Bachelor's Thesis

Verallgemeinerung der Diskretisierung von Jahresdauerlinien für mehrere Energieträger

Generalizing the discretization of load duration curves
for multiple energy carriers

verfasst von

Timea Harmat

Mart.Nr. 03628380

eingereicht am

Lehrstuhl für Energiewirtschaft und Anwendungstechnik

Technische Universität München,

bei

Prof. Dr. rer. nat. Thomas Hamacher

Betreuer: Dipl-.Ing. Johannes Dorfner

Kurzfassung

Die vorliegende Arbeit beschäftigt sich mit der Diskretisierung von mehreren zeitgleichen betrachteten Lastprofilen unterschiedlicher Energieträger. Die Reduktion der Auflösung des Lastprofils führt zur Komplexitätsreduktion bei weiterführenden energetischen Analysen. Somit können zeitnahen und kostengünstigen Analysen entstehen. Gegenstand der Arbeit ist ein verallgemeinerter Prozess, der anhand von zeitgleicher Betrachtung mehrere Eigenschaften einer Datenreihe eine Diskretisierung der Datenreihe erzielt. Dafür werden zunächst allgemeine Darstellungs- und Diskretisierungsmöglichkeiten vermittelt, um anschließend eine zielführende Auswahl zu treffen. Die Ergebnisse von mehreren Methoden der ausgewählten Diskretisierungsmöglichkeit *Cluster-Analyse* und der damit verbundenen Darstellung *Streudiagramm* werden abschließend im Hinblick auf die Verwendung für Energiesystemmodelle verglichen und ausgewertet. Der Vergleich zeigt, dass die Wahl der Methode abhängig ist von der Optimierungsart des Energiesystemmodells.

Abstract

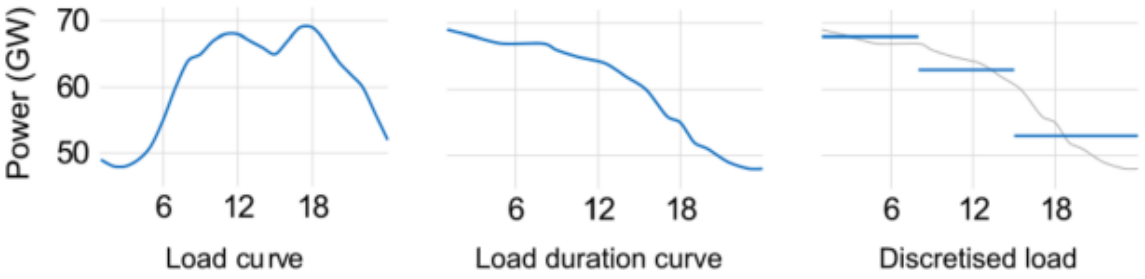
The following thesis deals with the discretization of multiple simultaneous observed load profiles of differential energy carriers. The reduction of the resolution of a load profile leads too a reduction of complexity in continuative energy analysis. Thus realtime and cost-efficient analysis can be developed. The subject of the thesis is a generalized process, which discretizes a stream of data based on a simultaneous consideration of multiple qualities of a stream of data. For this purpose general options of display and discretization will be imparted for the start. A constructiv selection from the options will be taken afterwards. The results of several methodes from the selected option of discretisation c*luster analysis* and the associated option of display *scatterplot* will be compared and evaluated with regard to usage for energy system models. The comparison shows that the selection of a methode depends on the optimization type of an energy sytem model.

Lehrstuhl für
Erneuerbare und Regenerative Energiesysteme

Themenstellung Bachelorarbeit

Verallgemeinerung der Diskretisierung von Jahresdauerlinien für mehrere Energieträger

Zur Charakterisierung des zeitlichen Verslaufs von Energieverbräuchen werden Jahresdauerlinien verwendet. Sie erlauben es, die Verteilung der Energienachfrage nach Strom, Wärme oder Kälte über das Jahr graphisch zu erfassen und zu vergleichen. Sie werden daher gerne als Eingangsgröße für Energiemodelle verwendet. Um die Datenmenge zu reduzieren, kann die Dauerlinie in wenige charakteristische Werte (Spitzenlast, Grundlast, saisonale Last) aggregiert.



Diskretisierte Lastkurve für einen Energieträger

Im Zuge dieser Arbeit sollen zunächst bestehende Methoden zur Diskretisierung von Jahresdauerlinien implementiert werden. Eine anschließende Literaturrecherche nach Wegen zur Erstellung, Darstellung und Diskretisierung von Jahresdauerlinien mehrerer Energieträger vergleicht deren Stärken und Schwächen. Abschließend werden mindestens zwei Methoden an einem bereitgestellten Datensatz mit simultanen Nachfragezeitreihen von Strom und Wärme angewandt und das Ergebnis interpretiert und verglichen im Hinblick auf die Verwendung für Energiesystemmodelle.

Voraussetzungen

- Interesse an Energiewirtschaft und mathematischen Methoden
- Kenntnisse in mind. einem von Excel VBA, MATLAB, GNU Octave, Python (NumPy)…

Ansprechpartner

Dipl.-Ing. Johannes Dorfner
Lehrstuhl für Erneuerbare und Nachhaltige Energiesysteme (Prof. Dr. rer. nat. T. Hamacher)
Theresienstraße 90, 80333 München, Gebäude N8, Raum N2825
Telefon +49 (0) 89 289-23948, E-Mail johannes.dorfner@tum.de

Erklärung

Hiermit erkläre ich,
Name: Harmat
Vorname: Timea
Mart. Nr.: 03628380

dass ich die beiliegende Bachelor's Thesis zum **Thema**:
Verallgemeinerung der Diskretisierung von Jahresdauerlinien für mehrere Energieträger

selbstständig verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, sowie alle wörtlichen und sinngemäß übernommenen Stellen in der Arbeit gekennzeichnet und die entsprechenden Quelle angegeben habe.
Vom Lehrstuhl und von seinen Mitarbeitern zur Verfügung gestellten Hilfsmittel, wie Modelle oder Programme, sind ebenfalls angegeben. Diese Hilfsmittel sind Eigentum des Lehrstuhls bzw. des jeweiligen Mitarbeiters. Ich werde sie nicht über die vorliegende Arbeit hinaus weiter verwenden oder an Dritte weitergeben.

Einer weiteren Nutzung dieser Arbeit und deren Ergebnisse (auch Programme und Methoden) zu Zwecken der Forschung und Lehre, stimme ich zu.

Ich habe diese Arbeit noch nicht zum erwerben eines anderen Leistungsnachweises eingereicht.

München, den 23.02.2015

_____

(Bearbeiter: Name, Vorname)

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

1.1 Motivation

…der vorliegenden Arbeit ist es, komplexe Analysen in energetischen Systemen basierend auf gemessenen oder simulierten Daten kostengünstig, zeitnah, … und in manchen Fällen überhaupt durchführen zu können.

Energie spielt eine große Rolle im alltäglichen Leben. Durch ihren Einsatz sind viele Techniken und Komforts im Leben möglich geworden. Die Erhaltung des Fortschritts durch die Energie ist nur möglich durch eine Sicherstellung von Energiesystemen. Die Sicherstellung beinhaltet mehrere Aspekte. Einer der größeren Aspekte ist eine kostengünstige und effiziente Infrastruktur von Energiesystemen. Um diese so kostengünstig und so effizient wie möglich zu errichten, werden Optimierungsprozesse von Energieinfrastrukturen unternommen. Für die Optimierungsprozesse werden Energiesystemmodelle entwickelt [EDE00]. Die Weiterentwicklung oder Optimierung der Energiesystemmodelle, und damit am Ende auch der Energiesysteme selbst wird unter anderem anhand von Lastprofilen der zu analysierenden energetischen Objekte durchgeführt. In einer allgemeinen Modellierung der Energiesysteme können verschiedene Energieträger (Elektrizität, Gas, Öl usw.) betrachtet werden. Die Anzahl und auch die Auflösung der Lastprofile beeinflussen SEHR deutlich die Komplexität des Energiesystemmodells selbst wie auch die Komplexität der Analyse [KUH12, S. 63].

Diese Arbeit befasst sich bei den verschiedenen Analyse-/Optimierungsmöglichkeiten „nur“ mit den Potentialen, die sich in der Auflösung des Lastprofils befinden.

Bei einer stündlichen Messung innerhalb eines Jahres entstehen 8 760 Messwerte, bei einer sekündlichen Messung über dreißigmillionen Messwerte pro Objekt und pro Energieträger. Für viele weitere energetische Analysen, die für Jahrzehnte geplant sind, ist diese Auflösung des Lastprofils meistens zu detailliert. Es entsteht eine Scheingenauigkeit der Lastprofile.

Um eine Komplexitätsreduktion zu erlangen und damit eine schnellere und kostengünstigere Analyse zu ermöglichen, müssen Lastprofile mit geringerer Auflösung entstehen. Die geringe Auflösung kann auf verschiedene Wege erreicht werden, wie z. B.:

- über eine geringere Messfrequenz, z. B. eine monatliche Messung
- über eine Diskretisierung

Bei einer geringeren Messfrequenz ist die Wahrscheinlichkeit sehr hoch, dass wichtige charakteristische Werte wie der Peak, die Durchschnittslast usw. eines Lastprofils falsch erfasst werden. Im Gegensatz zur „geringen Messfrequenz“ werden bei der Diskretisierung die Werte des Lastprofils weiterhin stündlich/sekündlich erfasst. Bei der Diskretisierung geht es grundsätzlich um mathematische Methoden, die eine Datenreihe mit hoher Auflösung auf einen repräsentativen Wert reduzieren.

Das Verfahren der Diskretisierung wird in Zusammenhang mit Lastprofilen beispielhaft für die Modellierung von Ausbauplänen von Kraftwerken angewendet [KUH12, S. 61ff.]. Hierbei wird z. B. eine geordnete Jahresdauerlinie für einen Energieträger diskretisiert und als Eingangsdaten der Modellierung benutzt. Existieren mehrere Energieträger für eine Modellierung, so wird der Prozess „Diskretisierung von Jahresdauerlinien eines Energieträgers“ für jedes Lastprofil der unterschiedlichen Energieträger einzeln angewendet. Somit wird dieselbe Anzahl an Diskretisierungen durchgeführt wie an Lastprofilen betrachtet wird. Um diese Vorgehensweise zu verkürzen, sollen im Zuge dieser Arbeit für eine Betrachtung mehrerer Lastprofile die mehrfach stattfindenden Diskretisierungsprozesse in einen allgemeinen Diskretisierungsprozess zusammengefasst werden.

1.2 Zielsetzung

...der vorliegenden Arbeit ist die Diskretisierung von mehreren zeitgleich betrachteten Lastprofilen unterschiedlicher Energieträger eines energetischen Objektes in einem Diskretisierungsprozess.

Um dieses Ziel zu erreichen, werden als Grundlage Wege zur Erstellung und Darstellung von Lastprofilen für mehrere Energieträger untersucht. Als einen weiteren Teil der Grundlage werden allgemeine Möglichkeiten zur Diskretisierung von Datenreihen betrachtet. Dabei werden sowohl konventionelle Diskretisierungen sowie Cluster-Analysen als Verfahren zur

Diskretisierung einer Datenreihe untersucht. Durch eine anschließende Auswahl anhand der Vorgehensweisen der Diskretisierungsmöglichkeiten wird der Diskretisierungsprozess für die Diskretisierung zeitgleich betrachteten Lastprofilen unterschiedlicher Energieträgern festgelegt. Mit Hilfe von Implementierungen in Programmen und Vergleichskriterien werden abschließend die Ergebnisse der ausgewählten Diskretisierung im Hinblick auf die Verwendung für Energiesystemmodelle ausgewertet.

1.3 Aufbau der Arbeit

Die vorliegende Arbeit untergliedert sich in acht Kapitel. Nachdem in der Einleitung die Motivation und die Zielsetzung der Arbeit erklärt wurden, wird in den kommenden Kapiteln die Ausführung der Arbeit zur Zielerreichung dargestellt.

In Kapitel 2 wird durch eine allgemeine Definition von Lastprofil und Datensatz der verwendete Datensatz für die Bachelorarbeit vorgestellt. Der Datensatz dient als eine Grundlage für die weiteren Schritte der Arbeit.

Anschließend werden in Kapitel 3 visuelle Darstellungsmöglichkeiten für Lastprofile gezeigt. Hierbei wird der Fokus vor allem auf die Darstellungsmöglichkeit der Lastprofile durch Diagramme gelegt.

In Kapitel 4 werden allgemeine Möglichkeiten zur Diskretisierung von Datenreihen beschrieben. Es erfolgt zusätzlich eine Klassifizierung der Möglichkeiten anhand ihrer Strategie in zwei Gruppen: konventionelle Diskretisierungsmethoden und Cluster-Analyse als Diskretisierungsmethode.

In Kapitel 5 wird eine Auswahl an Diskretisierungsmethoden anhand der Zielsetzung der Arbeit getroffen.

In Kapitel 6 erfolgt die Implementierung der ausgewählten Methoden in Programme. Es werden ebenfalls die Ergebnisse der Methoden für weitere Auswertungen festgehalten.

Die Auswertung der Ergebnisse der Methoden im Hinblick auf die Verwendung für Energiesystemmodelle wird in Kapitel 7 vorgenommen. Dafür werden Vergleichskriterien definiert, anhand derer die Auswertung erfolgt.

Abschließend wird in Kapitel 8 eine Zusammenfassung der gewonnenen und erarbeiteten Arbeitsinhalte gegeben.

2 Die Daten

Grundsätzlich können verschiedenen Daten, Datenabläufe und Funktionen diskretisiert werden. Innerhalb von dieser Arbeit werden, wie in der Einleitung beschreiben und festgelegt, die Diskretisierungsmöglichkeiten von energetischen Daten betrachtet bzw. analysiert. Die Daten liegen als Lastprofile in Datensätzen vor.

2.1 Lastprofil

Ein Lastprofil (bezeichnet auch als Lastkurve oder Lastgang) zeigt den zeitlichen Verlauf der benötigten Leistung eines energetischen Objektes über einen Zeitabschnitt an. Die benötigte Leistung eines energetischen Objektes kann z. B. elektrische Leistung oder Gasleistung usw. sein. Sie ist von zeitlichen Faktoren wie der Tageszeit, dem Wochentag oder der Jahreszeit abhängig. Auch andere Faktoren wie Standort, Gewohnheiten des Verbrauchers, Wetter etc. haben einen Einfluss auf den Verlauf des Lastprofils [RP00].

Je nach betrachtetem Zeitabschnitt kann das Lastprofil auch als Tages-, Wochen- oder Jahres-Lastprofil bezeichnet werden. Die Erfassung ihrer Messwerte kann unterschiedlich detailliert ausfallen. Sie können z. B. stündlich, minütlich, täglich etc. gemessen werden.

2.2 Die Lastprofildaten – Datensätze

„Ein Datensatz ist die Zusammenfassung von Daten, die in einer direkten Beziehung zueinander stehen oder gemeinsame Merkmale haben. Daten, die in einem Sinnzusammenhang stehen, können dabei in einem Ordnungssystem zusammengefasst sein. Ein solches Ordnungssystem besteht aus mehreren Datenfeldern, die jedes für sich die Daten für bestimmte Attribute enthält.“ [IW00]

Die Datenfelder eines energetischen Datensatzes können allgemein folgende Informationen beinhalten: die Verbrauchswerte (Strom, Spannung, Energie usw.) eines Verbrauchers, die bereitgestellte Energie eines Erzeugers, den zeitlichen Rahmen der durchgeführten Beobachtungen usw. Die Attribute eines energetischen Datensatzes sind dementsprechend die Oberbegriffe der sinngemäß gruppierten Datenfelder: Stromverbrauch, Energieverbrauch, Zeit usw.

Für Lastprofile werden vor allem Leistungswerte unterschiedlicher Energieträger abhängig von zeitlichen Werten in den Datenfeldern erfasst. Somit sind die zugehörigen Attribute eines Datensatzes für Lastprofile:

- Zeit
- benötigte Leistung des jeweiligen Energieträgers

2.3 Das verwendete Lastprofil

Die Lastprofile für die Bachelorarbeit sind als Datensätze vorhanden. Die Datensätze sind für alle TMY3 Orte (Typical Meteorological Yaers bzw. Testreferenzjahr) in den USA erhältlich [OEI13]. Die Werte der Lastprofile sind nicht durch Messungen entstanden. Die Werte wurden mit Hilfe von charakteristischen Bauarten und Energiekonsumwerten erstellt. Die Leistungswerte wurden stündlich für ein Jahr ausgearbeitet.

Die Datensätze sind als csv-Dateien abgespeichert. Die verwendete csv-Datei „USA_AK_Anchorage.Intl.AP.702730_TMY3_BASE.csv“ für die Arbeit beinhaltet die Lastprofildaten für ein Wohngebäude in Anchorage (siehe Anhang B). Sie ist folgendermaßen aufgebaut:

- Die erste Zeile enthält die Spaltennamen (Attribute des Datensatzes). Die Spalten sind nach ihrem Inhalt benannt.
- Spalten:
  - *1. „Date/Time“* – enthält die Zeitpunkte der Leistungswerte. Sie sind in der Form von *„01/01 01:00:00“* notiert.
  - *2. „Electricity:Facility [kW](Hourly)“* – enthält die Leistungswerte der abgenommenen elektrischen Leistung des gesamten Wohngebäudes.
  - *3. „Gas:Facility [kW](Hourly)“* – enthält die Leistungswerte der abgenommenen Gasleistung des gesamten Wohngebäudes.
  - *4. –15.* Die Spalten enthalten den Aufbau der elektrischen und Gasgesamtleistung des Wohngebäudes aufgeteilt auf einige typische Verbrauchsarten (Heizung, Kühlung, Licht usw.).

3 Visuelle Darstellungsmöglichkeiten

Die zu analysierenden energetischen Daten können grundsätzlich in vielen verschiedenen Formen dargestellt werden. Ganz allgemein betrachtet werden bei Darstellungsmöglichkeiten von Daten *non-visuelle* bis *visuelle* Darstellungen verwendet. Im Weiteren werden mögliche visuelle Darstellungsmöglichkeiten für Lastprofile betrachtet:

- tabellarische Darstellung
- verschiedene Diagramme

3.1 Tabellarische Darstellung

Die tabellarische Darstellung besteht grundsätzlich aus Zeilen und Spalten. In vielen technischen Bereichen werden die zu analysierenden Daten (Messwerte, simulierte Werte usw.) meistens in tabellarischer Form geliefert/bereitgestellt. Der Umfang einer tabellarischen Darstellung kann prinzipiell unterschiedlich ausfallen. Je größer und komplexer der Umfang einer Tabelle wird, desto unübersichtlicher ist ihre Darstellung für das menschliche Fassungsvermögen. In solchen Fällen sind andere visuelle Darstellungsmöglichkeiten von Vorteil.

Die Daten der Arbeit stehen in tabellarischer Form zu Verfügung. Eine detaillierte Beschreibung wird in Kapitel 2.3 durchgeführt.

3.2 Diagramme

Gewissermaßen sind Diagramme grafische Schnittstellen zwischen den Daten und uns Menschen. Um den Inhalt einer größeren Datenmenge besser veranschaulichen / verstehen zu können, werden Diagramme eingesetzt.

Für die Darstellung von unterschiedlichen Daten sind unterschiedliche Diagramme geeignet. Daher werden nur Diagramme betrachtet, die vorwiegend für Lastprofile verwendbar sind:

- Liniendiagramm
- Paralleldiagramm
- Streudiagramm

3.2.1 Liniendiagramm

Liniendiagramme (bezeichnet auch als Kurvendiagramm) dienen zur Darstellung des Verlaufes einer Datenreihe, die aus Wertepaaren besteht. Sie zeigt somit den funktionellen Zusammenhang der zwei Variablen des Wertepaares. Bei energetischen Lastprofilen werden die Lastwerte als eine Funktion abhängig von der Zeit festgehalten. Zwei mögliche Liniendiagramme für energetische Lastprofile werden als Ganglinie und Dauerlinie bezeichnet [WM89].

Bei der Ganglinie befinden sich die Messwerte in ihrer zeitlichen Reihenfolge der Messung. Die Wertepaare bestehen aus Zeit und Leistung. Auf der x-Achse kann der Zeitpunkt der Messungen abgelesen werden, auf der y-Achse die Werte der Messungen.

Bei der Dauerlinie befinden sich die Messwerte absteigend nach ihrer Größe geordnet. Auf der x-Achse kann die Zeitdauer eines Wertes abgelesen werden, auf der y-Achse die Werte der Messungen.

In der Abbildung 3.1 werden die Jahresganglinie und die Jahresdauerlinie für den Energieträger Gas des verwendeten Datensatzes dargestellt. Da die Verbrauchsart bei Gasleistung vorwiegend eine Wärmeleistung ist, wird die Beschriftung der y-Achse für die Gasleistung im Weiteren durch „P-Heat“ gekennzeichnet.

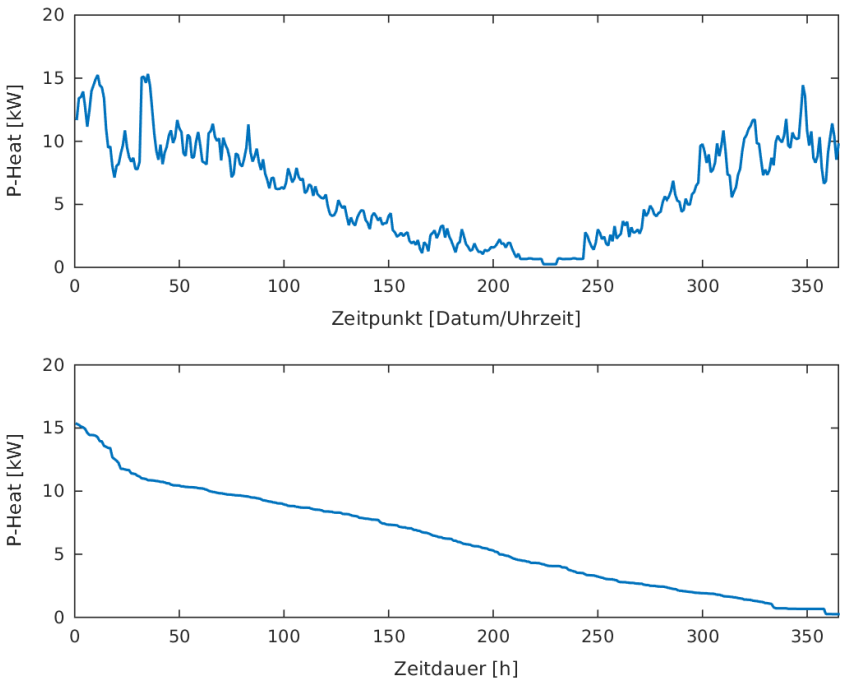


Abbildung 3.1: Ganglinie (oben) und Dauerlinie (unten)

3.2.2 Paralleldiagramm

Unter „Paralleldiagramm“ wird eine Erweiterung des Liniendiagramms verstanden. Während im Liniendiagramm die Darstellung einer Datenreihe von Wertepaaren erfolgt, werden im Paralleldiagramm die Wertepaare mehrerer Liniendiagramme gleichzeitig dargestellt. Die Wertepaare bilden weiterhin für sich einzelne, unabhängig darstellbare Liniendiagramme. Für die Erstellung eines Paralleldiagramms ist es wichtig, dass eine Variable des Wertepaares für jedes Liniendiagramm identisch ist. Bei Lastprofilen ist dies die Zeit. Diese Darstellung kann sowohl mit den Ganglinien als auch mit den Dauerlinien erfolgen.

In der Abbildung 3.2 werden beispielhaft zwei Ganglinien des verwendeten Datensatzes aufgezeigt. Die rote Jahresganglinie beschreibt die abgenommene Leistung des Energieträgers Gas, die blaue Jahresganglinie beschreibt die abgenommene Leistung des Energieträgers Elektrizität.

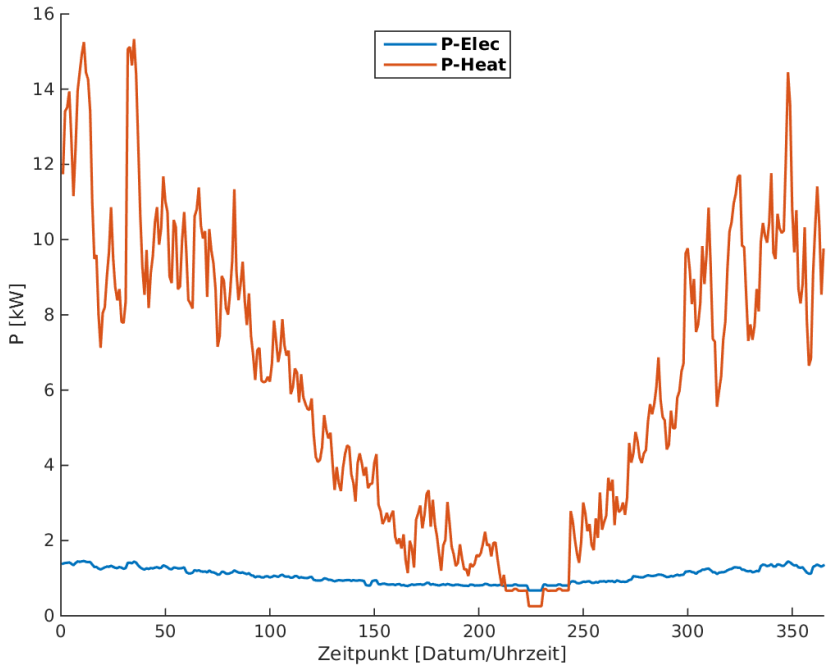


Abbildung 3.2: Paralleldiagramm

3.2.3 Streudiagramm

Streudiagramme (bezeichnet auch als Scatterplot) dienen zur Darstellung von funktionell zusammenhängenden oder auch nicht funktionell zusammenhängenden Datenreihen von

Wertepaaren. Im Gegensatz zu Liniendiagrammen können bei einem Streudiagramm Wertepaare mit einem oder beiden identischen Werten in der Datenreihe auftreten. Diese Darstellung kann bei den Lastprofilen jeweils mit den Ganglinien oder mit den Dauerlinien erfolgen.

Die Entstehung des Streudiagramms für den verwendetem Datensatz erfolgt durch die Parameterdarstellung der Wertepaare von den Lastprofilen der beiden Energieträger Elektrizität und Gas. Es ist wichtig, dass eine Variable des Wertepaares für beide Energieträger identisch ist. Bei Lastprofilen ist dies durch die Variable Zeit gegeben. Die Zeit *t* ist der Parameter der die Ganglinien oder Dauerlinien beider Energieträger durchläuft. Hierbei werden für beide Energieträger zum selben Zeitpunkt die Leistungswerte abgelesen. Sie ergeben die neuen x-Koordinaten und y-Koordinaten im Streudiagramm.

$$p(t)=\begin{pmatrix} x \\ y \end{pmatrix}=\begin{pmatrix} P_{gangelec}(t) \\ P_{ganggas}(t) \end{pmatrix} \tag{3.1}$$

Somit enthält ein Wertepaar im Streudiagramm die Werte der elektrischen Leistung und der Gasleistung. Es entsteht das in Abbildung 3.3 dargestellte Streudiagramm für den verwendeten Datensatz.

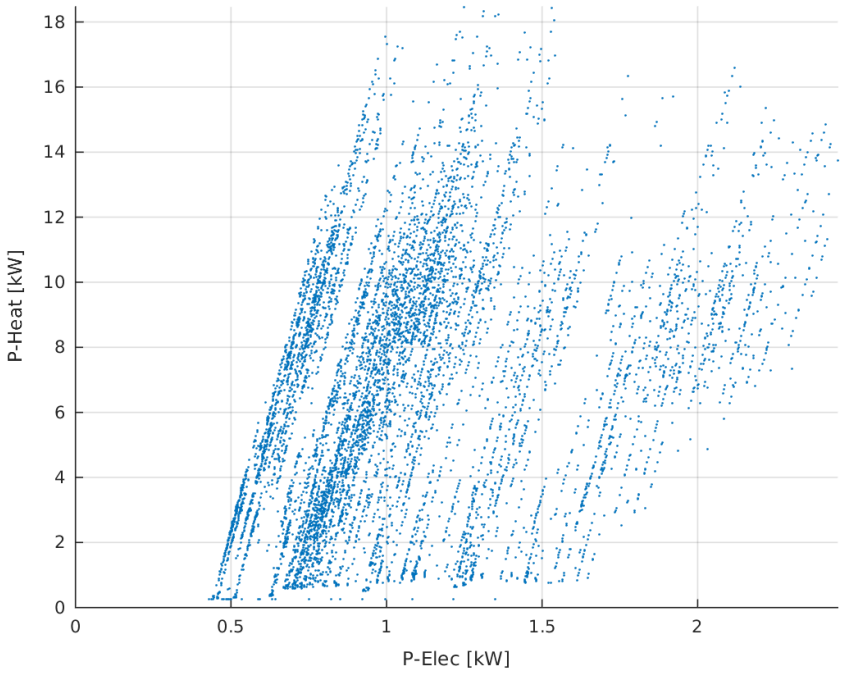


Abbildung 3.3: Streudiagramm mit normierten Achsen

Das Streudiagramm zeigt allgemein die möglichen energetischen Zustände eines Wohngebäudes in Anchorage. Der Parameter *Zeit* ist visuell im Streudiagramm nicht sichtbar, kann jedoch in einer externen Tabelle für eine eventuelle Weiterverwendung für jedes Wertepaar festgehalten werden. Die Koordinatenachsen des Streudiagramms beinhalten jeweils die Leistungswerte eines Energieträgers. Des Weiteren sind die Achsen der Abbildung normiert. In Abbildung 3.4 ist das selbe Streudiagramm wie in der Abbildung 3.3 dargestellt. Die Achsen sind in dieser Abbildung jedoch nicht normiert.

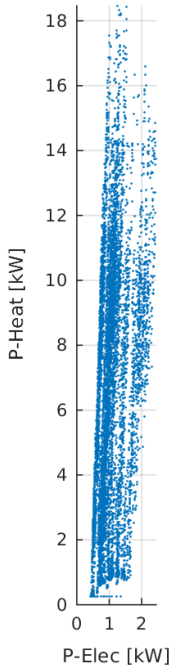


Abbildung 3.4: Streudiagramm mit nicht normierten Achsen

4 Diskretisierungsmöglichkeiten und deren Klassifizierung

Unter Diskretisierung wird die Abbildung einer größeren Datenmenge auf eine kleinere, repräsentative Teilmenge verstanden. Im extremen Fall kann das Argument, die Eingangsmenge für den Diskretisierungsprozess, unendlich sein, jedoch muss die Teilmenge immer endlich sein. Somit erfolgt durch die Diskretisierung einer Datenmenge eine Verringerung an Daten. Die Verringerung von Daten bedeutet auch einen Verlust an Informationen.

Die Diskretisierung wird bei Lastprofilen angewendet, um vor allem eine Datenreduktion zu erreichen. Für weiterführende energetische Analysen von Lastprofilen können z. B. charakteristische Repräsentanten wie Spitzen-, Grund- und Mittellast verwendet werden. Von der weiterführenden Analyse hängt es ab, welche Informationen nach einer Diskretisierung erhalten bleiben und wie detailliert die Informationserhaltung ausfallen soll.

Im Folgenden werden die Diskretisierungsmöglichkeiten in zwei Gruppen aufgeteilt und beschrieben:

- Konventionelle Diskretisierungsmethoden
- Cluster-Analyse als Diskretisierungsmethode

4.1 Konventionelle Diskretisierungsmethoden

Die konventionellen Diskretisierungsmethoden sind univariate Diskretisierungen. Die Diskretisierung einer Datenreihe wird durch die Bewertung einer Eigenschaft/Variable durchgeführt. Grundsätzlich existieren mehrere Diskretisierungsmethoden. Die Methoden können anhand ihrer unterschiedlichen Vorgehensweisen zur Diskretisierung einer Datenreihe in eine hierarchische Struktur eingeordnet werden. Eine mögliche Struktur mit mehreren Methoden zur Diskretisierung ist in der Abbildung 4.1 gegeben (ähnlich wie *Figure 2* aus [LHT+02]).

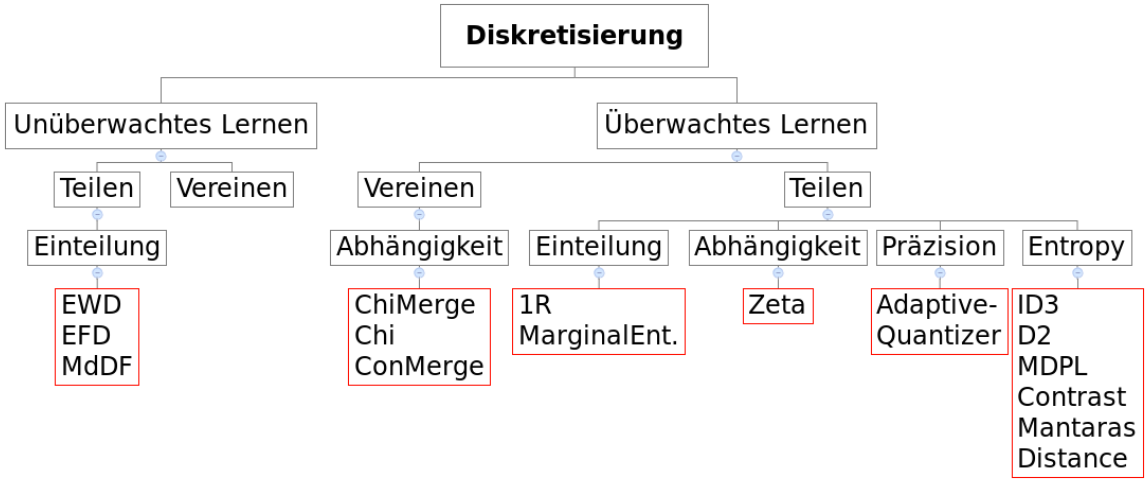


Abbildung 4.1: Kategorisierung der Diskretisierungsmethoden

Die Diskretisierungsmethoden sind in der Abbildung durch eine rote Umrandung gekennzeichnet. Die Methoden unterscheiden sich in ihren Vorgehensweisen. Ihre Kategorisierung erfolgt auf drei Ebenen. Jede Ebene steht für eine Kernvorgehensweise. Die Verzweigung jeder Ebene zeigt die grundlegenden Unterschiede in der Kernvorgehensweise der Methoden:

- überwachtes oder unüberwachtes Lernen
- Teilen oder Vereinen
- Diskretisierungstyp

Die diskrete Form für Diskretisierungsmethoden ist eine Stufenfunktion. Für die Entstehung der Stufenfunktion werden im Weiteren zwei Diskretisierungsmethoden erläutert:

- Equal Width Discretization
- Minimierung des Diskretisierungsfehlers

4.1.1 Equal Width Discretization (EWD)

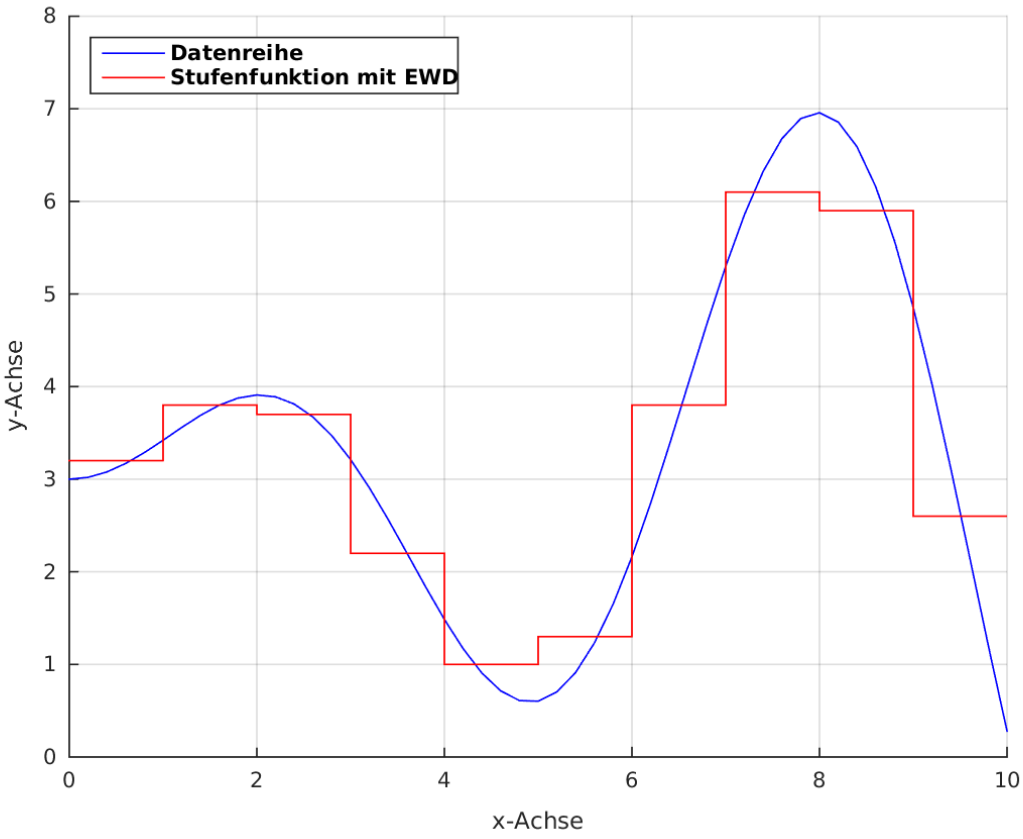


Abbildung 4.2: Diskretisierung mit EWD

EWD ist ein Repräsentant der Kategorie „unüberwachtes Lernen/Teilen/Einteilung“. EWD teilt eine Datenreihe in eine vom Benutzer definierte Anzahl k von gleich großen Intervallen ein. Der Algorithmus von EWD ist folgendermaßen aufgebaut: [LHT+02, S. 401]

1. Im ersten Schritt wird die Länge l der Intervalle mit folgender Formel berechnet:

$$l = \frac{x_{min} - x_{max}}{k} \tag{4.1}$$

wobei gilt:

x_{min}	kleinster Wert der Variable x
x_{max}	größter Wert der Variable x
k	vorgegebene Anzahl der Intervalle

2. Im zweiten Schritt werden die Trennungslinien *cp* mit Hilfe der berechneten Intervalllänge der Datenreihe festgelegt:

$$cp_1 = x_{min} + l, cp_2 = x_{min} + 2l, cp_k = x - x_{min} + (k-1)l \quad (4.2)$$

3. Im letzten Schritt werden die Repräsentanten der Intervalle berechnet. Dies könnte z.B. der Mittelwert des Intervalls sein.

Schließlich entsteht eine Stufenfunktion mit gleich langen Stufen. Sie ist in der Abbildung 4.2 mit der Farbe rot dargestellt. Die originale Datenreihe ist durch die Farbe blau abgebildet.

4.1.2 Minimierung des Diskretisierungsfehlers (MdDF)

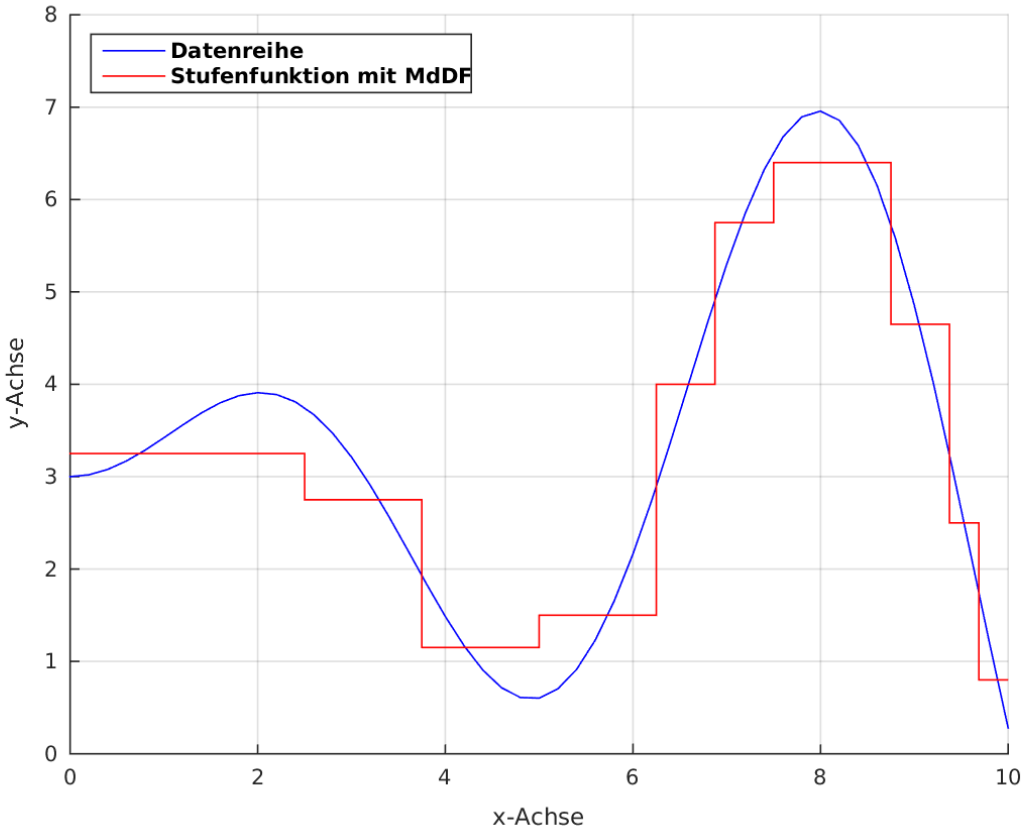


Abbildung 4.3: Diskretisierung mit MdDF

MdDF ist ein Repräsentant der Kategorie „unüberwachtes Lernen/Teilen/Einteilung“. MdDF teilt einen Datensatz in unterschiedlich große Intervalle ein. Die Intervalle werden generiert,

indem der Diskretisierungsfehler in einem Intervall minimiert werden soll. Der Algorithmus von MdDF ist folgendermaßen aufgebaut: [KUH12, S. 64ff.]

1. Im ersten Schritt wird die Datenreihe in zwei gleich große Intervalle unterteilt und der Mittelwert beider Intervalle ermittelt.
2. In Schritt zwei wird nach einem Intervall *best* gesucht, der den größten Diskretisierungsfehler aufweist. Der Diskretisierungsfehler wird mit folgender Formel berechnet:

$$best = max\left(\sum_{j=1}^{|Intervall|} \left|\left(D_j - SW_j\right)\right|\right) \tag{4.3}$$

   wobei gilt:

   D_j Wert des Datensatzes zum Zeitpunkt j

   SW_j Wert der Stufe zum Zeitpunkt j

3. Das gefundene Intervall *best* wird in zwei kleinere, gleich große Intervalle geteilt.
4. Die Schritte zwei und drei werden iterativ wiederholt, bis eine Abbruchbedingung erfüllt ist. Die Abbruchbedingung kann das Erreichen einer vorgegebenen Anzahl *k* von Intervallen sein oder das Erreichen eines minimalen Diskretisierungfehlers innerhalb der Intervalle.

Schließlich entsteht eine Stufenfunktion mit unterschiedlichen Stufenlängen. Sie ist in der Abbildung 4.3 mit der Farbe rot dargestellt. Die originale Datenreihe ist durch die Farbe blau abgebildet. Auffällig an der Abbildung ist, dass steilere Bereiche der Datenreihe durch mehrere, kürzere Intervalle nachgebildet werden als flachere Bereiche der Datenreihe. Somit wird der Verlauf der Datenreihe nach der Diskretisierung mit der Methode MdDF besser erhalten als mit der Methode EWD.

4.2 Cluster-Analyse als Diskretisierungsmethode

Die Cluster-Analyse ist eine Technik des Data-Minings. Ziel der Cluster-Analyse ist es, Ähnlichkeitsstrukturen und Zusammenhänge in einer Datenmenge zu finden. Die entstandenen Cluster sollen möglichst ähnliche Objekte aufweisen. Die Findung der Cluster kann mit unterschiedlichen Methoden durchgeführt werden. Anhand ihrer Vorgehensweise

können die Methoden in Kategorien eingeteilt werden: [CL14]

- hierarchische Clusterbildung
  - divisive Clusterbildung
  - agglomerative Clusterbildung
- partitionierende Clusterbildung
- dichtebasierte Clusterbildung

In der Abbildung 4.4 ist eine hierarchische Struktur der Methoden der Cluster-Analyse zu sehen. Die Methoden sind durch eine rote Umrandung gekennzeichnet.

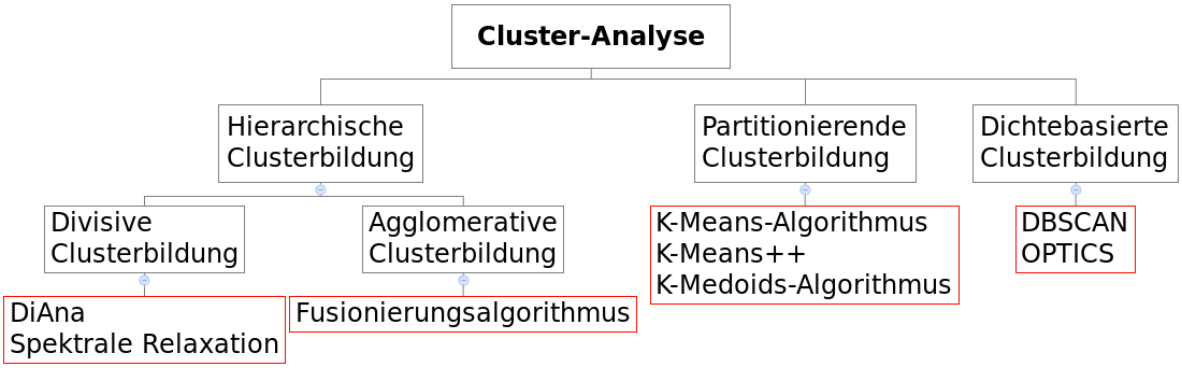


Abbildung 4.4: Kategorisierung der Methoden der Cluster-Analyse

Die Methoden der Cluster-Analyse sind keine kompletten Prozesse zur Diskretisierung einer Datenreihe. Die Cluster-Analyse als Diskretisierungsmethode unterteilt sich in zwei Schritte. Die Datenreihe wird durch die Cluster-Analyse noch nicht reduziert, sondern in Cluster/Gruppen eingeteilt. Dieser erste Schritt bildet die Grundlage für die Diskretisierung. Um alle Objekte einer Datenreihe auf eine Teilmenge abbilden zu können, werden nach der Durchführung der Clusterisierung die Repräsentanten der jeweiligen Cluster ausgerechnet. Die Repräsentanten bilden die Teilmenge der ursprünglich größeren Datenmenge. Mit der Berechnung der Repräsentanten wird die Diskretisierung einer Datenreihe erreicht.

Die Cluster-Analyse als eine Diskretisierung ist im Fall der gleichzeitigen Betrachtung von zwei Eigenschaften/Variablen einer Datenreihe eine bivariate Diskretisierung. Werden mehr als zwei Eigenschaften/Variablen einer Datenreihe gleichzeitig betrachtet bei der Cluster-

Analyse, dann handelt es sich um eine multivariate Diskretisierung.

Die verwendete diskrete Form für die Cluster-Analyse unterscheidet sich von der diskreten Form der Diskretisierungsmethode. Sie ist keine Stufenfunktion (siehe Abbildung 4.2), sondern eine punktuelle Darstellung (siehe Abbildung 4.5 [HL00]).

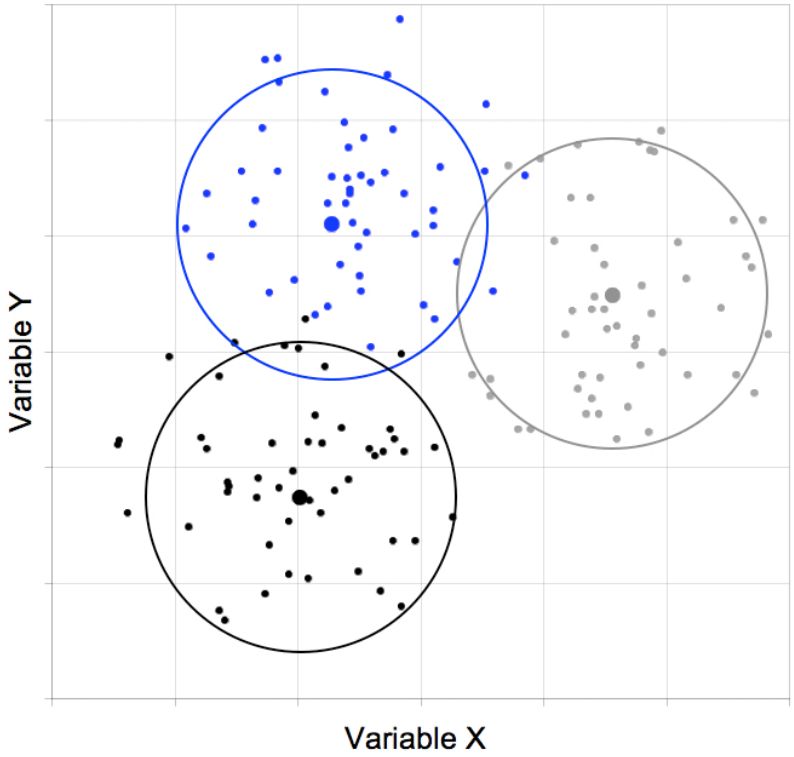


Abbildung 4.5: Clusterbildung

In der Abbildung 4.5 sind drei Cluster mit den Farben blau, schwarz und grau zu erkennen. Die Einteilung der Punkte/Objekte zu ihren jeweiligen Clustern sind durch die unterschiedlichen Farben gegeben. Der Mittelpunkt eines Kreises ist der diskrete Wert, durch den die Cluster repräsentiert werden. Der Repräsentant ist durch seine x-Koordinate und y-Koordinate definiert. Zusätzlich kann in einer dritten Variable *Gewichtung* die Anzahl der Objekte eines Cluster festgehalten werden.

Die Cluster-Analysen als Diskretisierungsmethoden unterteilen sich grundsätzlich in die zwei beschriebenen Bereiche Clusterbildung und Repräsentantenberechnung. Im Weiteren werden folgende Methoden der Clusterbildung erklärt:

- DiAna
- Fusionierungsalgorithmus
- K-Means-Algorithmus
- DBSCAN

Anschließend werden die zwei gängigsten Methoden zur Berechnung des Repräsentanten eines Clusters erklärt:

- Medoid
- Centroid

4.2.1 DiAna

DiAna ist ein Repräsentant der Kategorie „hierarchischen Clusterbildung/divisive Clusterbildung“. Die Kategorie ist das Komplement zur Kategorie „agglomerative Clusterbildung“. Der Zusammenhang zwischen den beiden Kategorien kann der Abbildung 4.6 entnommen werden [SAY15]. Die divisive Clusterbildung teilt eine Datenreihe in ihre einzelnen Objekte auf, die agglomerative Clusterbildung fügt die einzelne Objekte einer Datenmenge zusammen.

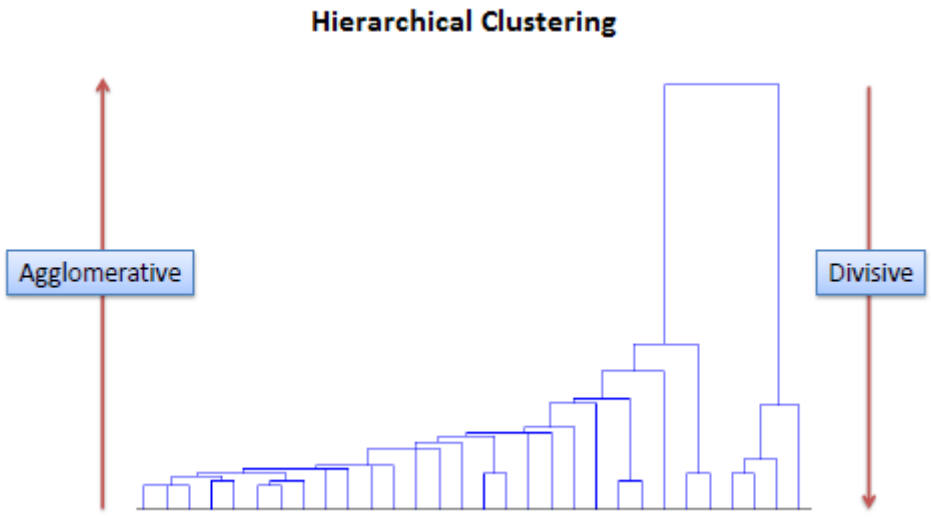


Abbildung 4.6: Hierarchische Clusterbildung

DiAna steht ausgeschrieben für „Divisive Analysis Clustering“. Zum Startzeitpunkt befinden sich alle *n* Objekte der Datenreihe in einem Cluster. Das Ziel des Algorithmus ist es, Cluster anhand der Unähnlichkeit ihrer Objekte aufzuteilen und damit neue Cluster zu erhalten. Die neu gewonnenen Cluster sollen nun Objekte mit größerer Ähnlichkeit enthalten als im vorherigen gemeinsamen Cluster. Die Aufspaltung der Datenreihe kann maximal bis zur Anzahl der Objekte durchgeführt werden. In diesem Fall ist in jedem Cluster nur ein Objekt vorhanden. Der Algorithmus von DiAna ist durch folgenden Aufbau beschrieben: [SHR00]

1. Im ersten Schritt wird der Cluster *A* mit dem größten Unähnlichkeitsgrad gesucht. Der Unähnlichkeitsgrad wird mit folgender Formel berechnet:

$$unähnlichkeit = \frac{1}{|A|-1} \cdot \sum_{i,j \in A,\, j \neq i} d(i,j) \tag{4.4}$$

   wobei gilt:

   d(i,j) Abstandsmaß zwischen den Objekten *i* und *j*

2. Anschließend erfolgt die Aufteilung des Clusters *A*. Hierfür wird als zweiter Schritt im Cluster *A* ein Objekt *k* gesucht. Dieses Objekt weist den größten Unähnlichkeitsgrad zu den anderen Objekten innerhalb des Clusters *A* auf. Das Objekt *k* bildet den Kern des neuen Clusters *B* und wird mit folgender Formel berechnet:

$$k = max\left(\frac{1}{|A|} \cdot \sum_{j \in A} d(i,j)\right) \tag{4.5}$$

   wobei gilt:

   i betrachteter Kandidat für Objekt *k*

3. Im dritten Schritt erfolgt die Aufteilung der restlichen Objekte von Cluster *A*. Hierfür wird iterativ aus Cluster *A* das Objekt *p* mit dem größten Ähnlichkeitsgrad zu Cluster *B* entfernt und Cluster *B* hinzugefügt. Objekt *p* wird mit folgender Formel ermittelt:

$$p = max(a-b) \tag{4.6}$$

   wobei gilt:

   a Unähnlichkeitsgrad eines Objektes *i* zu Cluster *A*

   b Unähnlichkeitsgrad eines Objektes *i* zu Cluster *B*

4. Der Abbruch von der Aufteilung des Clusters *A* und somit des dritten Schrittes erfolgt, sobald das ausgesuchte Objekt *p* eine größere Ähnlichkeit zu Cluster *A* aufweist als zu Cluster *B* oder Cluster *A* nur noch ein Objekt enthält.

   4.1. $p>0$ Objekt wird Cluster *B* hinzugefügt. Algorithmus wird bei Schritt 3 fortgesetzt.

   4.2. $p<0$ Objekt wird Cluster *B* nicht hinzugefügt. Algorithmus wird bei Schritt 1 fortgesetzt.

   4.3. $|A|=1$ Objekt wird Cluster *B* nicht hinzugefügt. Algorithmus wird bei Schritt 1 fortgesetzt.

5. DiAna wird iterativ wiederholt, bis eine Abbruchbedingung für den Algorithmus erfüllt ist. Die Abbruchbedingung kann das Erreichen einer vom Benutzer definierten Anzahl von Iterationen/Clustern oder der Unterschreitung eines vordefinierten Unähnlichkeitsgrades innerhalb der Cluster sein.

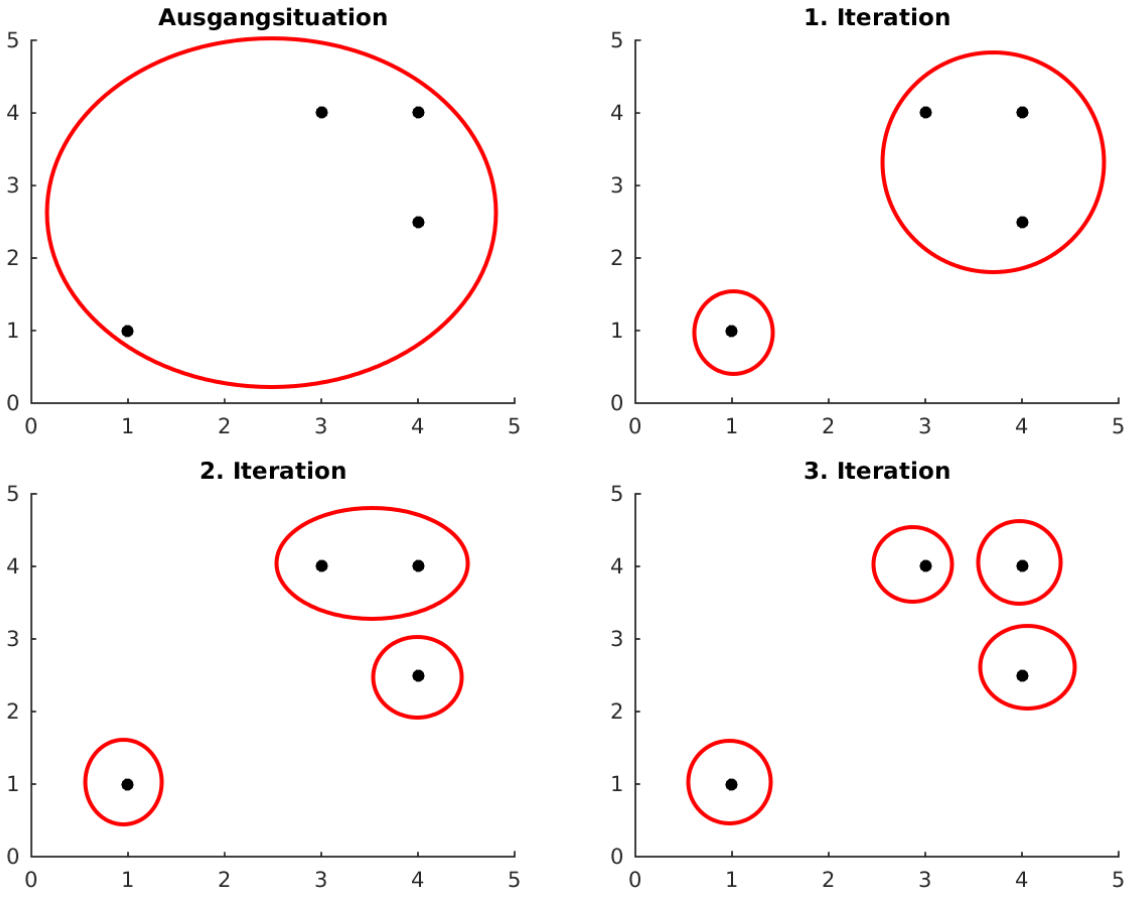


Abbildung 4.7: Clusterbildung mit DiAna

In der Abbildung 4.7 wird die Aufteilung einer Datenreihe anhand des Algorithmus von DiAna dargestellt. Die Objekte der Datenreihe sind durch die schwarzen Punkte veranschaulicht. Die gefundenen Cluster sind durch die roten Ellipsen dargestellt. Hierbei umkreist jede Ellipse die Objekte, die im Cluster enthalten sind. Jedes Bild ist das Zwischenergebnis einer Iteration. Die Abbruchbedingung für die beispielhafte Aufteilung der Datenreihe ist das Erreichen derselben Anzahl an Clustern wie an Objekten in der Datenreihe vorhanden sind.

4.2.2 Fusionierungsalgorithmus

Fusionsalgorithmus ist ein Repräsentant der Kategorie „hierarchscihe Clusterbildung/agglomerative Clusterbildung“. Agglomerative Clusterbildung wird auch als „Bottom-up-Clusterbildung“ bezeichnet. Die Methoden der agglomerativen Clusterbildung beginnen mit der entgegengesetzten Betrachtungsweise als Methoden der divisiven Clusterbildung (siehe Abbildung 4.6). Als Ausgang wird jedes Objekt der Datenreihe als ein einzelnes Cluster angenommen. Das Ziel ist es, ähnliche Cluster in ein gemeinsames übergreifendes Cluster zusammenzufassen. Sobald ein Cluster mehr als ein Objekt beinhaltet, kann der Ähnlichkeitsgrad zwischen den Clustern auf unterschiedliche Arten berechnet werden. Aufgrund dieser Tatsache gibt es mehrere, verschiedene Fusionierungsalgorithmen. Bei jedem Fusionierungsalgorithmus gilt, dass der Prozess der Clusterbildung maximal solange wiederholt werden kann, bis alle Objekte in einem Cluster zusammengefasst sind.

Bevor der Grundablauf des Algorithmus erklärt wird, werden die unterschiedlichen Berechnungsarten von Ähnlichkeitsgraden betrachtet. Grundsätzlich ist hier die Rede von einer Abstandsberechnung zwischen Cluster A und Cluster B: [CL14, S. 155 ff.]

- Single-Linkage ist der kleinste Abstand zwischen einem Paar von Objekten $a \in A$ und $b \in B$ der beiden Cluster.

$$d_{SL}(A,B) = min(d(a,b)) \tag{4.7}$$

- Complete-Linkage ist der größte Abstand zwischen einem Paar von Objekten $a \in A$ und $b \in B$ der beiden Cluster.

$$d_{CL}(A,B) = max(d(a,b)) \tag{4.8}$$

- Die Centroid-Methode ist der Abstand zwischen den zwei Schwerpunkten $s_a \in A$ und $s_b \in B$ von den Clustern.

$$d_{CM}(A,B) = d(s_a, s_b) \tag{4.9}$$

- Average-Linkage ist der durchschnittliche Abstand aller möglichen Paarbildungen der Objekte $a \in A$ und $b \in B$ zwischen den beiden Clustern.

$$d_{AL}(A,B) = \frac{1}{|A| \cdot |B|} \cdot \sum_{a \in A,\, b \in B} d(a,b) \tag{4.10}$$

- Average-Group-Linkage ist der durchschnittliche Abstand aller Objekte $a, b \in A \cup B$ zueinander, nach der Vereinigung der beiden Cluster zu einem Cluster.

$$d_{AGL}(A,B) = \frac{1}{(|A|+|B|) \cdot (|A|+|B|+1)} \cdot \sum_{a,\, b \in A \cup B} d(a,b) \tag{4.11}$$

Mit Hilfe der Berechnung der Ähnlichkeitsgrade zwischen den Clustern sieht der Fusionsalgorithmus wie folgt aus:

1. Im ersten Schritt wird zwischen jedem Cluster ein Ähnlichkeitsgrad festgestellt.
2. Anschließend werden die zwei Cluster ausgesucht, die den höchsten Ähnlichkeitsgrad aufweisen. Die gefundenen Cluster werden zu einem gemeinsamen Cluster zusammengefasst.
3. Die Schritte 1 und 2 werden iterativ wiederholt, bis eine Abbruchbedingung erfüllt ist. Die Abbruchbedingung kann das Erreichen einer Anzahl von Clustern/Iterationen oder die Unterschreitung einer Grenze des Ähnlichkeitsgrades sein.

In der Abbildung 4.8 wird die Clusterbildung einer Datenreihe anhand des Fusionierungsalgorithmus dargestellt. Die Objekte der Datenreihe sind durch die schwarzen Punkte veranschaulicht. Die gefundenen Cluster sind durch die roten Ellipsen gekennzeichnet. Hierbei umkreist jede Ellipse die Objekte, die in dem Cluster enthalten sind. Jedes Bild ist das

Zwischenergebnis einer Iteration. Die Cluster wurden mit der Centroid-Methode gebildet. Die Abbruchbedingung für die beispielhafte Aufteilung der Datenreihe ist das Erreichen der Anzahl eins an Clustern.

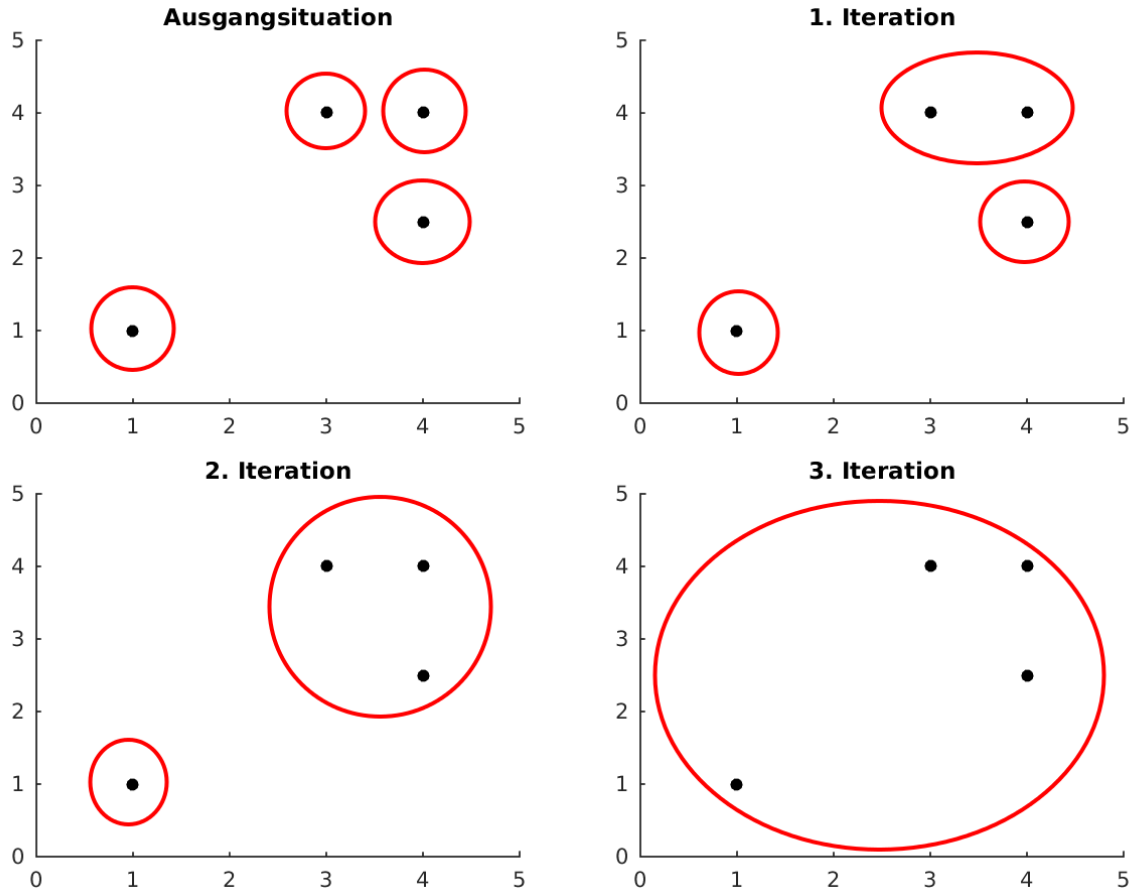


Abbildung 4.8: Clusterbildung mit Fusionsalgorithmus

4.2.3 K-Means-Algorithmus

Der K-Means-Algorithmus ist ein Repräsentant der Kategorie „partitionierende Clusterbildung“. Bei dieser Methode wird die Anzahl der gesuchten Cluster vom Benutzer vorgegeben. Die Objekte der Datenreihe befinden sich zum Startzeitpunkt keinem Cluster zugeteilt. Das Ziel des Algorithmus ist es, durch eine Neueinteilung aller Objekte die Position des Clusters iterativ zu verändern. Durch die Findung der neuen Clusterpositionen sollen immer ähnlichere Objekte innerhalb einer Zuteilung abgedeckt werden.

Der prinzipielle Aufbau des K-Means-Algorithmus besteht aus drei Schritten und einer Abbruchbedingung: [CL14, S.139ff.]

1. Im ersten Schritt werden zufällig *k* Objekte aus der Datenreihe bestimmt. Die Objekte repräsentieren vorläufig die *k* Zentren der Cluster, die Clusterzentren.

2. Im zweiten Schritt werden alle Objekte des Datensatzes den Clusterzentren zugeordnet. Die Zuordnung erfolgt anhand eines Ähnlichkeitsgrades zwischen den Objekten und den Clusterzentren. Jedes Objekt wird dem Clusterzentrum mit dem größten Ähnlichkeitsgrad zugeteilt.

3. Im dritten Schritt werden die neuen Positionen der Cluster berechnet. Sie ersetzen schließlich für die nächste Iteration die alten Positionen. Die neue Position des Clusters kann mit der Formel für Schwerpunktberechnung ermittelt werden:

$$poistion(C) = \frac{1}{|C|} \cdot \sum_{i \in C} i \tag{4.12}$$

   wobei gilt:

   C Cluster, dessen Position berechnet wird

4. Die Schritte zwei und drei des Algorithmus werden solange iterativ wiederholt, bis eine Abbruchbedingung erfüllt ist. Die Abbruchbedingung kann das Erreichen einer Anzahl von Iterationen oder einer Nicht-Neupositionierung der Clusterzentren sein. Letzteres bedeutet, dass die Clusterzentren sich nicht mehr im Raum bewegen.

In der Abbildung 4.9 wird die Clusterbildung einer Datenreihe anhand des K-Means-Algorithmus dargestellt. Die Objekte sind durch Punkte gekennzeichnet. Die Positionen der Clusterzentren sind durch die farbigen Quadrate gekennzeichnet. Mit der Annahme, dass nach zwei Clustern gesucht wird, sind die Quadrate rot und blau. Die Zuteilung der Objekte zu den Clustern ist ebenfalls durch die zwei Farben gegeben. Jedes Bild stellt das Zwischenergebnis einer Iteration dar. Als Abstandsmaß wurde für die Abbildung der euklidische Abstand genommen. Als Clusterzentrum wurde der Schwerpunkt eines Clusters berechnet. Die Abbruchbedingung für die beispielhafte Aufteilung ist die Unveränderung der Position von den Clusterzentren.

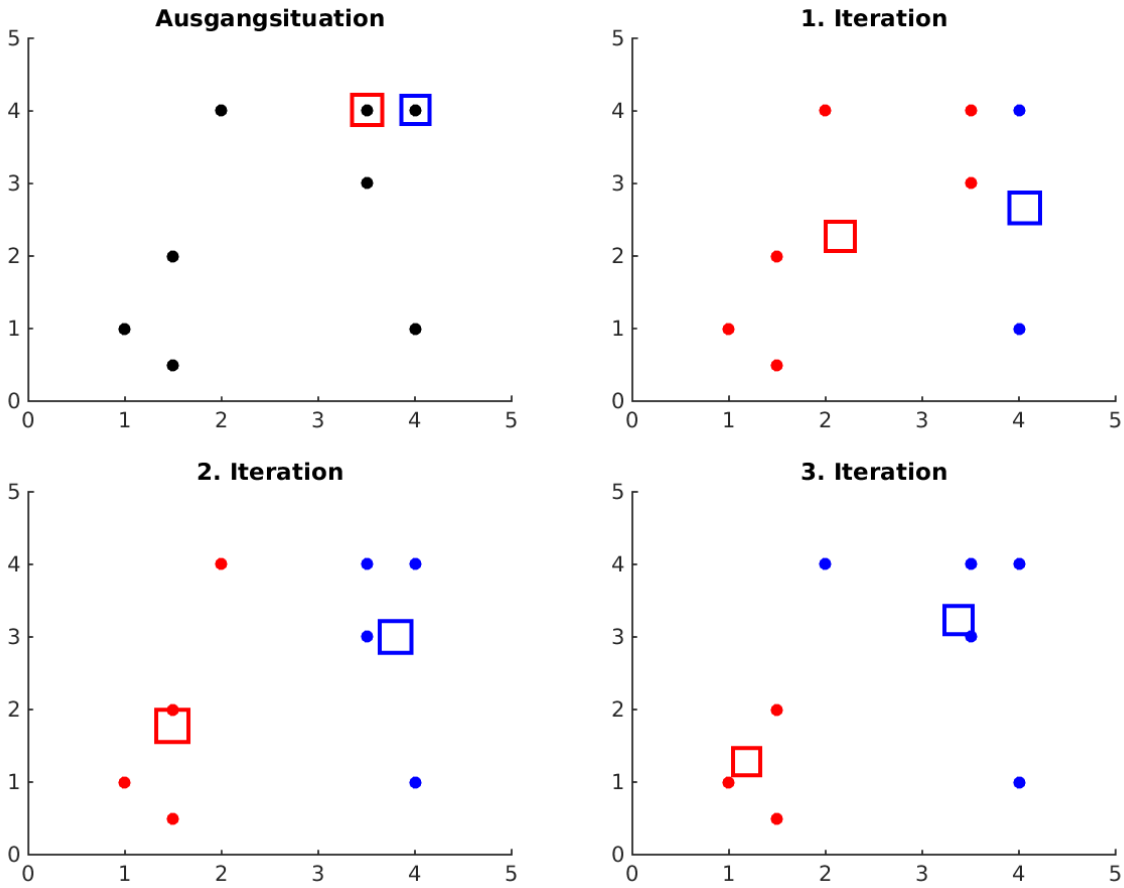


Abbildung 4.9: Clusterbildung mit K-Means-Algorithmus

Abschließend soll für dieses Verfahren notiert sein, dass es in den einzelnen Bereichen des Algorithmus mehrere Möglichkeiten der Umsetzung oder Berechnung von Aspekten gibt:

- In Schritt zwei und drei des Algorithmus können die Zuordnung eines Elementes und die damit verbundene Berechnung des Clusterzentrums einzeln gekoppelt erfolgen. Es wird immer ein Wert einem Zentrum zugeordnet und dieser *anschließend sofort* aktualisiert. Erst nach der Aktualisierung wird der nächste Wert einem Zentrum zugeordnet.
- Der Mittelpunkt eines Clusters wird mit Beachtung des alten Clusters berechnet. Dies geschieht, indem z. B. der Mittelpunkt zwischen dem neuen und dem alten Cluster kalkuliert wird. Dies führt zu einer langsameren Annäherung an das Endergebnis, jedoch auch zu einer präziseren.

4.2.4 DBSCAN-Algorithmus

Der DBSCAN-Algorithmus wird häufig bei der Erfassung von nicht konvexen Gruppenstrukturen einer Datenmenge eingesetzt. Das Ziel des DBSCAN-Algorithmus ist es, die Objekte anhand ihrer Dichteverbundenheit zueinander in Cluster einzuordnen. Die Regionen der einzelnen Cluster werden identifiziert, indem eine bestimmte minimale Dichte zwischen den Objekten bestehen muss. Durch diese Dichte werden die einzelnen Objekte verbunden und zu einem Cluster definiert. Die Trennung zwischen den einzelnen Clustern wird anhand der Tatsache gefunden, dass der Zwischenraum zwischen den Clustern eine geringere Dichte der Objekte aufweist als die vorgegebene minimale Dichte der Cluster [CL14, S. 160 ff.].

Ausgangspunkt des Algorithmus ist, dass jedes Objekt der Datenmenge als „nicht eingeteilt“ gekennzeichnet ist. Auch werden im Voraus zwei Parameter spezifiziert: der Radius $r > 0$ und die minimale Anzahl von Nachbarn *minNach*, die ein betrachtetes Objekt für eine gute Dichteverbundenheit haben sollen. Der DBSCAN kann somit drei Arten von Punkten aufweisen:

- **Kernobjekte** sind Punkte, die mindestens eine Anzahl von *minNach*-Nachbarn innerhalb des Radius *r* aufzeigen.
- **Randobjekte** sind Punkte, die als Nachbar eines Kernobjektes aufgefasst werden, aber selber nicht genug Nachbarn haben, um ein Kernobjekt zu sein. Diese Punkte bilden den Rand des Clusters.
- **Noise** sind Rauschpunkte, die von Anfang an nicht genug Nachbarn haben, um als ein Kernobjekt erfasst zu werden. Sie sind auch nicht als ein Nachbar eines Kernobjektes erfasst.

In den folgenden Schritten wird der Aufbau des DBSCAN-Algorithmus beschrieben.

1. Im ersten Schritt wird ein „nicht eingeteiltes“ Objekt *i* gesucht. Dieses Objekt wird als neuer Cluster *A* definiert und mit „eingeteilt“ gekennzeichnet.
2. Im Umkreis des Objektes *i* mit Radius *r* wird die Anzahl der Nachbarn *n* gezählt. Nun erfolgt anhand *n* eine Einordnung des Objektes *i* in eine der folgenden Kategorien:

2.1. $n<minNach$ Das Objekt *i* wird als Noise gekennzeichnet. Anschließend wird der Algorithmus bei Schritt 1 fortgesetzt.

2.2. $n \geq minNach$ Das Objekt *i* wird als Kernobjekt gekennzeichnet. Seine Nachbarn werden Cluster A hinzugefügt. Der Algorithmus wird bei Schritt 3 fortgesetzt.

3. Im dritten Schritt werden die Nachbarn der Kernobjekte betrachtet. In diesem Schritt wird entschieden, ob der Nachbar *k* eines Kernobjekts ebenfalls ein Kernobjekt oder ein Randobjekt ist. Aus dem Nachbarn *k* wird jetzt das Objekt *k*. Im Umkreis des Objekts *k* mit einem Radius *r* wird die Anzahl der Nachbarn *n* gezählt. Nun erfolgt anhand *n* eine Einordnung des Objektes *k* in eine Kategorie:

   3.1. $n<minNach$ Das Objekt *k* wird als Randobjekt gekennzeichnet. Seine Nachbarn werden Cluster *A* nicht hinzugefügt.

   3.2. $n \geq minNach$ Das Objekt *k* wird als Kernobjekt gekennzeichnet. Seine Nachbarn werden Cluster *A* hinzugefügt.

4. Schritt 3 wird solange wiederholt, bis keine neuen Nachbarn Cluster *A* hinzugefügt werden und alle gefundenen Nachbarn der Kernobjekte betrachtet worden sind.

5. Der Algorithmus wird iterativ solange wiederholt, bis alle Punkte „eingeteilt“ wurden.

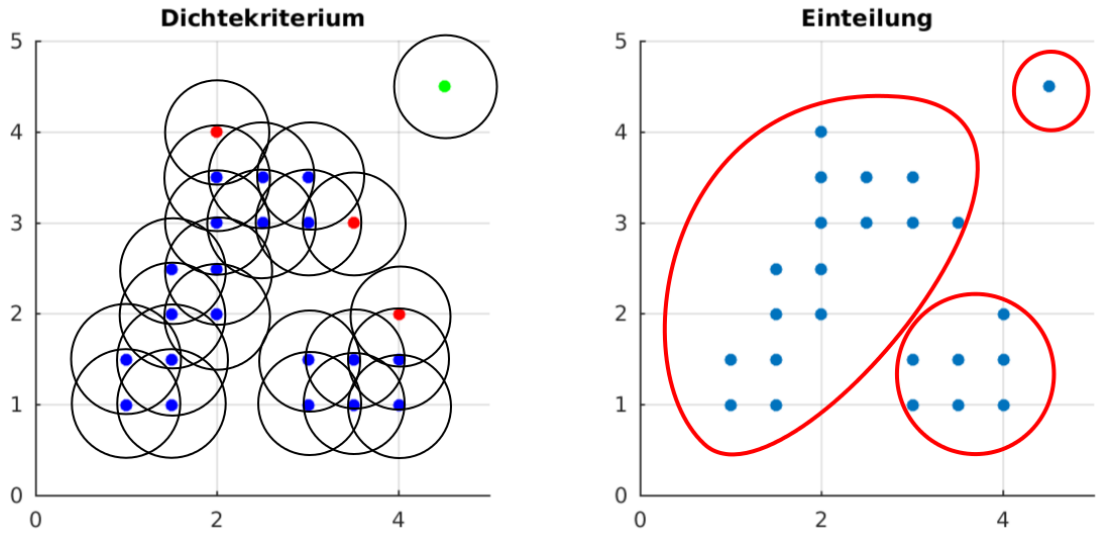


Abbildung 4.10: Clusterbildung mit DBSCAN-Algorithmus

In der Abbildung 4.10 wird die Clusterbildung einer Datenreihe anhand des DBSCAN-Algorithmus dargestellt. Da der Algorithmus keine iterative Neubildung der Cluster

durchführt, wird im linken Bild der Abbildung das Dichtekriterium gezeigt, anhand dessen der Algorithmus die Objekte einteilt. Der Radius *r* beträgt 0.6, und die Nachbaranzahl *minNach* beträgt 2. Die Objekte sind durch Punkte gekennzeichnet. Die Einteilung in eine der drei Kategorien ist durch die Farbgebung veranschaulicht. Die Kernobjekte sind durch blaue Punkte, die Randobjekte durch rote Punkte und Noise durch grüne Punkte dargestellt. Im rechten Bild sind die fertigen Cluster durch die roten Ellipsen gekenzeichnet. Hierbei umkreist jede Ellipse die Objekte, die in dem Cluster enthalten sind.

4.2.5 Repräsentantenberechnung

Die Berechnung der Clusterrepräsentanten ist ein weiterführender Schritt von der Clusterbildung zur Diskretisierung von Datenmengen. Die Repräsentanten können auf mehrere Arten bestimmt werden [CL14, S. 148].

Centroid:

Der Centroid eines Clusters ist der Mittelpunkt bzw. Schwerpunkt des Clusters. Er wird mit der Formel der Schwerpunktberechnung ermittelt (siehe Formel 4.12). Der Centroid muss nicht als ein Objekt der Datenmenge existieren, er kann fiktiv sein.

Medoid:

Der Medoid eines Clusters ist im Gegensatz zum Centroid immer ein Objekt der Datenmenge. Er wird mit Hilfe des Centroids für einen Cluster bestimmt. Der Medoid eines Clusters ist das Objekt, das dem Centroid des Clusters am ähnlichsten ist.

5 Auswahl von Diskretisierungsmöglichkeiten für Lastprofile

In Kapitel 4 wurden allgemein Möglichkeiten zur Diskretisierung von Datenreihen besprochen. Die Diskretisierungen sind grundsätzlich auf unterschiedliche Datentypen und Datenreihen für unterschiedliche Zwecke anwendbar. Jedoch ist nicht jede Diskretisierung für jede Art von Daten gleich gut geeignet.

Die Eingangsdaten der Arbeit sind energetische Lastprofile. Für den weiteren Verlauf der Arbeit wird eine Auswahl an Möglichkeiten zur Diskretisierung von Lastprofilen für mehrere Energieträger getroffen. Dafür werden die zwei Diskretisierungsmöglichkeiten, Diskretisierungsmethoden und Cluster-Analyse als Diskretisierungsmethode, im Hinblick auf die mögliche Anzahl der beachteten Eigenschaften/Variablen zur Diskretisierung einer Datenreihe betrachtet:

- univariate Diskretisierung
- multivariate Diskretisierung

5.1 Univariate Diskretisierung

Bei der univariaten Diskretisierung wird die Diskretisierung einer Datenreihe durch die Bewertung einer Eigenschaft/Variable durchgeführt. Bei den energetischen Lastprofilen ist die bewertete Variable der Datenreihe der Leistungswert *P(t)* eines Energieträgers. Somit müssen bei einer Betrachtung von mehreren Lastprofilen für unterschiedliche Energieträger die Diskretisierungen der Lastprofile unabhängig von einander durchgeführt werden. Obwohl die diskretisierten Lastprofile in einem Paralleldiagramm (siehe Kapitel 2) zusammen gefasst werden können, beschreiben die diskreten Werte weiterhin unabhängige Lastverläufe. Diese Unabhängigkeit wird auch in weiterführenden energetischen Analysen zu unabhängigen Teilanalysen führen. Als Resultat wird das komplette energetische System quasi in seine Einzelteile (basierend auf Energieträger) „zerlegt“.

Univariate Diskretisierungen sind auf Liniendiagramme und Paralleldiagramme besonders gut anwendbar. Die univariate Diskretisierung von mehreren Lastprofilen beinhaltet dieselbe Anzahl an Diskretisierungen wie an Lastprofilen betrachtet wird. Univariate Diskretisierungen sind die konventionellen Diskretisierungsmethoden.

5.2 Multivariate Diskretisierung

Bei multivariaten Diskretisierungen wird die Diskretisierung einer Datenreihe durch die gleichzeitige Bewertung mehrerer Eigenschaften/Variablen durchgeführt. Bei den energetischen Lastprofilen können somit Leistungswerte mehrerer Energieträger zeitgleich und abhängig voneinander diskretisiert werden.

Für die gleichzeitige Betrachtung mehrerer Lastprofile unterschiedlicher Energieträger kann es insbesondere dann vorteilhaft sein, wenn die Lastprofile der jeweiligen Energieträger nicht unabhängig voneinander sind. Die Abhängigkeiten können die weiterführenden energetischen Analysen (z.B. Verlustoptimierung) wesentlich beeinflussen. Wie in Kapitel 3 beschrieben kann die Variable *Zeit* der zeitlich abhängigen Lastprofile als Parameter verstanden werden. Nach einer Transformation/Eliminierung des Parameters Zeit beinhalten die entstandenen Datensätze die Informationen und Abhängigkeiten aller Energieträger. Die neuen Datensätze stellen energetische Zustände der energetischen Objekte dar. Ein Zustand beinhaltet alle zu diesem Zustand gehörenden Lastwerte der Energieträger. Im einfachsten Fall werden nur zwei Energieträger betrachtetet (siehe Abbildung 3.3), und es wird eine bivariate Diskretisierung durchgeführt.

Multivariate und bivariate Diskretisierungen sind auf Streudiagramme besonders gut anwendbar. Die Diskretisierung mehrerer Lastprofile erfolgt zeitgleich in einem Diskretisierungsprozess. Multivariate und bivariate Diskretisierungen sind die Methoden der Cluster-Analyse.

5.3 Auswahl

Die Auswahl der Diskretisierungsmöglichkeiten erfolgt im Hinblick auf die Zielsetzung der Arbeit. Die Zielsetzung ist die Diskretisierung von mehreren, zeitgleich betrachteten Lastprofilen unterschiedlicher Energieträger in einem Prozess. Somit soll die Diskretisierung der Lastprofile unter Beachtung von mehreren Eigenschaften/Variablen durchgeführt werden.

Schlussfolgernd ist die passende Diskretisierung für eine Betrachtung von mehreren zeitgleichen Lastprofilen die multivariate Diskretisierung, die Cluster-Analyse als Diskretisierungsmethode. Die dazu passend verwendete Darstellung ist das Streudiagramm.

6 Programme und Ergebnisse

Für weitere Untersuchungen, Bewertungen und Vergleiche der Methoden der Cluster-Analyse sind die Ergebnisse der Methoden von größerer Bedeutung. Der Fokus dieses Kapitels liegt auf der Darstellung der Ergebnisse der diskretisierten Lastprofile. Um die Ergebnisse berechnen zu können, wurden Programme anhand der Methoden der Cluster-Analyse aus Kapitel 4.2 erstellt. Die Programme wurden alle unter denselben Rahmenbedingungen entwickelt.

6.1 Rahmenbedingungen für alle Programme

Die Umsetzung der Methoden der Cluster-Analyse erfolgt in der Umgebung MATLAB. MATLAB ist eine „*Höhere Programmiersprache für nummerische Berechnungen, Visualisierung und Anwendungsentwicklung*“ [MAT00]. Mit der MATLAB-Umgebung können Programme entwickelt und graphisch visualisiert werden.

Bei der Entwicklung der Programme wurde auf eine einheitliche Struktur geachtet. Die Programme wurden im „Strukturierte-Programmierung“-Stil geschrieben. Dadurch wurde die Schnelligkeit der Programme auf Kosten der Lesbarkeit vom Code gesteigert. Diese Steigerung war notwendig, da die Ausführungsgeschwindigkeit von MATLAB beim „Prozedurale-Programmierung“-Stil signifikant langsamer war.

Die Entwicklung der Programmiercodes dient zum Verständnis der Methoden der Cluster-Analyse. Es wurden keine vorhandenen speziellen Funktionen für die Cluster-Analyse von MATLAB verwendet. Ein weiterer Grund für die eigenständige Entwicklung der Programmiercodes ist die Möglichkeit der Modifizierung der Codes.

Die Programme wurden auf der Betrachtung von zwei Lastprofilen für unterschiedliche Energieträger entwickelt. Durch das Einfügen einer Abstandsnorm höherer Dimensionen können jedoch auch mehr als zwei Lastprofile betrachtet werden.

Die verwendeten Eingangsdaten für die Programme sind die Daten des verwendeten Lastprofils (siehe Kapitel 2.3) in Form des Streudiagramms. Die Leistungswerte beider Energieträger können sowohl nicht normiert als auch normiert an das Programm weitergegeben werden. Die Normierung der Daten wird berechnet, indem die Daten durch den Maximal-Wert der Datenreihe geteilt werden.

Zur Berechnung des Abstandmaßes *d(x,y)* für zwei Energieträger wird der euklidische Abstand benutzt.

$$d(x,y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2} \quad \text{mit} \quad x,y\in\mathbb{R}^2 \tag{6.1}$$

Der Repräsentant eines Clusters ist durch den Centroid des Clusters gegeben.

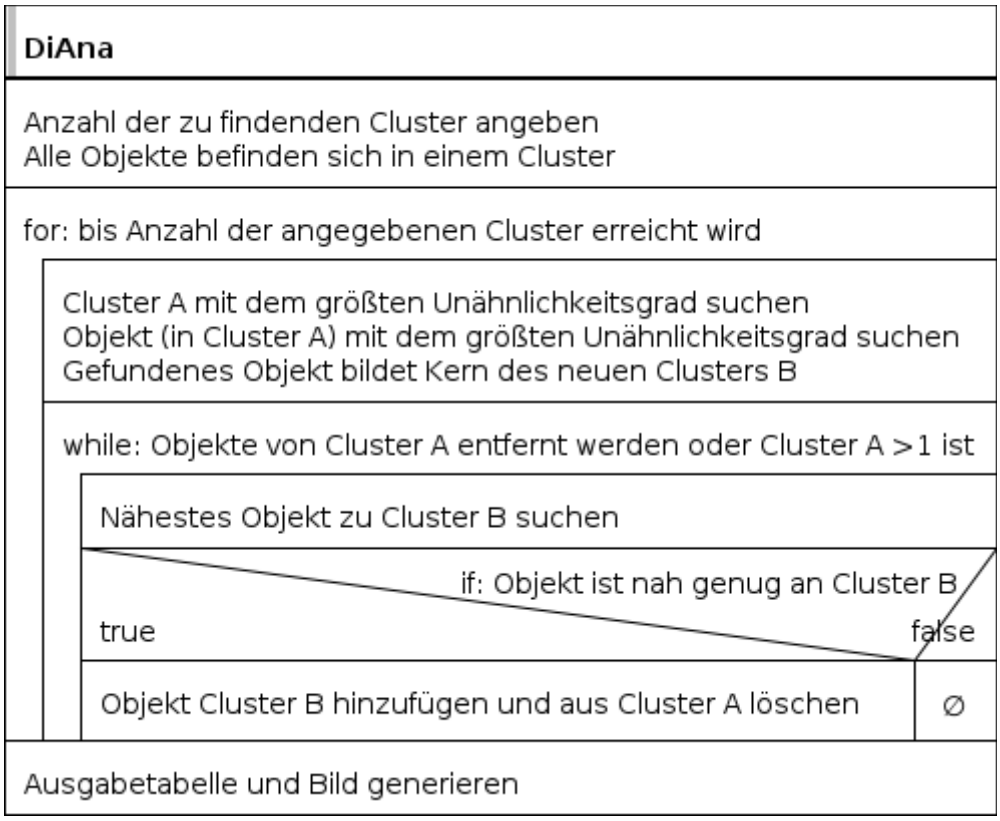
6.2 Ergebnisse

Im Weiteren werden die Ergebnisse der Programme in Abbildungen dargestellt. Für jede Methode wird ein Ergebnis mit den nicht nominierten Leistungswerten als Eingangsdaten für die Programme generiert. Die Vorgehensweise/Struktur der Programme wird durch Struktogramme wiedergegeben. Alle Programme können dem Anhang entnommen werden. Für die Verständlichkeit der Abbildungen müssen folgende Punkte beachtet werden:

- Die gefundene Abhängigkeitsstruktur der Objekte wird durch die Verwendung mehrerer Farben dargestellt. Jedes gefundene Cluster mit seinen Objekten wird durch eine Farbe gekennzeichnet. In der Tabelle wird jedes Cluster durch die Informationen einer Zeile wiedergegeben.
- Die Anzahl der zu suchenden Cluster bzw. die Anzahl der Repräsentanten in den Abbildungen ist vorgegeben. Die gesuchte Anzahl der Cluster beträgt fünf. Sie kann nur aufgrund der eigenen Strategie von einer Methode abweichen.
- Die Repräsentanten der Cluster werden durch schwarze Punkte in der Abbildung gekennzeichnet. Ihre Positionen können den beigefügten Tabellen der jeweiligen Abbildungen entnommen werden. Die Positionen sind durch die x- und y-Koordinaten in den jeweiligen Spalten gegeben.
- Die Anzahl der Objekte bzw. die Gewichtung der Cluster kann den beigefügten Tabellen der jeweiligen Abbildungen entnommen werden. Sie sind in der Spalte *Anzahl* festgehalten.

6.2.1 DiAna

Die Methode DiAna wurde mit dem Programm „DiAna“ durchgeführt (siehe Anhang B).



Struktogramm 6.1: DiAna

Tabelle 6.1: DiAna mit fünf Clustern

X	Y	Anzahl
0.87	1.31	2458
0.95	4.44	2045
1.16	7.40	1448
1.23	9.51	1637
1.28	12.71	1172

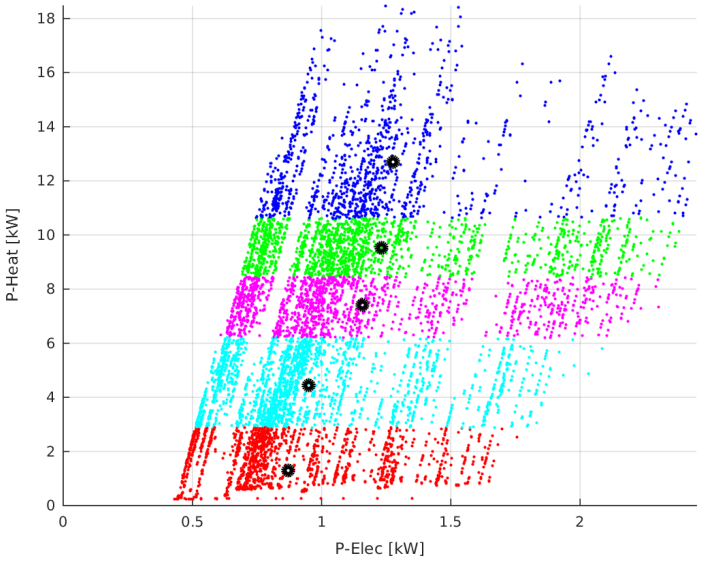


Abbildung 6.1: DiAna mit fünf Clustern

6.2.2 Fusionierungsalgorithmus

Die Methode Fusionierungsalgorithmus wurde mit dem Programm „Fusion“ durchgeführt (siehe Anhang B).

Fusion

Anzahl der zu findenden Cluster angeben
Alle Objekte befinden sind in einem Cluster

while: Anzahl der gegebenen Cluster nicht erreicht ist

for: bis alle Cluster durchlaufen sind

Zwei Cluster mit dem minimalsten Abstand ihrer Zentren suchen

Beide Cluster zu einem gemeinsammen Cluster zusammenfassen
Anzahl der Cluster um eins verringern

Ausgabetabelle und Bild generieren

Struktogramm 6.2: Fusion

Tabelle 6.2: Fusionierungsalgorithmus mit fünf Clustern

X	Y	Anzahl
0.87	1.22	2318
0.94	4.30	2140
1.19	8.83	3589
1.34	13.21	601
1.40	16.32	112

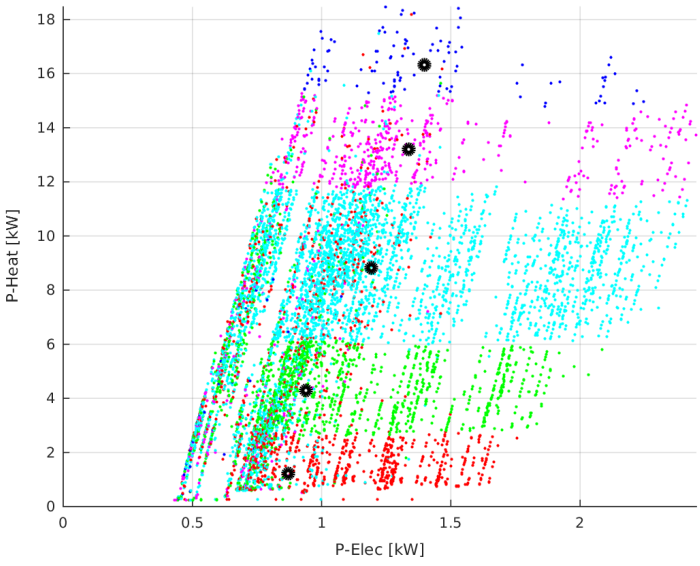


Abbildung 6.2: Fusionierungsalgorithmus mit fünf Clustern

6.2.3 K-Means-Algorithmus

Die Methode K-Means-Algorithmus wurde mit dem Programm „K-Means“ durchgeführt (siehe Anhang B).

K-Means

Anzahl der zu findenden Cluster angeben
Zufällige Startpunkte der Clusterzentren initialisieren

while: Clusterzentren sich bewegen
- Alte Position der Clusterzentren merken
- for: bis alle Objekte durchlaufen werden
  - Objekte werden den Zentren mit der minimalsten Entfernung zugeordnet
- for: bis alle Objekte durchlaufen werden
  - Neue Position der Zentren anhand der Neueinteilung der Objekte berechnen
- Bewegung der Zentren anhand alter und neuer Position der Zentren berechnen

Ausgabetabelle und Bild generieren

Struktogramm 6.3: K-Means

Tabelle 6.3: K-Means mit fünf Clustern

X	Y	Anzahl
0.88	1.14	2202
0.92	3.93	1916
1.13	7.10	1854
1.22	9.91	2097
1.33	13.76	691

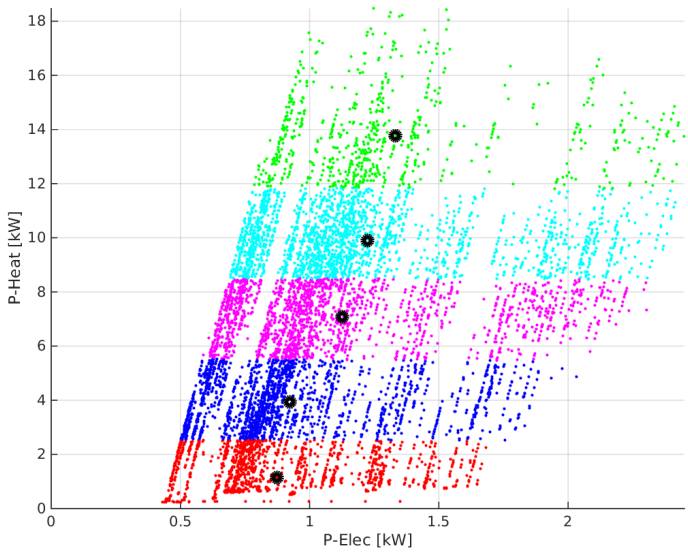


Abbildung 6.3: K-Means mit fünf Clustern

6.2.4 K-Means-Algorithmus-Modifiziert

Das Programm „K-Means-M“ (siehe Anhang B) ist eine modifizierte Version des klassischen K-Means-Algorithmus. Die Modifizierung wird im Struktogramm durch eine blaue Unterstreichung gekennzeichnet. Das Programm wurde aufgrund des Themengebietes entwickelt. Es berücksichtigt die wichtige Eigenschaft Spitzenlast eines Lastprofils.

K-Means-M

Anzahl der zu findenden Cluster angeben
Startpunkte der Clusterzentren initialisieren
-> Zwei den max. Werten der Energieträger zuordnen, Rest zufällig

while: Clusterzentren sich bewegen

- Alte Position der Clusterzentren merken
- for: bis alle Objekte durchlaufen werden
  - Objekte werden den Zentren mit der minimalsten Entfernung zugeordnet
- for: bis alle Objekte durchlaufen werden
  - Neue Position der Zentren anhand der Neueinteilung der Objekte berechnen
    Zentren bei den max. Werten bleiben beim max. Wert der Energieträger
- Bewegung der Zentren anhand alter und neuer Position der Zentren berechnen

Ausgabetabelle und Bild generieren

Struktogramm 6.4: K-Means-M

Tabelle 6.4: K-Means-M mit fünf Clustern

X	Y	Anzahl
0.87	1.56	2851
1.00	5.25	2211
1.20	9.10	2774
1.33	18.46	89
2.45	12.82	835

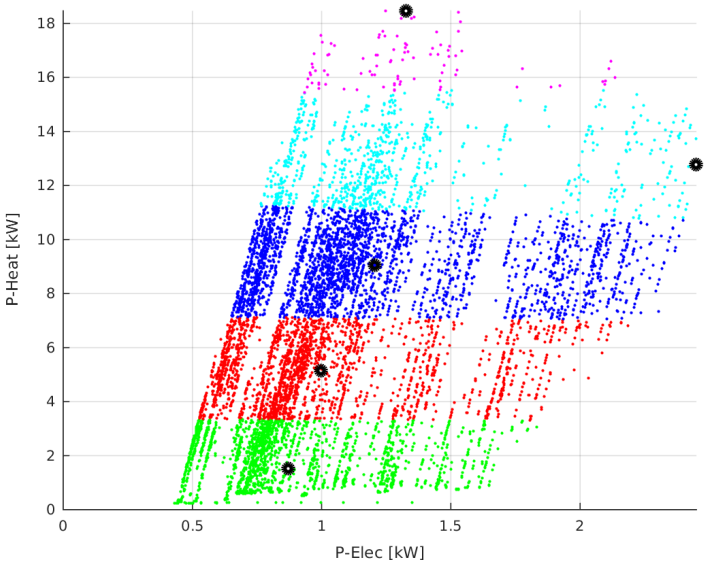
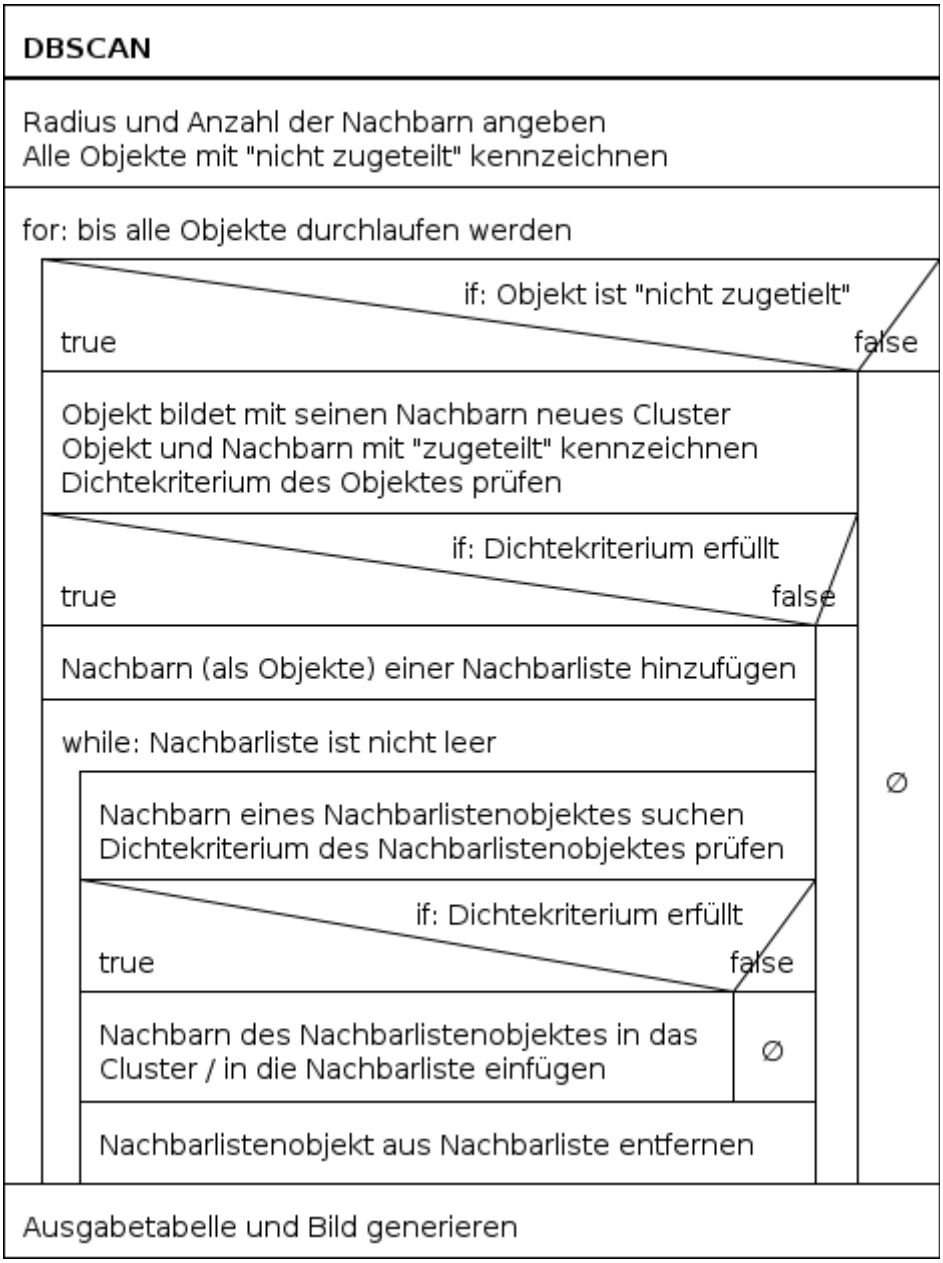


Abbildung 6.4: K-Means-M mit fünf Clustern

6.2.5 DBSCAN

Die Methode DBSCAN-Algorithmus wurde mit dem Programm „DBSCAN“ durchgeführt (siehe Anhang B). Für DBSCAN wird keine Anzahl von Clustern gegeben, sondern ein Radius und die Anzahl an Nachbarn, anhand derer der Algorithmus selber die Anzahl der Cluster bestimmt. Da die gefundene Clusteranzahl sehr groß ist, werden nur die drei Repräsentanten der größten Cluster in der Ausgabetabelle und in der Abbildung dargestellt. Die Kategorie „Noise“ wird im Programm vernachlässigt, da keine Rauschfilterung der Objekte stattfinden soll. Es werden die Objekte mit allen gefundenen Nachbarn (unabhängig von der Anzahl) in ein Cluster zusammengefasst.



Struktogramm 6.5: DBSCAN

Tabelle 6.5: DBSCAN (mit den drei größten Clustern)

X	Y	Anzahl
0.50	0.32	254
0.91	5.50	6272
1.23	12.70	64

Anzahl der Cluster = 664

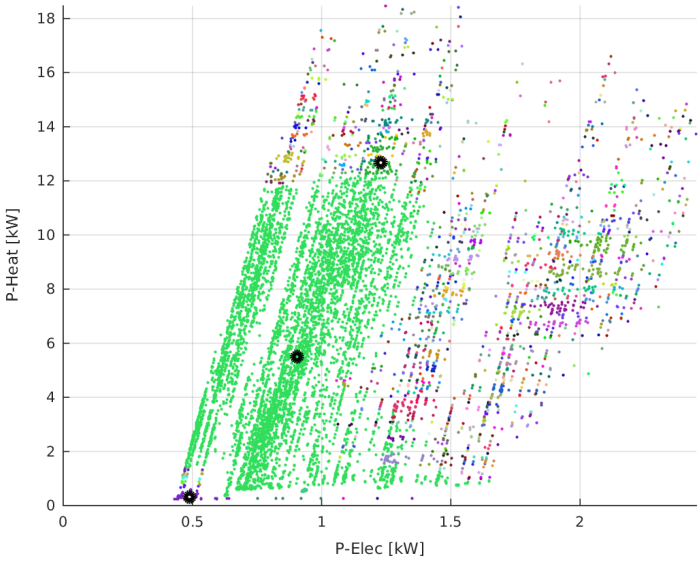


Abbildung 6.5: DBSCAN (mit den drei größten Clustern)

6.3 Fazit

Zum Schluss dieses Kapitels ist feststellbar, dass die Ergebnisse der Programme DiAna, Fusion, K-Means und K-Means-M sehr ähnlich sind (siehe Abbildung 6.6). Das Ergebnis des DBSCAN wird aufgrund seiner hohen Anzahl von Clustern und damit seiner großen Unterscheidung von den anderen Ergebnissen vernachlässigt.

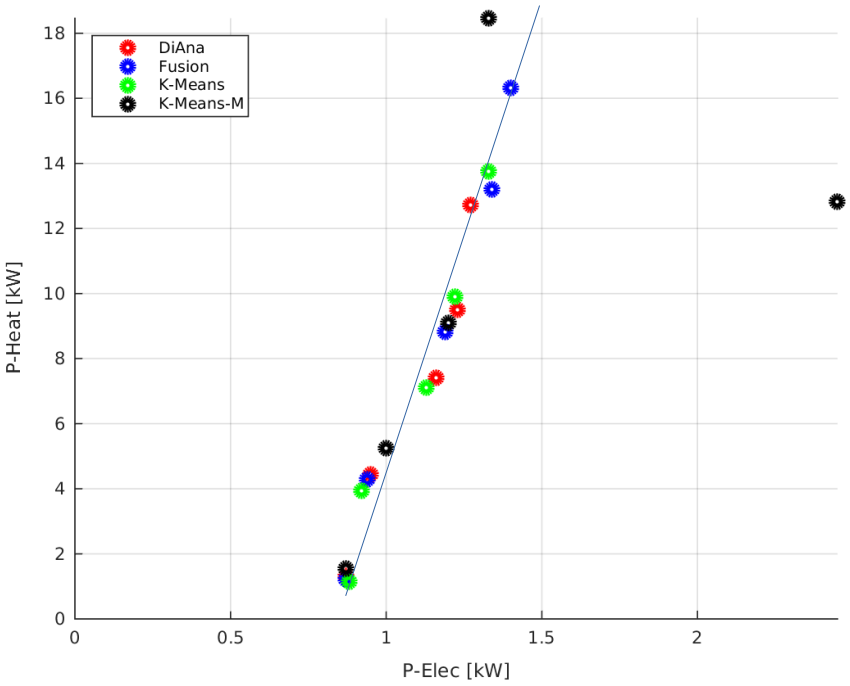


Abbildung 6.6: Ergebnisse für nicht normierte Daten

Die Ähnlichkeit der Ergebnisse der Programme ist durch die Abbildung 3.4 begründbar. Da die Programme auf Abstandsnormen basieren, beeinflussen höhere Werte die Einteilung der Objekte mehr als kleine Werte. Durch die viel höheren Leistungswerte des Energieträgers Gas entsteht bei der Diskretisierung eine zeilenhafte Einteilung der Objekte. Die entstandenen Repräsentanten können somit durch eine annähernde Diagonale in der Abbildung 6.6 dargestellt werden.

Werden die Programme auf normierte Eingangsdaten angewendet, entstehen andere Ergebnisse für die Programme (siehe Abbildung 6.7). Auch hier ist jedoch eine Ähnlichkeitsstruktur der einzelnen Ergebnisse erkennbar. Die entstandenen Repräsentanten können annähernd durch zwei Diagonalen dargestellt werden.

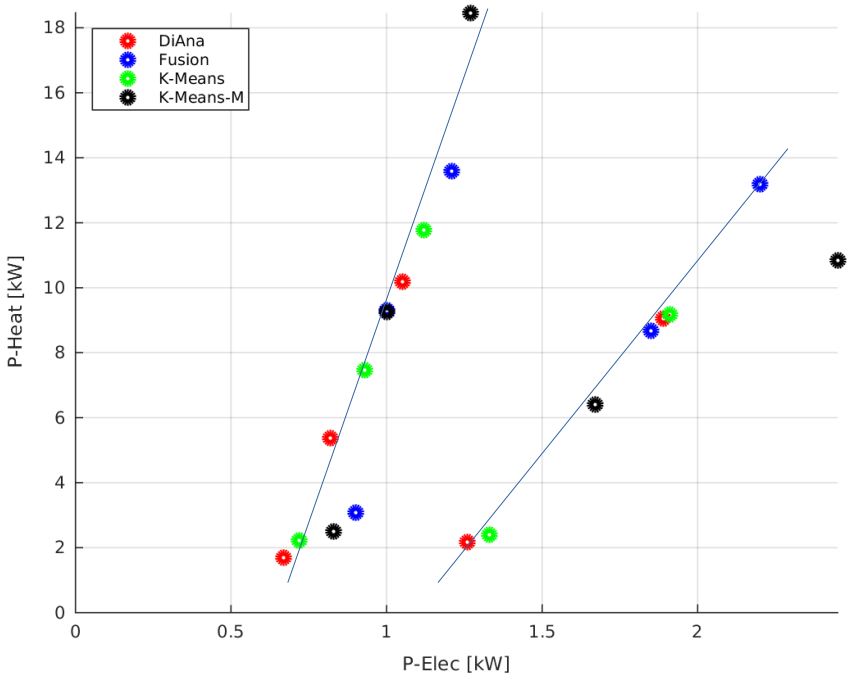


Abbildung 6.7: Ergebnisse für normierte Daten

Für eine weitere interessante Sichtweise können dem Anhang Diskretisierungen des Streudiagramms durch Menschenhand entnommen werden (siehe Anhang A und Anhang B).

7 Vergleich und Auswertung der Ergebnisse im Hinblick auf die Verwendung für Energiesystemmodelle

Durch die Diskretisierung wurde eine deutliche Reduktion der Komplexität und Anzahl der Eingangsdaten, der Lastprofile, ermöglicht. Mit einer Reduktion von Komplexität ist jedoch auch ein Verlust von Informationen der Lastprofile verbunden. Die Höhe und die Art des Verlusts an Informationen werden sehr stark durch die Wahl der Methode zur Diskretisierung, sowie auch durch die Wahl der späteren weiterführenden Analyse beeinflusst.

Im Folgenden werden die Ergebnisse aus Kapitel 6 verglichen. Hierfür werden zunächst Vergleichskriterien definiert und berechnet. Anschließend werden die Methoden in Bezug auf die Verwendung für Energiesystemmodelle ausgewertet.

7.1 Vergleichskriterien

Die möglichen Vergleichskriterien für die Programme sind im Sinne der energetischen Aufgabenstellung und der Darstellungsform der Datenreihen gewählt. Im Folgenden werden die Erstellung und die Ergebnisnotation für jedes Vergleichskriterium aufgeführt.

Ausführungsdauer :

Bei diesem Vergleichskriterium wird die zeitliche Dauer der Programme erfasst. Dauert ein Programm länger als eine Stunde, sind die Sekunden aufgrund ihres geringen Einflusses vernachlässigbar.

Peak-Erhaltung :

Dieses Vergleichskriterium zeigt an, ob durch die Programme die Spitzenwerte der beiden Energieträger nach der Diskretisierung geändert werden. Hierfür wird der Vergleichswert als eine relative Veränderung des Spitzenwertes vor und nach der Diskretisierung verstanden.

Energieerhaltung:

Dieses Kriterium gibt wieder, wie weit sich die verbrauchte Energie des Systems durch die Diskretisierung geändert hat. Der Vergleichswert wird hierbei als eine relative Veränderung der Energie vor und nach der Diskretisierung verstanden.

Oberfläche:

Dieses Vergleichskriterium gibt wieder, wie weit sich die aufgespannte Oberfläche der Daten durch die Diskretisierung geändert hat. In der Abbildung 7.1 wird beispielhaft eine Visualisierung des Vergleichskriteriums gezeigt. Die blaue Fläche stellt die aufgespannte Oberfläche der Eingangsdaten dar, die rote Fläche die aufgespannte Fläche der diskretisierten Daten. Der Vergleichswert wird als eine relative Veränderung der aufgespannten Oberfläche der Daten vor und nach der Diskretisierung verstanden.

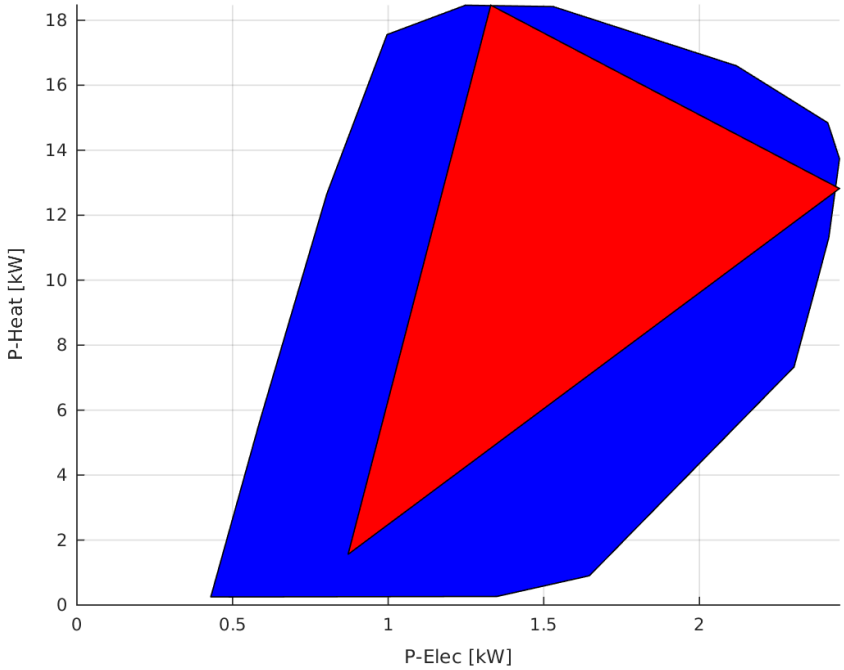


Abbildung 7.1: Oberflächenvergleich anhand von K-Means-M

Determinismus (Algorithmus):

Dieses Vergleichskriterium gibt wieder, ob bei der Durchführung der Programme bei derselben Anzahl an Clustern immer dieselben Ergebnisse entstehen. Ein Algorithmus ist deterministisch, wenn die Ausgabedaten bei den selben Eingabedaten immer identisch sind.

7.2 Ergebnis der Vergleichskriterien

Die Ergebnisse der Vergleichskriterien wurden mit den Ergebnissen aus Kapitel 6 berechnet. Sie sind in der Tabelle 6 festgehalten:

Tabelle 7.1: Ergebnisse der Vergleichskriterien

	Ausführungs-dauer	Peak-Erhaltung (Elec/Gas)	Energieerhaltung (Elec/Gas)	Oberfläche	Determinismus (Algorithmus)
DiAna	3 h 22 min	0.5/0.7	1.00/1.00	0.008	1
Fusion	1 h 56 min	0.5/0.7	1.00/1.00	0.009	1
K-Means	0.65 s – 0.8 s	0.5/0.9	1.00/1.00	0.004	0
K-Means-M	0.65 s – 0.8 s	1.0/1.0	1.54/1.02	0.413	0
DBSCAN	25 s – 35 s	1.0/1.0	1.00/1.00	0.992	1

Für die Interpretation der Tabelle müssen folgende Punkte beachtet werden:

- DBSCAN ist sehr schwer mit den anderen Programmen zu vergleichen. Es muss beachtet werden, dass die Clusteranzahl von DBSCAN ein Vielfaches von der Clusteranzahl der anderen Programme beträgt. Somit basieren die Vergleichswerte nicht auf denselben Grundlagen.
- Die Ausführungsdauer eines Programms soll so kurz wie möglich sein. *Gewinner: K-Means und K-Means-M*
- Die Vergleichswerte für die Peak-Erhaltung sind *zwischen 0 und 1*. Bei einer Erhaltung der Spitzenlast beträgt der Vergleichswert *1*. Für manche weiterführenden Analysen ist die Erhaltung der Spitzenlast sehr wichtig. *Gewinner: K-Means-M (,DBSCAN)*
- Die Vergleichswerte für die Energieerhaltung sind *zwischen 0 und* ∞. Bei einer Erhaltung der Energie beträgt der Vergleichswert *1*. Für mehr Energie wird der Vergleichswert > 1, für weniger Energie < 1. Für manche weiterführenden Analysen ist die Erhaltung der Energie sehr wichtig. *Gewinner: alle außer K-Means-M*
- Die Vergleichswerte für die Oberfläche sind *zwischen 0 und 1*. Bei einer kompletten Erhaltung der aufgespannten Oberfläche der Daten beträgt der Vergleichswert *1*. *Gewinner: (DBSCAN)*
- Die Vergleichswerte für den Determinismus sind *entweder 0 oder 1*. Determiniert der Algorithmus, so beträgt der Vergleichswert *1*. *Gewinner: DiAna, Fusion (,DBSCAN)*

7.3 Auswertung der Ergebnisse im Hinblick auf die Verwendung für Energiesystemmodelle

Grundsätzlich existiert eine Vielzahl an Energiesystemmodellen. Ein Energiesystemmodell wird meistens zur Optimierung eines energetischen Verhaltens innerhalb eines Systems entworfen. Hierbei können verschiedene Optimierungsarten betrachtet werden. Im Folgenden werden zwei mögliche Optimierungsarten erklärt. Im Zuge der Erklärung erfolgt auch eine Auswahl der geeigneten Methoden für die jeweilige Optimierungsart:

- Verlustoptimierung
- Spitzenlastabdeckung

7.3.1 Verlustoptimierung

Das Ziel der Verlustoptimierung ist es, Energie effizient und mit geringem Verlust zur Verfügung zu stellen. Für solch eine energetische Betrachtung sind vor allem die Energiezustände von Bedeutung, die innerhalb einer Zeitperiode am meisten auftreten. Es ist geschickt, bei diesen Zuständen die Verlustoptimierung vorzunehmen, da sich dadurch über eine längere Zeit die eingesparte Energie aufsummiert. Wird dagegen die Optimierung bei einem kurz anhaltenden Energiezustand durchgeführt, so wird nur die eingesparte Energie innerhalb dieser kurzen Dauer aufsummiert. Eine Übertragung der Überlegung auf diskretisierte Lastprofile lässt einen zu der Schlussfolgerung kommen, dass das Cluster mit der größten Anzahl an Objekten die höchste Priorität für das Energiesystemmodell hat.

Ein weiterer wichtiger Aspekt muss bei den diskretisierten Lastprofilen beachtet werden. Um eine korrekte Verlustoptimierung nach einer Diskretisierung der Lastprofile durchführen zu können, müssen die diskretisierten Daten dieselbe Gesamtenergie aufweisen wie die nicht diskretisierten Daten. Somit können die Programme DiAna sowie Fusion (und DBSCAN) für eine vorherige Diskretisierung der Lastprofile angewendet werden. Anschließend wird das Cluster mit der größten Anzahl an Objekten betrachtet.

7.3.2 Spitzenlastabdeckung

Das Ziel der Spitzenlastabdeckung ist die Sicherstellung der Energieverfügbarkeit und somit eine Vermeidung von Überlastung des Energiesystems. Für solch eine energetische Betrachtung sind vor allem die größten Spitzenlasten unterschiedlicher Energieträger eines Energiesystems von Bedeutung. Können diese durch Energiezufuhr abgedeckt werden, entsteht keine Überlastung des Energiesystems. Eine Übertragung der Überlegung auf diskretisierte Lastprofile lässt einen zu der Schlussfolgerung kommen, dass die Cluster mit dem Repräsentanten des größten Leistungswerts eines Energieträgers die höchste Priorität für das Energiesystemmodell haben.

Ein weiterer wichtiger Aspekt muss bei den diskretisierten Lastprofilen beachtet werden. Um eine korrekte Spitzenlastabdeckung nach einer Diskretisierung der Lastprofile durchführen zu können, müssen die diskretisierten Daten dieselbe Spitzenlast aufweisen wie die nicht diskretisierten Daten. Somit kann das Programm K-Means-M (und DBSCAN) für eine vorherige Diskretisierung der Lastprofile angewendet werden. Anschließend wird das Cluster des Repräsentanten mit dem höchsten Leistungswert betrachtet.

7.4 Fazit

Zum Schluss dieses Kapitels ist allgemein festzuhalten, dass es keinen optimalen Diskretisierungsprozess für alle betrachteten Aspekte gibt. Jedoch hängt die Wahl eines geeigneten Diskretisierungsprozesses im Hinblick auf die Verwendung für Energiesystemmodelle hauptsächlich von den Vergleichskriterien *Peak-Erhaltung und Energieerhaltung* ab. Grundsätzlich erfüllt kein Programm beide Kriterien gleichzeitig. Erfüllen hingegen mehrere Programme ein Kriterium, so kann die engere Auswahl für die Diskretisierung mit Hilfe der restlichen Vergleichskriterien *Ausführungsdauer, Oberfläche* und *Determinismus (Algorithmus)* getroffen werden.

8 Zusammenfassung und Ausblick

Ziel der Arbeit war die Verallgemeinerung mehrerer, unabhängiger Diskretisierungsprozesse für mehrere Lastprofile in einen gemeinsamen Diskretisierungsprozess. Es wurden hierfür unterschiedliche Darstellungs- und Diskretisierungsmöglichkeiten für Lastprofile erläutert. Durch die Betrachtung der beiden Grundlagen in Bezug auf die Zielsetzung wurde der in Abbildung 8.1 dargestellte, mögliche Pfad für die Verallgemeinerung mehrerer Diskretisierungsprozesse zu einem Diskretisierungsprozess gefunden.

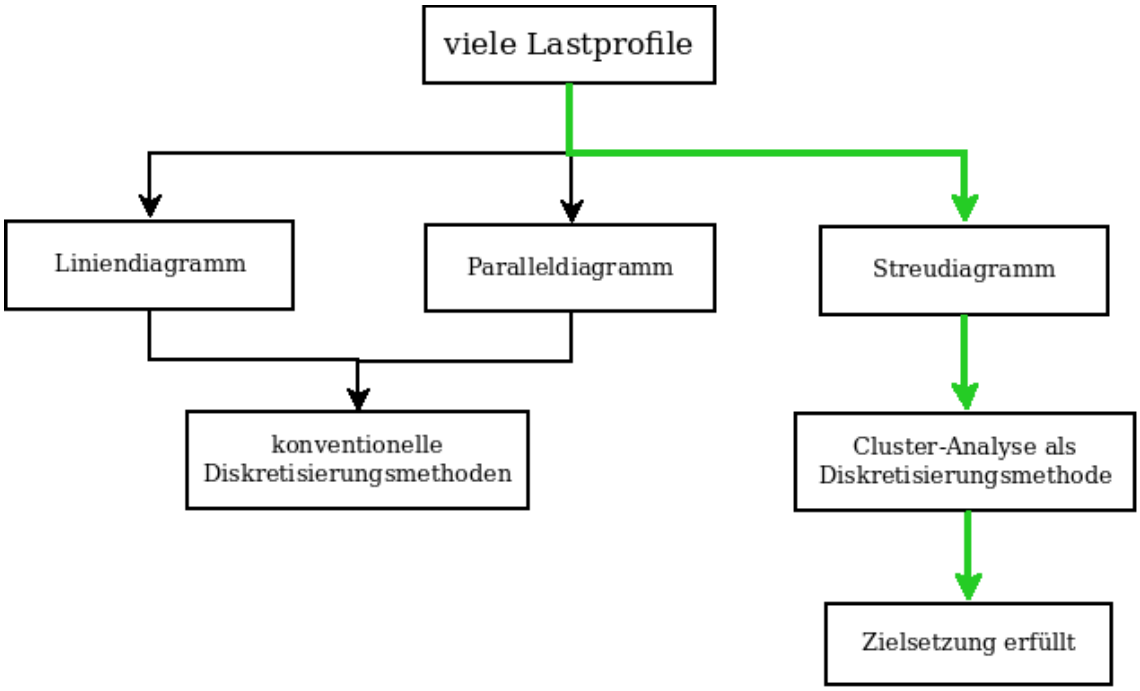


Abbildung 8.1: Zusammenfassung

Aufgrund der zahlreichen Methoden der Cluster-Analyse sind mehrere Möglichkeiten einer verallgemeinerten Diskretisierung vorhanden. Durch einen Vergleich der Ergebnisse der Methoden konnte festgestellt werden, dass die Ergebnisse im Hinblick auf die Verwendung für Energiesystemmodelle nicht durch eine „Königsmethode“ erstellt werden können. Mit Hilfe der Betrachtung der Optimierungsart des Energiesystemmodells kann jedoch eine Auswahl der geeigneten Methode getroffen werden.

Eine Ausweitung bzw. ein Vergleich der in dieser Arbeit erzielten Ergebnisse mit weiteren Vergleichsmöglichkeiten für Diskretisierungen von Lastprofilen stellt einen spannenden Ausblick dar. Ein in der Arbeit angesprochener wichtiger Faktor der Diskretisierung ist die Anzahl der Repräsentanten einer Diskretisierung. Eine geringe Repräsentantenbildung bei einer Diskretisierung führt zu einem höheren Informationsverlust der Lastprofile. Sowohl die Balance zwischen der Anzahl an Repräsentanten und der Informationserhaltung, als auch die Auswirkung der Anzahl an Repräsentant auf die Vergleichsergebnisse kann in weiterführenden Arbeiten untersucht werden. Mit Hilfe weiterer Untersuchungen können die Diskretisierungsprozesse für mehrere zeitgleich betrachtete Lastprofile optimiert werden.

Anhang A: Beispiel für Diskretisierung durch Menschenhand

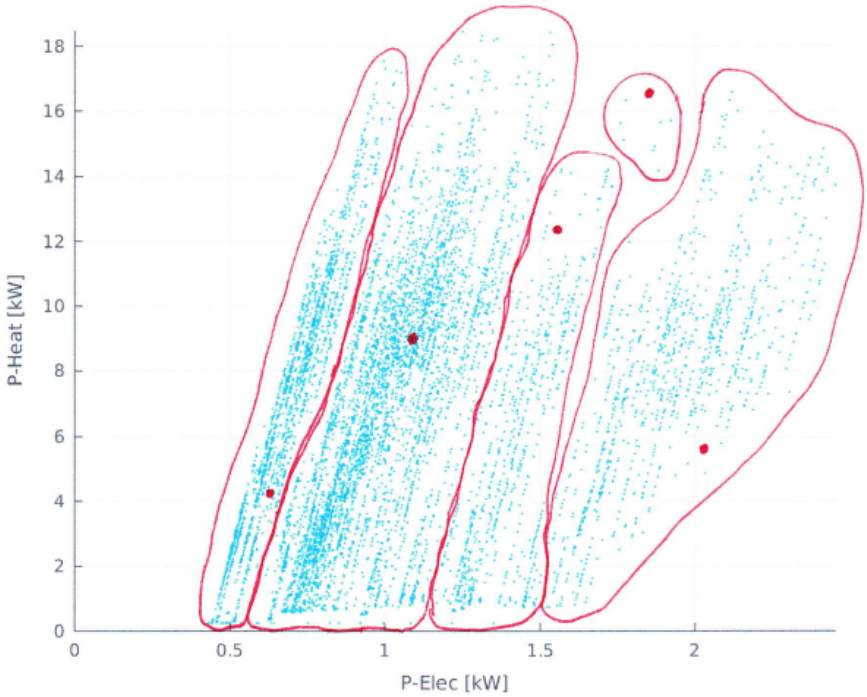
Recherche für Bachelorarbeit:

Verallgemeinerung der Diskretisierung von Jahresdauerlinien für mehrere Energieträger

Aufgabe:

In der folgenden Abbildung können Sie 8760 blaue Punkte sehen.

1. Teilen Sie die Punkte in **fünf** Mengen ein. Umranden Sie hierfür alle Punkte, die zu einer Menge gehören sollen. Jeder Punkt muss innerhalb eines Ihrer gezeichneten Einteilungen sein.
2. Zeichnen Sie in jede Menge einen Punkt. Der Punkt soll die Punkte einer Menge am besten repräsentieren.



Anhang B: CD

Inhalt:

- Bachelorarbeit
- Abschlusspräsentation
- USA_AK_Anchorage.Intl.AP.702730_TMY3_BASE.csv
- Diskretisierung durch Menschenhand
- Programme
- Literatur

Literaturverzeichnis

[EDE00]: EDER, Tobias, (o. Jahr): Energiesystemmodelierung. Paper, Max-Planck-Institut für Plasmaphysik, Garching

[RP00]: PASCHOTTA, Rüdiger, (o. Jahr): Lastprofil, in: RP-Energie-Lexikon, https://www.energie-lexikon.info/lastprofil.html, [Stand 11.02.15]

[IW00]: O. Verf., (o. Jahr): Datensatz, in: IT Wissen. Das große Online-Lexion für Informationstechnologie, http://www.itwissen.info/definition/lexikon/Datensatz-data-record.html, [Stand 08.02.15]

[OEI13]: Commercial and Residential Reference Building Models , (2013): Commercial and Residential Hourly Load Profiles for all TMY3 Locations in the United States , in: Open Energie Information, http://en.openei.org/doe-opendata/dataset/commercial-and-residential-hourly-load-profiles-for-all-tmy3-locations-in-the-united-states, [Stand 11.02.15]

[WM89]: WOHINZ, W. Josef; MOOR Michael, (1989): Betriebliches Energiemanagment: Aktuelle Investion in die Zukunft. Springer Verlag Wien New York, S. 73

[LHT+02]: LIU, Huan; HUSSAIN, Farhad; TAN, Chew Lim; DASH, Manoranjan, (2002): Discretization: An Enabling Technique. Kluwer Academic Publishers

[KUH12]: KUHN, Philip, (2012): Iteratives Modell zur Optimierung von Speicherausbau und -betrieb in einem Stromsystem mit zunehmend fluktuierender Erzeugung. Dissertation, Technsiche Universität München

[CL14]: CLEVE, Jürgen; LÄMMEL, Uwe, (2014): Data Mining. Oldenbourg Wissenschaftsverlag GmbH

[HL00]: Hochschule Luzern, (o. Jahr): Clusteranalyse. http://www.google.de/imgres?imgurl=http%3A%2F%2Fwww.empirical-methods.hslu.ch%2Fclusteranalyse-abb3.jpg&imgrefurl=http%3A%2F%2Fwww.empirical-methods.hslu.ch%2Fh-clusteranalyse&h=625&w=654&tbnid=RTBhtLfN1NRGnM%3A&zoom=1&docid=EEHrjp9OESJ0oM&ei=TiSsVOTvCYXaOL-OgZAL&tbm=isch&iact=rc&uact=3&dur=811&page=1&start=0&ndsp=18&ved=0CC4QrQMwAg [Stand 08.02.15]

[SAY15]: SAYAD, Saed, (2015): Hierarchical Clustering. http://www.saedsayad.com/clustering_hierarchical.htm [Stand 12.02.15]

[SHR00]: STRUYF, Anja; HUBERT, Mia; ROUSSEEUW, Peter J., (o. Jahr): Clustering in an Object-Oriented Environment. Paper, University of Antwerpen, S. 20 - 23

[MAT00]: O. Verf., (o. Jahr): MATLAB. Hauptmerkmale. http://de.mathworks.com/products/matlab/features.html [Stand 08.02.15]