TÉCNICO LISBOA

Clustering Applications in Electric Mobility

Marcelo José da Silva Braço Forte

Thesis to obtain the Master of Science Degree in

Electrical and Computer Engineering

Supervisors: Prof. Hugo Gabriel Valente Morais
Dr. Cindy Paola Guzman Lascano

Examination Committee

Chairperson: Prof. Célia Maria Santos Cardoso de Jesus
Supervisor: Prof. Hugo Gabriel Valente Morais
Member of the Committee: Dr. Alexios Lekidis

November 2023

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgments

I want to thank my parents and sister for their friendship, encouragement, and support throughout these years. My parents gave me the opportunity to have the best academic education, and I am extremely grateful to them. Thank you. I would also like to mention my grandparents, who helped raise me and make me the person I am today.

To my supervisors, Dr. Hugo Morais and Dr. Cindy P. Guzman, for their invaluable guidance and support throughout this research journey. Their experience and insightful feedback were fundamental in shaping this thesis and helped me feel motivated and always supported. A special thanks for their trust and guidance in writing scientific papers resulting from this work. Thank you very much.

I would like to acknowledge the EV4EU team, specifically the Greek partners, for providing access to an extraordinary dataset that allowed me to accomplish the objectives of this thesis.

Last but not least, thanks to my friends and colleagues who helped me grow and have been there for me through this five-year academic journey.

To each and every one of you - thank you.

Abstract

The continuous growth of electric vehicles (EVs) has been boosted by the need to achieve society's decarbonization targets. The mass adoption of EVs introduces new challenges in power systems planning and operation. Clustering has emerged as a powerful tool to help better understand and categorize the uncertain behavior of EV users and the electric vehicle supply equipment (EVSE) needs. However, previous studies lack empirical European EV charging data and relevance for practical applications. In this thesis, different clustering techniques were evaluated to identify typical groups of EV charging processes to support characterizing EV charging profiles, EV user behavior profiles, and EVSE accessibility. The defined methodology comprises three major stages: data preprocessing, clustering application, and validation of results. A use case considering both open and private EV charging data (Caltech University and the publicly operated EVSEs in Greece, respectively) has been utilized to test the proposed methods, closing the gap verified in the literature. The experimental results demonstrated that Caltech features highly flexible charging sessions with routine users, while Greece exhibits more frequent EV users and quick-stay sessions. Additionally, there is an excellent opportunity to expand the charging network in Greece at specific locations. This information unlocks the potential for future studies, enabling distribution system operators and charge point operators to intelligently and successfully integrate EVs into the energy system.

Keywords

Charging Flexibility; Clustering; Data Analysis; Electric Mobility; Typical profiles; User Behavior;

Resumo

O crescimento contínuo dos veículos elétricos (VEs) tem sido impulsionado pela necessidade de atingir os objetivos de descarbonização da sociedade. A adoção em massa de VEs introduz novos desafios no planeamento e operação de sistemas de energia. O clustering tornou-se uma ferramenta poderosa para ajudar a compreender e categorizar melhor o comportamento incerto dos utilizadores de VEs e as necessidades dos equipamentos de abastecimento de veículos elétricos (EAVEs). No entanto, os estudos anteriores carecem de dados empíricos de carregamento europeus e de relevância para aplicações práticas. Nesta tese, diferentes técnicas de clustering foram avaliadas para identificar grupos típicos de processos de carregamento e ajudar na caraterização de perfis de carregamento de VEs, de comportamento do utilizador de VEs e a acessibilidade de EAVEs. A metodologia definida compreende três etapas principais: pré-processamento dos dados, aplicação de clustering e validação dos resultados. Para testar os métodos propostos, foram utilizados dados de carregamento de livre acesso (Universidade Caltech) e também privados (EAVEs públicos na Grécia), colmatando a lacuna identificada na literatura. Os resultados experimentais demonstraram que a Caltech apresenta sessões de carregamento altamente flexíveis com utilizadores regulares, enquanto a Grécia apresenta utilizadores mais frequentes com sessões de estadia rápida. Além disso, existe uma excelente oportunidade para expandir a rede de carregamento na Grécia em locais específicos. Esta informação revela o potencial para estudos futuros, permitindo que os operadores de sistemas de distribuição e os operadores de pontos de carregamento integrem de forma inteligente e com sucesso os VEs no sistema de energia.

Palavras Chave

Análise de Dados; Clustering; Comportamento do Utilizador; Flexibilidade de carga; Mobilidade Eléctrica; Perfis Típicos;

Contents

List of Figures

List of Tables

List of Algorithms

Acronyms

ACN	Adaptive Charging Network
AC	Alternating Current
BEV	Battery Electric Vehicle
CPO	Charge Point Operator
CP	Charging Pool
CS	Charging Station
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DC	Direct Current
DSO	Distribution System Operator
EV	Electric Vehicle
EVSE	Electric Vehicle Supply Equipment
EU	European Union
EM	Expectation-Maximization
GMM	Gaussian Mixture Model
GHG	Greenhouse Gas
ICEV	Internal Combustion Engine Vehicle
IQR	Interquartile Range
JPL	NASA's Jet Propulsion Laboratory
KDE	Kernel Density Estimation
ML	Machine Learning
OPTICS	Ordering Points To Identify Cluster Structure
PHEV	Plug-in Hybrid Electric Vehicle

RES	Renewable Energy Sources
SoC	State of Charge
STEPS	Stated Policies Scenario
SSE	Sum of Squared Errors
UK	United Kingdom
UN	United Nations
USA	United States of America
WCSS	Within-Cluster Sum-of-Squares

1

Introduction

Contents

1.1 Motivation

The world is an ever-changing place, but in between this revolution, one thing seems well-defined: fight climate change. People are facing a dramatic transformation in their lifestyle to become carbon neutral, with the United Nations (UN) placing the fight against climate change under one of the goals of Sustainable Development [1]. At the 2015 UN Climate Change Conference (COP 21), 196 countries reached the first-ever universal and legally binding climate change agreement that sets out a worldwide action plan to "limit global warming to well below 2°C, preferably to 1.5°C, compared to pre-industrial levels" [2]. This ambitious plan requires a significant reduction in Greenhouse Gas (GHG) emissions.

Transport is the only sector where GHG emissions have increased in the past three decades in Europe [3]. This sector was responsible for more than a quarter of Europe's total GHG emissions in 2019, of which approximately 71% came from road transportation, increasing 33% between 1990 and 2019, according to a 2022 report by the European Environment Agency [4]. In addition to GHG, burning fossil fuels, whether in power plants or in Internal Combustion Engine Vehicles (ICEVs), releases harmful pollutants that can significantly degrade air quality.

In this context, the adoption of Electric Vehicles (EVs) is increasing rapidly in the 21st century due to the urgency for global energy demand to shift away from fossil fuels, particularly in the last decade. Even though the production and disposal of EVs are currently less eco-friendly than those of an ICEV (mainly due to the production of its batteries [5]), an analysis of the entire life cycle of an EV shows that it is still cleaner than an ICEV, as revealed by Zhang et al. [6]. Their study demonstrated that EVs could potentially provide a 45% reduction in GHG emissions compared to ICEVs, considering the energy cost of production, assembly, transportation, and usage (the authors assumed 300 000 km as the average lifetime of a passenger vehicle).

As the share of electricity from Renewable Energy Sources (RES) is set to increase in the future, as well as making batteries more sustainable, EVs should become even less harmful to the environment [7]. According to the World Energy Transitions Outlook 2023 [8], the share of RES in electricity generation should increase from 28% in 2020 to 91% in 2050. With that in mind, car manufacturers and governments have been investing in new models and tax incentives for the purchase and adoption of EVs, whose popularity has significantly increased over the past five years [9].

The European Union (EU) aims to be carbon-neutral by 2050. This objective is the heart of the European Green Deal and in line with the EU's commitment to global climate action under the Paris Agreement [10]. To achieve this goal, in 2022 the EU's environment ministers approved a new "Fit for 55 in 2030" package [11], which orders that only zero-emission vehicles can be sold in Europe from 2035 [12]. The United States of America (USA) and the United Kingdom (UK) are also targeting net-zero emissions by 2050, China and Russia by 2060, and India by 2070 [13], together with the EU, the biggest polluters in the world.

Due to all these factors, the number of EVs will certainly increase in the upcoming years. However, the EV rise poses several challenges in the power systems, mainly at the distribution level [14]. Uncontrolled EV charging has negative impacts on the existing power grid, including high load peaks, voltage instabilities, higher energy use, and degradation of power quality, among others [15]. Therefore, reliable control and understanding of the EV charging process will be essential.

Utilities, Distribution System Operators (DSOs), and Charge Point Operators (CPOs) need to quantify the impacts on grid infrastructure and network reinforcement demands to address future challenges and opportunities associated with EV mass adoption. EV batteries also represent a flexibility potential that may become increasingly valuable to the energy system as RES increase in prevalence. The identification of typical profiles is of great relevance for these entities to perform a successful and intelligent integration of EVs in the energy system.

1.2 Objectives

The main aim of this work is to investigate the possibility of identifying different groups of EV charging processes, through clustering, to provide support in characterizing EV charging profiles, EV user behavior profiles, and Electric Vehicle Supply Equipment (EVSE) accessibility, based on comprehensive datasets of empirical charging processes. A detailed insight into the complexity of EV charging behavior has enormous significance for the future sizing of distribution grids and charging infrastructures. It can also be helpful in future studies, particularly in the coordination of EVs with solar and wind renewable energies. Beyond this goal, this thesis aims to answer the following research questions:

- Can EVSEs and EV users be classified according to their charging behavior?
- What are the most suitable clustering methods to classify the EV charging and EV user behavior?
- Which applications can benefit from the information extracted through clustering methods?

1.3 Related Projects and Scientific Outputs

The work carried out as part of this thesis was developed under the scope of the following research project:

- **Horizon Europe EV4EU** – Electric Vehicles Management for carbon neutrality in Europe, funded by the European Union under grant agreement no. 101056765. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the grating authority can be held responsible for them.

The developed work resulted in the following outputs:

Scientific articles:

- M. Forte, C. P. Guzman, and H. Morais, “A Comprehensive Review of Clustering Methods Applications in Electric Mobility”, submitted to *Renewable and Sustainable Energy Reviews*, under review since Jul. 2023;
- M. Forte, C. P. Guzman, A. Lekidis, P. M. S. Carvalho, and H. Morais, “Clustering Methodologies for Electric Vehicles Supply Equipment Flexibility Characterization”, submitted to *Electric Power Systems Research*, Special Issue *23rd Power Systems Computation Conference (PSCC 2024)*, under review since Oct. 2023.

Others:

- Participation in the *Workshop* “The Consumer’s Role in the Energy Transition” organized by *Universidade Federal de Juiz de Fora, Brasil*, on July 21st, 2023, with the presentation “Clustering Methodologies for the Identification of EV Patterns”;
- Participation in EV cluster analysis, writing of content, and scientific review of the *Horizon Europe Project EV4EU* Deliverable D3.3 - EVs use Clustering results report (*under internal review*).

1.4 Organization of the Document

This document is structured as follows. The present chapter introduces the main motivation and objectives of this thesis. Chapter 2 addresses the background history and current state of EVs with some insights into the charging process. Chapter 3 starts by exploring the definition of clustering, followed by a state-of-the-art review of the clustering applications in EV-related data. Additionally, the main objectives of the thesis are outlined along with their differences. Chapter 4 presents the definition and proposal of methodologies to achieve the intended objectives, along with a complete description of the chosen datasets, clustering methods, and evaluation metrics. Chapter 5 details the obtained results for each objective and summarizes the main findings of the work. It also proposes practical applications. Finally, Chapter 6 contains the conclusions and possible future work.

2

Background

Contents

2.1 History & Current State of EVs

It is hard to pinpoint the invention of the electric car to one inventor or country. Instead, it was a series of breakthroughs (from the battery to the electric motor) in the 1800s that led to the first EV on the road. Around 1832, Robert Anderson developed the first crude EV, but it wasn't until 1881 that Gustave Trouve, a French electrical engineer, reportedly created the first battery-powered EV, a tricycle. It weighed 160 kg and was powered by lead-acid batteries [16]. The taxi "Electroboat" was the first EV in the USA, introduced by William Morris in 1889. It had a top speed of 32 km/h and 40 km range, which was a significant advance over earlier models. In 1900, 22% of the 4200 vehicles sold in the USA were ICEVs, while 38% were EVs [17]. But EVs began to lose some relevance and their use declined as Henry Ford decided to mass-produce the Model T, which made gasoline-powered cars broadly accessible and reasonably priced in 1908, dominating the market.

However, the 1973 Arab oil crisis encouraged the development of alternative energy sources, and the interest in EVs returned [18]. Even NASA contributed to increasing awareness of EVs when, in 1971, its electric Lunar rover became the first piloted vehicle to drive on the moon. The invention of more powerful and durable motors, Direct Current (DC) to Alternating Current (AC) inverters, and effective battery management systems, together with the breakthrough in microprocessors in the 1980s and 1990s, all helped to revive interest in EVs. Since then, automakers have been developing prototypes in response to new transportation emissions restrictions. First-generation EVs from this modern era include Toyota Prius (1997 Plug-in Hybrid Electric Vehicle (PHEV)), Tesla Roadster (2008 Battery Electric Vehicle (BEV)), and Nissan Leaf (2011 BEV). Moreover, with the Paris Agreement [2], reducing GHG emissions has been a priority, and EVs became part of the solution.

Following the nomenclature of the International Energy Agency [19], in this thesis, the term "EV" includes BEVs and PHEVs unless otherwise specified. Recent reports reveal that the EV stock has increased exponentially in recent years, as shown by Figure 2.1, especially BEVs.

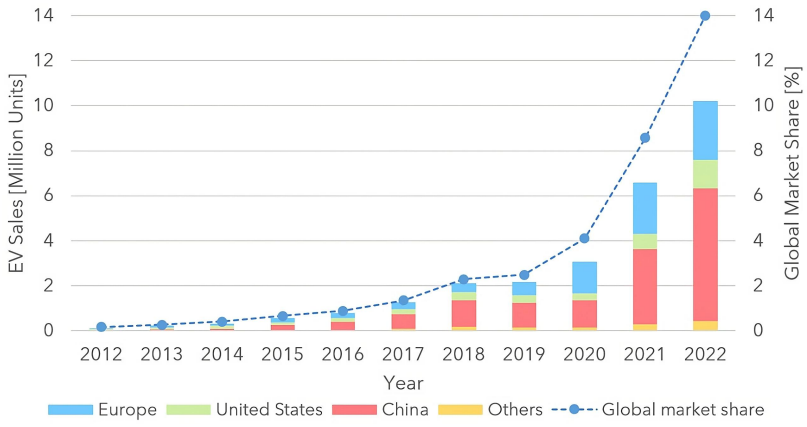


Figure 2.1: Global sales and market share of EVs, 2012-2022 (Data from [19]).

In 2022, global EV sales achieved 10.2 million units (up 55% relative to 2021), representing around 14% of the market share, and the EV car stock represented 2.1% of the global fleet [19]. More than USD 425 billion were spent on EVs globally in 2022, a rise of around 50% from 2021, with a strong focus on SUVs (or Sports Utility Vehicles) and big cars, similar to the pattern seen in ICEV markets.

EV sales in Europe follow the worldwide trend. In 2022, 2.7 million EVs were sold, an increase of more than 15% compared with 2021. The European market for EVs represented 2.4% (7.8 million) of the global automobile stock [19]. Despite the slower growth in 2022 (63.5% growth in 2021 compared with 2020), EV sales are still increasing in the context of continued contraction in the European car market. In Stated Policies Scenario (STEPS) [20], it is foreseen a high growth in EV sales, possibly reaching 36% sales share and 15% stock market share in 2030 [19].

Although there are different technologies to power electric motors, battery packs are the primary power source for this new EVs [17]. Nowadays, BEVs can travel up to 700 km on a single charge, unlike early models that would often last less than 100 km due to battery constraints. The capacity of the batteries, the overall efficiency of the EV, and the management strategies directly impact the range [21]. Figure 2.2 presents the Top-5 best-selling EVs in 2022 with their respective maximum ranges.

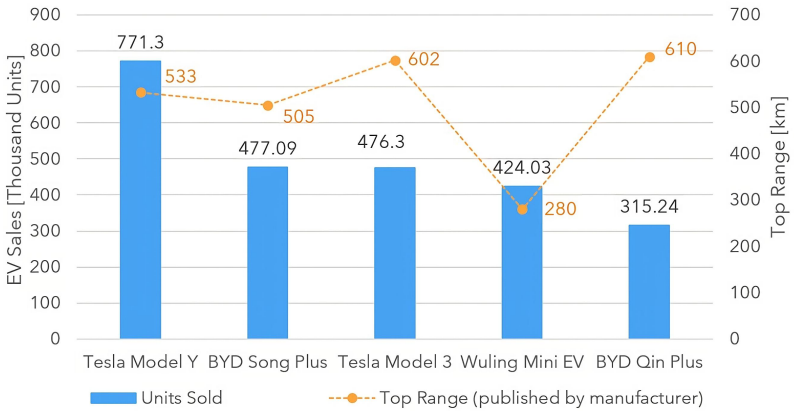


Figure 2.2: Range of best-selling EVs worldwide in 2022 (Data from [21] and [22]).

The battery pack of these BEVs needs to be recharged. Thus, there is an urgent demand to build Charging Stations (CSs) to meet the needs of drivers. According to [23], the number of public EVSEs reached more than 2.7 million in 2022, of which around one-third were fast chargers (Figure 2.3).

The EU had 13 EVs per EVSE in 2022, above the 2014 Alternative Fuel Infrastructure Directive [24] suggestion of 10 EVs/EVSE by 2020. The “Fit for 55 in 2030” package [11] includes a proposal to repeal this directive and turn it into a regulation, the Alternative Fuels Infrastructure Regulation. Article 3 [25] mandates 1 kW of publicly available charger per BEV and 0.66 kW per PHEV by 2030.

In addition to light-duty personal vehicles, electrification is reaching other categories, including two or three-wheelers, commercial and heavy-duty vehicles [26]. In fact, 27 nations (including the USA and EU) have committed to achieving 100% sales of zero-emission buses and trucks by 2040 [27].

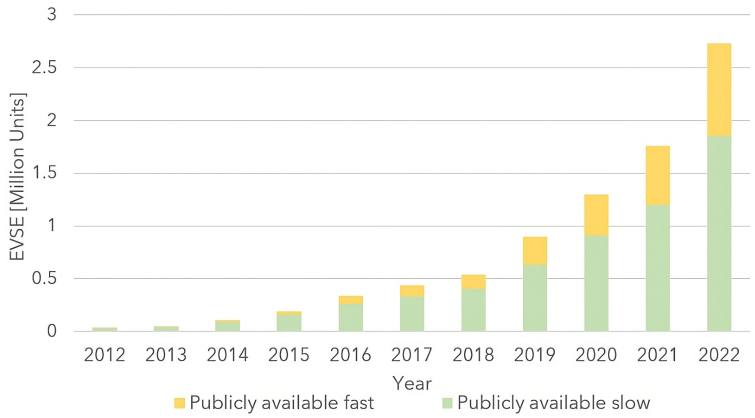


Figure 2.3: Evolution of public EVSEs worldwide, 2012-2022 (Data from [23]).

2.2 EV Charging Process

In the context of EV charging, the terms EVSE and CS are often used interchangeably. This thesis follows the terminology outlined by the EU - Sustainable Transport Forum [28]. While commonly referred to as a "charger" or a "charging point", an **EVSE** is technically the equipment that provides electricity to an EV. On the other hand, a **CS** is a physical object that includes one or more EVSEs sharing a common user identification interface (similar to a gas pump with multiple refueling hoses for ICEVs). These definitions collide with the general idea of an *EV charging station* as the equivalent of an *ICEV gas station*. However, according to [28], a site with one or more CSs and the associated parking lots is known as a **Charging Pool (CP)** and is operated and managed by a CPO.

Regarding the charging process, conductive EV charging can be divided into three categories: **Level 1** (slow process, uses a regular 120-volt wall plug, found in all houses and garages), **Level 2** (requires a dedicated 240-volt charger, but it's 15 times faster than Level 1), and **Level 3** (mostly known as DC fast charging, uses 480+ volts, found in public places). An EV receives AC from Level 1 and Level 2 chargers, which is then converted to DC internally by the EV (slow process). EV batteries only support DC power. In contrast, no conversion is necessary when using a DC fast EVSE. Level 1 and Level 2 chargers use Type 1 connectors typically in America (SAE J1772), and for European and Asian vehicles, Type 2 connectors are standard. Dimitriadou et al. [29] present an overview of the current status of the infrastructure utilized for the realization of both conductive and wireless charging of an EV battery, presenting a detailed exposition of the respective standards and charging levels, as well as future challenges and opportunities.

3

Related Work

Contents

The present chapter starts by exploring the definition of clustering, followed by a literature review of clustering applications in EV-related data. Additionally, the main objectives of the thesis are outlined, as well as their differences and characteristics.

3.1 Clustering Methods

In the literature, cluster analysis has received a lot of attention and has been researched extensively. There are papers such as [30], published in 1969, that helped to investigate and develop various mathematical and classification techniques. Nevertheless, it is important to first give a brief introduction.

Cluster analysis, often known as **clustering**, is not a specific algorithm, but rather the general problem of partitioning a dataset into natural subgroups called **clusters** [31]. Objects within the same group should be as similar as possible (based on a similarity measure), while objects between different groups should be as dissimilar as possible. Clustering uses almost no information to evaluate the data and does not require a separate training dataset to determine the model parameters (unsupervised learning approach). It is the main objective of exploratory data analysis, a popular statistical analysis technique that is applied in a variety of domains, including pattern recognition, image analysis, bioinformatics, and Machine Learning (ML) [32]. Figure 3.1 provides a simple illustration of clustering.

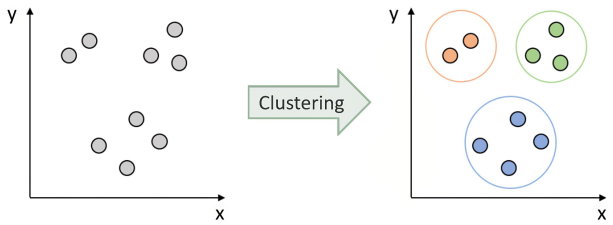


Figure 3.1: Simple illustration of clustering in two dimensions (Adapted from [33]).

Since there is no clear definition of the term “cluster”, numerous clustering methods for distinct strategies have been developed, further discussed in the following sections. In this document, the notation and nomenclature follow the ones defined by Zaki and Meira [31], presented in Figure 3.2.

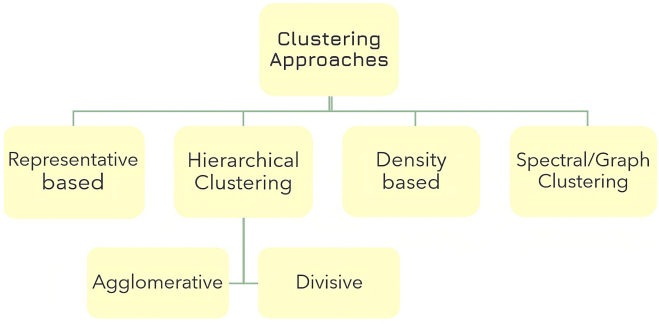


Figure 3.2: A summary of clustering methods (Based on [31]).

3.1.1 Representative-based Clustering

Representative-based clustering aims to divide a dataset into k clusters. Each cluster is characterized by a representative point (called **centroid**), commonly chosen as the mean of within-cluster points. The K-means and Expectation-Maximization (EM) algorithms are examples of representative-based clustering approaches:

- K-means [34] is a greedy technique that minimizes the squared distance between points and their corresponding cluster means. It also conducts hard clustering, meaning that each point is assigned to only one cluster;
- EM [35] generalizes K-means by modeling the data as a mixture of normal distributions (Gaussian Mixture Model (GMM)) and maximizing the likelihood of the data to find the cluster parameters (the mean and covariance matrix). It is a soft clustering approach since it returns the probability of a point belonging to each cluster. EM is the algorithm utilized by the GMM clustering method.

3.1.2 Hierarchical Clustering

Hierarchical clustering techniques create a sequence of nested partitions, which can be visualized as a tree, also called *dendrogram*, indicating the merging process and the intermediate clusters. The highest level (root) of the tree consists of all points in one single cluster, whereas the lowest level (leaves) consists of clusters of individual points, each point in its cluster. If the desired number of clusters is known, one can graphically see the level at which k clusters exist. There are two algorithmic approaches to get Hierarchical clusters [36]:

- **Agglomerative**: Start with the points as individual clusters and, at each step, merge (or agglomerate) the most similar or closest pair of clusters until the desired number of clusters has been found. This requires a definition of cluster similarity or distance;
- **Divisive**: Start with one cluster (all points), and at each step, divide a cluster until only clusters of individual points remain. In this case, it is required to decide, at each stage, which cluster to split and how to perform it. It works just the opposite of the Agglomerative approach.

3.1.3 Density-based Clustering

Density-based clustering methods use the density or connectedness properties to find nonconvex clusters. This type of clustering employs the local density of points to determine the clusters rather than using only the distance between points, such as in K-means or EM. The most popular methods are Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [37] and Ordering Points To Identify Cluster Structure (OPTICS) [38].

3.1.4 Spectral/Graph Clustering

The goal of **Graph clustering** is to cluster the nodes by using the edges and their weights, which represent the similarity between the incident nodes. Graph clustering can be viewed as an optimization problem over a k-way cut in a graph, with different objectives represented as spectral decompositions of various graph matrices, derived from the original graph data or the kernel matrix, such as the adjacency matrix and Laplacian matrix [39]. The graph can then be split into connected components using a specific graph cut method, and those components are referred to as clusters.

3.2 Applications of Clustering in EV Data

EV charging data has been submitted to clustering to identify the most common and recurrent profiles [40–42]. The literature often considers EV charging and EV user behavior profiles synonyms. Authors name their work depending on the dataset and the chosen fields. The same does not happen for EVSE accessibility, whose studies utilize EVSE location data and not EV charging data.

For example, Shen et al. [43] grouped the charging sessions by each user and then performed clustering, naming his work *EV user charging behavior identification*. Shahriar and Al-Ali [44] also utilized the same dataset but clustered the features without grouping the data by user, naming it *charging behavior clusters*. Ultimately, the two studies found groups with similar charging behavior characteristics. Thus, this section presents the most relevant work done in each of these areas, divided into the subsections EV Charging & User Behavior Profiles, EVSE Accessibility, and Summary of Literature Review.

3.2.1 EV Charging & User Behavior Profiles

Working with a large dataset from metropolitan areas of the Netherlands, Helmus et al. [45] carried out a two-step, bottom-up data clustering approach that first employs GMM to cluster charging sessions and then portfolios of charging sessions per user using K-Medoids (comparable to K-means clustering). The study considers starting time, connection duration, the distance between two sessions, and hours between sessions as features. From the first step, thirteen clusters were found: 7 types of daytime and 6 types of nighttime charging sessions. The second step resulted in nine distinct clusters: 3 clusters contained daytime users, 3 nighttime, and the other 3 featured unusual users. The study is well-detailed, yet perhaps too complex. It requires careful reading and prior knowledge of some of the methods.

On the other hand, Märtz et al. [46] claim they used the most extensive (private) dataset on charging patterns from an EV perspective known in the literature, containing approximately 21 000 BMW i3 BEVs and about 2.6 million charging processes during one year (2019). The authors performed GMM clustering on the EV charging behavior, utilizing plug-in time and duration as features, and discovered seven distinct clusters: 3 overnight and 4 daytime. The authors conducted a second analysis with K-means

clustering to identify switching EV users between clusters. They also made known the flexibility potential of the EV charging processes, concluding that there was a huge potential: on average, the temporal flexibility was 8 hours. The methods are well described, and the decisions are thoroughly justified, leading to outstanding illustration and understanding of the characteristics of the clusters found, turning this analysis into one of the most complete in the literature.

Shahriar and Al-Ali [44] conducted one of the most interesting analyses found. They performed cluster analysis with K-means, Hierarchical clustering, and GMM to identify similar groups of charging behavior, based on arrival and departure times, on real public EV charging activity during the COVID-19 pandemic. K-means produced the best results, followed by Hierarchical clustering. The authors only discovered three clusters corresponding to the knee of the elbow method curve. The study's drawbacks include only employing a single method for establishing the appropriate number of clusters and selecting only two features to group the data (arrival and departure times), resulting in generic results.

The K-means technique was employed by Shen et al. [43] to identify charging behavior clusters. The authors supervised the clustering results and adjusted them to achieve the best possible outcomes, something fundamental when data is sparse and/or irregular. To obtain typical user behavior, the data was grouped by user, leading to the average charging time, the standard deviation of charging time, and the standard deviation of connection time as the basis for the clustering. Three clusters were discovered. Two groups were identified as stable and predictable users, but the third cluster comprised unexpected users. Similarly, Xiong et al. [47] attempted to find EV user behavior by organizing the data by the user. Thus, the tuple (average arrival time, average departure time, standard deviation of arrival time, standard deviation of departure time) represented each user. In addition to these, the authors also incorporated another field, the so-called Pearson correlation coefficient, between stay duration and energy consumption. With this data, they performed clustering with K-means, obtaining four profiles.

Van Kriekinge et al. [48] proposed a methodology to simulate the charging demand for different types of drivers. Typical EV driver profiles with similar charging habits are needed to accomplish this goal. All charging sessions from a private dataset were replaced by one specific theoretical charging session per EV driver (average value of the plug-in times, parking times, and charged energy) with the goal of obtaining user behavior profiles. The result is a mean behavior for each driver. The clustering proposed in this study works in two stages: cluster the average characteristics per EV user and then analyze the frequency of charging, always with the K-means algorithm. The results indicated five clusters, with big differences in behavior between the EV drivers. In addition, the Kernel Density Estimation (KDE) process allows capturing the details of each cluster, helping in the final simulation stage, which demonstrated a strong impact on power and energy demand when adding new EV users to the population.

Gerossier et al. [49] employed Hierarchical clustering to identify four groups of EV charging behavior. The authors received data in time-series format, which they processed to extract individual sessions

categorized by start-up time (initial plug-in time) and duration of the charging process, following a method well described and presented in the study. Most customers belonged to the first group, where charging was typically performed during the evening and morning.

3.2.2 EVSE Accessibility

Given the previously mentioned studies, it may seem that clustering is only applied to EV charging data. However, the focus of the literature goes beyond charging patterns. As EV sales increase, the location and accessibility of EVSEs may become an issue, which clustering can help to address.

Carlton and Sultana [50] performed spatial clustering of public EVSEs to analyze the characteristics of their land use, and how these characteristics impact EVSE accessibility. The authors applied DBSCAN to identify spatially clustered Level 1, Level 2, and DC fast-charging infrastructures in the Chicago Metropolitan Area. The results indicated that access to EVSEs is unequal between suburban and urban neighborhoods, bringing social inequalities into view and preventing the widespread of EVs. Similarly, Borlaug et al. [51] conducted an extensive study on the accessibility and utilization of 3 705 public Level 2 and Level 3 EVSEs in the USA over 2.5 years (2019-2022), observing usage patterns over time. They also performed a regression analysis to evaluate the correlation between CS utilization and various contextual and environmental factors. The study concluded that the presence of DC fast chargers resulted in decreased utilization of Level 2 EVSEs. Furthermore, as of March 2022, EVSE utilization was still below pre-pandemic levels.

Finding the most appropriate location to place EVSEs is a big problem. Most research papers concentrate on building placement models and positioning CSs based on objective functions and restrictions [52]. To prove that clustering can produce more accurate and understandable results, Li et al. [53] proposed a technique for finding EVSE locations in Qingdao, China, in response to the expanding demand for electric taxis. Electric taxi information and charging needs are derived from the extensive GPS trajectories of gasoline taxis. To find the ideal site for the CSs, the Qingdao area is subjected to multiple same-type clustering and multi-type clustering methods. The K-means and Agglomerative Hierarchical clustering revealed that the positioning results of the multi-type clustering are more credible.

Kalakanti and Rao [54] addressed two main problems related to EVSE: the EVSE location and the EVSE need estimation. This work investigated different explainable solutions based on ML and simulation. For the problem of EVSE location, the authors utilized a geolocation dataset of EV households to perform a comprehensive analysis with different classes of clustering methods, namely K-means, GMM, OPTICS (similar to DBSCAN), and Spectral clustering. The results were compared with the existing EVSE location in Austin, USA (to show the improvement over the existing setup), and with a greenfield area like Bengaluru, India, where synthetic EVSE data were used. Silhouette coefficient, Calinski-Harabasz, and Davies-Bouldin index were the chosen metrics to evaluate the clustering results.

3.2.3 Summary of Literature Review

As previously mentioned, there are some recent studies conducted with the specific intent of obtaining typical profiles for EV data. However, most of these studies lack practical relevance to help DSOs and CPOs with grid management, and due to the uncertainty of the charging data and the employed methods, it is difficult to achieve a generalized result across all the analyzed studies. It is also important to note that most of these (few) studies utilized datasets from countries outside of Europe. Thus, there is an opportunity to conduct studies that close this gap. Table 3.1 summarizes the analyzed studies.

Table 3.1: Summary of the most relevant EV and EVSE clustering papers reviewed.

Study	Brief summary	Clustering method	Dataset	Conclusions
Helmus et al. [45] Amsterdam, Netherlands 2020	Provides a realistic analysis of charging behavior and EV user types based on clustering, differing from the typical literature that is frequently oversimplified	GMM for clustering and Partition Around Medoids to find portfolios of charging sessions per user	5.82 million charging transactions (January 2017- March 2019) from the Dutch metropolitan area	13 clusters were found: 7 types of daytime charging sessions (4 short, 3 medium duration) and 6 types of overnight charging sessions
Märtz et al. [46] Germany 2022	Investigates the possibility of identifying different clusters of EV charging processes, validating the results against synthetic load profiles and the original data	GMM and K-means	2.6 million private charging processes of 21 000 BMW's i3 model from 2019 in Germany	High number of charging opportunities during day, as well as user exchange between charging clusters, to reduce localized energy demand. Found 7 clusters
Shahriar and Al-Ali [44] UAE 2022	Investigates the impacts of COVID-19 on EV charging behavior by analyzing the charging activity during the pandemic	K-means, Hierarchical clustering, and GMM	ACN dataset, from Caltech University Campus	Identified 3 groups of charging behavior. The best clustering was obtained using K-means followed by Hierarchical clustering
Shen et al. [43] USA & Canada 2020	To manage (dis)charging behavior of EVs in the smart grid, proposes a communication network for analysis and prediction of user behavior	HITL-based K-means clustering and K-NN algorithm for prediction	ACN dataset, from Caltech University Campus	Identified 2 clusters of stable, predictable users, but the third cluster was found to be unexpected users
Xiong et al. [47] Los Angeles, USA 2018	Proposes an EV user behavior technique, using unsupervised and deep learning techniques, applied to historical EV data to make the day-ahead parking and charging prediction	K-means for clustering, multilayer perceptron for classification	More than 4 years data of the UCLA SMERC smart charging network infrastructure	Identified 4 clusters, with 3 relatively predictive behavior, but one cluster represented random traveling schedule and energy consumption
Kriekinge et al. [48] Brussels, Belgium 2023	Proposes a methodology to simulate charging demand for different EV driver types. The identification of similar profiles is performed using clustering	K-means for clustering and KDE to better capture details for the simulation stage	8 755 private EV charging sessions (Jul 2018 - Jan 2022)	Identified 5 clusters, with distinct and different characteristics, showing good clustering results
Gerossier et al. [49] Texas, USA 2019	Models the consumption profile of EVs from raw power measurements. The charging habits model is then used for forecasting short-term (1 day ahead) and long-term (2030)	Hierarchical clustering with Ward's method	46 private EV charging data recorded every minute of the year 2015 in Texas	Identified 4 clusters. Simulating the projected demand in 2030, it appears that the growth in EVs will have little effect on the load curve's shape
Carlton et al. [50] North Carolina, USA 2022	Performs spatial clustering of public EVSE to analyze their associated land use tendency, and how these can impact EVSE accessibility	Hierarchical clustering based on DBSCAN	Public EVSE location data from the Alternative Fuel Data Center (AFDC)	Majority of level 2 EVSE, only 26% of clusters with mixed land uses (residential, commercial and recreational)
Kalakanti and Rao [54] India 2022	Aims to solve two problems: the EVSE placement and the EVSE need estimation, to guide the urban planners in making better decisions	K-means, GMM, OPTICS, Spectral clustering, and other ML methods	Austin Charging Station Network real geolocations	K-means and GMM consistently yielded the best results, with OPTICS and Spectral clustering often wrong or nonsense

3.3 EV Charging profiles vs EV User Behavior profiles vs EVSE Accessibility

The literature often considers EV charging and EV user behavior profiles synonyms, as previously discussed in Section 3.2. However, in this thesis, the two types of profiles are not synonymous as they represent and characterize different information.

An ***EV charging profile*** aims to characterize the times of day when more or fewer charging sessions occur, whether the sessions are high energy, low energy, long or short-term, with high or low flexibility potential. On the other hand, an ***EV user behavior profile*** intends to give an understanding of whether the user's behavior is recurrent, routine, or, on the other hand, random and without a typical charging frequency. For these studies, it is necessary to use a dataset of charging information about the EVSEs and the users. Regarding ***EVSE accessibility***, the required information is the location of the EVSEs since the aim is to understand the geographic distribution of the corresponding CPs, whether the current supply is in line with the demand, and whether there are inequalities in access to EVSEs that prevent the widespread use of EVs.

4

Methodology

Contents

This chapter presents the defined methodology and the proposed solution. The main characteristics of the chosen datasets are presented, followed by a description of each of the defined stages necessary to achieve the objectives of this thesis.

4.1 Solution Proposal

Few studies investigate EV real-world charging sessions due to the lack of this open data, resulting in typical EV profiles with high uncertainty. In this thesis, only real-world data will be used to avoid these uncertainties when performing clustering. Additionally, the goal is to find EV charging profiles, EV user behavior profiles, and the overall EVSE accessibility (Section 3.3). The overview of the methodological approach for the proposed solution is illustrated in Figure 4.1.

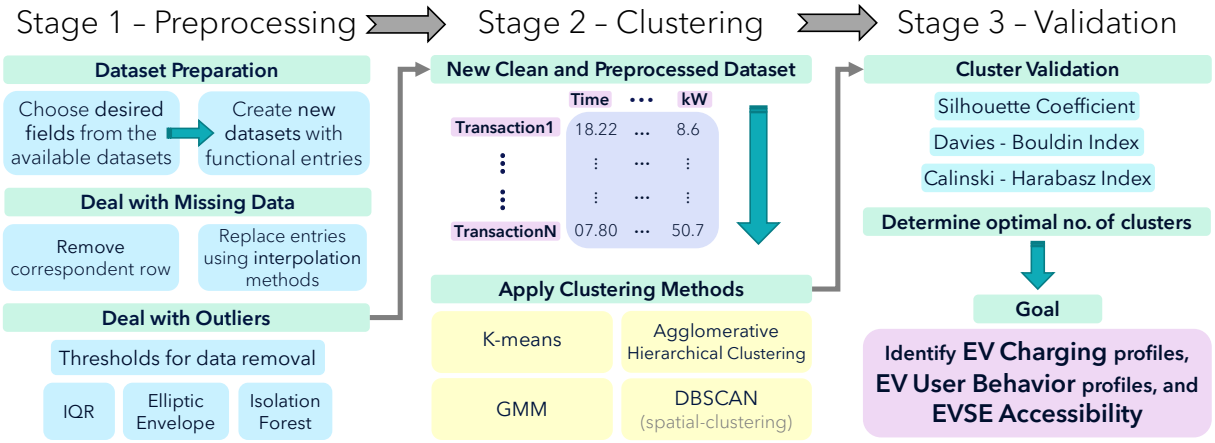


Figure 4.1: Overview of methodological approach.

Starting with the data, a thorough description of the selected datasets is provided in Section 4.2. The data preprocessing and cleaning approaches are explained in Section 4.3 and Section 4.4 describes, in detail, the chosen clustering algorithms to perform EV data analysis. Finally, the selected cluster validation techniques are described in Section 4.5.

4.2 Data Description and Analysis

There is no cluster analysis without a dataset. Therefore, it is essential to have an adequate EV charging dataset. Amara-Ouali et al. [55] perform an outstanding study of the best EV open data available, providing the community with a structured and carefully selected list of open datasets ready to be used to foster data-driven research in this field. Furthermore, Calearo et al. [56] present a review of data sources for EVs, categorized into different classes by the type of data and its availability.

Based on these papers, an open dataset was found and will be utilized: **ACN-Data**, to identify EV charging profiles and EV user behavior profiles. In addition to open data, this thesis had access to private datasets from several European partners in the context of the **EV4EU** project [57], from countries like Denmark, Greece, Slovenia, and Portugal. Thus, by performing deep analyses of this data, this thesis contributes to filling the lack of European studies identified in Section 3.2. The private dataset of public EVSEs in Greece (**GR-Data**) is one of the most complete and was therefore chosen to be studied to find EV charging profiles, EV user behavior profiles, and the EVSEs' accessibility. Both datasets are in the *charging event* format (1 row of the dataset, 1 EVSE transaction). Table 4.1 presents a summary of the characteristics of the chosen datasets.

Table 4.1: Summary of the main characteristics and available fields in the chosen datasets.

Datasets	ACN-Data	GR-Data
File Format	*Charging Event* JSON file	*Charging Event* CSV file
Time Interval	25 Apr 2018 - 14 Sep 2021	01 Jul 2021 - 05 May 2022
Total Sessions	31 424	22 412
Number of different EVSEs	55	312
EVSE ID and Location	Only Identification	Both
Plug-in and Plug-out Times	Yes	Yes
Start and End Charging Times	Yes	Only Start Charging Time
Charging Duration	No	No
Energy Consumed	Yes	Yes
EVSEs' Max Power	No	Yes
Customer ID	Yes	Yes

4.2.1 ACN-Data Dataset

Zachary J. Lee, Tongxin Li, and Steven H. Low are responsible for the public release of the **ACN-Data** dataset [58]. In [59], they describe the characteristics of the dataset, how they manage to get the data and prove that it has several possible applications, including clustering of EV charging data using GMM.

ACN-Data was collected from two Adaptive Charging Networks (ACNs) located in California, USA, namely in Caltech and in NASA's Jet Propulsion Laboratory (JPL). The ACN on the Caltech campus is located in a parking garage, containing 54 EVSEs Level 2 with rated 6.656 kW and one 50 kW DC fast charger. The JPL campus is closed to visitors, and only employees are permitted to use the EVSEs, unlike the Caltech ACN, which is accessible to everyone and frequently used by drivers not affiliated with Caltech. Caltech is a cross between workplace and public use charging, whereas the JPL site is an example of workplace charging. Thus, only the Caltech ACN dataset is going to be used in this thesis, as it presents more comprehensive data that does not solely focus on workplace behavior. At the time of writing, ACN-Data has 31 424 EV charging sessions. The first session was Apr 2018 and the last was Sep 2021. The monthly evolution of charging sessions is presented in Figure 4.2.

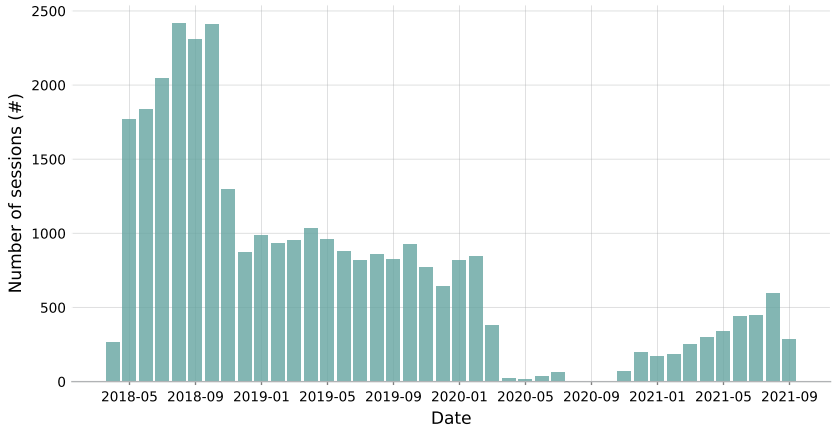


Figure 4.2: Charging activity per month in the ACN-Data dataset.

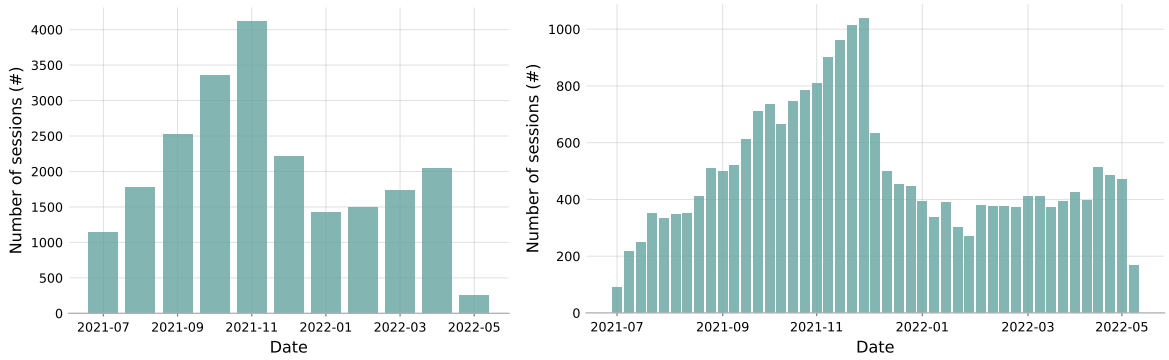
High charging activity can be observed at the beginning of the data recording period, with a peak of sessions from August to October 2018, with 2350 sessions per month on average during these months. In fact, until September 2018, there was a regular increase in EV charging sessions, which gradually decreased during the rest of the year. This behavior may indicate a higher number of users who attend this CP because it was a new and innovative space open to the general public. In 2019, there was a stabilization in the number of sessions, with an average of around 900 per month. This value remained constant in January and February of 2020, after which there was a significant drop in charging activity due to COVID-19, multiple electricity blackouts, and forest fires during this time [44]. EV charging activities resumed in November 2020, gradually growing from there. The number of sessions was already at pre-COVID levels in the last week of accessible data, in September 2021.

4.2.2 GR-Data Dataset

The private **GR-Data** dataset is courtesy of one of the Greek EV4EU project partners. This dataset was chosen since it is one of the most complete EV4EU datasets available in the *charging event* format, allowing for equal comparison with the ACN-Data dataset. It has a total of 22 412 charging sessions.

The GR-Data dataset was collected from public EVSEs in Greece, mainly located in high-traffic and quick-stay areas such as highways, gas stations, supermarkets, and stores. There are a total of 312 EVSEs with registered sessions in the dataset, of which only eight are Level 3: six EVSEs with 50 kW, one with 60 kW, and one with 120 kW. All the remaining chargers have 22 kW maximum power.

Figure 4.3 illustrates the monthly and weekly evolution of session numbers, from which one sees an increase until the end of November 2021, reaching a maximum around this time. However, there was a sharp decline in December 2021 and January 2022, coinciding with Greece's highest peak of COVID-19 cases due to the Omicron variant. The session numbers remain consistent throughout the remainder of 2022, with an increasing trend.



(a) Monthly activity.

(b) Weekly activity.

Figure 4.3: Charging activity in the GR-Data dataset.

4.3 Stage 1: Data Preprocessing and Cleaning

According to earlier research [44, 46, 60], data cleaning and preprocessing are two key processes in obtaining interpretable results from cluster analysis, including handling missing data and outliers, creating new fields, and normalizing the data before applying any method.

4.3.1 Deal with Outliers and Missing Data

Some datasets' entries might have **missing information**, including the plug-in/plug-out times or energy consumed, for example, that prevents the correct implementation of clustering methods and the desired outcomes. Interpolation using nearby entries can be used to replace these absent values. Another possibility would be to remove the datasets' rows corresponding to missing entries, resulting in a dataset with solely accurate and unaltered data. The optimal alternative should be studied and evaluated for each dataset. On the other side, there might also be inaccurate information in some entries, such as an abnormal energy supply, or EV drivers with an excessive number of sessions. These points, known as **outliers**, should be handled and eliminated using, for instance, techniques like Interquartile Range (IQR), Elliptic Envelope, Isolation Forest, or by defining thresholds for data removal.

IQR [61] is the range between the first ($Q1$) and the third ($Q3$) quartiles: $IQR = Q3 - Q1$. The data points which fall below $Q1 - 1.5IQR$ or above $Q3 + 1.5IQR$ are considered outliers. **Elliptic Envelope**'s algorithm [62] creates an imaginary elliptical area around the given dataset. Values that fall outside the envelope are returned as outliers. This model performs best when the dataset has a Gaussian distribution. **Isolation Forest** [63] is based on the Decision Tree algorithm. It isolates the outliers by randomly selecting a feature from the given dataset and then selecting a split value between the max and min values of that feature. This random partitioning of features will produce shorter paths in trees for the anomalous data points, thus distinguishing them from the rest of the data.

The preprocessing stages of Shahriar and Al-Ali [44] included converting the arrival (plug-in) and departure (plug-out) values to a suitable numeric structure. The minute was divided by 60, converting, for instance, 10h17 to 10.28h. Furthermore, the length of charging sessions was calculated by subtracting the arrival time from the departure time since the ACN-Data dataset does not have the *charging duration* field. This may be a crucial and essential preprocessing step since this dataset will also be employed in this thesis' cluster analysis. Märtz et al. [46] found that the continuous values used to represent the date and time of EV plug-in and plug-out are difficult to cluster, as they are scattered throughout the year. Thus, the BMW i3 dataset was adjusted to allow useful clustering. The first step was to set all charging activities to start on the same day while maintaining the plug-in time instant. However, this approach had a drawback related to plug-in and plug-out spatial proximity loss. Consequently, adjustments had to be made to recover the spatial proximity, considering the period of lower charging activity.

One of the most crucial steps corresponds to the **normalization** of the data before clustering, especially when working with several fields/features. Clustering algorithms are sensitive to the scale of the data. These methods work with distances, densities, or both, and if distinct features have different scales, then some features may dominate over others. Normalizing the data ensures that each entry contributes equally to the distance calculation between data points, helping to improve the accuracy of the clustering algorithms and generate good-quality clusters. Consequently, each dataset field (column) should range from 0 to 1, allowing an overall normalization of the data. To achieve this, the MinMaxScaler method from the *scikit-learn* Python library [64] can be applied.

4.3.2 Feature Engineering

Another relevant step comprises creating features not previously included in the dataset that help to analyze and obtain more meaningful clustering, the so-called ***feature engineering*** [65].

As previously mentioned and presented in Table 4.1, the two datasets do not provide all the required fields to obtain EV charging profiles and EV user behavior profiles. Therefore, additional fields must be created, and their calculation adjusted according to the information provided in the datasets.

Two periods can be obtained, based on the study from Develder et al. [66]: the time (t) the EV was parked and plugged into the EVSE (*Sojourn Time*), and the fraction thereof that is effectively spent on charging (*Charging Time*). With these two indicators, the *Idle Time* can be determined, as a measure of flexibility of the charging process. More formally, these new features can be defined as

$$\text{Sojourn Time} = t^{\text{plug-out}} - t^{\text{plug-in}}, \tag{4.1}$$

$$\text{Charging Time} = t^{\text{end charging}} - t^{\text{start charging}}, \tag{4.2}$$

$$\text{Idle Time} = \text{ Sojourn Time} - \text{Charging Time}. \tag{4.3}$$

ACN-Data contains all the necessary information required to compute expressions (4.1) and (4.2), with the start charging time equal to the plug-in time field. However, GR-Data does not provide access to the session's end of charging, and consequently (4.2) cannot be employed. Instead, GR-Data offers information on the maximum power capacity of the EVSEs. As a result, it is possible to obtain an estimated value of the charging time for each session through (4.4). An adjustment factor (equal to 0.8) guarantees a more realistic charging time since this process is not carried out at a constant power rate; it depends on external factors such as temperature, high loads on the grid, and the State of Charge (SoC) (as the battery becomes fully charged, the charging rate decreases), among others [67]. Thus, this factor ensures a 20% safety margin for the maximum power value.

$$\text{Average Charging Time}_{session\ i} = \frac{Energy\ Delivered_i}{(max\ Power\ EVSE)_i \times Adjustment\ Factor} \tag{4.4}$$

Regarding the EV user behavior profiles, the datasets' sessions should be grouped by customer ID to obtain characteristic average values for each EV user, based on the papers reviewed in Section 3.2. Defining **standard deviations** for temporal fields, including *plug-in time*, *charging time*, or *sojourn time*, can offer valuable details into the variability and dispersion of these fields. For instance, a high standard deviation for plug-in time suggests that the driver usually starts charging at no specific time of day.

Additionally, since the main goal is to get insights into the frequency of charging, a new field must be associated with the users to differentiate regular EV drivers from occasional ones, defined by

$$\text{Frequency}_{user_i} = \frac{number\ of\ sessions_i}{Period,\ in\ weeks}, \tag{4.5}$$

where the period in the denominator comprises the number of weeks between the first and last session the EV user attended the Caltech or Greek EVSEs. Thus, the frequency values indicate the average number of days the driver charged its EV, per week.

4.4 Stage 2: Selected Clustering Methods

Three well-known clustering methods are the choice for identifying groups of similar charging patterns and EV user behavior: K-means, GMM, and Hierarchical clustering. To analyze the distribution and accessibility of public EVSEs, DBSCAN will be employed to perform spatial clustering.

These clustering methods are frequently employed in applications related to charging behavior, as revealed in Table 3.1. Moreover, Al-Ogaili et al. [40] and Shahriar et al. [41] precisely suggest the use of these methods for the analysis of EV charging patterns (their studies provide a comprehensive overview of the application of supervised and unsupervised ML techniques in EV and EVSE deployment data).

4.4.1 K-means Clustering

The goal of K-means [34] is to find a clustering that minimizes the Sum of Squared Errors (SSE) score, which measures the accuracy or goodness of the clustering, defined as

$$SSE(C) = \sum_{i=1}^{k} \sum_{\boldsymbol{x}_j \in C_i} \|\boldsymbol{x}_j - \boldsymbol{\mu}_i\|^2, \tag{4.6}$$

where $\boldsymbol{x}_j \in \mathbb{R}^d$ is a point from a given dataset $\mathbf{D}^{n \times d}$ and $\boldsymbol{\mu}_i \in \mathbb{R}^d$ is the centroid of the cluster C_i.

As stated in the pseudo-code given in **Algorithm 4.1**, the points are initially assigned to the clusters at random, with the integer k being the number of clusters. The **elbow method** is typically used to determine the optimal k [68]. The points are then iteratively assigned to new centroids based on how close they are (line 4). In each iteration, the centroids are updated based on the mean of the assigned points (line 6). The process repeats until the centroids stop changing (defined by a threshold), and the algorithm converges.

Algorithm 4.1: K-means

Input: $(\mathbf{D}, k, \epsilon)$

1 Initialize the cluster centroids $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k \in \mathbb{R}^d$ randomly

2 **repeat**

3 **foreach** *data point* $\boldsymbol{x}_j$ **do**

4 calculate distance and assign each $\boldsymbol{x}_j$ to the closest $\boldsymbol{\mu}_i$:

$$C_i := arg\,\min_i \|\boldsymbol{x}_j - \boldsymbol{\mu}_i\|^2$$

5 **foreach** *cluster* C_i **do**

6 compute and update centroids for each cluster:

$$\boldsymbol{\mu}_i := \frac{1}{|C_i|} \sum_{\boldsymbol{x}_j \in C_i} \boldsymbol{x}_j$$

7 **until** $\sum_{i=1}^{k} \|\boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^{t-1}\|^2 \leq \epsilon$;

K-means is typically run multiple times, with the run with the lowest SSE value being selected to report the final clustering. This happens because the method begins with a random guess for the initial centroids. In terms of computational complexity, from **Algorithm 4.1** and assuming t iterations, the total time for K-means is given as $\mathcal{O}(tnkd)$.

4.4.2 GMM Clustering

Given n points $\boldsymbol{x_j}$ in a $d-$dimensional space, let $\mathbf{X} = (X1, X2, \ldots, Xd)$ be the vector random variable across the $d-$attributes, with $\boldsymbol{x_j}$ being a data sample from $\mathbf{X}$. The EM algorithm [35] assumes that each cluster C_i is characterized by a multivariate normal distribution

$$f(\boldsymbol{x} \mid \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}) = \frac{1}{(2\pi)^{(d/2)} |\boldsymbol{\Sigma_i}|^{1/2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu_i})^T \boldsymbol{\Sigma_i}^{-1} (\boldsymbol{x} - \boldsymbol{\mu_i}) \right\}, \tag{4.7}$$

where the cluster C_i centroid $\boldsymbol{\mu}_i \in \mathbb{R}^d$ and the covariance $\boldsymbol{\Sigma_i} \in \mathbb{R}^{d \times d}$ are both unknown parameters and $f(\boldsymbol{x} \mid \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})$ is the probability density at $\boldsymbol{x}$ attributable to cluster C_i.

A Gaussian Mixture Model over all k clusters defines the probability density function of **X**, given as

$$f(\boldsymbol{x}) = \sum_{i=1}^{k} f(\boldsymbol{x} \mid \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}) P(C_i), \tag{4.8}$$

where the prior probabilities $P(C_i)$ satisfy $\sum_{i=1}^{k} P(C_i) = 1$.

Thus, the Gaussian Mixture Model is characterized by the mean $\boldsymbol{\mu}_i$, the covariance $\boldsymbol{\Sigma}_i$, and the *mixture parameters* for each of the k clusters, written compactly as

$$\boldsymbol{\theta} = \{\boldsymbol{\mu}_1, \boldsymbol{\Sigma_1}, P(C_i), \ldots, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, P(C_k)\}. \tag{4.9}$$

After all the key points described, moving forward is thus doable. The goal of EM is to find the maximum likelihood estimates for the parameters $\boldsymbol{\theta}$. To achieve that, EM executes a two-step iterative algorithm (**Algorithm 4.2**) that starts from an initial guess for the parameters $\boldsymbol{\theta}$.

Algorithm 4.2: Expectation-Maximization (EM)

Input: (**D**, k, ϵ)

Initialize the cluster centroids $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^d$ randomly

$\boldsymbol{\Sigma}_i \leftarrow \mathbf{I},\ P(C_i) \leftarrow \frac{1}{k},\ \forall i = 1, \ldots, k$

repeat

  for *i = 1, . . . , k and j = 1, . . . , n* **do**

    Expectation Step (calculate posterior probability):

$$w_{ij} := \frac{f_i(\boldsymbol{x}_j) P(C_i)}{\sum_{a=1}^{k} f_a(\boldsymbol{x}_j) P(C_a)}$$

  for *i = 1, . . . , k* **do**

    Maximization Step (recalculate $\boldsymbol{\theta}$):

$$\boldsymbol{\mu}_i := \frac{\sum_{j=1}^{n} w_{ij} \cdot \boldsymbol{x}_j}{\sum_{j=1}^{k} w_{ij}}$$

$$\boldsymbol{\Sigma}_i := \frac{\sum_{j=1}^{n} w_{ij} (\boldsymbol{x}_j - \mu_i)(\boldsymbol{x}_j - \mu_i)^T}{\sum_{j=1}^{k} w_{ij}}$$

$$P(C_i) := \frac{\sum_{j=1}^{n} w_{ij}}{n}$$

until $\sum_{i=1}^{k} \|\boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^{t-1}\|^2 \leq \epsilon$;

In the **Expectation Step**, given the current estimates for $\boldsymbol{\theta}$, EM computes the cluster posterior probabilities through the Bayes theorem

$$w_{ij} = P(C_i|\boldsymbol{x}_j) = \frac{P(\boldsymbol{x}_j|C_i)P(C_i)}{\sum_{a=1}^{k} P(\boldsymbol{x}_j|C_a)P(C_a)} = \frac{f_i(\boldsymbol{x}_j)P(C_i)}{\sum_{a=1}^{k} f_a(\boldsymbol{x}_j)P(C_a)}, \tag{4.10}$$

since each cluster is modeled as a multivariate normal distribution [31]. Therefore, $P(C_i|\boldsymbol{x}_j)$ can be considered the weight contribution of $\boldsymbol{x}_j$ to cluster C_i.

Next, in the **Maximization Step**, EM recalculates $\boldsymbol{\theta}$ using the weights w_{ij}. The algorithm ends when $\sum_{i=1}^{k} \|\boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^{t-1}\|^2 \leq \epsilon$, where ϵ is the convergence threshold, and t denotes the iteration.

For the Expectation Step, inverting $\boldsymbol{\Sigma}_i$ and computing its determinant takes $\mathcal{O}(kd^3)$, and evaluating the density $f_i(\boldsymbol{x})$ takes $\mathcal{O}(nkd^2)$. For the Maximization Step, the time is dominated by the $\boldsymbol{\Sigma}_i$ update. Assuming t iterations, the computational complexity of the EM method is $\mathcal{O}(t(kd^3 + nkd^2))$.

4.4.3 Agglomerative Hierarchical Clustering

Agglomerative Hierarchical clustering starts with each of the n points in a separate cluster. Then, the two closest clusters are repeatedly merged until all points are members of the same cluster, as shown in the pseudo-code given in **Algorithm 4.3**. Given a set of clusters $\mathcal{C} = \{C_1, C_2, \ldots, C_m\}$, first, the closest pair of clusters C_i and C_j are found and merged into a new cluster, C_{ij}. Next, the set of clusters is updated, removing C_i and C_j and adding C_{ij}. This process is repeated until $\mathcal{C}$ contains exactly k clusters.

Algorithm 4.3: Agglomerative Hierarchical Clustering

Input: ($\mathbf{D}$, k)
Initialize each cluster with a single point $\mathcal{C} \leftarrow C_i = \{\boldsymbol{x}_i\}, \forall i = 1, \ldots, n$
Compute the distance matrix $\boldsymbol{\Delta} \leftarrow \|\boldsymbol{x}_i - \boldsymbol{x}_j\|, \forall i = 1, \ldots, n\,; \forall j = 1, \ldots, n$
repeat
    Find the closest pair of clusters: $C_i, C_j \in \mathcal{C}$
    Merge clusters $C_{ij} \leftarrow C_i \cup C_j$
    Update $\mathcal{C} \leftarrow (C \setminus \{C_i, C_j\} \cup \{C_{ij}\}$ and $\boldsymbol{\Delta}$ to reflect new clustering
until $|\mathcal{C}| = k$;

Finding the closest pair of clusters is the algorithm's key step. For this, a variety of distance measures can be employed [69] (see **Figure 4.4**), including:

- **Single link**: The distance between two clusters is defined as the minimum distance between a point in C_i and a point in C_j. First developed by Florek et al. [70] and then independently by McQuitty (1957) and Sneath (1957) [71];

- **Complete link**: The distance between two clusters is defined as the maximum distance between a point in C_i and a point in C_j. Developed by Sørenson in 1948 [72];

- **Average link**: The distance between two clusters is defined as the average pairwise distance between points in C_i and C_j. Developed by Sokal and Michener (1958) [73] to avoid the extremes introduced by either single or complete link;

- **Mean distance**: The distance between two clusters is defined as the distance between the centroids of the two clusters. The earliest use known of this strategy is that of Sokal and Michener (1958) [73].

But, possibly the most employed measure is **Ward's Method**, introduced by Joe H. Ward, Jr. in 1963 [74]. The distance between two clusters is defined as the increase in the sum of squared errors when the two clusters are merged. The objective is to minimize the total within-cluster variance. It can be seen as a weighted version of the mean distance measure, as it weights the distance between centroids by half of the harmonic mean of the cluster size.

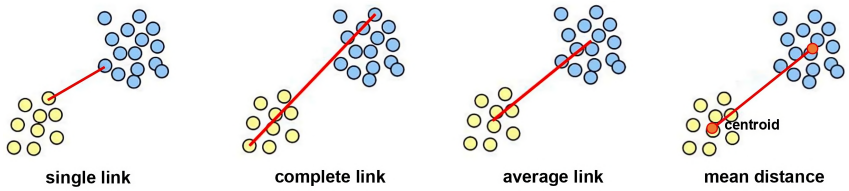


Figure 4.4: Different distance measures for Agglomerative Hierarchical clustering (Adapted from [75]).

When two clusters C_i and C_j combine to form C_{ij}, the distances between C_{ij} and each of the remaining clusters $C_r (r \neq i, r \neq j)$ must be updated in the matrix $\mathbf{\Delta}$. For all of the cluster proximity measures, the **Lance-Williams** [76] formula offers a general equation to recompute the distances:

$$\delta(C_{ij}, C_r) = \alpha_i \cdot \delta(C_i, C_r) + \alpha_j \cdot \delta(C_j, C_r) + \beta \cdot \delta(C_i, C_j) + \gamma \cdot |\delta(C_i, C_r) - \delta(C_j, C_r)|, \quad (4.11)$$

where the parameters $\alpha_i, \alpha_j, \beta$ and γ differ from one measure to another [76].

In terms of computational complexity, Agglomerative Hierarchical clustering initially takes $\mathcal{O}(n^2)$ time to create the distance matrix $\mathbf{\Delta}$, and updating/deleting distances from it takes $\mathcal{O}(\log(n))$ time for each operation, leading to a total of $\mathcal{O}(n^2 \log(n))$.

4.4.4 Density-based Clustering (DBSCAN)

DBSCAN [37], being a density-based clustering technique, can discover nonconvex clusters, unlike representative-based techniques that can only find Ellipsoid-shaped or convex clusters. DBSCAN uses the local density of points to determine the clusters, rather than using only the distance between points. The neighborhood of $\boldsymbol{x} \in \mathbb{R}^d$ is defined as

$$N_\epsilon(\boldsymbol{x}) = \delta(\boldsymbol{x}, \boldsymbol{y}) \leq \epsilon\,, \quad (4.12)$$

where $\delta(\boldsymbol{x}, \boldsymbol{y})$ represents the distance between points $\boldsymbol{x}$ and $\boldsymbol{y}$ (usually Euclidean distance, but it might be a different metric). The threshold ϵ needs to be specified.

In order to fully understand the algorithm, it is first necessary to define some important concepts:

- $\boldsymbol{x}$ is a *core point* if there are at least $minpts$ points in its ϵ-neighbourhood ($N_\epsilon(\boldsymbol{x}) \geq minpts$, with $minpts$ a user-defined threshold);

- A *border point* does not meet the $minpts$ threshold, but it belongs to the ϵ-neighbourhood of another point $\boldsymbol{z}$, $\boldsymbol{x} \in N_\epsilon(\boldsymbol{z})$;
- A *noise point* is neither a core nor a border point (outlier);
- $\boldsymbol{x}$ is *density reachable* from $\boldsymbol{y}$ if there is a set of core points leading from $\boldsymbol{y}$ to $\boldsymbol{x}$;
- Two points $\boldsymbol{x}$ and $\boldsymbol{y}$ are *density connected* if there exists a core point $\boldsymbol{z}$ such that both $\boldsymbol{x}$ and $\boldsymbol{y}$ are density reachable from $\boldsymbol{z}$.

Having the key concepts defined, understanding the pseudo-code for the DBSCAN method shown in **Algorithm 4.4** is thus possible.

Algorithm 4.4: DBSCAN

1 **DBSCAN** *(**D**, ϵ, minpts)*:
2 $Core \leftarrow \emptyset, k \leftarrow 0$ // initialize core points and cluster id
3 **foreach** $\boldsymbol{x}_i \in \boldsymbol{D}$ **do** // find core points
4 Compute $N_\epsilon(\boldsymbol{x}_i)$
5 $id(\boldsymbol{x}_i) \leftarrow \emptyset$
6 **if** $N_\epsilon(\boldsymbol{x}_i) \geq minpts$ **then** Core $\leftarrow$ Core $\cup \{\boldsymbol{x}_i\}$
7 **foreach** $\boldsymbol{x}_i \in Core$, *with* $id(\boldsymbol{x}_i) = \emptyset$ **do**
8 $k \leftarrow k + 1$
9 $id(\boldsymbol{x}_i) \leftarrow k$ // assign $\boldsymbol{x}_i$ to cluster id k
10 **DENSITYCONNECTED**$(\boldsymbol{x}_i, k)$
11 $\mathcal{C} \leftarrow \{C_i\}_{i=1}^k, C_i \leftarrow \{\boldsymbol{x} \in \boldsymbol{D} \,|\, id(\boldsymbol{x}) = i\}$ // define clusters
12 $Noise \leftarrow \{\boldsymbol{x} \in \boldsymbol{D} \,|\, id(\boldsymbol{x}) = \emptyset\}$
13 $Border \leftarrow \boldsymbol{D} \setminus \{Core \cup Noise\}$
14 **DENSITYCONNECTED** $(\boldsymbol{x}, k)$:
15 **foreach** $\boldsymbol{y} \in N_\epsilon(\boldsymbol{x})$ **do** // density connected points to $\boldsymbol{x}$
16 $id(\boldsymbol{y}) \leftarrow k$ // assign $\boldsymbol{y}$ to cluster k
17 **if** $\boldsymbol{y} \in Core$ **then** DENSITYCONNECTED $(\boldsymbol{y}, k)$

First, DBSCAN computes the ϵ-neighborhood for each point $\boldsymbol{x}_i$ in the dataset, checks if it is a core point (lines 3–6) and sets the cluster id null for all points. Next, starting from each unassigned core point, the method finds all its density-connected points recursively, which are assigned to the same cluster (line 11). Some border points may be accessible from core points in more than one cluster. As DBSCAN is a sequential algorithm, they will be arbitrarily assigned to the first created cluster that can incorporate that specific border point.

Regarding the computational complexity, it takes $\mathcal{O}(n^2)$ to compute the neighborhood for each point when the dimensionality is high. With $N_\epsilon(\boldsymbol{x})$ computed, the algorithm needs only a single pass over all points to find the density connected clusters, leading to the overall complexity $\mathcal{O}(n^2)$.

4.5 Stage 3: Clustering Validation Techniques

Since no ground truth is available, internal validation should be used to quantify the performance of the clustering [31]. Three internal validation metrics, Silhouette coefficient [77], Davies-Bouldin index [78], and Calinski-Harabasz index [79] can be employed, based on the studies presented and reviewed in Section 3.2.

4.5.1 Silhouette Coefficient

For each point $\boldsymbol{x}_i$, the silhouette coefficient is

$$s_i = \frac{\mu_{out}^{min}(\boldsymbol{x}_i) - \mu_{in}(\boldsymbol{x}_i)}{\max\{\mu_{out}^{min}(\boldsymbol{x}_i), \mu_{in}(\boldsymbol{x}_i)\}}, \tag{4.13}$$

where $\mu_{out}^{min}(\boldsymbol{x}_i)$ is the mean of the distances from $\boldsymbol{x}_i$ to points in the closest cluster, and $\mu_{in}(\boldsymbol{x}_i)$ is the mean distance from $\boldsymbol{x}_i$ to points in its own cluster.

The total **Silhouette coefficient** [77] is defined as the mean s_i value across all points, given by (4.14), where a value close to $+1$ denotes good clustering.

$$SC = \frac{1}{n}\sum_{i=1}^{n} s_i \tag{4.14}$$

4.5.2 Davies-Bouldin Index

The Davies-Bouldin measure for a pair of clusters C_i and C_j is defined as

$$DB_{ij} = \frac{\sigma_{\mu_i} + \sigma_{\mu_j}}{\delta(\mu_i, \mu_j)}, \tag{4.15}$$

where μ_i denotes the centroid of cluster C_i, $\sigma_{\mu_i} = \sqrt{var(C_i)}$ represents the dispersion of the points around the respective centroid (square root of the total variance) and $\delta(\mu_i, \mu_j)$ is the distance between the centroids.

The **Davies-Bouldin index** [78] is thus defined as

$$DB = \frac{1}{k} \cdot \sum_{i=1}^{k} \max_{i \neq j}\{DB_{ij}\}, \tag{4.16}$$

meaning that for each cluster C_i it is chosen the cluster C_j that returns the largest DB_{ij} ratio. Therefore, smaller DB values, closer to zero, mean better clustering (clusters are well separated and each one is well represented by its centroid).

4.5.3 Calinski-Harabasz Index

Given the dataset **D** $= \{\boldsymbol{x}_i\}_{i=1}^{n}$, the **Calinski-Harabasz index** [79] is given by

$$CH(k) = \frac{tr(\boldsymbol{S}_B)}{tr(\boldsymbol{S}_W)} \cdot \frac{n-k}{k-1}, \tag{4.17}$$

where $tr(\boldsymbol{S}_B)$ is the trace of the within-cluster scatter matrix, and $tr(\boldsymbol{S}_W)$ is the trace of the between-cluster scatter matrix. Those matrices are defined by (4.18) and (4.19), respectively, where $\boldsymbol{\mu}$ is the dataset's mean and $\boldsymbol{\mu}_i$ is the mean for cluster C_i.

$$\boldsymbol{S}_B = \sum_{i=1}^{k} n_i \, (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T, \tag{4.18}$$

$$\boldsymbol{S}_W = \sum_{i=1}^{k} \sum_{\boldsymbol{x}_j \in C_i} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T, \tag{4.19}$$

A good value k (number of clusters) should result in a high $CH(k)$. The intuition is to determine the value of k for which $CH(k)$ is higher than $CH(k-1)$ and there is a slight improvement or a decrease in the $CH(k+1)$ value. This way, the Calinski-Harabasz index can be also used to choose the number of clusters that maximize $CH(k)$, an alternative to the elbow method typically used for K-means [68].

5

Results

Contents

This chapter is divided into several sections, which are equally relevant to this study. Starting with the datasets' preprocessing steps, it follows the presentation of EV charging profiles, EV user behavior profiles, and EVSE accessibility. It ends with a description of possible practical applications for this thesis's results. The code was written in Python 3.10.11 on a Jupyter Notebook using the Google Colab platform, and mainly the *scikit-learn* library [64] for the preprocessing, clustering, and evaluation methods (most parameters were left at default, while those modified are mentioned throughout the text).

5.1 Data Preprocessing

5.1.1 ACN-Data dataset

5.1.1.A Dataset preparation and feature engineering

The first step in obtaining superior clustering results is data preprocessing, according to the schematic in Figure 4.1. Since the dataset is provided in a JSON file, several conversions were necessary to obtain each field in the required format. With the help of the DateTime method of *Pandas* library [80], the fields *connectTime*, *disconnectTime*, and *doneChargingTime* were converted into DateTime values to allow further data processing, namely obtaining the *sojournTime*, *chargingTime*, and *idleTime* fields based on (4.1), (4.2) and (4.3), respectively. The *kWhDelivered*, *stationID*, and *userID* were converted to float format. All the additional fields were redundant or unnecessary for the work and thus discarded.

After determining the extra fields, the entries in DateTime format needed to be converted into a suitable numeric structure, as mentioned by Shahriar and Al-Ali [44] and by Märtz et al. [46]. The *connectionTime*, *disconnectTime*, *doneChargingTime*, *sojournTime*, *chargingTime*, and *idleTime* fields were therefore transformed into float format: the DateTime values were converted to seconds and then divided by 3 600, to get the instant of the day only in an hour scale, and not in date, hours and minutes. For example, 10h17 (10 hours and 17 minutes) becomes 10.28h (10.28 hours), consequently allowing full use of outlier removal approaches, clustering methods, and graphical representations. It is worth noting that it is necessary to choose only a subset of the available fields to obtain interpretable results.

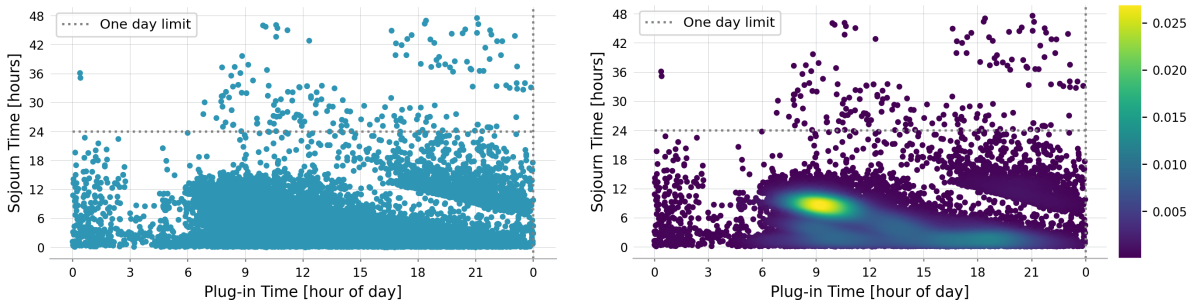
5.1.1.B Deal with Missing Data

There were no missing entries in the preprocessed dataset, except in the *userID* and *doneChargingTime* fields. The absence of these *doneChargingTime* entries indicates that the charging time was insufficient to obtain a fully charged battery. Thus, this field was assigned with the value of the *disconnectTime* entry in these sessions, leading to an idle time of zero. Regarding the *userID*, the lack of this information makes it impossible to discover or predict the corresponding session user. Hence, for the characterization of user behavior (presented in Section 5.3), only sessions with a *userID* will be considered.

5.1.1.C Outlier Detection

With a fully functional dataset, the next step involved setting thresholds to remove unwanted data. The limit defined allowed removing sessions with a *sojournTime* or *chargingTime* greater than 48 hours and less than 1 minute. Another threshold was also set to clear sessions with energy-supplied values greater than 100 kWh, selected considering the period of the data (2018-2021) and the characteristics of EVs available on the market during these years. All negative entries were also removed.

Figure 5.1 presents the distribution of the clean ACN-Data sessions regarding the Sojourn Time and Plug-in Time, demonstrating roughly three main groups of data: one from 00h00 to 03h00, with scattered sojourn times; another from 06h00 to the end of the day, with longer sojourn times when connecting in the morning; and finally, another between 17h00 and 23h59, with higher sojourn times. For better visualization of the data distribution, Figure 5.1(b) illustrates the density of the points.



(a) Scatter plot.

(b) Density Scatter plot.

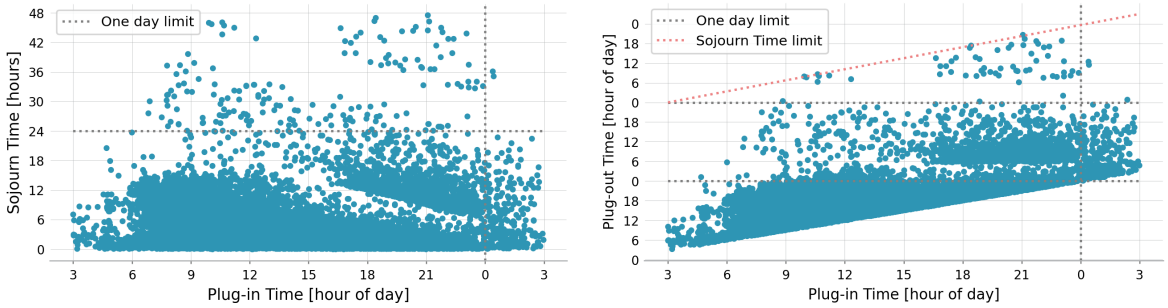
Figure 5.1: Clean ACN-Data distribution regarding Sojourn Time and Plug-in Time.

Observing Figure 5.1, some points are more scattered and distant from the three main groups identified previously. These points are the so-called **outliers**. However, when analyzing the data, one can see that these points represent behavior that could have happened and are not errors in the data, even though they are distant from most sessions. The grosser errors, effectively outliers, have already been identified and eliminated when defining the fields' thresholds. Therefore, no outlier removal method will be employed for the identification of the ACN-Data EV charging profiles and EV user behavior profiles.

5.1.1.D Data Adjustment

The plug-in time with day and hours became plug-in time at the hour of the day when the DateTime variables were converted into float values. The drawback of this strategy is that the time frame under consideration was 00h00 to 23h59. Due to their loss of spatial proximity, early and late plug-in times might be clustered separately. As shown in Figure 5.1, there is an instant when charging activity is at its lowest, fluctuating throughout the night and early morning, reaching its lowest point around 03h00. To restore the spatial proximity, all charging sessions with plug-in times less than this minimum were

relocated to the right side to continue the timeframe after 23h59. Figure 5.2(a) presents the adjusted distribution in terms of Sojourn Time and Plug-in Time, while Figure 5.2(b) displays the adjusted data distribution from a different perspective, regarding the Plug-out Time and Plug-in Time.



(a) Distribution regarding Sojourn and Plug-in Time. **(b)** Distribution regarding Plug-out and Plug-in Time.

Figure 5.2: Final adjusted ACN-Data scatter plot distribution regarding different fields.

From Figure 5.2(b), it is possible to see more noticeably the existence of sessions that start and end on different days. A limit indicating the defined threshold of 48 hours maximum sojourn time is also present, verifying that, in fact, no sessions surpass this threshold in the final adjusted dataset. Table 5.1 summarizes the final available fields present in the clean and preprocessed dataset.

Table 5.1: Summary of the final usable fields in the ACN-Data dataset.

Field name	Non-Null count	Dtype
connectionTime (Plug-in Time)	31318	float64
disconnectTime (Plug-out Time)	31318	float64
doneChargingTime (End Charging Time)	31318	float64
kWhDelivered	31318	float64
stationID	31318	float64
userID	16355	float64
chargingTime	31318	float64
sojournTime	31318	float64
idleTime	31318	float64

5.1.2 GR-Data dataset

5.1.2.A Dataset preparation and feature engineering

Since this dataset had the same format as the ACN-Data, the steps followed for the conversion of the entries were identical, only changing the fields' names according to the GR-Data's specific characteristics. The *sojournTime* is present in this dataset, unlike the *chargingTime* field. Consequently, it was necessary to determine the latter, utilizing the expression (4.4) previously mentioned. With the average charging time defined, it was then possible to obtain the idle time through (4.3). The entries in DateTime format were converted to float values according to the process previously described in Section 5.1.1.A.

The final dataset fields *Start_datetime*, *End_datetime*, *kWhDelivered*, *stationID*, *maxPowerEVSE*, *userID*, *sojournTime*, *averageChargingTime*, and *idleTime* are all available in float format (see Table 5.2).

5.1.2.B Deal with Missing Data

There were no missing entries in the dataset, except in the *kWhDelivered* and *maxPowerEVSE*, making it impossible to determine the average charging time and idle time. Thus, these sessions were discarded. Additionally, some sessions presented an average charging time higher than the sojourn time, indicating that the EV was effectively charging during the entire parking period and that the adjustment factor was too harsh for these particular sessions. Accordingly, the *averageChargingTime* was assigned with the value of the *sojournTime* entry in these sessions, leading to a corresponding idle time of zero.

5.1.2.C Outlier Detection

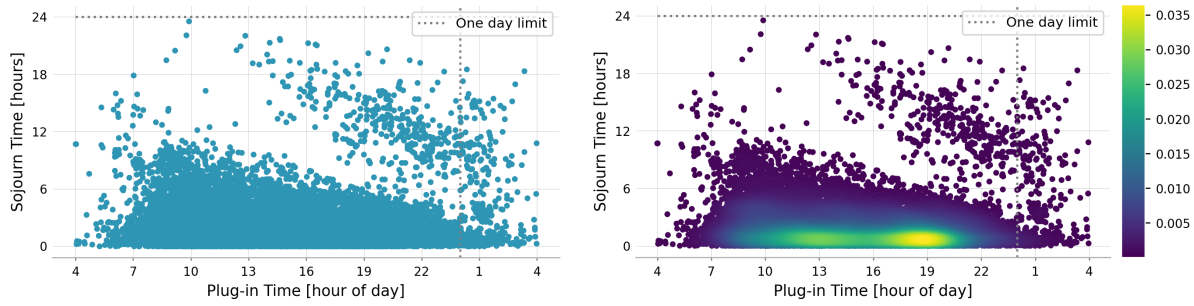
The defined thresholds match those specified for the previous dataset, with slight differences: 24-hour charging time and sojourn time limit, only sessions with more than 1 minute of sojourn time, and maximum energy delivered of 100 kWh (considering the 2021-2022 EV sales in Europe). There are only 32 sessions with more than 24 hours of parking stay, which does not correspond to an actual profile. Consequently, by removing these 32 sessions, the clustering results will be improved, yielding more meaningful clusters. All negative entries were also removed.

5.1.2.D Data Adjustment

Due to their loss of spatial proximity, early and late plug-in times might be clustered separately, as previously discussed. Similarly to the ACN-Data dataset, there is an instant when charging activity is at its lowest, around 04h00. To restore the spatial proximity, all charging sessions with plug-in time less than this minimum were relocated to the right side to continue the timeframe after 23h59. The final clean and preprocessed dataset is illustrated in Figure 5.3 regarding Sojourn Time and Plug-in Time. Table 5.2 contains all the usable fields from the final preprocessed GR-Data dataset.

Table 5.2: Summary of the final usable fields in the GR-Data dataset.

Field name	Non-Null count	Dtype
Start_datetime (Plug-in Time)	21801	float64
End_datetime (Plug-out Time)	21801	float64
kWhDelivered	21801	float64
stationID	21801	object
maxPowerEVSE	21081	float64
userID	21801	object
sojournTime	21801	float64
averageChargingTime	21801	float64
idleTime	21801	float64



(a) Scatter plot.

(b) Density Scatter plot.

Figure 5.3: Final adjusted GR-Data distribution regarding Sojourn Time and Plug-in Time.

5.2 EV Charging profiles

5.2.1 ACN-Data dataset

5.2.1.A Chosen fields and normalization of the data

Based on the paper by Shahriar and Al-Ali [44], analyzed in Section 3.2, the first step consisted of only choosing *connectionTime* and *disconnectTime* for the same period to verify if the obtained outcomes were identical. With three clusters, the results were, in fact, very similar to those presented in the study, indicating a correct use and handling of the code and the data processing tools. However, these clusters are far from representing the real behavior of users or EVSEs. Each cluster suggests a wide range of values, with no information about the energy consumed in each session. Therefore, the next step consisted of carrying out several studies to obtain more accurate and less generic clusters, even as suggested by the authors. Figure 5.4 presents the heatmap of the correlation between the features. The created correlation matrix reveals the single correlation between each field on the dataset. Positive and negative values indicate whether the features are directly or inversely related; e.g., a correlation of -0.7 between two fields denotes that if one variable increases, the other decreases strongly.

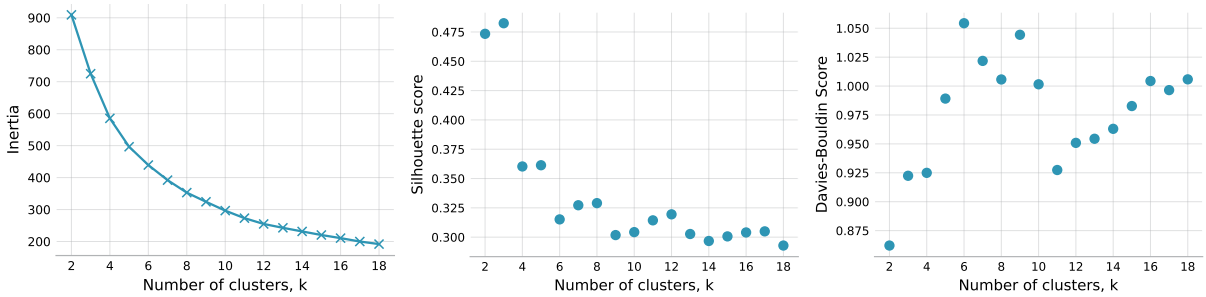
Figure 5.4: Correlation matrix of the fields in the clean ACN-Data dataset.

By analyzing the covariance matrix between the available features, interesting patterns arise. The highly correlated *connectionTime*, *disconnectTime*, and *doneChargingTime* fields are redundant, so only one is required (*connectionTime* provides intelligible information, and thus it must be chosen). The same reasoning applies to *kWhDelivered* and *chargingTime*. The *sojournTime* strongly correlates with *disconnectTime*, *chargingTime*, and *idleTime* (redundant, as expected), so only one should be selected. However, *connectionTime* exhibits an inverse relationship with *sojournTime*, so choosing the latter is a viable option to capture more diversity of the data. Additionally, the positive correlation between *userID* and *stationID* suggests that users may have favorite EVSEs they attend more frequently. Ultimately, it is crucial to select complementary, non-redundant features that align with the current objective.

Based on the multiple papers reviewed in Section 3.2 and the analysis on the correlation matrix, the *connectionTime*, *sojournTime*, and *kWhDelivered* fields were chosen to obtain EV charging profiles since this triplet yielded the best results in a first cluster analysis. The remaining fields were eliminated, and the data were normalized using the MinMaxScaler method [64] to obtain the best possible results. The final dataset is thus ready to perform clustering. Several methods were applied, and the optimal number of clusters was determined for each. The obtained results are detailed in the following sections.

5.2.1.B K-means Clustering

The number of clusters, k, was chosen based on the elbow method [68], applied using the inertia values from the K-means method of Python's *scikit-learn* library [64]. Inertia is the sum of the squared distances of samples to their closest cluster center. It is also known as the Within-Cluster Sum-of-Squares (WCSS). A study was also conducted to determine the values of the Silhouette, Davies-Bouldin, and Calinski-Harabasz scores based on the number of clusters. This approach enables the identification of the optimal k that leads to the highest scores. Figure 5.5 illustrates the different scores as a function of k, indicating that k=2 originates the best scores. However, as previously mentioned, a higher k is necessary to obtain interpretable and meaningful EV charging profiles. Table A.1 reveals the precise values of the Silhouette, Davies-Bouldin, and Calinski-Harabasz scores, according to k.

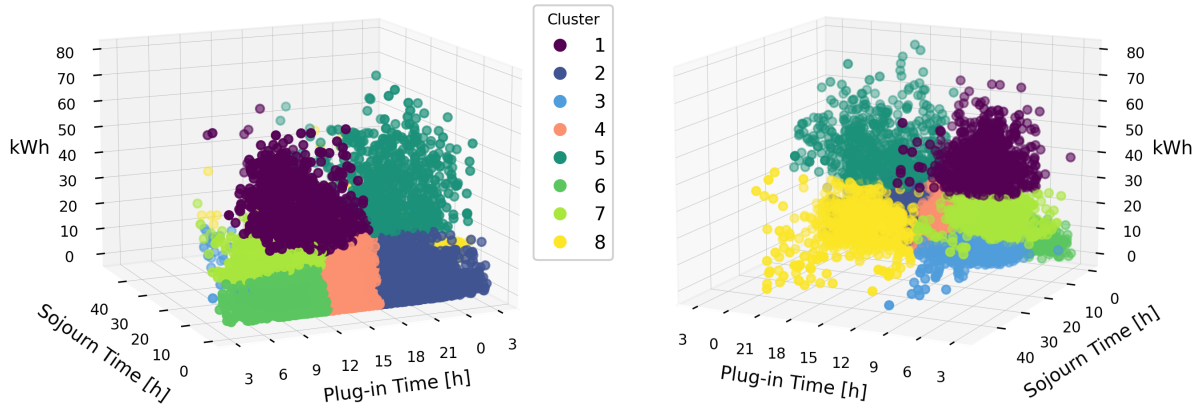


(a) Elbow Method. **(b)** Silhouette Coefficient. **(c)** Davies-Bouldin Index.

Figure 5.5: Different scores as a function of k for the ACN-Data K-means clustering.

The elbow method does not effectively display an elbow, making it insufficient for determining the ideal number of clusters. Nevertheless, the knee of the curve suggests k from 5 to 8. Within this range, by performing a more in-depth analysis, the best results are thus found for k=5 or k=8, with higher Silhouette and lower Davies-Bouldin scores, as seen in Figure 5.5(b) and 5.5(c), respectively. The Calinski-Harabasz scores behave similarly to inertia (elbow method), not helping in this study. For k=9, both scores are worse, indicating that the optimal value is indeed in the range previously defined.

Analyzing the profiles with 5 and 8 clusters, one realizes that choosing k=5 yields still quite generic profiles that comprise relatively different behaviors within the same clusters. With k=8, on the other hand, the clusters are better defined and identifiable. Figure 5.6 presents the distribution of the adjusted EV charging profiles regarding the Plug-in Time, Sojourn Time, and kWh (energy delivered) fields.



(a) Azimuth = -115°. **(b)** Azimuth = 115°.

Figure 5.6: 3D distribution of the adjusted K-means EV Charging profiles for the ACN-Data dataset.

From Figure 5.6, one can see that the profiles are well-defined and have little overlap. An intriguing result that is immediately apparent is the separation of the high consumption profiles (clusters 1 and 5), which are virtually divided by the plane defined by kWh $\approx$ 30, from the low and medium consumption profiles (clusters 2, 3, 4, 6, 7, and 8). Additionally, there are more short/medium-term sessions, which impacts the number of profiles. The longer sojourn time sessions are comprised in clusters 3 and 8.

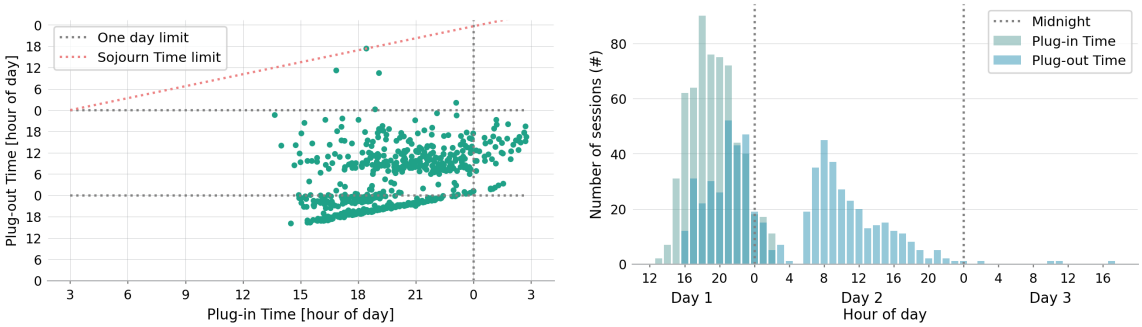
Table 5.3 lists the mean quantitative characteristics of the eight profiles, from which one can see that cluster 5 behaves slightly differently from the others, with a plug-out time close to 05h00 and around 600 sessions, indicating that this profile is the least common, as it contains sessions that start in the late afternoon and finish the following day (early morning). To get a better perspective on this behavior, Figure 5.7 illustrates two distinct representations of cluster 5 sessions. Roughly half of them start and end on the same day (late afternoon). The remaining sessions only end the next day, with a higher incidence in the morning, suggesting that EVs stay connected to the EVSE during the night. Since cluster 5 comprises two distinct behaviors, the average plug-out time does not fully reflect all sessions.

Table 5.3: Mean quantitative characteristics of the K-means EV Charging profiles for the ACN-Data dataset.

Cluster ID	No. of Sessions	Plug-in Time	Plug-out Time	Energy [kWh]	Sojourn Time	Charging Time	Idle Time	Profile*
1	1174	10h12	16h51	34.735	6h 38min	5h 28min	1h 10min	Morning to afternoon, high energy, long-term
2	5420	19h14	21h05	7.215	1h 51min	1h 28min	22min	Evening short-term stay, low energy
3	6305	09h30	18h22	4.765	8h 52min	3h 12min	5h 41min	Morning to afternoon long-term, low energy
4	6588	14h05	17h11	6.165	3h 06min	1h 52min	1h 15min	Afternoon medium-term stay, low energy
5	609	19h51	04h33	39.390	8h 42min	5h 16min	3h 26min	Evening to next morning, high energy, long-term
6	4671	09h15	11h43	4.987	2h 28min	1h 43min	45min	Morning medium-term stay, low energy
7	5399	09h12	17h11	14.322	7h 59min	4h 55min	3h 05min	Morning to afternoon, medium energy
8	1152	19h51	10h49	12.777	14h 58min	6h 01min	8h 57min	Evening to next morning, medium energy

*Note: "Low energy": below 10 kWh; "Medium energy": between 10 kWh and 30 kWh; "High energy": over 30 kWh.
"Short-term": sojourn time below 2h; "Medium-term": between 2h and 4h; "Long-term": over 4h.

The problem with cluster 5 does not occur with the remaining profiles, as they are more uniform and well-defined. For instance, despite the apparent spatial proximity verified in Figure 5.6 (high energy clusters), cluster 1 corresponds to sessions from the beginning until the end of the same day. Regarding the lower energy clusters with similar plug-in times, cluster 2 only contains sessions that start and end on the same day, and cluster 8 only contains sessions that start in the late afternoon and end on the following day (as confirmed by the high sojourn time in Table 5.3).



(a) Scatter plot of sessions.

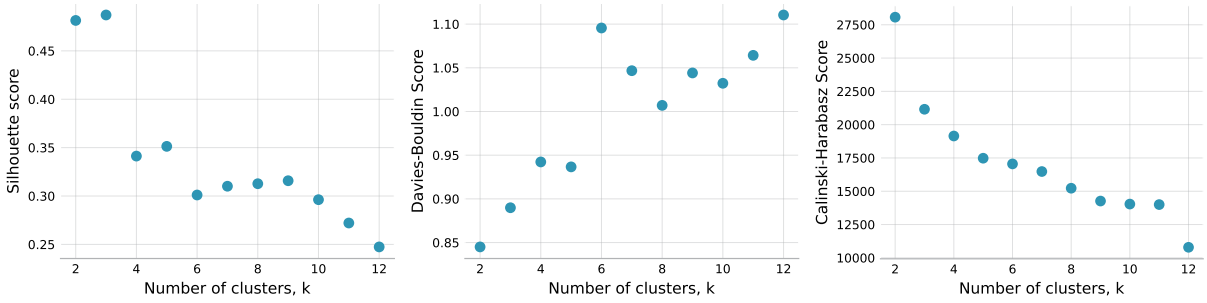
(b) Distribution of sessions.

Figure 5.7: Deep examination of K-means ACN-Data cluster 5, regarding the Plug-in and Plug-out times.

5.2.1.C GMM Clustering

According to the *scikit-learn* website [64], the GMM method includes four choices for the covariance type: *full covariance* (each component has its own overall covariance matrix), *tied covariance* (all components share the same overall covariance matrix), *diagonal covariance* (each component has its own diagonal covariance matrix), and *spherical covariance* (each component has its own unique variance).

Consequently, in addition to the number of clusters, it was also necessary to understand which type of covariance provides the best profiles and scores. Thus, a preliminary cluster analysis proved that *tied covariance* originates meaningful profiles, achieving the best Silhouette, Davies-Bouldin, and Calinski-Harabasz scores, regardless of the number of clusters. Figure 5.8 illustrates the plots of the different scores as a function of the number of clusters.

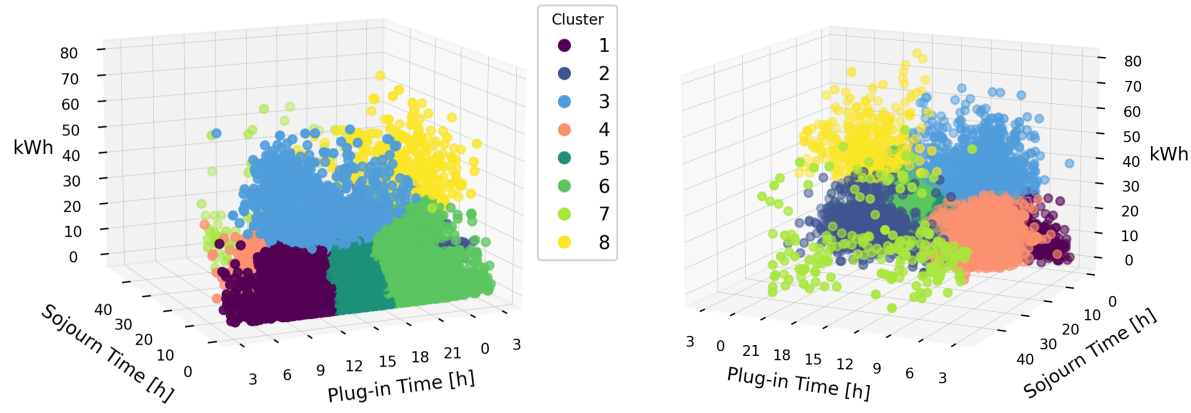


(a) Silhouette Coefficient. **(b)** Davies-Bouldin Index. **(c)** Calinski-Harabasz Index.

Figure 5.8: Different scores as a function of k for the ACN-Data GMM clustering, considering tied covariance.

According to the scores and considering the previous K-means study (Section 5.2.1.B), the optimal number of clusters should also be 5 or 8 since these k's have higher Silhouette and lower Davies-Bouldin scores than the k's immediately below or above (Figures 5.8(a) and 5.8(b), respectively). The Calinski-Harabasz scores did not assist in determining k since they exhibit hyperbolic behavior (Figure 5.8(c)).

Analyzing the results with 5 and 8 clusters, one realizes that choosing 5 clusters yields profiles that are still quite generic and comprise relatively different behaviors within the same cluster, just like seen with the K-means clustering. With k=8, on the other hand, the clusters are better defined and better identifiable. Figure 5.9 presents the distribution of adjusted EV charging profiles regarding the Plug-in Time, Sojourn Time, and kWh (energy delivered) fields.



(a) Azimuth = -115°. **(b)** Azimuth = 115°.

Figure 5.9: 3D distribution of the adjusted GMM EV Charging profiles for the ACN-Data dataset, considering tied covariance.

From Figure 5.9, one can see that the profiles are well-defined and have little overlap. A separation between high-energy and low-energy sessions is also present here. However, unlike K-means clustering, GMM cluster 3 contains most of the high-energy short sessions, regardless of the plug-in time. Cluster 8, on the other hand, contains only high-energy nighttime sessions. Consequently, with GMM clustering, the problem of K-means cluster 5 (Figure 5.7) does not occur, which is a good achievement.

Another intriguing result corresponds to cluster 7. This cluster contains all the most different sessions, characterized by high sojourn times, regardless of the plug-in time. These sessions would be the ones considered outliers, and the GMM method was able to group them all in a single cluster, allowing the remaining clusters to reveal typical average values more adjusted to the sessions they contain. Cluster 5 maintains roughly the same behavior as K-means cluster 4, which comprises the afternoon and low-energy sessions. Table 5.4 lists the mean quantitative characteristics of the eight profiles.

Table 5.4: Mean quantitative characteristics of the GMM EV Charging profiles for the ACN-Data dataset.

Cluster ID	No. of Sessions	Plug-in Time	Plug-out Time	Energy [kWh]	Sojourn Time	Charging Time	Idle Time	Profile*
1	5047	09h10	11h36	6.304	2h 26min	1h 49min	37min	Morning medium-term stay, low energy
2	1110	20h14	09h13	13.136	12h 59min	5h 37min	7h 22min	Evening to next morning, medium energy
3	1191	10h48	16h44	35.512	5h 56min	5h 08min	48min	Morning to afternoon high-term, high energy
4	11779	09h32	18h02	8.547	8h 30min	3h 55min	4h 34min	Morning to afternoon high-term, low energy
5	6281	14h14	16h57	6.134	2h 44min	1h 46min	57min	Afternoon medium-term stay, low energy
6	5393	19h17	21h13	8.267	1h 55min	1h 34min	22min	Evening short-term stay, low energy
7	197	15h00	21h57	15.867	30h 57min	10h 16min	20h 41min	Extreme sojourn times, medium energy
8	320	20h58	09h04	44.606	12h 06min	7h 11min	4h 56min	Evening to next morning, high energy

*Note: “Low energy”: below 10 kWh; “Medium energy”: between 10 kWh and 30 kWh; “High energy”: over 30 kWh.
“Short-term”: sojourn time below 2h; “Medium-term”: between 2h and 4h; “Long-term”: over 4h.

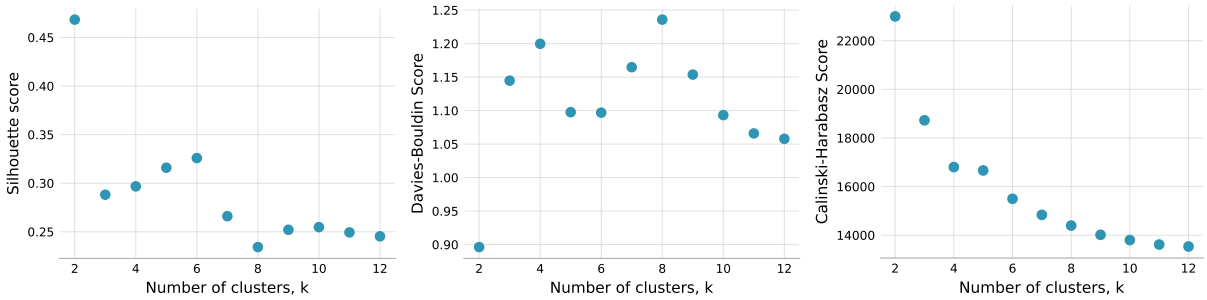
According to Table 5.4, cluster 4 comprises more than one-third of the total number of sessions, characterized by morning plug-in time and late afternoon plug-out time. K-means clustering distributed these sessions across clusters 3 and 7, while GMM clustering grouped all these sessions into a single profile. This merging of the clusters enabled the creation of the GMM cluster 7 with the most different sessions while maintaining the remaining profiles similar to those discovered in the K-means clustering.

Another noteworthy point concerns idle times. Specifically, cluster 4 exhibits an average idle time that surpasses the average charging time, meaning that the EVs spend more time parked without charging than actually charging. This indicates a high flexibility potential. Such flexibility characterization can provide significant insight to DSOs and CPOs seeking innovative approaches to integrate EVs into the power system (refer to Section 5.5 for further details). Table A.2 reveals the precise values of the Silhouette, Davies-Bouldin, and Calinski-Harabasz scores, according to the type of covariance and k.

5.2.1.D Agglomerative Hierarchical Clustering

According to the *scikit-learn* website [64], the Agglomerative Hierarchical clustering method allows the choice of the distance (linkage) measure, namely between Ward's method, complete-link, average-link, and single-link measures (remember Section 4.4.3). Consequently, in addition to the number of clusters, it was also necessary to understand which distance measure yields the best profiles and scores.

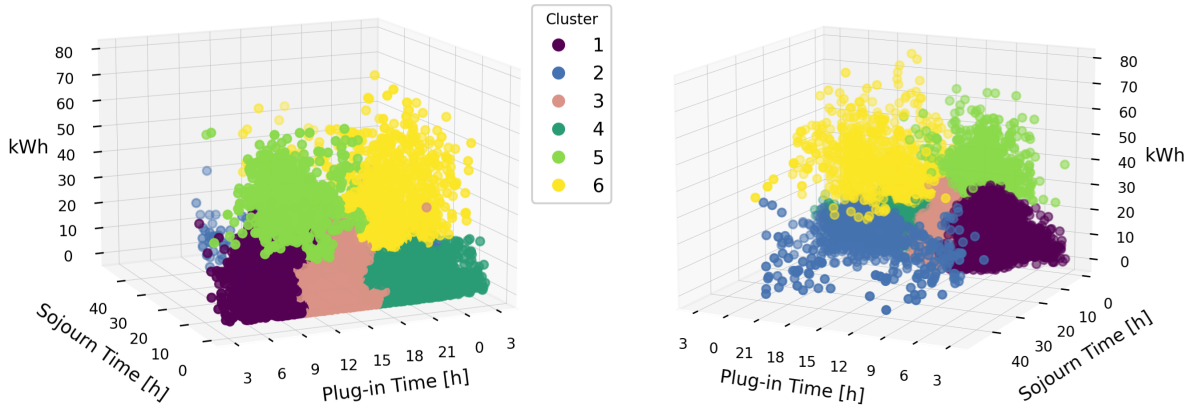
The results revealed that no metric performs consistently well across all scores. For instance, average-link produces higher Silhouette scores, whereas single-link obtains the lowest Davies-Bouldin scores. Concerning Calinski-Harabasz, Ward's method is the best. Thus, a more comprehensive analysis was necessary to determine which measure achieves a suitable balance between meaningful profiles and good scores. Complete-link, average-link, and single-link generate high scores since they tend to assign most sessions to one or two clusters while leaving the remaining clusters with fewer sessions. Conclusively, Ward's method undoubtedly achieved the best balance. Figure 5.10 illustrates the plots of the different scores as a function of the number of clusters, from which it is evident that k=6 yields the highest Silhouette and the lowest Davies-Bouldin scores, besides k=2. The Calinski-Harabasz scores did not assist since they display hyperbolic behavior (Figure 5.10(c)). Similarly, the *dendrogram* did not specify an optimal number of clusters other than k=2. Therefore, k=6 was selected.



(a) Silhouette Coefficient. **(b)** Davies-Bouldin Index. **(c)** Calinski-Harabasz Index.

Figure 5.10: Different scores as a function of k for the ACN-Data Hierarchical clustering, with Ward's method as distance measure.

Figure 5.11 presents the distribution of the adjusted EV charging profiles regarding the Plug-in Time, Sojourn Time, and kWh (energy delivered) fields, from which is clear that the profiles exhibit more overlap and less distinctness compared to the K-means (Section 5.2.1.B) or GMM clustering (Section 5.2.1.C) results. Specifically, cluster 3 is more irregular, with several points overlapping the clusters around it. However, the grouping of high and low-energy sessions into different clusters is also present, although with less visual separation than in the previously established profiles. Regarding the most different sessions, with higher sojourn time, they are distributed across the clusters instead of isolated as found in GMM clustering. Nevertheless, most of these sessions fall within cluster 2, which is characterized by plug-in times in the late afternoon or early evening, and low energy.



(a) Azimuth = -115°. **(b)** Azimuth = 115°.

Figure 5.11: 3D distribution of the adjusted Hierarchical EV Charging profiles for the ACN-Data dataset, with Ward's method as distance measure.

Table 5.5 lists the mean quantitative characteristics of the six profiles. From the results in Figure 5.11 and Table 5.5, one can see that cluster 6 includes multiple sessions during the late afternoon that end on the same day or the next with high energy delivered, yielding the same problem as K-means cluster 5 (recall Figure 5.7). Additionally, cluster 1 comprises most of the sessions that in GMM clustering were grouped into clusters 1 and 4 (remember Figure 5.9), resulting in a more generic and less distinctive cluster. Given that there are only six clusters, this merging of profiles was expected. A higher k would lead to the emergence of new meaningless clusters in terms of EV charging profiles, so k=6 is effectively the best number of clusters for this method, which proved inadequate in identifying meaningful profiles. K-means and GMM methods yielded superior outcomes and are more probable to be employed in real-world applications (refer to Section 5.5 for further details).

Table A.3 reveals the precise values of the Silhouette, Davies-Bouldin, and Calinski-Harabasz scores, according to the distance measure and the number of clusters.

Table 5.5: Mean quantitative characteristics of the Hierarchical EV Charging profiles for the ACN-Data dataset.

Cluster ID	No. of Sessions	Plug-in Time	Plug-out Time	Energy [kWh]	Sojourn Time	Charging Time	Idle Time	Profile*
1	13033	08h56	16h15	8.965	7h 20min	3h 40min	3h 40min	Morning to afternoon high-term, low energy
2	1178	19h20	10h51	11.893	15h 31min	6h 13min	9h 18min	Evening to next morning, medium energy
3	9888	12h53	16h24	5.847	3h 31min	1h 59min	1h 32min	Afternoon medium-term stay, low energy
4	5571	19h04	20h58	7.343	1h 54min	1h 32min	22min	Evening short-term stay, low energy
5	822	10h01	16h06	37.854	6h 05min	5h 19min	46min	Morning to afternoon high-term, high energy
6	826	19h24	03h57	35.356	8h 33min	4h 54min	3h 40min	Evening to next morning, high energy

*Note: "Low energy": below 10 kWh; "Medium energy": between 10 kWh and 30 kWh; "High energy": over 30 kWh. "Short-term": sojourn time below 2h; "Medium-term": between 2h and 4h; "Long-term": over 4h.

5.2.2 GR-Data dataset

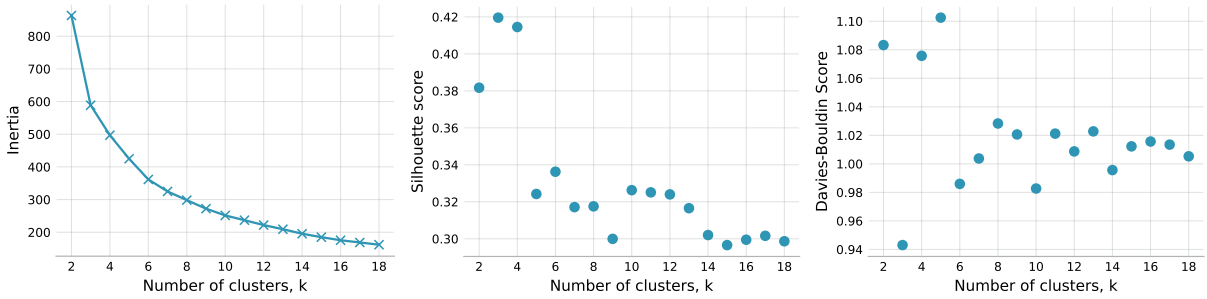
5.2.2.A Chosen fields and normalization of the data

The analysis described in Section 5.2.1.A for the ACN-Data dataset was also performed with the private dataset GR-Data, yielding highly similar results. Consequently, the chosen fields were *Start_datetime*, *sojournTime*, and *kWhDelivered*, allowing a comparable analysis between the profiles found in both datasets. The remaining features were removed, and the data were normalized using the MinMaxScaler method [64] to obtain the best possible outcomes, presented in more detail in the following sections.

5.2.2.B K-means Clustering

The number of clusters, k, was chosen based on the elbow method and the values obtained for the Silhouette, Davies-Bouldin, and Calinski-Harabasz scores, also considering the resulting profiles.

Figure 5.12 illustrates the plots of the different scores as a function of k. The elbow method does not effectively display an elbow, making it insufficient for determining the ideal k. Nevertheless, the knee of the curve suggests k from 5 to 8. Within this range, by performing a more in-depth analysis, the best results are thus found for k=6, supported by the scores in Figures 5.12(b) and 5.12(c). However, both plots indicate a turning point at k=10, with interesting scores compared with the remaining k's. Selecting k=6 yields quite generic profiles that comprise relatively different behaviors within the same cluster. With k=10, on the other hand, the clusters are better defined and identifiable. Table A.6 reveals the precise values of the Silhouette, Davies-Bouldin, and Calinski-Harabasz scores, according to k.

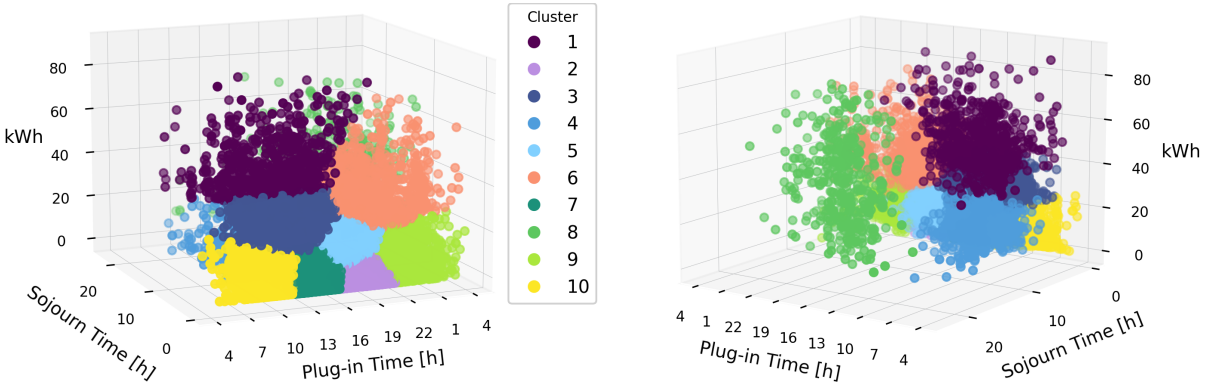


(a) Elbow Method. **(b)** Silhouette Coefficient. **(c)** Davies-Bouldin Index.

Figure 5.12: Different scores as a function of k for the GR-Data K-means clustering.

Figure 5.13 presents the distribution of the adjusted EV charging profiles regarding the Plug-in Time, Sojourn Time, and kWh (energy delivered) fields, from which one sees that the results are relatively similar to those obtained for the ACN-Data dataset (Section 5.2.1.B). There is, however, greater separation between the sessions as five clusters were found with plug-in times in the morning (clusters 1, 3, 4, 7, and 10) and only three with plug-in times in the evening (clusters 6, 8 and 9). There are also two clusters

during the middle/late afternoon (clusters 2 and 5). Therefore, one can conclude that, in this dataset, the sessions during the day differ significantly from each other, translating into a higher number of daily profiles when compared to ACN-Data. Additionally, the reduced number of clusters in the evening suggests that the sessions during this period exhibit more similar behavior than those during the day.



(a) Azimuth = -115°. **(b)** Azimuth = 125°.

Figure 5.13: 3D distribution of the adjusted K-means EV Charging profiles for the GR-Data dataset.

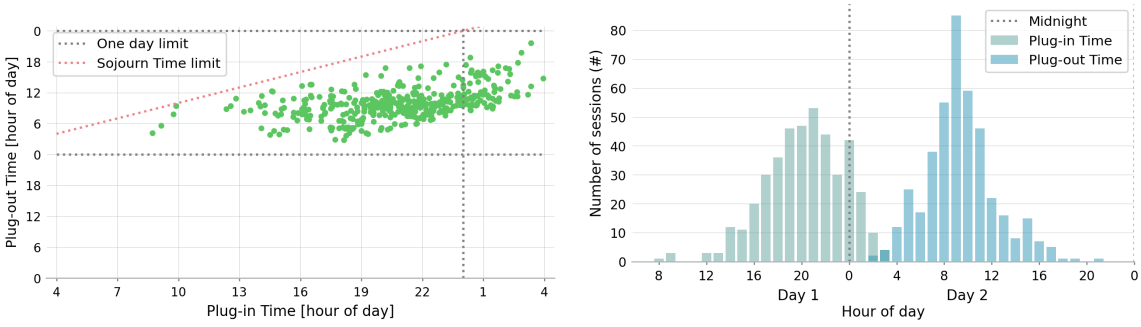
Another interesting point is that the most different sessions (higher sojourn times and, thus, higher flexibility potential) fall into distinct clusters: cluster 8 contains the sessions that only end the next day, regardless of the plug-in time, while cluster 4 comprises the sessions that start in the morning and only end in the afternoon of the same day. Table 5.6 lists the mean quantitative characteristics of the ten profiles.

Table 5.6: Mean quantitative characteristics of the K-means EV Charging profiles for the GR-Data dataset.

Cluster ID	No. of Sessions	Plug-in Time	Plug-out Time	Energy [kWh]	Sojourn Time	Charging Time	Idle Time	Profile*
1	704	11h39	16h33	49.497	4h 54min	2h 23min	31min	Morning to afternoon high energy, long-term stay
2	4297	18h25	19h07	3.957	42min	13min	29min	Early evening low energy, short-term stay
3	1500	11h49	14h20	26.807	2h 31min	1h 17min	1h 15min	Early afternoon medium energy, medium-term stay
4	1384	10h35	15h51	12.187	5h 16min	41min	4h 34min	Morning to afternoon medium energy, long-term
5	2520	17h05	19h18	14.759	2h 13min	45min	1h 28min	Afternoon to evening medium energy, medium-term
6	1154	19h58	22h43	35.992	2h 45min	1h 35min	1h 10min	Evening to night high energy, medium-term stay
7	4529	13h42	14h40	5.3141	58min	17min	41min	Early afternoon low energy, short-term stay
8	419	20h43	09h51	33.445	13h 08min	1h 54min	11h 14min	Evening to next morning medium energy, long-term
9	1888	21h45	23h06	9.788	1h 21min	29min	52min	Night low energy, short-term stay
10	3406	09h43	10h54	6.835	1h 10min	21min	49min	Morning low energy, short-term stay

*Note: “Low energy”: below 10 kWh; “Medium energy”: between 10 kWh and 30 kWh; “High energy”: over 30 kWh.
“Short-term”: sojourn time below 2h; “Medium-term”: between 2h and 4h; “Long-term”: over 4h.

According to Table 5.6, one verifies that clusters 2 and 7 are the most typical profiles, as they comprise the highest number of sessions, meaning that the short and low-energy sessions are the most frequent, and the later the drivers plug in, the more energy they consume. Morning and afternoon profiles are generally lower energy. Additionally, compared to cluster 5 of ACN-Data K-means clustering, cluster 8 of GR-Data is better defined since it effectively only contains sessions that end the next day (remember Figure 5.7). Consequently, the mean flexibility potential (idle time) of this profile is even greater, with more than eleven hours of parking stay without charging. To get a better perspective on this behavior, Figure 5.14 illustrates the distribution of the corresponding sessions.

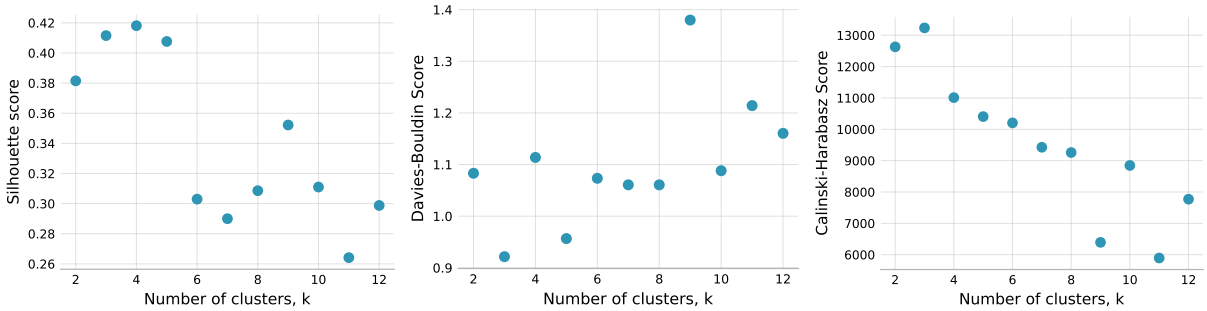


(a) Scatter plot of sessions. **(b)** Distribution of sessions.

Figure 5.14: Deep examination of K-means GR-Data cluster 8, regarding the Plug-in and Plug-out times.

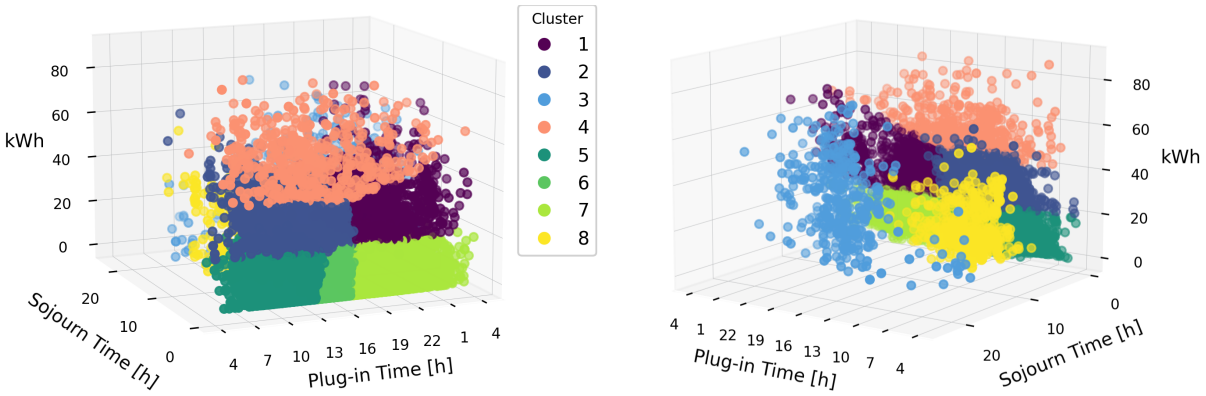
5.2.2.C GMM Clustering

For the GR-Data dataset, it was also found that the *tied covariance* originates meaningful profiles and the best Silhouette, Davies-Bouldin, and Calinski-Harabasz scores, regardless of the k value. However, contrary to the K-means study (Section 5.2.2.B), the obtained scores do not indicate a clear choice for the optimal k (Figure 5.15). Still, the best options are k={6,8,10}. Further analyses revealed that k=8 is the best choice, yielding a good balance between scores and meaningfulness. Figure 5.16 presents the distribution of the adjusted EV charging profiles regarding the Plug-in Time, Sojourn Time, and kWh.



(a) Silhouette Coefficient. **(b)** Davies-Bouldin Index. **(c)** Calinski-Harabasz Index.

Figure 5.15: Different scores as a function of k for the GR-Data GMM clustering, considering tied covariance.



(a) Azimuth = -115°.

(b) Azimuth = 125°.

Figure 5.16: 3D distribution of the adjusted GMM EV Charging profiles for the GR-Data dataset, considering tied covariance.

Figure 5.16 reveals a clear division concerning the energy delivered: profiles up to 20 kWh (clusters 5, 6, and 7), up to 40 kWh (clusters 1 and 2), and finally above 40 kWh (cluster 4). In fact, contrary to K-means, GMM groups all the highest energy sessions in cluster 4, without differentiating the plug-in time. Although fewer in number, the clusters differentiate the sessions with higher sojourn time, namely clusters 3 and 8. Cluster 3 contains the sessions with the highest sojourn times, most of which do not finish until the following day. However, it also includes some sessions with a plug-in time of around 07h00 that end on the same day, unlike K-means cluster 8 (remember Figure 5.13). Table 5.7 lists the quantitative mean characteristics of the eight profiles, demonstrating that profile 4, which has the highest energy delivered and relatively fast charging, contains few sessions and is the second least usual. Cluster 3 is the least common, corresponding to highly flexible night-time charging sessions.

Table 5.7: Mean quantitative characteristics of the GMM EV Charging profiles for the GR-Data dataset.

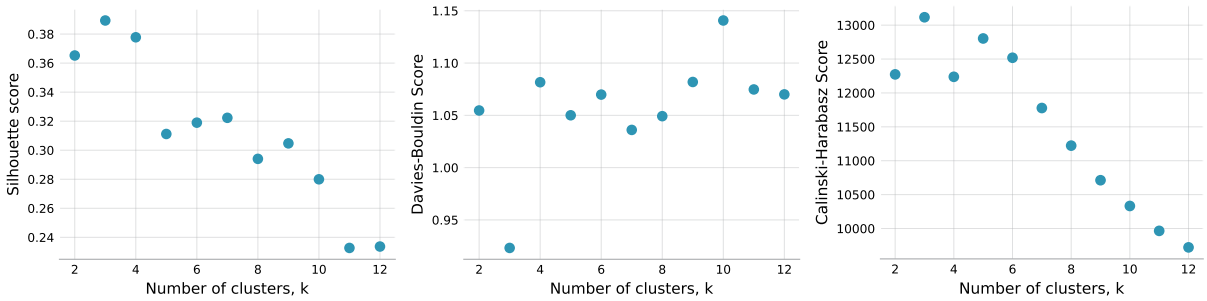
Cluster ID	No. of Sessions	Plug-in Time	Plug-out Time	Energy [kWh]	Sojourn Time	Charging Time	Idle Time	Profile*
1	1241	20h01	23h01	33.414	2h 59min	1h 35min	1h 24min	Evening to midnight high energy, medium-term stay
2	1521	11h57	14h58	31.430	3h 01min	1h 32min	1h 29min	Morning to afternoon high energy, medium-term stay
3	392	20h02	09h33	29.090	13h 32min	1h 39min	11h 52min	Evening to next morning medium energy, long-term
4	491	14h13	17h21	55.446	3h 08min	2h 04min	1h 04min	Afternoon high energy, medium-term stay
5	6093	10h38	12h06	7.771	1h 28min	25min	1h 04min	Morning low energy, short-term stay
6	3661	14h35	15h50	6.746	1h 15min	22min	53min	Afternoon low energy, short-term stay
7	7724	19h04	20h16	7.397	1h 12min	23min	48min	Evening low energy, short-term stay
8	678	10h56	18h01	15.367	7h 05min	52min	6h 13min	Morning to evening medium energy, long-term

*Note: “Low energy”: below 10 kWh; “Medium energy”: between 10 kWh and 30 kWh; “High energy”: over 30 kWh.
“Short-term”: sojourn time below 2h; “Medium-term”: between 2h and 4h; “Long-term”: over 4h.

As seen for K-means, the short duration and low energy profiles are the most frequent (clusters 5, 6, and 7). Clusters 3 and 8 offer significant flexibility potential due to their high idle times resulting from fast charging. This suggests that using such high charging power does not make sense since EV drivers tend to park longer than the car is effectively charging. Reducing the charging rate (maximum EVSE power) during those sessions would result in fewer power peaks on the grid. Table A.4 reveals the precise values of the Silhouette, Davies-Bouldin, and Calinski-Harabasz scores.

5.2.2.D Agglomerative Hierarchical Clustering

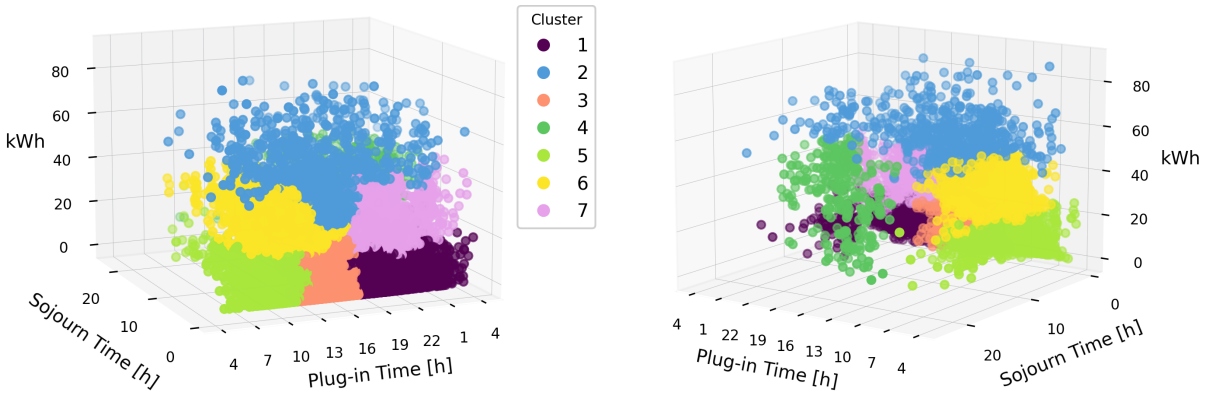
The values presented in detail in Table A.5 reveal very similar behavior to that observed in Section 5.2.1.D, in which no metric performs consistently well across all scores. Nevertheless, further analysis revealed that Ward's method also leads to meaningful profiles, proving to be the best choice among the optional linkage measures. Figure 5.17 illustrates the plots of the different scores as a function of k. According to the scores and considering the previous clustering results, it is clear that k=7 gives the highest Silhouette score (Figure 5.29(a)) and the lowest Davies-Bouldin score (Figure 5.29(b)) compared with the k's immediately below or above. The Calinski-Harabasz score (Figure 5.29(c)) did not contribute to determining the number of clusters since it displays a hyperbolic behavior for $k > 6$.



(a) Silhouette Coefficient. **(b)** Davies-Bouldin Index. **(c)** Calinski-Harabasz Index.

Figure 5.17: Different scores as a function of k for the GR-Data Hierarchical clustering, with Ward's method as distance measure.

Figure 5.18 presents the distribution of the adjusted EV charging profiles regarding the Plug-in Time, Sojourn Time, and kWh fields, from which one sees that the profiles exhibit more overlap and less clear definition than those found with the K-means (Section 5.2.2.B) or GMM clustering (Section 5.2.2.C), similar to the outcomes found for the ACN-Data Hierarchical clustering. Despite several overlapped points with neighboring clusters and the absence of GMM cluster 8, the obtained profiles visually resemble the results of the GMM clustering. As a result, clusters 5 and 6 contain sessions with differing behaviors due to the grouping of short and long morning sessions. Furthermore, cluster 1 includes some sessions that end the following day, affecting the profile characterization. Overall, the clusters are poorly defined, resulting in less accurate and reliable identification of typical profiles.



(a) Azimuth = -115°.

(b) Azimuth = 125°.

Figure 5.18: 3D distribution of the adjusted Hierarchical EV Charging profiles for the GR-Data dataset, with Ward's method as distance measure.

Table 5.8 lists the mean quantitative characteristics of the seven profiles. The results confirm that Hierarchical clustering produced significantly different results when compared with the K-means or GMM studies. The clustering method did not differentiate short-term sessions, which were grouped with sessions of longer duration, as seen in clusters 5 and 6, for example. Profile 4, typical of nighttime charging that only ends the next day, includes a reduced number of sessions, with a large part of these next-morning sessions incorporated in clusters 1 and 2. Nevertheless, the average characteristic values of each profile are still relevant. Increasing the number of clusters to solve these shortcomings is not viable since a higher k results in the emergence of new meaningless clusters in terms of EV charging profiles. Thus, k=7 is effectively the best number of clusters for this method, which proved to be the weakest method for identifying meaningful profiles. The K-means and GMM methods obtained superior results, in line with the analysis of the ACN-Data dataset.

Table 5.8: Mean quantitative characteristics of the Hierarchical EV Charging profiles for the GR-Data dataset.

Cluster ID	No. of Sessions	Plug-in Time	Plug-out Time	Energy [kWh]	Sojourn Time	Charging Time	Idle Time	Profile*
1	6515	19h25	20h38	6.750	1h 13min	21min	51min	Evening low energy, short-term stay
2	913	15h09	19h15	50.102	4h 06min	2h 11min	1h 55min	Afternoon to evening high energy, long-term stay
3	6172	14h20	15h39	6.711	1h 19min	21min	58min	Afternoon low energy, short-term stay
4	298	20h14	9h48	33.137	13h 34min	1h 53min	11h 41min	Evening to next-morning high energy, long-term stay
5	4787	09h57	12h00	8.007	2h 03min	26min	1h 37min	Morning low energy, medium-term stay
6	1604	11h18	14h45	28.635	3h 26min	1h 26min	2h	Morning to afternoon medium energy, medium-term
7	1512	19h56	22h15	27.923	2h 20min	1h 18min	1h 02min	Evening to midnight medium energy, medium-term

*Note: "Low energy": below 10 kWh; "Medium energy": between 10 kWh and 30 kWh; "High energy": over 30 kWh.
"Short-term": sojourn time below 2h; "Medium-term": between 2h and 4h; "Long-term": over 4h.

5.2.3 Summary of Results

The K-means and GMM methods delivered consistent and effective results for identifying meaningful EV charging profiles with practical applications. The K-means method produced the highest overall scores, performing better in GR-Data. GMM yielded more specific and extreme profiles (which can be particularly interesting for analyzing the most different sessions), achieving superior results in ACN-Data. On the other hand, Hierarchical clustering produced more generic, overlapped, and less visually defined clusters (better scores for fewer clusters). Nevertheless, it achieved the best values in some scores, but these did not translate into better typical profiles.

The ACN-Data's charging sessions tend to last longer, starting in the morning and ending in the evening. This allows for greater flexibility since EVs spend more time parked without charging than actually being charged. On the other hand, the GR-Data dataset is characterized by shorter charging sessions with less energy supplied and, therefore, less flexibility potential. Table 5.9 summarizes the metrics and parameters selected for each clustering method applied to both datasets.

Table 5.9: Summary of the selected metrics for each ACN-Data and GR-Data clustering method.

ACN-Data	**K-means**	**GMM**	**Hierarchical**
Best number of clusters	8	8	6
Parameters	-	Tied Covariance	Ward's Method
Elbow Method	k={5, 6, 7, 8}	-	-
Silhouette Coefficient	0.329	0.313	0.325
Davies-Bouldin Index	1.006	1.007	1.097
Calinski-Harabasz Index	17561.08	15226.62	15496.63
GR-Data	**K-means**	**GMM**	**Hierarchical**
Best number of clusters	10	8	7
Parameters	-	Tied Covariance	Ward's Method
Elbow Method	k={7, 8, 9, 10}	-	-
Silhouette Coefficient	0.326	0.309	0.322
Davies-Bouldin Index	0.983	1.061	1.036
Calinski-Harabasz Index	10715.45	9259.41	11777.57

5.3 EV User Behavior profiles

In the context of this thesis, EV user behavior profiles differ from EV charging profiles, as previously discussed in Section 3.3. Now, it is intended to understand the typical user behavior for charging an EV.

Due to the similarity in data preprocessing, Section 5.3.1 presents the steps performed for the two datasets under analysis. Then, the obtained results are detailed for each dataset separately.

5.3.1 ACN-Data & GR-Data: Chosen fields and normalization of the data

The first step consisted of creating new datasets from the previous preprocessed and clean ones, focusing solely on sessions with a *userID* (remember Tables 5.1 and 5.2). Based on the findings from Section 3.2, the sessions were grouped by user, replacing all individual driver sessions with a single theoretical charging session composed by the *mean* of the plug-in times, *mean* of the sojourn times, *standard deviation* of plug-in times, and *standard deviation* of sojourn times.

Then, a threshold was defined to eliminate users with less than three recorded sessions, which were random and unsuitable for finding behavior patterns. As a result, the ACN-Data dataset dropped from 571 to 338 users, and the GR-Data dataset from 3184 to 1228 users. The purpose is to represent each EV user with a single theoretical charging session consisting solely of mean values.

In addition to these fields, a new feature must be associated with the users to differentiate regular EV drivers from occasional ones: the *frequency* field, obtained through (4.5). A summary of the usable fields from the new user behavior datasets is presented in Table 5.10.

Table 5.10: Summary of the usable fields in the ACN-Data and GR-Data user behavior datasets.

Field name	ACN-Data Non-Null count	GR-Data Non-Null count	Dtype
mean Plug-in Time	338	1228	float64
mean standard deviation (Std) of Plug-in Time	338	1228	float64
mean sojournTime	338	1228	float64
mean standard deviation (Std) of sojournTime	338	1228	float64
frequency	338	1228	float64

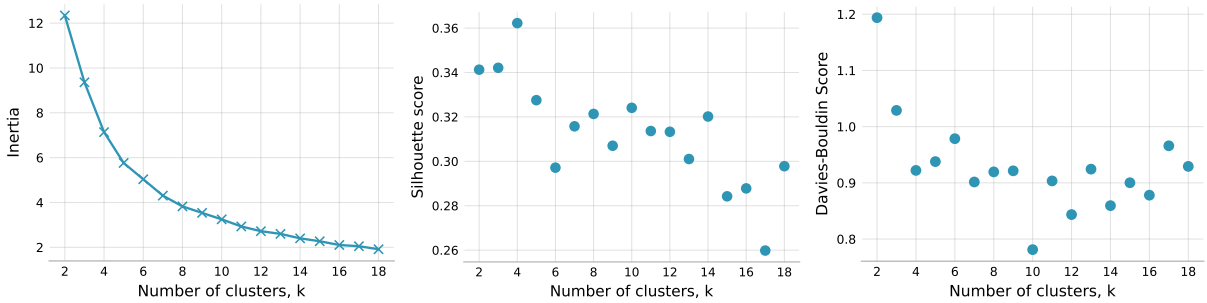
The following stage involved selecting fields for clustering, a comparable but more time-consuming process than the selection for EV charging profiles, mainly due to the subjectivity of *user behavior* in the literature. Nevertheless, in this thesis, the fields *std of plug-in time*, *std of sojourn time*, and *frequency* were chosen, as this triplet yielded the most interpretable profiles among the available fields, allowing deep insight into the typical behavior of EV users. The remaining fields were eliminated, and the data were normalized to obtain the best possible results, detailed next.

5.3.2 ACN-Data dataset

5.3.2.A K-means Clustering

Figure 5.19(a) represents the inertia's value as a function of the number of clusters. The elbow method does not effectively display an elbow, making it insufficient for determining the ideal k (similar to the previous studies). Nevertheless, the knee of the curve suggests k from 4 to 6. Within this range, by performing a more in-depth analysis, the best results are thus found for k=4, with higher Silhouette and lower Davies-Bouldin scores, as seen in Figures 5.19(b) and 5.19(c), respectively. Also, the Calinski-Harabasz score results in a high value for k=4.

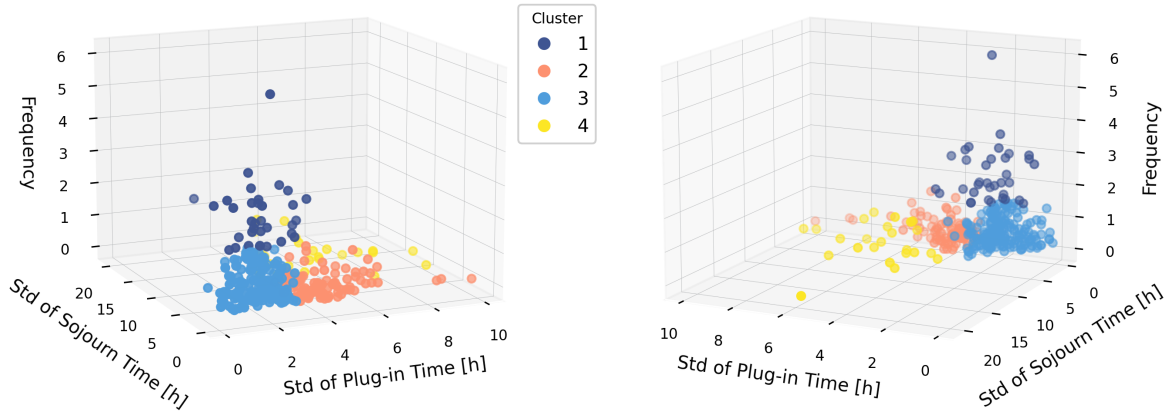
Selecting k=5 or k=6 produces profiles that are perhaps too closely fitted to the data (overfit problem) since the number of users per cluster reduces. Therefore, k=4 is the best option. Table B.1 reveals the precise values of the Silhouette, Davies-Bouldin, and Calinski-Harabasz scores, according to k.



(a) Elbow Method. **(b)** Silhouette Coefficient. **(c)** Davies-Bouldin Index.

Figure 5.19: Different scores as a function of k for the ACN-Data user behavior K-means clustering.

Figure 5.20 presents the distribution of the EV user behavior profiles regarding the standard deviation of the Plug-in Time, the standard deviation of the Sojourn Time, and the Frequency fields, from which one sees that the profiles are well-defined and have minimal overlap. An interesting result is the separation of high-frequency users (cluster 1), which are virtually divided by the plane defined by Frequency $\approx$ 1.5 from the low and medium frequency profiles. Additionally, most users exhibit relatively low standard deviations of plug-in and sojourn times. However, there are also users with high standard deviations, mainly present in cluster 4, indicating that their charging behavior is indeed random and lacks routine.



(a) Azimuth = -115°. **(b)** Azimuth = 115°.

Figure 5.20: 3D distribution of the K-means EV User Behavior profiles for the ACN-Data dataset.

Table 5.11 lists the quantitative characteristics of the four profiles. Cluster 3 comprises most users, characterized by routine behavior without high deviations. Therefore, the typical Caltech EV users start charging in the morning and only end in the late afternoon, with reduced deviations.

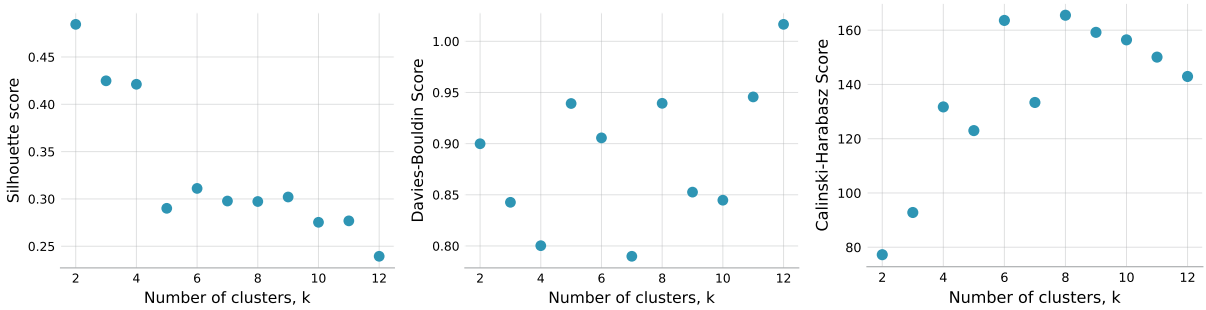
Table 5.11: Mean quantitative characteristics of the K-means EV User Behavior profiles for the ACN-Data dataset.

Cluster ID	No. of users	Plug-in Time: Mean	Plug-in Time: Std	Sojourn Time: Mean	Sojourn Time: Std	Frequency	Profile
1	34	11h11	2h 05 min	6h 23min	2h 47min	2.47	Morning to afternoon charging, with some deviation in sojourn time. Recharge more than 2 times per week
2	99	14h49	4h 14min	2h 34min	1h 36min	0.41	No specific time to recharge, with short sojourn time and low frequency
3	181	12h47	1h 50min	4h 28min	1h 48min	0.58	Morning to afternoon charging, with low deviations. Recharge more than once every 2 weeks approximately
4	24	15h16	5h 24min	11h 12min	8h 22min	0.55	Random behavior, recharge more than once every 2 weeks approximately

Clusters 1, 2, and 4 ultimately differentiate the most extreme users in each field. Another aspect worth mentioning concerns the frequency field. Clusters 2, 3, and 4 have values close to 0.5, indicating that users attend Caltech EVSEs approximately once every two weeks, while users from cluster 1 visit, on average, two times per week. However, one user stands out from the rest by utilizing the EVSEs about six times a week, visible in Figure 5.20.

5.3.2.B GMM Clustering

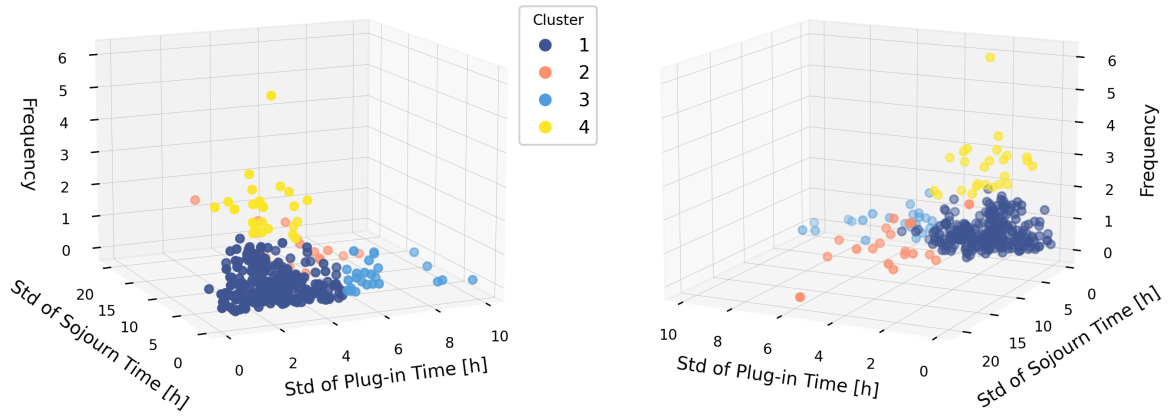
Similarly to the previous studies (Sections 5.2.1.C and 5.2.2.C), it was necessary to select the most suitable type of covariance for the data. Thus, in the first approach, it was verified that the *tied covariance* originates meaningful profiles and at the same time with the best Silhouette, Davies-Bouldin, and Calinski-Harabasz scores, regardless of the k value. Figure 5.21 illustrates the plots of the different scores as a function of the number of clusters. The optimal number of clusters should also be k=4 since it yields higher Silhouette and lower Davies-Bouldin scores than the k's immediately below or above, seen in Figures 5.21(a) and 5.21(b), respectively. Furthermore, the Calinski-Harabasz scores (Figure 5.21(c)) also follow this behavior. Table B.2 reveals the precise values of the Silhouette, Davies-Bouldin, and Calinski-Harabasz scores, according to the type of covariance and k.



(a) Silhouette Coefficient. **(b)** Davies-Bouldin Index. **(c)** Calinski-Harabasz Index.

Figure 5.21: Different scores as a function of k for the ACN-Data user behavior GMM clustering, considering tied covariance.

Figure 5.22 presents the distribution of the EV user behavior profiles regarding the standard deviation of the Plug-in Time, the standard deviation of the Sojourn Time, and the Frequency fields. It is visible that the separation between high-frequency and low-frequency users is also present here. However, unlike K-means clustering, GMM cluster 1 contains most of the low-frequency users, regardless of the standard deviation of the plug-in time. Cluster 3 includes users with random plug-in times and relatively low standard deviations of the sojourn time, while cluster 2 comprises only users with high standard deviations of plug-in and sojourn times, a profile not found in the K-means results (Section 5.3.2.A). These users would be the ones considered outliers, and GMM clustering successfully grouped them into a single cluster, similar to the results obtained in Section 5.2.1.C.



(a) Azimuth = -115°. **(b)** Azimuth = 115°.

Figure 5.22: 3D distribution of the GMM EV User Behavior profiles for the ACN-Data dataset, considering tied covariance.

Table 5.12 lists the quantitative characteristics of the four profiles, from which one sees that cluster 1 effectively comprises most lower frequency users, approximately 80% of the total drivers, resulting in a profile with a higher standard deviation of plug-in time. The remaining clusters have a smaller number of users, resulting in higher average frequencies and more fitted EV user behavior profiles.

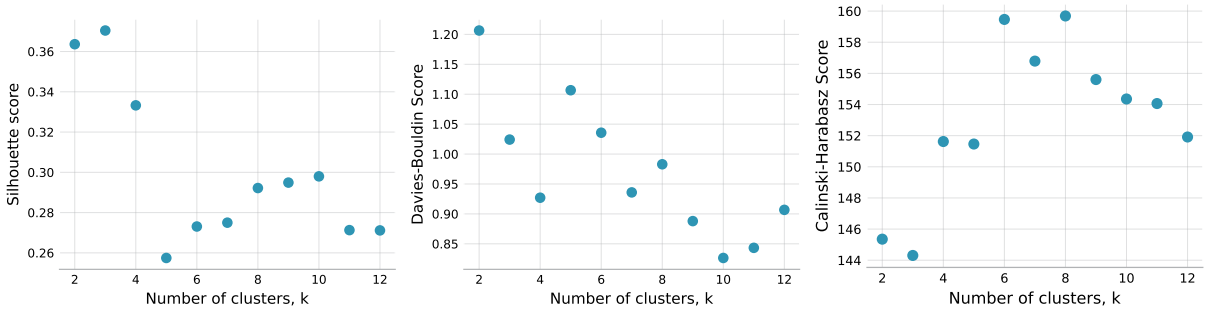
Table 5.12: Mean quantitative characteristics of the GMM EV User Behavior profiles for the ACN-Data dataset.

Cluster ID	No. of users	Plug-in Time: Mean	Plug-in Time: Std	Sojourn Time: Mean	Sojourn Time: Std	Frequency	Profile
1	269	13h19	2h 25min	4h 04min	1h 50min	0.56	Morning to afternoon charging, with low deviations. Recharge more than once every 2 weeks approximately
2	19	15h28	4h 38min	12h 22min	9h 37min	0.64	Random behavior, recharge more than once every 2 weeks approximately
3	25	15h08	6h 20min	2h 14min	1h 30min	0.52	No specific time to recharge, with short sojourn time and low frequency
4	25	10h57	2h 15min	6h 24min	2h 38min	2.72	Morning to afternoon charging, with some deviation in sojourn time. Recharge more than 2 times per week

5.3.2.C Agglomerative Hierarchical Clustering

As previously discussed, it is necessary to choose the most suitable distance (linkage) measure for the data. The values detailed in Table B.3 reveal very similar behavior to that observed in Section 5.2.1.D, in which no metric performs consistently well across all scores. However, further analysis revealed that Ward's method also leads to meaningful profiles, proving to be the best choice among the options.

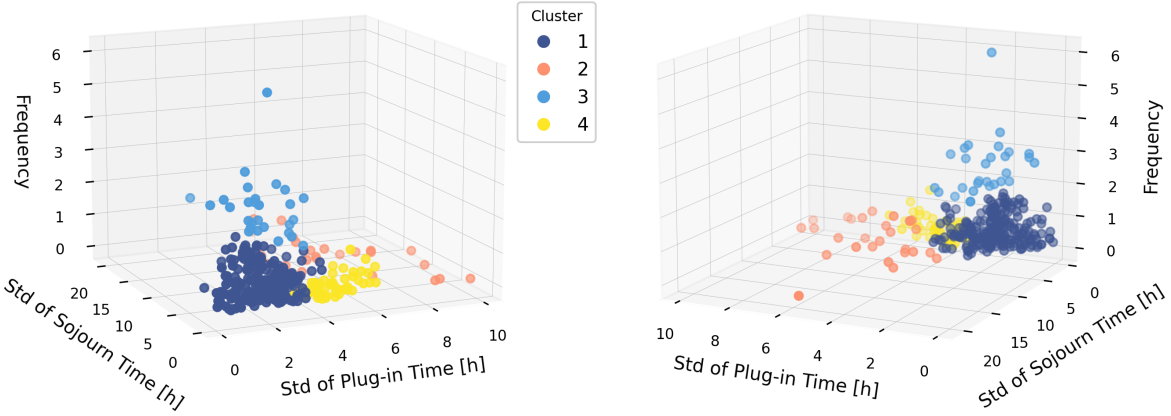
Figure 5.23 illustrates the plots of the different scores as a function of the number of clusters. According to the scores, one can see that k=4 gives the highest Silhouette score (Figure 5.23(a)) and the lowest Davies-Bouldin score (Figure 5.23(b)), in the range of k from 4 to 6. Also, the Calinski-Harabasz score (Figure 5.23(c)) agrees with this value for the number of clusters. Thus, k=4 was chosen.



(a) Silhouette Coefficient. **(b)** Davies-Bouldin Index. **(c)** Calinski-Harabasz Index.

Figure 5.23: Different scores as a function of k for the ACN-Data user behavior Hierarchical clustering, with Ward's method as distance measure.

Figure 5.24 presents the distribution of the EV user behavior profiles regarding the standard deviation of the Plug-in Time, the standard deviation of the Sojourn Time, and the Frequency fields, from which one verifies that the profiles are more overlapped and less precisely defined than in other clustering methods.



(a) Azimuth = -115°. **(b)** Azimuth = 115°.

Figure 5.24: 3D distribution of the Hierarchical EV User Behavior profiles for the ACN-Data dataset, with Ward's method as distance measure.

In fact, some low-frequency users are included in cluster 1 due to the poor cluster definition. Cluster 2 contains the most distant users from the remaining clusters (with higher standard deviations), allowing cluster 4 to be better defined. Table 5.5 lists the quantitative characteristics of the four profiles, indicating that the findings align with those validated by the previous studies with minimal cluster discrepancies, despite the overlapped clusters. Perhaps the most notable difference is precisely cluster 4, which includes the users with the lowest mean and standard deviation of sojourn time, an interesting result not achieved by K-means or GMM clustering.

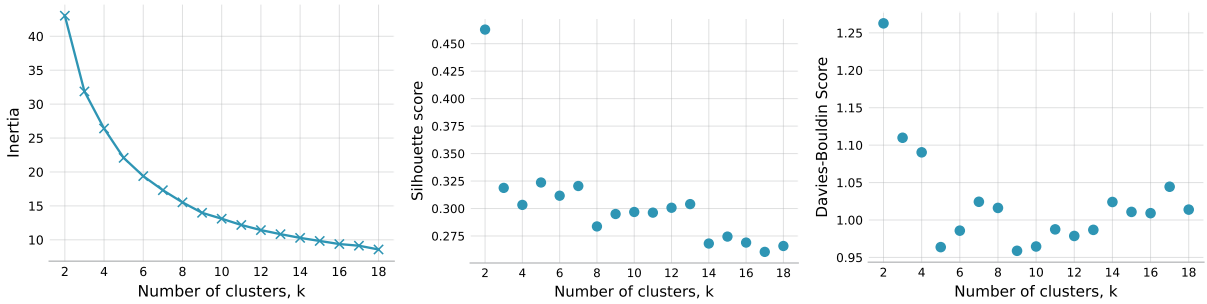
Table 5.13: Mean quantitative characteristics of the Hierarchical EV User Behavior profiles for the ACN-Data dataset.

Cluster ID	No. of users	Plug-in Time: Mean	Plug-in Time: Std	Sojourn Time: Mean	Sojourn Time: Std	Frequency	Profile
1	225	12h55	2h 06min	4h 25min	1h 56min	0.60	Morning to afternoon charging, with low deviations. Recharge more than once every 2 weeks approximately
2	27	15h35	5h 48min	10h 03min	7h 29min	0.51	Random behavior, recharge more than once every 2 weeks approximately
3	28	11h35	2h 16min	6h 15min	2h 50min	2.64	Morning to afternoon charging, with some deviation in sojourn time. Recharge more than 2 times per week
4	58	15h04	4h 29min	1h 48min	1h 05min	0.36	No specific time to recharge, with short sojourn time and low frequency

5.3.3 GR-Data dataset

5.3.3.A K-means Clustering

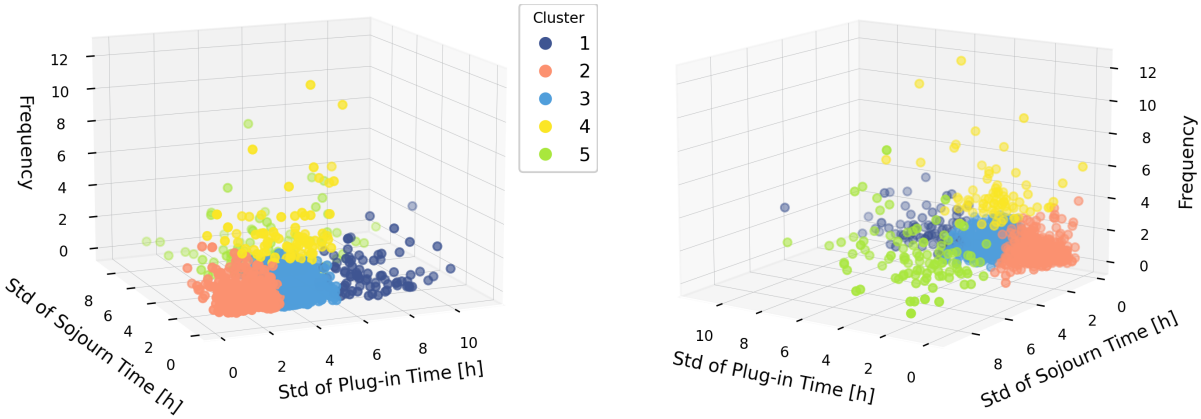
Figure 5.25(a) represents the inertia's value as a function of the number of clusters. The elbow method does not effectively display an elbow, but the knee of the curve suggests k from 4 to 7. Within this range, by performing a more in-depth analysis, the best results are thus found for k=5, with higher Silhouette and lower Davies-Bouldin scores compared to the k's immediately before and after, as seen in Figures 5.25(b) and 5.25(c), respectively. Also, the Calinski-Harabasz score results in a high value for k=5.



(a) Elbow Method. **(b)** Silhouette Coefficient. **(c)** Davies-Bouldin Index.

Figure 5.25: Different scores as a function of k for the GR-Data user behavior K-means clustering.

Figure 5.26 presents the distribution of the EV user behavior profiles regarding the standard deviation of the Plug-in Time, the standard deviation of the Sojourn Time, and the Frequency fields. Comparing the results to those obtained for the ACN-Data (Section 5.3.2.A), there is an increase in users that led to a growth in the number of clusters with distinct standard deviations of plug-in time (clusters 1-3). One profile with the most frequent users (cluster 4) remains, while cluster 5 consists of the users with the highest standard deviations of sojourn time (which do not follow a charging routine).



(a) Azimuth = -115°. **(b)** Azimuth = 125°.

Figure 5.26: 3D distribution of the K-means EV User Behavior profiles for the GR-Data dataset.

Table 5.14 lists the quantitative characteristics of the five profiles, from which one can verify that the mean value of the plug-in time does not change over the different clusters, all of them lying at the 14h00 mark. This means that K-means did not find a correlation between the plug-in time's standard deviation and the plug-in time's mean value. The Greek users do not demonstrate a specific routine of only charging in the morning or the evening, for example. Nonetheless, cluster 2 contains the most routine users who recharge at lunchtime, with minor deviations regarding the plug-in time and sojourn time fields.

Table 5.14: Mean quantitative characteristics of the K-means EV User Behavior profiles for the GR-Data dataset.

Cluster ID	No. of users	Plug-in Time: Mean	Plug-in Time: Std	Sojourn Time: Mean	Sojourn Time: Std	No. of EVSEs	Frequency	Profile
1	83	14h16	7h 06min	1h 31min	1h 05min	4	1.145	No specific time to recharge, with low deviation of short sojourn time, recharge once per week approximately
2	388	14h52	1h 45min	1h 32min	51min	3	0.755	Lunchtime charging, with low deviation of short sojourn time. Recharge at specific EVSEs once every week and a half
3	581	15h00	3h 43min	1h 33min	1h 01min	5	0.837	Morning or afternoon charging of short sojourn time. Recharge more than once every week and a half
4	80	14h37	3h 36min	1h 48min	1h 08min	7	4.074	Morning or afternoon charging, medium sojourn time. Recharge at different EVSEs more than four times per week
5	96	14h41	4h 34min	5h 03min	4h 49min	4	1.305	No specific time to recharge, with high deviation of long sojourn time. Recharge more than once per week

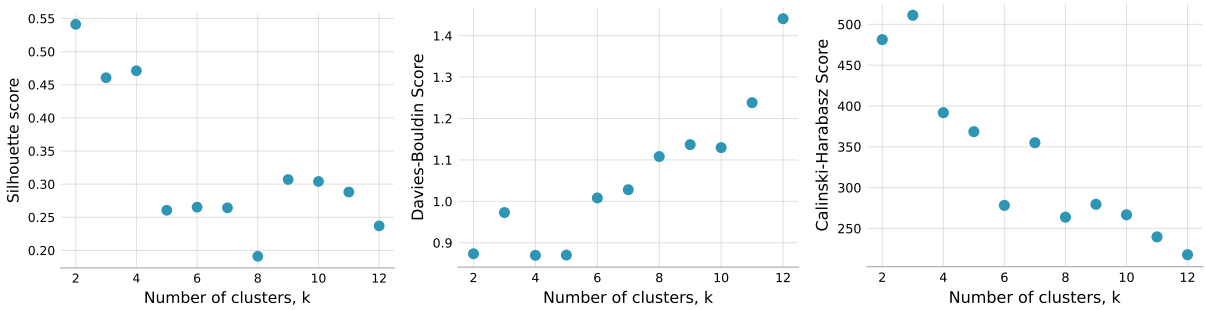
One noteworthy aspect is that the charging frequency in GR-Data user profiles is higher than in the profiles obtained for ACN-Data (recall Table 5.11). Further examination reveals that GR-Data user profiles demonstrate an overall higher frequency but a lower sojourn time, while ACN-Data user profiles demonstrate precisely the opposite behavior: low frequency but long sojourn times. This is justified by the location of the EVSEs and the consequent user behavior since the GR-Data's EVSEs are either situated along highways, gas stations, or in quick-stay areas like supermarkets. On the other hand, the ACN-Data's EVSEs are located in a garage, thus allowing for longer sojourn times.

Additionally, the GR-Data dataset contains EVSEs across Greece, which enables the introduction of a new feature to characterize the profiles: the number of different EVSEs attended by the users. The analysis of this information reveals that routine EV drivers (cluster 2) typically attend the lowest number of EVSEs (three, on average), while the most frequent users (cluster 4) rely on more EVSEs (seven, on average). One possible explanation for this trend is that these users travel extensively throughout the country and require recharging from distinct locations. Table B.6 reveals the precise values of the Silhouette, Davies-Bouldin, and Calinski-Harabasz scores, according to the number of clusters.

5.3.3.B GMM Clustering

Similarly to the previous analyses, GMM clustering delivered meaningful profiles and most often the best Silhouette, Davies-Bouldin, and Calinski-Harabasz scores with the *tied covariance* parameter. Figure 5.27 illustrates the plots of the different scores as a function of the number of clusters.

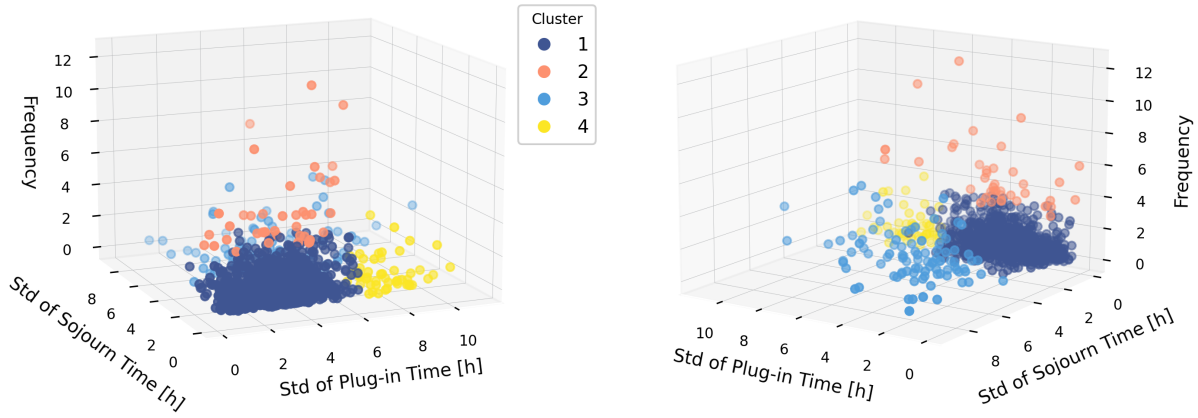
According to the scores and considering the K-means study (Section 5.3.3.A), the optimal number of clusters should be k=4 since it produces high Silhouette and low Davies-Bouldin scores, with the Calinski-Harabasz score (Figure 5.27(c)) also following this behavior. Table B.4 reveals the precise values of the Silhouette, Davies-Bouldin, and Calinski-Harabasz scores, according to the type of covariance and k. Figure 5.28 presents the distribution of the EV user behavior profiles regarding the standard deviation of the Plug-in Time, the standard deviation of the Sojourn Time, and the Frequency fields.



(a) Silhouette Coefficient. **(b)** Davies-Bouldin Index. **(c)** Calinski-Harabasz Index.

Figure 5.27: Different scores as a function of k for the GR-Data user behavior GMM clustering, considering tied covariance.

Based on Figure 5.28, it is evident that the GMM clustering demonstrates similar results on both the GR-Data and ACN-Data datasets (remember Section 5.3.2.B). Cluster 1 in both datasets comprises predominantly low-frequency users, regardless of the standard deviation of the plug-in time. The remaining clusters consist of extreme users in each field: cluster 2 comprises the highest charging frequency users, cluster 3 contains users with the highest standard deviation of sojourn time, and cluster 4 includes users with the highest standard deviation of plug-in time. These clusters explain the observed high scores but do not provide meaningful information regarding user behavior profiles.



(a) Azimuth = -115°. **(b)** Azimuth = 125°.

Figure 5.28: 3D distribution of the GMM EV User Behavior profiles for the GR-Data dataset, considering tied covariance.

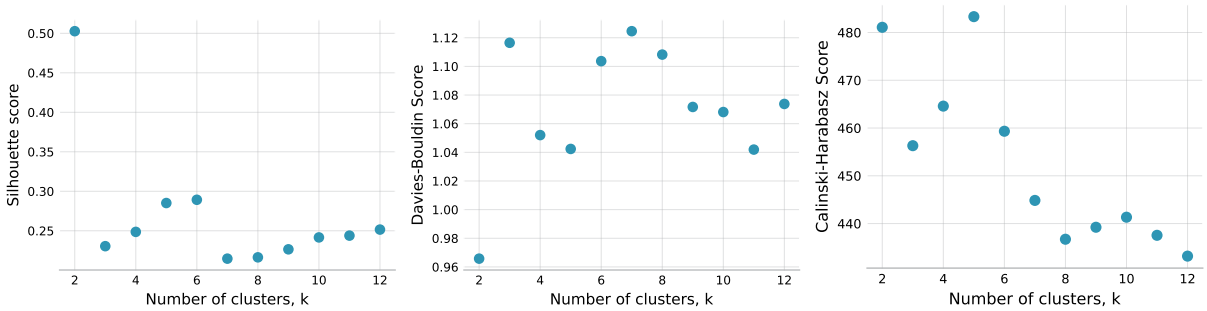
Table 5.15 lists the quantitative characteristics of the four profiles, revealing that cluster 1 comprises most low-frequency users, approximately 85% of the total users. This profile exhibits a higher standard deviation in plug-in time since it includes the users present in clusters 2 and 3 of K-means (recall Figure 5.26). Interestingly, GMM cluster 3 and K-means cluster 5 exhibit very similar mean characteristic values (remember Table 5.14). If k=5, the new cluster would consist of about 50 low-frequency users with extremely low standard deviations, i.e., another extreme profile that leads to worse overall scores.

Table 5.15: Mean quantitative characteristics of the GMM EV User Behavior profiles for the GR-Data dataset.

Cluster ID	No. of users	Plug-in Time: Mean	Plug-in Time: Std	Sojourn Time: Mean	Sojourn Time: Std	No. of EVSEs	Frequency	Profile
1	1041	14h58	3h 03min	1h 34min	58min	4	0.905	Morning or afternoon charging of short sojourn time. Recharge less than once per week
2	41	13h44	3h 22min	1h 44min	1h 15min	6	5.213	Morning or afternoon charging, medium sojourn time. Recharge at different EVSEs more than five times per week
3	96	14h40	4h 41min	5h 03min	4h 48min	4	1.242	No specific time to recharge, with high deviation of long sojourn time. Recharge more than once per week
4	50	13h53	7h 46min	1h 19min	57min	4	1.020	No specific time to recharge, with low deviation of short sojourn time, recharge once per week approximately

5.3.3.C Agglomerative Hierarchical Clustering

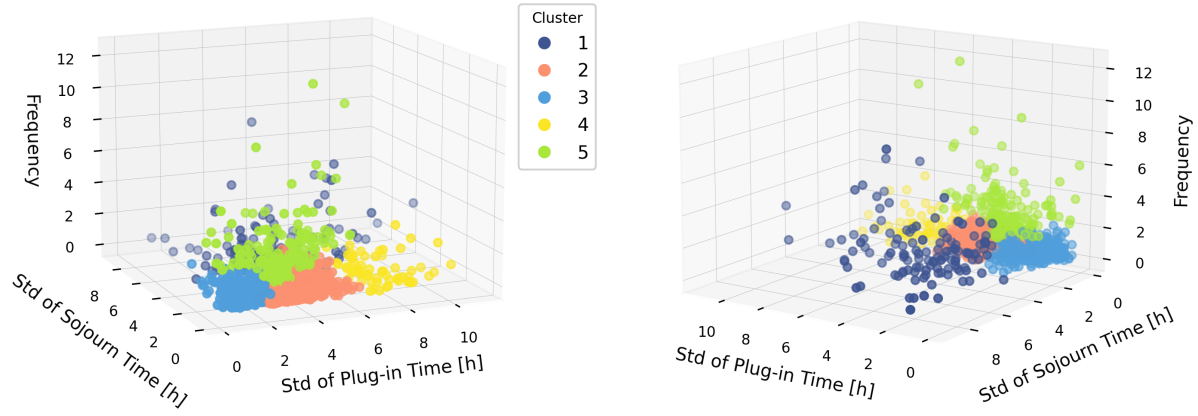
The resulting scores detailed in Table B.5 demonstrate that no distance measure consistently performs best across all scores. However, further analysis revealed that Ward's method leads to the best balance between meaningful profiles and relevant scores, similar to previous studies. Figure 5.29 illustrates the plots of the different scores as a function of the number of clusters. Based on the scores, k=5 yields one of the highest Silhouette scores (Figure 5.29(a)) and one of the lowest Davies-Bouldin scores (Figure 5.29(b)), while achieving the highest Calinski-Harabasz score (Figure 5.29(c)). This value for the number of clusters aligns with the previous studies, thus, k=5 was chosen.



(a) Silhouette Coefficient. **(b)** Davies-Bouldin Index. **(c)** Calinski-Harabasz Index.

Figure 5.29: Different scores as a function of k for the GR-Data Hierarchical clustering, with Ward's method as distance measure.

Figure 5.30 presents the distribution of the EV user behavior profiles regarding the standard deviation of the Plug-in Time, the standard deviation of the Sojourn Time, and the Frequency fields. The results are consistent with prior Hierarchical clustering studies: more overlapped and less precise profiles. For instance, the most frequent users are grouped with lower frequency users in cluster 5, influencing the remaining clusters.



(a) Azimuth = -115°. **(b)** Azimuth = 125°.

Figure 5.30: 3D distribution of the Hierarchical EV User Behavior profiles for the GR-Data dataset, with Ward's method as distance measure.

An intriguing observation relates to cluster 3. Unlike the K-means or GMM methods, Hierarchical clustering successfully grouped the most regular users (i.e., those with the lowest standard deviation of plug-in time and standard deviation of sojourn time) into one single cluster. This outcome holds great potential for practical applications (refer to Section 5.5). The remaining clusters, 1 and 4, resemble those obtained using the previous methods without noticeable dissimilarities.

Table 5.8 lists the quantitative characteristics of the five profiles. One can confirm that the results match those verified with the K-means clustering (Section 5.3.3.A), despite the slight differences in the characteristic values due to the higher number of users in cluster 5 (high-frequency profile) and the consequent loss of definition in the remaining clusters. Specifically, the profiles maintain a high standard deviation of plug-in time and a low standard deviation in sojourn time, except for cluster 1, which is representative of random users.

Table 5.16: Mean quantitative characteristics of the Hierarchical EV User Behavior profiles for the GR-Data dataset.

Cluster ID	No. of users	Plug-in Time: Mean	Plug-in Time: Std	Sojourn Time: Mean	Sojourn Time: Std	No. of EVSEs	Frequency	Profile
1	126	14h09	4h 31min	4h 43min	4h 21min	4	1.403	No specific time to recharge, with high deviation of long sojourn time. Recharge more than once per week
2	562	15h09	3h 43min	1h 22min	52min	5	0.679	Morning or afternoon charging of short sojourn time. Recharge once every 2 weeks approximately
3	312	14h52	1h 41min	1h 44min	1h	3	0.645	Lunchtime charging, with low deviation of short sojourn time. Recharge at specific EVSEs once every 2 weeks approx.
4	61	14h10	7h 24min	1h 15min	54min	4	1.148	No specific time to recharge, with low deviation of short sojourn time, recharge once per week approximately
5	167	14h37	3h 06min	1h 35min	56min	6	2.972	Morning or afternoon charging, medium sojourn time. Recharge at different EVSEs about 3 times per week

5.3.4 Summary of Results

The study on EV charging profiles concluded that GMM can identify the most extreme points. However, in the context of EV user behavior, this feature proved inadequate, as the method consistently grouped most users into a single cluster to identify the most extreme profiles despite achieving the best scores. Nonetheless, these profiles can be valuable for specific applications. K-means was the most consensual method, achieving the best results in terms of meaningful profiles, followed by Hierarchical clustering, whose clusters were similar but less defined for both datasets.

ACN-Data drivers exhibit predictable behaviors. Most users start charging in the morning and conclude in the evening. The Caltech EV drivers prefer longer charging sessions, typically every two weeks. On the other hand, Greek drivers demonstrate a preference for faster, lower energy, and more frequent charging sessions. The users lack a charging routine as they charge their EVs at varying times throughout the day, approximately once per week, and at different EVSEs.

Table 5.17 summarizes the metrics and parameters selected for each clustering method applied to the ACN-Data and GR-Data user behavior datasets.

Based on the analysis of Tables 5.9 and 5.17, it can be concluded that the Tied covariance (for GMM) and Ward's method (for Hierarchical clustering) were consistently the most effective options for achieving the best balance between meaningful profiles and relevant scores, across all studies and despite the varying characteristics of the datasets.

Table 5.17: Summary of the selected metrics for each ACN-Data and GR-Data user behavior clustering method.

ACN-Data	**K-means**	**GMM**	**Hierarchical**
Best number of clusters	4	4	4
Parameters	-	Tied Covariance	Ward's Method
Elbow Method	k={4, 5, 6}	-	-
Silhouette Coefficient	0.362	0.421	0.333
Davies-Bouldin Index	0.922	0.800	0.927
Calinski-Harabasz Index	175.08	131.72	151.62
GR-Data	**K-means**	**GMM**	**Hierarchical**
Best number of clusters	5	4	5
Parameters	-	Tied Covariance	Ward's Method
Elbow Method	k={4, 5, 6, 7}	-	-
Silhouette Coefficient	0.324	0.471	0.285
Davies-Bouldin Index	0.964	0.870	1.042
Calinski-Harabasz Index	558.37	391.75	483.32

5.4 EVSE Accessibility

One of the barriers noted for massifying EVs is the scarcity of Charging Pools (CPs), especially publicly available infrastructures for those who cannot recharge in apartments. This problem is not felt as much by families living in private homes, where they can install an EVSE. However, these are usually low-powered and take a long time to recharge the EVs. The lack of charging infrastructure is a major barrier to EV adoption [19]. Thus, it becomes imperative to build a public EVSE network to accommodate the increasing demands of EV drivers and enable travel without range anxiety.

This study precisely aims to analyze the geographical distribution and accessibility of EVSEs and understand whether the current supply is in line with the demand or whether there are inequalities that prevent the widespread use of EVs. The GR-Data dataset was chosen to conduct an in-depth study on this topic as it provides the locations of the publicly-operated Greek CPs, spread across Greece. Since no address field is present in the ACN-DATA dataset, no study of this kind can be conducted. Additionally, even if that information were available, analyzing distribution and location would be useless since all EVSEs are placed in a parking garage at Caltech University.

5.4.1 GR-Data dataset

The GR-Data dataset provides details on the location of the EVSEs, namely information on address, zip code, city (and country). However, the geographic coordinates are needed to analyze the spatial distribution of the EVSEs, i.e., in the format (*latitude*, *longitude*). Therefore, it was necessary to modify the preprocessed dataset (recall Table 5.2) by creating a new column with the location of the EVSEs in the format "*address, zip code, city, country*" for each session (row of the dataset). By removing the remaining fields and the duplicated entries, a new dataset was created with 124 unique CPs.

Then, the **OpenCageGeocode API** [81] was employed to obtain the geographic coordinates of each address. It is a service that provides geocoding and geosearch options through open data sources, returning the corresponding geographic coordinates accurately by inputting an address. Different geocoding techniques were evaluated, including Nominatim from the Python library GeoPy [82] and the Google Maps API [83]. The results revealed the poor performance of the previous tools in locating coordinates based on addresses, frequently producing null or undefined outputs. In contrast, OpenCageGeocode consistently yielded non-null outcomes. Nevertheless, a deeper analysis revealed that the API occasionally failed to retrieve the coordinates of the indicated street, giving the city center or region as output, resulting in some overlapped locations. Consequently, these outputs had to be manually modified to reflect the most accurate coordinates, obtained through individual web searches.

DBSCAN is particularly interesting for studying the accessibility and location of EVSEs as it allows finding irregularly shaped clusters. Specifically, it can find data points that do not fit into any group, labeling them for *cluster -1*. A small number of clusters indicates a uniform distribution over the territory of Greece, ideally equal to one for a perfectly uniform distribution.

5.4.2 Density-based Clustering (DBSCAN)

As previously mentioned (remember Section 4.4.4), the threshold ϵ and the $minpts$ value must be defined a priori. In the literature, the **Haversine distance metric** - a metric option in *scikit-learn*'s DBSCAN method [84] - is often employed to calculate the distance between coordinate pairs (*latitude, longitude*). However, to meet the requirements of this metric, the degree-based coordinates from the OpenCageGeocode API must be converted to radians. Consequently, the threshold ϵ must also be in radians.

Considering the studies reviewed in Section 3.2 and the number of existing EVSEs, $minpts = 1$ was defined, meaning that at least two different CPs (1 core point + 1 neighbor) are needed to form a cluster. The threshold ϵ was selected considering the size of the Greek cities. The Athens urban area, the largest in the country, spreads across 50 km from Agios Stefanos in the north to Varkiza in the south. Accordingly, ϵ was set to 5 km (converted to radians by dividing it by the Earth's radius, both in meters), meaning that a CP belongs to the neighborhood of another if it is less than 5 km away. Figure 5.31 illustrates the spatial distribution of the DBSCAN clusters, where black triangles indicate non-clustering.

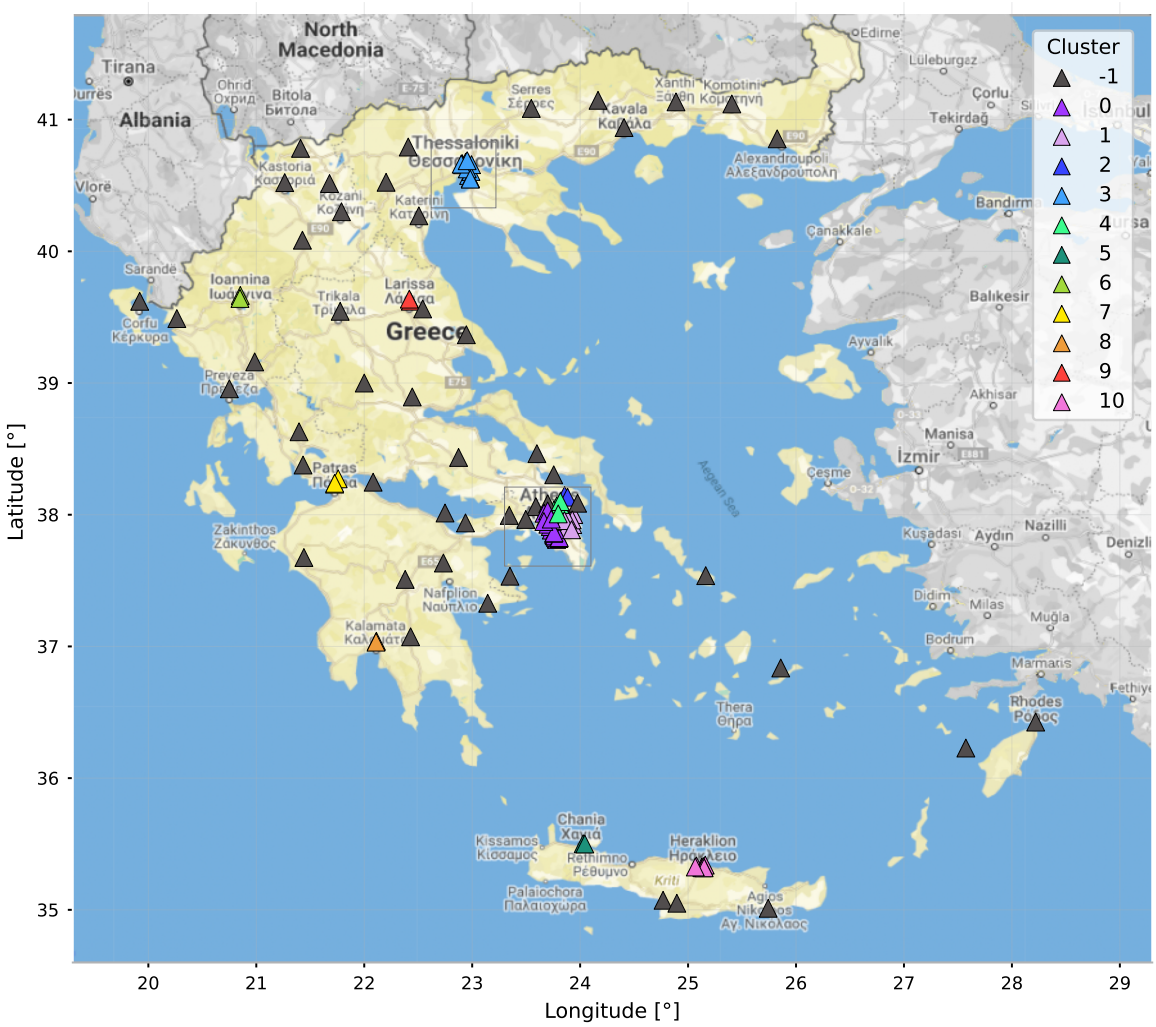
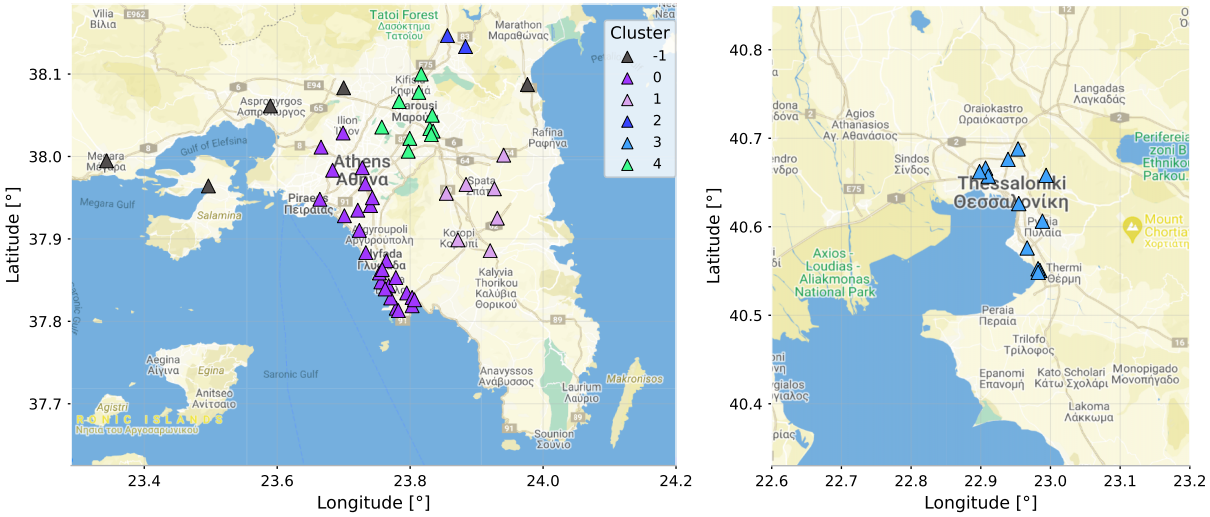


Figure 5.31: DBSCAN results on the location of CPs in Greece, from the GR-Data dataset.

Four clusters are in the Athens region, the most populous Greek urban area (about three million residents). Cluster 0, which has 26 CP locations, is the largest cluster and belongs to this area. Cluster 3 is the second-largest cluster, with 12 CP locations, found in Greece's second-largest region, Thessaloniki, with a population of over one million people. Figure 5.32 presents a closer look at these regions. The largest cluster outside Athens and Thessaloniki is cluster 10, which contains 7 CPs in Heraklion, on the Crete (or *Kríti*) island. The remaining comprise less than 4 CPs and are concentrated in city centers (areas with higher population density [85]). Cluster -1 corresponds to 49 non-clustered sites, located on highways or in areas with a reduced EVSE network despite the significant population density.

This concentration of CPs in city centers is an excellent representation of the disparity of EVSE accessibility. Smaller towns and even suburban areas lack a sufficient charging infrastructure for the imminent rise in EV sales (recall Chapter 2). However, it is worth noting that these locations are from 2021-2022 and solely correspond to publicly-operated EVSEs. According to the Global EV Outlook 2023 [19], Greece has seen an exponential increase in the deployment of public EVSEs since 2020.

Most people living in large cities inhabit apartments that lack private EVSEs. Therefore, it is logical to prioritize the most densely populated areas in the implementation of a more complete EVSE network.



(a) Athens Urban Area. **(b)** Thessaloniki Region.

Figure 5.32: Cluster distribution in the highest density areas of CPs, from the GR-Data dataset.

Nevertheless, it is possible to conduct further analysis on the EVSE use according to the obtained results. Table 5.18 reveals the most relevant information for each cluster, including the total number of sessions, the most popular EVSE per cluster, and its utilization rate (number of sessions divided by the number of weeks with recorded sessions).

Table 5.18: Key characteristics of the DBSCAN clusters on the GR-Data dataset.

Cluster ID	Region	No. of CPs	Total no. of sessions	Most popular EVSE	Max Power [kW]	Utilization rate [sessions/week]
-1	-	49	3118	A06211502136	22	6
0	Athens	26	9356	T54HU11021001	50	21
1	Athens	7	1475	T54HU11221001	50	11
2	Athens	2	1062	2067	22	7
3	Thessaloniki	12	2649	DC-IKEA-THES	50	31
4	Athens	10	2990	2193	22	11
5	Chania	2	74	2127	22	7
6	Ioannina	2	116	1775	22	3
7	Patras	3	335	2281	22	5
8	Kalamata	2	90	2337	22	2
9	Larissa	2	172	2291	22	3
10	Heraklion	7	364	2163	22	3

Table 5.18 reveals that the utilization rate of the most popular EVSE in each cluster is often low, regardless of the size of the areas they serve. Most sessions occur in cluster 0; however, the most requested EVSE is found in cluster 3. The “DC-IKEA-THES” experiences approximately 31 charging sessions per week, displaying significantly higher demand compared to other EVSEs in the same CP and

cluster. Its popularity may derive from its convenient placement within a large multinational retailer, *IKEA*, and its maximum power capacity of 50 kW, surpassing the 22 kW maximum capacity of the remaining EVSEs in the cluster. Expansion of the EVSE network within this location represents a potentially advantageous strategy regarding demand and economics.

Although they cost more, fast chargers are attended more frequently than 22 kW EVSEs in CPs with these facilities. Users seeking the convenience of fast charging seem undeterred by its utilization price. In fact, seen as a loss leader in the past, EV charging is expected to overtake fuel pumps in the profitability race as early as 2025, according to BP's head of customers [86]. Also, Wedbush Securities' Dan Ives estimates that the Supercharger business might contribute up to six percent (20 billion USD) of Tesla's total revenues by 2030 [87]. This indicator encourages CPOs and DSOs to install more fast chargers at high-demand CPs.

Table 5.19 reveals the Top-5 Greek EVSEs in the most relevant topics: utilization, energy delivered, and profitability, corroborating the previously stated points. For instance, "DC-IKEA-THES" stands out as the most utilized EVSE, supplying the most energy and generating the highest profits. The most attended EVSEs correspond to quick-stay locations, specifically restaurants and supermarkets. In contrast, the second and third most profitable EVSEs (with a significant lead over the fourth and fifth place) are located in parking lots and offer fast charging of up to 50 kW and 120 kW, respectively.

Table 5.19: GR-Data EVSE rankings by key metrics: Utilization, Energy Delivered, and Profitability.

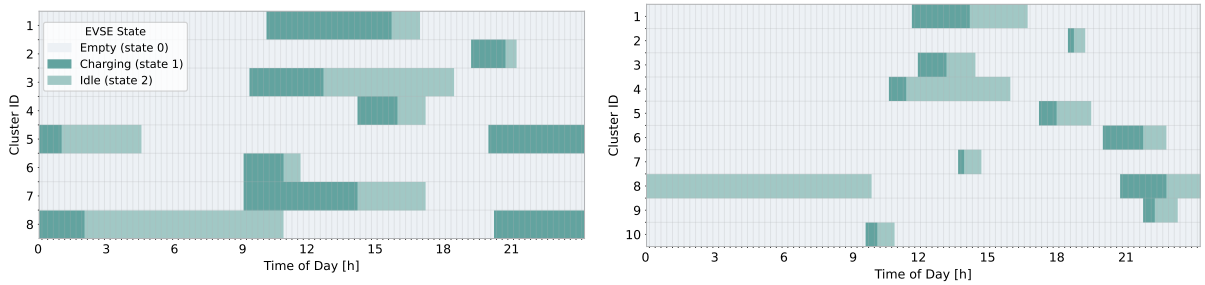
Top 5 Most Utilized EVSEs					
EVSE ID	**Max Power [kW]**	**Location Type**	**Cluster ID**	**Profit [€]**	**Total no. of sessions**
DC-IKEA-THES	50	Retail Store (IKEA)	3	2938.36	1360
2487	22	Restaurant/Urban Park	0	1045.60	736
1969	22	Supermarket	0	1318.27	531
2253	22	Supermarket	0	1252.61	466
2193	22	Supermarket	4	710.68	458
Top 5 EVSEs with Highest Energy Delivered					
EVSE ID	**Max Power [kW]**	**Location Type**	**Cluster ID**	**Profit [€]**	**Energy Delivered [kWh]**
DC-IKEA-THES	50	Retail Store (IKEA)	3	2938.36	28453.296
T54HU10321015	50	Parking Lot	3	2364.02	7907.409
2487	22	Restaurant/Urban Park	0	1045.60	7613.991
T124IT10521065	120	Parking Lot	1	2326.19	7398.797
1969	22	Supermarket	0	1318.27	6784.318
Top 5 Most Profitable EVSEs					
EVSE ID	**Max Power [kW]**	**Location Type**	**Cluster ID**	**Profit [€]**	**Profit/session [€]**
DC-IKEA-THES	50	Retail Store (IKEA)	3	2938.36	2.16
T54HU10321015	50	Parking Lot	3	2364.02	6.42
T124IT10521065	120	Parking Lot	1	2326.19	6.48
1969	22	Supermarket	0	1318.27	2.48
2253	22	Supermarket	0	1252.61	2.69

5.5 Practical Applications

The study conducted in this thesis provides various practical applications. The EV charging profiles, EV user behavior profiles, and EVSE accessibility yield valuable information for CPOs and DSOs that assists in grid management and the correct insertion of EVs into the energy system, providing powerful insights into the typical charging process, user behavior, and utilization and accessibility of the EVSEs.

In the literature, these outcomes have been exploited for various practical applications, as briefly mentioned in Section 3.2. For instance, Xiong et al. [47] discovered EV user behavior profiles and used them as input for a model that can apply to different **scheduling** EV charging algorithms. Nespoli et al. [88] also relied on clustering to identify typical charging profiles, a fundamental step in obtaining the **forecast** results. The authors focused on forecasting and reconstructing the aggregated power profile of the Caltech parking lot (ACN-Data), employing OPTICS with the *connection time* and *charging duration* fields. Similarly, Gerossier et al. [49] utilized the typical profiles obtained through clustering to forecast the consumption profile of EVs in the short-term (one day ahead) and in the long-term (2030).

Additionally, this thesis's results may lead to further applications not yet extensively explored in the literature. EV charging profiles provide valuable information about **flexibility**. The analysis of this information can be helpful in future projects, particularly in the coordination of EVs with solar and wind renewable energies to balance the network during wind curtailment or solar power gaps. Specifically, the EV4EU research project by Jerónimo et al. [89] could benefit from utilizing empirical profiles rather than relying on simulation algorithms for the generation of flexibility profiles for the ACN-Data EVSEs. The authors propose a new flexibility model for CPOs that requires the characterization of EVSE charging and occupancy rates as inputs to the model. The model is then incorporated into the network planning problem by defining a new flexibility cost function. Following the representation in [89], Figure 5.33 illustrates the temporal characterization of the typical profiles found, presented in a lattice format. In particular, this representation visually demonstrates the longer duration and greater flexibility of the ACN-Data sessions compared to the GR-Data sessions.



(a) ACN-Data K-means profiles.

(b) GR-Data K-means profiles.

Figure 5.33: Flexibility characterization of the EV Charging profiles for each dataset.

Regarding the EV user behavior profiles, one further application not yet seen in the literature is to utilize these profiles to help create **customized charging tariffs**. These tariffs can impact the charging behavior and result in advantages for the user, environment, and grid management [90]. User benefits may include adjusted tariffs that promote cost efficiency, while environmental advantages can be gained from collaboration with RES, resulting in reduced carbon emissions. Grid management can also be improved through adjusting prices to mitigate high peak loads. For instance, each EV user behavior profile can lead to an adjusted tariff based on the cluster's characteristics. Frequent users who utilize multiple EVSEs could have discounts by attending CPs with less demand, consequently increasing the availability of the most popular locations. Alternatively, regular users with longer sojourn times can receive incentives to adjust their charging power to avoid high load peaks and to help with solar gaps and wind curtailment. Additionally, this application would solve one of the main problems with the current charging tariffs: not considering the driver's needs [90].

Finally, the EVSE accessibility analysis can assist CPOs in the appropriate **EVSE placement** by providing practical details about their distribution and location. As previously discussed (remember Section 5.4.2), identifying the ideal sites for installing public EVSEs remains a significant concern in electric mobility. By analyzing information such as the total number of sessions, profit per session, and the location of the most frequently used EVSE in each cluster, CPOs can make informed decisions on where to install new public infrastructures, considering both geographic and economic factors.

6

Conclusions and Future Work

Contents

6.1 Conclusions

This thesis performed a robust assessment of the possible applications of clustering in EV-related data. In particular, EV charging profiles, EV user behavior profiles, and the EVSE's accessibility were found by applying clustering methods to datasets of empirical charging processes: ACN-Data (open data) and GR-Data (private data from one of the Greek EV4EU project partners). The experimental results demonstrated the feasibility of utilizing clustering techniques to extract comprehensive insights into the EV charging process, the behavior of EV drivers, and the accessibility of EVSEs, confirming all the objectives and research questions. The EV charging and EV user behavior profiles were obtained using K-means, GMM, and Hierarchical clustering, with subsequent comparison of methods and approaches. DBSCAN was employed to obtain information about the accessibility and distribution of Greek EVSEs.

The EV charging profiles provide information about the times of day when more or fewer charging sessions occur, whether the sessions are high energy, low energy, with high or low flexibility potential. GMM yielded more specific and superior profiles in ACN-Data, whereas K-means performed better in GR-Data. ACN-Data is characterized by highly flexible profiles since EVs spend more time parked than charging. GR-Data, on the other hand, predominantly contains quick-stay sessions, in line with the EVSE locations (publicly available infrastructures). The analysis of this information can be helpful in future activities related to power systems planning and in the coordination of EVs with RES.

EV user behavior profiles determine whether the user's behavior is routine, or random and without a typical charging frequency. K-means generated the most consensual profiles for the two datasets. Most ACN-Data users choose infrequent, long charging routines, typically every two weeks. In contrast, GR-Data includes more frequent users without a specific charging routine, as they prefer short sessions at no particular time of the day, about once a week, on different EVSEs. This information can be applied to create personalized charging tariffs that benefit the user, the environment, and the grid.

Furthermore, studying the accessibility of EVSEs revealed the geographic distribution of the corresponding publicly-operated Greek CPs, whether the current supply is in line with the demand, and whether there are inequalities in access to EVSEs that prevent the widespread use of EVs. Results confirmed that the most densely populated cities had the most extensive charging networks during the 2021-2022 data period and indicated the possibility for additional EVSEs in strategic locations.

It was necessary to perform an extensive study on the number of clusters that provided the optimal balance between best scores (Silhouette, Davies-Bouldin, and Calinski-Harabasz) and meaningful profiles to obtain the results of this thesis. The best scores often led to meaningless typical profiles, requiring a more in-depth analysis. Selecting the ideal covariance type for GMM clustering and distance measure for Hierarchical clustering was also crucial. Tied covariance and Ward's Method were consistently the most appropriate options for the different studies. Additionally, it was verified that the *seed* affects the results and their reproducibility since K-means and GMM are sensitive to the initialization

of their algorithms (the *seed* is responsible for these initializations in the *scikit-learn* library, defined through the parameter *random_state* [64]). As a result, multiple analyses were conducted to determine the optimal *random_state* for each study.

Governments throughout the world have already stated their commitment to lowering GHG emissions. EVs have become part of the solution, with record-breaking sales in 2022 and perspectives for growing market share in the upcoming years. Therefore, the results of this thesis seek to help Utilities, DSOs, and CPOs to perform a successful and intelligent integration of EVs into the energy system, providing them with valuable information about the charging behavior of EVs and EVSEs.

6.2 System Limitations and Future Work

The conducted study presents some limitations due to the available data. The EV charging profiles from ACN-Data and GR-Data exhibit similarities and differences, reflecting the data's nature since it only depicts charging patterns in a specific region/country. Moreover, obtaining a generalized result across all the analyzed studies is challenging due to the uncertainty of the data and employed methods. Hence, future work may include further clustering studies with newly available datasets from different regions/countries to increase knowledge about EVs and EVSEs.

Additionally, it is worth mentioning that the user behavior profiles found are limited since only EVSE data was utilized. Users may attend EVSEs beyond those in the current datasets, making it unclear whether the profiles found correspond to the total user behavior. For a more comprehensive analysis, it is recommended to use data from EVs or users instead of only from EVSEs.

Finally, in an era characterized by large volumes of data, smart cities, and an increasingly connected future, clustering has emerged as a powerful ally for processing and extracting valuable information for forthcoming studies. By seamlessly integrating simulated studies with real-world data enhanced through clustering (as suggested in Section 5.5), a giant leap can be made toward understanding and guiding a sustainable future that we aspire to share with everyone.

Bibliography

[1] U. Nations, “SDG Indicators,” 2017. [Online]. Available: https://unstats.un.org/sdgs/indicators/indicators-list/

[2] ——, “The Paris Agreement,” 2015. [Online]. Available: https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement

[3] E. E. Agency, “Is Europe reducing its greenhouse gas emissions?” 2022. [Online]. Available: https://www.eea.europa.eu/themes/climate/eu-greenhouse-gas-inventory

[4] ——, “Transport and environment report 2022 — European Environment Agency,” 2023. [Online]. Available: https://www.eea.europa.eu/publications/transport-and-environment-report-2022

[5] R. Lakshmi, “The Environmental Impact of Battery Production for Electric Vehicles,” *Earth.Org*, Feb. 2023. [Online]. Available: https://earth.org/environmental-impact-of-battery-production/

[6] X. Zhang, F. Gao, X. Gong, Z. Wang, and Y. Liu, “Comparison of Climate Change Impact Between Power System of Electric Vehicles and Internal Combustion Engine Vehicles,” in *Advances in Energy and Environmental Materials*, ser. Springer Proceedings in Energy, Y. Han, Ed. Singapore: Springer Singapore, 2018, pp. 739–747. [Online]. Available: http://link.springer.com/10.1007/978-981-13-0158-2_75

[7] K. Y. Yap, H. H. Chin, and J. J. Klemeš, “Solar Energy-Powered Battery Electric Vehicle charging stations: Current development and future prospect review,” *Renewable and Sustainable Energy Reviews*, vol. 169, p. 112862, Nov. 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364032122007444

[8] IRENA, *World Energy Transitions Outlook 2023: 1.5℃ Pathway; Preview*. International Renewable Energy Agency, Abu Dhabi., Mar. 2023. [Online]. Available: https://www.irena.org/Publications/2023/Mar/World-Energy-Transitions-Outlook-2023

[9] M. Kane, “Global Plug-In Electric Car Sales Increased 61% In July 2022 To 778,000,” Sep. 2022. [Online]. Available: https://insideevs.com/news/607856/global-plugin-car-sales-july2022/

[10] E. Union, " EU Action: 2050 long-term strategy," n.d. [Online]. Available: https://climate.ec.europa.eu/eu-action/climate-strategies-targets/2050-long-term-strategy_en

[11] European Council, "Fit for 55: towards more sustainable transport," Jun. 2023, publisher: General Secretariat of the European Council. [Online]. Available: https://europa.eu/!yfBkpH

[12] E. Parliament, "Fit for 55: MEPs back objective of zero emissions for cars and vans in 2035 | News | European Parliament," Aug. 2022. [Online]. Available: https://www.europarl.europa.eu/news/en/press-room/20220603IPR32129

[13] Z. Liu, Z. Deng, S. J. Davis, C. Giron, and P. Ciais, "Monitoring global carbon emissions in 2021," *Nature Reviews Earth & Environment*, vol. 3, no. 4, pp. 217–219, Mar. 2022. [Online]. Available: https://www.nature.com/articles/s43017-022-00285-w.pdf

[14] D. B. Richardson, "Electric vehicles and the electric grid: A review of modeling approaches, Impacts, and renewable energy integration," *Renewable and Sustainable Energy Reviews*, vol. 19, pp. 247–254, Mar. 2013. [Online]. Available: https://doi.org/10.1016/j.rser.2012.11.042

[15] C. B. Jones, M. Lave, W. Vining, and B. M. Garcia, "Uncontrolled Electric Vehicle Charging Impacts on Distribution Electric Power Systems with Primarily Residential, Commercial or Industrial Loads," *Energies*, vol. 14, no. 6, p. 1688, Jan. 2021. [Online]. Available: https://www.mdpi.com/1996-1073/14/6/1688

[16] R. Massey, "We test a replica of Trouve's 1881 rechargeable electric vehicle," Apr. 2021. [Online]. Available: https://www.thisismoney.co.uk/money/cars/article-9512103/We-test-replica-Gustave-Trouves-1881-rechargeable-electric

[17] M. Chandran, K. Palanisamy, D. Benson, and S. Sundaram, "A Review on Electric and Fuel Cell Vehicle Anatomy, Technology Evolution and Policy Drivers towards EVs and FCEVs Market Propagation," *The Chemical Record*, vol. 22, no. 2, Feb. 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/tcr.202100235

[18] U. D. of State, "Milestones: 1969–1976 - Office of the Historian," n.d. [Online]. Available: https://history.state.gov/milestones/1969-1976/oil-embargo

[19] I. E. A. (IEA), "Global EV Outlook 2023," IEA, Paris, Tech. Rep., 2023. [Online]. Available: https://www.iea.org/reports/global-ev-outlook-2023

[20] ——, "Stated Policies Scenario (STEPS) – Climate Model," 2023. [Online]. Available: https://www.iea.org/reports/global-energy-and-climate-model/stated-policies-scenario-steps

[21] E. Database, "Range of full EVs," 2023. [Online]. Available: https://bit.ly/45VQ28P

[22] Statista, “Worldwide electric vehicle sales by model 2022,” Feb. 2023. [Online]. Available: https://www.statista.com/statistics/960121/sales-of-all-electric-vehicles-worldwide-by-model/

[23] IEA, “Trends in charging infrastructure - Global EV Outlook 2023,” 2023. [Online]. Available: https://www.iea.org/reports/global-ev-outlook-2023/trends-in-charging-infrastructure

[24] European Commission , “European Alternative Fuels Observatory,” 2023. [Online]. Available: https://alternative-fuels-observatory.ec.europa.eu/

[25] European Parliament and Council, “Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL,” 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:52021PC0559

[26] J. A. Manzolli, J. P. Trovão, and C. H. Antunes, “A review of electric bus vehicles research topics – Methods and trends,” *Renewable and Sustainable Energy Reviews*, vol. 159, p. 112211, May 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364032122001344

[27] Drive to Zero, “The Program,” 2023. [Online]. Available: https://globaldrivetozero.org

[28] European Commission, “Recharging systems | European Alternative Fuels Observatory,” 2019. [Online]. Available: https://alternative-fuels-observatory.ec.europa.eu/general-information/recharging-systems

[29] K. Dimitriadou, N. Rigogiannis, S. Fountoukidis, F. Kotarela, A. Kyritsis, and N. Papanikolaou, “Current Trends in Electric Vehicle Charging Infrastructure; Opportunities and Challenges in Wireless Charging Integration,” *Energies*, vol. 16, no. 4, p. 2057, Feb. 2023. [Online]. Available: https://www.mdpi.com/1996-1073/16/4/2057

[30] E. H. Ruspini, “A new approach to clustering,” *Information and Control*, vol. 15, no. 1, pp. 22–32, Jul. 1969. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0019995869905919

[31] M. J. Zaki and W. Meira, Jr, *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*, 2nd ed. Cambridge University Press, 2020.

[32] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding, and C.-T. Lin, “A review of clustering techniques and developments,” *Neurocomputing*, vol. 267, pp. 664–681, Dec. 2017. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0925231217311815

[33] T. Amestoy, “Clustering basics and a demonstration in clustering infrastructure pathways,” Mar. 2022. [Online]. Available: https://waterprogramming.wordpress.com/2022/03/16/clustering-basics-and-a-demonstration-in-clustering

[34] L. M. L. Cam and J. Neyman, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability: Weather modification*. University of California, 1967, google-Books-ID: IC4Ku_7dBFUC.

[35] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977. [Online]. Available: http://www.jstor.org/stable/2984875

[36] M. Steinbach, G. Karypis, and V. Kumar, "A Comparison of Document Clustering Techniques," University of Minnesota Digital Conservancy, Report, May 2000. [Online]. Available: http://conservancy.umn.edu/handle/11299/215421

[37] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial Databases with Noise," *Knowledge Discovery and Data Mining*, pp. 226–231, Jan. 1996. [Online]. Available: https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf

[38] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "OPTICS: ordering points to identify the clustering structure," in *Proceedings of the 1999 ACM SIGMOD international conference on Management of data*. Philadelphia Pennsylvania USA: ACM, Jun. 1999, pp. 49–60. [Online]. Available: https://dl.acm.org/doi/10.1145/304182.304187

[39] H. Jia, S. Ding, X. Xu, and R. Nie, "The latest research progress on spectral clustering," *Neural Computing and Applications*, vol. 24, no. 7-8, pp. 1477–1486, Jun. 2014.

[40] Al-Ogaili, T. J. Tengku Hashim, N. A. Rahmat, A. K. Ramasamy, M. B. Marsadek, M. Faisal, and M. A. Hannan, "Review on Scheduling, Clustering, and Forecasting Strategies for Controlling Electric Vehicle Charging: Challenges and Recommendations," *IEEE Access*, vol. 7, pp. 128 353–128 371, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8825773/

[41] S. Shahriar, A. R. Al-Ali, A. H. Osman, S. Dhou, and M. Nijim, "Machine Learning Approaches for EV Charging Behavior: A Review," *IEEE Access*, vol. 8, pp. 168 980–168 993, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9194702/

[42] M. Nazari, A. Hussain, and P. Musilek, "Applications of Clustering Methods for Different Aspects of Electric Vehicles," *Electronics*, vol. 12, no. 4, p. 790, Feb. 2023. [Online]. Available: https://www.mdpi.com/2079-9292/12/4/790

[43] Y. Shen, W. Fang, F. Ye, and M. Kadoch, "EV Charging Behavior Analysis Using Hybrid Intelligence for 5G Smart Grid," *Electronics*, vol. 9, no. 1, p. 80, Jan. 2020. [Online]. Available: https://www.mdpi.com/2079-9292/9/1/80

[44] S. Shahriar and A. R. Al-Ali, “Impacts of COVID-19 on Electric Vehicle Charging Behavior: Data Analytics, Visualization, and Clustering,” *Applied System Innovation*, vol. 5, no. 1, p. 12, Jan. 2022. [Online]. Available: https://www.mdpi.com/2571-5577/5/1/12

[45] J. R. Helmus, M. H. Lees, and R. van den Hoed, “A data driven typology of electric vehicle user types and charging sessions,” *Transportation Research Part C: Emerging Technologies*, vol. 115, p. 102637, Jun. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0968090X19315414

[46] A. Märtz, U. Langenmayr, S. Ried, K. Seddig, and P. Jochem, “Charging Behavior of Electric Vehicles: Temporal Clustering Based on Real-World Data,” *Energies*, vol. 15, no. 18, p. 6575, Sep. 2022. [Online]. Available: https://www.mdpi.com/1996-1073/15/18/6575

[47] Y. Xiong, B. Wang, C.-C. Chu, and R. Gadh, “Electric Vehicle Driver Clustering using Statistical Model and Machine Learning,” in *2018 IEEE Power & Energy Society General Meeting (PESGM)*. Portland, OR: IEEE, Aug. 2018, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/8586132/

[48] G. Van Kriekinge, C. De Cauwer, N. Sapountzoglou, T. Coosemans, and M. Messagie, “Electric Vehicle Charging Sessions Generator Based on Clustered Driver Behaviors,” *World Electric Vehicle Journal*, vol. 14, no. 2, p. 37, Feb. 2023. [Online]. Available: https://www.mdpi.com/2032-6653/14/2/37

[49] A. Gerossier, R. Girard, and G. Kariniotakis, “Modeling and Forecasting Electric Vehicle Consumption Profiles,” *Energies*, vol. 12, no. 7, p. 1341, Apr. 2019. [Online]. Available: https://www.mdpi.com/1996-1073/12/7/1341

[50] G. J. Carlton and S. Sultana, “Electric vehicle charging station accessibility and land use clustering: A case study of the Chicago region,” *Journal of Urban Mobility*, vol. 2, p. 100019, Dec. 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2667091722000073

[51] B. Borlaug, F. Yang, E. Pritchard, E. Wood, and J. Gonder, “Public electric vehicle charging station utilization in the United States,” *Transportation Research Part D: Transport and Environment*, vol. 114, p. 103564, Jan. 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S136192092200390X

[52] Y. Lin, K. Zhang, Z.-J. M. Shen, B. Ye, and L. Miao, “Multistage large-scale charging station planning for electric buses considering transportation network and power grid,” *Transportation Research Part C: Emerging Technologies*, vol. 107, pp. 423–443, Oct. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0968090X18312105

[53] Q. Li, X. Li, Z. Liu, and Y. Qi, “Application of Clustering Algorithms in the Location of Electric Taxi Charging Stations,” *Sustainability*, vol. 14, no. 13, p. 7566, Jun. 2022. [Online]. Available: https://www.mdpi.com/2071-1050/14/13/7566

[54] A. K. Kalakanti and S. Rao, “Charging Station Planning for Electric Vehicles,” *Systems*, vol. 10, no. 1, p. 6, Jan. 2022. [Online]. Available: https://www.mdpi.com/2079-8954/10/1/6

[55] Y. Amara-Ouali, Y. Goude, P. Massart, J.-M. Poggi, and H. Yan, “A Review of Electric Vehicle Load Open Data and Models,” *Energies*, vol. 14, no. 8, p. 2233, Apr. 2021. [Online]. Available: https://www.mdpi.com/1996-1073/14/8/2233

[56] L. Calearo, M. Marinelli, and C. Ziras, “A review of data sources for electric vehicle integration studies,” *Renewable and Sustainable Energy Reviews*, vol. 151, p. 111518, Nov. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364032121007966

[57] P. Ferreira, “EV4EU,” Jun. 2022. [Online]. Available: https://www.inesc-id.pt/ev4eu-launches-today/

[58] Z. J. Lee, T. Li, and S. H. Low, “ACN-Data – A Public EV Charging Dataset,” 2021. [Online]. Available: https://ev.caltech.edu/dataset

[59] ——, “ACN-Data: Analysis and Applications of an Open EV Charging Dataset,” in *Proceedings of the Tenth ACM International Conference on Future Energy Systems*. Phoenix AZ USA: ACM, Jun. 2019, pp. 139–149. [Online]. Available: https://dl.acm.org/doi/10.1145/3307772.3328313

[60] A. Satre-Meloy, M. Diakonova, and P. Grünewald, “Cluster analysis and prediction of residential peak demand profiles using occupant activity data,” *Applied Energy*, vol. 260, p. 114246, Feb. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306261919319336

[61] J. W. Tukey, *Exploratory data analysis*, ser. Addison-Wesley series in behavioral science. Reading, Mass: Addison-Wesley Pub. Co, 1977.

[62] P. J. Rousseeuw and K. V. Driessen, “A Fast Algorithm for the Minimum Covariance Determinant Estimator,” *Technometrics*, vol. 41, no. 3, pp. 212–223, Aug. 1999. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/00401706.1999.10485670

[63] F. T. Liu, K. M. Ting, and Z.-H. Zhou, “Isolation Forest,” in *2008 Eighth IEEE International Conference on Data Mining*, Dec. 2008, pp. 413–422, iSSN: 2374-8486.

[64] “scikit-learn 1.3.0 documentation.” [Online]. Available: https://scikit-learn.org/stable/index.html#

[65] H. Patel, “What is Feature Engineering — Importance, Tools and Techniques for Machine Learning,” Sep. 2021. [Online]. Available: https://bit.ly/what-is-feature-engineering

[66] C. Develder, N. Sadeghianpourhamami, M. Strobbe, and N. Refa, “Quantifying flexibility in EV charging as DR potential: Analysis of two real-world data sets,” in *2016 IEEE International Conference on Smart Grid Communications (SmartGridComm)*. Sydney, Australia: IEEE, Nov. 2016, pp. 600–605. [Online]. Available: http://ieeexplore.ieee.org/document/7778827/

[67] C. Daake, M. Cammerer, and M. Hackmann, “P3 Charging Index Report 07/22 – Comparison of the fast charging capability of various electric vehicles,” P3 GROUP GMBH, Tech. Rep. 07/22, 2022. [Online]. Available: http://bit.ly/3KENHGJ

[68] T. Kodinariya and P. Makwana, “Review on determining of cluster in k-means clustering,” *International Journal of Advance Research in Computer Science and Management Studies*, vol. 1, pp. 90–95, 01 2013. [Online]. Available: https://bit.ly/44YeSU3

[69] P. H. A. Sneath and R. R. Sokal, *Numerical Taxonomy*. W H Freeman & Company, Jan. 1973.

[70] K. Florek, J. Łukaszewicz, J. Perkal, H. Steinhaus, and S. Zubrzycki, “Sur la liaison et la division des points d’un ensemble fini,” *Colloquium Mathematicum*, vol. 2, no. 3-4, pp. 282–285, 1951. [Online]. Available: http://eudml.org/doc/209969

[71] F. J. Rohlf, “12 Single-link clustering algorithms,” *Handbook of Statistics*, Jan. 1982.

[72] T. Sørenson, *A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons*, ser. Biologiske skrifter. I kommission hos E. Munksgaard, 1948. [Online]. Available: https://www.royalacademy.dk/Publications/High/295_S%C3%B8rensen,%20Thorvald.pdf

[73] R. Sokal, C. Michener, and U. of Kansas, *A Statistical Method for Evaluating Systematic Relationships*. University of Kansas, 1958. [Online]. Available: https://bit.ly/3LuRirh

[74] J. D. Ward, “Hierarchical Grouping to Optimize an Objective Function,” *Journal of the American Statistical Association*, vol. 58, no. 301, p. 236, Mar. 1963.

[75] P. B. G. Alvez, “Inference of a human brain fiber bundle atlas from high angular resolution diffusion imaging,” phdthesis, Université Paris Sud - Paris XI, Oct. 2011. [Online]. Available: https://theses.hal.science/tel-00638766

[76] G. N. Lance and W. T. Williams, “A General Theory of Classificatory Sorting Strategies: 1. Hierarchical Systems,” *The Computer Journal*, vol. 9, no. 4, pp. 373–380, Feb. 1967. [Online]. Available: https://academic.oup.com/comjnl/article-lookup/doi/10.1093/comjnl/9.4.373

[77] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, Nov. 1987. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0377042787901257

[78] D. L. Davies and D. W. Bouldin, “A Cluster Separation Measure,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979. [Online]. Available: http://ieeexplore.ieee.org/document/4766909/

[79] T. Calinski and J. Harabasz, “A dendrite method for cluster analysis,” *Communications in Statistics - Theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/03610927408827101

[80] “pandas 2.0.3 documentation,” 2023. [Online]. Available: https://pandas.pydata.org/docs/index.html

[81] “OpenCage - Geocoding and Geosearch.” [Online]. Available: https://opencagedata.com/

[82] “GeoPy 2.3.0 documentation,” 2018. [Online]. Available: https://geopy.readthedocs.io/en/stable/

[83] “Google Maps Platform.” [Online]. Available: https://developers.google.com/maps

[84] M. Amiruzzaman, R. Rahman, M. R. Islam, and R. M. Nor, “Evaluation of DBSCAN algorithm on different programming languages: An exploratory study,” in *2021 5th International Conference on Electrical Engineering and Information Communication Technology*. Dhaka, Bangladesh: IEEE, Nov. 2021, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/9667925/

[85] Imgur, “Population density map of Greece 2021,” Feb. 2021. [Online]. Available: https://imgur.com/a/8qP3kZD

[86] R. Bousso, “For BP, car chargers to overtake pumps in profitability race,” *Reuters*, Jan. 2022. [Online]. Available: https://reut.rs/46m9RG7

[87] Anubhav, “Tesla is Projected to Make Billions from its Supercharger Business,” Aug. 2023. [Online]. Available: https://www.gizmochina.com/2023/08/26/tesla-multi-billion-dollar-supercharger/

[88] A. Nespoli, E. Ogliari, and S. Leva, “User Behavior Clustering Based Method for EV Charging Forecast,” *IEEE Access*, vol. 11, pp. 6273–6283, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10014991/

[89] A. Jerónimo, P. Carvalho, C. Jesus, L. Dias, L. M. Ferreira, and H. Morais, “Modeling demand response of Charge Point Operators to consider flexibility in grid planning,” in *International Conference on Smart Energy Systems and Technologies (SEST)*. Mugla, Turkey: SEST, Sep. 2023. [Online]. Available: https://bit.ly/3RukHW5

[90] F. Daneshzand, P. J. Coker, B. Potter, and S. T. Smith, “EV smart charging: How tariff selection influences grid stress and carbon reduction,” *Applied Energy*, vol. 348, p. 121482, Oct. 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306261923008462

Appendix A - EV Charging profiles

Table A.1: Detailed results for the ACN-Data K-means clustering, according to the number of clusters.

No. of clusters	Silhouette	Davies-Bouldin	Calinski-Harabasz
2	0.47357	0.86186	28574.51
3	0.48246	0.92243	21892.78
4	0.36035	0.92501	20569.21
5	0.36134	0.98913	19588.04
6	0.31514	1.05453	18528.65
7	0.32717	1.02194	17926.27
8	**0.32897**	**1.00647**	**17561.08**
9	0.30136	1.04497	17070.46
10	0.30430	1.00262	16916.17
11	0.31370	0.92487	16804.79
12	0.31838	0.94807	16565.86

Table A.2: Detailed results for the ACN-Data GMM clustering, according to the no. of clusters and covariance type.

Covariance Type	**Full**	**Tied**	**Diagonal**	**Spherical**
No. of clusters	**Silhouette Coefficient**			
2	0.41251	0.48143	0.41536	0.43424
3	0.26984	0.48708	0.29609	0.44138
4	0.25661	0.34126	0.27398	0.32045
5	0.20343	0.35130	0.18475	0.34143
6	0.15032	0.30105	0.15702	0.30324
7	0.12251	0.31011	0.15753	0.25442
8	0.09940	**0.31271**	0.05926	0.22970
9	0.13783	0.31580	0.18210	0.25840
10	0.05792	0.29622	0.11346	0.27802
11	0.14120	0.27210	0.10384	0.28234
12	0.14202	0.24740	0.10932	0.27222
No. of clusters	**Davies-Bouldin Index**			
2	1.7801	0.84502	1.70151	1.01159
3	1.53589	0.88985	1.49403	1.40486
4	1.94109	0.94227	1.97802	1.27881
5	2.24157	0.93662	2.15110	1.12440
6	2.09343	1.09557	1.85613	1.16216
7	3.06487	1.04670	1.73417	1.25671
8	5.12387	**1.00700**	2.70676	1.28011
9	2.67068	1.04412	2.81587	1.18493
10	2.76541	1.03229	2.52624	1.04683
11	2.30663	1.06427	2.35165	1.29935
12	2.26651	1.11045	2.23481	1.15738
No. of clusters	**Calinski-Harabasz Index**			
2	4352.18	28070.54	4158.26	22183.52
3	9362.63	21155.34	10566.51	17549.73
4	6823.52	19147.85	7464.81	14927.15
5	6087.19	17479.96	5578.93	16791.32
6	5599.10	17052.13	6844.71	15248.35
7	3746.01	16480.51	6787.40	12251.94
8	4250.03	**15226.62**	4142.44	11151.71
9	3948.06	14264.09	4703.44	11726.81
10	3790.44	14036.52	5144.84	14354.09
11	6043.51	13999.08	4883.66	12652.41
12	5559.72	10790.71	4908.84	12403.60

Table A.3: Detailed results for the ACN-Data Agglomerative Hierarchical clustering, according to the no. of clusters and distance measure.

Distance measure	**Ward's method**	**Complete-link**	**Average-link**	**Single-link**
No. of clusters	**Silhouette Coefficient**			
2	0.46838	0.45748	0.71146	0.69274
3	0.28815	0.45845	0.69039	0.69039
4	0.29674	0.46232	0.55039	0.68689
5	0.31595	0.46254	0.52369	0.66590
6	**0.32528**	0.44501	0.43899	0.62566
7	0.26605	0.44176	0.43643	0.23637
8	0.23424	0.43608	0.40813	0.23278
9	0.25205	0.41690	0.40746	0.10077
10	0.25477	0.40847	0.39773	0.10088
11	0.24934	0.40846	0.34130	0.02158
12	0.24532	0.23787	0.33384	0.02137
No. of clusters	**Davies-Bouldin Index**			
2	0.89629	1.05526	0.56043	0.22550
3	1.14471	0.93233	0.29283	0.29283
4	1.19966	0.78233	0.64574	0.21800
5	1.09765	0.73445	0.72318	0.22100
6	**1.09692**	0.73987	0.77087	0.23361
7	1.16480	0.74012	0.75920	0.45237
8	1.23580	0.79069	0.73081	0.42916
9	1.15374	0.90221	0.71403	0.41300
10	1.09326	0.89850	0.71409	0.41120
11	1.06595	0.86438	0.72149	0.43642
12	1.05788	0.96564	0.75121	0.44275
No. of clusters	**Calinski-Harabasz Index**			
2	23001.49	19062.53	44.93	15.25
3	18725.57	10881.13	24.94	24.94
4	16801.9	12035.46	454.16	16.78
5	16663.67	9079.93	383.41	16.28
6	**15496.63**	7687.98	997.31	15.27
7	14835.33	6519.69	832.82	13.26
8	14395.74	5692.03	734.19	12.92
9	14016.14	5569.92	645.90	11.35
10	13799.17	5631.02	577.21	23.02
11	13614.26	5082.67	903.27	20.88
12	13533.4	6717.94	831.61	22.31

Table A.4: Detailed results for the GR-Data GMM clustering, according to the no. of clusters and covariance type.

Covariance Type	**Full**	**Tied**	**Diagonal**	**Spherical**
No. of clusters	**Silhouette Coefficient**			
2	0.26540	0.38150	0.27053	0.29178
3	0.01088	0.41154	0.11608	0.38189
4	0.02765	0.41811	0.11217	0.38925
5	0.09537	0.40764	0.11762	0.32026
6	0.02366	0.30298	0.06055	0.25645
7	0.02577	0.29001	0.03653	0.24752
8	0.02710	**0.30859**	0.03710	0.24831
9	-0.04504	0.35220	-0.01539	0.25787
10	0.01480	0.31102	0.01971	0.24783
11	-0.00800	0.26416	0.01043	0.23362
12	-0.16190	0.29880	0.04106	0.26341
No. of clusters	**Davies-Bouldin Index**			
2	1.73222	1.08312	1.68783	1.20949
3	4.13433	0.92159	2.32263	1.13754
4	4.31540	1.11376	2.41701	1.21673
5	2.36893	0.95664	2.08761	1.33165
6	3.84114	1.07336	3.25029	1.28886
7	4.21964	1.06072	1.55934	1.39125
8	4.57814	**1.06068**	1.52143	1.35632
9	3.60060	1.37983	2.86745	1.26315
10	3.21785	1.08808	3.14817	1.24080
11	3.66278	1.21414	2.81161	1.33142
12	3.38955	1.16043	2.39371	1.13697
No. of clusters	**Calinski-Harabasz Index**			
2	5807.40	12630.92	6097.47	7926.36
3	2987.32	13235.28	4627.68	11304.23
4	3141.06	11010.57	3738.46	11033.27
5	3317.73	10405.81	3927.49	9821.07
6	2808.50	10204.26	3082.10	8840.09
7	3030.12	9426.95	3734.92	8293.49
8	3255.93	**9259.41**	4168.92	7309.53
9	3097.35	6393.49	3119.55	7793.34
10	3929.59	8846.09	3920.02	7552.51
11	3139.23	5894.13	3826.79	7004.49
12	1815.72	7771.03	4063.66	7941.74

Table A.5: Detailed results for the GR-Data Agglomerative Hierarchical clustering, according to the no. of clusters and distance measure.

Distance measure	**Ward's method**	**Complete-link**	**Average-link**	**Single-link**
No. of clusters	**Silhouette Coefficient**			
2	0.36524	0.51560	0.67181	0.67181
3	0.38935	0.36598	0.53889	0.65820
4	0.37778	0.35997	0.47794	0.64444
5	0.31118	0.37035	0.46936	0.62952
6	0.31898	0.36903	0.40626	0.54340
7	**0.32231**	0.35887	0.40155	0.54374
8	0.29405	0.35085	0.29433	0.54361
9	0.30471	0.24693	0.29175	0.46728
10	0.27997	0.25433	0.29148	0.46270
11	0.23269	0.25177	0.28610	0.45317
12	0.23366	0.25167	0.27724	0.45132
No. of clusters	**Davies-Bouldin Index**			
2	1.05466	0.80714	0.24176	0.24176
3	0.92321	0.93118	0.55434	0.29659
4	1.08166	0.80693	0.72967	0.23260
5	1.05006	0.87422	0.72114	0.25642
6	1.06984	0.93020	0.81416	0.27744
7	**1.03606**	1.03833	0.78597	0.31382
8	1.04921	1.11920	0.77099	0.34381
9	1.08190	1.18052	0.77971	0.35276
10	1.14070	1.14694	0.77090	0.34205
11	1.07476	1.11700	0.77663	0.33598
12	1.07004	1.04482	0.81600	0.33639
No. of clusters	**Calinski-Harabasz Index**			
2	12273.73	293.08	14.30	14.30
3	13116.49	6823.81	911.63	24.13
4	12237.96	6504.19	727.16	16.20
5	12803.51	7428.55	548.13	36.15
6	12518.97	5970.72	912.94	30.35
7	**11777.57**	5051.53	768.13	37.85
8	11222.55	4504.13	667.22	32.57
9	10712.25	4945.95	584.48	29.20
10	10332.39	5673.12	521.24	27.51
11	9965.21	5231.59	489.75	25.87
12	9721.59	4761.15	1005.73	25.41

Table A.6: Detailed results for the GR-Data K-means clustering, according to the number of clusters.

No. of clusters	Silhouette	Davies-Bouldin	Calinski-Harabasz
2	0.38171	1.08329	12646.32
3	0.41961	0.94303	14357.92
4	0.41454	1.07574	12678.16
5	0.32422	1.10253	12045.72
6	0.33627	0.98594	12090.11
7	0.31712	1.00377	11619.31
8	0.31754	1.02830	11126.51
9	0.29995	1.02061	10931.56
10	**0.32630**	**0.98269**	**10715.45**
11	0.32511	1.02119	10370.62
12	0.32400	1.00872	10193.72

B

Appendix B - EV User Behavior profiles

Table B.1: Detailed results for the ACN-Data user behavior K-means clustering, according to the number of clusters.

No. of clusters	Silhouette	Davies-Bouldin	Calinski-Harabasz
2	0.34126	1.19387	163.95
3	0.34210	1.02889	161.08
4	**0.36223**	**0.92213**	**175.08**
5	0.32752	0.93782	182.01
6	0.29713	0.97856	175.74
7	0.31575	0.90153	180.11
8	0.32133	0.91951	179.17
9	0.30701	0.92140	172.45
10	0.32408	0.78121	169.65
11	0.31363	0.90343	172.20
12	0.31329	0.84355	170.46

Table B.2: Detailed results for the ACN-Data user behavior GMM clustering, according to the no. of clusters and covariance type.

Covariance Type	**Full**	**Tied**	**Diagonal**	**Spherical**
No. of clusters	**Silhouette Coefficient**			
2	0.27170	0.48448	0.32297	0.40483
3	0.28328	0.42482	0.28416	0.24552
4	0.26567	**0.42116**	0.20121	0.31802
5	0.24999	0.29009	0.14601	0.24710
6	0.15114	0.31111	0.11827	0.27105
7	0.14215	0.29779	0.12276	0.29974
8	0.09369	0.29729	0.11705	0.25300
9	0.12092	0.30205	0.11916	0.29226
10	0.08630	0.27532	0.12748	0.30970
11	0.12094	0.27683	0.07661	0.29759
12	0.02333	0.23940	0.10843	0.26210
No. of clusters	**Davies-Bouldin Index**			
2	1.33701	0.89989	1.26047	1.67718
3	1.20782	0.84265	1.19773	1.67693
4	1.17150	**0.80026**	1.18948	1.16293
5	1.20893	0.93921	1.68658	1.10425
6	1.71623	0.90562	1.55385	0.97053
7	1.44279	0.78990	1.57404	0.96608
8	1.31268	0.93940	1.26399	0.89929
9	1.31953	0.85260	1.31579	0.93398
10	2.11985	0.84471	2.05610	1.18140
11	1.30083	0.94565	2.12649	0.94792
12	2.18061	1.01658	1.94048	0.96927
No. of clusters	**Calinski-Harabasz Index**			
2	128.59	77.25	147.94	98.43
3	117.04	92.80	117.56	91.72
4	125.89	**131.72**	105.26	135.32
5	127.10	122.99	76.72	130.19
6	76.96	163.60	68.64	135.47
7	83.30	133.36	70.08	156.15
8	74.03	165.51	81.02	148.08
9	78.52	159.20	83.90	157.78
10	58.88	156.44	71.43	138.59
11	84.63	150.07	64.26	153.42
12	48.82	142.94	63.62	145.46

Table B.3: Detailed results for the ACN-Data user behavior Agglomerative Hierarchical clustering, according to the no. of clusters and distance measure.

Distance measure	**Ward's method**	**Complete-link**	**Average-link**	**Single-link**
No. of clusters	**Silhouette Coefficient**			
2	0.36363	0.56765	0.70028	0.70028
3	0.37043	0.55632	0.65914	0.65914
4	**0.33327**	0.53134	0.53134	0.49858
5	0.25747	0.29084	0.47729	0.47166
6	0.27309	0.31005	0.42627	0.45960
7	0.27500	0.27778	0.41796	0.40592
8	0.29220	0.27568	0.41250	0.41402
9	0.29490	0.27571	0.33238	0.41546
10	0.29800	0.26813	0.34046	0.40563
11	0.27130	0.26022	0.32207	0.39991
12	0.27116	0.25351	0.31936	0.21283
No. of clusters	**Davies-Bouldin Index**			
2	1.20647	0.78186	0.20446	0.20446
3	1.02421	0.59308	0.21058	0.21058
4	**0.92706**	0.43856	0.43856	0.28344
5	1.10668	0.73371	0.52487	0.34640
6	1.03568	0.74967	0.53776	0.35449
7	0.93593	0.72763	0.46802	0.40083
8	0.98294	0.74428	0.50886	0.40257
9	0.88796	0.69428	0.62899	0.39151
10	0.82636	0.72909	0.66090	0.35172
11	0.84329	0.74116	0.63267	0.32743
12	0.90675	0.76189	0.66013	0.37088
No. of clusters	**Calinski-Harabasz Index**			
2	145.35	70.85	17.10	17.10
3	144.31	46.34	16.82	16.82
4	**151.62**	37.28	37.28	13.25
5	151.46	74.72	40.24	22.92
6	159.46	86.43	52.20	18.72
7	156.79	111.28	44.73	17.42
8	159.69	100.10	40.00	17.84
9	155.60	91.22	77.56	16.89
10	154.36	113.04	86.34	15.03
11	154.06	112.79	79.72	13.51
12	151.91	113.41	73.75	12.72

Table B.4: Detailed results for the GR-Data user behavior GMM clustering, according to the no. of clusters and covariance type.

Covariance Type	**Full**	**Tied**	**Diagonal**	**Spherical**
No. of clusters	**Silhouette Coefficient**			
2	0.37202	0.54116	0.36811	0.46616
3	0.17636	0.46063	0.17869	0.24009
4	0.15901	**0.47112**	0.13998	0.22874
5	0.12348	0.26057	0.15226	0.27842
6	0.10752	0.26526	0.08950	0.30536
7	0.10284	0.26421	0.06540	0.30831
8	0.12045	0.19111	0.08586	0.25756
9	0.07170	0.30694	0.05274	0.26699
10	0.08036	0.30401	0.06837	0.28526
11	0.08848	0.28810	0.06939	0.28746
12	0.07741	0.23701	0.07491	0.28952
No. of clusters	**Davies-Bouldin Index**			
2	1.73441	0.87348	1.70895	1.49276
3	1.93741	0.97290	1.90969	1.53224
4	1.75839	**0.86970**	1.78246	1.72034
5	1.83537	0.87029	1.53420	1.38924
6	1.64499	1.00817	2.12001	1.05955
7	1.85225	1.02806	2.16321	1.26052
8	1.53939	1.10824	1.62902	1.37092
9	1.84682	1.13699	1.89767	1.32925
10	1.76484	1.12972	1.73592	1.04256
11	1.63437	1.23825	1.61173	0.97966
12	1.85089	1.44101	1.75280	0.96220
No. of clusters	**Calinski-Harabasz Index**			
2	327.22	481.33	336.44	442.81
3	268.84	511.16	293.02	326.22
4	268.85	**391.75**	287.66	291.52
5	261.64	368.55	348.89	409.81
6	293.32	278.24	306.80	474.68
7	263.09	355.00	235.21	418.22
8	288.87	263.82	267.90	358.64
9	240.88	279.55	216.27	343.45
10	246.67	266.70	257.45	462.04
11	246.66	239.66	256.29	457.17
12	208.04	217.81	239.03	461.97

Table B.5: Detailed results for the GR-Data user behavior Agglomerative Hierarchical clustering, according to the no. of clusters and distance measure.

Distance measure	**Ward's method**	**Complete-link**	**Average-link**	**Single-link**
No. of clusters	**Silhouette Coefficient**			
2	0.50269	0.57040	0.64589	0.67408
3	0.23050	0.48437	0.59210	0.67263
4	0.24859	0.48810	0.57743	0.60018
5	**0.28512**	0.37946	0.49113	0.59863
6	0.28924	0.37690	0.48920	0.57905
7	0.21471	0.37460	0.46024	0.56014
8	0.21636	0.37364	0.45312	0.52332
9	0.22650	0.37307	0.45110	0.47714
10	0.24158	0.31372	0.45074	0.48567
11	0.24387	0.31355	0.44711	0.48547
12	0.25151	0.31600	0.44454	0.42640
No. of clusters	**Davies-Bouldin Index**			
2	0.96578	0.60934	0.41484	0.22059
3	1.11648	0.89855	0.63462	0.28708
4	1.05203	0.82929	0.53776	0.28351
5	**1.04237**	0.84419	0.47404	0.23558
6	1.10367	0.90122	0.64101	0.23254
7	1.12458	0.90255	0.63853	0.24323
8	1.10824	0.86916	0.72124	0.25177
9	1.07168	0.77046	0.69445	0.30802
10	1.06813	0.81319	0.65022	0.34423
11	1.04192	0.81471	0.66908	0.33245
12	1.07378	0.78440	0.66551	0.33408
No. of clusters	**Calinski-Harabasz Index**			
2	481.09	186.68	81.18	13.82
3	456.30	337.88	78.82	23.01
4	464.59	260.39	55.03	18.45
5	**483.32**	249.45	42.52	13.93
6	459.33	246.86	136.77	14.05
7	444.84	289.55	199.75	15.99
8	436.72	252.84	190.83	14.83
9	439.23	224.27	168.95	14.41
10	441.33	320.89	154.12	20.53
11	437.54	298.57	140.95	20.11
12	433.18	283.78	129.15	18.86

Table B.6: Detailed results for the GR-Data user behavior K-means clustering, according to the number of clusters.

No. of clusters	Silhouette	Davies-Bouldin	Calinski-Harabasz
2	0.46285	1.26269	552.00
3	0.31877	1.10987	586.81
4	0.30330	1.09038	556.07
5	**0.32368**	**0.96377**	**558.37**
6	0.31165	0.98577	542.14
7	0.32049	1.02435	529.99
8	0.28370	1.01615	526.34
9	0.29500	0.95885	527.71
10	0.29685	0.96460	508.86
11	0.29625	0.98748	501.46
12	0.30069	0.97868	492.86