpunkten

Quelle: Scikit Learn, http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis (2016)

4 CRISP-DM

4.1 Referenzmodell

CRISP-DM ist ein hierarchisches Prozessmodell und stellt einen Entwurf zur Durchführung von Data-Mining-Projekten dar.[21] Um eine strukturierte Umsetzung der Problemlösung zu gewährleisten, wird CRISP-DM in dieser Arbeit angewendet. Es werden allerdings nur Arbeitspakete angewendet, die zur Durchführung dieses Projektes nötig sind. Arbeitspakete, welche nicht verwendet werden, sind in Tabelle 8 im Anhang, Kapitel A, aufgelistet. Insgesamt wird der Lebenszyklus eines Data-Mining-Projektes in sechs Phasen heruntergebrochen, welche wiederum aus mehreren Arbeitspaketen bestehen. Die folgende Abbildung zeigt die Phasen des Data-Mining-Prozesses. Die Pfeile deuten die wichtigsten und am häufigsten genutzten Abhängigkeiten zwischen den Phasen an. Der äußere Kreis symbolisiert die zyklische Natur eines Data-Mining-Projekts.

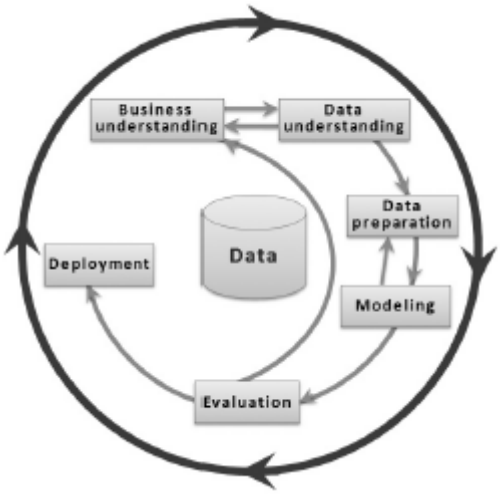


Abbildung 6: CRISP-DM Referenzmodell

Quelle: Chapman et al. (1999)

[21] Vgl. Collin (2000), S.14

Die Arbeitspakete im Prozessmodell sind vollkommen und generisch. Sie sind somit unabhängig von einer spezifischen Problemstellung. Das Prozessmodell ist weiterhin stabil in Bezug auf zukünftige Entwicklungen im Data Mining, wie beispielsweise neue Modellierungstechniken. Neben diesen Vorteilen wurde CRISP-DM in dieser Arbeit aufgrund seiner Flexibilität, beziehungsweise agilen Struktur als Prozessmodell ausgewählt. Wie in Abbildung 6 erkennbar, besteht die Möglichkeit bei Bedarf zwischen den Phasen zu wechseln. Im Vergleich zu alternativen Prozessmodellen, wie SEMMA oder KDD, ist CRISP-DM besser dokumentiert.

4.2 Phasen

Chapman et al. (1999) definieren die Phasen vom Referenzmodell CRISP-DM folgendermaßen:

1. **Business Understanding**

   Diese Phase konzentriert sich auf die Zielsetzung des Projektes und die externen Anforderungen aus betriebswirtschaftlicher Sicht. Dieses Wissen wird anschließend in eine Data-Mining-Problemstellung umformuliert.

2. **Data Understanding**

   Es werden erste Daten gesammelt, um ein Verständnis und erste Erkenntnisse über diese zu erlangen, die Datenqualität zu überprüfen oder Hypothesen über versteckte Information zu formulieren.

3. **Data Preparation**

   Die Datenvorverarbeitungsphase beinhaltet alle Aktivitäten zur Erzeugung einer finalen Datentabelle, welcher als Eingabe für das Datenmodell genutzt wird. Die Aktivitäten können wiederholt werden und einer beliebigen Reihenfolge unterliegen. Klassische Aktivitäten beinhalten das Selektieren relevanter Attribute, das Transformieren und Säubern der Daten.

4. **Modeling**

   Ein oder mehrere Datenmodelle werden ausgewählt, die optimalen Parameter bestimmt und anhand des finalen Datensatzes kalibriert. Es gibt häufig mehrere Techniken für das gleiche Data-Mining-Problem. Manche haben bestimmte Anforderungen an die Daten, weshalb ein Schritt zurück in die Datenvorverarbeitungsphase nötig sein kann.

5. **Evaluation**

   Die Qualität des Modells wird anhand statischer Kennzahlen beurteilt und vorangegangene Schritte können aufgearbeitet werden. Damit wird sichergestellt, dass das Modell den Anforderungen aus betriebswirtschaftlicher Sicht genügt.

6. **Deployment**

   Das erzeugte Modell kann nun genutzt werden, um automatische Realtime- oder Batchanfragen zu beantworten. Falls das ursprüngliche Ziel jedoch aus der reinen Wissensgenerierung bestand, kann die Deployment-Phase von einem einfachen Report bis hin zu der Implementation eines komplexen Data-Mining-Prozesses in dem Unternehmen reichen.

5 Bestimmung der Wirkungsweise dynamischer Preisberechnung auf die Verbrauchsmuster

5.1 Business Understanding

Das Geschäftsziel besteht darin zu überprüfen, ob eine dynamische Preisberechnung einen positiven Einfluss auf die Verbrauchsmuster von Verbrauchern elektrischer Fahrzeuge hat. Hierfür werden die Strommessdaten häuslicher Fahrzeugladestationen ausgewertet. Die folgenden Verbraucherklassen werden dabei in Betracht gezogen:

1. Klasse - Dynamic Pricing (DP): Verbraucher erhalten durch eine dynamische Preisberechnung einen wirtschaftlichen Anreiz, ihren Strom außerhalb der Spitzenlastzeit günstiger zu beziehen.
2. Klasse - Control (C): Verbraucher erhalten keinen wirtschaftlichen Anreiz, ihren Strom außerhalb der Spitzenlastzeit günstiger zu beziehen.

Um das Ergebnis messbar zu machen, werden die Verbrauchsmuster in Bezug auf ihre Gefährdung für die Netzstabilität bewertet und in kritische, sowie nicht kritische Zeiträume unterteilt. Als kritisch werden alle Zeiträume eines Verbrauchsmusters bezeichnet, die einen Verbrauch von 1,5 Kilowatt oder höher zu Spitzenlastzeiten verzeichnen. Die Spitzenlastzeit definieren Dennis R. Landsberg et al. (1980) folgendermaßen:

„Peak Demand, peak-load or on-peak … describing a period in which electrical power is expected to be provided for a sustained period at a significantly higher than average supply level. “

Die Spitzenlastzeit variiert hierbei in der Winter- und Sommerzeit, als auch zwischen Werks- und Feiertagen, und wird regional durch den jeweiligen Energieversorger bestimmt. Die in dieser Arbeit verwendeten Spitzenlastzeiten stammen von einem amerikanischen Energieversorger und sind in der folgenden Tabelle dargestellt:

Spitzenlastzeiten - Winter (Zeitraum: 01.10 – 31.05)	
Werktag	Feiertag und Wochenende
05:00 – 09:00 17:00 – 21:00	-
Spitzenlastzeiten - Sommer (Zeitraum: 01.06 – 30.09)	
Werktag	Feiertag und Wochenende
14:00 – 20:00	-

Tabelle 1: Spitzenlastzeiten im Winter und Sommer.

Quelle: Black Hills Electric Cooperative, http://www.bhec.com/content/peak-demand-times (o.J.)

Die kritischen Abschnitte aller Verbrauchsmuster einer Verbraucherklasse werden nun mit der jeweiligen Auftrittshäufigkeit eines Verbrauchmusters multipliziert und in das Verhältnis aller Auftrittshäufigkeiten gesetzt. Der relative Anteil kritischer Abschnitte aller Ladephasen einer Verbraucherklasse kann so bestimmt und verglichen werden. Verbraucher befinden sich zu Werkstagen, aufgrund ihrer sich oft überschneidenden Arbeitszeiten, zu ähnlichen Uhrzeiten zu Hause, weshalb sich der Verbrauch akkumuliert und zu einer Spitzenlast führt. Deshalb ist zu Wochenenden und Feiertagen nicht mit einer Spitzenlast zu rechnen, weshalb diese Tage nicht in der Analyse berücksichtigt werden. Außerdem muss der Variation der Spitzenlastzeit im

Sommer und Winter Rechnung getragen werden. Hierfür muss der Datensatz entsprechend aufgeteilt werden. Insgesamt sind die folgenden drei Szenarien möglich.

1. Szenario: Der relative Anteil kritischer Ladephasen aller Verbrauchsmuster ist bei Verbrauchern der Klasse DP (mit dynamischer Preisberechnung) um mindestens 5% höher als bei Verbrauchern der Klasse C (Ohne dynamische Preisberechnung). Die dynamische Preisberechnung hätte somit einen negativen Einfluss auf die Verbrauchsmuster.

2. Szenario: Der relative Anteil kritischer Ladephasen aller Verbrauchsmuster ist bei Verbrauchern der Klasse DP mindestens 5% niedriger als bei Verbrauchern der Klasse C. Die dynamische Preisberechnung hätte somit einen positiven Einfluss auf die Verbrauchsmuster.

3. Szenario: Der relative Anteil kritischer Ladephasen aller Verbrauchsmuster weicht bei Verbrauchern der Klasse DP und der Klasse C um maximal 5% ab. Die dynamische Preisberechnung hätte somit keinen signifikanten Einfluss auf die Verbrauchsmuster.

Der Schwellenwert von 5% wird hier eingeführt, um ausschließlich bei einer stärkeren Korrelation zwischen dynamischer Preisabrechnung und den Verbrauchsmustern von einem Einfluss auszugehen.

Um das gesetzte Geschäftsziel zu erreichen, steht eine PostgreSQL-Datenbank zur Verfügung. Diese wird von der Entwicklungs- und Forschungsorganisation PecanStreet Inc. zur Verfügung gestellt. Der Fokus dieser Organisation liegt in der Entwicklung und Prüfung von erweiterten Technologien, Geschäftsmodellen und dem Kundenverhalten in dem Bereich der fortschrittlichen Energiemanagementsysteme. Das Vorzeigeschild dieser Institution liegt in der Implementierung eines intelligenten Stromnetzes in Austin, Texas. Ein solches Netz umfasst die kommunikative Vernetzung und Steuerung

von Stromerzeugern, Speichern, elektrischen Verbrauchern und Netzbetriebsmitteln in Energieübertragungs- und Verteilungsnetzen der Elektrizitätsversorgung.[22] Insgesamt teilen 1000 Teilnehmer ihre Energieverbrauchsdaten über Smart-Meter. Diese intelligenten Messgeräte sind Bestandteil des intelligenten Netzes und können digitale Daten, wie beispielsweise Verbrauchsdaten, empfangen oder senden. Insgesamt wurden solche intelligenten Messgeräte in 750 Haushalten und 25 Geschäften installiert. Ein intelligentes Messgerät zeigt die folgende Abbildung:

Abbildung 7: Smart-Meter

Quelle: Clean Energy Ministerial, http://www.cleanenergyministerial.org/News/isgan-releases-advanced-metering-infrastructure-case-book-1093 (2013)

Die Datenbank umfasst neben den Energieverbrauchsdaten von Strom, Wasser und Gas auch Wetterdaten und Daten zu Umfragen und Audits. Zur Erreichung des Geschäftsziels werden allerdings ausschließlich die Energieverbrauchsdaten ($n = 2.489.461$) häuslicher Fahrzeugladestationen und Daten zu variablen Tarifen benötigt. Die Messdaten müssen eindeutig identifizierbar und mit einem Zeitschlüssel versehen sein, damit diese zu Zeitreihen zusammengesetzt werden können. Anschließend können diese Zeitreihen als Eingabe für das Datenmodell verwendet

[22] Vgl. Von Dollen D. (2005), S.6

werden. Die Daten werden mittels PostgreSQL durch das Tool PGAdmin3 extrahiert. Die Analyse wird in der Sprache Python (Version 3.5) in der Entwicklungsumgebung Spyder programmiert.

Aus dem Geschäftsziel lassen sich zwei Data-Mining-Ziele ableiten. Zum einen müssen die ungeordneten Messdaten der häuslichen Fahrzeugladestationen in tägliche Zeitreihen umgewandelt werden. Zum anderen müssen diese Zeitreihen, abhängig ihrer Distanz zueinander unterschiedlichen Clustern zugeordnet werden. Die Bildung der Zeitreihen als Ziel ist also unabdingbar zur Erfüllung des zweiten Ziels. Die Bildung der Cluster erfolgt durch unüberwachtes Lernen und fällt somit in den klassischen Bereich der Clusteranalyse. Dabei werden alle verfügbaren Daten genutzt, um die Cluster zu bilden, beziehungsweise den Algorithmus anzulernen. Die Anzahl der Cluster ist dabei zunächst unbestimmt, weshalb ein partitionierendes Verfahren angewendet wird. Das Verfahren erhält die Energiemessdaten in Form einer Zeitreihe und ordnet dieser eine Ausgabe in Form eines Clusters zu. Es wird also das Rohdaten-Verfahren angewendet. Die Mittelpunkte dieser Cluster werden hierbei als Verbrauchsmuster verwendet. Diese ermöglichen es, den Einfluss dynamischer Preisberechnung auf das Verbraucherverhalten zu erkennen. Der Erfolg des zweiten Data-Mining-Ziels hängt von der Qualität der gebildeten Cluster hinsichtlich ihrer Unterscheidbarkeit ab. Die Unterscheidbarkeit der Cluster wird durch den Sillhouettenkoeffizienten beschrieben. Dieser muss bei beiden Verbraucherklassen zumindest positiv sein.

5.2 Data Understanding

Für die Erzeugung täglicher Zeitreihen müssen die Strommessdaten eindeutig identifizierbar und mit einem Zeitstempel versehen sein. Die benötigten Daten zur Erreichung dieses Ziels sind in der Tabelle „electricity-egauge-15min“ enthalten. Dort sind die erhobenen Strommessdaten aller elektrischen Geräte eines Haushaltes im 15-Minuten-Takt hinterlegt. Der hinterlegte Wert entspricht jeweils der durchschnittlich gemessenen Wattzahl in dem entsprechenden Zeitintervall eines Gerätes. Die Energiemessdaten der häuslichen Fahrzeugladestationen sind in der Spalte „car1“

hinterlegt. Die Spalte „dataid“ enthält eine eindeutige Nummer zur Identifizierung eines Wohnsitzes und stellt zusammen mit dem Zeitstempel aus der Spalte „local_15min“ einen Primärschlüssel dar, um jeden Messwert eindeutig zu identifizieren. Es werden alle vorhandenen Strommessdaten vom 01.04.2013 bis zum 30.09.2014 (18 Monate) verwendet. In diesem Zeitraum wurde das Versuchsprogramm durch das Center for Commercialization of Electric Technologies (CCET) in Kooperation mit PecanStreet Inc. durchgeführt. Die Daten zu variablen Tarifen sind in der Spalte „program_ccet_group“ in der Tabelle „metadata“ gespeichert. Die folgende Anfrage selektiert alle verwertbaren Strommessdaten häuslicher Fahrzeugladestationen im entsprechenden Zeitraum aller Haushalte, die am $CCET$- Programm teilnehmen, und speichert diese in einer CSV-Datei. Jeder dieser Haushalte verfügt also über ein Elektroauto.

$$
\begin{aligned}
&SELECT\ dataid, local\_15min\ , car1\\
&FROM\ university.electricity\_egauge\_15min\\
&\quad where\ car1\ >=\ 0\ and\ '2013.04.01'\ <=\ local\_15min\\
&\quad and\ local\_15min\ <\ '2014.10.01'\\
&\quad and\ dataid\ in\ (SELECT\ distinct(dataid)\\
&\qquad FROM\ university.metadata\\
&\qquad where\ car1 = \ 'yes'\\
&\qquad and\ program\_ccet\_group\ is\ NOT\ NULL)\\
&\quad and\ dataid\ in\ (SELECT\ dataid\\
&\qquad FROM\ university.electricity\_egauge\_15min\\
&\qquad WHERE\ car1\ >\ 0\\
&\qquad group\ by\ dataid\\
&\qquad having\ stddev\_pop(car1)\ >\ 0.1);
\end{aligned}
$$

Im Folgenden werden die Tabellen sowie die darin enthaltenen und relevanten Attribute näher beschrieben.

Tabelle „electricity-egauge-15min“			
Attribut	Dataid	Local_15min	Car1
Typ	Natürliche Zahl	Zeitstempel	Gleitkommawert
Bedeutung	Identifiziert den Haushalt, in welchem die Messdaten angefallen sind.	Gibt Datum und Uhrzeit einer Messung an.	Gibt die Strommessdaten der jeweiligen Fahrzeugladestation zum entsprechenden Zeitstempel wider.
Anzahl der Datensätze	2.489.461	2.489.461	2.489.461
Minimum	26	01.04.2013 00:00:00	0 (n = 1.668.980)
Maximum	9937	30.09.2014 23:45:00	9,61
Durchschnitt	-	-	0,26
Standardabweichung	-	-	0,88
75% - Quartil	-	-	0,00

Tabelle 2: Beschreibung der Attribute aus der Tabelle " electricity-egauge-15min "

Tabelle „metadata“		
Attribut	Dataid	Program_ccet_group
Typ	Natürliche Zahl	Aufzählungstyp
Wertebereich	26 – 9937	{CCET – Pricing Trial, CCET - Control}
Bedeutung	Identifiziert den Haushalt, in welchem die Messdaten angefallen sind.	Gibt an, ob der entsprechende Haushalt in dem Zeitfenster vom 01.04.2013 bis zum 01.10.2014 an einer dynamischen Preisberechnung partizipierte. Dabei gilt die folgende Semantik: „CCET – Pricing Trial“- Haushalt erhält variablen Tarif. (n=33) „CCET – Control“- Haushalt erhält keinen variablen Tarif. (n = 8)
Anzahl der Datensätze	41	41

Tabelle 3: Beschreibung der Attribute aus der Tabelle "metadata"

Das Attribut „Dataid“ stellt zwischen den gegebenen Tabellen jeweils eine 1:1-Beziehung her. Die Informationen können miteinander verknüpft werden, um die Datensätze später entsprechend nach Verbraucherklassen aufzuteilen. Das 75%-

Quartil der Strommessdaten liegt nahe bei 0. Dies bedeutet, dass die Fahrzeuge die meiste Zeit über nicht geladen werden. Rund 80% der Haushalte erhalten in diesem Zeitraum einen variablen Tarif. Dieses Ungleichgewicht wird ausgeglichen, indem die relative Anzahl der kritischen Ladephasen aller Verbrauchsmuster der beiden Verbraucherklassen miteinander verglichen werden.

Um ein genaueres Verständnis der Daten zu erlangen, werden mittels Visualisierungstechniken die Verteilungen der Strommessdaten der beiden Verbraucherklassen visualisiert und können somit gegenüber gestellt werden. Die Wochenenden wurden für diese Darstellungen bereits entfernt. Aus Gründen der Übersichtlichkeit wurden nur Werte größer als 1 kW genommen. Es werden hier also ausschließlich die aktiven Ladephasen betrachtet.

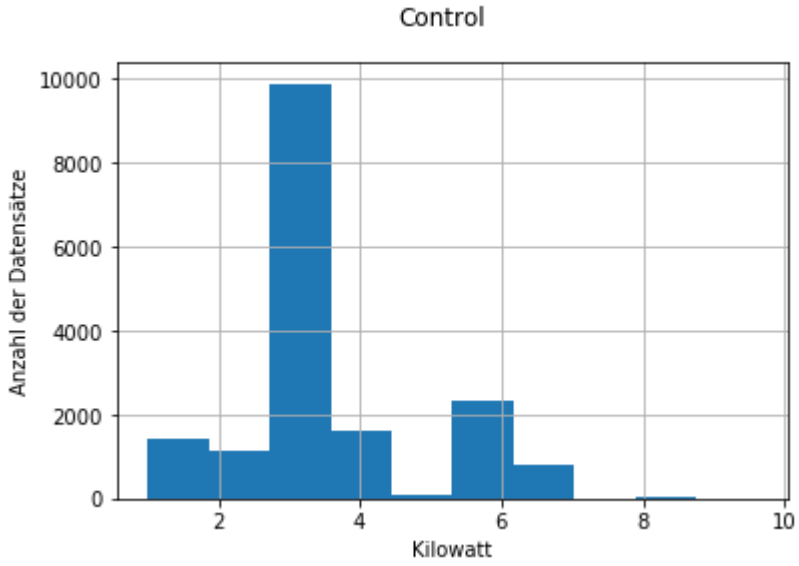


Abbildung 8: Verteilung der Strommessdaten von der Verbraucherklasse C

Der überwiegende Teil der Strommessdaten der Verbraucherklasse C liegt in den aktiven Ladephasen im Intervall [1; 4]. Ein weiterer Cluster befindet sich im Intervall [5; 7]. Außerdem scheinen einige wenige Extremwerte im Intervall [8; 9] vorzuliegen. Die Daten erscheinen zunächst plausibel hinsichtlich ihrer Verteilung und des vorliegenden Wertebereichs. Schließlich verändern sich die Strommesswerte zum Anfang und Ende einer aktiven Ladephase schrittweise. Diese steigen kontinuierlich bis zum Maximum an, halten sich auf diesem Level und fallen dann wieder schrittweise auf 0. Liegt das

Maximum einer aktiven Ladephase beispielsweise bei 3 Kilowatt, so werden am Anfang und Ende einer Ladephase Werte zwischen 0 und 3 erreicht. Die Ladestation erreicht oder verlässt in einer aktiven Ladenphase ihr Maximum also nicht umgehend. Die Ladekapazität variiert hierbei durch das verwendete Kabel oder Automodell. Somit ist eine Abweichung in den Strommessdaten realistisch. Dennoch sollten die Verteilungen pro Haushalt analysiert werden, um auszuschließen, dass es sich im Einzelfall um Ausreißer handelt.

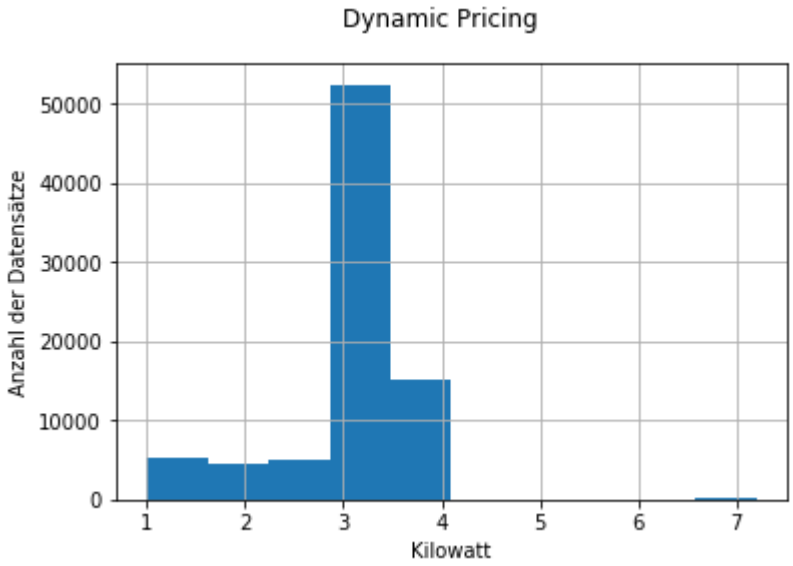


Abbildung 9: Verteilung der Strommessdaten von der Verbraucherklasse DP

Bei der Verbraucherklasse DP befinden sich die Strommessdaten in den aktiven Ladephasen fast ausschließlich in dem Intervall [1; 4]. Im Gegensatz zu der Verbraucherklasse C ist kein weiterer Cluster im Intervall [5; 6] vorhanden. Einige Extremwerte scheinen sich in dem Intervall [6,5; 7,1] zu befinden. Auch in diesem Fall erscheinen die Daten zunächst plausibel. Eine genauere Analyse der Verteilung eines Haushaltes kann weitere Erkenntnisse liefern. Insgesamt erhalten rund 80% der Haushalte in der Testphase einen variablen Tarif. Dies spiegelt sich in der Anzahl an Messdaten bei den Verbraucherklassen wider. Der Stromverbrauch während des gesamten Zeitraumes akkumuliert sich bei der Klasse DP auf 268166 Kilowatt. Bei der Klasse C sind es 65398 Kilowatt. Vergleicht man diese Werte im Verhältnis der Haushalte pro Verbraucherklasse an der Gesamtzahl der Haushalte wird ersichtlich,

dass der Stromverbrauch beider Klassen zunächst unverändert ist. Die variablen Tarife haben also zunächst keinen Einfluss auf die Höhe des gesamten Strombedarfs.

Neben den Verteilungen pro Verbraucherklasse können statistische Eigenschaften pro Haushalt weitere wichtige Erkenntnisse für die nächsten Schritte liefern. Aus Gründen der Übersichtlichkeit wird hier erneut auf eine tabellarische Darstellungsform zurückgegriffen. Für jede Verbraucherklasse wird eine eigene Tabelle angelegt. Diese befinden sich aufgrund ihrer Größe im Anhang, Kapitel A. Ein Vergleich zwischen den Kennzahlen der Verbraucherklassen macht erneut deutlich, wie ähnlich sich diese sind. Die Anzahl der Datensätze, als auch der Durchschnitt, die Standardabweichung und das 75%-Quartil weichen im Durchschnitt wenig voneinander ab. Die geringe Abweichung der statistischen Kennzahlen innerhalb der Verbraucherklassen zeigt, dass die Haushalte in Ihrem Verbrauch sehr ähnlich sind. Alle Haushalte weisen ein 75%-Quartil nahe 0 auf. Die Fahrzeuge werden im Durchschnitt also die meiste Zeit über nicht aufgeladen. Um einen größtmöglichen Informationsgewinn zu erzielen, muss in der Datenvorverarbeitungsphase ein Filter auf Ebene der Zeitreihen eingesetzt werden, um nur solche Zeitreihen als Eingabe für das Datenmodell zu nutzen, die ein vorher festgelegtes Maximum oder eine vorher festgelegte Standardabweichung überschreiten. Auffällig ist weiterhin, dass einige Haushalte ein weitaus größeres Maximum als andere Haushalte aufweisen. Dies könnte ein Zeichen für Ausreißer sein und wird im Folgenden genauer überprüft. Tabelle 10 im Anhang zeigt die Verteilungen der Strommessdaten, in denen Extremwerte vermutet werden. Im Einzelfall wird entschieden, ob es sich bei den Werten um Ausreißer handelt. Diese werden bereinigt, falls es sich um solche handelt.

Wie die vorherigen Untersuchungen gezeigt haben, ist die Datenqualität für die Erreichung der Data-Mining-Ziele ausreichend.

- Die Schlüssel der Tabellen „electricity-egauge-15min“ und „metadata“ stimmen überein und können so zusammengefasst werden.
- Die Verteilung der Haushalte hat gezeigt, dass die Strommessdaten plausibel sind. Schließlich sind diese nicht negativ, in den Haushalten ähnlich verteilt und decken einen gültigen Wertebereich ab. Außerdem scheinen aktive Ladephasen nur einen Bruchteil der gesamten Strommessungen zu bestimmen. Das entspricht den Erwartungen, weil die Fahrzeuge einen Teil der Zeit in Benutzung sind und in dieser Zeit nicht aufgeladen werden können. Es kann aber auch sein, dass ein Fahrzeug an einem bestimmten Tag überhaupt nicht gebraucht wird und deshalb auch nicht aufgeladen werden muss, oder an einer öffentlichen Ladestation aufgeladen wurde.
- Die Bedeutung der Attribute und die enthaltenen Werte passen zusammen. In der Spalte „local_15min“ sind ausschließlich Zeitstempel hinterlegt. In der Spalte „dataid“ ausschließlich Nummern. Zusammen bilden diese eine eindeutige Kombination.

Die folgenden Probleme wurden hierbei erkannt. Es ist jeweils ein Lösungsvorschlag angegeben.

- Es herrscht ein Ungleichgewicht bezüglich der Anzahl der Stichproben beider Verbraucherklassen. Durch die Nutzung eines relativen Vergleichskriteriums zur Bestimmung des Geschäftsziels wird dieses Ungleichgewicht ausgeglichen.
- Das 75%-Quartil der Strommessdaten aller Haushalte liegt nahe bei 0. Für die Datenanalyse sind jedoch nur Zeitreihen, die eine aktive Ladephase beinhalten, von Bedeutung. Um einen möglichst großen Informationsgewinn zu erreichen, wird ein Filter auf Ebene der Zeitreihen eingesetzt. Es werden nur solche Zeitreihen als Eingabe für das Datenmodell verwendet, die ein gewisses

Maximum und eine gewisse Standardabweichung überschreiten. Zeitreihen mit einem niedrigen Informationsgehalt bleiben also unberücksichtigt.

- Fehlende Werte können erst mit Bildung der Zeitreihen bestimmt werden, diese werden entsprechend bereinigt.
- Die Maxima einiger Haushalte bezüglich der Stromessdaten sind deutlich höher als bei anderen Haushalten. Diese Ausreißer wurden aufgrund ihrer geringen Häufigkeit manuell bereinigt.

5.3 Data Preparation

Zunächst sollen die Strommessdaten mit den Informationen zu variablen Tarifen verknüpft werden. Hierzu werden die beiden Tabellen „electricity-egauge-15min“ und „metadata“ über das Attribut „dataid“ zusammengefasst.

dataid	local_15min	car1	program_ccet_group
6101	2014-09-28 23:45:00	3.309400	CCET - Pricing Trial
6101	2014-09-29 00:00:00	3.311400	CCET - Pricing Trial
6101	2014-09-29 00:15:00	3.309600	CCET - Pricing Trial
6101	2014-09-29 00:30:00	1.686667	CCET - Pricing Trial
6101	2014-09-29 00:45:00	0.357267	CCET - Pricing Trial
6101	2014-09-30 11:00:00	1.833533	CCET - Pricing Trial
6101	2014-09-30 11:15:00	3.317467	CCET - Pricing Trial
6101	2014-09-30 11:30:00	3.305133	CCET - Pricing Trial
6101	2014-09-30 11:45:00	3.308133	CCET - Pricing Trial
6101	2014-09-30 12:00:00	2.173733	CCET - Pricing Trial
4336	2013-08-14 00:00:00	5.989600	CCET - Control
4336	2013-08-14 00:15:00	6.006267	CCET - Control
4336	2013-08-14 00:30:00	6.021467	CCET - Control
4336	2013-08-14 00:45:00	6.028867	CCET - Control
4336	2013-08-14 01:00:00	6.041933	CCET - Control
4336	2013-08-14 01:15:00	6.056533	CCET - Control
4336	2013-08-14 01:30:00	6.067400	CCET - Control
4336	2013-08-14 01:45:00	6.065200	CCET - Control
4336	2013-08-14 02:00:00	5.845733	CCET - Control
4336	2013-08-14 07:30:00	0.034200	CCET - Control
4336	2013-08-14 17:15:00	0.001000	CCET - Control

Abbildung 10: Zusammengefasste Tabelle mit Informationen zu variablen Tarifen

Die Strommessdaten können nun einer Verbraucherklasse zugeordnet werden. Um die Cluster für jede Verbraucherklasse einzeln ausfindig zu machen, muss der Datensatz entsprechend aufgeteilt werden. Über die Spalte „program_ccet_group“ werden die entsprechenden Untermengen ausgewählt und in zwei kleineren Tabellen separiert gespeichert. Eine solche Tabelle ist in der folgenden Abbildung dargestellt.

dataid	local_15min	car1	program_ccet_group
4336	2013-08-14 00:00:00	5.989600	CCET - Control
4336	2013-08-14 00:15:00	6.006267	CCET - Control
4336	2013-08-14 00:30:00	6.021467	CCET - Control
4336	2013-08-14 00:45:00	6.028867	CCET - Control
4336	2013-08-14 01:00:00	6.041933	CCET - Control
4336	2013-08-14 01:15:00	6.056533	CCET - Control
4336	2013-08-14 01:30:00	6.067400	CCET - Control
4336	2013-08-14 01:45:00	6.065200	CCET - Control
4336	2013-08-14 02:00:00	5.845733	CCET - Control
4336	2013-08-14 07:30:00	0.034200	CCET - Control

Abbildung 11: Strommessdaten der Verbraucherklasse C

Die Spalte „program_ccet_group“ wird nun gelöscht, weil diese nicht mehr benötigt wird. Im Folgenden werden aus der Spalte „local_15min“ drei neue Attribute abgeleitet.

- „Weekday“. Enthält den Wochentag (Mo - So), an dem die Strommessung angefallen ist. Wird benötigt um die Wochenenden (Sa + So) aus dem Datensatz zu entfernen.
- „Date“. Enthält das Datum, an dem die Strommessung angefallen ist. Wird benötigt um die Zeitreihen zu erzeugen.
- „Hour“. Enthält die Uhrzeit, zu der die Strommessung angefallen ist. Wird benötigt um die Zeitreihen zu erzeugen.

dataid	local_15min	car1	weekday	hour	date
4336	2013-08-14 00:00:00	5.989600	Wed	00:00	2013-08-14
4336	2013-08-14 00:15:00	6.006267	Wed	00:15	2013-08-14
4336	2013-08-14 00:30:00	6.021467	Wed	00:30	2013-08-14
4336	2013-08-14 00:45:00	6.028867	Wed	00:45	2013-08-14
4336	2013-08-14 01:00:00	6.041933	Wed	01:00	2013-08-14
4336	2013-08-14 01:15:00	6.056533	Wed	01:15	2013-08-14
4336	2013-08-14 01:30:00	6.067400	Wed	01:30	2013-08-14
4336	2013-08-14 01:45:00	6.065200	Wed	01:45	2013-08-14
4336	2013-08-14 02:00:00	5.845733	Wed	02:00	2013-08-14
4336	2013-08-14 07:30:00	0.034200	Wed	07:30	2013-08-14

Abbildung 12: Strommessdaten der Klasse C mit abgeleiteten Zeitvariablen

Es können nun alle Wochenenden aus dem Datensatz über die Spalte „weekday“ entfernt werden. Anschließend werden die Spalten „weekday“ und „local_15min“ nicht mehr benötigt. Aufgrund der abweichenden Spitzenlastzeiten im Sommer und Winter müssen die Daten erneut anhand ihres Datums aufgeteilt werden. Für den Versuchszeitraum sind die Jahreszeiten wie folgt bestimmt.

- 01.04.2013 – 31.05.2013 : Winterzeit (2 Monate)
- 01.06.2013 – 30.09.2013: Sommerzeit (4 Monate)
- 01.10.2013 – 31.05.2014: Winterzeit (8 Monate)
- 01.06.2014 – 30.09.2014: Sommerzeit (4 Monate)

Der 18 Monate andauernde Versuchszeitraum teilt sich also in 10 Monate Winterzeit und 8 Monate Sommerzeit auf. Der Datensatz wird entsprechend aufgeteilt. Außerdem werden die Strommessdaten auf 2 Nachkommastellen abgerundet. Abbildung 8 zeigt eine finalisierte Tabelle mit Strommessdaten der Klasse C in seinen Winterphasen.

dataid	car1	hour	date
4336	5.89	22:00	2013-10-01
4336	6.00	22:15	2013-10-01
4336	6.01	22:30	2013-10-01
4336	6.00	22:45	2013-10-01
4336	6.02	23:00	2013-10-01
4336	6.01	23:15	2013-10-01
4336	6.03	23:30	2013-10-01
4336	6.03	23:45	2013-10-01
4336	6.01	00:00	2013-10-02
4336	6.02	00:15	2013-10-02

Abbildung 13: Abgerundete Strommessdaten der Verbraucherklasse C in den Winterphasen ohne Wochenenden.

Die folgende Tabelle soll nun einen kurzen Überblick die Anzahl der jeweiligen Datensätze pro Verbraucherklasse und Jahreszeit geben. Die Gesamtzahl der Datensätze ist nun reduziert, weil die Wochenenden entfernt wurden.

Verbraucherklasse	Jahreszeit	Anzahl der Datensätze
C	Winter	130.857
C	Sommer	90.103
DP	Winter	559.279
DP	Sommer	463.224
Gesamt	-	**1.243.463**

Tabelle 4: Anzahl der Datensätze pro Verbraucherklasse

Die Strommessdaten können nun zu Zeitreihen zusammengesetzt werden. Dies geschieht über einen „Pivot“-Befehl aus der Panda-Bibliothek. Die Spalten „dataid“ und „date“ bilden hier den eindeutigen Multiindex der neuen Tabelle, um jede Zeitreihe zu identifizieren. Außerdem bilden die möglichen Merkmalsausprägungen der Spalte „hour“ nun alle Zeitintervalle eines Tages im 15-Minuten-Takt ab. Die Strommessdaten wurden korrekt dem entsprechenden Zeitintervall zugeordnet.

```
hour              00:00  00:15  00:30  00:45  01:00  01:15  01:30  01:45
dataid date
1642   2013-10-01   0.0    0.0    0.0    0.0    0.0    0.0   0.00    0.0
       2013-10-02   0.4    0.0    0.0    0.0    0.0    0.0   0.00    0.0
       2013-10-03   0.0    0.0    0.0    0.0    0.0    0.2   0.19    0.0

hour              02:00  02:15  ...    21:30  21:45  22:00  22:15  22:30
dataid date                     ...
1642   2013-10-01   0.0    0.0  ...     0.00   0.01   0.01   0.00   0.00
       2013-10-02   0.0    0.0  ...     0.01   0.01   0.00   0.26   0.17
       2013-10-03   0.0    0.0  ...     0.01   0.01   0.00   0.00   0.00

hour              22:45  23:00  23:15  23:30  23:45
dataid date
1642   2013-10-01   0.0   0.01    0.0   0.00   0.01
       2013-10-02   0.0   0.00    0.0   0.00   0.00
       2013-10-03   0.0   0.00    0.0   0.46   0.00
```

Abbildung 14: Zeitreihen der Verbraucherklasse C in den Winterphasen.

Die Zeitreihen liegen in kontinuierlicher, gleichmäßig abgetasteter und univariater Form vor. Alle Zeitreihen sind dabei gleicher Länge und müssen nicht mehr entsprechend angepasst werden. Nicht für alle Zeitintervalle gibt es entsprechende Messungen. Fehlende Werte machen die Zeitreihen unvollständig und müssen

bereinigt werden. Die folgende Tabelle gibt die Anzahl der Zeitreihen, als auch die Anzahl fehlender Werte pro Datensatz wieder.

Datensatz	Anzahl der Zeitreihen	Fehlende Werte - Absolut	Fehlende Werte – Relativ	Unvollständige Zeitreihen (>=50%)
C – Winter	1.376	375	0,2%	4
C - Sommer	940	137	0,1%	1
DP - Winter	5.850	2321	0,4%	27
DP - Sommer	4.845	1.896	0,4%	20

Tabelle 5: Übersicht - Zeitreihen und Fehlende Werte

Nach Konstruktion der Zeitreihen sind nur sehr wenig fehlende Werte aufgetreten. Die relative Quote an fehlenden Werte liegt in alle Datensätzen unter 1%. Unvollständige Zeitreihen, in denen mindestens die Hälfte der Werte fehlt, sind kaum vorhanden. Diese werden von der Analyse ausgeschlossen. Alle übrigbleibenden fehlenden Werte werden durch 0 ersetzt. Anschließend werden nur diejenigen Zeitreihen als Eingabe für das Datenmodell verwendet, die ein Maximum und eine Standardabweichung von 1 überschreiten.

5.4 Modeling

Die aufbereiteten und selektierten Zeitreihen können nun als Eingabe für das Datenmodell verwendet werden. Die Modellierung soll in dieser Arbeit durch den Algorithmus k-Means erfolgen. Um das Geschäftsziel zu erreichen, werden 2-dimensionale Verbrauchsmuster benötigt. Anhand der Überschneidungen mit den Spitzenlastzeiten können die Verbrauchsmuster in Bezug auf die Gefährdung der Netzstabilität bewertet werden. Ausschließlich k-Means hat die Möglichkeit Repräsentanten der Cluster in Form von 2-dimensionalen Mittelwertkurven zu erzeugen. Wegen der Zielsetzung kommen keine Alternativen zu k-Means für die Modellierung der Daten in Betracht. Zwar müssen die Anzahl der Cluster zu Beginn der Berechnungen vorher bestimmt werden, allerdings können die Zeitreihen dafür

während der Verschiebung ihrer Clusterzentren ihre Clusterzugehörigkeit wechseln. Neben der geringen Parameterzahl liegt der Stärke von k-Means in seiner schnellen Performance.[23] Außerdem kann k-Means auf Verteilungen jeder Art angewendet werden. K-Means verwendet standardmäßig die euklidische Distanz, welche nach Iglesias et al (2013) sehr gute Ergebnisse in der Erzeugung von Lastprofilen erzielt hat. Die nachfolgende Tabelle beschreibt die Parameter von k-Means sowie die ausgewählten Werte für jeden Parameter. Die angegebenen Parameter wirken sich auf die Effektivität nicht auf die Effizienz des Algorithmus aus. Auf den Effizienz-Aspekt soll hier nicht weiter eingegangen werden. Schließlich hat dieser keinen Einfluss auf das Ergebnis.

Parameter	**Bedeutung**	**Wert**
n_clusters	Die Anzahl der zu formenden Cluster.	1 – 5. Mehrere Durchläufe.
n_init	Anzahl der Durchläufe mit jeweils unterschiedlicher Initialisierung der Clusterzentren. Der beste Durchlauf wird hierbei ausgewählt.	10
tol	Relatives Toleranzkriterium zur Bestimmung der Konvergenz.	0.0001

Tabelle 6: Parametereinstellung von k-Means

5.5 Evaluation

Als Metrik zur Messung der Clusterqualität wird der Silhouettenkoeffizient verwendet. Die Anzahl der zu formenden Cluster wird in mehreren Durchläufen inkrementiert und jeweils der durchschnittliche Silhouettenkoeffizient berechnet. Die optimale Anzahl von Clustern wird anhand des höchsten durchschnittlichen Silhouettenkoeffizienten

[23] Vgl. MacQueen (1967), S. 281

bestimmt. Der Nachteil der vorherigen Festlegung der Clusteranzahl wird somit eliminiert. Für jeden Datensatz werden mithilfe von Silhouetten-Darstellungen die Einteilungen der Zeitreihen in die Cluster sichtbar. Die folgende Auflistung zeigt die Silhouetten-Darstellungen jedes Datensatzes, mit dem höchsten durchschnittlichen Silhouettenkoeffizienten.

Verbraucherklasse: C

Jahreszeit: Winter

Allgemeiner Silhouettenkoeffizient: 0,29

Cluster: 2

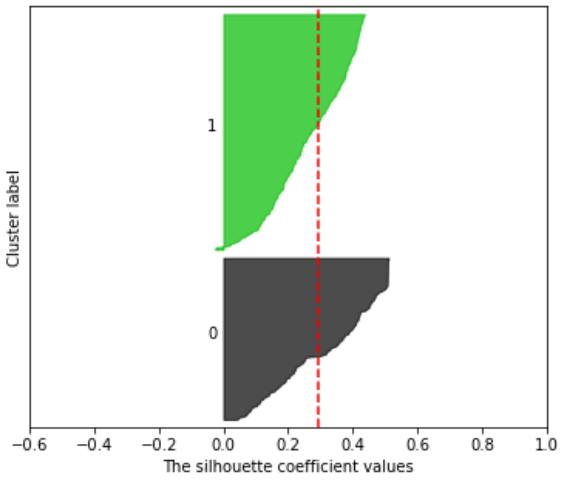


Abbildung 15: Silhouetten-Darstellung der Klasse C im Winter

Verbraucherklasse: C

Jahreszeit: Sommer

Allgemeiner Silhouettenkoeffizient: 0,27

Cluster: 2

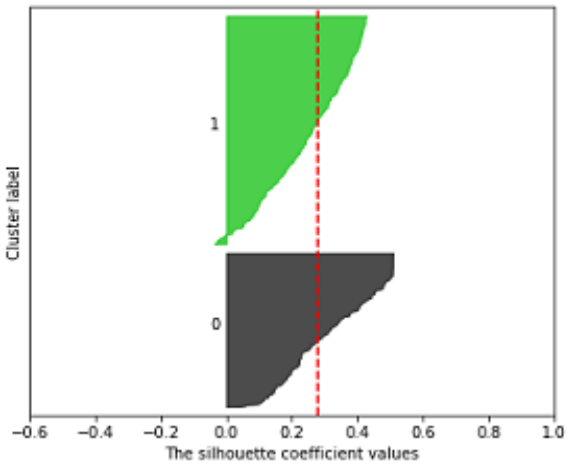


Abbildung 16: Silhouetten-Darstellung der Klasse C im Sommer

Verbraucherklasse: DP

Jahreszeit: Winter

Allgemeiner Silhouettenkoeffizient: 0,24

Cluster: 5

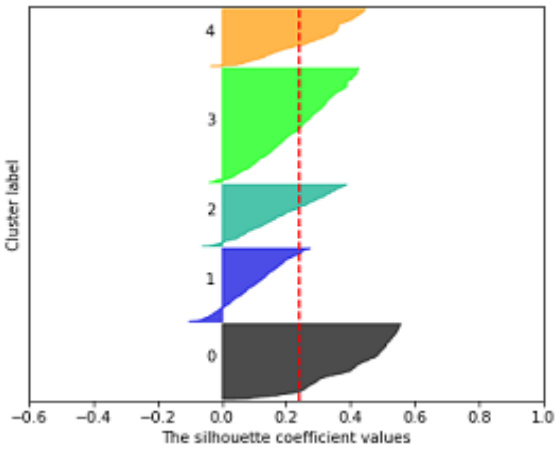


Abbildung 17: Silhouetten-Darstellung der Klasse DP im Winter

Verbraucherklasse: DP

Jahreszeit: Sommer

Allgemeiner Silhouettenkoeffizient: 0,22

Cluster: 5

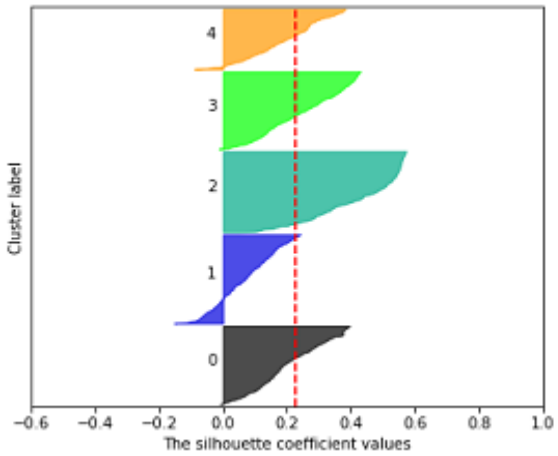


Abbildung 18:Silhouetten-Darstellung der Klasse DP im Sommer

Es wurde für jeden Datensatz ein positiver durchschnittlicher Silhouettenkoeffizient erzielt. Somit ist Qualität der Clusterbildung als ausreichend im Hinblick auf die Zielsetzung zu betrachten. Interessanter Weise variiert die Anzahl der Cluster nur im Hinblick auf die Verbraucherklassen, nicht aber in Bezug auf die Jahreszeit. Die Cluster müssen aber genauer untersucht werden, um hier eine endgültige Aussage über den Einfluss der Jahreszeit auf die Verbrauchsmuster zu treffen. Eine falsche Einordnung der Zeitreihen in einen Cluster ist bei allen Datensätzen sehr selten aufgetreten. Allerdings ist der Silhouettenkoeffizient bei einigen Zeitreihen sehr niedrig. Dies ist ein Indiz dafür, dass diese Zeitreihen nicht eindeutig einem Cluster zugehörig sind, bzw.

diese nicht markant voneinander zu trennen sind. Dies ist auf die Tatsache zurückzuführen, dass zu jedem Zeitpunkt Strom bezogen wird, wenn man alle Zeitreihen aggregiert über 24 Stunden betrachtet. Es können also zu jedem Zeitpunkt kleine lokale Clusterzentren entstehen. Das hat einen negativen Einfluss auf den Silhouettenkoeffizient zur Folge. Eine endgültige Aussage über die Unterscheidbarkeit der Cluster kann auch hier erst mit Visualisierung der Cluster getroffen werden.

5.6 Deployment

Die Cluster werden nun zunächst für jeden Datensatz visualisiert und der Einfluss dynamischer Preisberechnung auf die Verbrauchsmuster bestimmt.

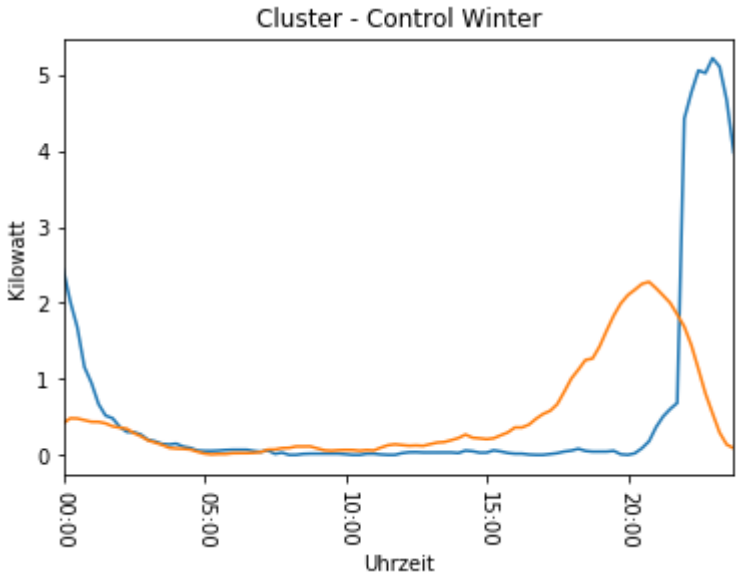


Abbildung 19: Visualisierung der Cluster von Verbraucherklasse C im Winter

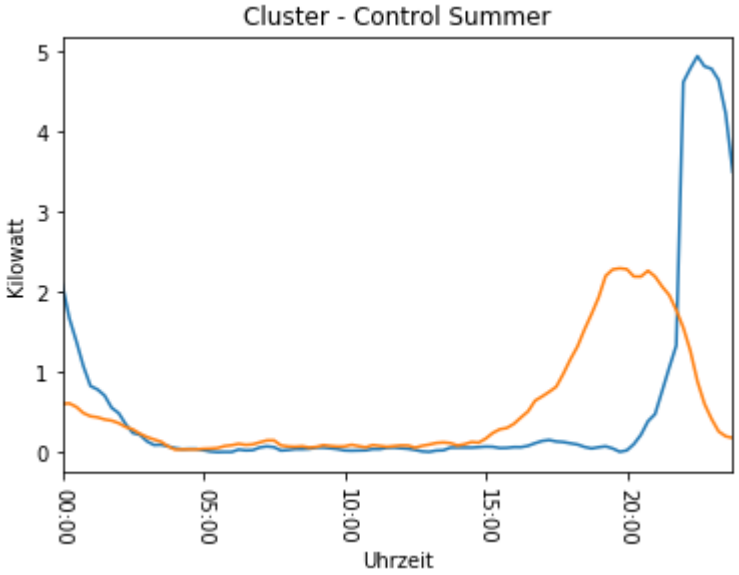


Abbildung 20: Visualisierung der Cluster von Verbraucherklasse C im Sommer

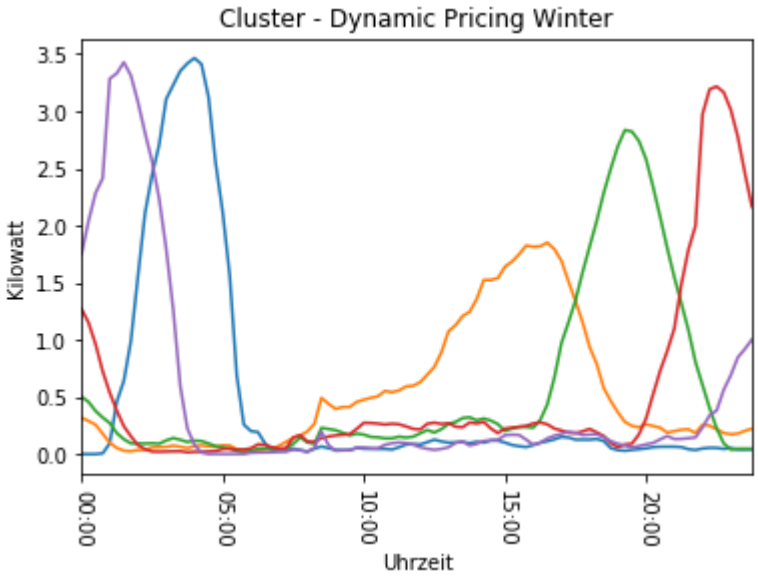


Abbildung 21:Visualisierung der Cluster von Verbraucherklasse DP im Winter

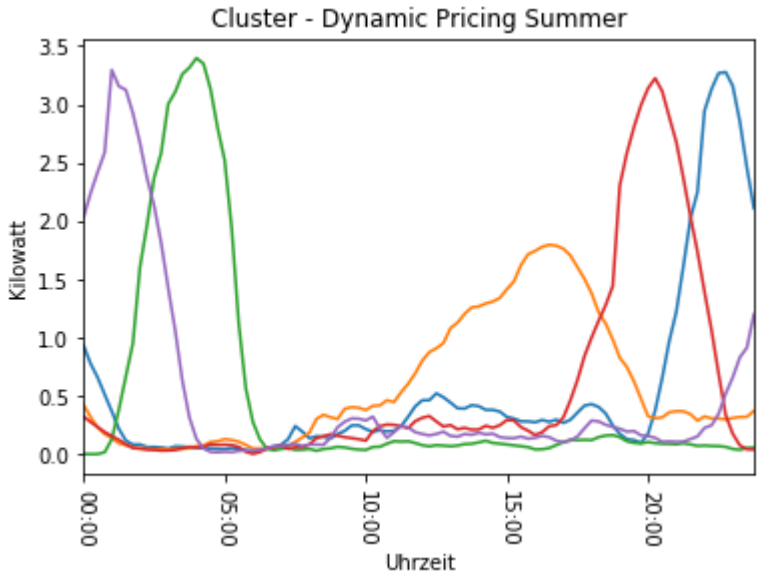


Abbildung 22: Visualisierung der Cluster von Verbraucherklasse DP im Sommer

Die Visualisierung der Cluster erweckt zunächst den Eindruck, dass die Jahreszeit einen geringen Einfluss auf die Struktur der Verbrauchsmuster hat. Allerdings variieren die Häufigkeiten, als auch die Zeitspannen der Cluster bedeutsam. Die Cluster überschneiden sich teilweise, stellen aber dennoch markante, leicht unterscheidbare Strukturen dar. Das zweite Data-Mining-Ziel, die Clusterbildung wurde somit erreicht. Auffällig ist weiterhin, dass sich in der Verbraucherklasse DP zwei Cluster in der Nachtzeit befinden. Dies stellt ein erstes Indiz dafür dar, dass eine dynamische Preisberechnung einen positiven Effekt auf die Verbrauchsmuster hat. Weiterhin wird deutlich, dass die Spitzenlast in der Gruppe DP deutlich niedriger ist und gleichmäßiger verteilt ist.

Es lassen sich für die weitere Untersuchung die folgenden Schlüsse aus der Visualisierung ziehen.

- Die aktiven Ladephasen der Cluster sind unterschiedlich lang und erreichen unterschiedlich hohe Spitzenwerte. Die Relevanz eines Clusters in Bezug auf die Gefährdung der Netzstabilität hängt von seiner Struktur ab. Es werden der aktiven Ladephase eines Clusters nur Uhrzeiten zugeordnet, die einen festgelegten Wert überschreiten.
- Die aktive Ladephase eines Cluster überschneidet sich oft nur teilweise mit der Spitzenlastzeit. Die relative Abdeckung eines Clusters wird als kritisch gewertet, nicht aber ein Cluster als Ganzes.

Die Kennzahlen zur Bestimmung der Wirkungsweise dynamischer Preisberechnung auf die Verbrauchsmuster sind in der folgenden Tabelle dargestellt.

Klasse	**Jahreszeit**	**Cluster**	**Aktive Ladephase (> 1,5 Kilowatt)**	**Dauer**	**Abdeckung - Absolut**	**Abdeckung - Relativ**	**Häufigkeit**
C	Winter	1	22:00 – 00:30	2,5H	0H	0%	216
		2	19:15 – 22:00	2,75H	1,75H	63%	314
	Sommer	1	22:00 – 00:15	2,25H	0H	0%	182
		2	18:30 – 21:00	2,5H	1,5H	60%	269
DP	Winter	1	02:00 – 05:15	3,25H	0,25H	7%	376
		2	14:15 – 17:15	3H	0,25H	8%	465
		3	17:45 – 21:00	3,25H	3,25H	100%	307
		4	21:30 – 23:45	2,25H	0H	0%	568
		5	00:00 - 03:00	3H	0H	0%	285
	Sommer	1	21:15 - 23:45	2,5H	0H	0%	346
		2	15:30 - 17:45	2,25H	2,25H	100%	393
		3	02:00 – 05:15	3,25H	0H	0%	356
		4	19:00 – 21:45	2,75H	1H	36%	341
		5	00:00 – 02:45	2,75H	0H	0%	266

Tabelle 7: Kennzahlen pro Cluster zur Bestimmung der Wirkungsweise dynamischer Preisberechnung auf die Verbrauchsmuster

Zunächst werden für jeden Cluster die aktiven Ladephasen definiert. Zu diesen Zeiten überschreiten die Strommessdaten eines Clusters einen festgelegten Wert. Den unterschiedlichen Strukturen der Cluster wird somit Rechnung getragen. Anschließend wird die Dauer der aktiven Ladephase, als auch die absolute Abdeckung bestimmt. Diese gibt an, wie lange sich die aktive Ladephase eines Clusters mit der Spitzenlastzeit überschneidet. Die relative Abdeckung gibt hierbei einen prozentualen Wert wieder, der die absolute Abdeckung und die Dauer ins Verhältnis setzt. Zuletzt wird die Anzahl der Zeitreihen angeben, welche einem Cluster zugeordnet wurden. Die Auftrittshäufigkeiten kritischer Ladephasen können für jede Verbraucherklasse in das Verhältnis aller Auftrittshäufigkeiten gesetzt werden.

Rechnung – Klasse C

(314 * 0,63 + 269 * 0,6) / (216 + 314 + 182 + 269) = 359 / 981 = **36,5%**

Rechnung – Klasse DP

(376 * 0,07 + 465 * 0,08 + 307 * 1 + 393 * 1 + 341 * 0,36) / (376 + 465 + 307 + 568 + 285 + 346 + 393 + 356 + 341 + 266) = 886 / 3703 = **23,9%**

Die relative Anzahl kritischer Ladephasen ist bei der Verbraucherklasse C um 12,6% höher als bei der Verbraucherklasse DP und hat somit den festgelegten Schwellenwert deutlich überschritten. Weiterhin wurde gezeigt, dass sich die Cluster bei der Gruppe DP in die Nacht verlagern und eine niedrigere maximale Spitzenlast aufweisen. Eine dynamische Preisberechnung hat somit einen positiven Einfluss auf die Verbrauchsmuster.

6 Schluss

Inwieweit eine dynamische Preisberechnung einen positiven Effekt auf das Nutzungsverhalten von Autofahrern elektrischer Fahrzeuge hat, und somit einem hohen Durchdringungsniveau entgegenwirken kann, ist Gegenstand dieser Arbeit gewesen. Zunächst wurde die Problemstellung erläutert und der aktuelle Stand der Forschung in den Bereichen der dynamischen Preisberechnung, sowie der Clusteranalyse zeitbasierter Energieverbrauchsdaten wiedergegeben. Technische Grundlagen wurden erläutert, welche für das Verständnis der Clusteranalyse essentiell sind. Außerdem wurde das Referenzmodell CRISP-DM für eine strukturierte Umsetzung der Problemstellung vorgestellt und seine Vorteile aufgezeigt. Aufgrund seiner umfangreichen Dokumentation war es die richtige Entscheidung dieses Prozessmodell zu verwenden, obwohl die iterative Natur dieses Modells nicht genutzt worden ist. In der Phase Business Understanding im Hauptteil wurden zunächst eine exakte, messbare Zielsetzung, sowie die Verbraucherklassen, mögliche Ausgangsszenarien, als auch die Spitzenlastzeiten definiert. Ein kurzer Überblick über den Umfang und die Abstammung der Datenquelle, als auch der dazugehörigen Organisation wurde gegeben. Zusätzlich wurden die formalen Voraussetzungen an die Datenqualität definiert. Das Geschäftsziel wurde in die entsprechenden Data-Mining-Ziele übersetzt und erste Erkenntnisse zur Erreichung und Qualitätsbestimmung dieser Ziele gewonnen. In der Phase Data Understanding wurde zunächst abgegrenzt, welche Daten für die Clusteranalyse benötigt werden. Diese wurden anhand ihrer statistischen Merkmale beschrieben. Weitere Erkenntnisse zur Bestimmung der Datenqualität wurden durch Verteilungsgraphen der gesamten Strommessdaten pro Verbrauchergruppe und statistischen Kennzahlen pro Haushalt gesammelt. Die Datenqualität wurde in Bezug zur Erzeugung der Zeitreihen als ausreichend bestimmt. Probleme in der Datenqualität wurden ausfindig gemacht und Lösungsansätze zur

Verbesserung aufgezeigt. In der Phase Data Preparation wurden die Zeitreihen zunächst erzeugt und den entsprechenden Kombinationen aus Verbraucherklasse und Jahreszeit zugeordnet. Hierfür kamen typische Data-Cleansing-Ansätze, wie beispielsweise dem Ableiten, Löschen, der Zusammenführung oder Pivotisierung von Attributen, zum Einsatz. Außerdem wurden fehlende Werte und Ausreißer bereinigt. Die Datenvorverarbeitungsschritte haben qualitativ hochwertige Datensätze erzeugt. Allerdings wären weitere Dimensionsreduktionsschritte nötig gewesen, um markantere Verbrauchsmuster herauszuarbeiten. In der Phase Modeling wurde auf die verwendeten Parameter des Algorithmus eingegangen, beziehungsweise warum dieser verwendet wurde. Aufgrund seiner Eigenschaft, Repräsentanten der Cluster in Form von 2-dimensionalen Mittelwertkurven zu erzeugen, war es die richtige Entscheidung den Algorithmus k-Means für die Modellierung zu verwenden. Schließlich wurde in der Evaluation-Phase die Clusterqualität mittels des Silhouettenkoeffizienten bestimmt. Dieses Evaluierungskriterium hat sehr gut funktioniert und lieferte wichtige Erkenntnisse in Bezug auf die Daten- und Modellierungsqualität. Für zukünftige Arbeiten könnten hier allerdings noch weitere Evaluierungskriterien angewendet werden. Zuletzt wurden in der Phase Deployment die Verbrauchsmuster für jeden Datensatz visualisiert und ihre Gefährdung in Bezug auf die Netzstabilität bestimmt. Das hierbei verwendete Verfahren könnte weiterhin verbessert werden. So ist die Bestimmung der aktiven Ladephase eines Verbrauchsmusters etwas ungenau. Außerdem kann es, neben der Auftrittshäufigkeit eines Verbrauchsmusters und der Zeitüberschneidung eines Musters mit einer Spitzenlastzeit, weitere Faktoren zur Bestimmung der Gefährdung in Bezug auf der Netzstabilität geben. Für zukünftige Arbeiten kann ein weitaus komplexeres und genaueres Verfahren zur Bewertung der Verbrauchsmuster entwickelt werden.

Dieses Kapitel präsentiert weiterhin die empirischen Ergebnisse in Kapitel 6.1 und anschließend die Limitationen, welche für diese Arbeit gelten, in Kapitel 6.2.

6.1 Empirische Ergebnisse

Zusammenfassend lassen sich die folgenden Ergebnisse aus Kapitel 5 festhalten:

Datenqualität

- Die beiden Verbrauchergruppen sind sich bezüglich ihrer statistischen Merkmale sehr ähnlich. Lediglich die maximalen Werte weichen in der Verbrauchergruppe C stark von der Gruppe DP ab. Ausreißer wurden dabei manuell entfernt.
- Das 75%-Quartil der Strommessdaten aller Haushalte liegt nahe bei 0. Viele Zeitreihen enthalten also nur wenige Informationen. Diese wurden herausgefiltert.
- Es wurde ein Ungleichgewicht bezüglich der Anzahl der Datensätze festgestellt, ein relatives Vergleichskriterium hat dieses ausgeglichen. Eine weitere Möglichkeit dieses Ungleichgewicht auszugleichen besteht darin, eine zufällige Stichprobe aus der Datenmenge der Verbraucherklasse DP zu entnehmen, sodass diese der Datenmenge der Verbraucherklasse C entspricht.
- Fehlende Werte kamen nur wenige vor und haben die Datenqualität nicht merklich beeinflusst. Unvollständige Zeitreihen wurden von der Analyse ausgeschlossen, andernfalls wurden diese Werte substituiert.

Modellierung

- Die Anzahl der Cluster musste im Voraus bekannt sein. Die optimale Anzahl an Clustern wurde durch den durchschnittlichen Silhouettenkoeffizient bestimmt. Es wurden 5 Cluster bei der Gruppe DP, sowie 2 Cluster bei der Gruppe C entdeckt.
- Die Cluster unterscheiden sich nicht markant voneinander. Dies wird durch die sich überschneidenden Clusterschwerpunkte deutlich. Dennoch ist die Qualität der Analyse ausreichend, um ein Fazit zu ziehen.
- Die Anzahl der Cluster variiert nur in Bezug auf die Verbrauchergruppe, nicht aber in Bezug auf die Jahreszeit.

Wirkungsweise dynamischer Preisberechnung

- Die variablen Tarife haben keinen Einfluss auf die Höhe des gesamten Strombedarfs, dieser verlagert sich ausschließlich zeitlich.
- Die Spitzenlast ist in der Gruppe DP deutlich geringer und gleichmäßiger verteilt, außerdem bilden sich in der Nachtzeit deutliche Cluster heraus.
- Die relative Anzahl kritischer Ladephasen ist bei der Verbraucherklasse C um 12,6% höher als bei der Verbraucherklasse DP.
- Es ist von einem positiven Einfluss dynamischer Preisberechnung auf das Nutzungsverhalten auszugehen. Dies unterstützt die Thesen von Faruqui et al. (2009).

6.2 Limitierungen

Die folgenden Limitierungen gelten für diese Arbeit:

- Obwohl unterschiedliche Möglichkeiten bestehen, fehlende Werte zu ersetzen, wurden diese ausschließlich durch 0 ersetzt.
- Aufgrund der zeitlichen Limitierung wurde kein automatisiertes Verfahren zur Erkennung von Ausreißern implementiert.
- Neben dem Rohdaten-Ansatz könnte der Feature-basierte Ansatz weitere wichtige Erkenntnisse liefern.
- k-Means wurde ausschließlich mit der euklidischen Distanz angewendet. Andere Distanzmaße wurden nicht getestet. Weitere Konvergenzkriterien waren nicht Teil dieser Analyse. Es wurden nur bis zu 5 Cluster pro Verbrauchergruppe getestet.
- Komplexe Verfahren zur Auswertung von Clustern in Bezug die Gefährdung der Netzstabilität konnten nicht gefunden und somit angewendet werden.
- Der Wert zur Definierung der aktiven Ladephasen eines Clusters Verbrauchsmuster wurde willkürlich gewählt, gilt aber dennoch für beide Verbraucherklassen gleichermaßen und wurde in Bezug auf die Ergebnisse bestimmt.

Des Weiteren gelten die folgenden äußeren Beschränkungen:

- Insbesondere der Silhouettenkoeffizient hat gezeigt, dass einige Zeitreihen nicht besonders gut einem Cluster zuzuordnen waren. Die Datenqualität ist formal betrachtet nicht schlecht. Allerdings tritt über den Tag verteilt immer ein gewisser Stromverbrauch auf. Die Verbrauchsmuster überschneiden sich dadurch und sind teilweise nicht markant voneinander zu trennen. Weitere Dimensionsreduktionsschritte könnten markantere Verbrauchsmuster herausarbeiten.
- Es sind keine Details zu den Tarifen bekannt, weshalb keine genaueren Schlüsse aus der Beziehung, zwischen Tarif und den Verbrauchsmustern, gezogen werden können.

6.3 Ausblick

Anhand der in Kapitel 6.2 aufgeführten Limitierungen lassen sich weitere Schritte ableiten, um die Wirkungsweise dynamischer Preisberechnung auf das Ladeverhalten der Autofahrer genauer zu bestimmen. Im Folgenden werden mögliche Schritte erläutert, welche aufbauend auf dieser Arbeit durchgeführt werden können.

- Forward-Filling: In der Datenvorverarbeitungsphase besteht die Möglichkeit fehlende Werte statt mit 0, durch den letzten bekannten Wert zu ersetzten.
- LocalOutlierFactor: Dieser Algorithmus wird in der Python-Bibliothek sklearn.neighbors zur Verfügung gestellt. Er ist in der Lage Ausreißer in Daten zu erkennen und somit die Datenqualität zu erhöhen. Diese Strategie eignet sich insbesondere für Zeitreihen, weil ein Ausreißer anhand seiner benachbarten Datenpunkte, also den Vorgängern und Nachfolgern, bestimmt werden kann.
- Feature-basierte Ansatz: Wie bereits in Kapitel 3.2 vorgestellt, besteht die Möglichkeit unterschiedliche Kennzahlen aus den Zeitreihen zu errechnen und

diese Kennzahlen anschließend einem Cluster zuzuordnen. Die Datenqualität ist weniger anfällig gegenüber Ausreißern, allerdings sind die Repräsentanten der Cluster selbst nicht mehr in Form von 2-dimensinalen Zeitreihen verfügbar. Hierbei müsste ebenfalls ein komplett neuer Ansatz für die Evaluierung des Ergebnisses genutzt werden.

- Es besteht außerdem die Möglichkeit, die Zeitreihen in eine markantere Form mittels Dimensionsreduktionsschritten zu bringen. Die Repräsentanten der Cluster würden hierbei als Zeitreihen erhalten bleiben.
- K-Means kann mit weiteren Distanzmaßen, Konvergenzkriterien oder Clustergrößen getestet werden.
- Es können neben dem Sillhouettenkoeffizient weitere Evaluierungskriterien zur Bestimmung der Clusterqualität angewendet werden.
- Um die Cluster besser analysieren zu können, werden komplexer Verfahren zur Auswertung von Clustern in Bezug die Gefährdung der Netzstabilität benötigt.
- Der Fokus in dieser Arbeit liegt, aufgrund der hohen benötigten Energiemenge dieses Postens, auf den Strommessdaten häuslicher Fahrzeugladestation. Die verwendeten Techniken lassen sich selbstverständlich auch auf andere Messdaten anwenden.

A. Anhang

Phase	Arbeitspaket	Inhalt	Grund des Verzichts
Business Understanding	Ressourceninventar	Beschreibt die personellen und informellen Projektressourcen, als auch die zur Verfügung stehende Hardware.	Sowohl die Daten, als auch die verwendete Software werden beschrieben. Eine Auflistung personeller Projektressourcen oder der verwendeten Hardware ist zu Erreichung der Zielsetzung nicht nötig.
Business Understanding	Risiken, Kontingente	Listet mögliche Projektrisiken auf und stellt mögliche Kontingente zur Beseitigung dieser Risiken auf.	Projektrisiken können maximal in Bezug auf die Datenqualität auftreten und werden deshalb nicht zusätzlich aufgelistet.
Business Understanding	Kosten-/Nutzenanalyse	Vergleicht Projektkosten und potentielle Gewinne miteinander.	Diese Arbeit dient der Wissensgenerierung. Projektkosten und potentielle Gewinne sind schwer bis überhaupt nicht abschätzbar.
Business Understanding	Projektplan	Beschreibt die einzelnen Schritte in ihrer zeitlichen Reihenfolge zur Erreichung der Ziele.	Unangemessen in Bezug auf die Projektressourcen und den Umfang dieser Arbeit. Das Referenzmodell reicht für ein geplantes Vorgehen aus.
Deployment	Wartungsplan	Spezifiziert, wie die Ergebnisse gewartet werden sollen.	Das Modell dient der Wissensgenerierung und wird nicht in das tägliche Geschäft eines Unternehmens integriert. Von daher ist keine Wartung nötig.

Tabelle 8: Nicht verwendete Arbeitspakete des Referenzmodells CRIPS-DM

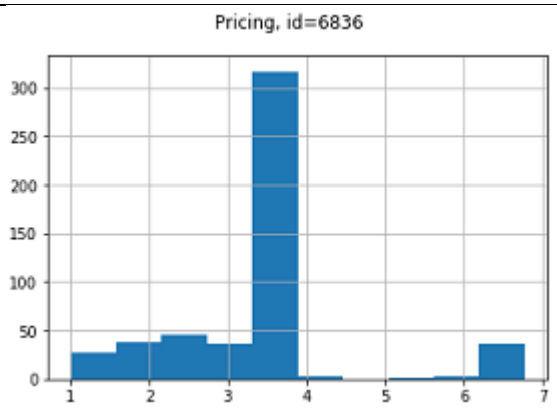
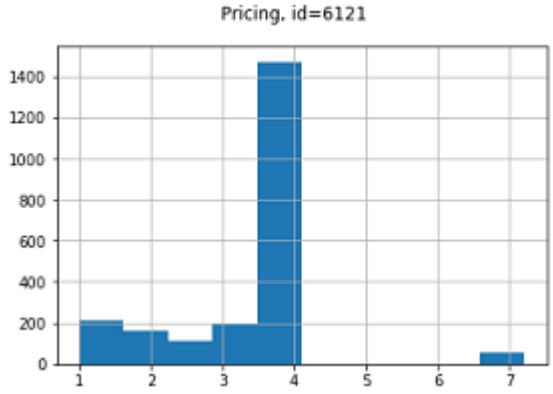
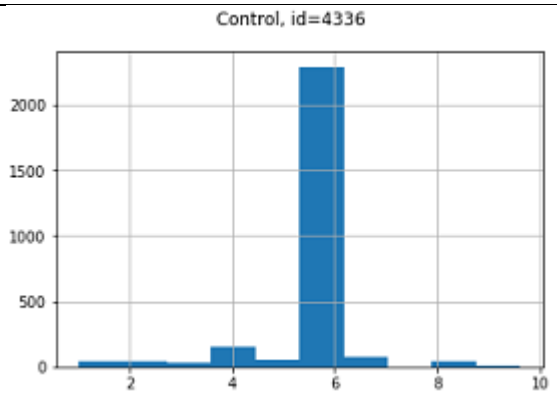
Haushalt	Anzahl	Durch-Schnitt	Standard-Abweichung	75%-Quartil	Maximum
2814	37.632	0,17	0,98	0,00	6,73
9499	30.410	0,22	0,80	0,00	3,41
1714	22.120	0,23	0,84	0,01	3,71
6941	37.632	0,25	0,85	0,00	3,36
1642	26.871	0,25	0,84	0,01	3,41
3044	12.000	0,37	1,06	0,01	3,74
4505	26.071	0,38	1,02	0,00	3,36
4336	28.224	0,56	1,74	0,00	9,62
Durch-Schnitt	**27.620**	**0,30**	**1,02**	**0,00**	**4,66**
Standard-abweichung	**8.330**	**0,12**	**0,30**	**0,00**	**2,30**

Tabelle 9: Statistische Eigenschaften der Strommessdaten häuslicher Fahrzeugladestationen pro Haushalt der Verbrauchergruppe C.

Haushalt	Anzahl	Durch-Schnitt	Standard-Abweichung	75%-Quartil	Maximum
9932	8.352	0,05	0,42	0,01	3,84
6072	35.232	0,07	0,46	0,00	3,32
3192	33.290	0,10	0,56	0,00	3,35
6836	11.040	0,15	0,74	0,00	6,77
3482	23.060	0,16	0,69	0,00	3,29
2470	37.632	0,17	0,72	0,00	3,37
114	18.910	0,17	0,72	0,00	3,38
3723	37.632	0,19	0,67	0,00	3,39
1185	37.632	0,19	0,73	0,00	3,33
661	25.536	0,20	0,77	0,00	3,44
7982	37.632	0,20	0,82	0,01	3,85

Haushalt	Anzahl	Durch-Schnitt	Standard-Abweichung	75%-Quartil	Maximum
5357	37632	0,21	0,85	0,00	3,78
26	22.008	0,21	0,76	0,01	3,53
9729	37.632	0,22	0,89	0,00	4,08
4526	37.632	0,23	0,82	0,00	3,39
2335	37.632	0,23	0,83	0,00	3,39
4352	31.621	0,23	0,83	0,00	3,39
3967	37.316	0,23	0,79	0,00	3,35
7940	32.840	0,24	0,91	0,00	3,78
4135	35.809	0,25	0,87	0,00	3,40
5403	37.632	0,27	0,89	0,00	3,41
370	37.632	0,28	0,92	0,01	3,68
8645	33.916	0,29	0,89	0,01	3,37
6990	37.632	0,31	1,01	0,00	3,80
6139	37.612	0,31	0,94	0,01	3,41
4998	24.212	0,32	0,97	0,00	3,42
6101	37.632	0,35	0,98	0,00	3,39
3036	24.580	0,35	1,07	0,00	3,80
6871	5.658	0,37	0,98	0,00	3,65
8669	37.632	0,39	1,05	0,00	3,38
6121	17.202	0,44	1,19	0,00	7,20
4373	37.461	0,44	1,02	0,00	3,38
4641	37.632	0,53	1,17	0,00	3,44
Durch-Schnitt	**30.984**	**0,25**	**0,84**	**0,00**	**3,71**
Standard-abweichung	**9.602**	**0,10**	**0,17**	**0,00**	**0,86**

Tabelle 10: Statistische Eigenschaften der Strommessdaten häuslicher Fahrzeugladestationen pro Haushalt der Verbrauchergruppe DP.

Haushalt	Verteilung	Bemerkung
6836		Die Messdaten sind nur bis zu Werten kleiner als 4 durchgehend frequentiert vertreten. Es treten allerdings vereinzelt Messungen zwischen den Intervallen [3; 4] und [6,1;] auf. Es ist daher nicht von Ausreißern auszugehen.
6121		Die Messdaten sind nur bis zu Werten kleiner als 4 durchgehend frequentiert vertreten. Es scheint sich bei den Messungen im Intervall [6,5; 7,1] um Ausreißer zu handeln. Diese Messungen werden auf den Wert 4 gesetzt.
Haushalt	**Verteilung**	**Bemerkung**
4336		Die Messungen häufen sich in dem Intervall [5; 6,1]. Andere Messungen verteilen sich gleichmäßig um dieses Cluster. Es ist deshalb nicht von Ausreißern auszugehen.

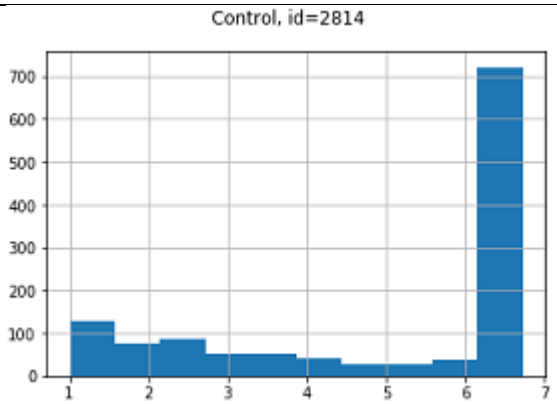
2814		Die Messdaten häufen sich in dem Intervall [6; 6,8] an und sind auch in vorangehenden Intervallen frequentiert vertreten. Es scheint sich deshalb nicht um einen Ausreißer zu handeln.

Tabelle 11: Verteilungen verdächtiger Haushalte in Bezug auf Extremwerte

Literaturverzeichnis

1. *Assessment of Impact of Charging Infrastructure for Electric Vehicles On Distribution Networks.* **Villafafila-Robles, R., et al.** o.O : 15th European Conference on Power Electronics and Applications (EPE), 2013, September. 10.1109/EPE.2013.6634671.

2. *Impact of Electric Vehicles on Power Distribution Networks.* **Putrus, G. A., et al.** o.O. : IEEE Vehicle Power and Propulsion Conference, 2009, Oktober. 10.1109/VPPC.2009.5289760.

3. **MacDonald, J.** Industry Research. *https://about.bnef.com/.* [Online] Bloomberg L.P. [Cited: 9 März 2017.] https://about.bnef.com/blog/electric-vehicles-to-be-35-of-global-new-car-sales-by-2040/.

4. **Todd, J., Chen, J. and Clogston, F.** *Analysis of the Electric Vehicle Industry.* Washington, DC 20005 : International Economic Development Council, 2013.

5. **Ruder, Adam.** *Electricity Pricing Strategies to Reduce Grid Impacts from Eletric Vehicle Charging in New York State.* New York : New York State Energy Research and Development Authority, 2015.

6. **Salameh, Ziyad.** The Impact of Electric Vehicles on Utilities. *International Journal of Power and Renewable Energy Systems (IJPRES).* 2015, Vol. 2, 2.

7. **Umweltbundesamt.** Daten zum Verkehr. Ausgabe 2012. [Online] 2012. [Cited: 25 03 2017.] https://www.umweltbundesamt.de/sites/default/files/medien/publikation/long/4364.pdf.

8. **Black Hills Electric Cooperative.** Black Hills Electric Cooperative is an equal opportunity provider, employer, and lender. *Black Hills Electric Cooperative.* [Online] [Cited: 28 3 2017.] http://www.bhec.com/content/peak-demand-times.

9. **Landsberg, Dennis R. and Stewart, Ronald.** *Improving Energy Efficiency in Buildings.* State University of new York : State University of new York Press, Albany, 1980. 0-87395-400-9.

10. **von Dollen, Don.** *Report to NIST on the Smart Grid Interoperability.* 2009.

11. **Borenstein, Severin, Jaske, Michael and Rosenfeld, Arthur.** *Dynamic Pricing, Advanced Metering, and Demand Response in Electricity Markets.* Kalifornien : s.n., 2002.

12. **Sweeney, James.** The California Electricity Crisis: Lessons for the Future. *The Bride.* 2002, Vol. 32, 2.

13. **Faruqui, Ahmad and Sergici, Sanem.** *Household Response To Dynamic Pricing Of Electricity - A Survey Of The Experimential Evidence.* San Francisco : s.n., 2009.

14. **Popeangă, Janina.** Data Mining Smart Energy Time Series. *Database System Journal.* 2015, Vol. VI, 1.

15. **Figueiredo, V., Rodrigues, F. and Gouveia, J. G.** An Electric Energy Consumer Characterization Framework Based On Data Mining Techniques. *IEEE Transactions on Power Systems.* 2005, Vol. 20, 2.

16. *The Application Of A Data Mining Framework To Energy Usage Profiling In Domestic Residences Using UK Data.* **Dent, Ian, Aickelin, Uwe and Rodden, Tom.** Bath, UK : s.n., 2011.

17. **Momtazpour, Marjan, et al.** *Coordinated Clustering Algorithms to Support Charging Infrastructure Design for Electric Vehicles.* Department of Computer Science, Virginia Tech, VA, 24060, USA : NEC Laboratories America, Inc., CA, 95014, USA, 2012.

18. **Meshram, Rupali A., Deorankar, A. V. and Chatur, P. N.** *Predicting Electricity Load Pattern Of Customers Using Clustering Technique.* Amravati, Maharashtra : Department Of Computer Science and Engineering, Government, 2013.

19. *Analysis of Similarity Measures in Times Series Clustering for the Discovery of Building Energy Patterns.* **Iglesias, Felix and Kastner, Wolfgang.** Vienna University of

Technology, Treitlstr. 1-3/ 4. Floor, Vienna A-1040, Austria : s.n., 2013. doi:10.3390/en6020579.

20. **Robinson, A. P., et al.** *Use of cluster analysis to identify electric vehicle driver recharging profiles in a region with a high density public recharging infrastructure.* Barcelona, Spain : s.n., 2013.

21. **Haben, Stephen, Singleton, Colin and Grindrod, Peter.** *Analysis and Clustering of Residential Customers Energy Behavioral Demand Using Smart Meter Data.* [PDF-Dokument] Mathematical Institute, University of Oxford, Oxford, OX2 6GG, UK : Oxford, 2015.

22. **Aswin, Raj. C., et al.** Smart Meter Based on Real Time Pricing. *Procedia Technology.* 2015, Vol. 1, 21.

23. **Abreau, Joana M., Pereira, Francisco C. and Ferrão, Paulo.** Using pattern recognition to identify habitual behavior in residential electricity consumption. *Energy and Buildings.* 2012, Vol. 49.

24. **Hernández, Luis, et al.** Classification and Clustering of Electricity Demand Patterns in Industrial Parks. *Energies.* 2012, 5.

25. **Xydas, Erotokritos, et al.** A data-driven approach for characterising the charging demand of electric vehicles: A UK case study. *Applied Energy.* 2016, 162.

26. **Liao, T. Warren.** Clustering of time series data - a survey. *Pattern Recognition.* 2005, 38.

27. **Han, J. and Kamber, M.** *Data Mining: Concepts and Techniques.* San Fransisco : Morgan Kaufmann, 2001. 978-1-55860-901-3.

28. **Wang, Wei, Yang, Jiong and Muntz, Richard.** *STING: A Statistical Information Grid Approach to Spatial Data Minging.* [PDF Dokument] Los Angeles : University of California, 1997.

29. **MacQueen, J.** Some Methods for Classification and Analysis of Multivariate Observations. [book auth.] J. Neyman L.M. LeCam. *Proceedings of the Fifth Berkeley Symposium on.* Berkley, USA : s.n., 1967, Vol. 1.

30. **Ester, Martin and Sander, Jörg.** *Knowledge Discovery in Databases: Techniken und Anwendungen.* Hamburg/Berlin : Springer, 2000. 3-540-67328-8.

31. **Rousseeuw, Peter J.** Silhouettes: A Graphical Aid To The Interpretation And Validation Of Cluster Analysis. *Journal of Computational and Applied Mathematics.* 1987, 20.

32. **Shearer, Collin.** The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing.* 2000, Vol. 5, 4.

33. **Chapman, Pete, et al.** *CRISP-DM 1.0 - Step by Step Data Mining Guide.* [PDF Dokument] USA : s.n., 1999.

Versicherung über Selbstständigkeit

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbstständig verfasst und nur die angegebenen Hilfsmittel benutzt habe.

Hamburg, den _____ _____