TUHH
Technische
Universität
Hamburg

INSTITUT ieet
FÜR ELEKTRISCHE ENERGIETECHNIK

Kategoriebasierte Lastprognose in Niederspannungsnetzen zur Day-Ahead-Koordinierung haushaltsnaher Flexibilität

Joost Henning Lindner

Joost Henning Lindner

54119

M.Sc. Elektrotechnik

Masterarbeit

25.008

November 2025

Prüfer: Prof. Dr.-Ing. Christian Becker

Zweitprüfer: Prof. Dr. rer. pol. Kathrin Fischer

Betreuer: Finn Nußbaum

TUHH
Technische
Universität
Hamburg

INSTITUT ieet
FÜR ELEKTRISCHE ENERGIETECHNIK

Masterarbeit Joost Lindner

Datum: 12.05.2025

Kategoriebasierte Lastprognose in Niederspannungsnetzen zur Day-Ahead-Koordinierung haushaltsnaher Flexibilität

Vor dem Hintergrund der Energiewende werden im Bereich von Haushalten bzw. Wohngebäuden zunehmend flexible Anlagen wie Wärmepumpen und Ladestationen für E-Autos sowie dezentrale Erzeugung in Form von PV installiert. Dies wird zu einer Zunahme von Engpässen im elektrischen Verteilnetz führen. Diese zu lösen, wird zu einer immer wichtigeren Aufgabe der Netzbetreiber. Neben kurativen Maßnahmen wie nach § 14a EnWG gewinnen präventive Maßnahmen an Bedeutung. Im Rahmen des Forschungsprojektes „KoLa“ wird hierzu eine sog. Koordinierungsfunktion entwickelt, die am Vortag Leistung von Haushaltkunden verschieben soll.

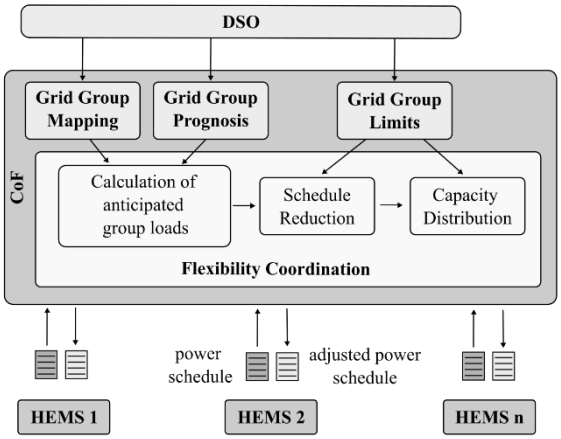


Abbildung 1: Konzept einer Koordinierungsfunktion zur Nutzung haushaltsnaher Flexibilität

Ein wesentlicher Bestandteil einer erfolgreichen Flexibilitätskoordinierung ist eine zuverlässige Prognose der Netzlast, auf deren Basis Leistung verschoben wird. Neben der Fahrplaninformation der an der Koordinierung beteiligten Kunden sind dazu Prognosen der übrigen Lasten der Niederspannungsabgänge notwendig. Im Rahmen dieser Arbeit soll daher ein Verfahren entwickelt werden, das unter Kenntnis der Netztopologie, der anfragenden Kunden und ihrer Planungen sowie ggfs. vorhandenen historischen Messdaten der übrigen Netzteilnehmer die Last einzelner Niederspannungsabgänge vorhersagt. Die Prognose soll auf Basis möglichst weniger zusätzlicher Messdaten funktionieren. Hierzu wird eine vorherige Kategorisierung der Netzkunden und eine Entwicklung von Kategorie-spezifischen Vorhersagen vorgeschlagen. Im Anschluss ist das Prognoseverfahren hinsichtlich der Güte zu bewerten und es sind Empfehlungen bezüglich der notwendigen Datengrundlage abzugeben.

Folgende Strukturierung der Arbeit wird vorgeschlagen:

- Literaturrecherche zur Vorhersage von Niederspannungsabgängen, insbesondere zu Kategorie-basierten Ansätzen
- Kategorisierung der Netzkunden auf Basis bekannter Verbrauchermerkmale
- Entwicklung einer Prognose auf Basis der vorhandenen Informationen
- Erweiterung der Prognose um zusätzliche Messdaten der Netzkunden
- Bewertung der Güte der Netzprognose und der Anwendbarkeit im Rahmen der Koordinierung

Im Anschluss an diese Arbeit ist in einem Vortrag über die Ergebnisse zu berichten.

Ansprechpartner: Finn Nußbaum, Tel.: 040 428 78 4092, Mail: finn.nussbaum@tuhh.de

Kurzfassung

Der zunehmende Anteil dezentraler Erzeuger und flexibler Verbraucher erschwert die Gewährleistung der Netzstabilität und erhöht die Anforderungen an die Steuerung und Überwachung von Verteilnetzen. Im Rahmen eines Forschungsprojekts wird eine Koordinierungsfunktion (KOF) entwickelt, die flexible Haushaltsleistungen am Vortag verschiebt, um Netzengpässe präventiv zu vermeiden. Da Fahrplandaten vieler Haushalte unvollständig vorliegen, entstehen Unsicherheiten in der Netzbewertung. Diese Arbeit setzt an dieser Stelle an und erweitert das bestehende Konzept um ein Prognoseverfahren, das die KOF mit einer vollständigen und belastbaren Datengrundlage unterstützt. Hierzu wird ein datenbasierter Ansatz entwickelt, der mithilfe des HDBSCAN-Clustering-Verfahrens Haushalte mit ähnlicher Ausstattung und Verbrauchsverhalten kategorisiert, um daraus repräsentative Lastprofile zu bilden. Fehlende Fahrpläne können so plausibel ergänzt werden. Anschließend verfeinert eine Gradient Boosting Regression die clusterbasierten Prognosen durch zusätzliche Einflussgrößen wie Wetter-, Preis- und Netzdaten. Die Ergebnisse zeigen, dass die Kombination aus Clustering und Regressionsanalyse eine effektive Methode zur Prognose von Haushaltsfahrplänen darstellt. Dadurch wird die Datengrundlage für die KOF deutlich verbessert und eine robuste Lastprognose auch bei eingeschränkter Datenverfügbarkeit ermöglicht. Die Clusterprognose liefert auf aggregierter Ebene eine hohe Prognosegüte, während die Regressionsanalyse die Genauigkeit auf der Haushaltsebene weiter erhöht. Einzelne Haushalte mit stark individuellen Verbrauchsmustern lassen sich nur begrenzt abbilden, was zwar die effiziente Nutzung von Flexibilität im Netz beeinträchtigen kann, die Gesamtaussagekraft der Prognose jedoch nur geringfügig beeinflusst. Künftige Arbeiten sollten reale Smart-Meter-Daten zur Validierung und Verbesserung der Modellgüte einbeziehen sowie den Einsatz Deep-Learning-basierter Verfahren prüfen, um komplexe zeitliche Muster und kurzfristige Verbrauchsänderungen noch präziser zu erfassen.

Abstract

The increasing share of decentralized producers and flexible consumers complicates the maintenance of grid stability and raises the requirements for the control and monitoring of distribution networks. As part of a research project, a coordination function (COF) is being developed that performs day-ahead scheduling of flexible household loads to proactively prevent grid congestion. Since the schedule data of many households are incomplete, uncertainties arise in grid assessment and flexibility planning. This work addresses this issue and extends the existing concept by developing a forecasting method that provides the COF with a complete and reliable data basis. To this end, a data-driven approach is developed that applies the HDBSCAN clustering algorithm to categorize households with similar technical characteristics and consumption behavior, thereby forming representative load profiles. Missing schedules can thus be plausibly reconstructed. Subsequently, a gradient boosting regression refines the cluster-based forecasts by incorporating additional influencing factors such as weather, price, and grid data. The results demonstrate that the combination of clustering and regression analysis provides an effective approach for forecasting household schedules. This highly improves the data foundation for the COF and enables robust load forecasting even under limited data availability. The cluster forecast provides a high level of predictive accuracy at the aggregated level, while the regression analysis further improves accuracy at the household level. Individual households with highly specific consumption patterns can only be represented to a limited extent, which may affect the efficient use of flexibility in the grid but has only a minor impact on the overall reliability of the forecast. Future work should integrate real smart meter data to validate and improve the model quality and examine the use of deep learning-based methods to capture complex temporal patterns and short-term consumption changes even more accurately.

Inhaltsverzeichnis

Abkürzungen

PV Photovoltaik

KOF Koordinierungsfunktion

KOLA Koordinierungsfunktion des Verteilnetzes und Lastmanagement des elektrifizierten Personennahverkehrs

HEMS Home Energy Management System

VNB Verteilnetzbetreiber

DBSCAN Density-Based Spatial Clustering of Applications with Noise

HDBSCAN Hierarchical Density-Based Spatial Clustering of Applications with Noise

ARIMA Auto-Regressive Integrated Moving Average

LSTM Long Short-Term Memory

EnWG Energiewirtschaftsgesetz

DTW Dynamic Time Warping

GMM Gaussian Mixture Model

PCA Principal Component Analysis

UMAP Uniform Manifold Approximation and Projection

DBI Davies-Bouldin-Index

CHI Calinski-Harabasz-Index

MAE Mean Absolute Error

RMSE Root Mean Squared Error

GBM Gradient Boosting Machines

SVM Support Vector Machines

KNN k-Nearest Neighbor

BEV Battery Electric Vehicle

Formelzeichen

Kapitel 2

λ_t^{dem}	Strompreis für den Netzbezug zum Zeitpunkt t [€/kWh]
λ_t^{sup}	Vergütung für die Netzeinspeisung zum Zeitpunkt t [€/kWh]
$\mathrm{P}_t^{\text{Grid,dem}}$	Leistung für den Strombezug aus dem Netz [kW]
$\mathrm{P}_t^{\text{Grid,sup}}$	Leistung für die Einspeisung ins Netz [kW]
$\mathrm{P}_t^{\text{infl}}$	Inflexible elektrische Lasten des Haushalts [kW]
$\mathrm{P}_t^{\text{Bat}}$	Leistung der Batterie [kW]
$\mathrm{P}_{t,n}^{\text{BEV}}$	Leistung des n-ten Elektrofahrzeugs [kW]
P_t^{HP}	Elektrische Leistungsaufnahme der Wärmepumpe [kW]
P_t^{PV}	Leistung der Photovoltaikanlage [kW]
Δt	Zeitliche Auflösung des Optimierungsmodells [h]
λ_{BEV}	Strafkostenfaktor für Komfortverletzungen Elektrofahrzeug [€/kWh]
λ_{HP}	Strafkostenfaktor für Komfortverletzungen der Wärmepumpe [€/K]
v_{BEV}	Verletzungsvariable für den Ladezustand des Elektrofahrzeugs [kWh]
v_{HP}	Verletzungsvariable für Temperaturabweichungen der Wärmepumpe
C_{Bat}	Kostenanteil der Batterieverschleiß- bzw. Nutzungskosten [€]
N	Menge aller Elektrofahrzeuge im betrachteten Haushalt
T	Menge aller Zeitschritte des Optimierungshorizonts

Kapitel 3

K	Anzahl der Cluster
n_k	Anzahl der Datenpunkte im k-ten Cluster
$x_i^{(k)}$	i-ter Datenpunkt im Cluster k
u_k	Zentrum (Mittelwert) des k-ten Clusters
$\left\|x_i^{(k)} - u_k\right\|^2$	Quadrierter euklidischer Abstand zwischen Punkt und Clusterzentrum
ε	Radius des maximalen Punktabstands im Cluster
$MinPts$	Minimale Anzahl von Punkten, um ein Cluster zu bilden
x_{neu}	Neuer, zu klassifizierender Datenpunkt
x_i	Bekannter Datenpunkt aus der Menge X_{bekannt}

X_{bekannt}	Menge aller bekannten Datenpunkte
$N_k(x_{\text{neu}})$	Menge der k nächsten Nachbarn des Punktes x_{neu}
k	Anzahl der berücksichtigten nächsten Nachbarn
C	Menge aller existierenden Cluster
$C(x_{\text{neu}})$	Zugewiesenes Cluster des neuen Datenpunktes x_{neu}
$C(x_i)$	Cluster, dem der bekannte Punkt x_i angehört
$I(\cdot)$	Indikatorfunktion, die 1 liefert, wenn $C(x_i) = c$, andernfalls 0
DB_c	Davies-Bouldin-Index
c	Gesamtzahl der Cluster
c_i	Zentrum (Mittelwert) des i-ten Clusters
c_j	Zentrum (Mittelwert) des j-ten Clusters
$dia(c_i)$	Durchschnittliche Streuung der Datenpunkte innerhalb des Clusters c_i
n_i	Anzahl der Datenpunkte im Cluster c_i
x	Datenpunkt im Cluster c_i
$\|x - c_i\|$	Euklidischer Abstand zwischen Datenpunkt x und Clusterzentrum c_i
$CH(c)$	Calinski-Harabasz-Index
B_m	Zwischen-Cluster-Streumatrix
W_m	Innere Cluster-Streumatrix
N	Gesamtzahl der geclusterten Datenpunkte
$trace(\cdot)$	Spur einer Matrix
$(x - c_i)(x - c_i)^T$	Streuungsanteil eines Datenpunkts innerhalb seines Clusters
$(c_i - k)(c_i - k)^T$	Streuungsanteil zwischen Clusterzentren
$S(i)$	Silhouette-Koeffizient des Datenpunkts i
$a(i)$	Durchschnittlicher Abstand des Punkts i im Cluster
$b(i)$	Kleinster mittlerer Abstand zu anderem Cluster
y_i	Vorhergesagter Wert der Zielvariable für Beobachtung i
$x_{i1}, \ldots, x_{ik}$	Werte der unabhängigen Variablen für Beobachtung i
β_0	Achsenabschnitt der Regressionsgeraden
$\beta_1, \ldots, \beta_k$	Regressionskoeffizienten der unabhängigen Variablen
ε_i	Fehlerterm der Beobachtung i

n	Anzahl der Beobachtungen
k	Anzahl der unabhängigen Variablen
σ^2	Varianz der Fehlerterme
$m_{M,n}(x)$	Gesamte Vorhersage des Random Forest
M	Gesamtzahl der Entscheidungsbäume im Random Forest
$m_n(x; \Theta_j, D_n)$	Vorhersage des j-ten Baums
Θ_j	Zufallsparameter des j-ten Baums
D_n	Trainingsdatensatz mit n Beobachtungen
$m_{\infty,n}(x)$	Theoretische Vorhersage eines unendlich großen Random Forests
$\mathbb{E}_\Theta[\cdot]$	Erwartungswert über alle möglichen Zufallsparameter Θ
x	Eingabevektor bzw. Merkmalsvektor für die Vorhersage
$F_m(x)$	Modell nach der m-ten Iteration
$F_{m-1}(x)$	Modell aus der vorherigen Iteration
$\hat{F}_M(x)$	Finale Modellapproximation nach M Iterationen
$F^*(x)$	Unbekannte Zielfunktion, die approximiert werden soll
$h_m(x)$	Basisfunktion bzw. Entscheidungsbaum in der m-ten Iteration
ρ_m	Gewichtungsfaktor des m-ten Baums (Schrittweite)
$L(y, F(x))$	Verlustfunktion zur Bewertung der Modellgüte
r_{mi}	Pseudo-Residuum der Beobachtung i in Iteration m
$D = \{(x_i, y_i)\}_{i=1}^N$	Trainingsdatensatz
N	Anzahl der Trainingsbeispiele
MAE	Mittlerer absoluter Fehler
y_i	Tatsächlicher Wert der Zielvariablen für Beobachtung i
$\hat{y}_i$	Vom Modell vorhergesagter Wert für Beobachtung i
RMSE	Wurzel des mittleren quadratischen Fehlers
$\cos(\alpha)$	Kosinusähnlichkeit
A, B	Zu vergleichende Vektoren im Merkmalsraum
A_i, B_i	i-te Komponenten der Vektoren A und B
$\|A\|$	Betragsnorm (euklidische Länge) des Vektors A
ρ_p	Pearson-Korrelationskoeffizient

x_i, y_i	Wertepaare der zu vergleichenden Variablen X und Y
μ_X, μ_Y	Mittelwerte der Variablen X bzw. Y
$\mathrm{var}(X), \mathrm{var}(Y)$	Varianzen der Variablen X bzw. Y
σ_X, σ_Y	Standardabweichungen der Variablen X bzw. Y
$\mathrm{cov}(X, Y)$	Kovarianz zwischen den Variablen X und Y

Kapitel 4

$\tilde{L}_C(t, h)$	Skalierte Clusterlastkurve für Haushalt h zum Zeitpunkt t
$\bar{L}_C(t)$	Gemittelte (unskalierte) Clusterlastkurve zum Zeitpunkt t
E_h	Jahresenergieverbrauch des Haushalts h
$\bar{E}_C$	Durchschnittlicher Jahresenergieverbrauch aller Haushalte im Cluster
P_C	Repräsentativer Spitzenwert ähnlicher Haushalte mit Elektrofahrzeug
$L_{C,\max}$	Maximale Last der unskalierten Clusterkurve
t	Zeitindex
h	Index des betrachteten Haushalts
hour	Stunde des Tages
$\mathrm{hour}_{\sin}$	Sinus-transformierte Stundenkomponente
$\mathrm{hour}_{\cos}$	Kosinus-transformierte Stundenkomponente
day	Wochentag
$\mathrm{day}_{\sin}$	Sinus-transformierte Tageskomponente
$\mathrm{day}_{\cos}$	Kosinus-transformierte Tageskomponente

1 Einleitung

Die Energiewende in Deutschland führt zu einem grundlegenden Wandel des Energiesystems. Die zunehmende Dezentralisierung der Energieerzeugung, der starke Ausbau von fluktuierenden erneuerbaren Energien sowie die fortschreitende Elektrifizierung der Endverbraucher sind charakteristisch für diesen Umbruch der Energieversorgung [1]. Durch diese strukturellen Veränderungen entstehen neue Herausforderungen für den Betrieb und die Planung elektrischer Verteilnetze. Die bisher vorwiegend passiv ausgelegten Niederspannungsnetze stoßen im Zuge der vermehrten Integration von verbrauchernahen Technologien wie Photovoltaik (PV), Wärmepumpen und Ladeinfrastruktur für Elektrofahrzeuge zunehmend an ihre technischen und operativen Grenzen [2]. Ein zentrales Problem dabei ist, dass diese neue Technologien nicht nur zu einer stärkeren Auslastung der Netze führen, sondern auch eine deutlich dynamischere Last- und Einspeisesituation erzeugen [1]. Dies erschwert eine vorausschauende Netzplanung und stellt hohe Anforderungen an die Betriebsführung. Dadurch entsteht ein zunehmender Bedarf an Flexibilität im Netzbetrieb, um auf kurzfristige Belastungsspitzen reagieren und Netzengpässe vermeiden zu können [1, 2, 3]. Diese Flexibilität beschreibt die Fähigkeit, Stromverbrauch oder -erzeugung kurzfristig an die aktuelle Situation im Netz anzupassen [4]. Sie wird zu einem entscheidenden Instrument, um Netzengpässe zu vermeiden, die Netzstabilität zu sichern und die Integration volatiler erneuerbarer Energien effizient zu ermöglichen [3].

Im herkömmlichen Stromnetz wurde Flexibilität vor allem durch große zentrale Kraftwerke bereitgestellt, die ihre Einspeisung bei Bedarf erhöhen oder verringern konnten [5]. Mit dem Wandel zu einem dezentralen Energiesystem verlagert sich diese Flexibilität zunehmend auf die Niederspannungsebene und wird direkt an die Verbraucher adressiert [3]. Flexibilität wird heute auf allen Spannungsebenen benötigt. Auf Systemebene unterstützt sie die Integration volatiler erneuerbarer Energien und die Aufrechterhaltung der Netzfrequenz [5]. In

Verteilnetzen, insbesondere in der Niederspannung, rückt hingegen die lokale Netzstabilität in den Fokus. Hier können flexible Verbrauchseinrichtungen, wie Wärmepumpen, Elektrofahrzeuge oder Batteriespeicher kurzfristig gesteuert werden, um ihren Verbrauch zeitlich zu verschieben oder gezielt zu reduzieren [2].

Ziel dieser Flexibilitätsnutzung ist es, lokale Netzengpässe zu vermeiden, Lastspitzen zu verringern und die vorhandene Netzkapazität optimal auszunutzen [3]. Im Zuge der aktuellen regulatorischen Entwicklungen, insbesondere mit dem §14a Energiewirtschaftsgesetz (EnWG), erhält die Nutzung dieser Flexibilität einen rechtlichen Rahmen [6]. Netzbetreiber haben die Möglichkeit, auf dezentrale steuerbare Verbrauchseinrichtungen zuzugreifen und diese im Bedarfsfall netzdienlich zu steuern. Diese Maßnahme ist überwiegend reaktiv ausgerichtet, da sie erst bei drohenden Netzengpässen greift. Ziel ist es jedoch, Engpässe im Netz proaktiv und nicht reaktiv zu verhindern, ohne teure Netzausbauten sofort realisieren zu müssen [7]. Flexibilität wird damit zur virtuellen Netzerweiterung, um vorhandene Infrastrukturen besser auszunutzen.

Ein Ansatz in diesem Kontext ist die Koordinierungsfunktion (KOF), welche im Projekt Koordinierungsfunktion des Verteilnetzes und Lastmanagement des elektrifizierten Personennahverkehrs (KOLA) entwickelt wird [8]. Diese Funktion dient dazu, flexible Stromverbräuche so zu steuern, dass sie im Einklang mit den Kapazitäten des Stromnetzes stehen [8]. Die KOF ermöglicht es, Verbrauchsvorgänge präventiv zu verschieben [8]. Dadurch können Verbraucher ihren Stromverbrauch gezielt an ökologischen Kriterien, wie der Nutzung regenerativen Stroms, oder an ökonomischen Aspekten, wie günstige Strompreise ausrichten, ohne dabei die Stabilität des Stromnetzes zu gefährden [8]. Diese Umgebung übernimmt dabei eine koordinierende Rolle auf Verteilnetzebene. Sie sorgt dafür, dass flexible Stromverbräuche netzverträglich eingesetzt werden und trägt so dazu bei, den Ausbau des Stromnetzes zu minimieren und gleichzeitig die Integration erneuerbarer Energien zu fördern [8]. In dieser Arbeit verfolgt die KOF das Ziel, die verfügbare Flexibilität von Haushalten mit steuerbaren Verbrauchseinrichtungen so zu steuern, dass Netzengpässe bereits im Vorfeld vermieden werden können [2]. Die Datengrundlage der KOF bilden die von den Haushalten übermittelten Day-Ahead-Fahrpläne, welche bei Bedarf angepasst werden.

In der Praxis ist davon auszugehen, dass sich nicht alle Haushalte an der Übersendung von entsprechenden Fahrplänen beteiligen. Zum einen steht die benötigte Datenbasis in vielen Verteilnetzbereichen nur eingeschränkt zur Verfügung, da flächendeckende intelligente Messsysteme noch im Aufbau sind [9]. Zum anderen besteht im aktuellen regulatorischen Rahmen keine Verpflichtung zur Fahrplananmeldung [10], was unter anderem auf das Fehlen geeigneter Kommunikations- und Steuerungstechnologien in der bestehenden Infrastruktur zurückzuführen ist [7]. Diese lückenhafte Datenbasis stellt die Lastprognose auf der Ebene von Niederspannungsabgängen vor erhebliche Herausforderungen [11]. Die Identifikation und gezielte Steuerung von Flexibilitätspotenzialen durch die KOF wird erschwert, da ungemeldete Haushalte mit pauschalen oder statistischen Lastprofilen geschätzt werden müssen. Dadurch steigt die Unsicherheit in der Netzzustandsprognose und das Risiko, dass entweder zu viel oder zu wenig Flexibilität bereitgestellt wird, was die Effizienz und Sicherheit des Netzes beeinträchtigen kann.

1.1 Ziel der Arbeit

Diese Arbeit setzt an dieser Problematik an und entwickelt einen Lösungsansatz, der trotz unvollständiger Fahrplanmeldungen eine möglichst vollständige und belastbare Datengrundlage für die Koordinierungsfunktion schafft. Daraus ergibt sich folgende Forschungsfrage:

Forschungsfrage 1

Wie kann die Datenlücke von fehlenden Fahrplänen für die Koordinierungsfunktion vervollständigt werden?

Ein Lösungsansatz besteht darin, Haushalte mit ähnlicher technischer Ausstattung und vergleichbarem Verbrauchsverhalten mithilfe von Clustering-Verfahren zu kategorisieren, um daraus typische Lastverläufe zu identifizieren. Diese repräsentativen Lastprofile ermöglichen es der KOF, auch für Haushalte mit unvollständigen Daten eine gute Abschätzung zu erhalten. Auf dieser Basis kann die verfügbare Flexibilität im Netz zuverlässiger abgeschätzt und koordiniert werden. Diese Arbeit berücksichtigt die praktische Notwendigkeit, Haushaltsfahrpläne nicht nur zu gruppieren, sondern aus den erkannten Merkmalen auch

konkret fehlende oder nicht gemessene Lastverläufe zu generieren. Dieser Schritt wurde in bisherigen Studien kaum behandelt, ist für netzdienliche Anwendungen oder Day-Ahead-Prognosen jedoch entscheidend. In diesem Kontext stellt sich die zweite Forschungsfrage:

Forschungsfrage 2

Wie gut können die Fahrplandaten von unbekannten Haushalten mit einem Clustering-Algorithmus prognostiziert werden?

Die Prognose elektrischer Lasten auf der Niederspannungsebene ist geprägt von einem zunehmenden Fokus auf kategoriebasierte Ansätze [12]. Diese Methoden verfolgen das Ziel, den heterogenen Charakter von Verbrauchern, insbesondere flexible Lasten, differenziert abzubilden. Grundlage solcher Prognosen sind reale Smart-Meter-Daten mit einer hohen zeitlichen Auflösung, die in Deutschland bislang nur eingeschränkt verfügbar sind, oder synthetisch generierte Lastprofile [9]. In der praktischen Anwendung stoßen viele clusterbasierte Prognoseverfahren jedoch auf mehrere Einschränkungen, insbesondere im Hinblick auf ihre operative Anschlussfähigkeit. Sie passen sich nur begrenzt an veränderte Netzzustände oder neue Verbraucher an und nutzen bereitgestellte Haushaltsinformationen kaum [13]. Da sie bekannte Verbrauchsmuster auf ähnliche, ungemessene Haushalte übertragen, reagieren clusterbasierte Ansätze zudem nur eingeschränkt auf dynamische Entwicklungen. Vor diesem Hintergrund stellt sich die dritte Forschungsfrage:

Forschungsfrage 3

Wie gut lässt sich die Genauigkeit der clusterbasierten Lastprognose durch die Einbindung zusätzlicher Netzdaten verbessern?

Darauf aufbauend wird in dieser Arbeit ein Prognosealgorithmus entwickelt, der Clusterinformationen mit weiteren verfügbaren Datenquellen kombiniert, um eine genauere Lastvorhersage zu ermöglichen. Neben netzseitigen Messdaten werden auch Wetter- und Preisdaten einbezogen. Letztere sind zwar bereits in den von Home Energy Management System (HEMS) gesteuerten Fahrplänen enthalten, liefern jedoch auch für die Prognose wertvolle Zusatzinformationen. In Kombination mit historischen Lastdaten ermöglichen diese Quellen dem Modell, zeitliche und saisonale Muster im Verbrauchsverhalten zu erkennen

und insbesondere in Bereichen ohne eigene Fahrplananmeldung verborgene Zusammenhänge zwischen äußeren Einflussfaktoren und aggregierten Lastverläufen zu identifizieren. Der Ansatz bietet insbesondere für Verteilnetzbetreiber eine potenziell skalierbare Lösung, auch unter eingeschränkter Datenverfügbarkeit oder begrenzter Kommunikationsinfrastruktur. Abschließend wird die Güte der Netzprognose und die Anwendbarkeit im Rahmen der Koordinierung mit der vierten Forschungsfrage bewertet:

Forschungsfrage 4

Wie gut unterstützt die entwickelte Lastprognose die Koordinierungsfunktion im Verteilnetz und reduziert Engpässe im Vergleich zu einer idealen Prognose?

Dabei wird analysiert, ob die Prognose einen Beitrag zur frühzeitigen Erkennung und Vermeidung von Engpässen leisten kann, indem sie Anpassungen bestehender Fahrpläne ermöglicht. Die Bewertung erfolgt im Vergleich zu einer idealisierten, perfekten Prognose, um den praktischen Nutzen und die Leistungsfähigkeit des Ansatzes zu bewerten.

1.2 Aufbau der Arbeit

Der Aufbau dieser Arbeit orientiert sich an den formulierten Forschungsfragen und führt schrittweise von den theoretischen Grundlagen bis zur Bewertung der entwickelten Prognosemethoden. Das Kapitel 2 beschreibt die relevanten Entwicklungen im Verteilnetz, einschließlich der zunehmenden Dezentralisierung und der Bedeutung der Koordinierungsfunktion im Netzbetrieb. Das Kapitel 3 behandelt die methodischen Grundlagen und gibt einen Überblick über den aktuellen Stand der Technik im Bereich der Lastprognose, insbesondere zu kategoriebasierten Ansätzen. In Kapitel 4 wird die methodische Vorgehensweise dieser Arbeit erläutert, einschließlich der verwendeten Datengrundlage und der Modellierung der entwickelten Prognosemethoden. Das Kapitel 5 präsentiert und diskutiert die erzielten Ergebnisse im Hinblick auf die Forschungsfragen und deren Beitrag zur Verbesserung der Netzkoordination. Das Kapitel 6 schließt die Arbeit mit einer Zusammenfassung der zentralen Erkenntnisse sowie einem Ausblick auf mögliche Weiterentwicklungen ab.

2 Relevante Entwicklungen des Verteilnetzes

In den folgenden Abschnitten werden die relevanten Entwicklungen im Verteilnetz vorgestellt. Zunächst wird in Kapitel 2.1 die Entwicklung des Stromnetzes beschrieben. Anschließend folgt in Kapitel 2.2 das Netzengpassmanagement, das sich mit Strategien und Maßnahmen zur Vermeidung von Überlastungen im Stromnetz befasst. Darauf aufbauend wird in Kapitel 2.3 die Koordinierungsfunktion betrachtet, die eine effiziente sowie nachhaltige Nutzung der Energieinfrastruktur ermöglichen soll. Abschließend wird in Kapitel 2.4 das Home Energy Management System (HEMS) vorgestellt, welches den Energieverbrauch in Haushalten optimiert.

2.1 Entwicklung des Stromnetzes

Seit zwei Jahrzehnten befindet sich das Stromnetz in einem immer schneller werdenden Wandel. Während das traditionelle Stromnetz auf zentraler Stromerzeugung basierte, bei dem große Kraftwerke die Leistung über Hoch- und Mittelspannungsleitungen bis hin zu den Verbrauchern im Niederspannungsnetz transportierten, ist das heutige Stromnetz zunehmend komplexer [14]. Getrieben durch die Energiewende, den Klimaschutz und technologische Innovationen hat die Dezentralisierung der Stromerzeugung massiv zugenommen. Photovoltaikanlagen, Wärmepumpen, Batteriespeicher und zunehmend auch Elektromobilität führen dazu, dass Haushalte nicht mehr nur passive Verbraucher, sondern zunehmend auch aktive Teilnehmer am Energiesystem sind [15]. Diese Entwicklung hat weitreichende Folgen für die Struktur und den Betrieb der Stromnetze, insbesondere auf der Niederspannungsebene, da diese die Schnittstelle zwischen den höheren Verteilnetzebenen bzw. dem

Übertragungsnetz und den Endverbrauchern bildet. Durch die wachsende Anzahl dezentraler Erzeugungsanlagen wird Leistung nicht mehr unidirektional zu den Verbrauchern, sondern zusätzlich bidirektional von den Verbrauchern zurück in das Netz gespeist [14]. Das ursprüngliche passive Verteilnetz wird zu einem aktiven System, das Last- und Erzeugungsspitzen flexibel ausgleichen muss. Dies bringt zahlreiche technische und organisatorische Herausforderungen mit sich.

Eines der zentralen Ziele besteht darin, die Netzstabilität aufrechtzuerhalten. Das Stromnetz muss zu jeder Zeit im Gleichgewicht zwischen Erzeugung und Verbrauch gehalten werden [5]. Diese Aufgabe wurde primär durch zentrale Großkraftwerke und Netzbetreiber sichergestellt. Heute schwanken Einspeisung und Verbrauch jedoch auf der Niederspannungsebene stärker, da viele dezentrale Anlagen, wie PV-Anlagen oder Wärmepumpen, witterungsabhängig arbeiten [15]. Dies kann zu lokalen Leistungsspitzen führen, bei denen zeitweise mehr Energie in einen Netzabschnitt eingespeist wird, als dort verbraucht oder weitertransportiert werden kann. Diese Überproduktion kann zu einer Überschreitung der zulässigen Spannungsgrenzen führen und das Risiko von Netzüberlastungen oder sogar Netzabschaltungen erhöhen [5]. Umgekehrt kann bei einer schwachen Einspeisung aus erneuerbaren Quellen und gleichzeitig einer hohen Nachfrage der Leistungsbedarf nicht gedeckt werden, wodurch ein erhöhter Leistungsfluss im Netz erforderlich wird [5]. Das Netz ist dadurch zunehmend Schwankungen und Unsicherheiten ausgesetzt, was die Aufrechterhaltung von Netzstabilität und Versorgungssicherheit erschwert.

Eine weitere Herausforderung liegt in den Kapazitätsgrenzen vieler bestehender Niederspannungsnetze. Diese wurden ursprünglich weder für die Einspeisung dezentraler Erzeuger noch für die heutige Menge und Größe elektrischer Lasten ausgelegt [5]. Abhängig vom jeweiligen Netzgebiet können daher in beide Richtungen, sowohl bei der Einspeisung als auch beim Verbrauch, Engpässe auftreten. Insbesondere in Regionen mit einem hohen Anteil an erneuerbaren Energien kann es vorkommen, dass die Netze aufgrund von günstigen Wetterlagen mit dezentral eingespeister Energie überlastet werden, was zu Netzengpässen und im Extremfall zu Netzabschaltungen führen kann [5].

2.2 Netzengpassmanagement

Ein Netzengpass im Stromnetz bezeichnet das tatsächliche oder drohende Überschreiten der Kapazität eines Netzbetriebsmittels, beziehungsweise die Gefahr, dass technische Grenzwerte nicht mehr eingehalten werden. Im Niederspannungsnetz kann ein Engpass einerseits strombedingt auftreten, wenn die maximale Wirkleistungskapazität eines Betriebsmittels überschritten wird, oder andererseits spannungsbedingt, wenn die Spannung außerhalb der vorgeschriebenen Toleranzbereiche liegt. Besonders relevant ist, dass der Betrieb des Stromnetzes aus Sicherheitsgründen nach dem sogenannten (n-1)-Prinzip erfolgt. Dieses Kriterium schreibt vor, dass das Netz so robust ausgelegt sein muss, dass auch beim Ausfall eines einzelnen Betriebsmittels die Versorgungssicherheit weiterhin gewährleistet bleibt. Deshalb nutzen Netzbetreiber im Normalbetrieb ihre Anlagen nie bis zur rechnerischen Maximalgrenze aus, sondern halten bewusst Reserven vor. [5]

Das Engpassmanagement, also die Bewältigung solcher Netzengpässe, lässt sich in langfristige und kurzfristige Maßnahmen unterteilen. Die Abbildung 2.1 zeigt schematisch die Klassifizierung dieser Maßnahmen. Langfristig gehört dazu vor allem der Ausbau oder Umbau der Netzinfrastruktur, zum Beispiel durch Verstärkung von Leitungen oder Austausch von Transformatoren mit höherer Nennleistung. Kurzfristige Engpassmanagement-Maßnahmen werden noch einmal unterteilt in präventive und kurative Ansätze. Präventive Methoden greifen bereits vor der Entstehung eines Engpasses und teilen zum Beispiel über Marktsysteme wie Auktionen die begrenzte Kapazität unter den Marktteilnehmern auf. Kurative Maßnahmen hingegen setzen erst dann ein, wenn ein Engpass tatsächlich entstanden oder unmittelbar bevorsteht. [16]

Eine wichtige Maßnahme in diesem Kontext ist der Redispatch, der sowohl präventiv als auch kurativ eingesetzt werden kann. Dabei werden die Einspeiseleistungen bestimmter Kraftwerke gezielt verändert, um die Stromflüsse im Netz so umzuleiten, dass überlastete Netzabschnitte entlastet werden [17]. Seit der Einführung von Redispatch 2.0 im Jahr 2021 sind nicht mehr nur große konventionelle Kraftwerke, sondern auch kleinere, dezentrale Anlagen, insbesondere aus dem Bereich der erneuerbaren Energien, verpflichtet, an diesem Mechanismus teilzunehmen [17]. Der Netzbetreiber weist dabei betroffenen Anlagen eine neue Einspeiseleistung zu und gleicht die Differenz finanziell aus [18]. Ein solches Einspei-

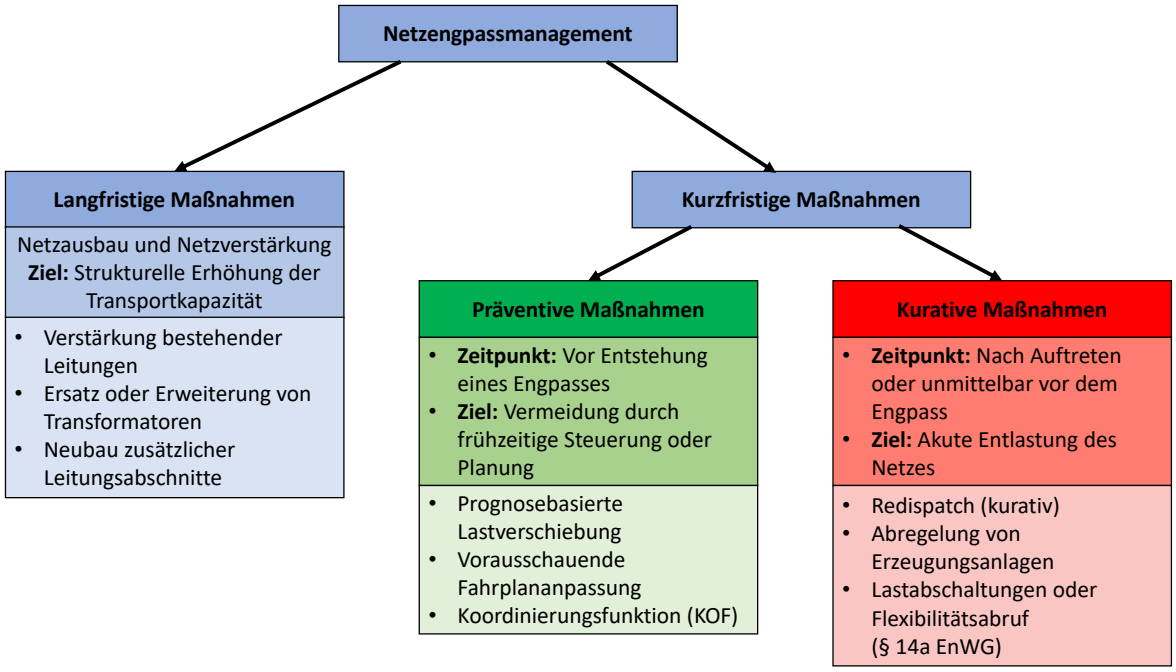


Abbildung 2.1: Klassifizierung von Engpassmanagement-Maßnahmen

semanagement regelt also Einspeiser im Bedarfsfall ab und begrenzt die Leistungsbereitstellung, um das Netz zu entlasten. Diese Maßnahme ist jedoch als kritisch zu betrachten, da sie einerseits dem Ziel der Dekarbonisierung widerspricht und andererseits hohe Kosten verursacht. Anlagenbetreiber müssen entschädigt werden, obwohl der erzeugte Strom nicht genutzt wird, was zu einer ineffizienten Systemführung führt. In extremen Situationen kann es darüber hinaus zu Netztrennungen oder kontrollierten Lastabwürfen kommen, bei denen einzelne Netzteile vom Gesamtsystem getrennt oder Verbraucher vom Netz genommen werden, um einen Blackout zu verhindern. [5]

Die Maßnahmen des Redispatch finden zwar überwiegend im Übertragungsnetz und damit auf den hohen Spannungsebenen statt, dennoch sind diese Engpässe eng mit den Entwicklungen in den unteren Netzebenen verknüpft. Mit dem wachsenden Anteil dezentraler Erzeugung und dem zunehmenden Anschluss steuerbarer Verbraucher, etwa Wärmepumpen oder Ladeeinrichtungen, verschiebt sich die Entstehung von Engpässen zunehmend in die unteren Spannungsebenen. Dadurch wird deutlich, dass Flexibilität nicht nur auf der Erzeugungsseite, sondern auch auf der Verbraucherseite eine zentrale Rolle spielt. Insbesondere durch die zunehmende Anzahl dezentraler, steuerbarer Anlagen und Verbraucher auf der Niederspannungsebene bieten sich neue Möglichkeiten, Engpässe intelligenter, fle-

xibler und kostengünstiger zu bewältigen [3]. Statt zunehmend mehr Erzeugungsmöglichkeiten kostenpflichtig abregeln zu müssen, kann Flexibilität auf der Verbraucherseite genutzt werden, um Lasten gezielt zu verschieben oder temporär zu begrenzen. So lassen sich Netzüberlastungen bereits im Vorfeld auf der Niederspannungsebene entschärfen.

Genau an diesem Punkt setzt § 14a des EnWG an [6]. Es bildet die gesetzliche Grundlage dafür, dass Netzbetreiber steuerbare Verbrauchseinrichtungen, wie Wärmepumpen, Ladeeinrichtungen für Elektrofahrzeuge oder Batteriespeicher, im Bedarfsfall gezielt steuern dürfen [6]. Ziel ist es, durch die zeitweise Reduzierung oder Verschiebung des Stromverbrauchs in Haushalten Engpässe im Verteilnetz zu vermeiden und das Netz effizienter auszulasten. § 14a EnWG ergänzt damit die bestehenden Maßnahmen wie Redispatch und Einspeisemanagement um eine kurative, dezentrale Komponente. Schaltmaßnahmen dürfen also ausschließlich dann ergriffen werden, wenn ein Engpass unmittelbar bevorsteht. Es werden steuerbare Verbraucher aktiv in das Engpassmanagement einbezogen, bevor kostenintensivere zentrale Maßnahmen auf der Erzeugerseite notwendig werden [6].

Für die Nutzer dieser steuerbaren Verbrauchseinrichtungen bringt § 14a EnWG Vorteile, denn es werden Anreize geschaffen, Flexibilität im Sinne des Gesamtsystems zur Verfügung zu stellen [3]. Nutzer erhalten reduzierte Netzentgelte oder andere Vergünstigungen als Ausgleich für ihre Bereitschaft zur Steuerung, um diese Flexibilität flächendeckend im Verbundnetz zu manifestieren [3].

2.3 Koordinierungsfunktion

Netzengpässe entstehen nicht nur durch strukturelle Defizite, sondern auch durch eine unkoordinierte Nutzung dezentraler Erzeugungs- und Verbrauchseinheiten. Um solche Situationen bereits im Vorfeld zu vermeiden, ist eine übergeordnete Instanz notwendig, die dezentrale Flexibilitäten systematisch zusammenführt und netzdienlich koordiniert. Genau hier setzt die Koordinierungsfunktion (KOF) an, welche parallel zu den Entwicklungen um § 14a im Rahmen des Forschungsprojektes KOLA entwickelt wurde [8]. Während § 14a kurativ auf akute Engpässe reagiert, verfolgt die KOF einen präventiven Ansatz, der versucht den Netzbetrieb vorausschauend innerhalb seiner technischen Grenzen zu halten [2]. Sie verfolgt das Ziel, Netzengpässe präventiv zu vermeiden, indem sie die von den HEMS übermittel-

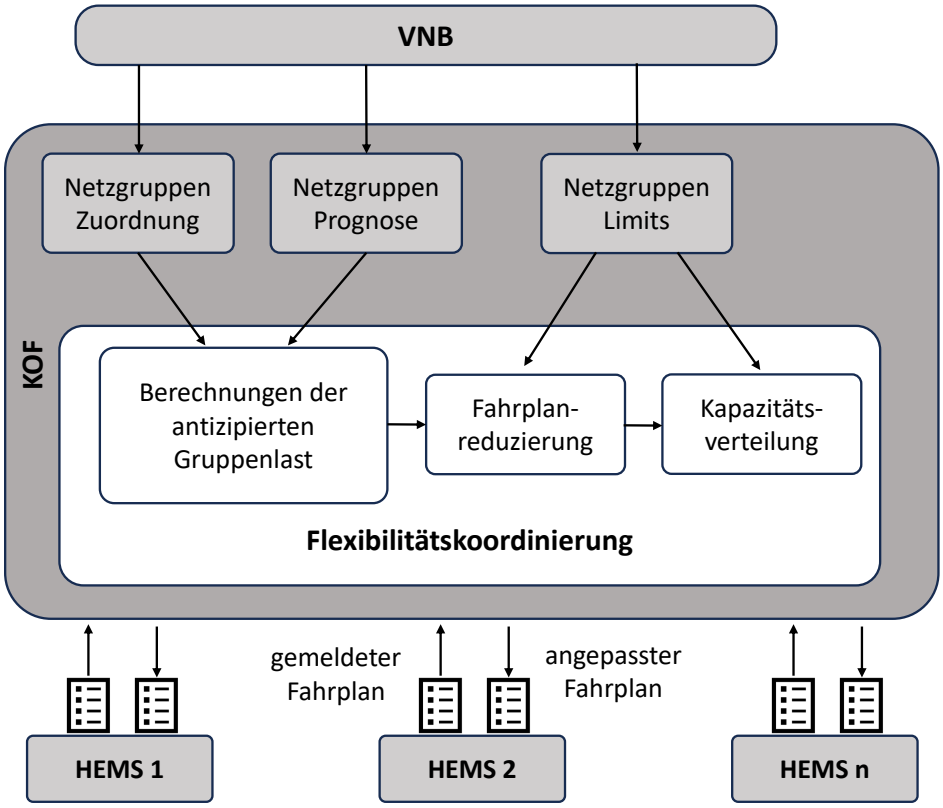


Abbildung 2.2: Vorgeschlagener Rahmen für die Koordinierungsfunktion aus [2]

ten Fahrpläne aggregiert, mit den technischen Netzgrenzen abgleicht und bei drohenden Überlastungen entsprechende Anpassungen vornimmt.

Die in Abbildung 2.2 dargestellte Grafik zeigt die vorgeschlagene Struktur der Koordinierungsfunktion zur präventiven Steuerung von Flexibilitäten im elektrischen Verteilnetz. Die KOF arbeitet als zentrale Instanz im Verteilnetz. Sie empfängt von allen angeschlossenen Kunden, die über HEMS verfügen, die geplanten Leistungsverbräuche und Einspeisungen für den kommenden Tag, welche in 15-Minuten-Intervallen aufgeschlüsselt sind [2]. Ein HEMS übernimmt dabei im Haushalt die zentrale Rolle der intelligenten Steuerung und Koordination sämtlicher energierelevanter Geräte und Anlagen. Die Gesamtheit der Informationen werden von der KOF aggregiert und einer Netzgruppenstruktur zugeordnet, die sich an den verschiedenen Spannungsebenen und Netzanschlusspunkten orientiert.

Die KOF nutzt diese Fahrpläne zusammen mit zusätzlichen Informationen, die vom Verteilnetzbetreiber (VNB) bereitgestellt werden. Wie in Abbildung 2.2 abgebildet ist, umfassen diese Informationen die Netzgruppenzuordnung der einzelnen Haushalte, Pro-

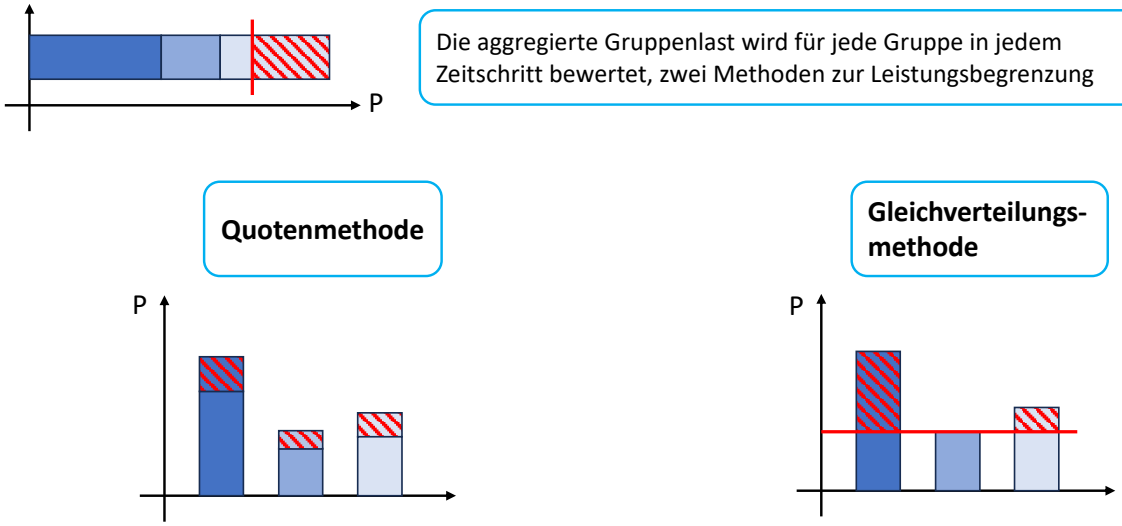


Abbildung 2.3: Möglichkeiten der Leistungsbegrenzung durch die KOF

gnosen zur erwarteten Netzbelastung sowie technische Kapazitätsgrenzen der jeweiligen Netzgruppen, etwa durch Leitungskapazitäten oder Transformatorengrenzen.

Auf Basis dieser Planungsdaten erstellt die KOF eine zu erwartenden Gruppenlast für jede Netzgruppe. Diese Prognosen werden anschließend mit den jeweiligen Kapazitätsgrenzen, also den maximal zulässigen Stromflüssen auf Leitungen oder Transformatoren, verglichen. Zeigen sich dabei Zeiträume, in denen die prognostizierte Last die zulässigen Grenzen übersteigt, greift die KOF regulierend ein und initiiert eine Reduktion der Lastpläne. Sie passt die zulässigen Leistungsgrenzen der flexiblen Verbraucher so an, dass die vorgegebenen Netzgrenzen eingehalten werden. Die konkrete Anpassung, wie etwa die zeitliche Verschiebung von Ladevorgängen oder die Reduktion von Leistungen, erfolgt anschließend durch die HEMS auf Basis der von der KOF vorgegebenen Grenzwerte. Es gibt unterschiedliche Methoden, mit denen diese Anpassungen vorgenommen werden können. [2]

Zwei der im Forschungsprojekt untersuchten Ansätze sind die quoten-basierte Methode sowie die gleichverteilende Methode, welche in Abbildung 2.3 dargestellt werden. Die Quotenmethode basiert auf einer proportionalen Lastverteilung. Ziel dieses Verfahrens ist es, die notwendige Reduktion proportional zu den ursprünglich geplanten Verbrauch der einzelnen Kunden zu verteilen [2]. Somit wird die maximal zulässige Leistung pro Gruppe anteilig auf die einzelnen flexiblen Lasten verteilt, gewichtet nach ihrem Anteil an der gesamten flexiblen Last der Gruppe.

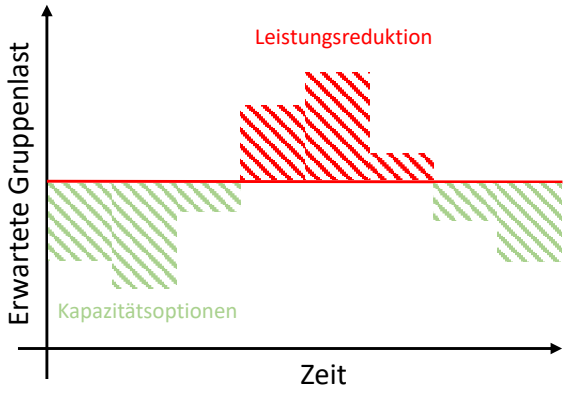


Abbildung 2.4: Zeiträume von Netzkapazitäten

Im Gegensatz dazu teilt die Gleichverteilungsmethode die verbleibende Netzkapazität gleichmäßig unter allen Teilnehmern auf, unabhängig von deren ursprünglichem Verbrauch [2]. Somit wird die im Netz verfügbare Kapazität während einer Überlastphase gleichmäßig unter allen Kunden aufgeteilt, die in diesem Zeitintervall flexible Leistung geplant hatten, unabhängig von der Höhe ihres ursprünglichen Leistungsbedarfs [2].

Beide Ansätze haben spezifische Vor- und Nachteile hinsichtlich Fairness, Effizienz und Kundenzufriedenheit, und je nach Netzsituation kann eine Kombination beider Methoden sinnvoll sein. Ein wesentlicher Aspekt der KOF ist jedoch nicht nur die Reduktion von Lasten in kritischen Zeiträumen, sondern auch die aktive Umverteilung von ungenutzten Netzkapazitäten. Wenn das Netz durch äußere Bedingungen weniger belastet ist, kann die KOF diese Kapazitäten gezielt den Kunden anbieten [2]. In Abbildung 2.4 werden Zeiträume von Netzkapazitäten beispielhaft dargestellt. Dadurch wird deutlich, dass verschobene Verbrauchs- oder Ladevorgänge nachgeholt werden können, was den Komfort für die Kunden erhält und die Netzauslastung insgesamt optimiert.

Um freie Netzkapazitäten möglichst fair und effizienter unter den Gruppen zu verteilen, zeigt die Abbildung 2.5 ebenfalls zwei Methoden zur zusätzlichen Leistungszuteilung, die in jedem Zeitschritt angewendet werden können [2]. Die zuvor beschriebenen Methoden zur Leistungsbegrenzung lassen sich in gleicher Weise auch auf die Zuteilung freier Netzkapazitäten anwenden.

Die Notwendigkeit einer solchen KOF ergibt sich daraus dass ohne eine übergeordnete Steuerung flexible Verbraucher häufig synchron auf Preissignale reagieren würden [2, 3].

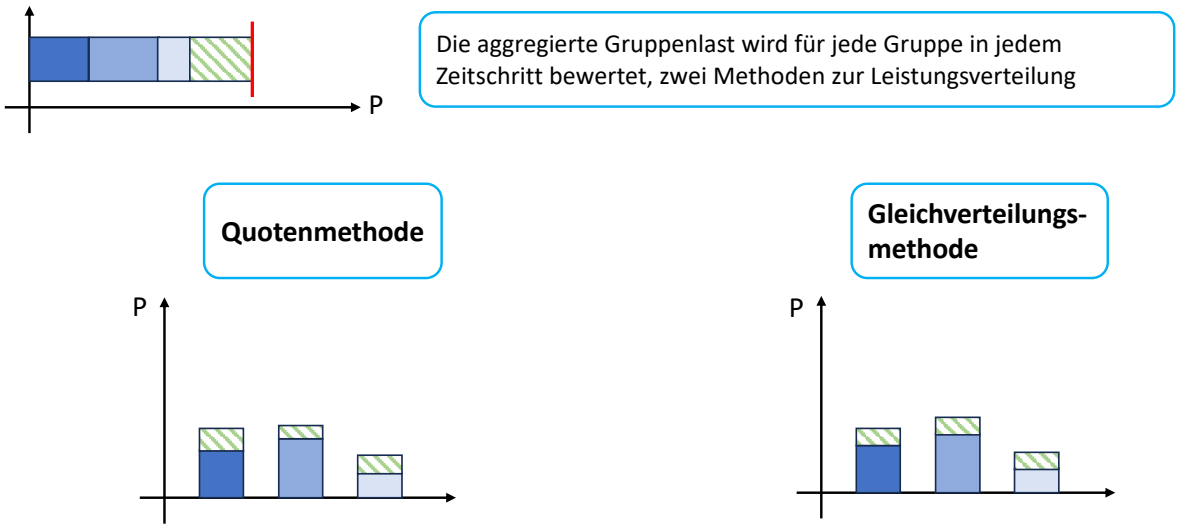


Abbildung 2.5: Möglichkeiten der Kapazitätsverteilung durch die KOF

Dies könnte beispielsweise zur Folge habe, dass viele Elektroautos zeitgleich zu günstigen Preisen laden wollen und dadurch neue Lastspitzen entstehen, die das Netz schnell an seine Grenzen bringen. Marktbasierte Lösungen, wie etwa dynamische Tarife, sind zwar theoretisch denkbar, jedoch in der Praxis komplex umzusetzen und stoßen oft an regulatorische oder Akzeptanzgrenzen [3]. Direktsteuerungen durch den Netzbetreiber sind hingegen technisch einfacher, können aber die Kundenakzeptanz beeinträchtigen und sind nicht immer ausreichend effektiv. Die KOF stellt ein Verfahren dar, mit dem Netzsicherheit, Effizienz und Kundeninteressen berücksichtigt werden. Sie ermöglicht den stabilen Betrieb des Stromnetzes auch bei einem steigenden Anteil flexibler und dezentraler Akteure, ohne dass ein umfangreicher Netzausbau notwendig wird. Durch die KOF kann das vorhandene Flexibilitätspotenzial genutzt und die Eingriffe bei den Kunden gleichzeitig auf ein Minimum beschränkt werden [3].

In der Realität ist allerdings davon auszugehen, dass eine nicht unerhebliche Zahl an Haushalten keine oder nicht regelmäßig ihre Fahrplandaten an die Koordinierungsfunktion meldet. Ein wesentlicher Faktor für solche Datenlücken ist die fehlende technische Infrastruktur [9]. Nicht jeder Haushalt ist zukünftig mit einem modernen HEMS oder anderen kompatiblen Geräten ausgestattet, die eine kontinuierliche und automatisierte Fahrplanmeldung ermöglichen. Für viele bedeutet die Nachrüstung einen zusätzlichen finanziellen und organisatorischen Aufwand, der sich besonders dann nicht lohnt, wenn das Flexibilitätspotenzial des eigenen Haushalts als gering eingeschätzt wird [19]. Hinzu kommen Bedenken beim Da-

tenschutz und der Privatsphäre [19]. Die detaillierte Übermittlung von Energiedaten lässt oft Rückschlüsse auf das alltägliche Verhalten und die Lebensgewohnheiten der Bewohner zu. Viele Verbraucher sehen darin einen erheblichen Eingriff in ihre Privatsphäre und sind unsicher, wie mit ihren Daten umgegangen wird und wer letztlich Zugriff darauf hat. Ein weiterer Punkt ist der wirtschaftlichen Nutzen für die Verbraucher. Die potentiellen finanziellen Anreize, die für die Teilnahme angeboten werden, sind mit einem hohen Aufwand verbunden, der sich aufgrund der hohen Kosten nicht rechnet. Außerdem besteht die Notwendigkeit, sich mit den zusätzlichen Anwendungen auseinanderzusetzen, was zu einer gewissen Überforderung oder Desinteresse führen kann, insbesondere, wenn der persönliche Nutzen nicht sofort ersichtlich ist [19]. Zudem ist die rechtliche Situation von Bedeutung, da in vielen Regionen keine klare Verpflichtung zur Datenmeldung besteht und häufig unklar ist, wie solche Daten rechtlich geschützt und verwendet werden. Die Teilnahme ist daher freiwillig und oft unverbindlich.

All diese Faktoren könnten dazu führen, dass ein erheblicher Anteil der Haushalte keine oder nur sporadisch Fahrpläne an die KOF übermittelt. Die KOF würde ein unvollständiges Bild über den tatsächlichen und geplanten Energiefluss im Niederspannungsnetz erhalten. Es würden wichtige Informationen über flexible Lasten und Erzeuger fehlen, die für die präzise Prognose, das rechtzeitige Erkennen von Lastspitzen oder Engpässen und die gezielte Nutzung von Flexibilitätspotenzialen unerlässlich wären. Ohne vollständige Daten kann die KOF ihre Aufgaben, wie das Vermeiden von Überlastungen, die effiziente Integration erneuerbarer Energien und die wirtschaftliche Steuerung der Energieflüsse, nur eingeschränkt erfüllen.

Die fehlende flächendeckende Fahrplanmeldung wirft die Frage nach einer zuverlässigen und effizienten Koordinierung im Stromnetz auf. Da eine direkte und vollständige Erfassung der individuellen Fahrpläne sämtlicher Haushalte in absehbarer Zeit nicht realisierbar ist, ergibt sich die Notwendigkeit alternativer Ansätze zur Schließung der bestehenden Datenlücken und zur Bereitstellung einer hinreichenden Datengrundlage für die KOF. Zu diesem Zweck werden Prognoseverfahren in dieser Arbeit thematisiert und modelliert, um das Verhalten von unbekannten Haushalten möglichst genau abzubilden.

Fehlprognosen sind dabei jedoch besonders problematisch für die KOF, da sie unmittelbar Einfluss auf die Effizienz und Stabilität des Netzbetriebs nehmen. Zu hohe Vorhersagen

führen zu einer Überpufferung im Netzbetrieb. Es wird zusätzliche Leistung oder Flexibilität vorgehalten, um vermeintlich bevorstehende Lastspitzen abzufangen, die in der Praxis gar nicht auftreten. Dadurch werden Netzressourcen ineffizient genutzt und andere Haushalte oder Anlagen unter Umständen unnötig abgeregelt, um eine vermeintliche Überlastung zu vermeiden. Dies kann zu einem künstlich konservativen Fahrplan führen, der die Gesamteffizienz des Energiesystems reduziert. Umgekehrt können zu niedrige Prognosen ebenfalls schwerwiegende Folgen haben. Werden reale Lastspitzen unterschätzt, so können diese nicht rechtzeitig erkannt und kompensiert werden. In der Folge kann es zu lokalen Netzengpässen kommen, da die KOF keine proaktiven Gegenmaßnahmen, wie Lastverschiebung, Zuschaltung von Speichern oder temporäre Leistungsreduktion, einleitet. Das Netz reagiert dann erst im Betrieb auf unerwartete Lastanstiege, was die Stabilität gefährdet und im Extremfall zu Überlastsituationen oder Regelenergieeinsätzen führt. Somit kann eine ungenaue Prognose direkte Auswirkungen auf die Netzsicherheit und Systemeffizienz haben. Für eine zuverlässige und wirtschaftliche Netzkoordination ist daher entscheidend, dass die entwickelte Prognose weder künstliche Lastspitzen erzeugen noch reale Laständerungen unterschätzen. Nur so kann die KOF die verfügbare Flexibilität optimal einsetzen und ein stabiles, vorausschauend gesteuertes Energiesystem gewährleisten.

2.4 Home Energy Management System

Um die erforderlichen Daten für die KOF bereitzustellen und die Flexibilität dezentraler Anlagen nutzbar zu machen, kommen HEMS zum Einsatz. Sie erfassen und optimieren Verbrauch, Erzeugung und Speicher auf Haushaltsebene und schaffen damit die Grundlage für eine koordinierte Integration in das Netzmanagement.

2.4.1 Grundlegender Nutzen eines HEMS

Ein HEMS dient grundsätzlich dazu, Energieflüsse innerhalb eines Haushalts intelligent zu steuern. Es verknüpft Erzeugungsanlagen, Verbraucher und Speicher und stimmt deren Betrieb optimal aufeinander ab.

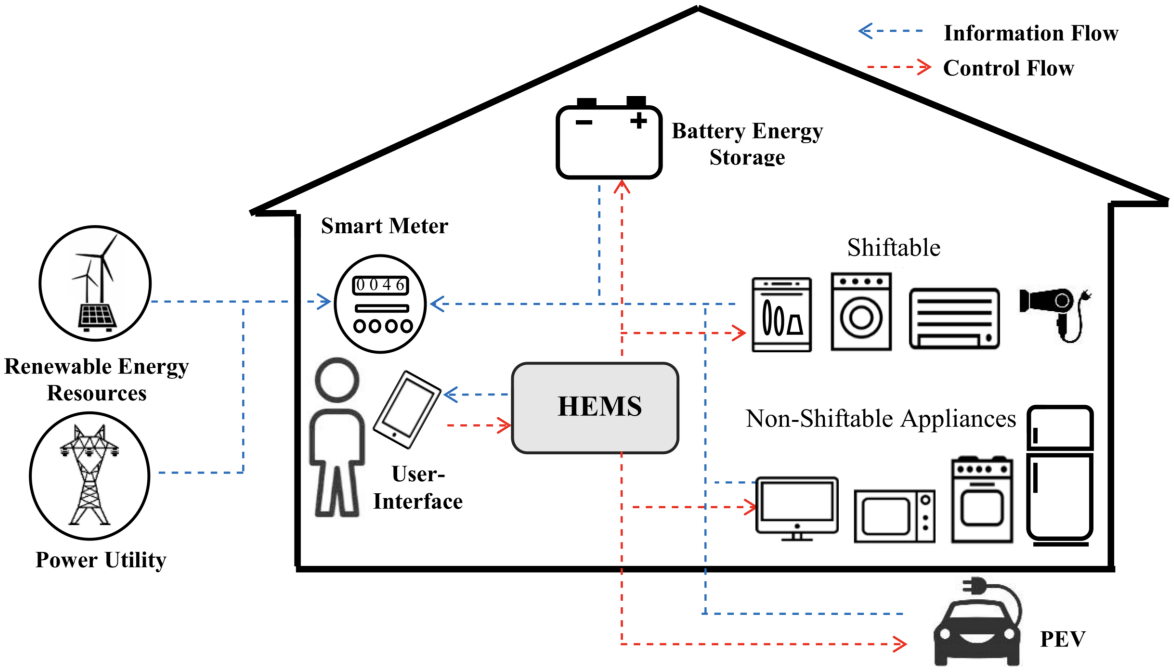


Abbildung 2.6: Aufbau eines Home Energy Management Systems aus [20]

Die Abbildung 2.6 zeigt den Aufbau eines solchen intelligenten Energiemanagementsystems [20]. Ziel des HEMS ist es, Energiequellen, Verbraucher und Speicher intelligent miteinander zu verknüpfen und aufeinander abzustimmen, um Energie effizient zu nutzen und Kosten zu senken [20]. Die Funktionsweise basiert darauf, dass es kontinuierlich Daten über den aktuellen Stromverbrauch, die Eigenstromerzeugung, den Ladezustand von Speichern und Elektrofahrzeugen, Wetterprognosen sowie aktuelle und prognostizierte Strompreise sammelt und verarbeitet. Basierend auf diesen Informationen trifft das HEMS eigenständig Entscheidungen darüber, wann welche Geräte im Haushalt ein- oder ausgeschaltet, Energie gespeichert, verbraucht oder ins öffentliche Netz eingespeist werden sollen. Das HEMS analysiert beispielsweise, wann es sinnvoll ist, das Elektroauto zu laden, den Batteriespeicher aufzufüllen oder die Wärmepumpe für Warmwasser oder Heizung einzusetzen. Hierbei berücksichtigt es nicht nur die Bedürfnisse und Präferenzen der Bewohner, sondern auch externe Faktoren wie Strompreis-Signale oder Wetterprognosen, die für die Solarstromerzeugung relevant sind. Ziel der HEMS-Optimierung ist die Minimierung der Betriebskosten, wozu stündliche Strompreise und Einspeisetarife in die Optimierung einfließen. Gleichzeitig stellt das System sicher, dass Komfortbedingungen wie Mindesttemperaturen im Haushalt oder ein gewünschter Ladezustand von Elektrofahrzeugen eingehalten werden. [2]

2.4.2 Einsatz und Funktion des HEMS in dieser Arbeit

Im Rahmen dieser Arbeit dienen HEMS als Grundlage, um Fahrpläne für eine vorausschauende Koordinierung flexibler Verbraucher und Erzeuger zu erstellen. Dabei werden feste Lastprofile für nicht verschiebbare Verbraucher sowie technische Parameter flexibler Anlagen wie Photovoltaik mit Batteriespeicher, Wärmepumpen und Elektrofahrzeuge berücksichtigt. Die gesendeten Fahrpläne beinhalten sowohl unveränderbare, also inflexible Lasten, wie Beleuchtung oder Grundlastgeräte, als auch flexible Lasten und Erzeuger, wie E-Autos, Wärmepumpen oder PV-Anlagen mit Batteriespeichern.
Charakteristisch für das eingesetzte Modell ist die integrierte Betrachtung von Strom- und Wärmebedarf. Neben den klassischen elektrischen Komponenten wird auch die Wärmeerzeugung über die Wärmepumpe und den Pufferspeicher modelliert, wodurch das Zusammenspiel von Stromverbrauch, Wärmespeicherung und Gebäudethermik erfasst werden kann. Auf diese Weise entstehen standardisierte, reproduzierbare und physikalisch konsistente Tagesfahrpläne, die das Betriebsverhalten moderner Haushalte realitätsnah widerspiegeln und die Datengrundlage dieser Arbeit bilden. [21]

2.4.3 Eingangs- und Ausgangsgrößen des HEMS

Die Berechnung der Tagesfahrpläne basiert auf verschiedenen Eingangsgrößen, die den energetischen Zustand des Haushalts und seiner Umgebung abbilden. Zu den zentralen Eingabedaten zählen die Preissignale für Strombezug und -einspeisung, die entweder konstant oder als variable Preise vorliegen, sowie Wetterdaten des Deutschen Wetterdienstes, aus denen insbesondere Temperatur und Globalstrahlung für die Berechnung der Photovoltaikerzeugung herangezogen werden. Ergänzend werden inflexible Haushaltslasten berücksichtigt, die entweder aus gemessenen Lastprofilen oder aus Prognosen stammen. Für die flexiblen Komponenten wie Elektrofahrzeuge und Wärmepumpe fließen technische Parameter ein, beispielsweise Kapazitäten, Lade- und Entladeleistungen, Wirkungsgrade oder thermische Kennwerte. Optional können zeitvariable Netzleistungsgrenzen einbezogen werden, um lokale Netzrestriktionen zu berücksichtigen. [21]
Auf Basis dieser Eingangsdaten ermittelt das HEMS optimierte Tagesfahrpläne über einen Zeitraum von 24 Stunden mit einer zeitlichen Auflösung von einer Stunde oder 15 Minuten.

Die Optimierung erfolgt als gemischt-ganzzahliges lineares Programm (MILP) über einen 24-Stunden-Horizont und wird mithilfe des Open-Source-Solvers HiGHS oder Gurobi effizient gelöst. Im Mittelpunkt dieser Optimierung steht die Leistungsbilanz des Haushalts, die sicherstellt, dass zu jedem Zeitschritt die Summe aus Photovoltaikerzeugung, Battery Electric Vehicle (BEV) und Wärmepumpenleistung sowie den inflexiblen Lasten der Netzeinspeisung entspricht. Eine Binärvariable verhindert dabei, dass gleichzeitig Strom bezogen und eingespeist wird. Komfortverletzungen, wie ein unzureichender Ladezustand des Elektrofahrzeugs oder Abweichungen der Raumtemperatur vom definierten Komfortbereich, werden durch hohe Strafkosten vermieden, sodass das Modell bestrebt ist, einen realistischen und nutzerfreundlichen Betrieb sicherzustellen. [21]

$$\min \sum_{t \in T} \left(-\lambda_t^{\text{sup}} P_t^{\text{Grid,sup}} + \lambda_t^{\text{dem}} P_t^{\text{Grid,dem}} \right) \Delta t + \lambda_{\text{BEV}} v_{\text{BEV}} + \lambda_{\text{HP}} v_{\text{HP}} + C_{\text{Bat}} \tag{2.1}$$

$$P_t^{\text{infl}} + P_t^{\text{Bat}} + \sum_{n \in N} P_{t,n}^{\text{BEV}} + P_t^{\text{HP}} + P_t^{\text{Grid,sup}} = P_t^{\text{PV}} + P_t^{\text{Grid,dem}} \quad \forall t \in T \tag{2.2}$$

Die Zielfunktion in Formel 2.1 beschreibt die Minimierung der gesamten Energiekosten des Haushalts. Dabei stellen λ_t^{dem} und λ_t^{sup} die zeitabhängigen Strompreise für Netzbezug bzw. Netzeinspeisung dar, $P_t^{\text{Grid,dem}}$ und $P_t^{\text{Grid,sup}}$ die entsprechenden Leistungen für Bezug und Einspeisung, Δt die zeitliche Auflösung, λ_{BEV} und λ_{HP} die Strafkosten für Komfortverletzungen des Elektrofahrzeugs bzw. der Wärmepumpe sowie C_{Bat} die Batterieverschleißkosten. Die Formel 2.2 stellt die Leistungsbilanz des Haushalts dar. Hierbei bezeichnet P_t^{infl} die inflexiblen Lasten, P_t^{Bat} die Leistung der Batterie, $P_{t,n}^{\text{BEV}}$ die Leistung des n-ten Elektrofahrzeugs, P_t^{HP} die Leistung der Wärmepumpe, P_t^{PV} die Leistung der Photovoltaikanlage sowie $P_t^{\text{Grid,dem}}$ und $P_t^{\text{Grid,sup}}$ den Strombezug bzw. die Einspeisung. [21]

In Abbildung 2.7 ist ein beispielhafter, durch das HEMS erstellter Fahrplan dargestellt. Die Grafik zeigt den zeitlichen Verlauf der Gesamtleistung des Haushalts, aufgeteilt in eine inflexible und eine flexible Leistungskomponente. Die inflexible Leistung beschreibt dabei den unveränderbaren Grundlastanteil, der aus kontinuierlich betriebenen Geräten resultiert. Die flexible Leistung hingegen umfasst steuerbare Verbraucher wie Wärmepumpen oder Elektrofahrzeuge und bildet somit den Anteil, der durch das HEMS gezielt verschoben oder angepasst werden kann. Der Verlauf der flexiblen Leistung prägt maßgeblich die Form der

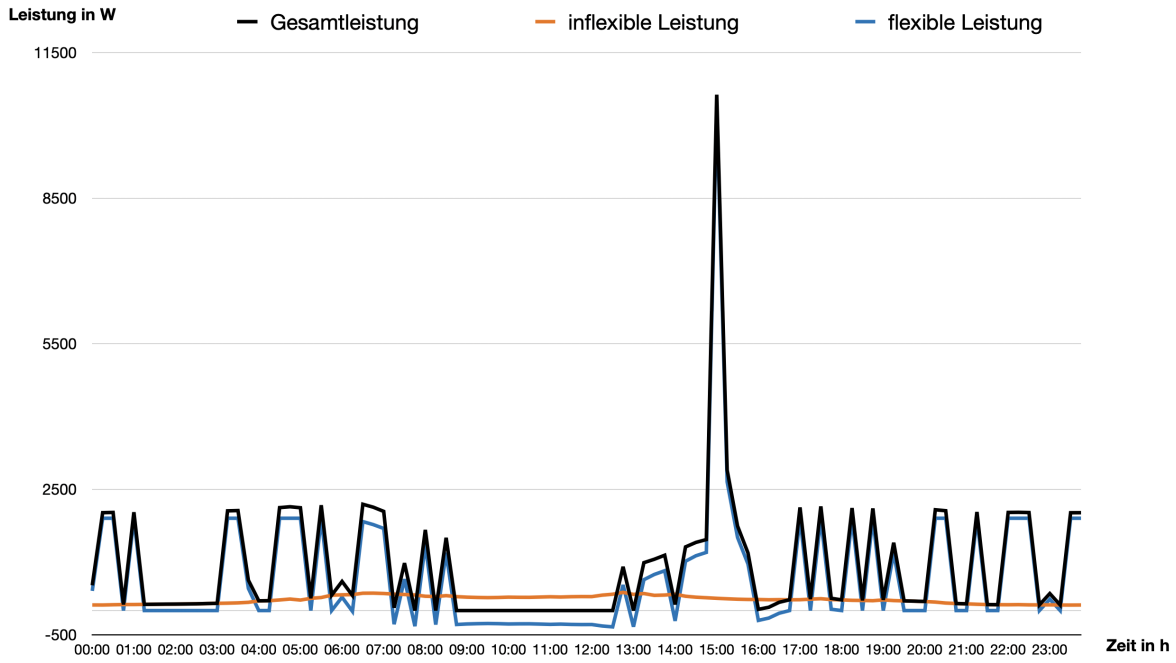


Abbildung 2.7: Simulierter Tagesfahrplan eines Haushalts durch das HEMS

gesamten Lastkurve und zeigt, wie das System auf veränderte Randbedingungen oder Optimierungsziele reagiert. Positive Leistungswerte kennzeichnen den Strombezug aus dem Netz, während negative Werte eine Einspeisung darstellen. Das dargestellte Ergebnis verdeutlicht, wie durch die koordinierte Steuerung der flexiblen Komponenten das elektrische Verhalten des Haushalts gegenüber dem Netz beeinflusst werden kann.

3 Methodische Grundlagen

Im folgenden Abschnitt werden die wesentlichen methodischen Grundlagen dieser Arbeit vorgestellt. Zunächst wird in Kapitel 3.1 ein Überblick über den aktuellen Forschungsstand und relevante Verfahren gegeben. Anschließend folgt in Kapitel 3.2 eine Darstellung gängiger Techniken zur Reduktion hochdimensionaler Datensätze, die zur Vereinfachung der nachfolgenden Clusteralgorithmen dienen, welche in Kapitel 3.3 thematisiert werden. Das Kapitel 3.4 stellt die Leistungsfähigkeit und Anwendungsbereiche von den benannten Clustermethoden gegenüber. In Kapitel 3.5 wird die Qualität der gebildeten Cluster mithilfe geeigneter Validierungsmetriken überprüft. Darauf aufbauend werden in Kapitel 3.6 unterschiedliche Regresionsansätze vorgestellt und in Kapitel 3.7 hinsichtlich ihrer Vorhersagegüte und Robustheit analysiert. Abschließend werden in Kapitel 3.8 Metriken vorgestellt, um die Ergebnisse der Analysen bewerten zu können.

3.1 Stand der Technik

In der aktuellen Forschung zur Analyse und Modellierung von Haushaltsstromverbräuchen stellt das Clustering von Lastprofilen einen zentralen Ansatz dar, um typische Verbrauchsmuster zu identifizieren und diese zur weiteren Verarbeitung nutzbar zu machen.

Clustering

Beim Clustering von Haushaltslastkurven wird in vielen Studien zwischen merkmalsbasierten und verlaufsbasierten Ansätzen unterschieden [22, 23]. Beim merkmalsbasierten Ansatz werden aus den Zeitreihen zunächst charakteristische Kenngrößen extrahiert, etwa Tagesmittelwerte, Spitzenzeiten, Verbrauchsverteilungen zwischen Wochentagen und Wochenenden, um daraus Cluster zu ermitteln [23]. Die Autoren von [23] zeigen, dass sich durch

funktionale Datenanalyse und Hauptkomponentenauswertung aus solchen Kenngrößen eine kompakte und dennoch repräsentative Struktur der Lastverläufe ableiten lässt. Diese Methoden sind besonders effizient in der Rechenzeit und robust gegenüber Messrauschen, da sie extreme Werte glätten und die Dimensionalität stark reduzieren. In der Praxis eignen sie sich daher gut für große Datenmengen oder den Einsatz in Echtzeitanalysen [23].

Demgegenüber steht das direkte Clustering der Lastkurvenverläufe, das ohne vorherige Transformation die vollständigen Zeitreihen nutzt [23]. Dies geschieht entweder direkt über Punkt-zu-Punkt-Distanzen, also euklidisch, oder mittels komplexerer Metriken wie Dynamic Time Warping (DTW), die auch zeitlich versetzte Muster korrekt erfassen können. Die Autoren von [24] zeigen in ihrer Arbeit, dass insbesondere hierarchisches Clustering mit DTW eine hohe Clusterqualität liefert und sich besser für Haushalte mit variierendem Tagesverhalten eignet als klassische K-Means-Verfahren. Verlaufsbasierte Methoden haben jedoch den Nachteil, dass sie im Vergleich zu merkmalbasierten Verfahren eine hohe Rechenkomplexität und schlechtere Skalierbarkeit aufweisen, was sie bei sehr großen Datensätzen einschränkt [23].

Eine vielversprechende Weiterentwicklung stellt die Klasse der hybriden Verfahren dar, die Elemente beider Ansätze miteinander kombiniert. Die Autoren von [25] schlagen ein zweistufiges Verfahren vor, das zunächst ein Clustering über K-Means und anschließend eine Feinstrukturierung über spektrales Clustering durchführt. Auch die Verfasser von [26] implementieren ein zwei-stufiges Clustering, bei dem in der ersten Phase mit einer hohen Clusteranzahl gearbeitet wird, um alle möglichen Muster zu erfassen, während in der zweiten Phase ähnliche Cluster mittels komplexitätsinvarianter DTW zusammengeführt werden. Ziel dieser hybriden Methoden ist es, eine gute Balance zwischen Strukturgenauigkeit und Rechenaufwand zu erreichen. Zunehmend wird auch die Integration von kontextuellen Merkmalen untersucht. Die Autoren in [27] zeigen, dass die Einbeziehung von demografischen Informationen, etwa Haushaltsgröße oder Anzahl steuerbarer Geräte, zu einer signifikanten Verbesserung der Clusterqualität führen kann.

Neben den methodischen Fortschritten zeigen viele Studien jedoch eine inhaltliche Lücke. Clustering wird primär zur Typisierung und Segmentierung eingesetzt, jedoch kaum zur aktiven Rekonstruktion oder Ergänzung unvollständiger Datenbestände. Obwohl Clusterkennungen ein hohes Potenzial bieten, um Haushalte mit fehlenden Fahrplan- oder Verbrauchs-

daten strukturell einzuordnen, wird diese Möglichkeit bislang selten ausgeschöpft [22, 23]. Auch in [28] wird Clustering nur als Vorverarbeitungsschritt für Prognosemodelle genutzt, ohne explizit eine Rückprojektion auf fehlende Daten vorzunehmen. Lediglich die Autoren in [29] nutzen das Clustering, um typische Verbrauchsmuster für individuelle Haushalte abzuleiten. Hier werden repräsentative Clusterkurven gebildet und unbekannte Haushalte anhand ihrer Merkmale einem Cluster zugeordnet [29]. Durch eine anschließende Skalierung auf den individuellen Jahresverbrauch lassen sich für neue Verbraucher Prognosen erstellen, ohne dass für jeden Haushalt ein separates Modell trainiert werden muss [29]. Die Autoren beschränken sich auf Clusterprognosen, bei denen repräsentative Lastprofile gebildet und anschließend skaliert werden [29]. In dieser Arbeit wird dieser Ansatz aufgegriffen, jedoch weiterentwickelt, indem die Clusterstruktur durch ein Regressionsmodell mit zusätzlichen Kontextdaten wie Wetter, Preisen und historischen Daten erweitert wird. Dadurch lassen sich dynamische und individuellere Prognosen erstellen, die insbesondere Spitzenlasten und kurzfristige Schwankungen realistischer abbilden.

Lastprognose

Zur Lastprognose in der Niederspannungsebene werden verschiedene Modellierungsmethoden eingesetzt. Statistische Verfahren wie Auto-Regressive Integrated Moving Average (ARIMA) oder multivariate Regressionen kommen zum Einsatz, wenn einfache zeitliche Muster mit externen Einflussgrößen, wie Wetter oder Tageszeit, korreliert werden sollen [12, 15]. Diese Techniken überzeugen vor allem durch ihre Transparenz und geringen Rechenaufwand, stoßen jedoch bei stark nichtlinearen und dynamischen Systemen oft an ihre Grenzen [30]. Für komplexere und nichtlineare Zusammenhänge, insbesondere bei kurz- bis mittelfristigen Prognosen, werden moderne Methoden des maschinellen Lernens genutzt. Besonders künstliche neuronale Netze, Support Vector Machines (SVM) und Gradient Boosting Machines (GBM) sind in der Lage, komplexe, nichtlineare Zusammenhänge aus großen Mengen historischer Daten zu extrahieren [12, 15].
Neuere Arbeiten nutzen auch Deep Learning-Modelle wie Long Short-Term Memory (LSTM)-Netze, um zeitliche Dynamiken in hochaufgelösten Zeitreihen, etwa beim Ladeverhalten von Elektrofahrzeugen, präzise vorherzusagen [12, 15]. Die Herausforderung

liegt hier jedoch oft in der datenintensiven Modellierung und dem Bedarf an kontinuierlicher Nachkalibrierung.

Neben der reinen Lastprognose rücken auch kombinierte Modelle in den Fokus, die weitere Eingangsdaten, wie Gebäudemerkmale, Clusterzugehörigkeiten, oder auch Strompreise, integrieren. Während Wetter- und Preisdaten in dieser Arbeit bereits in den generierten Fahrplänen der HEMS verarbeitet sind, können sie dennoch bei der Prognose für nicht-planende Haushalte durch ihren Einfluss auf das kollektive Verbrauchsverhalten relevante Informationen liefern und Lerneffekte erzielen. Hier zeigen vor allem hybride Modelle mit Clustering-Elementen, was die Autoren in [31] angewendet haben, einen deutlichen Vorteil in der Generalisierbarkeit und Robustheit. Durch die vorgelagerte Clustereinteilung lässt sich die Prognosegüte eines Feedforward-Deep-Neural-Networks signifikant verbessern [31]. Diese Kombination von Segmentierung und Vorhersage gilt als vielversprechender Ansatz, um sowohl Strukturinformationen als auch individuelle Verläufe zu berücksichtigen.

Einen weiteren Ansatz zur Prognoseverbesserung haben die Autoren in [32] erarbeitet, in dem Lastkurven von Verteiltransformatoren zunächst geclustert und anschließend mit Gradient-Boosting-Regressoren modelliert werden. Durch die Kombination von Clustering, das homogenere Gruppen erzeugt, und einer gezielten Feature-Selektion kann die Prognosegüte deutlich gesteigert werden, da komplexe Abhängigkeiten differenziert berücksichtigt werden [32]. Die Autoren in [31] schlagen ebenfalls eine Kombination von Clustering mit weiteren Verfahren vor. Hier werden Lastdaten zunächst gruppiert und anschließend über Dimensionsreduktionstechniken wie Principal Component Analysis (PCA) oder Uniform Manifold Approximation and Projection (UMAP) verdichtet. Auf dieser kompakten Repräsentation werden Prognosemodelle trainiert, die stabiler und weniger anfällig für Rauschen sind [31]. Einen anderen Weg verfolgen die Verfasser aus [33], die ein Ensemble-Framework einsetzen. Dabei werden die Daten zunächst in Cluster aufgeteilt, bevor mehrere Regressionsmodelle parallel trainiert und anschließend zu einer Gesamtprognose kombiniert werden [33]. Dieser ensemblebasierte Ansatz erweist sich als besonders robust, da Stärken verschiedener Verfahren gebündelt werden und sowohl Spitzenlasten als auch Grundlasten zuverlässiger erfasst werden können [33].

Insgesamt zeigt sich ein klarer Trend zur Kombination kategorialer Vorverarbeitung, etwa durch Clustering oder Feature Engineering, mit leistungsfähigen Vorhersagemodellen aus

dem Bereich des Deep Learnings. Diese hybriden Ansätze ermöglichen es, aus begrenzten oder heterogenen Datenbeständen belastbare Prognosen zu generieren und gleichzeitig strukturelle Besonderheiten der Lastprofile zu berücksichtigen. Die vorliegende Arbeit knüpft genau an diesen Stand der Technik an, indem sie eine Cluster-basierte Strukturierung mit einem anschließenden Prognosealgorithmus verbindet, der auf minimaler, aber gezielter Messdatengrundlage eine Lastprognose für Haushalte ermöglichen soll.

3.2 Dimensionsreduktion

Clustering-Verfahren stoßen in hochdimensionalen Datensätzen häufig an ihre Grenzen, da mit steigender Dimensionalität die Abstände zwischen Datenpunkten an Trennschärfe verlieren und viele Dimensionen Rauschen oder überflüssige Informationen enthalten. Eine Dimensionsreduktion kann daher ein entscheidender Schritt sein, um die Datenstruktur zu verdichten, wesentliche Muster hervorzuheben und damit die Clusterbildung zu erleichtern. Sie dient somit als erster Schritt, um die Daten in eine niedrigere Dimension zu überführen und durch die entstehende Darstellung potenzielle Cluster sichtbar zu machen. Dadurch können sowohl visuelle Analysen als auch automatische Clustering-Verfahren präziser und interpretierbarer durchgeführt werden. Ziel ist es, die Komplexität der Daten zu verringern, ohne deren innere Struktur zu verfälschen, sodass anschließende Clustering-Methoden zuverlässigere Ergebnisse liefern. [34]

3.2.1 Principal Component Analysis

Ein klassisches lineares Verfahren ist die Hauptkomponentenanalyse, welche im Englischen auch Principal Component Analysis (PCA) genannt wird. Dieses Verfahren projiziert die Daten in einen neuen Merkmalsraum, dessen Hauptkomponenten sukzessive die größte Varianz abbilden [35]. Durch diese Transformation werden Korrelationen zwischen Variablen aufgelöst und die wichtigsten Strukturen extrahiert, wodurch Clustering-Algorithmen oft stabilere und besser interpretierbare Ergebnisse erzielen [34]. Jede Hauptkomponente ist eine lineare Kombination der ursprünglichen Variablen, wobei die erste Hauptkomponente die Richtung der größten Varianz und die zweite Hauptkomponente die zweitgrößte Varianz unter der Bedingung orthogonaler Unabhängigkeit beschreibt, was auf beliebig viele Haupt-

komponenten erweitert werden kann [35]. Auf diese Weise entsteht eine geordnete Menge von Achsen, die die Daten zunehmend komprimieren. In vielen Datensätzen erklären schon wenige Komponenten den größten Teil der Varianz [35].
Für Clustering-Verfahren bietet dies zwei entscheidende Vorteile. Zum einen werden Rauschen und redundante Informationen entfernt, wodurch Cluster klarer hervortreten. Zum anderen erleichtert die Projektion auf zwei oder drei Komponenten die Visualisierung und damit die Interpretation der Clusterstruktur. Allerdings ist PCA ein linearer Ansatz, sodass komplexe, nichtlineare Zusammenhänge verborgen bleiben. Somit sind die Cluster zum Teil nur verzerrt oder unvollständig sichtbar. [34]

3.2.2 Uniform Manifold Approximation and Projection

Für Datensätze mit komplexeren, nichtlinearen Strukturen bieten sich hingegen Verfahren wie Uniform Manifold Approximation and Projection (UMAP) an. UMAP basiert auf der Annahme, dass hochdimensionale Daten auf einer zugrunde liegenden niedrigdimensionalen Struktur liegen. Es konstruiert zunächst einen gewichteten Graphen, in dem die Knoten Datenpunkte und die Kanten deren Nachbarschaften repräsentieren. Anschließend wird versucht, diese Struktur in einen Raum niedriger Dimension so zu projizieren, dass Nachbarschaften und Dichten möglichst gut erhalten bleiben. Dadurch ist UMAP in der Lage, nichtlineare Strukturen wie verschachtelte Cluster oder Cluster unterschiedlicher Form und Dichte darzustellen, ein Szenario, in dem PCA an ihre Grenzen stößt [34]. [36]
Durch UMAP können im Clustering komplexe Strukturen sichtbar gemacht und für Verfahren wie Density-Based Spatial Clustering of Applications with Noise (DBSCAN) oder spektrales Clustering aufbereitet werden. Im Kontext von Clustering eröffnet UMAP somit die Möglichkeit, komplexe Strukturen sichtbar zu machen und diese für Algorithmen wie DBSCAN oder spektrales Clustering zu nutzen. Gleichzeitig hat UMAP aber auch klare Grenzen. Ein Nachteil dieser Methode besteht darin, dass UMAP stärker von Hyperparametern abhängt und die Ergebnisse weniger direkt interpretierbar sind als die klar definierten Hauptkomponenten der PCA. Auch die Größen und Dichten von Clustern entsprechen nicht zwingend der Realität, da UMAP lokale Dichteunterschiede aktiv ausgleicht. Zudem handelt es sich um ein stochastisches Verfahren, sodass verschiedene Läufe teilweise unterschiedliche Er-

gebnisse liefern können. Deshalb empfiehlt es sich, mehrere Durchläufe zu vergleichen und stabile Strukturen zu identifizieren. Außerdem haben die Achsen im UMAP-Plot keine feste Bedeutung und einzelne scheinbare Strukturen können Optimierungsartefakte sein, weshalb Ergebnisse stets kritisch interpretiert werden sollten. [36]
Insgesamt dient die Dimensionsreduktion im Kontext des Clusterings also nicht nur der Vereinfachung und Visualisierung, sondern auch der Verbesserung der Clusterqualität. Die Wahl der Methode hängt dabei maßgeblich von den Eigenschaften des Datensatzes sowie den Zielen der Analyse ab [34, 35, 36].

3.3 Clustering-Algorithmen

Clustering zählt zu den grundlegenden Verfahren des unüberwachten maschinellen Lernens und dient der Gruppierung von Datenpunkten auf Basis ihrer Ähnlichkeit. Im Kontext der elektrischen Energiesysteme, insbesondere im Bereich der Analyse elektrischer Lastprofile, ermöglicht Clustering die Identifikation typischer Verbrauchsmuster und die Reduktion der Komplexität heterogener Smart-Meter-Daten [37]. Ziel ist es, Gruppen von Haushalten oder Verbrauchern zu identifizieren, die ein vergleichbares zeitliches Lastverhalten aufweisen, um daraus repräsentative Profile für Netzplanung, Tarifgestaltung oder Prognosemodelle abzuleiten.

3.3.1 Partitionierendes Clustering

Beim partitionierenden Clustering ist die Grundidee, dass der Algorithmus die Daten in genau eine Partition aufteilt, das heißt in eine feste Anzahl voneinander abgegrenzter Gruppen. Diese gebildeten Gruppen werden auch als Cluster bezeichnet. Ziel ist es, dass die Punkte innerhalb eines Clusters möglichst ähnlich zueinander sind, während sich die Gruppen untereinander möglichst stark unterscheiden. Das Ergebnis ist eine sogenannte harte Partitionierung der Daten, bei der jeder Punkt eindeutig einer Gruppe zugewiesen wird. [38]
Die bekannteste Methode der partitionierende Klasse ist der K-Means-Algorithmus, welcher sich auch in der Analyse elektrischer Lastprofile etabliert hat [37]. Ziel des Verfahrens ist es, eine gegebene Menge von Datenpunkten, wie beispielsweise Tageslastverläufe, in eine Anzahl von k Clustern zu unterteilen [37]. Die daraus resultierenden Cluster sollen intern

möglichst homogen und zueinander möglichst verschieden sein. So können Strukturen und Muster in großen Datensätzen erkannt werden, ohne dass die Daten vorher beschriftet oder kategorisiert wurden [38].

Formal basiert das Verfahren auf einer Optimierung der Streuung von Datenpunkten innerhalb eines Clusters [39]. Dabei wird die Zugehörigkeit der Punkte zu Clustern so gewählt, dass die Summe der quadrierten Abstände zwischen den Punkten und den zugehörigen Clusterzentren minimiert wird [38]. Diese Zielgröße wird durch die Formel 3.1 beschrieben.

$$d = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \left\| x_i^{(k)} - u_k \right\|^2 \tag{3.1}$$

Dabei bezeichnet K die Anzahl der Cluster, n_k die Anzahl der Datenpunkte im k-ten Cluster, $x_i^{(k)}$ den i-ten Datenpunkt im Cluster k und u_k das Zentrum, also den Mittelwert dieses Clusters [39]. Der Ausdruck $\left\| x_i^{(k)} - u_k \right\|^2$ steht für den quadrierten euklidischen Abstand zwischen einem Punkt und dem zugehörigen Clusterzentrum [39]. Das Ziel des Algorithmus ist es, diesen Wert, also die Streuung innerhalb der Cluster, zu minimieren, um möglichst kompakte und voneinander unterscheidbare Gruppen zu bilden [38]. Der Algorithmus arbeitet iterativ. Zunächst werden k zufällige Startpunkte als vorläufige Zentren gewählt. Anschließend werden alle Datenpunkte dem jeweils nächstgelegenen Zentrum zugewiesen. Danach werden die Zentren als Mittelwerte der jeweils zugewiesenen Punkte neu berechnet. Dieser Schritt wird so lange wiederholt, bis sich die Clusterzuweisungen nicht mehr ändern oder eine Abbruchbedingung erreicht ist [39].

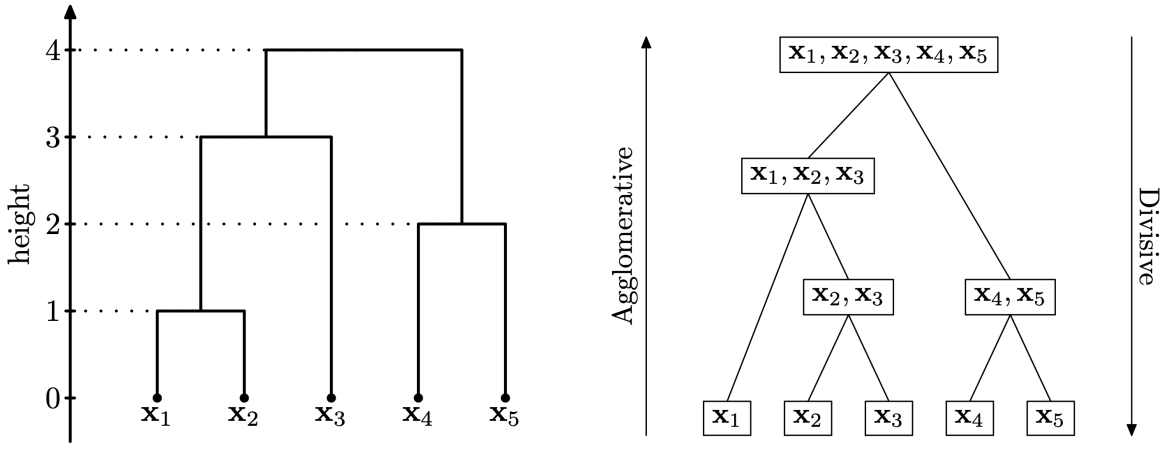
Die Vorteile von K-Means liegen in der einfachen Implementierbarkeit, guten Skalierbarkeit bei großen Datensätzen sowie der Fähigkeit, bei gut separierten Clustern schnell sinnvolle Gruppierungen zu liefern [40]. Dies macht es insbesondere für Anwendungen für Netzbetreiber attraktiv, etwa zur Typisierung von Haushaltslastprofilen oder zur Identifikation flexibler Verbraucher [37]. Der Algorithmus arbeitet in der Praxis schnell, vor allem bei kleineren bis mittelgroßen Datensätzen [40].

Allerdings bringt das Verfahren auch Einschränkungen mit sich. Ein grundlegender Nachteil ist die Abhängigkeit von der Anfangswahl der Cluster-Zentren [40]. Schlechte Initialisierungen können dazu führen, dass der Algorithmus in einem lokalen Minimum statt im globalen Optimum konvergiert [38]. Eine weitere Einschränkung ist die Bestimmung der Anzahl der

Cluster, welche a priori festzulegen ist. Das kann oftmals in praktischen Anwendungen problematisch sein, insbesondere ohne Vorwissen über die zugrunde liegende Datenstruktur. Typischerweise wird die Anzahl der Cluster k daher mit Hilfe heuristischer Methoden wie der Elbow-Methode oder der Silhouettenanalyse bestimmt [38], was in Kapitel 3.5 weiter behandelt wird. Zudem ist der Algorithmus empfindlich gegenüber Ausreißern und nichtsphärischen Clustern [40].
Es existieren mehrere Abwandlungen und Erweiterungen des ursprünglichen Verfahrens, die gezielt auf typische Probleme wie ungünstige Initialisierungen, mangelnde Skalierbarkeit, lineare Einschränkungen oder die starre Clusterzuordnung reagieren. Diese Varianten behalten die Grundidee von K-Means bei, erweitern jedoch das Verfahren, um es flexibler und robuster gegenüber verschiedenen Anforderungen und Datentypen zu machen [41].

3.3.2 Hierarchische Custering-Verfahren

Hierarchisches Clustering ist eine Methode der Clusteranalyse, bei der eine hierarchische Struktur in den Daten aufgebaut wird, entweder durch wiederholtes Zusammenfügen kleinerer Cluster (bottom-up) oder durch schrittweises Aufteilen einer großen Gruppe (top-down). Im Gegensatz zu partitionierenden Verfahren, bei denen eine feste Anzahl von Clustern vorab definiert werden muss, erzeugt das hierarchische Clustering eine verschachtelte Baumstruktur. [41]
Diese Struktur wird auch Dendrogramm genannt und ist in Abbildung 3.1a dargestellt. Dieses Grundkonzept stellt die Clusterbeziehungen auf verschiedenen Detailebenen dar. Dies ermöglicht es, erst nachträglich zu entscheiden, wie viele Cluster gebildet werden sollen, indem das Dendrogramm ab einer bestimmten Höhe begrenzt wird. Beim hirarchischen Clustering wird zwischen agglomerativem und divisivem Clustering unterschieden. [41]
Das agglomerative Clustering, welches auch als Bottom-up-Ansatz bezeichnet wird, beginnt damit, dass jeder Datenpunkt als eigenes Cluster betrachtet wird. In jedem Schritt werden dann die zwei ähnlichsten Cluster zusammengeführt, bis schließlich alle Punkte in einem einzigen, übergeordneten Cluster vereint sind. [41]
Das divisive Clustering, auch als Top-down-Ansatz bekannt, funktioniert genau umgekehrt. Hier wird mit allen Datenpunkten in einem einzigen großen Cluster begonnen, welches dann



(a) Prinzipieller Aufbau eines Dendrogramms mit fünf Datenpunkten aus [41]

(b) Agglomeratives und divisives hierarchisches Clustering aus [41]

Abbildung 3.1: Darstellung der hierarchischen Clusterbildung aus [41]

schrittweise in kleinere Untergruppen aufgeteilt wird. Dieser Ansatz wird in der Praxis seltener eingesetzt als das agglomerative Verfahren, da die rekursive Aufspaltung rechnerisch aufwendiger ist [41]. In Abbildung 3.1b wird schematisch der Unterschied beider Methoden dargestellt.

Ein großer Vorteil hierarchischer Verfahren besteht darin, dass sie keine vorherige Festlegung der Clusteranzahl erfordern. Stattdessen liefern sie eine vollständige Hierarchie, aus der sich verschiedene Clustering-Ebenen ableiten lassen. Durch das Dendrogramm kann die Clusterstruktur visuell erfasst und auf verschiedenen Auflösungsebenen analysiert werden. [37]

Allerdings hat auch das hierarchische Clustering einige Nachteile. Es ist im Allgemeinen weniger skalierbar als partitionierende Verfahren wie K-Means, da es eine höhere Rechenkomplexität aufweist, insbesondere bei großen Datensätzen [41]. Zudem sind die Ergebnisse irreversibel. Ist ein Zusammenfügen oder Aufteilen einmal erfolgt, kann es im weiteren Verlauf nicht rückgängig gemacht werden. Trotz Einschränkungen ist das hierarchische Clustering besonders nützlich zur Analyse von Haushaltslastverläufen im Hinblick auf technische Ausstattungsmerkmale [37].

3.3.3 Dichtebasierte Verfahren

Dichtebasiertes Clustering zeichnet sich dadurch aus, dass es Cluster beliebiger Form und Größe identifizieren kann [37]. Der Algorithmus untersucht dabei die lokale Punktdichte innerhalb eines Datenraums [41]. Ziel dieses Verfahrens ist es, Bereiche mit hoher Dichte als zusammengehörige Cluster zu identifizieren. Hingegen werden dünn besiedelte Bereiche, die keine ausreichende Dichte aufweisen, als Rauschen bzw. Ausreißer klassifiziert. Damit ist dieser Ansatz besonders gut geeignet für reale, komplexe Datensätze mit unregelmäßiger Verteilung, verrauschten Daten und unterschiedlich geformten Clustern [37].

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

Das bekannteste Verfahren innerhalb dieser Kategorie ist DBSCAN. Die Funktionsweise des Algorithmus wird schematisch in Abbildung 3.2 dargestellt [42]. DBSCAN arbeitet mit zwei zentralen Parametern. Der ϵ-Wert definiert den Radius um einen Punkt, innerhalb dessen andere Punkte als Nachbarn gelten [41]. Der zweite Parameter, ursprünglich als $Nmin$ bezeichnet und in neueren Quellen meist als $MinPts$ bekannt, gibt an, wie viele Punkte sich mindestens im ϵ-Umkreis eines Punktes befinden müssen, damit dieser als Kernpunkt klassifiziert wird [41, 37]. Basierend auf diesen Kriterien unterscheidet DBSCAN drei Arten von Punkten. Kernpunkte, die über genügend Nachbarn verfügen, um als Ausgangspunkt für ein Cluster zu dienen, Randpunkte, die zwar im ϵ-Radius eines Kernpunkts liegen, aber selbst nicht genügend Nachbarn besitzen, und Rauschpunkte, die keinem Cluster zugeordnet werden können, weil sie isoliert im Raum liegen [41].

Der Algorithmus beginnt mit einem beliebigen Punkt im Datensatz. Wird dieser als Kernpunkt identifiziert, wird ein neues Cluster begonnen. Alle Punkte, die direkt oder indirekt von diesem Kernpunkt aus über eine Folge von Nachbarschaftsbeziehungen erreichbar sind, werden diesem Cluster zugeordnet [41]. Dieser Prozess wird iterativ fortgesetzt, bis keine weiteren Punkte mehr hinzugefügt werden können. Anschließend wird der Algorithmus mit dem nächsten nicht besuchten Punkt fortgesetzt. Durch diese Vorgehensweise entstehen Cluster dort, wo lokal eine ausreichende Dichte vorhanden ist, während dünn besiedelte Bereiche automatisch als Rauschen erkannt werden [41].

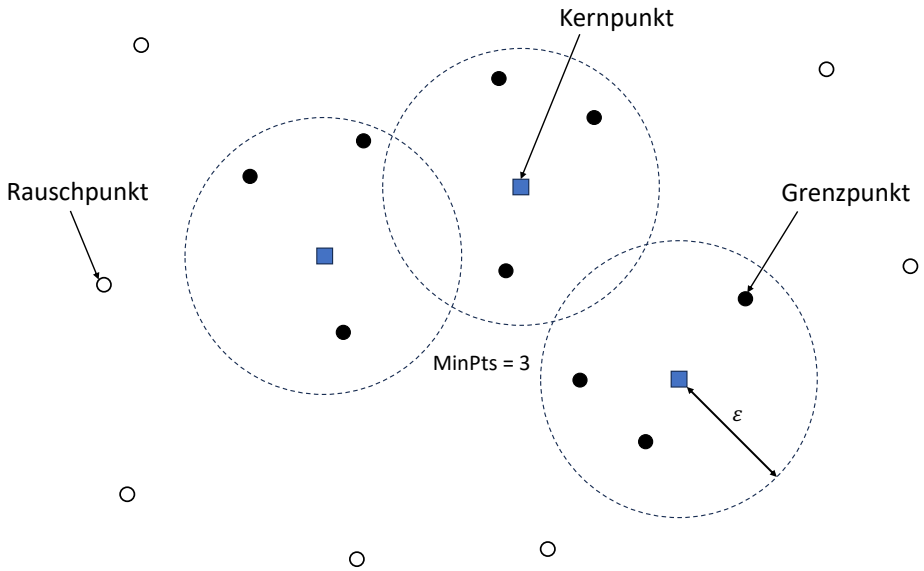


Abbildung 3.2: Schematische Darstellung der Funktionsweise des DBSCAN-Algorithmus nach [42]

Ein wesentlicher Vorteil von DBSCAN liegt darin, dass keine Anzahl der Cluster im Vorfeld angegeben werden muss, im Gegensatz zum K-Means-Verfahren [37]. Stattdessen entstehen die Cluster basierend auf der lokalen Struktur der Daten. Darüber hinaus kann DBSCAN Cluster beliebiger geometrischer Form erkennen, beispielsweise langgezogene, gebogene oder ringförmige Cluster [41]. Auch die explizite Erkennung von Ausreißern macht DBSCAN geeignet für reale Szenarien mit unvollständigen oder verrauschten Daten [37].
Eine Schwäche des DBSCAN ist, dass die Wahl der Parameter ϵ und $MinPts$ nicht trivial ist [37]. Ein zu kleiner ϵ-Wert führt dazu, dass viele kleine Cluster oder zu viele Rauschpunkte entstehen. Ein zu großer Wert kann dazu führen, dass verschiedene reale Cluster zu einem einzigen zusammengefasst werden. Dieses Problem verstärkt sich, wenn der Datensatz unterschiedliche Dichtebereiche enthält, da DBSCAN mit einem globalen ϵ-Wert arbeitet und somit nicht zwischen dichter und weniger dichter Struktur differenzieren kann.

Hierarchical Density-Based Spatial Clustering of Applications with Noise

Neben DBSCAN existieren weitere dichtebasierte Verfahren, wie Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), das hierarchische Elemente mit der Dichtebasiertheit kombiniert [44]. Es ermöglicht die Identifikation von Clustern auf unterschiedlichen Dichteebenen und liefert zusätzlich Stabilitätsmaße, mit denen die

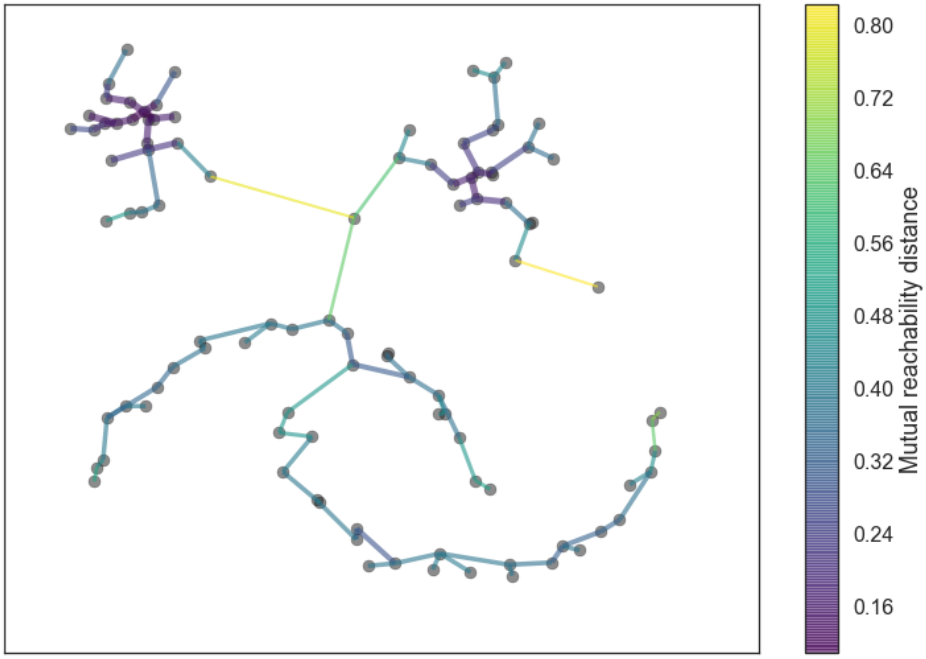


Abbildung 3.3: Darstellung des gegenseitigen Erreichbarkeitsgraphen im HDBSCAN-Algorithmus aus [43]

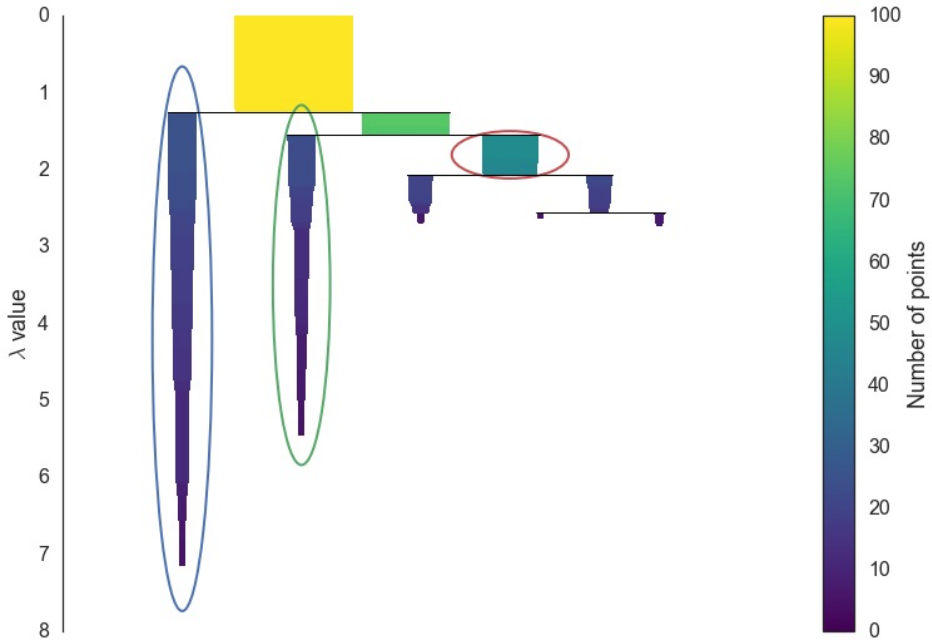


Abbildung 3.4: Komprimierter Hierarchiebaum des HDBSCAN-Algorithmus aus [43]

Verlässlichkeit eines Clusters beurteilt werden kann. Während DBSCAN einen festen Dichte-Schwellenwert liefert, erstellt HDBSCAN eine vollständige hierarchische Clusterstruktur, aus der anschließend durch ein Auswahlverfahren Cluster extrahiert werden können [44]. Dabei berechnet HDBSCAN dieselben Dichtelevel wie DBSCAN aus Abbildung 3.2, ver-

zichtet jedoch auf die Festlegung eines einzelnen ϵ-Wertes [44]. Stattdessen basiert der Algorithmus auf dem Konzept der gegenseitigen Erreichbarkeitsdistanz, einem Distanzmaß, das die lokalen Dichteverhältnisse berücksichtigt [44].

Die Abbildung 3.3 zeigt einen solchen Erreichbarkeitsgraphen, in dem alle Datenpunkte durch Kanten verbunden werden. Dieses Maß kombiniert den tatsächlichen Abstand zwischen zwei Punkten mit Informationen über die lokale Dichte ihrer Umgebung [45]. Dadurch werden Dichteunterschiede im Datensatz berücksichtigt. In der Abbildung 3.3 kennzeichnen dunkle Farben geringe Distanzen und damit eine hohe lokale Dichte, während helle Farben größere Distanzen und somit Bereiche geringerer Dichte anzeigen. Diese Distanzwerte bestimmen die Gewichtung der Kanten im Graphen und bilden die Grundlage für den anschließenden Aufbau des gewichteten minimalen Spannbaums [45]. In Abbildung 3.4 ist der komprimierte Spannbaum dargestellt, der aus dem in Abbildung 3.3 gezeigten gegenseitigen Erreichbarkeitsgraphen abgeleitet wird. Diese Struktur beschreibt, wie sich Cluster bei zunehmenden Dichteschwellen aufspalten oder auflösen. Auf der y-Achse ist der Parameter λ dargestellt, der die jeweilige Dichteschwelle angibt. Je größer der λ-Wert, desto dichter ist der betrachtete Bereich. Die Farbskala zeigt die Anzahl der Punkte pro Cluster, wobei dunkle Farben kleinere und hellere Farben größere Cluster repräsentieren. Zur Auswahl der bedeutendsten Cluster wird ein Stabilitätsmaß eingeführt, das bewertet, wie lange ein Cluster über verschiedene Dichtelevel hinweg bestehen bleibt [44]. Die Cluster mit der höchsten Stabilität gelten als besonders konsistent und repräsentieren die strukturell relevantesten Gruppierungen im Datensatz. In der Abbildung 3.4 sind jene Cluster markiert, die lokal die höchste Stabilität aufweisen. Diese drei Cluster bleiben über einen besonders großen Dichtebereich hinweg bestehen und gelten daher als die strukturell relevantesten Gruppierungen im Datensatz. Der rechts erkennbare Cluster entsteht, weil in diesem Bereich eine kleinere Gruppe von Punkten mit lokal erhöhter Dichte vorliegt. Er bleibt jedoch nur über einen begrenzten Dichtebereich hinweg bestehen und ist weniger kompakt, was auf eine geringere Punktdichte und damit eine niedrigere Stabilität im Vergleich zu den linken Clustern hinweist [44]. Die anderen, weniger stabilen Cluster werden verworfen und nicht in das Endergebnis übernommen.

Ein zentraler Vorteil von HDBSCAN gegenüber von DBSCAN liegt darin, dass keine ϵ-Schwelle als Eingabeparameter erforderlich ist. Stattdessen ist nur ein einziger Parameter

nötig, der die minimale Anzahl von Punkten angibt, die ein Cluster enthalten muss. Dadurch eignet sich HDBSCAN besonders für komplexe Datensätze mit Clustern variabler Dichte, die durch klassische Verfahren nur unzureichend erkannt würden. Wie bei DBSCAN erkennt auch HDBSCAN Ausreißer explizit und klassifiziert sie als Rauschen. Darüber hinaus bestimmt HDBSCAN ebenfalls die optimale Anzahl von Clustern automatisch anhand eines Stabilitätsmaßes und liefert zusätzlich eine hierarchische Darstellung der Clusterstruktur, wodurch Analysen auf unterschiedlichen Dichteebenen möglich sind. [44]
Somit ermöglichen dichtebasierte Clustering-Verfahren es, typische Verbrauchsmuster zu identifizieren, ohne dass die Anzahl der Cluster im Voraus bekannt sein muss, und bieten gleichzeitig eine automatische Erkennung von Ausreißern wie Messfehlern oder atypischen Lastspitzen. Durch ihre Fähigkeit, auch komplexe, nichtlineare Strukturen in verrauschten Datenräumen abzubilden, sind sie besonders geeignet für die Erkennung von Lastanomalien oder Lastprofilmustern. [37]

3.3.4 K-Nearest-Neighbor für die Clusterzuweisung

Um die bestehende Datenlücke durch fehlende Fahrpläne einzelner Haushalte zu schließen und gleichzeitig die bereits vorliegenden Informationen der bekannten Haushalte zu nutzen, wird das k-Nearest Neighbor (KNN)-Verfahren eingesetzt. Ziel ist es, unbekannte Haushalte anhand ihrer Merkmalsausprägungen den bestehenden Clustern zuzuordnen, damit eine repräsentative Lastkurve entsteht. Das Verfahren basiert auf der Annahme, dass ähnliche Beobachtungen im Merkmalsraum auch ähnlichen Clustern angehören.
Das KNN-Verfahren ist ein einfaches, aber wirkungsvolles Verfahren zur Klassifikation bzw. Zuweisung von Datenpunkten auf Basis ihrer Ähnlichkeit zu bereits bekannten Beobachtungen. Das Grundprinzip besteht darin, dass neue oder unbekannte Datenpunkte denjenigen Klassen bzw. Clustern zugeordnet werden, denen ihre k nächsten Nachbarn im Merkmalsraum angehören. Die Bestimmung der Nachbarn erfolgt anhand eines Distanz- oder Ähnlichkeitsmaßes, das die Nähe zwischen den Datenpunkten beschreibt. [46]
Für einen neuen Datenpunkt $\mathbf{x}_{\text{neu}}$ werden zunächst die Distanzen zu allen bekannten Punkten $\mathbf{x}_i \in X_{\text{bekannt}}$ berechnet. Ein häufig verwendetes Maß zur Bestimmung dieser Ähnlichkeit ist die Cosine-Distanz, die in Kapitel 3.8 näher thematisiert wird. Die k Punkte mit der

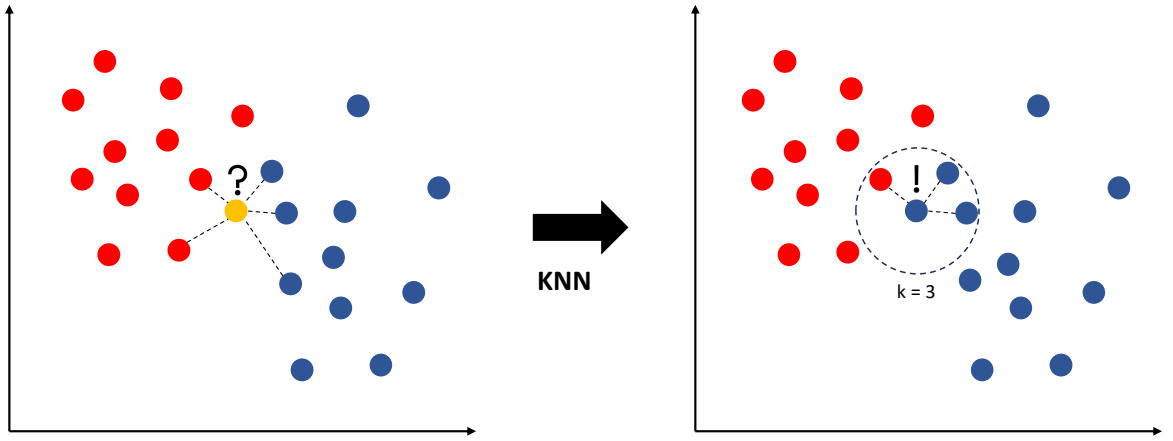


Abbildung 3.5: Schematische Darstellung des KNN-Verfahrens

geringsten Distanz bilden die Menge der nächsten Nachbarn $\mathcal{N}_k(\mathbf{x}_{\text{neu}})$. Auf Basis dieser Nachbarschaft wird die Zugehörigkeit des neuen Punktes zu einem Cluster bestimmt. [46]

$$C(\mathbf{x}_{\text{neu}}) = \arg\max_{c \in \mathcal{C}} \sum_{i \in \mathcal{N}_k(\mathbf{x}_{\text{neu}})} \mathbb{I}(C(\mathbf{x}_i) = c)\,, \tag{3.2}$$

Die Gleichung 3.2 beschreibt die Zuweisung, welche nach dem Mehrheitsprinzip erfolgt, wobei $\mathcal{C}$ die Menge aller existierenden Cluster und $\mathbb{I}(\cdot)$ die Indikatorfunktion ist, die den Wert 1 annimmt, wenn der Nachbar $\mathbf{x}_i$ dem Cluster c angehört, und andernfalls den Wert 0. Das Verfahren ordnet den neuen Punkt somit dem Cluster zu, dem die Mehrheit seiner k nächsten Nachbarn angehört. [46]

Die Abbildung 3.5 zeigt schematisch die Funktionsweise des KNN. Auf der linken Seite ist ein neuer, unbekannter Datenpunkt in gelb dargestellt, dessen Zugehörigkeit zu einem Cluster unklar ist. Im dargestellten Beispiel werden die Distanzen zu den umliegenden Punkten berechnet, wobei die $k = 3$ nächsten Nachbarn durch gestrichelte Linien gekennzeichnet sind. Da die Mehrheit dieser Nachbarn zum blauen Cluster gehört, wird der neue Punkt ebenfalls dieser Gruppe zugeordnet.

Ein zentraler Aspekt des Verfahrens ist die Wahl des Parameters k, also der Anzahl der berücksichtigten Nachbarn. Ein kleiner Wert von k führt zu einer hohen Empfindlichkeit gegenüber Ausreißern, da bereits einzelne fehlerhafte oder atypische Beobachtungen die Zuordnung des neuen Punktes beeinflussen können. Ein großer Wert von k hingegen bewirkt eine Glättung der Entscheidungsgrenzen und kann dazu führen, dass lokale Strukturen im Datensatz verloren gehen. [46]

3.4 Vergleich der Clustering Methoden

Beim Vergleich der verschiedenen Clustering-Methoden zeigen sich deutliche Unterschiede in ihrer Vorgehensweise und ihren Anwendungsgebieten. Um die Unterschiede zwischen den verschiedenen Clustering-Ansätzen zu verdeutlichen, bietet es sich an, die jeweiligen Ergebnisse anhand künstlich erzeugter Beispieldatensätze miteinander zu vergleichen. Solche Datensätze weisen gezielt unterschiedliche Strukturen auf und stellen damit typische Herausforderungen für Clustering-Verfahren dar. Während einige Methoden vor allem für klar getrennte, annähernd kugelförmige Cluster geeignet sind, zeigen andere ihre Stärken bei nichtlinearen oder dichtebasierten Strukturen. Die Abbildung 3.6 illustriert, wie die vorgestellten Verfahren mit diesen unterschiedlichen Situationen umgehen. Hier wird ein Überblick über die vorgestellten Clustering-Verfahren anhand von synthetischen Datensätzen mit jeweils unterschiedlichen Strukturen aus [47] gegeben. Die dargestellten Algorithmen sind K-Means (partitionierend), Agglomeratives Clustering (hierarchisch), Spectral Clustering (graphenbasiert), HDBSCAN (dichtebasiert) und Gaussian Mixture (modellbasiert). Diese wurden auf identische Daten angewendet, um ihre jeweiligen Stärken und Schwächen im direkten Vergleich zu veranschaulichen.

Bei komplexen, nichtlinear trennbaren Strukturen, wie konzentrischen Kreisen oder halbmondförmigen Clustern, versagt K-Means, da es auf lineare Trennung und kugelförmige Cluster ausgelegt ist [38]. Das spektrale Clustering hingegen erkennt die Struktur korrekt, indem es die zugrunde liegende Ähnlichkeitsstruktur im Graphen nutzt [48]. Auch HDBSCAN schneidet bei den Bögen sehr gut ab, während Gaussian Mixture Model (GMM) auf die Annahme elliptischer Clusterformen angewiesen ist [46]. Agglomeratives Clustering liefert in diesen Fällen ebenfalls überzeugende Ergebnisse, da es keine feste Form annimmt und durch die hierarchische Zusammenführung flexibel auf die Datenstruktur reagieren kann.

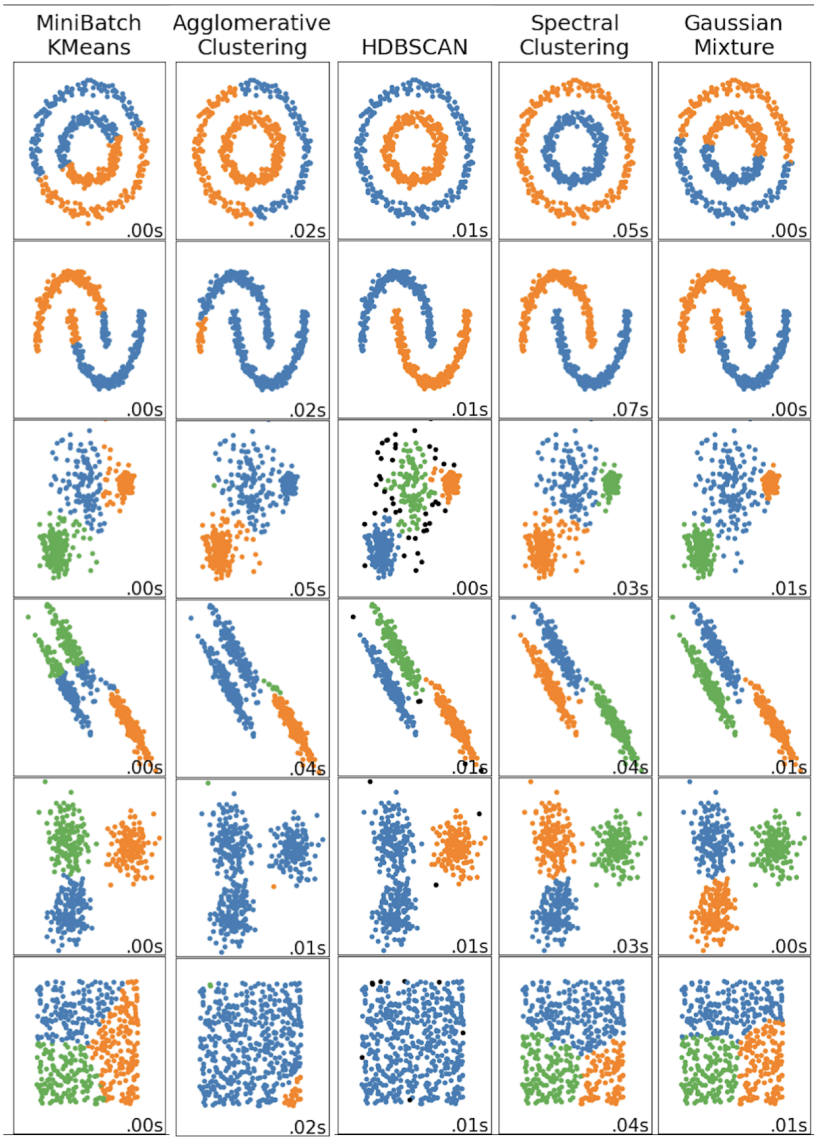


Abbildung 3.6: Vergleich der Clustering Methoden anhand von Clusterformen aus [47]

Bei unregelmäßig geformten Clustern mit Ausreißern eignet sich ein dichtebasiertes Verfahren wie HDBSCAN besonders gut. Es identifiziert sowohl die Clusterstruktur als auch die Ausreißer zuverlässig. MiniBatch K-Means und GMM trennen zwar die Hauptgruppen, erkennen aber keine Ausreißer. Das spektrale Clustering liefert ebenfalls gute Ergebnisse. Das agglomerative Clustering kann die grobe Struktur der Daten erfassen, unterscheidet jedoch nicht explizit zwischen Ausreißern, da alle Punkte in die Hierarchie integriert werden. In Fällen mit gut separierten, länglichen Clustern funktionieren alle Verfahren weitgehend zuverlässig. Ähnlich ist es bei klar voneinander getrennten, kompakten Clustern, bei denen alle fünf Methoden korrekte Ergebnisse liefern.

In der letzte Zeile sind die Punkte zufällig verteilt und weisen keine reale Clusterstruktur auf. Hier werden die Unterschiede besonders deutlich. Während K-Means, GMM und das spektrale Clustering eine künstliche Gruppierung erzeugen, erkennt HDBSCAN, dass keine sinnvolle Clusterstruktur vorhanden ist, und weist alle Punkte dem gleichen Cluster zu. Das agglomerative Clustering bildet zwei Cluster, da der Algorithmus zwangsläufig eine vollständige Hierarchie erzeugt, obwohl keine echte Struktur vorhanden ist.

Clustering-Methode	**Stärken**	**Schwächen**
Partitionierend (z.B. K-Means)	• Einfach und effizient bei großen Datensätzen • Gut bei konvexen, kugelförmigen Clustern • Leicht verständlich und implementierbar	• Anzahl der Cluster muss vorgegeben werden • Sensitiv gegenüber Ausreißern • Funktioniert schlecht bei nicht-konvexen oder unterschiedlich großen Clustern
Hierarchisch (z.B. Agglomerativ)	• Keine Angabe der Clusteranzahl im Voraus nötig • Visualisierung über Dendrogramm möglich • Liefert eine vollständige Clusterhierarchie	• Hoher Rechenaufwand bei großen Datensätzen • Entscheidungen beim Fusionieren sind nicht reversibel • Sensitiv gegenüber Rauschdaten und Ausreißern
Dichtebasiert (z.B. DBSCAN)	• Erkennt beliebig geformte Cluster • Robust gegenüber Ausreißern • Keine Angabe der Clusteranzahl erforderlich	• Schwierigkeiten bei variierender Dichte • Wahl geeigneter Parameter (ϵ, MinPts) ist kompliziert • Höhere Komplexität bei hohen Dimensionen
Modellbasiert (z.B. GMM)	• Liefert Wahrscheinlichkeiten für Clusterzugehörigkeit • Flexibler als K-Means (z.B. elliptische Cluster) • Statistisch fundiert	• Annahmen über Verteilung (z.B. Normalverteilung) notwendig • Rechenintensiv • Empfindlich gegenüber Initialisierung und Ausreißern
Graphenbasiert (z.B. Spectral Clustering)	• Funktioniert gut bei nicht-konvexen Strukturen • Kann komplexe Clusterformen erkennen • Nutzt globale Struktur der Daten	• Hoher Rechenaufwand bei großen Datenmengen • Erfordert Wahl einer Ähnlichkeitsfunktion • Anzahl der Cluster muss meist vorgegeben werden

Tabelle 3.1: Stärken und Schwächen der vorgestellten Clustering-Methoden nach [37, 40, 41, 46, 48, 47]

Die Wahl einer geeigneten Clustering-Methode hängt somit von den Eigenschaften der Daten und den Zielen der Analyse ab. In Tabelle 3.1 werden die fünf behandelten Clustering-Ansätze hinsichtlich ihrer jeweiligen Stärken und Schwächen miteinander verglichen. Insgesamt zeigt der Vergleich, dass kein Verfahren universell geeignet ist und seine Leistungsfähigkeit maßgeblich vom jeweiligen Anwendungsfall abhängt.
In dieser Arbeit wird HDBSCAN als Clustering-Methode für die Lastkurven gewählt, da dieses Verfahren im Gegensatz zu klassischen Methoden wie K-Means keine Vorabfestlegung der Clusteranzahl erfordert. Zudem ist HDBSCAN robust gegenüber Ausreißern und kann Cluster unterschiedlicher Form und Dichte identifizieren. Damit eignet sich HDBSCAN besonders für die heterogenen Verbrauchsmuster der Haushalte, die typischerweise durch starke zeitliche Schwankungen, unterschiedliches Verbrauchsverhalten und nichtlineare Strukturen in den Lastkurven charakterisiert sind.

3.5 Cluster-Validierung

Eine zentrale Herausforderung der Clusteranalyse besteht darin, die Güte einer Clusterlösung objektiv zu bewerten und die angemessene Anzahl von Clustern zu bestimmen. Ziel ist es, zu prüfen, ob die durch einen Algorithmus erzeugten Cluster tatsächlich eine sinnvolle Struktur in den Daten widerspiegeln oder zufällig zustande gekommen sind. Interne Validierungsmethoden bewerten die Clusterqualität ausschließlich anhand der Daten selbst, indem sie zum Beispiel die Kompaktheit innerhalb der Cluster und die Trennschärfe zwischen den Clustern messen [49]. Externe Validierungsmethoden hingegen prüfen anhand vorhandener Referenzlabels, inwieweit die ermittelten Cluster mit den bekannten Klassen übereinstimmen. Die folgenden Validierungsmetriken sind eng mit Verfahren wie K-Means verknüpft, da diese kompakte, annähernd sphärische Cluster erzeugen und eine Vorauswahl der Clusteranzahl erfordern [49]. Dennoch lassen sich diese Metriken auch auf andere Clustering-Methoden anwenden, da diese lediglich die Abstände zwischen Datenpunkten und Clusterzuweisungen berücksichtigen [49]. Aus diesem Grund werden folgende Metriken in dieser Arbeit vorgestellt, um unterschiedliche Clustering-Verfahren bewerten zu können.

3.5.1 Davies-Bouldin-Index

Der Davies-Bouldin-Index (DBI) ist eine klassische Validierungsmethode, die die Qualität einer Clusterlösung anhand von Kompaktheit und Trennschärfe bewertet [49]. Dabei beurteilt der Index nicht die einzelnen Cluster isoliert, sondern die Beziehungen zwischen ihnen, also wie stark sie voneinander getrennt sind, unter Berücksichtigung ihrer jeweiligen Streuung. Für jedes Cluster wird zunächst die durchschnittliche Distanz der Datenpunkte zu ihrem jeweiligen Clusterzentrum berechnet, was die Clusterstreuung beschreibt [49]. Anschließend wird für jedes Cluster ein Verhältnis gebildet, das seine Streuung ins Verhältnis zur Distanz zum ähnlichsten Nachbarcluster setzt. Für eine Lösung mit c Clustern wird der DBI in Formel 3.3 berechnet, wobei c_i das Zentrum des i-ten Clusters bezeichnet.

$$DB_c = \frac{1}{c} \sum_{i=1}^{c} \max_{j=1,\, j \neq i}^{c} \left\{ \frac{dia(c_i) + dia(c_j)}{\|c_i - c_j\|} \right\} \tag{3.3}$$

Die Clusterstreuung $dia(c_i)$ wird in Formel 3.4 definiert, wobei n_i die Anzahl der Punkte im Cluster c_i ist.

$$dia(c_i) = \left(\frac{1}{n_i} \sum_{x \in c_i} \|x - c_i\| \right)^{1/2} \tag{3.4}$$

Der DBI wird berechnet, indem für jedes Cluster i das maximale Verhältnis der Cluster-Streuung zur Distanz zum Zentrum eines anderen Clusters j ermittelt wird. Anschließend wird über alle Cluster der Mittelwert dieser Werte gebildet. Ein niedriger DBI-Wert weist auf kompakte und gut getrennte Cluster hin, während ein hoher Wert auf überlappende oder wenig trennscharfe Cluster hindeutet. Der Vorteil des DBI liegt in seiner einfachen Berechnung und breiten Anwendbarkeit auf verschiedene Clustering-Algorithmen. Allerdings reagiert er empfindlich auf Ausreißer und kann bei stark unterschiedlichen Clustergrößen oder -dichten zu weniger verlässlichen Ergebnissen führen. [49]

3.5.2 Calinski–Harabasz-Index

Der Calinski-Harabasz-Index (CHI) ist eine der am häufigsten eingesetzten Metriken zur Bewertung von Clusterlösungen. Er beruht auf dem Verhältnis von Streuung zwischen den Clustern zur Streuung innerhalb der Cluster [49]. Der CHI wird in Formel 3.5 definiert, wobei

B_m die Zwischen-Cluster-Streumatrix und W_m die innere Cluster-Streumatrix beschreibt. Die Gesamtzahl der geclusterten Datenpunkte wird mit N bezeichnet und c gibt die Anzahl der Cluster an [49].

$$CH(c) = \frac{\mathrm{trace}(B_m) \cdot (N - C)}{\mathrm{trace}(W_m) \cdot (C - 1)} \tag{3.5}$$

Die Funktion $\mathrm{trace}(\cdot)$ bezeichnet die Spur einer Matrix, also die Summe ihrer Diagonalelemente. Die Matrix der inneren Clusterstreuung wird in Formel 3.6 beschrieben, wobei die Summe über alle Punkte x läuft, die zum i-ten Cluster gehören, und c_i das Zentrum dieses Clusters ist. Die Matrix der Streuung zwischen den Clustern wird in Formel 3.7 dargestellt, wobei n_i die Anzahl der Punkte im i-ten Cluster angibt und k das Zentrum des gesamten Datensatzes ist. Werden die Werte aus den Formeln 3.6 und 3.7 in Gleichung 3.5 eingesetzt, ergibt sich der CHI. [49]

$$W_m = \sum_{i=1}^{c} \sum_{x \in c_i} (x - c_i)(x - c_i)^T \tag{3.6}$$

$$B_m = \sum_{i} n_i (c_i - k)(c_i - k)^T \tag{3.7}$$

Ein hoher CHI deutet darauf hin, dass die Cluster einerseits kompakt und andererseits gut voneinander getrennt sind. Die optimale Clusteranzahl wird daher als der Wert von k interpretiert, für den der CHI maximiert wird. [49]

3.5.3 Silhouette-Score

Der Silhouette-Score bewertet die Qualität einer Clusterlösung, indem er sowohl die Dichte innerhalb der Cluster als auch die Trennung zwischen den Clustern berücksichtigt. Der Silhouette-Koeffizient für einen Datenpunkt i ist in Formel 3.8 definiert, wobei $a(i)$ der durchschnittliche Abstand des Punktes i zu allen anderen Punkten desselben Clusters ist und damit die Kohäsion innerhalb des Clusters misst. Der Wert $b(i)$ bezeichnet den kleinsten durchschnittlichen Abstand des Punktes i zu allen Punkten eines anderen Clusters und misst somit die Trennung von benachbarten Clustern. [49]

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i),\, b(i)\}} \tag{3.8}$$

Als Ergebnis können folgende Fallunterscheidungen gemacht werden.

$$S(i) = \begin{cases} 1 - \dfrac{a(i)}{b(i)}, & \text{falls } a(i) < b(i), \\ 0, & \text{falls } a(i) = b(i), \\ \dfrac{b(i)}{a(i)} - 1, & \text{falls } a(i) > b(i). \end{cases}$$

Der Wert $S(i)$ liegt stets im Intervall $[-1, 1]$. Ein Wert nahe 1 bedeutet, dass der Punkt i gut in sein eigenes Cluster eingebettet ist und klar von anderen Clustern getrennt liegt. Werte nahe 0 deuten darauf hin, dass sich der Punkt an der Grenze zwischen zwei Clustern befindet. Negative Werte weisen darauf hin, dass der Punkt vermutlich dem falschen Cluster zugeordnet wurde. Der Silhouette-Score für das gesamte Clustering ergibt sich als Durchschnitt aller $S(i)$ über sämtliche Datenpunkte. Ein hoher Wert des Silhouette-Scores signalisiert daher eine gute Clusterqualität. [49]

3.6 Regressionsmodelle

Wie in Kapitel 3.3 erläutert wurde, ist Clustering ein geeigentes Verfahren, um Strukturen in Daten zu erkennen und Gruppen mit ähnlichen Eigenschaften zu identifizieren. Es erlaubt, Datensätze auf eine verständliche Weise zu segmentieren und typische Muster sichtbar zu machen. Allerdings ist Clustering primär ein beschreibendes Verfahren, was Ähnlichkeiten und Unterschiede aufzeigt, jedoch keine direkten Vorhersagen über zukünftige Werte oder die Größenordnung bestimmter Merkmale liefert. Zudem basieren viele Clusterverfahren auf normierten Daten, wodurch die ermittelten Cluster zwar vergleichbar, aber oftmals nicht direkt auf konkrete Skalierungen oder absolute Werte übertragbar sind.

An diesem Punkt können Regressionsmodelle einen wesentlichen Beitrag leisten. Während Clustering eine Grundlage zur Strukturierung und Mustererkennung schafft, ermöglichen Regressionsverfahren die Vorhersage kontinuierlicher Zielgrößen [50]. Sie können Abhängigkeiten zwischen Clustern, Datenmerkmalen und externen Einflussfaktoren modellieren und damit die Ergebnisse der Clusteranalyse erweitern. Insbesondere bei stark variierenden Datensätzen ermöglichen Regressionsmodelle, die statische Gruppierung aus dem Clustering zu Nutzen und eine verbesserte Prognosen zu erstellen [50]. Auf diese Weise können

sich beide Methoden ergänzen. Clustering liefert die Struktur und einen vollständigen Datensatz, Regression liefert die Anpassungsfähigkeit und Vorhersagekraft.

3.6.1 Klassische Regressionsmodelle

Die klassische Regression gehört zu den grundlegenden Verfahren der Statistik und des maschinellen Lernens. Die Modelle verfolgen das Ziel, Abhängigkeiten zwischen einer oder mehreren unabhängigen Variablen und einer abhängigen Zielgröße zu modellieren. Besonders etabliert ist die lineare Regression, bei der ein linearer Zusammenhang angenommen wird. Formal beschreibt das Modell den Zusammenhang zwischen einer Zielvariablen y und einer Menge unabhängiger Variablen $x_1, \ldots, x_k$. Dabei wird eine Gerade ermittelt, die den Abstand zwischen den vorhergesagten und den tatsächlich beobachteten Werten minimiert. Für Beobachtungen $i = 1, \ldots, n$ wird angenommen, dass die Zielvariable mit Formel 3.9 beschrieben werden kann, wobei β_0 den Achsenabschnitt, $\beta_1, \ldots, \beta_k$ die Regressionskoeffizienten und ε_i den Fehlerterm darstellen. [50]

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i, \tag{3.9}$$

Für die Fehler wird angenommen, dass sie unabhängig verteilt und im Mittel null sind mit konstanter Varianz σ^2. Die Schätzung der Parameter erfolgt in der Regel mit der Methode der kleinsten Quadrate. In diesem Fall werden jene Werte für $\beta_0, \ldots, \beta_k$ bestimmt, die die Summe der quadrierten Abweichungen zwischen den beobachteten und den vorhergesagten Werten minimieren. [50]
Die klassischen Regressionsverfahren zeichnen sich durch eine hohe Transparenz und geringe Rechenanforderungen aus. Daher sind sie auch für kleine Datensätze und vor allem für Szenarien mit deutlichen Abhängigkeiten geeignet. Die Grenzen hingegen liegen bei stark nichtlinearen, hochdimensionalen oder verrauschten Daten. In diesen Situationen braucht es bessere Techniken, die in den folgenden Abschnitten erklärt werden. [50]

3.6.2 Baumbasierte Regressionsmethoden im maschinellen Lernen

Viele reale Datensätze sind durch nichtlineare Beziehungen oder Ausreißer geprägt, die von linearen Ansätzen nur unzureichend erfasst werden können. Um diesen Herausforde-

rungen zu begegnen, haben sich im Bereich des maschinellen Lernens leistungsfähigere Verfahren etabliert. Besonders hervorzuheben sind dabei die baumbasierten Regressionsmodelle. Die Grundidee besteht darin, den Merkmalsraum rekursiv in immer kleinere Teilbereiche zu zerlegen, sodass innerhalb dieser Teilbereiche die Zielvariable möglichst gut durch einfache Vorhersagen approximiert werden kann. Das entstehende Modell wird als Regressionsbaum bezeichnet. Ausgehend von der Wurzel entscheidet das Verfahren an einem Knoten über eine Bedingung. Diese Splits werden so gewählt, dass sie die Daten im Hinblick auf die Zielgröße möglichst homogen aufteilen. Die Auswahl erfolgt meist durch ein Kriterium wie die Minimierung der mittleren quadratischen Abweichung innerhalb der entstehenden Teilknoten, sodass die Zielwerte innerhalb eines Knotens möglichst homogen sind. Am Ende jedes Zweiges steht ein Blattknoten, in dem eine Vorhersage getroffen wird, meist der Durchschnitt der Zielwerte in diesem Teilbereich. Auf diese Weise können sie komplexe, nichtlineare Muster abbilden und sind zugleich robust gegenüber Ausreißern und fehlenden Werten. Solche Entscheidungsbäume sind leicht interpretierbar, neigen jedoch bei großer Tiefe zur Überanpassung an die Trainingsdaten. [50]

Random Forest

Um diese Schwächen auszugleichen, wurden Ensemble-Methoden entwickelt, die viele Bäume kombinieren und so robustere Modelle erzeugen. Eine der bekanntesten Methoden ist der Random Forest. Das Training eines Random Forest erfolgt nach dem Prinzip des Baggings. Dabei wird jeder Baum auf einer zufälligen Stichprobe der Trainingsdaten trainiert. Diese Stichproben entstehen durch Ziehen mit Zurücklegen, sodass einzelne Datenpunkte mehrfach oder gar nicht in einem Baum vorkommen können. Zusätzlich wird bei jedem Split im Baum nicht auf alle Merkmale zurückgegriffen, sondern nur auf eine zufällige Teilmenge davon. Diese zusätzliche Zufälligkeit sorgt dafür, dass sich die einzelnen Bäume stärker voneinander unterscheiden und die Vorhersagen weniger stark korrelieren. [50]
Die Zielvorhersage erfolgt durch das Zusammenfassen der Ergebnisse aller Bäume. Bei Regressionsaufgaben wird der Mittelwert aller Einzelvorhersagen gebildet.

$$m_{M,n}(\mathbf{x}) = \frac{1}{M} \sum_{j=1}^{M} m_n(\mathbf{x}; \Theta_j, \mathcal{D}_n), \tag{3.10}$$

Mathematisch wird dieses Vorgehen in Gleichung 3.10 dargestellt, wobei $m_n(\mathbf{x}; \Theta_j, \mathcal{D}_n)$ die Vorhersage des j-ten Baums und M die Gesamtzahl der Bäume bezeichnet. Die Varianz wird dabei reduziert und die Stabilität des Modells gleichzeitig erhöht. Bei Klassifikationsaufgaben entscheidet dagegen die Mehrheit der Bäume über das Ergebnis. [51]
Bei einer Anzahl der Bäume M gegen unendlich, ergibt sich die theoretische Definition des Random Forest als Erwartungswert über alle Zufallsparameter Θ [51].

$$m_{\infty,n}(\mathbf{x}) = \mathbb{E}_\Theta[m_n(\mathbf{x}; \Theta, \mathcal{D}_n)]\,. \tag{3.11}$$

Diese Formulierung in Gleichung 3.11 beschreibt den Grenzfall eines unendlich großen Waldes, bei dem die Vorhersage dem Erwartungswert über alle möglichen Zufallsbäume entspricht. In der Praxis zeichnet sich der Random Forest durch seine Robustheit aus, da er im Vergleich zu einzelnen Entscheidungsbäumen weniger zur Überanpassung neigt und komplexe, nichtlineare Zusammenhänge effektiv modellieren kann. Die Nachteile liegen hingegen in der geringeren Interpretierbarkeit im Vergleich zu einem einzelnen Baum und in einem höheren Rechenaufwand, da viele Modelle gleichzeitig trainiert werden müssen. [50, 51]

Gradient Boosting Machine (GBM)

Es gibt aber auch sequentielle Verfahren, wie GBM. Hier werden die Bäume nicht unabhängig voneinander, sondern nacheinander aufgebaut. Jeder neue Baum versucht, die Fehler des bisherigen Modells zu korrigieren, indem er die Residuen, also die Abweichungen zwischen den Vorhersagen und den tatsächlichen Zielwerten, approximiert.
Ziel ist es, eine Näherung $\hat{F}(\mathbf{x})$ der unbekannten Zielfunktion $F^*(\mathbf{x})$ zu finden, indem die erwartete Verlustfunktion $L(y, F(\mathbf{x}))$ minimiert wird. Für ein Trainingsdatenset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ wird das Modell mit der Formel 3.12 schrittweise aufgebaut, wobei $h_m(\mathbf{x})$ den m-ten Entscheidungsbaum und ρ_m ihr Gewicht darstellt. [52]

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \rho_m h_m(\mathbf{x}), \tag{3.12}$$

Zu Beginn startet das Verfahren mit einer konstanten Anfangsapproximation, beispielsweise dem Mittelwert der Zielvariablen bei Verwendung des quadratischen Fehlers als Verlustfunktion. In jeder weiteren Iteration wird eine neue Basisfunktion $h_m(\mathbf{x})$ trainiert, die die Fehler des bisherigen Modells $F_{m-1}(\mathbf{x})$ reduziert. Hierzu werden sogenannte Pseudo-Residuen

berechnet, welche den negativen Gradienten der Verlustfunktion in Bezug auf die Modellvorhersage darstellen. [52]

$$r_{mi} = -\left[\frac{\partial L(y_i, F(\mathbf{x}))}{\partial F(\mathbf{x})}\right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}. \tag{3.13}$$

In Gleichung 3.13 wird somit die Richtung der stärksten Fehlerreduktion bestimmt. Das Modell $h_m(\mathbf{x})$ wird anschließend auf die Residuen trainiert und die optimale Gewichtung ρ_m durch Minimierung des Fehlers berechnet. Nach M Iterationen ergibt sich das finale Modell als Summe aller gewichteten Basisfunktionen. [52]

$$\hat{F}_M(\mathbf{x}) = \sum_{m=0}^{M} \rho_m h_m(\mathbf{x}). \tag{3.14}$$

Die Gleichung 3.14 zeigt die additive Modellstruktur, die sich aus allen zuvor gelernten Entscheidungsbäumen zusammensetzt. Gradient Boosting kann mit unterschiedlichen Verlustfunktionen $L(y, F(\mathbf{x}))$ eingesetzt werden, etwa dem quadratischen Fehler für Regressionsprobleme oder dem logistischen Verlust für Klassifikationsaufgaben. Durch die sequentielle Minimierung der Fehlergradienten entsteht ein leistungsfähiges Modell, das komplexe, nichtlineare Zusammenhänge flexibel abbilden kann. [52]

In Abbildung 3.7 wird der prinzipielle Ablauf des GBM-Algortihmus dargestellt und gezeigt, dass durch diese iterative Fehlerkorrektur schrittweise ein leistungsfähiges Modell entsteht. GBM erzielt oft eine höhere Vorhersagegenauigkeit als Random Forest. Das Modell ist allerdings empfindlicher gegenüber falschen Einstellungen der Hyperparameter, wie etwa der Lernrate oder der maximalen Tiefe der Bäume. Werden diese nicht sorgfältig gewählt, kann das Modell zu komplex werden und die Trainingsdaten überanpassen. Mit der passenden Abstimmung liefert das Modell sehr präzise Ergebnisse und gehört daher heute zu den am häufigsten eingesetzten Methoden im Bereich des maschinellen Lernens, vor allem bei Vorhersageaufgaben mit hohen Genauigkeitsanforderungen. [54]

Moderne Varianten des GBM, wie XGBoost, LightGBM und CatBoost, haben die ursprüngliche Methode weiter verbessert. Sie bauen ebenfalls auf der Grundidee des sequentiellen Lernens von Entscheidungsbäumen auf, nutzen jedoch zusätzliche Optimierungen, um sowohl die Effizienz als auch die Genauigkeit der Modelle zu steigern. Ein zentrales Element dieser Verfahren ist die Regularisierung, also eine Art Bestrafung im Modell, das verhindert,

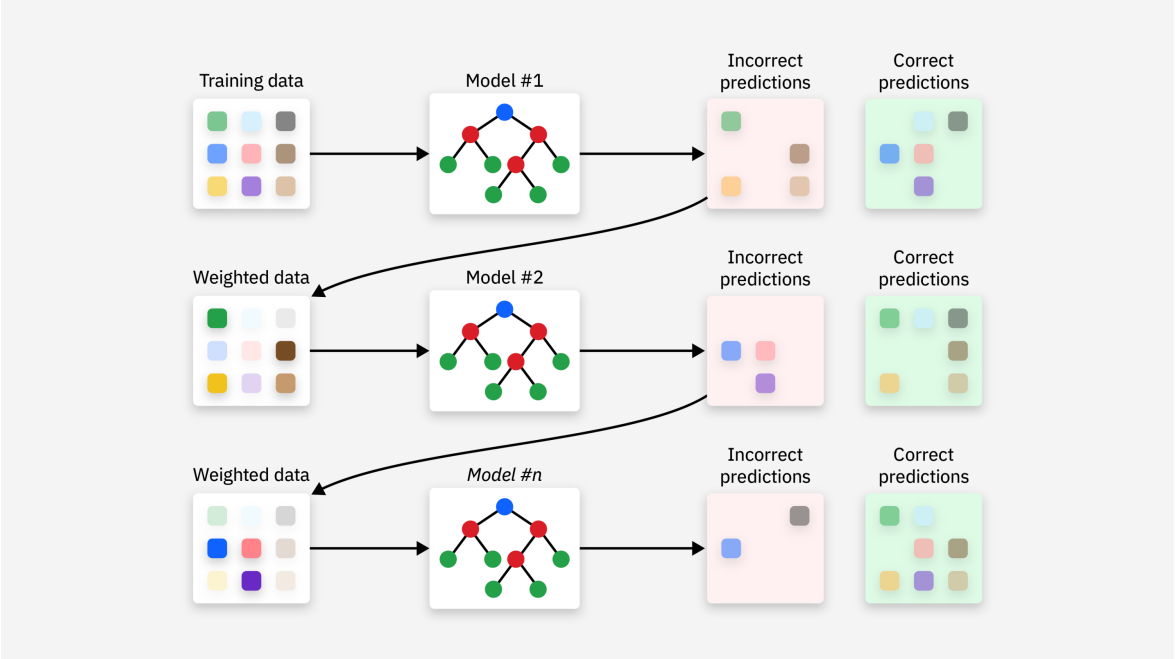


Abbildung 3.7: Funktionsablauf des Gradient-Boosting-Algorithmus aus [53]

dass die Vorhersagen zu stark an die Trainingsdaten angepasst werden. Damit lässt sich Überanpassung wirksam vermeiden, was bei klassischen Boosting-Ansätzen oft eine Herausforderung darstellt [50]. Darüber hinaus setzen diese modernen Implementierungen auf optimierte Splitting-Algorithmen, die den Prozess des Baumaufbaus deutlich beschleunigen und gleichzeitig dafür sorgen, dass die wichtigsten Strukturen in den Daten zuverlässig erkannt werden. Ein weiterer Vorteil ist die parallele Arbeitsweise. Während herkömmliche Boosting-Methoden oft nur schrittweise und damit relativ langsam arbeiten konnten, nutzen XGBoost, LightGBM und CatBoost Möglichkeiten, Berechnungen auf mehrere Kerne oder sogar Rechner zu verteilen. Insbesondere in Bereichen, in denen mit strukturierten Daten gearbeitet wird, also Tabellen mit Merkmalen und Zielwerten, haben sich diese Verfahren etabliert. [55, 56]

3.6.3 Deep Learning als Regressionsmethode

Das Deep Learning stellt eine Erweiterung der klassischen neuronalen Netze dar und zählt zu den zentralen Methoden des modernen maschinellen Lernens. Während Entscheidungsbäume oder klassische Regressionsmodelle primär für tabellarische, strukturierte Daten eingesetzt werden, sind tiefe neuronale Netze insbesondere für große Datenmengen und

komplexe Muster geeignet. Dies betrifft unter anderem zeitabhängige Daten, bei denen die Reihenfolge und Dynamik der Beobachtungen einen maßgeblichen Einfluss auf die Modellgüte haben. [50]

Im Kontext der Lastprognose können Deep-Learning-Modelle als Regressionsverfahren genutzt werden, um den Zusammenhang zwischen erklärenden Merkmalen und historischen Verbrauchsdaten abzubilden [50]. Spezialisierte Architekturen wie Recurrent Neural Networks (RNNs) oder Long Short-Term Memory-Netze (LSTMs) ermöglichen es, zeitliche Abhängigkeiten in den Daten explizit zu modellieren [57]. Dadurch lassen sich wiederkehrende Muster, saisonale Effekte und nichtlineare Beziehungen erfassen, die mit klassischen Verfahren nur eingeschränkt abgebildet werden können [57]. Ergänzend können tiefe Feedforward-Netze eingesetzt werden, wenn große Mengen an Merkmalen und Interaktionen berücksichtigt werden sollen [57].

Ein wesentlicher Vorteil von Deep Learning besteht in der automatischen Merkmalsextraktion. Die Modelle sind in der Lage, aus den Eingabedaten selbstständig repräsentative Strukturen zu lernen, sodass komplexe Abhängigkeiten ohne explizites Feature Engineering erfasst werden können. Demgegenüber stehen Herausforderungen wie eine erhöhte Rechenkomplexität, die Notwendigkeit großer Datenmengen und eine eingeschränkte Interpretierbarkeit der Ergebnisse. Zudem erfordert die Modellierung eine sorgfältige Abstimmung der Architektur und Hyperparameter, um Überanpassung zu vermeiden und eine robuste Generalisierbarkeit zu gewährleisten. [50]

3.7 Vergleich der Regressionsmodelle

Im Vergleich der drei dargestellten Modellklassen wird deutlich, dass jede Methode ihre eigenen spezifischen Stärken und Schwächen besitzt, die bei der Auswahl für die Verbesserung der Lastprognose berücksichtigt werden müssen. In Tabelle 3.2 werden klassische Regressionsverfahren, baumbasierte Methoden und Deep-Learning-Ansätze gegenübergestellt.

Klassische Regressionsverfahren zeichnen sich vor allem durch ihre Einfachheit, Transparenz und die geringen Rechenkosten aus. Sie bieten schnelle Schätzungen und sind statistisch fundiert, weshalb sie häufig als solide Baseline für viele Problemstellungen dienen [50].

Ihre Grenzen liegen jedoch in der Abbildung komplexer oder stark nichtlinearer Zusammenhänge, zudem reagieren sie empfindlich auf Ausreißer und sind stark von Modellannahmen abhängig [50].

Baumbasierte Verfahren, wie Random Forest oder Gradient Boosting, erweitern diese Möglichkeiten deutlich, da sie nichtlineare Abhängigkeiten abbilden können und in der Praxis eine hohe Prognosegenauigkeit bei strukturierten, tabellarischen Daten erreichen [50, 54]. Ein großer Vorteil besteht darin, dass sie relativ robust gegenüber Ausreißern und Rauschen sind und weniger stark auf umfangreiches Feature Engineering angewiesen sind [50]. Allerdings sind sie schwieriger zu interpretieren als klassische Regressionsmodelle, erfordern einen höheren Rechenaufwand bei großen Datensätzen und können durch die Vielzahl an Hyperparametern komplex sein, was insbesondere beim Tuning zeitaufwendig ist [54]. Den-

Modellklasse	**Stärken**	**Schwächen**
Klassische Regression (z.B. Lineare Regression)	• Einfach interpretierbar und transparent • Geringe Rechenkosten, schnelle Schätzung • Gute Baseline für viele Probleme	• Eingeschränkt bei stark nichtlinearen Zusammenhängen • Sensitiv gegenüber Ausreißern • Begrenzte Flexibilität bei komplexen Daten • Abhängigkeit von Modellannahmen
Baumbasierte Verfahren (z.B. Random Forest, Gradient Boosting)	• Abbildung nichtlinearer Zusammenhänge • Hohe Genauigkeit bei strukturierten Daten • Robust gegenüber Ausreißern und Rauschen • Geringer Bedarf an Feature Engineering	• Weniger interpretierbar als klassische Regression • Erhöhter Rechenaufwand bei großen Datensätzen • Hyperparameter-Tuning teilweise aufwendig • Kann bei sehr hoher Komplexität überanpassen
Deep Learning (z.B. Feedforward-Netze, RNNs, LSTMs)	• Sehr hohe Modellkapazität • Automatische Merkmalsextraktion • Besonders geeignet für große Datenmengen • Leistungsfähig bei zeitlichen und komplexen Mustern	• Hoher Rechenaufwand und lange Trainingszeiten • Benötigt große Datenmengen für stabile Ergebnisse • Geringe Interpretierbarkeit der Modelle • Empfindlich gegenüber Hyperparameter-einstellungen

Tabelle 3.2: Stärken und Schwächen verschiedener Modellklassen aus [50, 54, 55, 56, 57]

noch stellen sie einen Mittelweg dar, der einerseits die Vorzüge klassischer Modelle übertrifft und andererseits nicht die extremen Anforderungen von Deep-Learning-Ansätzen hat.
Deep-Learning-Modelle verfügen über eine extrem hohe Modellkapazität und sind in der Lage, automatisch relevante Merkmale zu extrahieren. Sie sind besonders leistungsfähig, wenn sehr große Datenmengen vorliegen oder komplexe, zeitliche und hochdimensionale Muster erkannt werden sollen. Diese Vorteile gehen jedoch mit erheblichen Nachteilen einher. Das Training ist rechenintensiv, zeitaufwendig und erfordert große Datenmengen, um stabile und generalisierbare Ergebnisse zu erzielen. Darüber hinaus sind Deep-Learning-Modelle im Vergleich zu den anderen Ansätzen deutlich schwerer zu interpretieren und stark von der Wahl geeigneter Hyperparameter abhängig. [57]
Insgesamt zeigt die Tabelle, dass klassische Regressionsverfahren vor allem als Baseline bei einfachen Zusammenhängen geeignet sind, während sich Deep-Learning-Ansätze insbesondere für sehr große und hochkomplexe Datenstrukturen eignen. Baumbasierte Regressionsmodelle stellen hingegen einen ausgewogenen Mittelweg dar, da sie Robustheit, Flexibilität und eine hohe Prognoseleistung mit einem moderaten Bedarf an Daten und Rechenressourcen verbinden. Besonders relevant ist in diesem Zusammenhang, dass für die vorliegende Problemstellung bereits eine Clusterstruktur der Daten existiert. Diese kann durch baumbasierte Modelle gezielt optimiert werden, da sie in der Lage sind, vorhandene Muster aufzunehmen und um zusätzliche, nichtlineare Abhängigkeiten zwischen den Merkmalen zu ergänzen. Dadurch lassen sich die bestehenden Cluster nicht nur stabilisieren, sondern auch inhaltlich erweitern, was zu einer Steigerung der Vorhersagequalität führen kann. [58]
Als Regressionsmethode wird in dieser Arbeit ein Gradient Boosting Regressor eingesetzt. Die Entscheidung für dieses Verfahren beruht auf den Eigenschaften der zugrunde liegenden Daten sowie auf den Anforderungen an die Prognosegüte. Die Datenbasis liegt in tabellarischer Form vor und umfasst kontinuierliche Variablen. Solche strukturierten Matrizen können vom Algorithmus direkt verarbeiten werden, ohne dass komplexe Vorverarbeitungsschritte notwendig sind [58]. Gleichzeitig ist GBM in der Lage, nichtlineare Abhängigkeiten zwischen den Spalten zu modellieren und Interaktionen zwischen verschiedenen Merkmalen automatisch zu berücksichtigen [58]. Zudem ist diese Regressionsmethode im Vergleich zu rein baumbasierten Ansätzen wie Random Forest weniger anfällig für übermäßig raue

Prognosen und liefert glattere Verläufe, was im Kontext von Spitzenlasten von Vorteil ist. Im Gegensatz zu Deep-Learning-Ansätzen erfordert GBM keine großen Datenmengen und keine tiefgreifende Hyperparameteroptimierung. Bei tabellarisch strukturierten Datensätzen zeigen sich neuronale Netze häufig nicht überlegen, während Boosting-Verfahren in der Regel effizienter trainieren, datenökonomischer arbeiten und eine bessere Interpretierbarkeit bieten [58].

3.8 Evaluation

Für die Bewertung der Clusterzuordnung sowie der anschließenden Regressionsprognosen werden in dieser Arbeit verschiedene Gütemaße herangezogen, die sowohl die Qualität des Clusterings als auch die Genauigkeit der numerischen Vorhersagen erfassen.

Mean Absolute Error

Der Mean Absolute Error (MAE) ist eine der am häufigsten verwendeten Metriken zur Evaluation von Regressionsmodellen [59]. Diese Metrik beschreibt die mittlere absolute Abweichung zwischen tatsächlichen Werten y_i und Prognosen $\hat{y}_i$ und wird in Formel 3.15 definiert [60].

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{3.15}$$

Der Vorteil des MAE liegt in seiner leichten Interpretierbarkeit, da dieser direkt in der Einheit der Zielvariablen angegeben wird und so unmittelbar verständlich macht, wie groß die durchschnittliche Abweichung ist [60]. Darüber hinaus ist diese Metrik robust gegenüber Ausreißern, da große Fehler nicht zusätzlich durch Quadrierung verstärkt werden. Somit eignet sich der MAE besonders gut, wenn eine gleichmäßige Gewichtung aller Abweichungen erwünscht ist und keine überproportionale Bestrafung von Extremwerten angestrebt wird. Der MAE ist zudem intuitiv interpretierbar, da er die mittlere Fehlergröße in den ursprünglichen Einheiten der Last wiedergibt [59].

Root Mean Squared Error

Der Root Mean Squared Error (RMSE) ähnelt dem MAE, gewichtet jedoch größere Abweichungen aufgrund der Quadrierung der Fehler stärker [59]. Mathematisch wird diese Methode in Formel 3.16 definiert und ergibt sich, indem die Differenzen zwischen Vorhersage $\hat{y}_i$ und Realität y_i quadriert, gemittelt und anschließend die Quadratwurzel gezogen wird [60].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.16}$$

Der RMSE betont große Fehler und gewichtet sie deutlich stärker als kleine, wodurch er besonders in Anwendungsfällen sinnvoll ist, in denen große Abweichungen vermieden werden sollen [59]. Ähnlich wie der MAE ist auch der RMSE in derselben Einheit wie die Zielvariablen angegeben, was seine Verständlichkeit erleichtert [59]. Da dieser jedoch immer größer oder gleich dem MAE ist, wird er oft als konservativere Kennzahl angesehen, die das Modell strenger bewertet [59].

Cosine Similarity

Zur Analyse der Clusterzuordnung wird die Cosine Similarity eingesetzt. Die Cosine Similarity ist keine klassische Fehlermetrik, sondern ein Maß zur Bewertung der Ähnlichkeit zweier Vektoren. Diese Methode misst die Winkelähnlichkeit zwischen zwei Vektoren A und B im Merkmalsraum und wird in Formel 3.17 definiert [61].

$$\cos(\alpha) = \frac{A \cdot B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{3.17}$$

Der resultierende Wert liegt zwischen -1 und 1, wobei ein Wert von 1 eine perfekte Übereinstimmung, ein Wert von 0 keine Ähnlichkeit und ein Wert von -1 eine entgegengesetzte Ausrichtung bedeutet. Im Gegensatz zu MAE oder RMSE bewertet die Cosine Similarity nicht die numerische Nähe von Vorhersagen zu Zielwerten, sondern die strukturelle Nähe im Merkmalsraum. Damit eignet sie sich besonders, um die inhaltliche oder semantische Ähnlichkeit von Datenpunkten zu erfassen. Ein weiterer Vorteil ist, dass die Länge der Vektoren keine Rolle spielt, sondern lediglich ihre Richtung, was die Metrik unempfindlich gegenüber Skalierungen macht. [61]

Pearson-Korrelation

Zur Bewertung der Formähnlichkeit zwischen der tatsächlichen und der rekonstruierten Lastkurve wird die Pearson-Korrelation herangezogen. Sie misst den Grad der linearen Übereinstimmung zwischen zwei zeitabhängigen Größen und ist damit ein geeignetes Maß, um Ähnlichkeiten in den zeitlichen Verbrauchsmustern zu quantifizieren. Damit lässt sich beurteilen, ob die rekonstruierten Lastkurven nicht nur im Mittelwert, sondern auch im Verlauf mit den realen Profilen übereinstimmen. Für den tatsächlichen Wert x_i und den rekonstruierten Wert y_i wird der Korrelationskoeffizient ρ_p nach Gleichung 3.18 berechnet. [62]

$$\rho_p = \frac{\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)}{N \cdot \sqrt{\mathrm{var}(\mathbf{X})}\,\sqrt{\mathrm{var}(\mathbf{Y})}} = \frac{\mathrm{cov}(\mathbf{X}, \mathbf{Y})}{\sigma_X \sigma_Y} \tag{3.18}$$

Hierbei bezeichnen μ_X und μ_Y die Mittelwerte, $\mathrm{var}(\mathbf{X})$ und $\mathrm{var}(\mathbf{Y})$ die Varianzen sowie $\mathrm{cov}(\mathbf{X}, \mathbf{Y})$ die Kovarianz der jeweiligen Größen. Der Korrelationskoeffizient ρ_p kann Werte zwischen -1 und 1 annehmen, wobei $\rho_p = 1$ eine perfekt positive lineare Beziehung, $\rho_p = -1$ eine vollständig negative Beziehung und $\rho_p = 0$ das Fehlen eines linearen Zusammenhangs beschreibt. In diesem Kontext signalisiert ein hoher positiver Korrelationswert, dass die modellierte Lastkurve den zeitlichen Verlauf der gemessenen Kurve zuverlässig reproduziert, auch wenn absolute Werte abweichen können. [62]

4 Methodische Vorgehensweise

In diesem Kapitel wird die methodische Vorgehensweise erläutert, die der vorliegenden Arbeit zugrunde liegt. In Kapitel 4.1 wird das allgemeine Vorgehen beschrieben, mit dem auf der Grundlage vorhandener Fahrplandaten und Messwerte Prognosen für das Lastverhalten unbekannter Haushalte abgeleitet werden. Hierfür wird ein zweistufiger Ansatz verfolgt. Im ersten Schritt wird der zugrunde liegende Datensatz sowie die Modellierung des Clustering-Verfahrens in Kapitel 4.2 beschrieben. Diese dienen der Identifikation der charakteristischen Verbrauchs- und Erzeugungsprofile aus den heterogenen Fahrplandaten. Haushalte mit ähnlichen technischen, netzseitigen und verhaltensbezogenen Merkmalen werden dabei zu Clustern zusammengefasst, deren repräsentative Profile als Grundlage für Haushalte ohne eigene Planungsdaten dienen. Aufbauend auf diesen Clustern wird im zweiten Schritt die Modellierung der Regressionsmethode in Kapitel 4.3 vorgestellt. Das Ziel dieser Schritte ist es, aus den zuvor abgeleiteten Profilen sowie weiteren verfügbaren Einflussgrößen eine belastbare Lastprognose zu erstellen. Dieses Modell bildet die methodische Basis für die anschließende Koordination von Flexibilitäten durch die KOF.

4.1 Vorgehen

Der Schwerpunkt dieser Arbeit liegt auf drei zentralen Aspekten. Zunächst wird das Clustering der bekannten Haushalte durchgeführt, um charakteristische Lastprofile zu identifizieren und typische Verbrauchsmuster innerhalb des Gesamtdatensatzes abzubilden. Der Ablauf des Clusteringalgorithmus wird schematisch in Abbildung 4.1 zusammengefasst. Um die Trennschärfe der Clusterbildung und die repräsentativen Lastprofile zu erhöhen, wird diesem Schritt ein vorgelagertes Preisclustering hinzugefügt. Dabei werden die Haushalte nach ihrer Preiskategorie gruppiert, wobei zwischen Festpreishaushalten und Haushalten

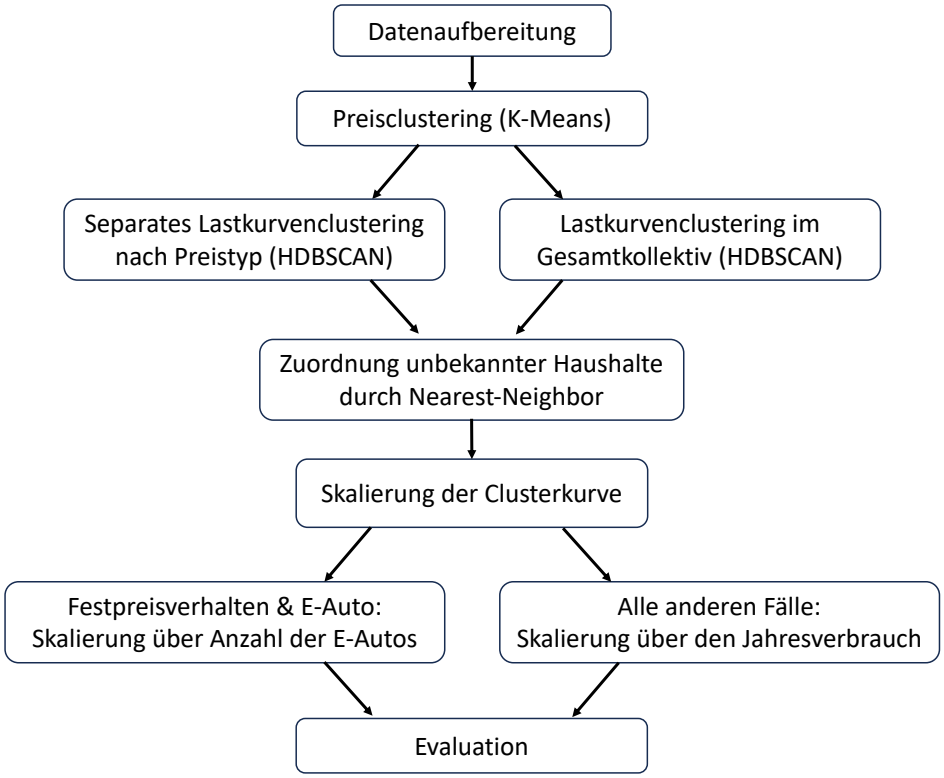


Abbildung 4.1: Schematischer Aufbau des Modells vom Clustering

mit variablem, zeitabhängigem Preisverhalten unterschieden wird. Diese Vorstrukturierung ermöglicht eine differenziertere Abbildung des Verbrauchsverhaltens, da sich Preissignale nachweislich auf die Lastverteilung und das Flexibilitätsverhalten der Haushalte auswirken. Innerhalb der so entstandenen Preisgruppen erfolgt anschließend das Lastkurven-Clustering mithilfe des Verfahrens HDBSCAN, um die unterschiedlichen Verbrauchs- und Erzeugungsprofile zu identifizieren.

Auf Grundlage einer Zuordnung wird für Haushalte, die keinen Fahrplan übermitteln, eine Clusterprognose erstellt. Ziel ist es, deren erwartete Last für den kommenden Tag möglichst zuverlässig zu schätzen und so eine vollständige Datengrundlage für die anschließende Flexibilitätskoordination zu schaffen.

Im zweiten Schritt wird diese Clusterprognose durch den Einsatz zusätzlicher historischer Messdaten und Kontextinformationen erweitert. Das Gradient Boosting Regressionsmodell soll die Prognosegüte weiter verbessern, indem es nichtlineare Zusammenhänge zwischen Verbrauch, Umweltbedingungen und tageszeitlichen Strukturen abbildet. Insgesamt ergibt sich ein hybrider Ansatz, welcher schematisch in Abbildung 4.2 dargestellt wird.

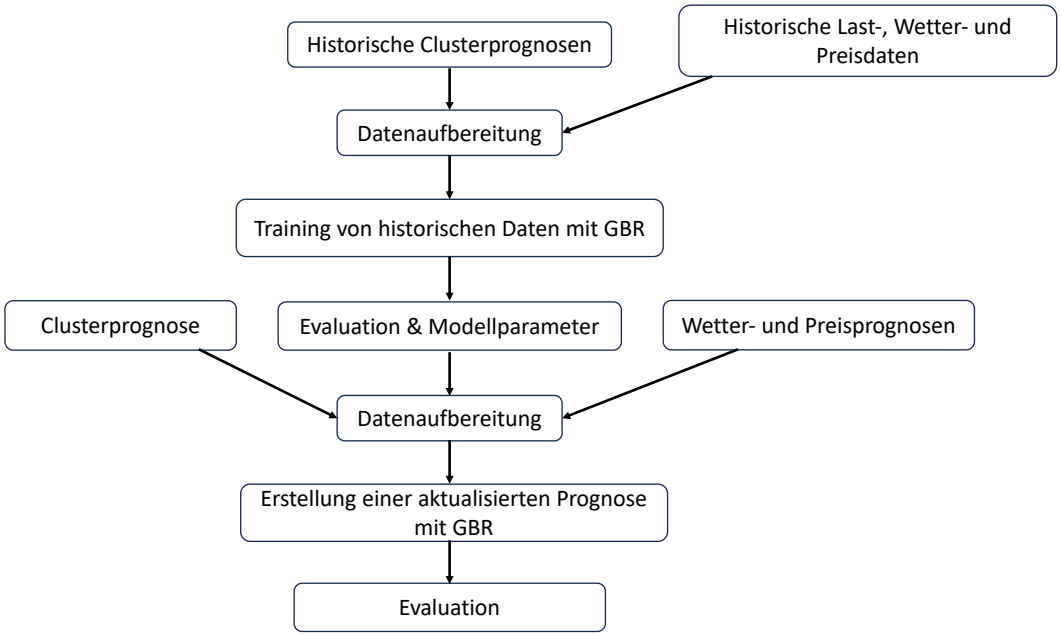


Abbildung 4.2: Schematischer Aufbau des Regressionsmodells pro Haushalt

Abschließend wird die resultierende Prognose mit einer Referenzmethode verglichen, die unter idealisierten Annahmen eine perfekte Vorhersage liefert. Auf diese Weise lässt sich quantitativ bewerten, wie präzise die entwickelte Prognosemethode arbeitet und in welchem Ausmaß die darauf basierende Koordination von Flexibilitäten durch die KOF potenzielle Engpässe in der Lastverteilung vermeiden kann.

Zur Beurteilung der Robustheit und Generalisierbarkeit der Ansätze werden die beschriebenen Szenarien unter unterschiedlichen Ausgangsbedingungen analysiert. Dabei wird der Anteil der bekannten Haushalte sukzessiv verringert. Es wird gestartet mit einem Anteil von 70%, dann wird der Anteil auf 50% und abschließend auf 30% verringert, während das Verhältnis von dynamischen Preishaushalten zu Festpreishaushalten konstant 50% beträgt. Dadurch können sowohl die Auswirkungen des Anteils bekannter Haushalte auf die Prognosegüte als auch die Stabilität der entwickelten Verfahren unter verschiedenen Datensituationen systematisch untersucht werden.

4.2 Clustering

Die Grundlage des Clusterings sind die bekannten Lastverläufe der Haushalte, die einen Fahrplan übermitteln. Wie in Kapitel 2.3 erläutert, steht in dieser Arbeit nur für einen Teil der

Haushalte ein vollständiger Fahrplan zur Verfügung. Diese Fahrpläne werden genutzt, um ähnliche Verbrauchsprofile zu gruppieren und charakteristische Lastmuster zu identifizieren.

4.2.1 Preis-Clustering

Parallel zu den Lastdaten werden Preisdaten für den Prognosezeitraum eingebunden, sodass die Korrelation zwischen Verbrauch und Strompreis berechnet werden kann. Daraus lassen sich Kennzahlen ableiten wie die Stärke des Preisbezugs, der durchschnittliche Verbrauch in Hoch- und Tiefpreiszeiten und die Empfindlichkeit gegenüber Preisänderungen. Damit kann ein optionales, vorgeschaltetes Preisclustering realisiert werden, um die bekannten Haushalte zunächst in Gruppen mit ähnlichem Preisverhalten einzuordnen. So wird eine erste Segmentierung geschaffen, die es ermöglicht, Unterschiede zwischen Haushalten mit variablen und Festpreistarifen zu berücksichtigen. Damit soll transparent gemacht werden, ob ein Haushalt auf Preismotivationen reagiert oder ob er unabhängig vom Preis Lasten zu- oder abschaltet.

Die Tabelle 4.1 zeigt beispielhaft eine Übersicht der abgeleiteten preis- und lastbezogenen Merkmale pro Haushalt. Jede Zeile repräsentiert einen einzelnen Haushalt, während die Spalten verschiedene Kennzahlen enthalten, die sowohl aus den Lastkurven als auch aus den Preisdaten berechnet werden. Die Korrelation zwischen Strompreis und Last dient als Indikator für das Reaktionsverhalten eines Haushalts auf Preisänderungen. Ein negativer Wert weist darauf hin, dass der Haushalt bei hohen Preisen tendenziell weniger Strom

HH	Korr.	Leistung Hoch-preis [kW]	Leistung Tief-preis [kW]	Ratio Hoch / Tief	Std	Peak Mean	Peak Zeit-punkt	Leistung morgens [kW]	Leistung abends [kW]	Ratio
H2	-0.42	0.85	1.10	0.77	0.24	1.48	28	0.93	1.05	1.13
H4	0.10	1.02	0.98	1.04	0.19	1.31	34	0.88	0.92	1.05
H5	-0.25	0.90	1.00	0.90	0.22	1.39	41	0.91	1.02	1.12
H9	0.02	1.05	1.03	1.02	0.17	1.28	37	0.86	0.89	1.03
H10	-0.30	0.88	1.15	0.77	0.25	1.55	26	0.92	1.20	1.30
...	...	...	...	...	...	...	...	...	...	...
H104	0.05	0.95	0.93	1.02	0.21	1.40	31	0.90	0.91	1.01

Tabelle 4.1: Schematischer Aufbau der Datenmatrix für das Preis-Clustering auf Grundlage der bekannten Haushalte

verbraucht, während ein positiver Wert auf ein gegenläufiges Verhalten hindeutet. Zur Ermittlung der Hoch- und Tiefpreiszeiten werden die Preiswerte zunächst in Quartile unterteilt. Zeitpunkte, deren Strompreise oberhalb des 75. Perzentils liegen, werden als Hochpreisphasen definiert, während Zeitpunkte, deren Strompreise unterhalb des 25. Perzentils liegen, den Tiefpreisphasen zugeordnet werden. Für jeden Haushalt wird anschließend der durchschnittliche Verbrauch innerhalb dieser Preisbereiche berechnet. Dadurch ergeben sich Merkmale, die den durchschnittlichen Leistungsverbrauch eines Haushalts während teurer beziehungsweise günstiger Preisphasen abbilden. Aus dem Verhältnis dieser Werte lässt sich ableiten, ob ein Haushalt bei hohen Preisen seinen Verbrauch reduziert oder weitgehend konstant hält. Ergänzend beschreibt die Standardabweichung die zeitliche Schwankung des Verbrauchs, während der mittlere Spitzenwert sowie der Zeitpunkt des maximalen Verbrauchs charakteristische Informationen über das Lastverhalten liefern. Darüber hinaus werden typische Verbrauchsniveaus in den Morgen- und Abendstunden betrachtet, indem das Verhältnis dieser Verbrauchsniveaus berechnet wird. So können beide Zeitfenster direkt verglichen werden und es kann zwischen tendenziell morgendlichen oder abendlichen Lastprofilen unterschieden werden. Diese Merkmale bilden die Grundlage für das anschließende Preis-Clustering, da sie zentrale Informationen über das Verbrauchs- und Preisreaktionsverhalten enthalten.

Zur Identifikation der Tariftypen wird ein Partitionierungsverfahren eingesetzt, da das Preisverhalten das Verbrauchs- und insbesondere das Ladeverhalten stark beeinflusst. Während variable Tarife Anreize zur Lastverschiebung setzen, führen Festpreistarife häufig zu ungesteuerten Ladevorgängen und ausgeprägten Lastspitzen. Beides erfordert unterschiedliche Betrachtungsweisen im Clustering und in der Skalierung. Da sich variable und Festpreistarife grundsätzlich im Verbrauchsverhalten unterscheiden, kann die Kategorisierung mit einer vorgegebenen Clusterzahl von zwei durchgeführt werden. Wie in Kapitel 3.4 thematisiert wurde, eignet sich für diesen Anwendungsfall das K-Means-Verfahren besonders gut, da es eine vorher festgelegte Anzahl von Clustern effizient trennen kann. Auf Grundlage der Korrelation zwischen Preis- und Belastungsdaten kann das Preisclustering durchgeführt werden, wodurch bereits vor der weiteren Analyse zwischen variablen und Festpreistarifen unterschieden wird. Das Verfahren ist effizient, leicht interpretierbar und erlaubt durch die Berechnung von Clusterzentren eine direkte Vergleichbarkeit mit den bekannten Tariflabels.

Zur Visualisierung wird eine zweidimensionale Hauptkomponentenanalyse PCA eingesetzt, in der die hochdimensionalen Verhaltensmerkmale in ein anschauliches Koordinatensystem projiziert werden.

Da sich Haushalte mit Festpreistarifen häufig anders verhalten als solche mit variablen Preismodellen, wird geprüft, ob das zuvor durchgeführte Preisclustering für die weitere Analyse geeignet ist oder ein Clustering auf dem vollständigen Haushaltskollektiv durchgeführt werden sollte. Unterschiede im Preisverhalten wirken sich direkt auf das Lastprofil aus, da variable Preise ein flexibleres und stärker preisgesteuertes Verbrauchsverhalten begünstigen, während Festpreiskunden tendenziell ein individuelleres Nachfrageverhalten zeigen. Werden diese Cluster nicht getrennt betrachtet oder entsteht ein Mischcluster aus beiden Gruppen, kann dies zu verzerrten Ergebnissen führen. Die Cluster bilden dann nicht mehr homogene Verbrauchsmuster ab, sondern überlagern unterschiedliche Verhaltensmotive. Dadurch sinkt sowohl die interpretative Aussagekraft als auch die Stabilität der Cluster. Die Entscheidung, ob die Ergebnisse des Preisclusterings verwendetet werden können, erfolgt datenbasiert anhand mehrerer Kriterien, die sowohl die Qualität als auch die Stabilität der Clusterbildung einbeziehen. Zunächst wird die Mindestgröße der Cluster überprüft, um sicherzustellen, dass jedes Cluster eine ausreichende Anzahl von Haushalten enthält. Kleine Cluster mit sehr wenigen Elementen weisen eine hohe Varianz und geringe Repräsentativität auf, was zu instabilen oder zufälligen Gruppierungen führen kann. Durch die Festlegung einer minimalen Clustergröße wird somit gewährleistet, dass nur ausreichend große und damit statistisch aussagekräftige Cluster berücksichtigt werden. Darüber hinaus wird das Größenverhältnis der Cluster analysiert. Ein stark unausgewogenes Verhältnis zwischen den Clustern würde die Aussagekraft der Ergebnisse einschränken. Ein solches Ungleichgewicht kann darauf hinweisen, dass keine klar trennbaren Gruppen im Datenraum existieren und das Preisclustering somit keine sinnvolle Segmentierung liefert. Wird eine dieser Bedingungen nicht erfüllt, so wird auf das vollständige Clustering zurückgegriffen. Dadurch werden künstliche oder wenig aussagekräftige Trennungen vermieden und die Robustheit der Gesamtanalyse sichergestellt. Nur wenn beide Cluster ausreichend groß und ausgewogen sind, wird das Ergebnis des Preisclusterings für die weiteren Analyseschritte übernommen.

4.2.2 Lastkurven-Clustering

Nach der Preissegmentierung werden die Haushalte auf Basis ihrer Lastverläufe gruppiert. Dazu werden die Daten so transformiert, dass die Haushalte die Zeilen und die Zeitpunkte die Spalten bilden. In Tabelle 4.2 ist der daraus resultierende Aufbau beispielhaft dargestellt. Jede Zeile entspricht einem Haushalt H_i, und jede Spalte $t = 1, \ldots, 96$ repräsentiert die jeweilige, auf 15-Minuten-Intervalle bezogene Leistung innerhalb eines Tages. Dadurch entsteht eine Datenmatrix, in der jede Zeile den vollständigen zeitlichen Lastverlauf eines Haushalts abbildet. Die Lastdaten werden mit dem $MaxAbsScaler$ normiert. Dieses Verfahren teilt alle Werte durch den größten Absolutwert innerhalb jeder Spalte, sodass sich die Werte im Bereich zwischen -1 und 1 befinden. Dabei bleibt die Form der Lastkurve vollständig erhalten, lediglich die Höhe der Kurve wird angepasst. Das ist besonders wichtig, weil die Lastprofile unterschiedlicher Haushalte unterschiedliche Größenordnungen aufweisen können, die Kurven aber trotzdem miteinander vergleichbar sein sollen. Durch diese Normierung werden insbesondere Unterschiede im Verlauf der Leistung hervorgehoben, jedoch nicht in der absoluten Höhe des Verbrauchs. Neben den zeitlich aufgelösten Leistungswerten enthält die Tabelle auch weitere binäre Merkmale, wie das Vorhandensein einer PV-Anlage, einer Batterie, einer Wärmepumpe oder eines BEV. Diese Zusatzinformationen dienen dazu, mögliche Zusammenhänge zwischen technologischer Ausstattung und typischen Verbrauchsmustern zu erkennen. Damit kann bestimmt werden, welche Flexibilität ein Haushalt grundsätzlich bereitstellen kann. In der Tabelle wird exemplarisch eine Aus-

Haushalt	skalierte Leistung t=1	skalierte Leistung t=2	...	skalierte Leistung t=96	PV	Batterie	Wärme-pumpe	BEV
H2	0.82	0.90	...	0.88	1	1	0	1
H4	0.60	0.58	...	0.65	0	0	1	0
H5	0.91	0.93	...	0.89	1	1	1	1
H9	0.72	0.74	...	0.79	1	1	0	0
H10	0.75	0.77	...	0.81	0	0	1	0
H11	0.70	0.68	...	0.73	0	0	0	1
...	...	...	...	...	...	...	...	...
H104	0.78	0.80	...	0.84	1	1	0	1

Tabelle 4.2: Schematischer Aufbau der Datenmatrix für das Lastkurven-Clustering auf Grundlage der bekannten Haushalte

wahl bekannter Haushalte dargestellt, um eine klare Abgrenzung gegenüber unbekannten Lastverläufen zu verdeutlichen. Da der Anteil der bekannten Haushalte variabel gestaltet ist, wirkt sich die Auswahl der betrachteten Daten unmittelbar auf die Zielrichtung und die methodische Grundlage des Clustering-Prozesses aus.
Für die Zuordnung der Haushalte ohne gemeldeten Fahrplan auf die gelernten Cluster wird neben den erwähnten technischen Merkmalen auch das Preisverhalten der Haushalte eingelesen. So wird berücksichtigt, welchem Tarifmodell ein Haushalt zugeordnet ist, wodurch sich das Lastverhalten im Kontext von Preissignalen bewerten lässt. Auf Basis dieser Segmentierung können auch Haushalte ohne gemeldeten Fahrplan einem geeigneten Cluster zugewiesen und damit mit einem typischen Lastprofil abgebildet werden.
Für die Clusteranalyse der Lastprofile wird ein zweistufiges Verfahren eingesetzt. Zunächst wird UMAP zur Dimensionsreduktion genutzt, anschließend HDBSCAN zur Clusterbildung. Wie in Kapitel 3.2 erläutert, eignet sich UMAP, weil es lokale Strukturen in hochdimensionalen Daten bewahrt und dabei nicht-lineare Beziehungen berücksichtigt werden, was ein Vorteil bei der Analyse von komplexen Lastprofilen ist. Im Anschluss an die Dimensionsreduktion erfolgt die Clusterbildung mit HDBSCAN. Die Mindestgröße der Cluster wird auf fünf Datenpunkte festgelegt, um sicherzustellen, dass nur ausreichend große Gruppen als eigenständige Cluster betrachtet werden und kleine, zufällige Lastkurven ausgeschlossen bleiben. Durch die Kombination von UMAP und HDBSCAN entsteht somit ein datenbasierter Ansatz, der sowohl lokale Muster als auch globale Zusammenhänge im Lastverhalten erfasst. Zur Bewertung der Qualität der resultierenden Cluster werden die in Kapitel 3.5 beschriebenen Metriken, der Silhouette-Score, der Davies–Bouldin-Index und der Calinski–Harabasz-Index, herangezogen. Diese Kennzahlen ermöglichen eine objektive Beurteilung der Trennschärfe, Homogenität und Kompaktheit der gebildeten Cluster und dienen somit der Validierung der gewählten Parameter und der Aussagekraft der gebildeten Gruppen.

Zuordnung der Haushalte zu den Clustern

Unbekannte Haushalte werden den Clustern durch den KNN-Algorithmus anhand der technischen Ausstattung zugewiesen. In dieser Arbeit wird $k = 1$ gewählt, da für jeden unbekannten Haushalt ausschließlich der jeweils ähnlichste bekannte Haushalt im Merkmalsraum als

Referenz herangezogen werden soll. Somit nutzt dieses Vorgehen die Ähnlichkeit in ausgewählten Haushaltsmerkmalen und überträgt die Clusterzugehörigkeit des nächsten bekannten Nachbarns. Ergänzend wird die Konsistenz durch die Berechnung des Cosine-Similarity zwischen Haushaltsmerkmalen und Clusterprofilen überprüft.

Saklierung der Clusterkurven

Zur Interpretation werden für jedes Cluster repräsentative Lastkurven gebildet, indem die Durchschnittslast zu den einzelnen Zeitpunkten über alle Haushalte eines Clusters aggregiert wird. Dadurch werden Einzelschwankungen geglättet und typische Muster hervorgehoben. Um Vergleiche auf Haushaltsebene interpretieren zu können, werden die repräsentativen Kurven skaliert. Die Skalierung der Clusterkurven erfolgt so, dass die repräsentative Kurve eines Clusters auf das jeweilige Haushaltsniveau angepasst wird. Im Standardfall, insbesondere bei variablen Tarifen, wird dafür der Jahresverbrauch des einzelnen Haushalts E_h ins Verhältnis zum durchschnittlichen Jahresverbrauch aller Haushalte $\bar{E}_C$ im Cluster gesetzt. Dieses Vorgehen wird mit Formel 4.1 beschrieben, wobei $L_h(t)$ die Lastkurve des Haushalts h und $\bar{L}_C(t)$ die gemittelte Clusterkurve zum Zeitpunkt t bezeichnet.

$$\tilde{L}_C(t,h) = \bar{L}_C(t) \cdot \frac{E_h}{\bar{E}_C} \tag{4.1}$$

Auf diese Weise stimmt die skalierte Clusterkurve im Energiegehalt mit dem echten Lastprofil überein. Bei Festpreis-Haushalten wird zusätzlich berücksichtigt, dass diese keine preislichen Anreize haben, ihr Ladeverhalten zeitlich zu verschieben. Es wird angenommen, dass Elektroautos daher häufig ohne Steuerung geladen werden, was zu deutlichen Spitzenlasten führt. Insbesondere bei einer steigenden Zahl von Elektroautos wären diese Peaks für die Netzbelastung entscheidend und können nicht über den Jahresverbrauch abgebildet werden. Um diese Ladespitzen dennoch abbilden zu können, wird bei Festpreis-Haushalten mit Elektroauto eine alternative Skalierung anhand eines typischen Spitzenwerts durchgeführt, wie in Gleichung 4.2 dargestellt.

$$\tilde{L}_C(t,h) = \bar{L}_C(t) \cdot \frac{P_C}{L_{C,\max}} \tag{4.2}$$

Hierbei steht P_C für einen repräsentativen Spitzenwert ähnlicher Haushalte mit Elektrofahrzeug im selben Cluster, während $L_{C,\max}$ die maximale Last der unskalierten Clusterkurve

bezeichnet. Der Spitzenwert P_C wird als Median der maximalen Lasten dieser Haushalte bestimmt, da der Median, im Gegensatz zum arithmetischen Mittel, unempfindlicher gegenüber Ausreißern ist und somit den typischen Lastpeak robuster beschreibt. Wenn in einem Cluster keine ausreichende Anzahl solcher Vergleichshaushalte vorhanden ist, wird die Standard-Skalierung gemäß Gleichung 4.1 verwendet. Dadurch wird gewährleistet, dass einerseits das jährliche Verbrauchsniveau, andererseits aber auch die für Festpreis-Haushalte charakteristischen Lastspitzen angemessen berücksichtigt werden.

Evaluation

Der Abgleich zwischen den individuellen Lastkurven der Haushalte und den repräsentativen Clusterkurven erfolgt über die thematisieren Fehler- und Ähnlichkeitsmaße in Kapitel 3.8. Mit dem MAE und RMSE wird die durchschnittliche Differenz zwischen den beiden Verläufen bewertet. Ergänzend dazu wird die Pearson-Korrelation herangezogen, die die Formähnlichkeit zwischen den Kurven unabhängig vom absoluten Niveau beschreibt. Durch die Kombination dieser Kennzahlen lässt sich nicht nur die durchschnittliche Passung innerhalb eines Clusters beurteilen, sondern auch die Eignung der skalierten repräsentativen Kurven für Prognosezwecke einschätzen.

4.3 Regressionsanalyse

Um die Prognosegüte der Clusterzuordnung zu erhöhen, wird der ursprüngliche Datensatz systematisch um weitere Informationsquellen und Merkmale erweitert und für eine Regressionsanalyse aufbereitet. Während das ursprüngliche Clustering vor allem auf den Fahrplänen der Haushalte und deren zeitlich normierten Lastprofilen beruht, wird durch die Einbindung zusätzlicher Datenquellen eine vielseitigere Datenbasis geschaffen.
Ein zentraler Bestandteil dieser Erweiterung sind die historischen Lastmessungen der Haushalte, die als absolute Netzlast in Kilowatt [kW] vorliegen und die Zielgröße der Regressionsmodelle bilden. Sie stellen damit die zu prognostizierende Größe dar und sind in Tabelle 4.3 exemplarisch aufgeführt. Ergänzend wird die aus dem Clustering abgeleitete Grundprognose als zusätzliches Merkmal in das Regressionsmodell integriert. Diese beschreibt das typische, auf den Clusterzuordnungen basierende Lastverhalten eines Haushalts und dient

der besseren Erfassung der erwarteten Verbrauchsstruktur. So wird vermieden, dass Spitzenlasten durch eine zu starke Glättung der Zielwerte verloren gehen, während zugleich ein stabiler Referenzverlauf als erklärende Eingangsgröße zur Verfügung steht.

Darüber hinaus werden Preisdaten für den Prognosezeitraum integriert, um auch dynamische Effekte variabler Stromtarife abzubilden. Aus den absoluten Strompreisen in Euro pro Megawattstunde [€/MWh] werden tagesbezogene Kontextmerkmale abgeleitet, die die relative Stellung eines Preises im zeitlichen Verlauf eines Tages beschreiben. Der relative Preis setzt den momentanen Strompreis ins Verhältnis zum jeweiligen Tagesmittelwert und zeigt damit an, ob der aktuelle Zeitpunkt eher über- oder unterdurchschnittlich teuer ist. Der Preisrang ordnet jeden Preiswert innerhalb eines Tages auf einer Skala von 0 bis 1 ein, wobei 0

Zeit	HH	**Ziel:** Last [kW]	Solar	Temp. [K]	Preis [€ / MWh]	Rel. Preis	Preis-rang	Preis-flag	Dist. min. Day [€ / MWh]	Zeit (sin / cos)	Grund-pro-gnose	Last ges-tern [kW]	Last letzte Wo-che [kW]
01.01. 00:00	H1	0.42	0.0	279.2	25.28	1.08	0.79	0	3.20	0.00 / 1.00	0.95	0.91	0.88
01.01. 00:00	H3	0.37	0.0	279.2	25.28	1.08	0.79	0	3.20	0.00 / 1.00	0.39	0.36	0.35
01.01. 00:00	H7	0.51	0.0	279.2	25.28	1.08	0.79	0	3.20	0.00 / 1.00	0.39	0.42	0.41
...	...	...	...	...	...	...	...	...	...	...	...	...	...
01.01. 00:00	H102	0.48	0.0	279.2	25.28	1.08	0.79	0	3.20	0.00 / 1.00	0.95	0.92	0.89
01.01. 00:15	H1	0.44	0.0	279.4	24.91	1.05	0.76	0	2.83	0.26 / 0.97	0.94	0.90	0.87
01.01. 00:15	H3	0.40	0.0	279.4	24.91	1.05	0.76	0	2.83	0.26 / 0.97	0.40	0.37	0.36
...	...	...	...	...	...	...	...	...	...	...	...	...	...
01.01. 00:15	H102	0.46	0.0	279.4	24.91	1.05	0.76	0	2.83	0.26 / 0.97	0.95	0.91	0.89
...	...	...	...	...	...	...	...	...	...	...	...	...	...
24.02. 23:45	H1	0.47	0.0	278.7	21.34	0.96	0.63	1	0.00	0.00 / 1.00	0.92	0.90	0.88
24.02. 23:45	H3	0.41	0.0	278.7	21.34	0.96	0.63	1	0.00	0.00 / 1.00	0.37	0.35	0.34
...	...	...	...	...	...	...	...	...	...	...	...	...	...

Tabelle 4.3: Schematischer Aufbau der Trainingsdaten mit Zielvariable für die Regressionsanalyse

den günstigsten und 1 den teuersten Zeitpunkt repräsentiert. Dadurch lassen sich Preisverläufe unabhängig vom absoluten Preisniveau zwischen verschiedenen Tagen vergleichen. Das Preisflag kennzeichnet Zeitpunkte, an denen der Strompreis im unteren Dezil des jeweiligen Tages liegt, und dient somit als binärer Indikator für besonders günstige Phasen. Ergänzend beschreibt die Distanz zum Tagestief den Abstand des aktuellen Preises zum niedrigsten Wert desselben Tages. Dieses Merkmal ermöglicht es, zu quantifizieren, wie weit der aktuelle Preis vom günstigsten verfügbaren Preisniveau entfernt ist.
Durch die Kombination dieser Preismerkmale kann das Modell nicht nur absolute Preisniveaus berücksichtigen, sondern auch deren tageszeitliche Struktur und Dynamik erfassen.
Ebenfalls berücksichtigt werden Wetterdaten, die aus Prognosen oder historischen Messungen stammen. Nach einer systematischen Bereinigung fließen ausschließlich numerische Größen wie Temperatur und solare Einstrahlung ein. Diese Größen beeinflussen insbesondere den Betrieb von Wärmepumpen und Photovoltaikanlagen und wirken sich somit unmittelbar auf die Lastverläufe der Haushalte aus.
Um zyklische Muster im Verbrauchsverhalten abzubilden, werden Zeitmerkmale in Form trigonometrischer Transformationen einbezogen. Eine rein numerische Kodierung von Stunden oder Wochentagen würde künstliche Sprünge erzeugen, da beispielsweise bei der Uhrzeit die Werte 23 und 0 in der Realität direkt aufeinanderfolgen, numerisch jedoch weit voneinander entfernt liegen. Um dieses Problem zu vermeiden, werden Tageszeit und Wochentag mithilfe von Sinus- und Kosinusfunktionen auf einen Einheitskreis projiziert. Die Transformation der Tageszeit wird in Formel 4.3 gezeigt und die der Wochenzeit wird in Formel 4.4 dargestellt. Durch die Kombination von Sinus und Kosinus erhält jeder Schritt eines Zeitmerkmals eine eindeutige Position innerhalb des jeweiligen Zyklus, ohne dass Brüche an den Übergängen entstehen. Auf diese Weise liegen benachbarte Zeitpunkte auch im Merkmalsraum eng beieinander, während gegenüberliegende Zeitpunkte klar voneinander abgegrenzt werden. Diese Transformation ermöglicht es, periodische Strukturen von Tages- und Wochenzyklen zu modellieren und erleichtert es dem Regressionsverfahren, wiederkehrende Muster im Verbrauchsverhalten zu identifizieren.

$$hour_{sin} = \sin\left(\frac{2\pi \cdot hour}{24}\right), \quad hour_{cos} = \cos\left(\frac{2\pi \cdot hour}{24}\right) \tag{4.3}$$

$$day_{sin} = \sin\left(\frac{2\pi \cdot day}{7}\right), \quad day_{cos} = \cos\left(\frac{2\pi \cdot day}{7}\right) \tag{4.4}$$

Ein weiterer Schwerpunkt liegt auf der Integration lastbezogener Vergangenheitswerte. Hierbei werden die Lasten derselben Viertelstunde des Vortags sowie der Vorwoche ergänzt. Diese Merkmale ermöglichen es dem Modell, wiederkehrende Verbrauchsmuster und typische Lastverläufe zu erkennen und abzubilden. Dadurch kann die Prognose auch kurzfristige Lastspitzen oder periodische Schwankungen präziser antizipieren.

Die Tabelle 4.3 zeigt den schematischen Aufbau der verwendeten Trainingsdaten für die Regressionsanalyse. Jeder Zeitpunkt enthält die vollständige Merkmalskombination für alle Haushalte. Dadurch wiederholt sich die Haushaltsstruktur zu jedem Zeitintervall, wodurch über den gesamten Trainingszeitraum Daten aus mehreren Wochen zur Verfügung stehen. Insgesamt entsteht so eine umfassende Merkmalsmatrix, die historische Lasten, Prognosen, Preis- und Wetterinformationen sowie zyklische Zeitstrukturen integriert und dem Regressionsmodell eine fundierte Grundlage für die Vorhersage individueller Haushaltslasten bietet.

Das Training erfolgt haushaltsspezifisch und umfasst einen Zeitraum von etwa zwei Monaten. Für jeden Haushalt wird ein eigener Datensatz aus den vorbereiteten Merkmalen und der zugehörigen Lastreihe gebildet. Diese Daten werden mit einem $StandardScaler$ normalisiert, um Unterschiede in den Wertebereichen der einzelnen Variablen auszugleichen. Jeder Haushalt wird im Trainingsprozess separat betrachtet, sodass das Modell jeweils spezifisch auf die individuellen Verbrauchscharakteristika angepasst werden kann. Die Modelle werden somit nacheinander für jeden Haushalt trainiert, wobei die zugrunde liegende Struktur der Eingangsgrößen identisch bleibt. Dieses Vorgehen gewährleistet, dass Unterschiede im Verbrauchsverhalten, in der technischen Ausstattung oder im Reaktionsverhalten auf Preissignale explizit berücksichtigt werden können.

Die Modellierung und Vorhersage erfolgt ebenfalls haushaltsspezifisch, sodass für jeden Haushalt ein individuelles Vorhersagemodell entsteht. Wie in Kapitel 3.7 beschrieben, wird als Lernverfahren ein Gradient-Boosting-Regressor eingesetzt. Die maximale Tiefe der Entscheidungsbäume wird auf fünf Ebenen begrenzt, um komplexe Wechselwirkungen zwischen den Eingangsgrößen zu erfassen und gleichzeitig vorzubeugen, dass das Modell beginnt, aus zufälligen Mustern aus den Trainingsdaten lernt. Die Gesamtzahl der Iterationen

wurde auf 400 festgelegt, um eine ausreichende Modellkapazität zur Abbildung nichtlinearer Zusammenhänge zu gewährleisten und gleichzeitig einer Überanpassung vorzubeugen. Die Lernrate ist mit einem Wert von 0,05 niedrig gewählt, um den Beitrag einzelner Iterationen zu begrenzen und den Lernprozess zu stabilisieren. Zur Sicherstellung der Reproduzierbarkeit der Ergebnisse wird ein fester Startwert für die Zufallsinitialisierung gesetzt. Das Training benötigt auf der eingesetzten Hardware ungefähr 15 Minuten pro Simulationstag. Nach dem Training werden die Modelle mit den zurückgehaltenen Testdaten überprüft. Die Vorhersagen werden dabei den tatsächlichen Lastwerten gegenübergestellt und mit den Fehlermaßen aus Kapitel 3.8 bewertet. Die Kombination erlaubt eine Einschätzung der Prognosequalität. Sowohl die Abbildung typischer Verläufe als auch die Erfassung kritischer Peaks werden überprüft.

5 Ergebnisse und Diskussion

In diesem Kapitel werden die Ergebnisse der entwickelten Methoden vorgestellt und interpretiert. In Kapitel 5.1 wird zunächst das Szenario erläutert, das den Datensatz dieser Arbeit bildet. In Kapitel 5.2 werden die Resultate des Clustering-Verfahrens beschrieben, mit dem die Haushalte anhand ihrer technischen Merkmale und Verbrauchscharakteristika kategorisiert werden. In Kapitel 5.3 folgen die Ergebnisse der Regressionsprognose, die auf der clusterbasierten Analyse aufbaut und durch zusätzliche Einflussgrößen ergänzt wird. Abschließend wird in Kapitel 5.4 untersucht, wie gut die KOF auf Basis der entwickelten Prognose Engpässe im Verteilnetz vermeiden kann und in welchem Maß dadurch die Netzstabilität unterstützt wird.

5.1 Szenario

In diesem Szenario basiert der verwendete Datensatz auf einem Verteilnetz aus dem Simbench-Benchmark und umfasst insgesamt 104 Haushalte, die über sechs Niederspannungsabgänge an einen Mittel-/Niederspannungstransformator angeschlossen sind.
Die Abbildung 5.1 zeigt den Aufbau des betrachteten Niederspannungsnetzes in Form einer Baumstruktur, bei der der Mittel-/Niederspannungstransformator der Ausgangspunkt ist. Von diesem Knoten aus verzweigen sich die sechs Niederspannungsabgänge (LV-Feeder), die in der Abbildung als Hauptleitungen dargestellt sind. An diesen Leitungen sind die einzelnen Haushalte angeschlossen. Jeder Haushalt ist mit einem blauen Knoten markiert und die Linien repräsentieren die jeweilige Verbindungsleitung im Verteilnetz. Der Baum verdeutlicht, dass die Haushalte nicht sternförmig direkt am Transformator angeschlossen sind, sondern über gestufte Leitungsabschnitte verteilt werden. Teilweise hängen viele Haushalte hintereinander an einem Strang, wodurch sich lange Zweige mit serieller Struktur ergeben. An

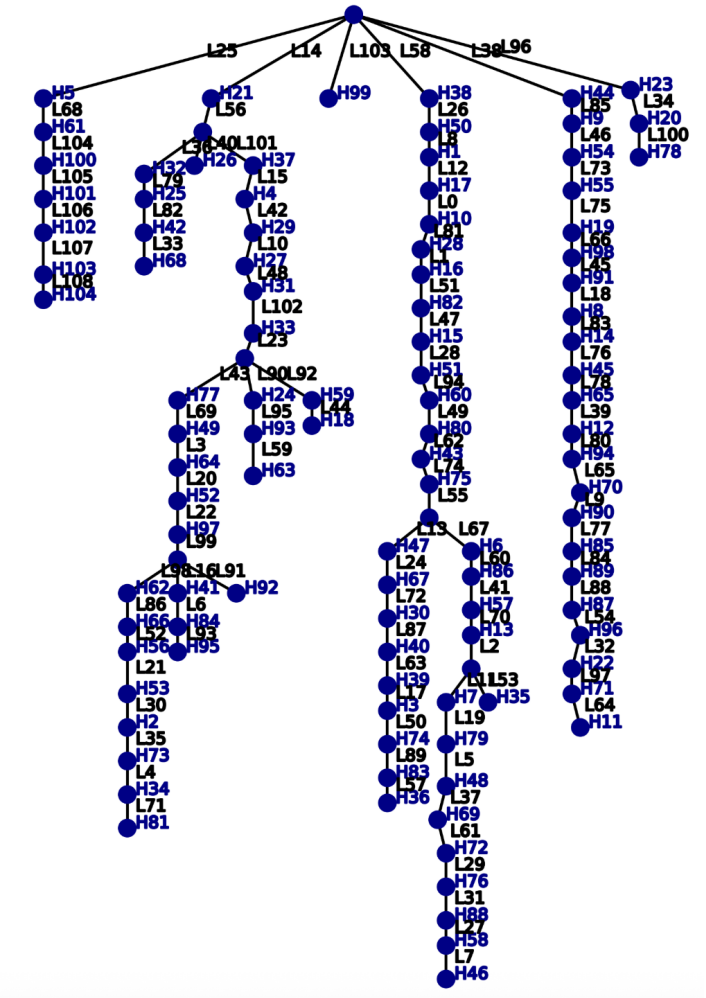


Abbildung 5.1: Aufbau des betrachteten Niederspannungsnetzes nach [2]

anderen Stellen zweigen kleinere Stränge ab, die nur wenige Haushalte versorgen. Damit spiegelt die Netzstruktur sowohl lange, linienförmige Abschnitte als auch kleine Abzweige wider. Ein solcher Aufbau ist typisch für ländliche bis semistädtische Niederspannungsnetze, die sich durch vergleichsweise große Leitungslängen und damit eine starke Abhängigkeit der Lastverteilung von der jeweiligen Topologie auszeichnen [63].

Alle Haushalte sind mit einem HEMS ausgestattet, das für jeden Tag einen Fahrplan erstellt. Diese Fahrpläne liegen in 15-Minuten-Intervallen vor und enthalten den geplanten Stromverbrauch bzw. die geplante Erzeugung. Ein Teil der Haushalte verfügt über zusätzliche Flexibilitätsoptionen. 60 Haushalte sind mit PV und Batteriespeichern ausgestattet, 66 Haushalte nutzen Wärmepumpen und 93 Haushalte besitzen mindestens ein batterieelektrisches Fahrzeug, wobei einige Haushalte sogar mehrere Fahrzeuge haben [2].

5.2 Clustering

Zunächst wird das Preisclustering betrachtet, bei dem die Haushalte anhand ihres Tarifs in flexible Preishaushalte und Festpreishaushalte segmentiert werden. Ziel ist die Bewertung der Trennschärfe dieser Gruppierung, wobei analysiert wird, unter welchen Bedingungen eine eindeutige Zuordnung möglich ist.
Darauf aufbauend erfolgt das Lastkurven-Clustering, das die Haushalte anhand ihrer charakteristischen Verbrauchsprofile gruppiert. Hierbei wird mittels HDBSCAN untersucht, in welchem Maße sich die Profile innerhalb der Cluster ähneln und welche Unterschiede zwischen den Clustern bestehen. Auf diese Weise lässt sich ableiten, inwieweit die ermittelten Clusterstrukturen für die Prognose unbekannter Haushalte herangezogen werden können. Zusätzlich werden die Merkmalsausprägungen der Cluster analysiert, da die technische Ausstattung der Haushalte ein wichtiger Faktor ist, nach dem die unbekannten Haushalte den jeweiligen Clustern zugewiesen werden.
Abschließend wird die resultierende Clusterprognose auf Ebene der Niederspannungsabgänge dargestellt. Hierbei wird die Prognosegüte untersucht, um den Einfluss der verfügbaren Informationsbasis auf die Genauigkeit der Lastschätzung zu erfassen. Dadurch kann bewertet werden, ob die Clusterprognose eine belastbare Grundlage für die Koordinierungsfunktion darstellt und welche Einschränkungen sich in den einzelnen Szenarien zeigen.

5.2.1 Preisclustering

Um die Resultate des Preisclusterings darstellen zu können, werden ausgewählte Tage betrachtet. Anhand dieser Beispiele lassen sich die Zusammenhänge übersichtlich präsentieren, ohne dass alle untersuchten Tage vollständig dargestellt werden. Dabei werden jeweils ein repräsentativer Sommer- und Wintertag untersucht, um saisonale Unterschiede in den Preisstrukturen zu berücksichtigen. Diese Auswahl ermöglicht es, typische Verbrauchs- und Erzeugungsmuster abzubilden und deren Einfluss auf die Preisbildung und Clusterzuordnung nachvollziehbar zu machen.

Simulationstag Winterszenario

Für die Analyse des Winterzeitraums wird exemplarisch der 27.02.2024 herangezogen. Die Ergebnisse des Preisclusterings lassen sich zunächst anhand der Projektion des Merkmalsraums nach PCA interpretieren. In Abbildung 5.2 ist zu erkennen, dass sich die Haushalte im zweidimensionalen Raum nach ihrem tatsächlichen Tariftyp weitgehend voneinander abgrenzen lassen. Während die Festpreishaushalte überwiegend im rechten Bereich der Darstellung liegen, konzentrieren sich die variablen Preishaushalte auf den linken Bereich. Obwohl sich einzelne Punkte im Merkmalsraum überlagern und keine kugelförmigen Cluster erkennbar sind, gelingt es dem Modell eine konsistente und weitgehend nachvollziehbare Abgrenzung der beiden Preistypen darzustellen.
Eine quantitative Bewertung liefert die in Abbildung 5.3 dargestellte Datenmatrix. Von insgesamt 74 betrachteten Haushalten werden 67 Haushalte korrekt klassifiziert, während 7 Fehlzuordnungen auftreten. Dies entspricht einer sehr hohen Genauigkeit und verdeutlicht, dass das Preisclustering an diesem Tag eine robuste und verlässliche Trennung der Tarifgruppen ermöglicht. Ähnliche Ergebnisse zeigen sich an nahezu allen betrachteten Wintertagen, was darauf hinweist, dass die grundsätzliche Trennbarkeit zwischen den beiden Tarifgruppen an Wintertagen robust ist. Die Ausprägung der Überschneidungen im Merkmalsraum variiert von Tag zu Tag, sodass auch die Trennschärfe der Clusterzuordnung tagesabhängig schwankt, die grundlegende Einteilung wird jedoch stabil erkannt.
Das Entscheidungskriterium für das Lastkurvenclustering basiert, wie in Kapitel 4.2 beschrieben, auf der Prüfung der Clustergrößen und ihres Verhältnisses. Nur wenn beide Cluster eine ausreichende Mindestgröße erreichen und zugleich ein ausgewogenes Größenverhältnis aufweisen, wird das Preisclustering als stabil eingestuft, andernfalls erfolgt die Wahl des vollständigen Clusterings. Da in diesem Beispiel die Cluster sowohl eine ausreichende Mindestgröße als auch ein ausgewogenes Größenverhältnis aufweisen, wird das Preisclustering korrekt als stabil eingestuft und die Lastkurven separat geclustert.

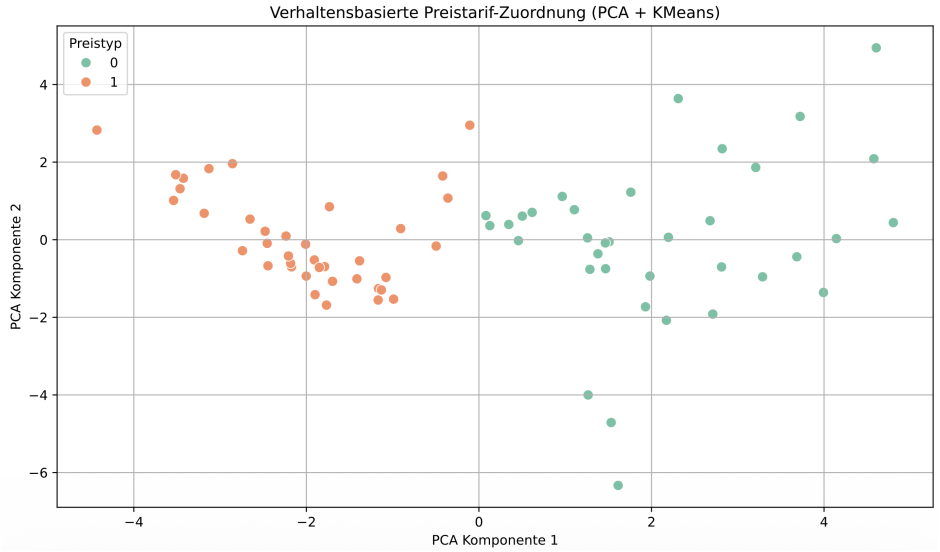


Abbildung 5.2: Darstellung des Merkmalsraums nach PCA (Winter)

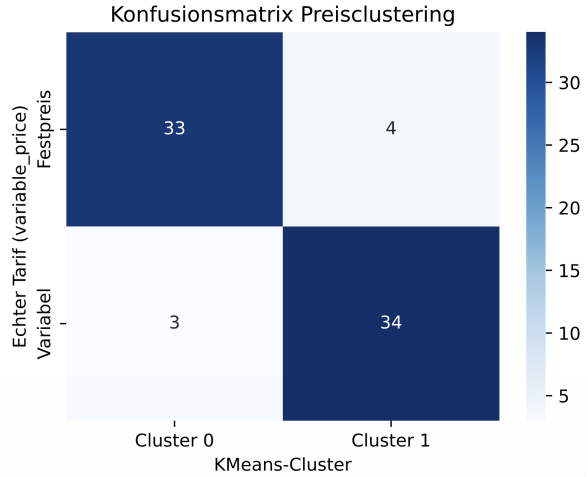


Abbildung 5.3: Datenmatrix Preisclustering (Winter)

Simulationstag Sommerszenario

Als weiteres Beispiel wird ein Sommertag herangezogen, um die Leistungsfähigkeit des Preisclusterings unter veränderten saisonalen Bedingungen zu untersuchen. Repräsentativ dient der 30.07.2024 als Simulationstag, anhand dessen die Ergebnisse exemplarisch dargestellt werden. In Abbildung 5.4 ist ebenfalls die Projektion der Haushalte in den zweidimensionalen Merkmalsraum nach der PCA-Transformation dargestellt. Auch hier zeigt die farbliche Markierung die Zuordnung der Haushalte zu den mutmaßlichen Preisgruppen. Die beiden Preistypen lassen sich deutlich besser voneinander abgrenzen als im Winterszenario. Während die variablen Preishaushalte fast vollständig im rechten Bereich des Merk-

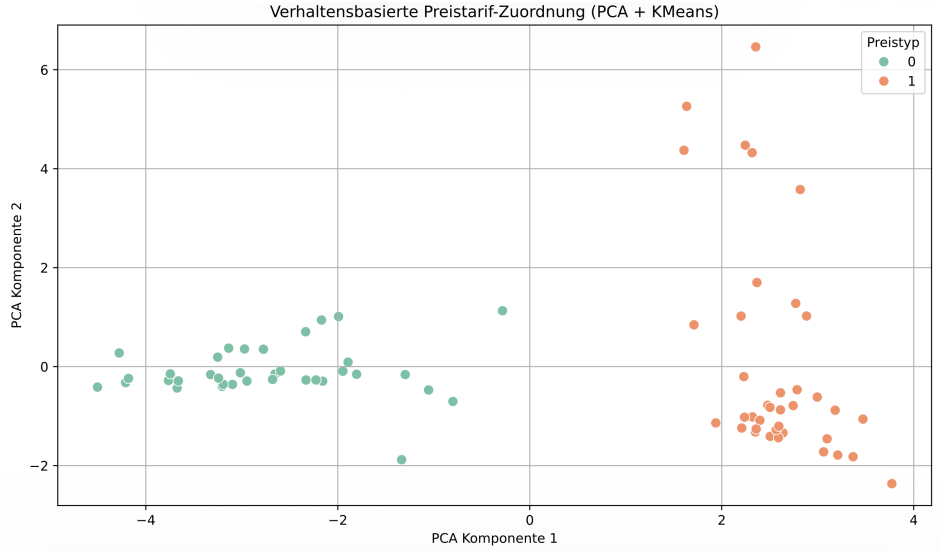


Abbildung 5.4: Darstellung des Merkmalsraums nach PCA (Sommer)

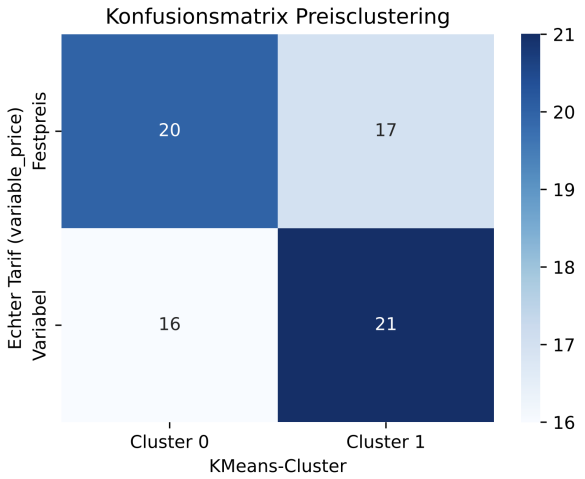


Abbildung 5.5: Datenmatrix Preisclustering (Sommer)

malsraums konzentriert sind, liegen die Festpreishaushalte weitgehend im linken Bereich. Nur wenige Ausreißer weichen von dieser Struktur ab, sodass insgesamt eine klare Trennung sichtbar wird. Allerdings zeigt diese scheinbar eindeutige Aufteilung auch die Grenzen des Preisclusterings auf. Die Abbildung verdeutlicht, dass sich die Gruppen für diesen Tag geometrisch gut unterscheiden lassen, doch diese klare Zweiteilung entspricht nicht der Merkmalsausprägung.

Die Datenmatrix in Abbildung 5.5 verdeutlicht diese Beobachtung. Der Vergleich zeigt, dass 20 Festpreishaushalte und 21 variable Preishaushalte korrekt zugeordnet werden konnten, während 16 Festpreishaushalte und 17 variable Preishaushalte falsch klassifiziert wurden.

Dies entspricht einer ungenügenden Genauigkeit und verdeutlicht, dass die visuell erkennbare Trennung im PCA-Merkmalsraum nur bedingt mit der tatsächlichen Tarifzugehörigkeit der Haushalte übereinstimmt. Das Preisclustering bildet demnach eher geometrische Muster im Datenraum ab, ohne die zugrunde liegenden tarifbedingten Unterschiede zuverlässig zu erfassen. An diesem Tag ist das Verhältnis der Clustergrößen zu unausgeglichen, weshalb das Modell folgerichtig auf das Clustering im Gesamtkollektiv entscheidet.
Dieses Verhalten beschränkt sich nicht nur auf diesen Tag, sondern wiederholt sich an nahezu allen betrachteten Sommertagen. Die Ursache liegt vor allem in den veränderten Laststrukturen. Während im Winter höhere Grundlasten und ein stärkerer Einfluss des Heizverhaltens dazu führen, dass sich Lastverschiebungen in Abhängigkeit vom Strompreis deutlicher abzeichnen, ist die Gesamtnachfrage im Sommer deutlich geringer. Hinzu kommt, dass die Einspeisung aus PV-Anlagen die Lastverläufe erheblich beeinflusst und teilweise überlagert, sodass preisinduzierte Muster im Verbrauch weniger sichtbar sind. In der Folge liefern die Merkmale, die auf der Korrelation zwischen Last und Preis beruhen, in den Sommermonaten nur eingeschränkt trennscharfe Informationen. Das Preisclustering kann daher in dieser Jahreszeit keine robuste Unterscheidung zwischen Festpreis- und variablen Preishaushalten gewährleisten, wohingegen im Winter eine deutliche Tendenz zur Abgrenzung erkennbar ist. Aus diesem Grund wird das Preisclustering in den Sommermonaten für die weitere Lastprognose nicht berücksichtigt. In den Übergangsmonaten zwischen den Jahreszeiten zeigt sich ein weniger eindeutiges Bild. Je nach Temperatur- und Sonneneinstrahlung können sich die Lastprofile stärker am Winter- oder am Sommerverhalten orientieren. Dadurch schwankt die Trennschärfe erheblich, sodass das Entscheidungskriterium in diesen Phasen keine konsistente Präferenz für Preis- oder vollständiges Clustering erkennen lässt.

5.2.2 Lastkurven-Clustering

Im Folgenden werden die Ergebnisse des Lastkurven-Clusterings vorgestellt. Als Ausgangspunkt wird das Szenario mit dem höchsten Bekanntheitsgrad gewählt, bei dem 70% der Haushalte bekannt sind. Dieses Szenario bietet die breiteste Datenbasis und erlaubt eine besonders detaillierte Betrachtung der Clusterstrukturen sowie ihrer charakteristischen Merkmale. Darauf aufbauend werden anschließend die Szenarien mit 50% und 30% be-

kannten Haushalten betrachtet, wodurch die Aussagekraft der Methoden unter zunehmend eingeschränkter Informationslage bewertet werden kann.

Es wird ebenfalls der Simulationstag 27.02.2024 aus Kapitel 5.2.1 herangezogen und die Ergebnisse vom Preisclustering werden angewendet und weitergeführt. Daher werden die Resultate im Folgenden getrennt nach Festpreis- und variablem Preisverhalten betrachtet. Dazu wird zunächst eine Dimensionsreduktion mit der UMAP-Methode durchgeführt, um die hochdimensionalen Lastprofildaten zweidimensional abbilden zu können.

Die Abbildung 5.6 zeigt das Ergebnis des HDBSCAN-Clusterings für Haushalte mit Festpreisverhalten auf Basis der UMAP-Darstellung. Entlang der beiden UMAP-Achsen wird deutlich, dass sich die Haushalte in klar voneinander getrennte Bereiche aufteilen und der Algorithmus vier nachvollziehbare Cluster bildet. Cluster 0 (blau) befindet sich im oberen Bereich der Darstellung, Cluster 1 (orange) im rechten Bereich, Cluster 2 (grün) im unteren Zentrum und Cluster 3 (rot) liegt links unten im Diagramm. Diese Trennung deutet darauf hin, dass sich die Haushalte mit Festpreis-Tarif in unterschiedliche, charakteristische Verbrauchsmuster gliedern lassen. Innerhalb der Cluster liegen die Punkte eng beieinander und spiegeln damit eine hohe Ähnlichkeit der zugehörigen Lastkurven wider, während zwischen den Clustern deutliche Unterschiede bestehen. Neben dem reinen Lastverlauf trägt die Einbeziehung technischer Merkmale der Haushalte zu einer kompakteren und klareren Clusterstruktur bei. Die UMAP-Darstellung macht somit die Heterogenität innerhalb der Gruppe der Festpreis-Haushalte sichtbar und liefert eine anschauliche Grundlage für die Clusterbildung.

Die Qualität der Clusterbildung wird zudem durch die vorgestellen Gütemaße aus Kapitel 3.5 bestätigt. Der Silhouette Score von 0,957 weist auf eine hohe Kohärenz innerhalb der Cluster hin. Der Davies-Bouldin Index von 0,056 weist auf eine sehr klare Trennung hin, da niedrigere Werte auf eine gute Separierbarkeit der Cluster hinweisen. Ergänzend zeigt der Calinski-Harabasz Index mit einem Wert von 11868 eine ausgeprägte Clusterstruktur und deutet auf eine hohe interne Homogenität bei gleichzeitig deutlicher Trennung zwischen den Clustern hin.

Die Abbildung 5.7 zeigt die Ergebnisse des HDBSCAN-Clusterings für Haushalte mit variablem Preisverhalten. Entlang der beiden UMAP-Achsen werden mehrere klar voneinander getrennte Cluster sichtbar, die eine deutliche Strukturierung der Daten erkennen lassen. Ins-

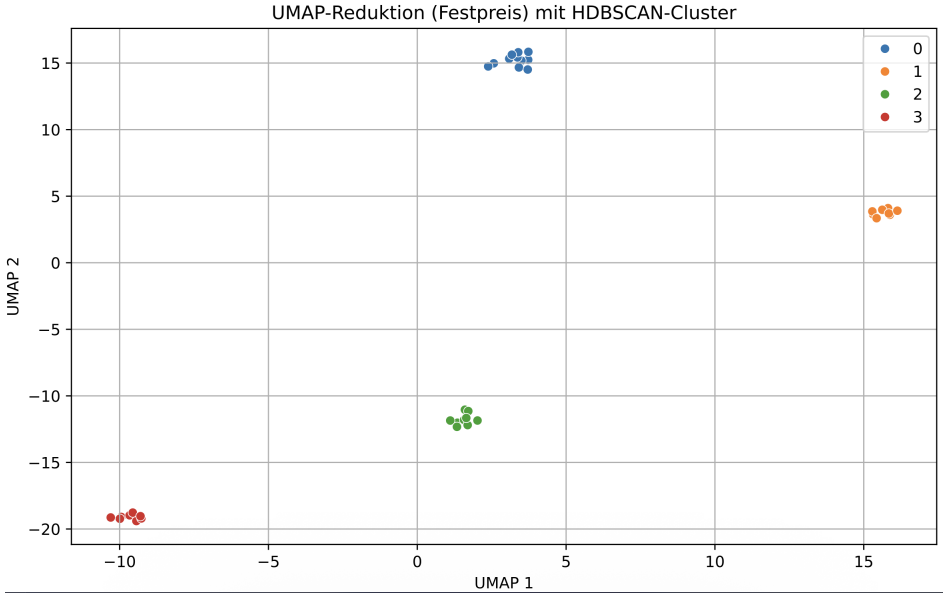


Abbildung 5.6: Darstellung des Merkmalsraums nach UMAP von Haushalten mit Festpreisverhalten

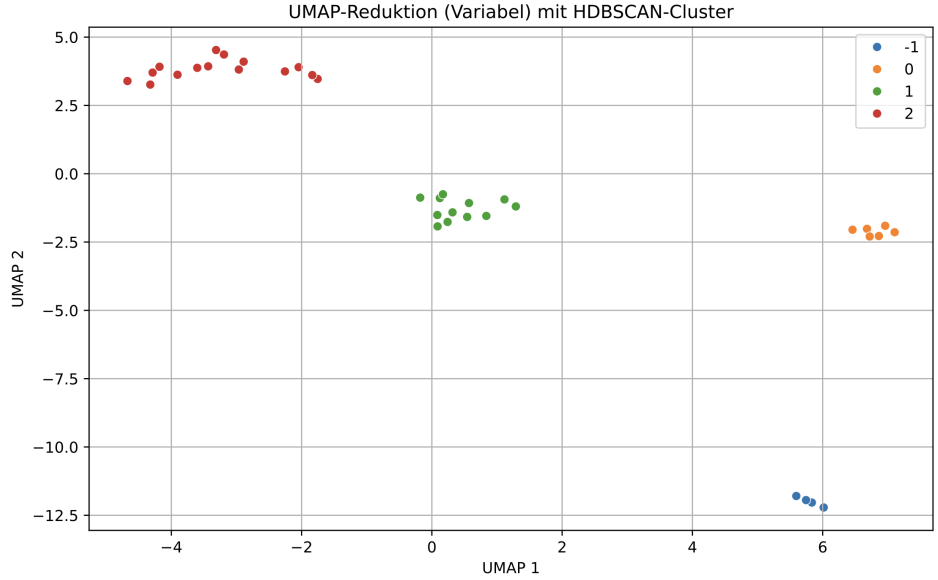


Abbildung 5.7: Darstellung des Merkmalsraums nach UMAP von Haushalten mit variablen Preisverhalten

gesamt ergeben sich drei stabile Cluster (orange, grün und rot) sowie ein kleineres Cluster (blau), das als Ausreißer klassifiziert wird. Die Cluster liegen in unterschiedlichen Bereichen der UMAP-Darstellung. Cluster 0 (orange) befindet sich im rechten Bereich, Cluster 1 (grün) im zentralen Bereich und Cluster 2 (rot) im oberen linken Bereich. Cluster -1 (blau) liegt isoliert im unteren rechten Bereich und umfasst wenige Ausreißer. Diese Einstufung erfolgt vermutlich dadurch, dass die Gruppe weniger als fünf Haushalte umfasst und aufgrund der

gewählten Hyperparameter kein eigenständiges Cluster bildet. HDBSCAN identifiziert Cluster auf Basis lokaler Dichteverteilungen und stuft Bereiche mit zu geringer Punktdichte als Rauschen ein, selbst wenn diese in der UMAP-Darstellung räumlich isoliert erscheinen. Die visuelle Distanz stellt somit nicht zwangsläufig eine rechnerische Trennung dar. Dennoch können die im Rausch-Cluster enthaltenen Haushalte inhaltlich relevant sein, da sie auf atypische oder besonders individuelle Verbrauchsmuster hinweisen können und werden daher als eigenständige Gruppe weiter betrachtet. Die Punkte innerhalb der Cluster liegen eng beieinander, was auf eine hohe interne Homogenität und eine deutliche Trennbarkeit zwischen den Gruppen hinweist. Auch hier wird die Qualität der Clusterbildung durch die Gütemaße bestätigt. Der Silhouette Score von 0,863 zeigt eine insgesamt gute interne Kohärenz der Cluster. Der Davies-Bouldin Index von 0,157 weist auf eine gute Trennung hin. Der Calinski-Harabasz Index von 807 ist im Vergleich zum Festpreis-Szenario niedriger und deutet auf eine kompaktere, weniger stark differenzierte Clusterstruktur hin. Diese Einschätzung wird auch durch die UMAP-Darstellung unterstützt. Im variablen Szenario zeigt sich eine kleinere Achsenskalierung, wodurch die Datenpunkte in einem engeren Wertebereich liegen und die Cluster kompakter sowie näher beieinander erscheinen. Dies deutet auf geringere Unterschiede zwischen den Haushalten mit variablem Tarif hin. Während die Punkte im Festpreis-Szenario über eine größere Fläche verteilt sind und stärkere Streuungen aufweisen, lassen die dichteren Strukturen im variablen Szenario auf tendenziell homogenere Verbrauchs- und Verhaltensmuster schließen. In der Interpretation der Ergebnisse ist jedoch zu beachten, dass UMAP eine nichtlineare Projektion der hochdimensionalen Daten darstellt und die resultierenden Strukturen keine eindeutigen Cluster im mathematischen Sinn wiedergeben. Dennoch lassen sich aus der Visualisierung qualitative Hinweise auf Unterschiede in der Streuung und Homogenität der Haushaltsverhalten ableiten.

Die Datenverteilung innerhalb der Cluster kann mit der durchschnittlichen technischen Ausstattung der Haushalte in den einzelnen Gruppen veranschaulicht werden. Diese Aufteilung wird in Abbildung 5.8 dargestellt. Dabei wird deutlich, dass sich die Cluster auch in ihren Merkmalsausprägungen klar voneinander unterscheiden. Während Cluster 0 vollständig elektrifizierte Haushalte mit PV-Anlage, Speicher, Wärmepumpe und E-Auto umfasst, zeigt Cluster 1 konventionell ausgestattete Haushalte mit ausschließlich E-Auto. Cluster 2 kombiniert hohe PV- und Speicheranteile mit E-Autos, und Cluster 3 enthält Haushalte mit

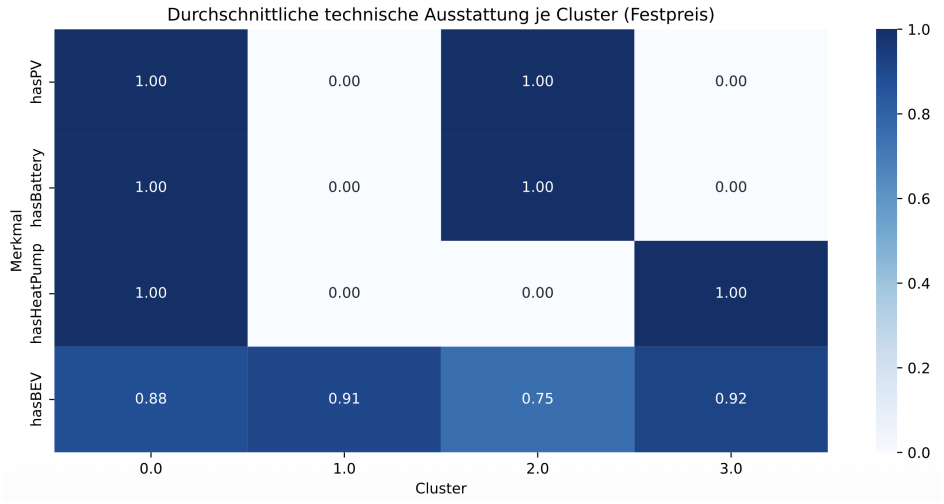


Abbildung 5.8: Merkmalsausprägung der gebildeten Cluster von Haushalten mit Festpreisverhalten

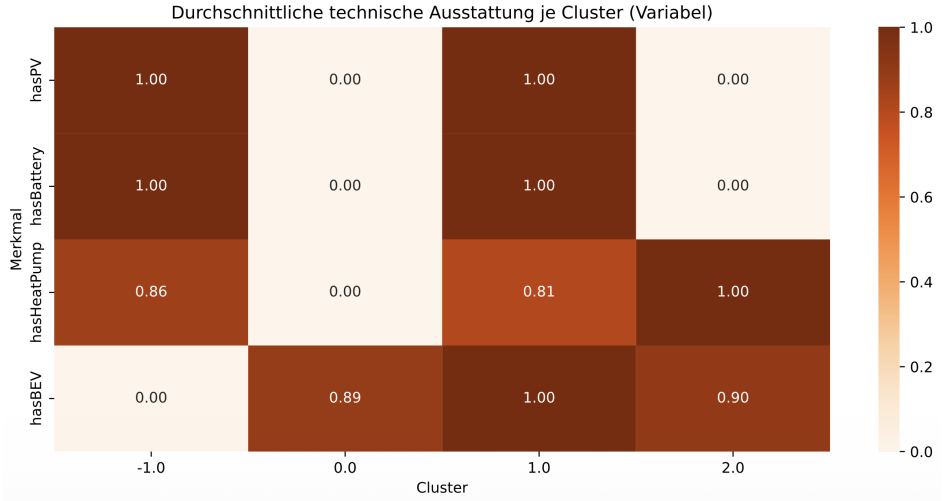


Abbildung 5.9: Merkmalsausprägung der gebildeten Cluster von Haushalten mit variablen Preisverhalten

Wärmepumpe und E-Auto ohne PV-Anlage. Damit bildet das Clustering klar differenzierbare technische Profile ab, wobei auch Verbrauchsmuster zur Trennung beitragen.

Im Szenario mit variablem Preisverhalten zeigt die Abbildung 5.9 eine differenziertere Verteilung der Merkmalsausprägungen über die vier Cluster hinweg. Cluster 0 umfasst Haushalte mit ausschließlich E-Auto und ohne Eigenerzeugung, deren Flexibilität vor allem aus dem Ladeverhalten resultiert. Cluster 1 vereint weitgehend vollständig elektrifizierte Haushalte mit PV-Anlage, Speicher, Wärmepumpe und E-Auto und steht damit für eine hohe Eigenverbrauchsoptimierung. Cluster 2 enthält Haushalte mit Wärmepumpe und E-Auto ohne PV-Anlage, während Cluster –1 Haushalte mit PV-Anlage, Speicher und Wärmepumpe

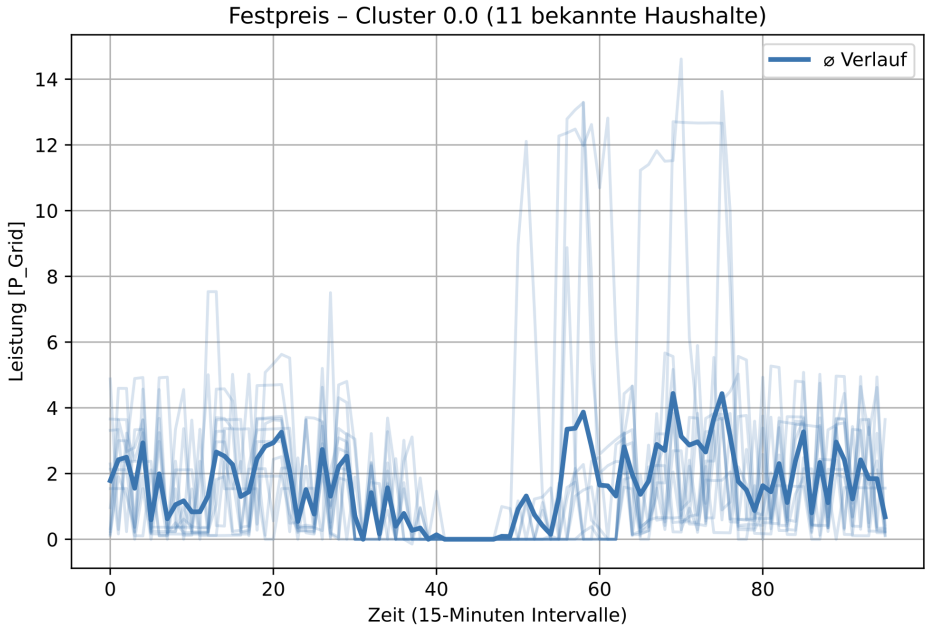


Abbildung 5.10: Darstellung der individuellen Lastkurven im Cluster 0 (Festpreisverhalten)

ohne E-Auto abbildet. Inhaltlich bleibt Cluster -1 relevant, da es ein spezifisches Ausstattungsprofil darstellt, das sich klar von den übrigen Gruppen unterscheidet.

Insgesamt zeigt die Analyse der technischen Merkmale, dass das Clustering eine sinnvolle und nachvollziehbare Gruppierung der Haushalte ermöglicht. Zwar liefert die technische Ausstattung der Haushalte wichtige Anhaltspunkte für die Charakterisierung der Cluster, sie ist jedoch nicht allein ausschlaggebend für deren Abgrenzung. Entscheidend sind vielmehr die tatsächlichen Verbrauchsverläufe, um Verhaltensmuster sichtbar zu machen. Um die gewonnenen Cluster daher inhaltlich zu bestätigen, werden im nächsten Schritt die Lastkurven der jeweils zugeordneten Haushalte betrachtet. Der Vergleich der Lastprofile innerhalb und zwischen den Clustern ermöglicht eine genauere Einschätzung, inwieweit die Gruppen tatsächlich ein konsistentes und unterscheidbares Verbrauchsverhalten aufweisen. Dabei werden insbesondere jene Lastprofile gegenübergestellt, die für die Mehrzahl der Haushalte innerhalb eines Clusters typisch sind und somit das charakteristische Verbrauchsverhalten der jeweiligen Gruppe widerspiegeln.

Die Abbildung 5.10 zeigt die Lastkurven der Haushalte im Cluster 0 des Festpreis-Szenarios. Auf der y-Achse ist die elektrische Leistung in Kilowatt, auf der x-Achse die zeitliche Entwicklung über den Tag hinweg in 15-Minuten-Intervallen dargestellt. Veranschaulicht sind die individuellen Lastverläufe der insgesamt 11 bekannten Haushalte sowie der entsprechende

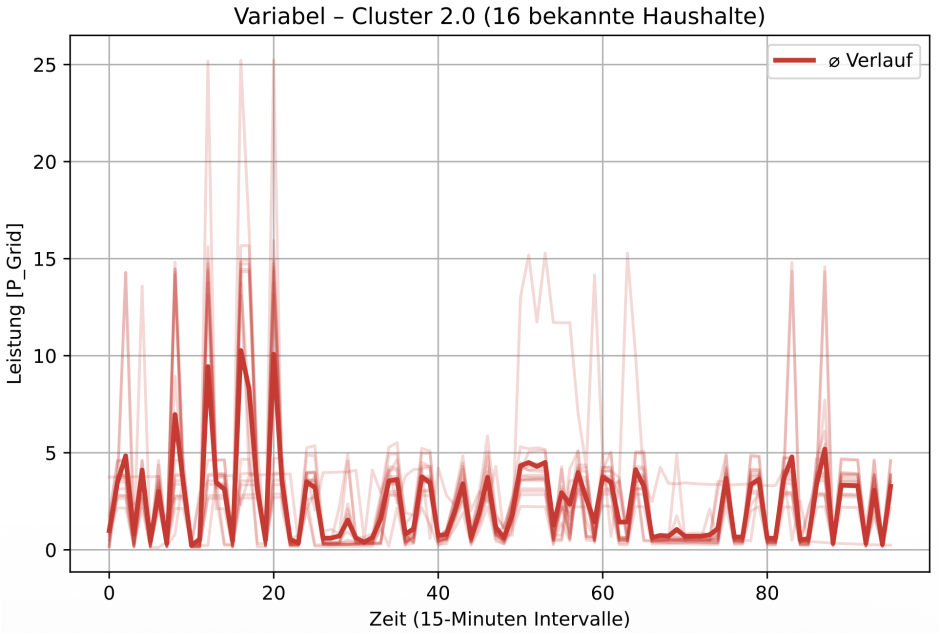


Abbildung 5.11: Darstellung der individuellen Lastkurven in Cluster 2 (variables Preisverhalten)

Mittelwertverlauf, der als dicke Linie hervorgehoben ist. Im Tagesverlauf ist ein grundsätzlich konsistentes Verbrauchsmuster erkennbar, das durch einen gleichmäßigen Grundlastanteil und mehrere deutliche Abendspitzen geprägt ist. Auffällig ist die hohe Streuung der Einzelverläufe, insbesondere in den Spitzenlastbereichen. Während einige Haushalte sehr hohe Verbrauchsspitzen aufweisen, bleibt der Verbrauch bei anderen nahezu konstant. Dies führt zu einer erhöhten Varianz innerhalb des Clusters und zeigt, dass die Haushalte zwar einem ähnlichen Grundmuster folgen, die individuelle Ausprägung des Verbrauchs jedoch stark variiert. Trotz dieser Unterschiede ist das allgemeine Tagesprofil erkennbar stabil, was auf eine gemeinsame, wenn auch locker definierte Struktur innerhalb des Clusters schließen lässt.

In der Abbildung 5.11 sind die Lastkurven der 16 Haushalte im Cluster 2 des variablen Preisszenarios dargestellt. Der mittlere Verlauf liegt insgesamt auf einem ähnlichen Leistungsniveau, weist jedoch klar abgegrenzte und zeitlich konzentrierte Peaks auf. Die Einzelverläufe streuen deutlich weniger als im Festpreis-Szenario, was auf ein homogeneres Verbrauchsverhalten hinweist. Die wiederkehrenden Spitzen deuten auf Lastverschiebungen in preisgünstigen Zeitfenstern hin.

Insgesamt wird deutlich, dass die Haushalte im Festpreis-Szenario eher ein konstantes aber individuelleres Verbrauchsverhalten aufweisen, während die Haushalte im variablen

Szenario flexibler und dynamischer reagieren. Die niedrige Streuung und die wiederkehrenden Lastspitzen deuten darauf hin, dass diese Haushalte ihren Stromverbrauch stärker an Preissignale anpassen. Damit bestätigt die Analyse der Lastkurven die zuvor beobachteten Unterschiede und verdeutlicht die unterschiedlichen Verhaltensmuster zwischen beiden Tarifgruppen.
Dieses Verhalten lässt sich auch bei weiteren Lastkurven zwischen beiden Szenarien beobachten. Da sich die Verläufe in ihrer Grundstruktur ähneln und lediglich in der Intensität der Spitzen unterscheiden, werden die zusätzlichen Darstellungen zur Vollständigkeit im Anhang A aufgeführt.

5.2.3 Clusterprognose

Im folgenden Abschnitt werden die Ergebnisse der skalierten Clusterkurven vorgestellt, die die prognostizierten Lastverläufe der jeweiligen Haushaltsgruppen abbilden. Dabei wird zunächst der Mittelspannungs-Transformator (MV-1) betrachtet, der die zusammengefassten Lastverläufe aller nachgelagerten Niederspannungsabgänge repräsentiert. Anschließend wird exemplarisch der Niederspannungsabgang LV-1-2 analysiert, um die Qualität und Aussagekraft der Prognosen auf Haushalts- bzw. Abgangsebene zu verdeutlichen. Bei einem Anteil von 70% bekannter Haushalte fasst der Niederspannungsabgang LV-1-2 drei unbekannte Haushalte zusammen. Damit werden unterschiedliche Aggregationsebenen dargestellt, welche einen wesentlichen Einfluss auf die Aussagekraft der Prognose haben. Bei hoch aggregierten Daten, wie bei der Transformator-Ebene, werden individuelle Verbrauchsschwankungen und Prognosefehler geglättet, sodass die resultierende Kurve meist sehr stabil und gleichmäßig verläuft. Das Modell muss hier weniger auf kurzfristige Einzelereignisse reagieren. Im Gegensatz dazu bilden Niederspannungsabgänge nur eine begrenzte Anzahl an Haushalten ab. Dadurch treten individuelle Verbrauchsdynamiken stärker hervor, was die Prognose auf dieser Ebene anspruchsvoller, aber aussagekräftiger macht. Abweichungen zwischen Prognose und Fahrplan sind hier direkter sichtbar und erlauben eine präzisere Bewertung der Modellgüte.
Exemplarisch wird auch hier der 27.02.2024 als Szenario herangezogen, um die Ergebnisse der clusterbasierten Lastprognose darzustellen. Da die Ergebnisse für die Sommerperiode

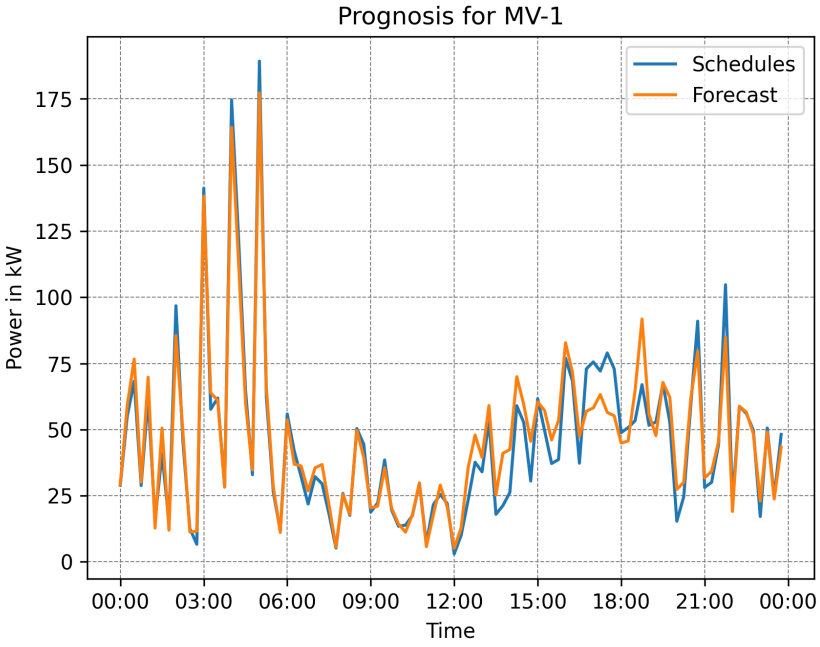


Abbildung 5.12: Vergleich von prognostiziertem und geplantem Lastverlauf für den Transformator MV-1

ein ähnliches Verhalten zeigen, werden diese für eine bessere Übersichtlichkeit im Anhang B dargestellt. Die Zuordnung der unbekannten Haushalte zu den bestehenden Clustern wird anhand ihrer Verbrauchsmerkmale durchgeführt. Die Güte dieser Zuordnung wird über die mittlere Cosine Similarity aus Kapitel 3.8 bewertet, die ein Maß für die Ähnlichkeit der Merkmalsvektoren zwischen bekannten und unbekannten Haushalten darstellt. Mit einem durchschnittlichen Score von 0,924 im Festpreis-Szenario und 0,916 im variablen Szenario zeigt sich eine sehr hohe Übereinstimmung, sodass die Zuordnung der unbekannten Haushalte zu einem Cluster als zuverlässig und konsistent einzustufen ist.

Die Abbildung 5.12 zeigt den Lastverlauf des Transformators, der die summierten Lasten der nachgelagerten Niederspannungsabgänge darstellt. Die Prognose, die in orange dargestellt ist, folgt der gemessenen bzw. geplanten Kurve in blau überwiegend präzise. Besonders in den Morgenstunden lassen sich die zeitlichen Verläufe und Leistungsniveaus gut reproduzieren. Im Nachmittags- und Abendzeitraum zeigen sich größere Abweichungen zwischen Prognose und Fahrplan. Insgesamt zeigt sich, dass das Modell auf Transformator-Ebene eine hohe Prognosegüte erreicht. Kleinere Abweichungen in den Spitzenzeiten resultieren vermutlich aus individuellen, kurzzeitigen Verbrauchsereignissen, die in den aggregierten

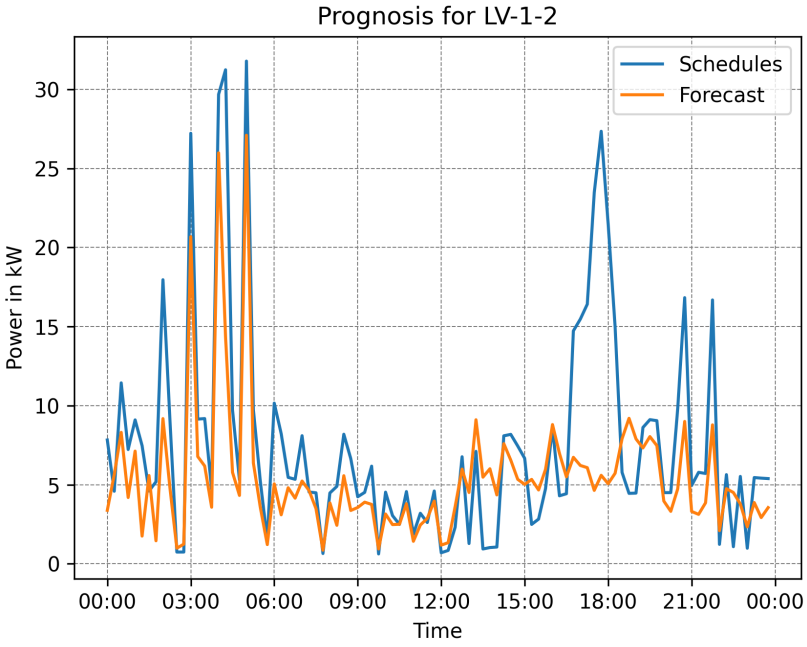


Abbildung 5.13: Vergleich von prognostiziertem und geplantem Lastverlauf für den Niederspannungsabgang LV-1-2

Clusterprofilen nicht vollständig abgebildet sind. Dennoch wird der allgemeine Tagesverlauf realitätsnah erfasst, und die Prognose bildet die saisonalen Besonderheiten des Winterverbrauchs sehr gut nach.

Die Abbildung 5.13 zeigt den Lastverlauf des Niederspannungsabgangs LV-1-2, welcher aufgrund der geringen Anzahl an Haushalten individueller ausfällt. Die Prognose bildet die grundsätzliche Struktur des Tagesgangs erkennbar ab, weist jedoch deutliche Unterzeichnungen in den Abend- und Nachtstunden auf. Insbesondere werden Spitzenwerte prognostiziert, die deutlich unter den tatsächlichen Fahrplanwerten liegen. Diese Abweichung deutet darauf hin, dass das verwendete Cluster ein zu gering ausgeprägtes Spitzenlastverhalten modelliert. Besonders kritisch fällt die Prognose im Tagesverlauf zwischen 17 und 19 Uhr aus. In dieser Zeit, in der die realen Lastwerte einen Anstieg aufweisen, bleibt die Prognose nahezu konstant auf einem niedrigeren Niveau. Das Modell unterschätzt hier die reale Tagesaktivität, was auf eine unzureichende Erfassung von Leistungsspitzen schließen lässt. Diese Kombination aus unterzeichneten Abend- und Nachtspitzen und unterschätztem Tagesverlauf zeigt, dass die Prognose auf dieser Aggregationsebene zwar die Haupt-

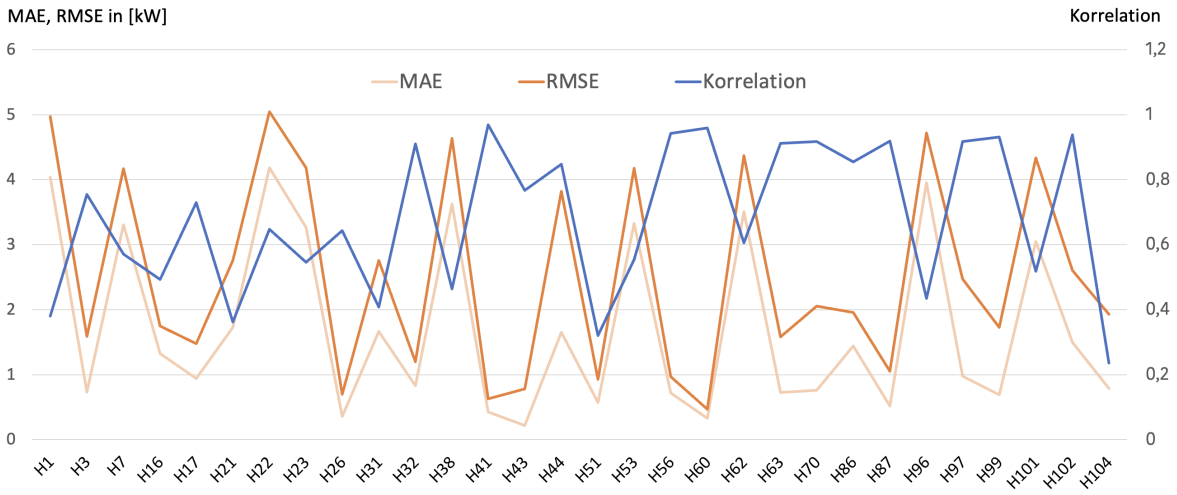


Abbildung 5.14: Fehlermetriken der Clusterprognose bei den einzelnen unbekannten Haushalten

verbrauchsphasen erkennt, jedoch Schwierigkeiten hat, die Leistungsintensität über den gesamten Tag realistisch nachzubilden.

In Abbildung 5.14 wird die Prognosegüte der Clustering-Methode auf Basis der einzelnen Haushalte dargestellt. Gezeigt werden die thematisierten Kennwerte aus Kapitel 3.8. Während die zuvor gezeigten Abbildungen die Ergebnisse auf Transformatorenebene zusammenfassen und damit bereits stark aggregierte Lastverläufe abbilden, zeigt die hier dargestellte Analyse die Prognosegüte auf Haushaltsebene. Dadurch wird sichtbar, wie stark sich die Modellgüte zwischen den einzelnen Verbrauchern unterscheidet, woraus abgeleitet werden kann, welche Einflussfaktoren für die Prognose relevant sind. Die orangefarbenen Linien zeigen die beiden Fehlermaße MAE und RMSE in [kW], während die in blau dargestellte Linie die Korrelation angibt und somit darstellt, wie stark die Form des prognostizierten Lastverlaufs mit dem tatsächlichen Verbrauchsverlauf übereinstimmt.

Insgesamt zeigt sich ein heterogenes Bild der Prognosegüte über die betrachteten Haushalte hinweg. Die Fehlerwerte schwanken stark zwischen den jeweiligen Haushalten, was auf erhebliche Unterschiede im individuellen Verbrauchsverhalten hindeutet. Auffällig ist, dass MAE und RMSE weitgehend parallel verlaufen und der RMSE etwas höhere Werte aufweist. Dies deutet darauf hin, dass keine ausgeprägten Ausreißer in den Prognosefehlern vorliegen und das Modell insgesamt ein gleichmäßiges Fehlerverhalten zeigt. Darüber hinaus besteht ein klarer Zusammenhang zwischen den Fehlerwerten und der Korrelation. Haushalte mit niedrigen Fehlern weisen in der Regel auch hohe Korrelationswerte auf, was bedeutet,

dass das Modell sowohl den Verlauf als auch die Höhe des Verbrauchs gut abbildet. Umgekehrt sind bei Haushalten mit hohen Fehlerwerten meist auch geringere Korrelationen zu beobachten, was auf eine schwächere zeitliche Übereinstimmung hinweist.

Zusätzlich wurden die Haushalte hinsichtlich ihres Preisverhaltens untersucht und es zeigt sich ein besonders deutlicher Zusammenhang. Haushalte mit variablem Preisverhalten erzielen tendenziell niedrigere Fehlerwerte und höhere Korrelationen, da ihr Verbrauch stärker auf Preissignale reagiert und somit regelmäßiger und strukturierter verläuft. Im Durchschnitt liegen die Fehlerwerte des MAE bei etwa 1 bis 2 kW, während Festpreishaushalte Werte zwischen 3 und 4 kW erreichen und damit rund 100% höhere Abweichungen aufweisen. Ihre Korrelationen fallen zugleich um etwa 40-50% geringer aus, was auf ein weniger geregeltes und schwerer vorhersagbares Verbrauchsverhalten hinweist. Diese systematische Preisabhängigkeit kann vom Prognosemodell gut abgebildet werden, während sich der Verbrauch der Festpreishaushalte stärker durch individuelle und zufällige Einflüsse bestimmen lässt, was zu höheren Prognosefehlern führt. Unter diesen Bedingungen zeigen sich die Schwächen der Clusterprognose deutlich. Da sich ein Teil der Haushalte nicht an Preissignalen orientiert, entstehen individuelle und wenig wiederkehrende Verbrauchsmuster, die sich nur schwer in homogene Cluster einordnen lassen. Das Clustering fasst häufig Haushalte zusammen, deren Tagesverläufe nur oberflächlich ähnlich sind, was zu unzureichenden Repräsentationen typischer Verbrauchsverläufe führt. Das Verhalten ist dadurch weniger durch externe Einflüsse gesteuert und damit für datenbasierte Modelle schwieriger zu erfassen.

Darüber hinaus verdeutlicht der Vergleich mit den zuvor gezeigten aggregierten Ergebnissen, dass sich individuelle Abweichungen auf Haushaltsebene stark bemerkbar machen, während sich Abweichungen auf Transformatorenebene gegenseitig ausgleichen. Dieser sogenannte Aggregationseffekt führt dazu, dass die Prognosequalität auf höheren Netzebenen stabiler und glatter ausfällt.

Auch unter einem geringeren Anteil bekannter Haushalte konnten die Ergebnisse grundsätzlich bestätigt werden. Sowohl im Szenario mit 50% als auch mit 30% bekannten Haushalten zeigen sich ähnliche Fehlerverläufe, sodass die Bewertung der Prognosegüte zur vereinfachten Darstellung hier über den MAE erfolgt.

Die Abbildung 5.15 zeigt den Vergleich des MAE für die Prognoseergebnisse bei unterschiedlichem Anteil bekannter Haushalte. Dargestellt sind die Fehlerwerte für drei Szenari-

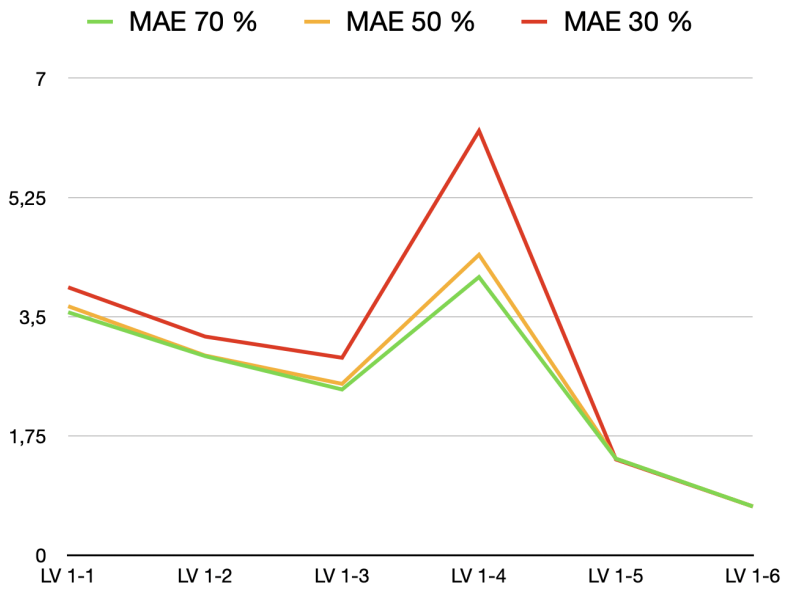


Abbildung 5.15: Vergleich der Prognosegüte (MAE) bei unterschiedlichem Anteil bekannter Haushalte

en mit 70% (grün), 50% (orange) und 30% (rot) bekannten Haushalten, wobei die x-Achse die sechs Niederspannungsabgänge des Transformators und die y-Achse den berechneten MAE in kW abbildet. Es ist erkennbar, dass die Fehlerverläufe über weite Teile des Datensatzes nahezu parallel verlaufen. Besonders zwischen den Szenarien mit 70% und 50% bekannten Haushalten bestehen nur geringe Abweichungen, was auf eine hohe Robustheit und Stabilität der Prognosemethode hinweist. Das Modell zeigt in diesen beiden Fällen ein konsistentes Verhalten, da die Fehlermetriken in ähnlicher Größenordnung liegen und keine systematischen Verschiebungen auftreten. Beim Szenario mit 30% bekannten Haushalten bleiben die allgemeinen Strukturen der Fehlerverläufe zwar erhalten, allerdings zeigen sich zunehmende Fehler und stärkere Diskrepanzen zu den anderen Szenarien. Dies weist darauf hin, dass die Modellgüte bei stark reduzierter Datengrundlage abnimmt und die Prognose empfindlicher auf zufällige Variationen oder untypische Verbrauchsmuster reagiert.

Insgesamt bestätigt die Abbildung, dass das Prognosemodell bis zur Reduktion auf 50% der bekannten Haushalte eine vergleichbare Vorhersagequalität beibehält. Bei einem Anteil von 30% ist ein Qualitätsverlust zu beobachten, der sich in einzelnen, stark erhöhten Fehlerwerten widerspiegelt. Dies zeigt, dass das Modell zwar robust gegenüber moderater

Datenreduktion ist, jedoch bei weiterer Reduktion die statistische Aussagekraft und Generalisierungsfähigkeit deutlich nachlässt.

5.3 Regressionsprognose

Um die Prognosegüte des Clusterings, besonders auf Haushaltsebene, weiter zu verbessern und die Modellierung gegenüber zufälligen Schwankungen robuster zu gestalten, werden im nächsten Schritt die Ergebnisse der Regressionsprognose bewertet. Auch hier erfolgt die Darstellung exemplarisch anhand des zuvor untersuchten Tages.

Die Regressionsprognose für den Transformator MV-1 in Abbildung 5.16 zeigt eine insgesamt sehr präzise Abbildung des tatsächlichen Lastverlaufs, unterscheidet sich allerdings in einigen Punkten von der Clusterprognose aus Abbildung 5.12. Auch hier verläuft die Regressionsprognose überwiegend annährend deckungsgleich mit der geplanten Leistungskurve. Die Vorhersage erfasst sowohl die Höhe als auch die zeitliche Lage vieler Peaks korrekt, was darauf hindeutet, dass das Modell kurzfristige Verbrauchsänderungen sehr gut darstellen kann. Die größten Abweichungen sind lediglich in der Amplitude der Spitzen punktuell und geringfügig vorhanden. Im direkten Vergleich mit der Clusterprognose fällt auf, dass beide Modelle auf dieser hohen Aggregationsebene eine ähnlich gute Prognosequalität liefern. Die Clusterprognose bildet die allgemeinen Tagesstrukturen ebenfalls zuverlässig ab. Kurzfristige Schwankungen werden jedoch tendenziell geglättet und es wird weniger dynamisch auf Laständerungen reagiert. Die Regressionsprognose hingegen zeigt eine feinere Auflösung und eine stärkere Dynamik, wodurch Peaks und Täler präziser abgebildet werden. Die Regressionsprognose bleibt jedoch tendenziell leicht oberhalb der tatsächlichen Werte und bildet eine systematische Überschätzung.

Um die Leistungsfähigkeit der Regressionsprognose zu beurteilen, können statt der aggregierten Transformator-Ebene einzelne Haushalte betrachtet werden. Auf der Transformator-Ebene wirken viele Glättungseffekte und individuelle Verbrauchsschwankungen einzelner Haushalte gleichen sich gegenseitig aus. Die Unterschiede zwischen dem Cluster- und Regressionsmodell sind zwar erkennbar, werden aber durch die hohe Aggregation weitgehend abgeschwächt. Auf der Haushaltsebene verlieren die natürlichen Mittelungseffekte ihre Wirkung und die Lastverläufe werden volatiler, unregelmäßiger und stärker von indi-

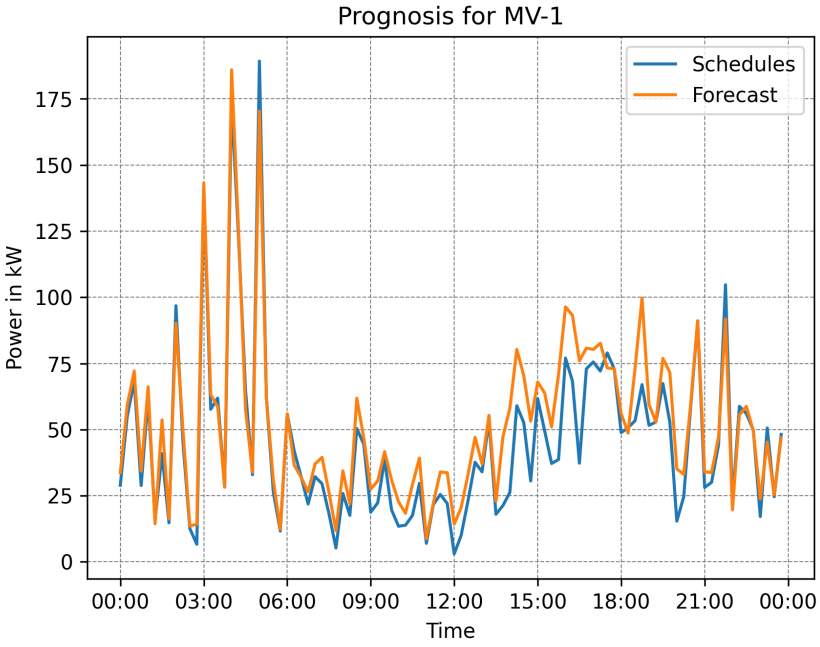
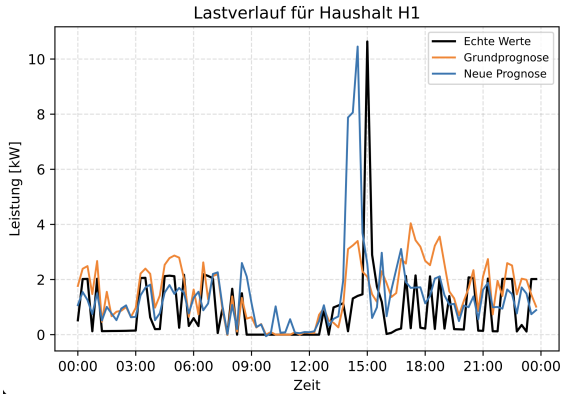


Abbildung 5.16: Vergleich von prognostiziertem und geplantem Lastverlauf für den Transformator MV-1

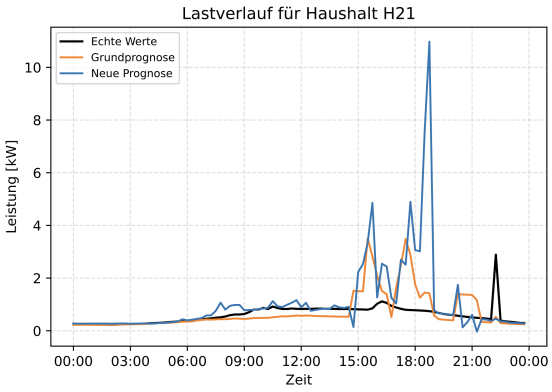
viduellen Verhaltensmustern beeinflusst. Im nächsten Schritt wird daher die Prognose auf einzelne Haushalte angewendet, um die Prognosen und die Verbesserung der Clusterprognose durch den Regressionsansatz besser bewerten zu können.

Die Abbildung 5.17 zeigt den Vergleich der Lastprognosekurven für einzelne Haushalte. In den Abbildungen sind jeweils die tatsächlich geplanten Fahrpläne (schwarz), die auf dem Clustering basierende Grundprognose (orange) sowie die daraus abgeleitete neue Prognose aus der Regressionsanalyse (blau) dargestellt. Ziel dieser Darstellung ist es, die Verbesserung der Prognosegüte durch die zusätzliche Regression auf Basis der Clusterprognosen zu veranschaulichen.

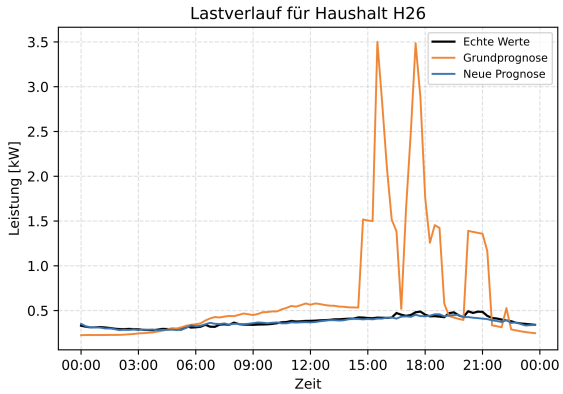
In den Abbildungen 5.17a bis 5.17d sind Haushalte dargestellt, die einem Festpreismodell folgen. Diese Haushalte zeigen meist ein stabiles, aber schwer vorhersagbares Verbrauchsverhalten, da ihre Lastprofile durch individuelle Gewohnheiten und zeitlich nicht klar strukturierte Verbrauchsspitzen geprägt sind. Bei diesen Haushalten zeigt sich deutlich, dass die reine Clusterprognose die tatsächlichen Lastverläufe nur unzureichend abbilden kann. Die Abweichungen zwischen tatsächlichem und prognostiziertem Verlauf sind teilweise erheblich, insbesondere bei plötzlichen Lastspitzen oder unregelmäßigen Verbrauchsereignissen.



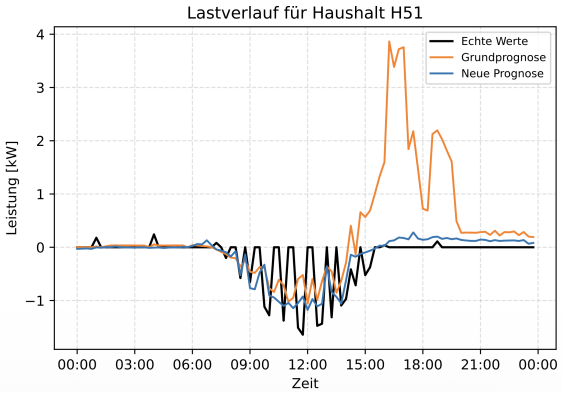
(a) Vergleich der Prognosekurven anhand des Haushalts H1



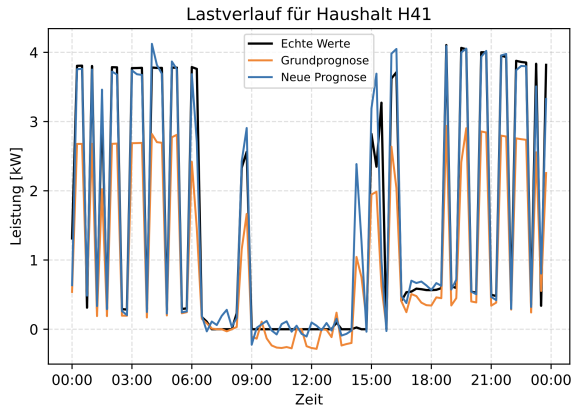
(b) Vergleich der Prognosekurven anhand des Haushalts H21



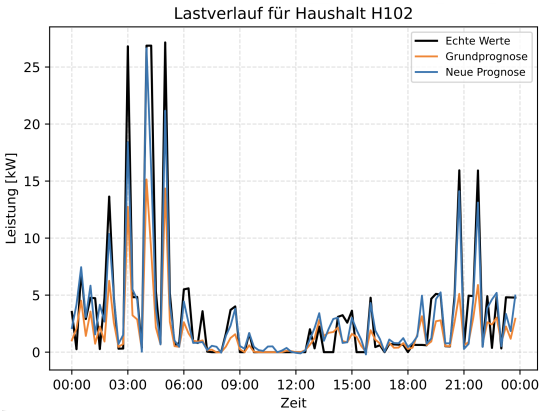
(c) Vergleich der Prognosekurven anhand des Haushalts H26



(d) Vergleich der Prognosekurven anhand des Haushalts H51



(e) Vergleich der Prognosekurven anhand des Haushalts H41



(f) Vergleich der Prognosekurven anhand des Haushalts H102

Abbildung 5.17: Vergleich der Prognosekurven anhand einzelner Haushalte

Die anschließende Regression auf Basis der Clusterprognose führt jedoch in fast allen Fällen zu einer sichtbaren Verbesserung der Prognosequalität. Während die Grundprognose in Abbildung 5.17a die Verbrauchsspitze um die Mittagszeit kaum erfasst, gelingt es der neuen Prognose, den Peak vorherzusagen, auch wenn er zeitlich leicht versetzt auftritt. Insgesamt

folgt die neue Prognose dem realen Verlauf deutlich präziser und bildet sowohl das Niveau als auch die Dynamik des Verbrauchs überzeugend ab. Bei Abbildung 5.17c zeigt sich ein anderes Muster. Während die Grundprognose noch ausgeprägte Peaks enthält, glättet die Regressionsanalyse den Verlauf stark und bildet das tatsächliche Verbrauchsprofil, das sich durch ein sehr niedriges und gleichmäßiges Niveau auszeichnet, wesentlich realistischer ab. Die Regression erkennt also die Tendenz zu einem konstant niedrigen Lastverlauf und korrigiert die Überschätzungen der Clusterprognose. Auch der Haushalt in Abbildung 5.17d profitiert von der Regressionsanpassung. Obwohl die Daten hier zum Teil unregelmäßige Werte aufweisen, gelingt es der neuen Prognose, den Verlauf zu stabilisieren und näher an den echten Werten auszurichten. Die Regression führt damit zu einer klaren Verbesserung der Vorhersagequalität gegenüber der Grundprognose.

Im Gegensatz dazu zeigt die Abbildung 5.17b ein gegenteiliges Verhalten. Hier verschlechtert die Regression das Ergebnis. Die neue Prognose sagt Peaks voraus, die in den tatsächlichen Werten nicht auftreten, während die echten Daten über den Tag hinweg auf einem relativ niedrigen Niveau bleiben. Dieses Verhalten ist bei Festpreishaushalten nicht ungewöhnlich, da dort Lastspitzen häufig ausbleiben. Die Regression scheint in diesem Fall eine Überanpassung an Clustertrends vorgenommen zu haben, die für diesen Haushalt nicht zutreffen.

Die Abbildungen 5.17e und 5.17f zeigen Haushalte, die einem variablen Preisverhalten folgen. Diese Haushalte reagieren dynamisch auf Strompreissignale, wodurch ihre Lastprofile strukturierter und stärker preissignalgetrieben sind. Bereits die Clusterprognose funktioniert in diesen Fällen besser, da sich ähnliche Verhaltensmuster innerhalb der Cluster stärker wiederholen. Dennoch kann die Regression auch hier eine weitere Feinabstimmung leisten. Die neue Prognose liegt nochmals näher an den echten Werten und bildet sowohl Lastspitzen als auch Lastsenken präziser ab.

Die Regressionsanalyse zeigt insgesamt ein deutliches Potenzial zur Verbesserung der Haushaltsprognosen in Kombination mit der vorgelagerten Clusteranalyse. In vielen Fällen gelingt es der Regressionsanalyse, die groben Trends der Clusterprognose zu verfeinern und systematische Abweichungen zu korrigieren.

Die Abbildung 5.18, die den Vergleich der Prognosegüte zwischen der Clusterprognose und der Regressionsprognose anhand des MAE zeigt, unterstreicht diesen positiven Effekt. Aus

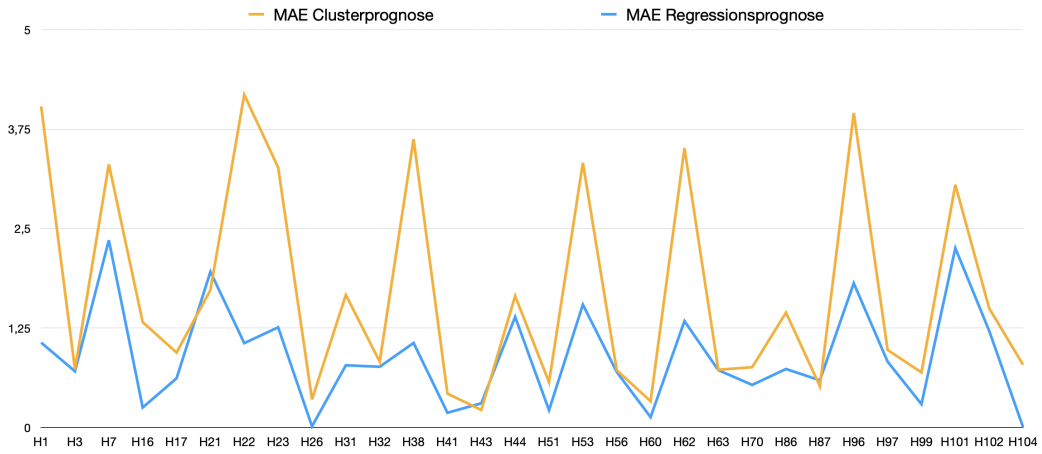


Abbildung 5.18: Vergleich der Prognosegüte (MAE) zwischen Cluster- und Regressionsprognose

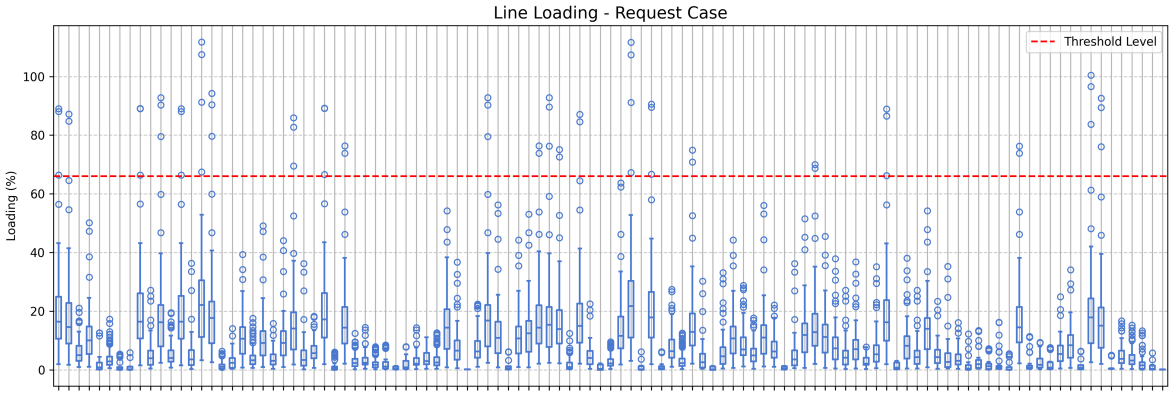
der Abbildung wird deutlich, dass die Regressionsanalyse in der Mehrheit der Fälle eine deutliche Verbesserung gegenüber der Clusterprognose erzielt. Bei nahezu allen Haushalten liegt die blaue Linie der Regressionsanalyse unterhalb der orangen Linie der Clusterprognose, was auf eine geringere mittlere Abweichung und damit auf eine höhere Prognosegenauigkeit hinweist. Lediglich in wenigen Fällen verläuft die Regressionslinie auf ähnlichem oder leicht höherem Niveau.

Ein weiterer Vorteil ist die Robustheit der Regressionsprognose, insbesondere im Zusammenhang mit dem variablen Preisverhalten. In diesen Fällen baut die Regression auf bereits gut strukturierten Clusterprognosen auf, da die Verbrauchsmuster dieser Haushalte stärker preissignalgetrieben und somit regelmäßiger sind. Dadurch kann die Regressionsanalyse die Prognosegüte weiter steigern und den tatsächlichen Fahrplan präziser abbilden. Gleichzeitig zeigt sich, dass die Regression auch bei Festpreishaushalten ein erhebliches Verbesserungspotenzial besitzt. Besonders in Abbildung 5.18 wird dieser Effekt sichtbar, da der zuvor stark variierende Verlauf der Prognosefehler der Clusteranalyse durch den Einsatz der Regression geglättet und stabilisiert wird. Während die Clusterprognose bei dieser Haushaltsgruppe aufgrund des individuellen und nicht preissensitiven Verbrauchsverhaltens häufig deutliche Abweichungen von den realen Fahrplänen aufweist, kann die Regression diese fehlerhaften Vorhersagen gezielt korrigieren. Dadurch wird eine Steigerung der Vorhersagequalität erreicht, insbesondere in Bezug auf das Verbrauchsniveau und die zeitliche Struktur der Lastverläufe.

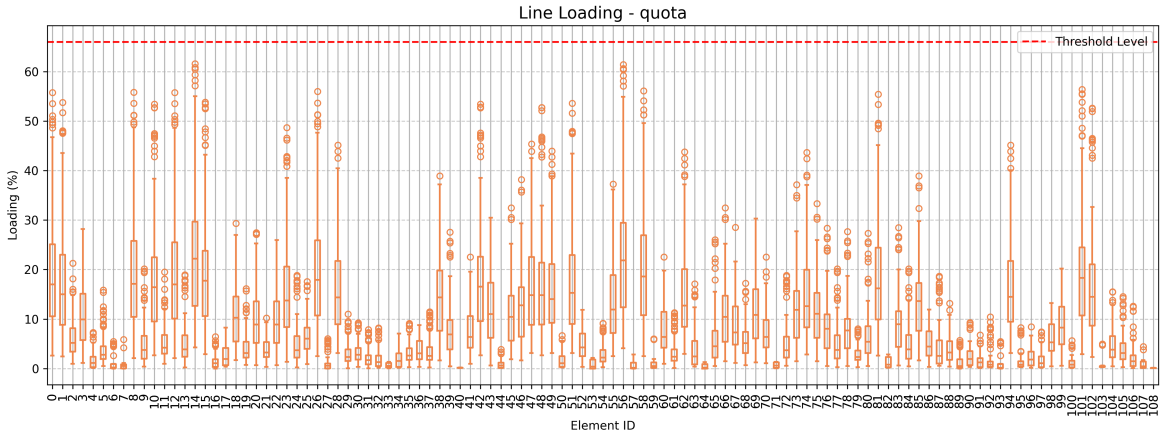
Trotz dieser positiven Ergebnisse weist der Prognosealgorithmus auch Grenzen auf. Die Regressionsanalyse zeigt die Tendenz, Clustertrends zu verallgemeinern, was insbesondere bei Haushalten mit atypischem oder schwach korreliertem Verbrauchsverhalten zu Überanpassungen führen kann. Dies äußert sich in zu hohen Prognosewerten, bei denen Lastspitzen vorhergesagt werden, die real nicht auftreten, oder in zu niedrigen Prognosewerten, welche tatsächliche Spitzen nicht erkennen. Beide Fehlerarten wirken sich unmittelbar auf die Genauigkeit der Lastplanung und damit auf die Netzkoordination aus. Die Ursachen für solche Fehlprognosen liegen meist in der begrenzten Generalisierungsfähigkeit des Regressionsmodelles. Wenn die Datenhistorie eines Haushalts nicht repräsentativ ist oder sich das Verbrauchsverhalten stark ändert, kann das Modell keine stabilen Zusammenhänge mehr erkennen. Auch nichtlineare Effekte oder kurzfristige Verhaltensänderungen, die sich aus individuellen Routinen ergeben, können durch diesen Regressionsansatz nur unzureichend erfasst werden. Besonders bei Festpreishaushalten, deren Lastprofile durch individuelle und schwer vorhersehbare Nutzungsmuster geprägt sind, stößt die Regressionsanalyse daher an ihre Grenzen.

5.4 Engpassmanagement Koordinierungsfunktion

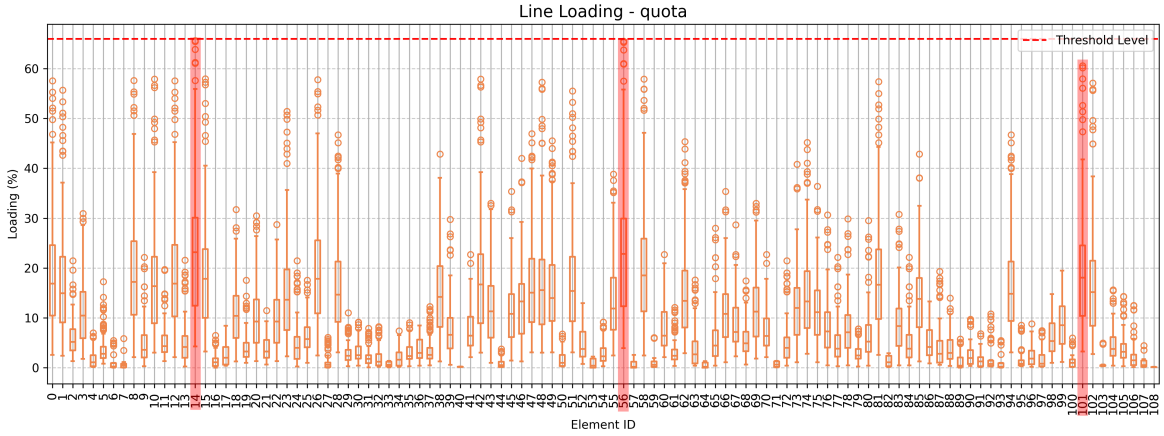
Die folgenden Ergebnisse veranschaulichen, inwiefern die KOF in der Lage ist, die bekannten Fahrpläne so anzupassen, dass trotz der bestehenden Prognoseunsicherheiten keine Engpässe im Netz entstehen. Um diese Fähigkeit zu bewerten, wird die Regressionsmethode mit einer perfekten Prognose verglichen. Während die Regression die reale Unsicherheit der Vorhersage abbildet und damit den praktischen Einsatzfall repräsentiert, dient die perfekte Prognose als theoretisches Optimum, bei dem die tatsächlichen Lastverläufe vollständig bekannt sind. Dadurch lässt sich beurteilen, wie stark Prognosefehler die Wirksamkeit der KOF beeinflussen. Damit wird untersucht, ob die Prognose selbst zusätzliche, künstliche Netzsituationen hervorruft, die im realen Betrieb ohne Vorhersagefehler gar nicht auftreten würden, oder ob die KOF die Unsicherheiten der Prognose ausreichend kompensieren kann. Ein solches Verhalten wäre für den praktischen Betrieb von zentraler Bedeutung, da die KOF in der Realität stets mit unvollständiger Information arbeitet und ihre Wirksamkeit daher stark von der Robustheit gegenüber Prognoseabweichungen abhängt.



(a) Leitungsauslastung im Ausgangszustand vor Anwendung der KOF



(b) Leitungsauslastung bei perfekter Prognose



(c) Leitungsauslastung bei regressionsbasierter Prognose

Abbildung 5.19: Vergleich der Leitungsauslastungen für angeforderte und angepasste Fahrpläne bei perfekter und regressionsbasierter Prognose

Die dargestellten Boxplots in Abbildung 5.19 zeigen die Verteilung der Leitungsauslastungen für alle Netzabschnitte unter zwei unterschiedlichen Szenarien. Auf der x-Achse sind die einzelnen Leitungen anhand ihrer Element-IDs dargestellt, während die y-Achse die jewei-

lige relative Auslastung in Prozent angibt. Die Boxplots veranschaulichen dabei die statistische Verteilung der Leitungsbelastungen, wobei die rote Linie die zulässige Auslastungsgrenze markiert. Die Abbildung 5.19a zeigt die Leitungsauslastungen im Ausgangszustand ohne Koordinierung. Die Auslastungen werden aus den gemeldeten Fahrplänen berechnet und bilden damit den Zustand vor Anwendung der KOF ab. Dabei zeigt sich, dass die von den Haushalten angeforderten Fahrpläne teilweise zu deutlichen Netzbelastungen führen. Mehrere Leitungen überschreiten den Schwellenwert, was auf lokal hohe Leistungsspitzen einzelner oder mehrerer Haushalte hinweist. Die breite Streuung der Boxplots verdeutlicht, dass diese Spitzen nicht gleichmäßig im Netz verteilt auftreten, sondern sich auf bestimmte Leitungen konzentrieren. Dies stellt den praktischen Ausgangszustand dar, in dem die Haushalte ihre individuellen Fahrpläne anmelden, ohne die resultierenden Netzbelastungen zu berücksichtigen. Die KOF passt die Fahrpläne anhand der in Kapitel 2.3 vorgestellten quoten-basierten Methode an. Der rote, gestrichelte Referenzstrich markiert die zulässige Auslastungsgrenze, oberhalb derer ein Engpass auftreten würde.

Der Referenzfall einer perfekten Prognose ist in Abbildung 5.19b dargestellt und zeigt das theoretische Optimum, bei dem die tatsächlichen Lastverläufe vollständig bekannt sind. Während die ursprünglich angeforderten Fahrpläne teils hohe Auslastungen aufweisen, reduziert die KOF nach der Anpassung die Belastung deutlich. Die resultierenden Werte der angepassten Fahrpläne liegen insgesamt auf einem niedrigen Niveau und Überschreitungen der zulässigen Auslastungsgrenze treten nicht auf. Besonders auffällig ist, dass die Streuung der Werte abnimmt, was darauf hinweist, dass die KOF nicht nur Überlastungen verhindert, sondern auch die Netzlast homogener verteilt. Dieses Ergebnis verdeutlicht, dass eine perfekte Prognose zu einer stabilen und gleichmäßigen Netzbelastung führt. Die KOF kann in diesem Fall die vorhandenen Netzkapazitäten optimal ausnutzen, ohne in kritische Betriebsbereiche zu gelangen.

Demgegenüber zeigt die Abbildung 5.19c das Szenario mit Prognoseunsicherheiten auf Basis der Regressionsmethode. Im Vergleich zur perfekten Prognose liegt die Auslastung vieler Leitungen insgesamt auf einem höheren Niveau. Dennoch kann die KOF die Netzbelastung deutlich reduzieren und eine stabile Betriebsführung gewährleisten. In einzelnen Bereichen, insbesondere bei den Leitungen mit den Element-IDs 14, 56 und 101, nähert sich die Auslastung jedoch dem kritischen Betriebsbereich an, verbleibt allerdings innerhalb der zuläs-

sigen Grenzwerte. Diese Leitungsabschnitte sind in der Abbildung durch rote Markierungen hervorgehoben und kennzeichnen potenziell sensible Netzbereiche. Solche Abweichungen entstehen, wenn Prognosefehler zu einer fehlerhaften Bewertung der tatsächlichen Netzsituation führen und die KOF dadurch suboptimale Fahrplananpassungen vornimmt. In der Folge kann es zu kurzzeitigen Überlastungen einzelner Leitungsabschnitte kommen, auch wenn die mittlere Auslastung insgesamt deutlich unter dem Grenzwert bleibt.

Das dargestellte Szenario stellt somit einen Grenzfall dar, bei dem die Netzbelastung nur knapp innerhalb der zulässigen Betriebsbereiche liegt. Schon geringfügig größere Prognoseabweichungen könnten dazu führen, dass Engpässe nicht mehr vollständig vermieden werden können und die Netzstabilität temporär gefährdet ist. Dies verdeutlicht die zentrale Bedeutung präziser Lastprognosen für den sicheren und effizienten Netzbetrieb. Eine hohe Prognosegüte ist entscheidend, da sie die Grundlage für eine verlässliche Fahrplananpassung und die koordinierte Steuerung dezentraler Leistungsflüsse bildet. Ungenaue Vorhersagen führen hingegen zu fehlerhaften Einschätzungen der verfügbaren Netzkapazitäten, wodurch sich die Auslastung ungleichmäßig verteilt und kritische Betriebspunkte wahrscheinlicher werden. Dennoch wird ersichtlich, dass die KOF in der Lage ist, die Unsicherheiten aus der Regressionsprognose teilweise zu kompensieren und auch unter nicht idealen Bedingungen ein stabiles Betriebsergebnis sicherzustellen.

6 Zusammenfassung und Ausblick

Die Arbeit zeigt, dass das modellierte Prognoseverfahren einen wesentlichen Beitrag zur Vorhersage von Haushaltslasten in Verteilnetzen leisten kann. Durch die Kombination von Clustering- und Regressionsansätzen können sowohl typische Verbrauchsmuster als auch individuelle Abweichungen präzise abgebildet werden. Besonders deutlich wird der Mehrwert der Methodik in der verbesserten Prognosegüte und der erhöhten Robustheit gegenüber unvollständigen oder heterogenen Eingangsdaten.

Forschungsfrage 1

Wie kann die Datenlücke von fehlenden Fahrplänen für die Koordinierungsfunktion vervollständigt werden?

Die Ergebnisse zeigen, dass das eingesetzte Clustering-Verfahren als eine tragfähige Grundlage für die Abbildung und Prognose typischer Lastverläufe dient und eine effektive Vervollständigung von fehlenden Fahrplänen für die KOF ermöglicht. Durch die Kategorisierung der Haushaltskurven aus bekannten Fahrplandaten zu charakteristischen Clustern können gute Repräsentativprofile gebildet werden, welche eine geeignete Basis für unbekannte Hauhsalte darstellen.

Für Haushalte, bei denen keine Fahrpläne vorliegen, ermöglicht dieses Verfahren eine datenbasierte Zuordnung zu demjenigen Cluster, dessen Profil dem bekannten Verbrauchsverhalten am ähnlichsten ist. So lässt sich für jeden unbekannten Haushalt ein plausibler Lastverlauf ableiten, der auf den Mustern ähnlicher Haushalte basiert. Dadurch wird die Datenbasis für die KOF erweitert und homogenisiert. Somit kann hinsichtlich der ersten Forschungsfrage geschlussfolgert werden, dass die systematische Bildung und Nutzung von Repräsentativprofilen aus Clustering-Ergebnissen eine effektive Methode zur Vervollständigung fehlender Fahrplandaten für die KOF darstellt.

Forschungsfrage 2

Wie gut können die Fahrplandaten von unbekannten Haushalten mit einem Clustering-Algorithmus prognostiziert werden?

Die Ergebnisse zeigen, dass das Clustering-Verfahren besonders auf aggregierter Ebene eine hohe Prognosegüte aufzeigt. Die gebildeten Repräsentativkurven erfassen dabei die typischen Verbrauchsdynamiken sehr zuverlässig und bilden das Gesamtverhalten der betrachteten Haushaltsgruppen realitätsnah ab. Hier kompensiert die Zusammenführung mehrerer Haushalte individuelle Verbrauchsabweichungen, wodurch sich ein stabiler Gesamtverlauf ergibt. Mit zunehmender Detaillierung auf der Haushaltsebene nimmt die Genauigkeit erwartungsgemäß ab, da individuelle Verbrauchsbesonderheiten und nichtlineare Effekte einen stärkeren Einfluss haben, der in der Clusterbildung nur begrenzt abgebildet werden kann.
Die Ergebnisse zeigen zudem, dass das Verbrauchsverhalten der Haushalte stark vom zugrunde liegenden Tarifszenario beeinflusst wird. Während im Festpreismodell ein gleichmäßiges und weitgehend trägheitsbehaftetes Lastverhalten dominiert, reagieren Haushalte im variablen Preismodell deutlich sensibler auf Preissignale. Das vorgelagerte Preisclustering erweist sich dabei als besonders gut geeignet für die Verbesserung der Trennschärfe innerhalb der Cluster. Durch die Unterscheidung zwischen Festpreis- und variablen Preishaushalten kann die Prognosegüte einzelner Haushalte deutlich gesteigert werden, da die jeweiligen Reaktionsmuster auf Preisschwankungen gezielter erfasst werden. Auch in den Sommermonaten, in denen die Haushalte als Kollektiv und nicht getrennt gruppiert werden, erzielt das Verfahren gute Ergebnisse. Das zeigt, dass die Clusterbildung auf Basis der Verbrauchsstruktur auch ohne ausgeprägte Preissignale robuste Prognosen liefern kann. Darüber hinaus zeigt das Modell auch bei reduzierter Datenverfügbarkeit eine hohe Stabilität. Auch wenn nur etwa 50% der Haushalte mit bekannten Fahrplandaten in die Clusterbildung einfließen, bleibt die Prognosequalität auf aggregierter Ebene weitgehend erhalten. Erst bei einem Anteil von rund 30% bekannter Haushalte nimmt die Genauigkeit sichtbar ab. Es kann angenommen werden, dass die verbleibende Datenbasis nicht mehr ausreicht, um die Vielfalt der Verbrauchsstrukturen angemessen abzubilden.

Forschungsfrage 3

Wie gut lässt sich die Genauigkeit der clusterbasierten Lastprognose durch die Einbindung zusätzlicher Netzdaten verbessern?

Die dritte Forschungsfrage wird durch eine nachgelagerte Regressionsanalyse untersucht, welche die bestehenden Clusterprognosen um zusätzliche Kontextinformationen erweitert. Durch diese Verknüpfung von Clusterergebnissen mit erklärenden Variablen wie Wetter, Preisen und zeitlichen Mustern gelingt eine deutliche Steigerung der Prognosequalität auf Haushaltsebene. Die Kombination beider Ansätze erweist sich als besonders effektiv, da die Regression auf den strukturierten Clusterdaten basiert und diese um individuelle Verbrauchseinflüsse erweitert. Auf diese Weise können systematische Muster im Lastverhalten besser erfasst und Unterschiede zwischen Haushalten innerhalb eines Clusters gezielter abgebildet werden. Besonders Haushalte mit variablem Preisverhalten profitieren von dieser Methodik, da ihre dynamischen Reaktionen auf Preissignale präziser nachgebildet werden können. Gleichzeitig zeigt sich, dass die Regressionsanalyse auch für Festpreishaushalte einen Mehrwert bietet, da sie deren meist träge und unregelmäßige Lastverläufe besser an die tatsächlichen Verbrauchsmuster anpasst. Während die clusterbasierte Prognose bei diesen Haushalten häufig zu starr oder ungenau ausfällt, kann die Regression durch die Einbeziehung zusätzlicher Einflussgrößen typische Verbrauchsmuster besser erfassen und die Vorhersage deutlich verbessern. Dadurch werden insbesondere die in der Clusteranalyse auftretenden systematischen Abweichungen reduziert, was zu stabileren und realitätsnäheren Lastverläufen führt.

Allerdings stößt das Verfahren besonders bei Festpreishaushalten auch an seine Grenzen. Da deren Verbrauchsverhalten weniger durch äußere Faktoren, sondern stärker durch individuelle Routinen oder spontane Nutzungsspitzen geprägt ist, lassen sich diese Schwankungen nur eingeschränkt modellieren. Unvorhersehbare Laständerungen, etwa durch unregelmäßige Haushaltsaktivitäten oder variierende Ladevorgänge, können von der Regressionsanalyse nicht zuverlässig vorhergesagt werden. In solchen Fällen glättet das Modell zwar extreme Ausreißer, erfasst jedoch die tatsächliche Dynamik nur bedingt. Damit verbleiben im Testnetz Haushalte, deren stark individuelles oder unregelmäßiges Verbrauchsverhalten eine zuverlässige Prognose nur eingeschränkt zulässt. Dennoch zeigt sich, dass die Kombi-

nation aus Clustering und Regression eine robuste Methode zur Lastprognose darstellt, die insbesondere auf Einzelhaushaltsebene eine signifikante Genauigkeitssteigerung durch die Einbindung zusätzlicher Netzdaten ermöglicht.

Forschungsfrage 4

Wie gut unterstützt die entwickelte Lastprognose die Koordinierungsfunktion im Verteilnetz und reduziert Engpässe im Vergleich zu einer idealen Prognose?

Die Ergebnisse der vierten Forschungsfrage zeigen, dass die entwickelte Lastprognose die KOF im Verteilnetz wirksam unterstützt und Engpässe weitgehend vermeiden kann. Trotz der unvermeidbaren Prognoseunsicherheiten, die durch den Regressionsansatz entstehen, bleibt die KOF in der Lage, die resultierenden Netzbelastungen gezielt zu glätten und kritische Überlastungen durch Fahrplananpassungen zu verhindern. Der Vergleich mit einer idealen, fehlerfreien Prognose verdeutlicht jedoch, dass die Regressionsmethode zu einer erhöhten Variabilität der Leitungsauslastungen führt und dass insbesondere in einzelnen Netzbereichen kritische Betriebszustände temporär auftreten können. Diese Situationen verdeutlichen, dass die KOF Prognosefehler nur eingeschränkt kompensieren kann und bei unzureichender Prognosequalität an ihre Grenzen stößt.

Gleichzeitig zeigen die Ergebnisse, dass die Wirksamkeit der KOF direkt mit der Prognosequalität korreliert. Je präziser die Vorhersagen sind, desto effizienter gelingt die Lastverteilung und desto besser wird die Netzstabilität gesichert. Im Gegensatz dazu können bei größeren Prognoseabweichungen kritische Netzsituationen vereinzelt bestehen bleiben oder sich durch die Regressionsprognose sogar verschärfen. Damit zeigt sich, dass eine hohe Prognosequalität eine zentrale Voraussetzung für die zuverlässige Funktion der KOF und die langfristige Netzstabilität darstellt.

6.1 Ausblick

Die in dieser Arbeit genutzten Fahrpläne basieren auf simulierten Verbrauchsverläufen, die unter Verwendung realer Preis- und Wetterdaten generiert wurden. Dadurch konnte ein weitgehend realistisches Szenario abgebildet werden, das jedoch nicht vollständig das Verhalten realer Haushalte widerspiegelt. Eine breitere Verfügbarkeit realer Messdaten, insbesondere durch den verstärkten Ausbau intelligenter Messsysteme, könnte in zukünftiger Forschung eine genauere Validierung und Kalibrierung der Modelle ermöglichen. Der zunehmende Smart-Meter-Rollout bietet hier eine entscheidende Grundlage, um die zeitliche und räumliche Auflösung der Datengrundlage zu verbessern und damit die Prognosequalität weiter zu steigern. Dies würde insbesondere die Nachvollziehbarkeit von kurzfristigen Verbrauchsänderungen und Reaktionen auf Preissignale verbessern.

Hinsichtlich der Prognosemethodik wurde sich in dieser Arbeit für ein baumbasiertes Regressionsverfahren auf Basis von Gradient Boosting entschieden, da die zugrunde liegenden Daten in strukturierter, tabellarischer Form vorliegen und der Algorithmus sowohl interpretierbare als auch robuste Ergebnisse liefert. Dennoch könnten weiterführende Untersuchungen prüfen, inwieweit Deep-Learning-Modelle zusätzliche Muster und nichtlineare Zusammenhänge erfassen können, die über klassische Regressionsverfahren hinausgehen. Diese Modelle können langfristige Abhängigkeiten und saisonale Muster automatisch erkennen und damit insbesondere bei träge reagierenden Haushalten eine deutlich höhere Prognosepräzision erreichen. Insbesondere wiederkehrende Routinen und tageszeitabhängige Verbrauchsmuster könnten so besser erlernt werden.

Literaturverzeichnis

[1] M. Vogel, *Flexibilität für das Netz Vergleich und Bewertung von Koordinationsmechanismen für den netzdienlichen Einsatz von Flexibilität*. Freiburg: Öko-Institut e.V., Apr. 2020. [Elektronische Quelle]. Adresse: https://www.oeko.de/fileadmin/oekodoc/Flexibilitaet-fuer-das-Netz.pdf

[2] F. Nußbaum, A.-L. Steen, P. T. Baboli, und C. Becker, “Coordination of Demand-Side Flexibility Resources for Preventive Congestion Management in Distribution Systems,” *ETG-Fachberichte*, 2025.

[3] P. Godron, M. Herrndorff, und S. Müller, “Haushaltsnahe Flexibilitäten nutzen: Wie Elektrofahrzeuge, Wärmepumpen und Co. die Stromkosten für alle senken können,” 2023. [Elektronische Quelle]. Adresse: https://www.agora-energiewende.de/publikationen/haushaltsnahe-flexibilitaeten-nutzen

[4] M. Wolter, “Flexibilisierung des Energiesystems,” 2023. [Elektronische Quelle]. Adresse: https://www.vde.com/de/etg/arbeitsgebiete/v2/flexibilisierung-des-energiesystems

[5] H. Schermeyer, “Netzengpassmanagement in regenerativ geprägten Energiesystemen,” Aug. 2018. [Elektronische Quelle]. Adresse: https://publikationen.bibliothek.kit.edu/1000086513

[6] Bundesnetzagentur, “Bundesnetzagentur - §14a EnWG Steuerbare Verbrauchseinrichtungen.” [Elektronische Quelle]. Adresse: https://www.bundesnetzagentur.de/DE/Beschlusskammern/BK06/BK6_83_Zug_Mess/841_SteuVE/BK6_SteuVE_node.html. (besucht am: 28.05.2025)

[7] Bundesministerium für Wirtschaft und Klimaschutz, “Roadmap Systemstabilität – Fahrplan zur Erreichung eines sicheren und robusten Betriebs des zukünftigen Stromversorgungssystems mit 100

[8] F. Nußbaum und A.-L. Steen, “IEET: KoLa (BMWK),” 2025. [Elektronische Quelle]. Adresse: https://www.tuhh.de/ieet/forschung/forschungsprojekte/kola-bmwk

[9] Bundesnetzagentur, “Monitoringbericht 2023,” *Bundesnetzagentur*, Nov. 2023. [Elektronische Quelle]. Adresse: https://data.bundesnetzagentur.de/Bundesnetzagentur/SharedDocs/Mediathek/Monitoringberichte/MonitoringberichtEnergie2023.pdf

[10] Bundesnetzagentur, “Flexibilität im Stromversorgungssystem Bestandsaufnahme, Hemmnisse und Ansätze zur verbesserten Erschließung von Flexibilität,” Apr. 2017. [Elektronische Quelle]. Adresse: https://www.bundesnetzagentur.de/DE/Fachthemen/ElektrizitaetundGas/VerteilerNetz/Flexibilitaet/BNetzA_Flexibilitaetspapier.pdf?__blob=publicationFile&v=1

[11] R. Brandalik, “Ein Beitrag zur Zustandsschätzung in Niederspannungsnetzen mit niedrigredundanter Messwertaufnahme,” *Fachbereich Elektrotechnik und Informationstechnik der Technischen Universität Kaiserslautern*, 2020. [Elektronische Quelle]. Adresse: https://kluedo.ub.rptu.de/frontdoor/deliver/index/docId/5977/file/Brandalik_Dissertation_Kluedo.pdf

[12] S. Haben, S. Arora, G. Giasemidis *et al.*, “Review of low voltage load forecasting: Methods, applications, and recommendations,” *Applied Energy*, Bd. 304, S. 117798, 2021.

[13] SINTEG - Schaufenster intelligente Energie - Digitale Agenda für die Energiewende, “6.2 Blaupause 16: Netzzustandsprognosen für das Verteilnetz.” [Elektronische Quelle]. Adresse: https://www.bmwk.de/Redaktion/DE/Dossier/Sinteg/6-2-blaupause.pdf?__blob=publicationFile&v=1

[14] F. Mahr, S. Henninger, M. Biller, und J. Jäger, *Elektrische Energiesysteme: Wissensvernetzung von Stromrichter, Netzbetrieb und Netzschutz*. Wiesbaden: Springer Fachmedien Wiesbaden, 2021. [Elektronische Quelle]. Adresse: https://link.springer.com/10.1007/978-3-658-34908-0

[15] W. Holderbaum, F. Alasali, und A. Sinha, *Energieprognose und Steuerungsmethoden für Energiespeichersysteme in Verteilungsnetzen: Prädiktive Modellierung und Kontrolltechniken*. Cham: Springer International Publishing, 2023. [Elektronische Quelle]. Adresse: https://link.springer.com/10.1007/978-3-031-45471-4

[16] T. Steffen, B. Wiegel, und C. Becker, "BEITRAG ZUR TRANSFORMATION DES DEUTSCHEN VERTEIL- NETZES UND AUSWIRKUNGEN GESETZLICHER ÄNDERUNGEN," *18. Symposium Energieinnovation*, 2024.

[17] Bundesnetzagentur. [Elektronische Quelle]. Adresse: https://www.bundesnetzagentur.de/DE/Fachthemen/ElektrizitaetundGas/Versorgungssicherheit/Netzengpassmanagement/start.html. (besucht am: 03.06.2025)

[18] V. Regener, N. Maas, M. Hinterstocker, und P. Mergner, *Anreizmechanismen zur Stromnetzentlastung*. FfE - Forschungsstelle für Energiewirtschaft e.V., Feb. 2025.

[19] M. Grosse, H. Send, und T. Loitz, "Smart Energy in Deutschland: Wie Nutzerinnovationen Die Energiewende Voranbringen (Smart Energy in Germany: The Benefits of User Innovation)," *SSRN Electronic Journal*, 2018. [Elektronische Quelle]. Adresse: https://www.ssrn.com/abstract=3249002

[20] F. Alfaverh, M. Denai, und Y. Sun, "Demand Response Strategy Based on Reinforcement Learning and Fuzzy Reasoning for Home Energy Management," *IEEE Access*, Bd. 8, S. 39310–39321, 2020.

[21] J. Rücker, "Optimal Scheduling of Flexible Components in Residential Neighborhoods Using Detailed Linear Programming," *Technische Universität Hamburg*, 2024.

[22] D. T. Kusuma, N. Ahmad, und S. S. S. Ahmad, "CLUSTERING ALGORITHM FOR ELECTRICAL LOAD PROFILING ANALYSIS: A SYSTEMATIC REVIEW OF MACHINE LEARNING APPROACHES FOR IMPROVED CLUSTERING ALGORITHMS," *. Vol.*, Nr. 10, 2024.

[23] K. Seo, H. S. Na, W. Lee *et al.*, "Clustering electricity consumption patterns using functional data analysis," *Sustainable Energy, Grids and Networks*, Bd. 43, S. 101742, 2025.

[24] F. AlMahamid und K. Grolinger, “Agglomerative Hierarchical Clustering with Dynamic Time Warping for Household Load Curve Clustering,” in *2022 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, Sep. 2022, S. 241–247, arXiv:2210.09523 [cs]. [Elektronische Quelle]. Adresse: http://arxiv.org/abs/2210.09523

[25] L. Sun, K. Zhou, und S. Yang, “An ensemble clustering based framework for household load profiling and driven factors identification,” *Sustainable Cities and Society*, Bd. 53, S. 101958, Feb. 2020.

[26] M. Afzalan, F. Jazizadeh, und H. Eldardiry, “Two-Stage Clustering of Household Electricity Load Shapes for Improved Temporal Pattern Representation,” *IEEE Access*, Bd. 9, S. 151667–151680, 2021.

[27] H. C. Jeong, M. Jang, T. Kim, und S.-K. Joo, “Clustering of Load Profiles of Residential Customers Using Extreme Points and Demographic Characteristics,” *Electronics*, Bd. 10, Nr. 3, S. 290, Jan. 2021.

[28] K. Yu, J. Cao, X. Chen *et al.*, “Residential load forecasting based on electricity consumption pattern clustering,” *Frontiers in Energy Research*, Bd. 10, S. 1113733, Jan. 2023.

[29] P. Laurinec und M. Lucká, “Clustering-based forecasting method for individual consumers electricity load using time series representations,” *Open Computer Science*, Bd. 8, Nr. 1, S. 38–50, Jul. 2018.

[30] A. K. Singh, Ibraheem, S. Khatoon *et al.*, “Load forecasting techniques and methodologies: A review,” in *2012 2nd International Conference on Power, Control and Embedded Systems*, 2012, S. 1–10.

[31] H.-J. Bae, J.-S. Park, J.-h. Choi, und H.-Y. Kwon, “Learning model combined with data clustering and dimensionality reduction for short-term electricity load forecasting,” *Scientific Reports*, Bd. 15, Nr. 1, S. 3575, Jan. 2025.

[32] G. Rouwhorst, E. M. S. Duque, P. H. Nguyen, und H. Slootweg, “Improving Clustering-Based Forecasting of Aggregated Distribution Transformer Loadings With Gradient Boosting and Feature Selection,” *IEEE Access*, Bd. 10, S. 443–455, 2022.

[33] D. Kontogiannis, D. Bargiotas, A. Daskalopulu *et al.*, “Structural Ensemble Regression for Cluster-Based Aggregate Electricity Demand Forecasting,” *Electricity*, Bd. 3, Nr. 4, S. 480–504, 2022.

[34] J. P. Bharadiya, “A Tutorial on Principal Component Analysis for Dimensionality Reduction in Machine Learning,” *International Journal of Innovative Science and Research Technology*, Bd. 8, Nr. 5, 2023.

[35] M. Greenacre, P. J. F. Groenen, T. Hastie *et al.*, “Principal component analysis,” *Principal Component Analysis*, 2023.

[36] J. Healy und L. McInnes, “Uniform manifold approximation and projection,” *Nature Reviews Methods Primers*, Bd. 4, Nr. 1, S. 82, Nov. 2024.

[37] C. Si, S. Xu, C. Wan *et al.*, “Electric Load Clustering in Smart Grid: Methodologies, Applications, and Future Trends,” *Journal of Modern Power Systems and Clean Energy*, Bd. 9, Nr. 2, S. 237–252, 2021.

[38] A. K. Jain, M. N. Murty, und P. J. Flynn, “Data clustering: a review,” *ACM Computing Surveys*, Bd. 31, Nr. 3, S. 264–323, Sep. 1999.

[39] C. Yuan und H. Yang, “Research on K-Value Selection Method of K-Means Clustering Algorithm,” *J*, Bd. 2, Nr. 2, S. 226–235, Jun. 2019.

[40] M. Ahmed, R. Seraj, und S. M. S. Islam, “The k-means Algorithm: A Comprehensive Survey and Performance Evaluation,” *Electronics*, Bd. 9, Nr. 8, S. 1295, Aug. 2020.

[41] G. Gan, C. Ma, und J. Wu, *Data Clustering: Theory, Algorithms, and Applications, Second Edition*. Philadelphia, PA: Society for Industrial and Applied Mathematics, Jan. 2020. [Elektronische Quelle]. Adresse: https://epubs.siam.org/doi/book/10.1137/1.9781611976335

[42] H. V. Singh, A. Girdhar, und S. Dahiya, “A Literature survey based on DBSCAN algorithms,” in *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*. Madurai, India: IEEE, 2022, S. 751–758. [Elektronische Quelle]. Adresse: https://ieeexplore.ieee.org/document/9788440/

[43] R. Campello, D. Moulavi, J. Sander *et al.*, “How HDBSCAN Works,” HDBSCAN Documentation, 2016. [Elektronische Quelle]. Adresse: https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html. (besucht am: 25.09.2025)

[44] R. J. G. B. Campello, D. Moulavi, und J. Sander, *Density-Based Clustering Based on Hierarchical Density Estimates*, Ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, Bd. 7819, S. 160–172. [Elektronische Quelle]. Adresse: http://link.springer.com/10.1007/978-3-642-37456-2_14

[45] G. Stewart und M. Al-Khassaweneh, “An Implementation of the HDBSCAN* Clustering Algorithm,” *Applied Sciences*, Bd. 12, Nr. 5, S. 2405, Feb. 2022.

[46] T. Hastie, R. Tibshirani, und J. Friedman, *The Elements of Statistical Learning*, Ser. Springer Series in Statistics. New York, NY: Springer New York, 2009. [Elektronische Quelle]. Adresse: http://link.springer.com/10.1007/978-0-387-84858-7

[47] scikit-learn developers, “Comparing different clustering algorithms on toy datasets,” https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html, 2024. (besucht am: 19.08.2025),

[48] U. v. Luxburg, “A Tutorial on Spectral Clustering,” *Statistics and Computing*, Nr. arXiv:0711.0189, Nov. 2007, arXiv:0711.0189 [cs]. [Elektronische Quelle]. Adresse: http://arxiv.org/abs/0711.0189

[49] I. K. Khan, H. B. Daud, N. B. Zainuddin *et al.*, “Determining the optimal number of clusters by Enhanced Gap Statistic in K-mean algorithm,” *Egyptian Informatics Journal*, Bd. 27, S. 100504, Sep. 2024.

[50] L. Fahrmeir, T. Kneib, S. Lang, und B. D. Marx, *Regression: Models, Methods and Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2021. [Elektronische Quelle]. Adresse: https://link.springer.com/10.1007/978-3-662-63882-8

[51] G. Biau und E. Scornet, “A Random Forest Guided Tour,” Nr. arXiv:1511.05741, Nov. 2015, arXiv:1511.05741 [math]. [Elektronische Quelle]. Adresse: http://arxiv.org/abs/1511.05741

[52] C. Bentéjac, A. Csörgő, und G. Martínez-Muñoz, “A Comparative Analysis of XGBoost,” *Artificial Intelligence Review*, Bd. 54, Nr. 3, S. 1937–1967, 2021, arXiv:1911.01914 [cs].

[53] B. Clark und F. Lee, “What is Gradient Boosting?” https://www.ibm.com/think/topics/gradient-boosting, IBM, 2025. (besucht am: 22.09.2025),

[54] A. Natekin und A. Knoll, “Gradient boosting machines, a tutorial,” *Frontiers in Neurorobotics*, Bd. 7, 2013. [Elektronische Quelle]. Adresse: http://journal.frontiersin.org/article/10.3389/fnbot.2013.00021/abstract

[55] L. Zhang und D. Jánošík, “Enhanced short-term load forecasting with hybrid machine learning models: CatBoost and XGBoost approaches,” *Expert Systems with Applications*, Bd. 241, S. 122686, 2024.

[56] S. Simaiya, M. Dahiya, S. Tomar *et al.*, “A transfer learning-based hybrid model with LightGBM for smart grid short-term energy load prediction,” *Energy Exploration Exploitation*, Bd. 42, Nr. 5, S. 1853–1876, Sep. 2024.

[57] X. Kong, Z. Chen, W. Liu *et al.*, “Deep learning for time series forecasting: a survey,” *International Journal of Machine Learning and Cybernetics*, Bd. 16, Nr. 7–8, S. 5079–5112, Aug. 2025.

[58] R. Shwartz-Ziv und A. Armon, “Tabular Data: Deep Learning is Not All You Need,” Nr. arXiv:2106.03253, Nov. 2021, arXiv:2106.03253 [cs]. [Elektronische Quelle]. Adresse: http://arxiv.org/abs/2106.03253

[59] T. O. Hodson, “Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not,” *Geoscientific Model Development*, Bd. 15, Nr. 14, S. 5481–5487, Jul. 2022.

[60] A. Mystakidis, P. Koukaras, N. Tsalikidis *et al.*, “Energy Forecasting: A Comprehensive Review of Techniques and Technologies,” *Energies*, Bd. 17, Nr. 7, S. 1662, Mar. 2024.

[61] A. R. Lahitani, A. E. Permanasari, und N. A. Setiawan, “Cosine similarity to determine similarity measure: Study case in online essay assessment,” in *2016 4th International Conference on Cyber and IT Service Management*. Bandung, Indonesia: IEEE, Apr. 2016, S. 1–6. [Elektronische Quelle]. Adresse: http://ieeexplore.ieee.org/document/7577578/

[62] C. Flygare, A. Wallberg, E. Jonasson *et al.*, “Correlation as a method to assess electricity users’ contributions to grid peak loads: A case study,” *Energy*, Bd. 288, S. 129805, Feb. 2024.

[63] J. Büchner, J. Katzfey, und O. Flörcken, “Moderne Verteilernetze für Deutschland,” Sep. 2014. [Elektronische Quelle]. Adresse: https://www.bundeswirtschaftsministerium.de/Redaktion/DE/Publikationen/Studien/verteilernetzstudie.pdf?__blob=publicationFile&v=1

Abbildungsverzeichnis

Tabellenverzeichnis

Anhang

A Lastkurven-Clustering

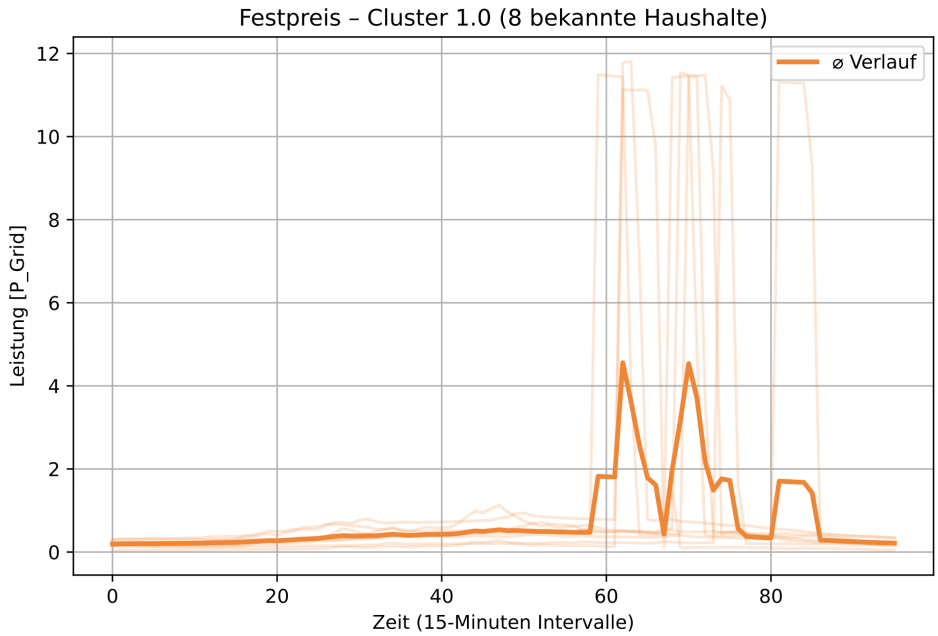


Abbildung A.1: Darstellung der individuellen Lastkurven Cluster 1 (Festpreisverhalten)

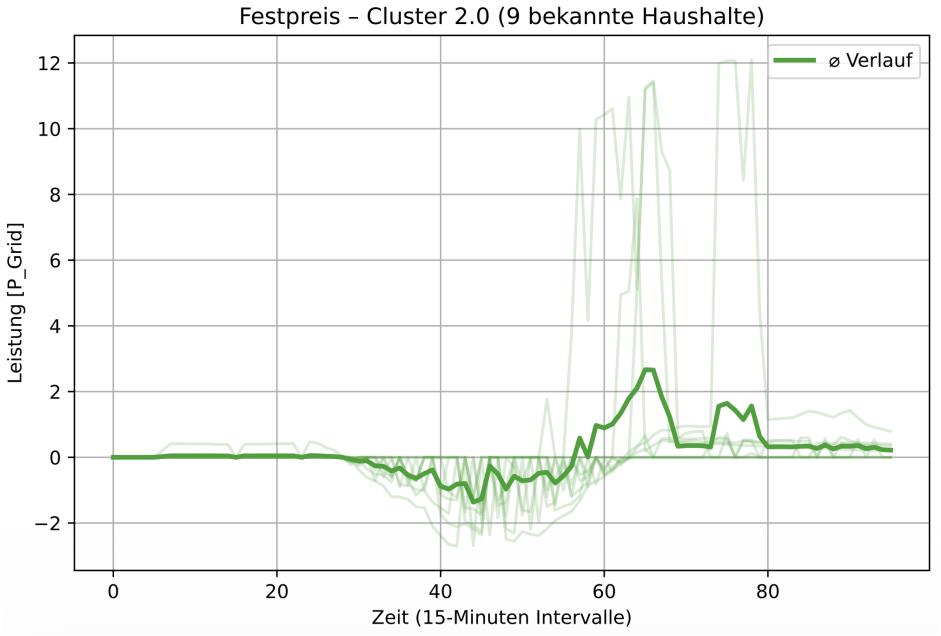


Abbildung A.2: Darstellung der individuellen Lastkurven Cluster 2 (Festpreisverhalten)

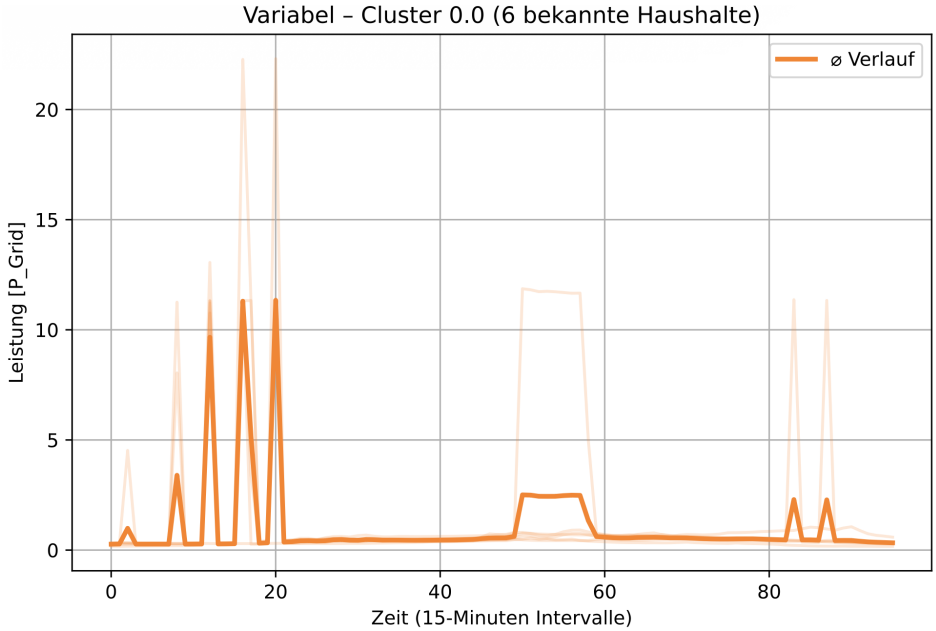


Abbildung A.3: Darstellung der individuellen Lastkurven Cluster 0 (variables Preisverhalten)

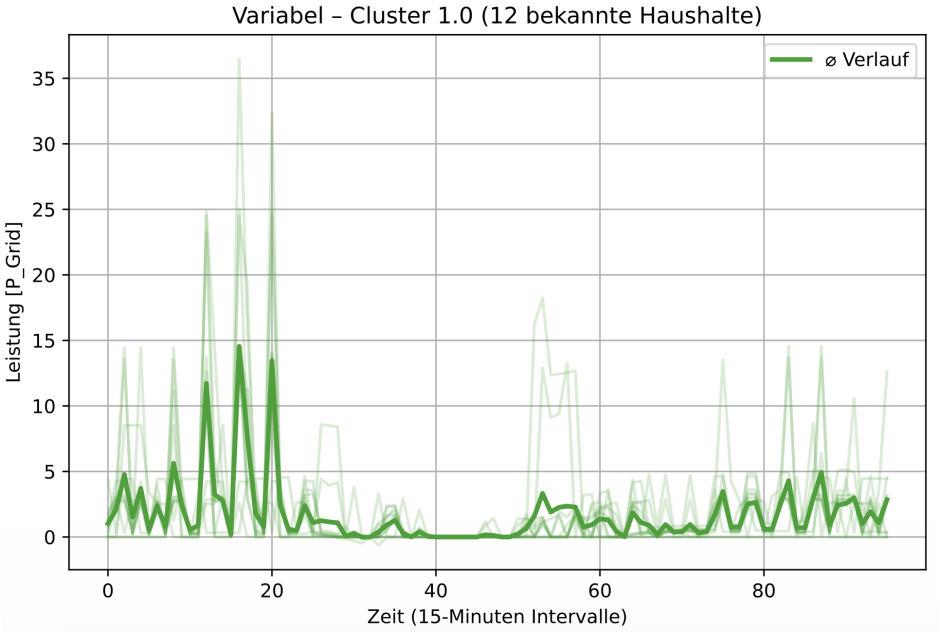


Abbildung A.4: Darstellung der individuellen Lastkurven Cluster 1 (variables Preisverhalten)

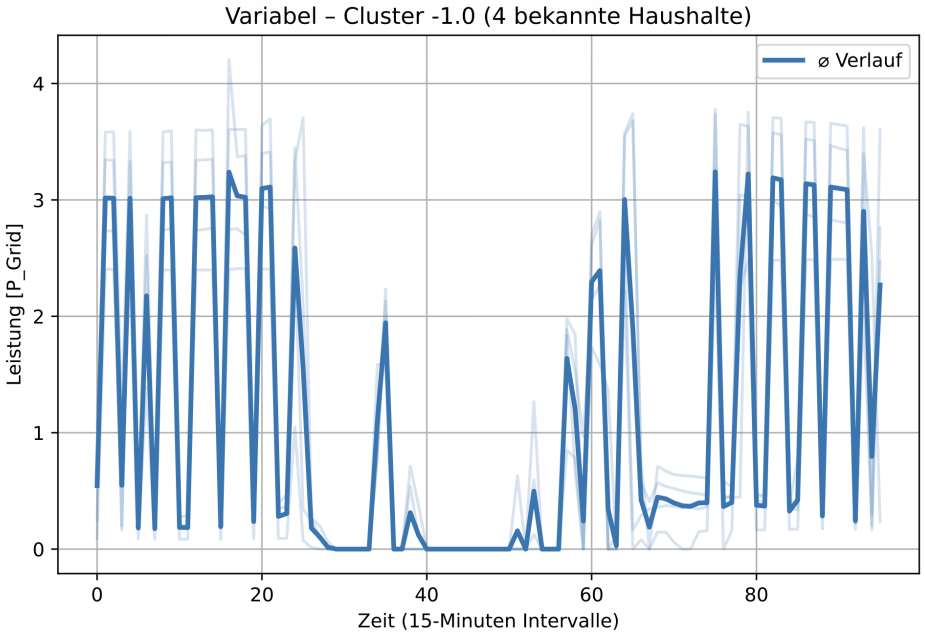


Abbildung A.5: Darstellung der individuellen Lastkurven Cluster -1 (variables Preisverhalten)

B Clusterprognose

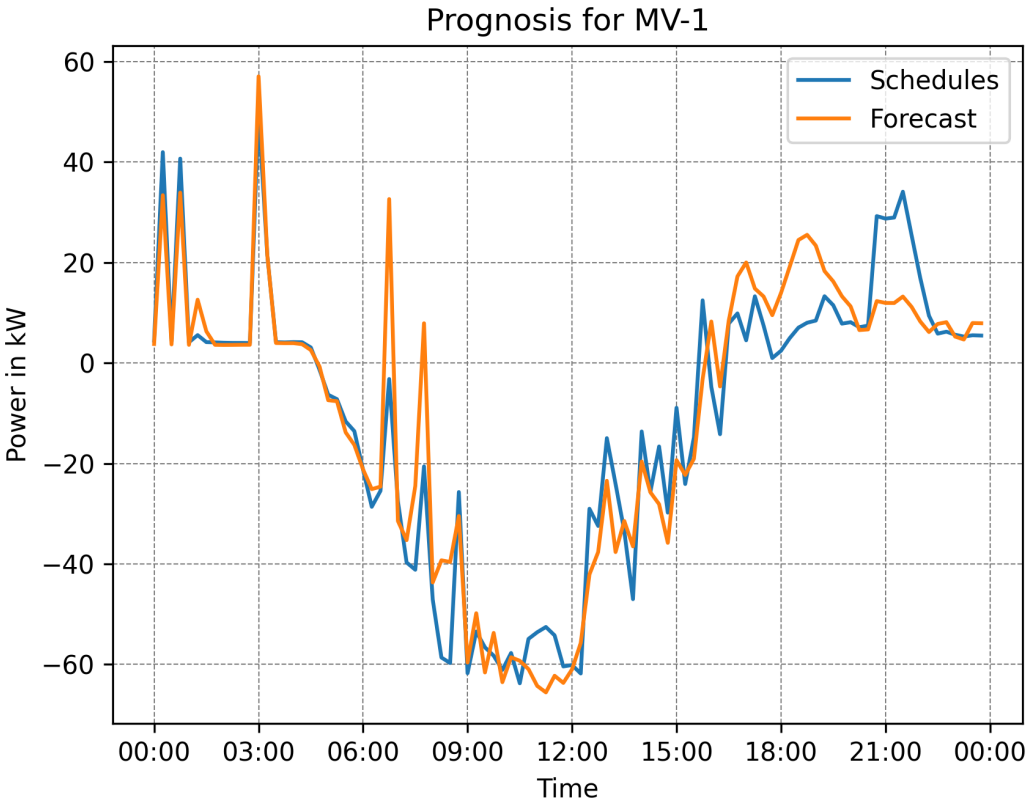


Abbildung B.1: Vergleich von prognostiziertem und geplantem Lastverlauf für den Transformator MV-1 (Clusterprognose, Sommer)

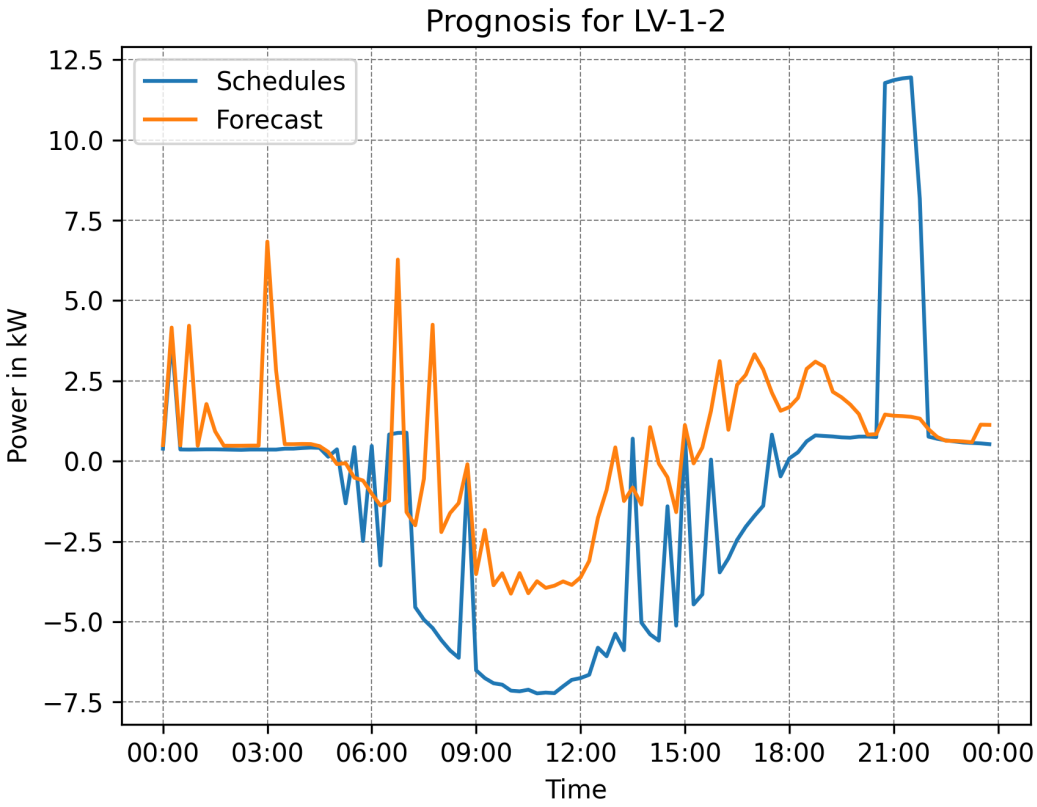


Abbildung B.2: Vergleich von prognostiziertem und geplantem Lastverlauf für den Niederspannungsabgang LV-1-2 (Clusterprognose, Sommer)

Eidesstattliche Erklärung

Hiermit erkläre ich, Joost Henning Lindner, dass ich die vorliegende Masterarbeit selbstständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe. Die aus fremden Quellen direkt oder indirekt übernommenen Stellen sind als solche kenntlich gemacht. Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungsbehörde vorgelegt und auch nicht veröffentlicht.

Kiel, 10. November 2025

Joost Henning Lindner