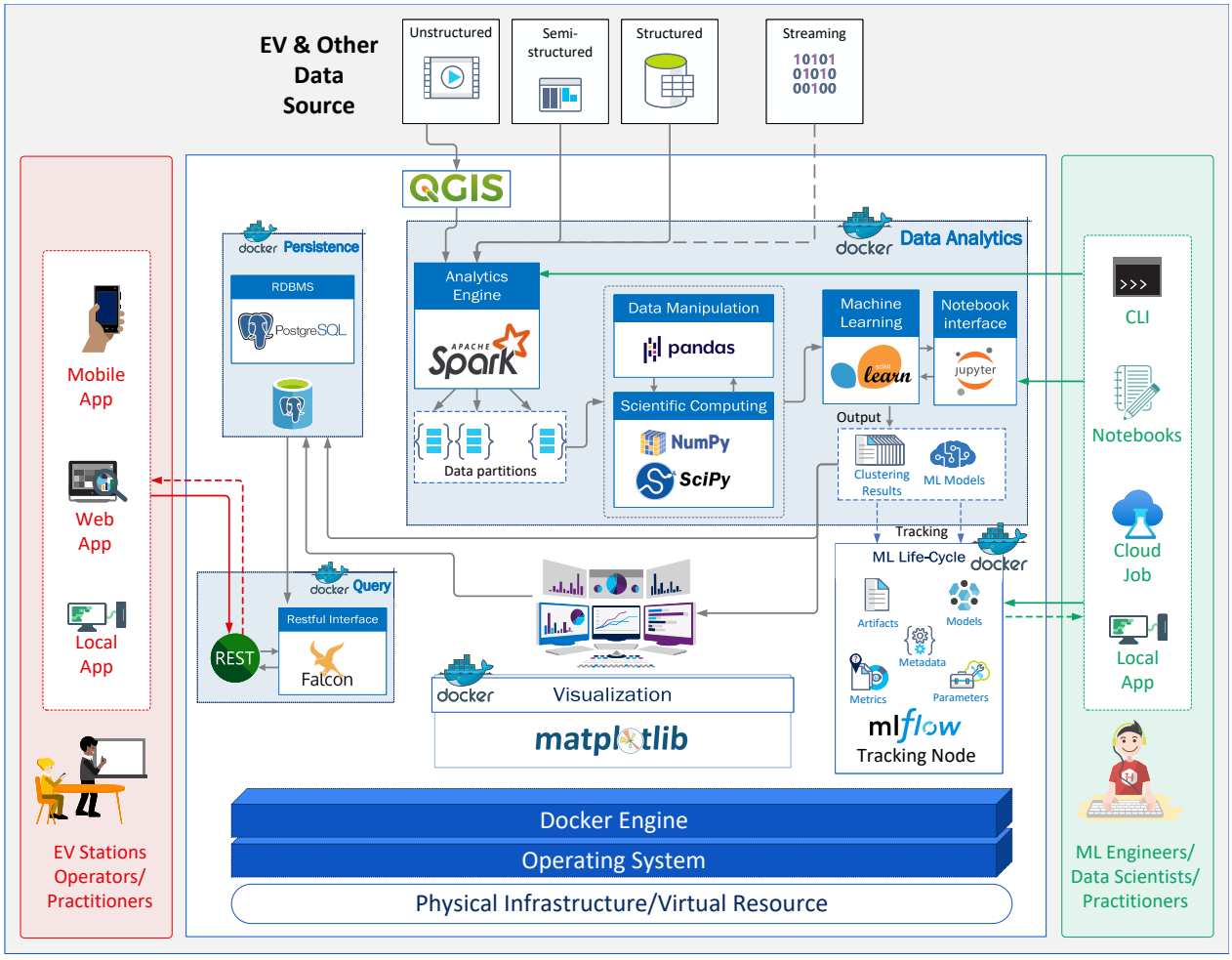
Graphical Abstract

EVStationSIM: An End-to-End Platform to Identify and Interpret Similar Clustering Patterns of EV Charging Stations Across Multiple Time Slices

René Richard, Hung Cao, Monica Wachowicz



Highlights

EVStationSIM: An End-to-End Platform to Identify and Interpret Similar Clustering Patterns of EV Charging Stations Across Multiple Time Slices

René Richard, Hung Cao, Monica Wachowicz

- A platform that facilitates the comparison of clustering results by practitioners.
- Enables the identification of similar clustering results across temporal partitions.
- Highlights utilization patterns, assisting in downstream analytical tasks.
- Leverages multiple data sources to describe EV charging station clustering results.

EVStationSIM: An End-to-End Platform to Identify and Interpret Similar Clustering Patterns of EV Charging Stations Across Multiple Time Slices

René Richard[a,c], Hung Cao[a], Monica Wachowicz[b]

[a] *University of New Brunswick, Fredericton, New Brunswick, Canada*
[b] *RMIT University, Melbourne, Australia*
[c] *Digital Technologies Research Centre, National Research Council of Canada, Fredericton, New Brunswick, Canada*

Abstract

Transport electrification introduces new opportunities in supporting sustainable mobility. Fostering Electric Vehicle (EV) adoption integrates vehicle range and infrastructure deployment concerns. An understanding of EV charging patterns is crucial for optimizing charging infrastructure placement and managing costs. Clustering EV charging events has been useful for ensuring service consistency and increasing EV adoption. However, clustering presents challenges for practitioners when first selecting the appropriate hyper-parameter combination for an algorithm and later when assessing the quality of clustering results. In a clustering process, the ground truth data is normally not available for practitioners to validate different modeling decisions. Consequently, it is difficult to judge the effectiveness of the discovered patterns because there is no objective method to compare them. This work proposes an end-to-end platform prototype named *"EVStationSIM"* that allows for the creation of relative rankings of similar clustering results. The ultimate goal is to support users/practitioners by allowing them to identify and interpret similar clustering patterns of EV charging stations using multiple time slices. The performance of this proposed platform is demonstrated with a case study using real-world EV charging event data from charging station operators in New Brunswick, Canada. The case study illustrates how generated results can assist in downstream analytical tasks such as planning infrastructure allocation expansions.

Keywords: Agglomerative Hierarchical Clustering, Usage Patterns, EV Charging Infrastructure, Traffic Counters, Geospatial Data, Clustering Process, Cluster Validity Indices

1. Introduction

The commitments of authorities around the world to electrify the transportation sector will have an impact on air quality, sustainable mobility and the management of natural resources. In 2016, New Brunswick had the 3rd highest per-capita Green House Gas (GHG) emissions in Canada and 30% of these emissions came from the transportation sector. Moreover, 24% of GHG emissions were attributable to vehicles in Canada [1]. By providing options for motorists to break from the immediate consumption of fossil fuels, opportunities

for supporting mobility from renewable resources and affecting environmental impacts will increase.

Widespread adoption of electric vehicles will require adequate public charging station infrastructure. The operation range of electric vehicles is often a major source of driver anxiety and a commonly cited barrier to widespread EV adoption [1, 2]. In light of early EV adoption concerns, increasing the availability of public charging infrastructure may assuage the hesitancy. However, EV charging infrastructure operators are reluctant to invest in new charging stations and expand charging networks as they tend to be less profitable in early adoption settings. In addition, EV uptake faces certain difficulties due to changing demographic factors in some regions. Enthusiasm around vehicle electrification is typically associated with a younger group of consumers and financial incentives. Less prosperous regions with aging populations and slow growth can struggle to provide the conditions to foster increased EV adoption.

A public EV charging network to serve a population with different lifestyles and parking habits (e.g., multi-tenant vs. single family dwellings) is needed to stimulate EV adoption [3]. The expensive capital investments required to install new public EV charging infrastructure and the use of public funds require well-informed decision making at all stages of the EV adoption life-cycle. In the context of early EV adoption, some EV charging stations may not be operating at full capacity, while others may serve a disproportionate number of users. Therefore, an understanding of charging infrastructure usage is paramount in optimizing investments and the placement of charging stations.

Grouping stations together based on similar utilization patterns, such as high versus low utilization groupings, is a useful planning tool for operators. A popular unsupervised machine learning method to assist practitioners in discovering hidden patterns from a data set is clustering. It has been utilized by users in the energy sector to group consumers with similar behaviors, predict future demand, and improve services. For instance, accurate load forecasting is one tool which can help operators ensure service consistency. Statistical machine learning models, built with data from EV charging stations having similar charging patterns (e.g. homogeneous clusters of stations) will reportedly have superior accuracy [4]. Usage behaviors and energy consumption patterns are crucial to improving services provided by utility companies, which are responsible for managing peaks and imbalances in infrastructure usage [5]. As vehicle electrification grows, so does the need for electricity and the possible strain on power grids. Utilities and other power generators can prepare for the changing electricity demand by initially deepening their understanding of energy usage behavior at the charging stations. Although there exists rich literature on different techniques and methods developed to determine EV charging station usage patterns, lacking is the use of real-world EV charging events from public infrastructure enriched with nearby traffic volumes and other sources of data to deepen our understanding of these patterns.

Clustering is currently applied in many interdisplinary and specialized knowledge domains. However, selecting an appropriate clustering algorithm with hyperparameter combination and then evaluating the performance and quality of clustering results can be challenging for practitioners. Without labeled data, the quality assessment of clustering results is highly subjective. This is likely the main reason why the state-of-the-art ML frameworks (i.e., AutoML) tend to focus on supervised learning tasks that require labeled data as input [6]. Identifying the clustering results that aligns with the needs of a practitioner is a

challenge since in many complex data sets, there is a variety of reasonable groupings, and practitioners might have multiple preferences and different priorities. [7]

This work proposes a platform prototype named "EVStationSIM" which supports identifying, exploring and comparing patterns present in EV charging data. In this work, many EV station clusterings are generated, results are made comparable using a combination of various internal cluster validity indices (CVI). These validity indices help practitioners identify a clustering result of interest, quantify different priorities and preferences and find similar clustering results in a group of results. A case study using real-world charging events from EV station operators in New Brunswick is used to evaluate the proposed process of identifying and exploring similar charging station clusters over multiple a priori determined temporal partitions in the data.

1.1. Research Objective

The main research objective in this work is, given the prospect of a clustering result of interest identified by a practitioner according to their preferences and priorities, to recommend similar clustering results over multiple, a priori selected temporal partitions in the data. The measurable research objectives can be described as follows:

- To build a system information flow which generates multiple clusterings of EV charging stations for different a priori selected temporal partitions of the same data (e.g weekly, monthly and seasonal partitions).
- To implement a clustering process which uses internal cluster validity indices to enable the identification of similar clustering results across these temporal partitions.
- To implement a mechanism to rank similar clustering results in order to assist practitioners in downstream analytical tasks such as improving regression or classification model performance.

1.2. Scientific Contributions

The scientific contributions of this work can be described as follows:

- An end-to-end system that facilitates the comparison of clustering results by practitioners with different priorities and preferences.
- The use of real-world event data from EV charging station operators combined with nearby traffic volumes and other data sources advances the understanding of EV charging behavior in New Brunswick.
- To the best of the author's knowledge, no other work has fused real-world EV charging events from station operators in New Brunswick with traffic volumes and geographic locations of nearby amenities in an attempt to describe EV charging station clustering results.
- The combination of eight internal cluster validity indices is used to compute a proximity measure (i.e. Euclidean distance) and reduce the cognitive demand on users in identifying, understanding and comparing similar clustering results.

1.3. Organization of Work

The structure of this work is organized as follows. Section 2 provides background on the regional context in which EV infrastructure investments occur. Section 3 summarizes related research work in this domain. Section 4 describes the proposed clustering process underpinning this work and Section 5 provides a summary of its implementation details. In section 6, the analytical results for a case study demonstrating the proposed approach and platform implementation are discussed. Finally, section 7 concludes and outlines future research work.

2. Background

Electrifying transportation will play a key role in Canada's commitment to impact climate change. Currently, EV uptake in Canada is relatively low compared to other countries. Fundamental to encouraging households in adopting EVs are national and regional consumer incentives. However, some regions struggle to provide basic government services and therefore need to optimize consumer incentives (if these are even an option) in addition to charging infrastructure investments in order to achieve maximal impact. To provide further context, this section focuses on EV charging infrastructure and describes the regional setting in which infrastructure investments occur. This section also introduces the various types of cluster analysis techniques available to data analysts and presents decision tree feature importance as a means of contextualizing and interpreting clustering results.

2.1. Electrifying Mobility

2.1.1. Electric Vehicle Charging Equipment

EV chargers or charging stations are commonly referred to as Electric Vehicle Supply Equipment (EVSE). The equipment sits between the power source and the vehicle's charging port. EVSE is comprised of cables, connectors and other devices that ensures a safe usage when transferring power to vehicles. In addition to power, EVSE is used to exchange data between charging equipment and the vehicle.

2.1.2. Charging Levels

EV charging opportunities are often grouped in three levels based on voltage, current and typical charging times. These levels are : Level 1 (L1), Level 2 (L2) and Level 3 (or DC Fast) [3, 8].

In Figure 1, from left to right, the first level of charger (L1) requires a standard three-prong, 120 volt alternating current (AC) household plug. The slowest among all charging connector types, it can take between 8 and 30 hours to fully recharge an EV battery with L1 charging. The use case associated with this type of charging is typically overnight parking or any other situation where the vehicle will be parked for a long period of time. The second level of charging (L2 charging) requires a 240 volt AC plug. It can take between 4 and 10 hours to fully recharge an EV with a level 2 charging station. The use case associated with this type of charging station typically involves parking at the workplace, home or short-term public locations such as stores or other public parking situations. The last type of charging station is the Level 3 charging station type. This type of charging station is also known as Direct Current Fast Chargers (DCFC). L3 chargers are the fastest amongst all charging

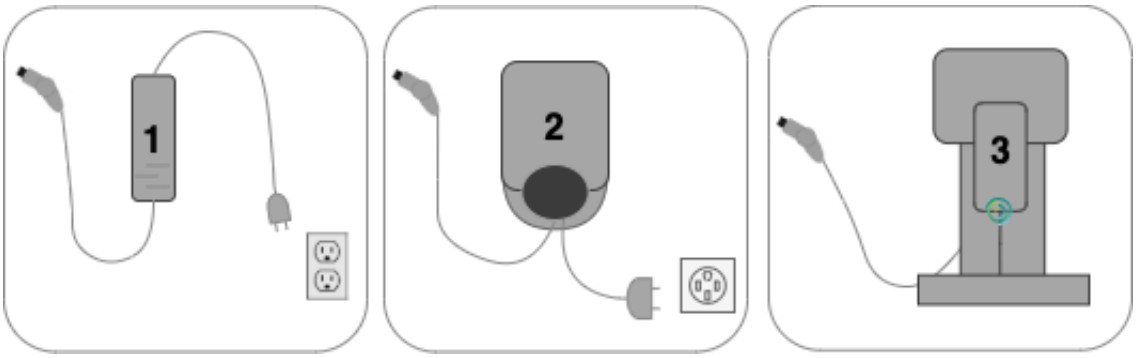


Figure 1: EV Charging Levels

station types. The power provided by L3 charging stations is up to 480 volts Direct Current (DC). L3 chargers can charge a vehicle's battery at 80% capacity in 25 to 30 minutes. The use case typically associated with this type of charging station is for quick charging during a prolonged road trip in areas such as along major highways.

2.1.3. Electric Vehicles and Connections

EVs can be powered either entirely from batteries or from the combination of batteries and an internal combustion engine (ICE). A Battery Electric Vehicle (BEV) is powered entirely from batteries. A BEV has a battery pack which is recharged by plugging the vehicle into an electrical outlet. The energy stored in the battery pack is used to power an electric motor and propel the vehicle. A Plug-in Hybrid Electric Vehicle (PHEV) has two motors. One is electric and the other is powered by gasoline. The PHEV's battery pack can be recharged by plugging the vehicle into an electrical outlet, regenerative breaking or the ICE. Depending on the PHEV, the vehicle can be propelled by the ICE or the electric motor or both systems simultaneously [8]. Stakeholders in the EV charging industry have agreed upon a handful of connector types when plugging a vehicle to a source of power for charging and communication.

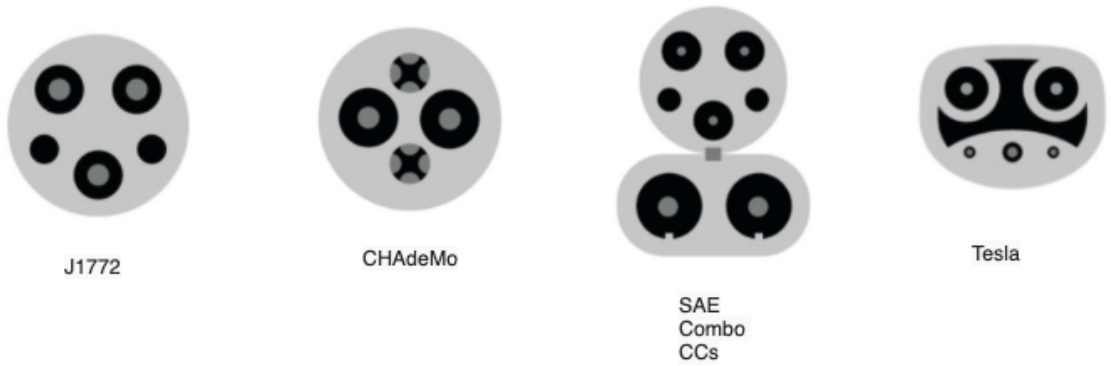


Figure 2: EV Station Connector Types

Figure 2 outlines common EV charging station connection types. From left to right, the J1772 connection type is the most common plug for level 1 and 2 charging in North America. For DCFC stations, CHAdeMO, SAE Combo CCS and Tesla are available and the vehicle connectivity depends on the manufacturer.

2.1.4. Charging Stations

Charging stations have different capabilities. Some charging stations operate on their own, do not have network connectivity and their main function is to charge the vehicle. Others can be networked and allow for operator controls such as load management, billing, monitoring, advertising and promotions etc. The basic, non-controllable chargers are also known as smart chargers. Smart chargers can automatically distribute power evenly in order to manage load. The more advanced and controllable chargers are often referred to as intelligent chargers. These chargers can be controlled and managed remotely in order to manage load and demand etc. Electric Vehicle Energy Management Systems (EVEMS) use the Open Charge Point Protocol (OCPP) to communicate with controllable charging stations. The OCPP protocol is vendor neutral, which enables communication with multiple EV charging hardware using a 'common language' regardless of the hardware vendor.

2.1.5. The EV Market in New Brunswick

New Brunswick is located in Eastern Canada and is one of the four Atlantic provinces (NB, NS, PEI and NL). The Atlantic region's ratio of people aged 65 and older is above, and average annual incomes are below, the national averages. The region has seen a steady out-migration and slow population growth over the years. Only a few large urban centers offer public transit meaning that the majority of the population depends on private vehicles for mobility [9]. New Brunswick has a surface area of 71,388 square kilometers, and according to the 2016 Canadian census, had 747,101 inhabitants. In 2016, there were 584,533 total road motor vehicle registrations in the province. The general trend in recent years with this statistic is upward [10, 11].

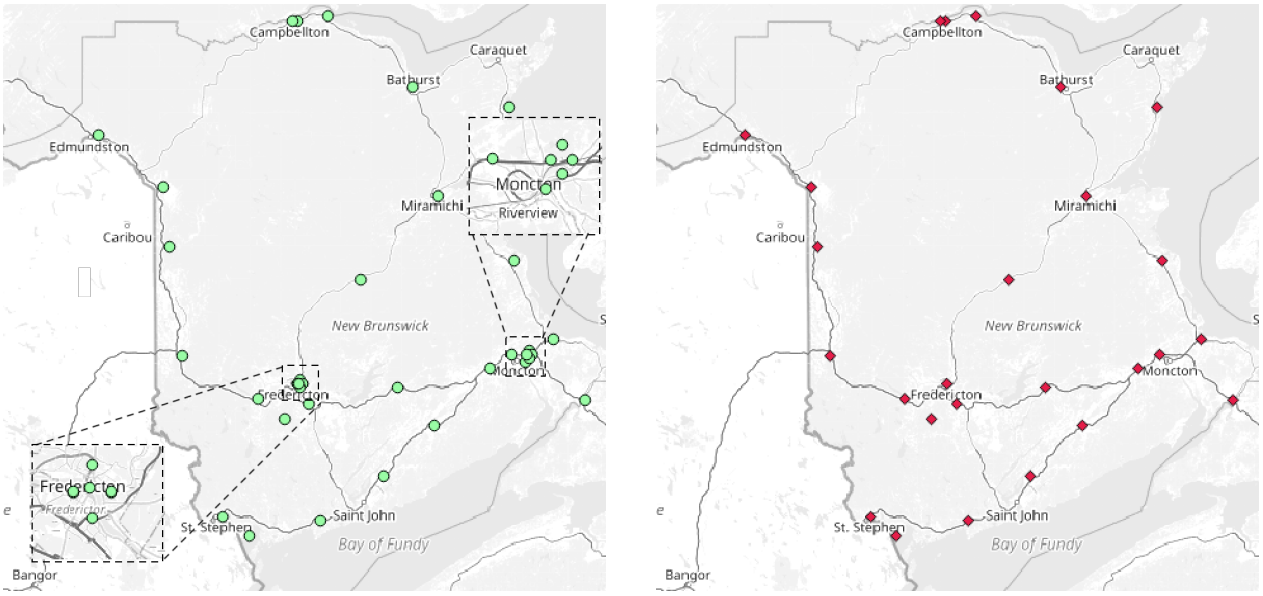
With respect to EV sales in Canada as a whole, the market share for EVs is relatively small (almost 2% in 2019). Comparatively, for the same period, the U.S. EV market was at 2.5%, it was 7% in the Netherlands, and 8% in Sweden. The largest provinces: British Columbia, Ontario and Quebec led in EV sales in Canada. The remaining provinces, which includes the Atlantic and prairie provinces are well behind in terms of EV sales; totalling 0.8 EVs per 1000 households [12].

According to [1], there were very few electric vehicle sales up to the year 2015. Electric vehicles represented just over 0.01 per cent of the light duty passenger vehicles registered in the province. Eastern Canadian Provinces (NB, NS, PEI, NL, and QC) had a total of 6,457 electric vehicles (with Québec counting for 6,166 of those).

Public EV charging stations in New Brunswick, operated by the New Brunswick Power Corporation, at the time this study was performed, are mapped in detail in Figure 3. Stations are located throughout the province where there is a distance of approximately 60 Kilometers between charging opportunities.

2.2. New Brunswick Road Network

The length of the two–lane equivalent, paved public road network in New Brunswick is 19.5 thousand kilometers. The province's road network is also comprised of the national highway system's 1,829 core and feeder routes [13]. Figure 4 maps major highways and EV charging stations in New Brunswick. As can be seen in the figure, charging stations are mostly located near arterial roads, expressways and freeways.



(a) L2 Stations (b) L3 Stations

Figure 3: Charging stations in New Brunswick, Canada.

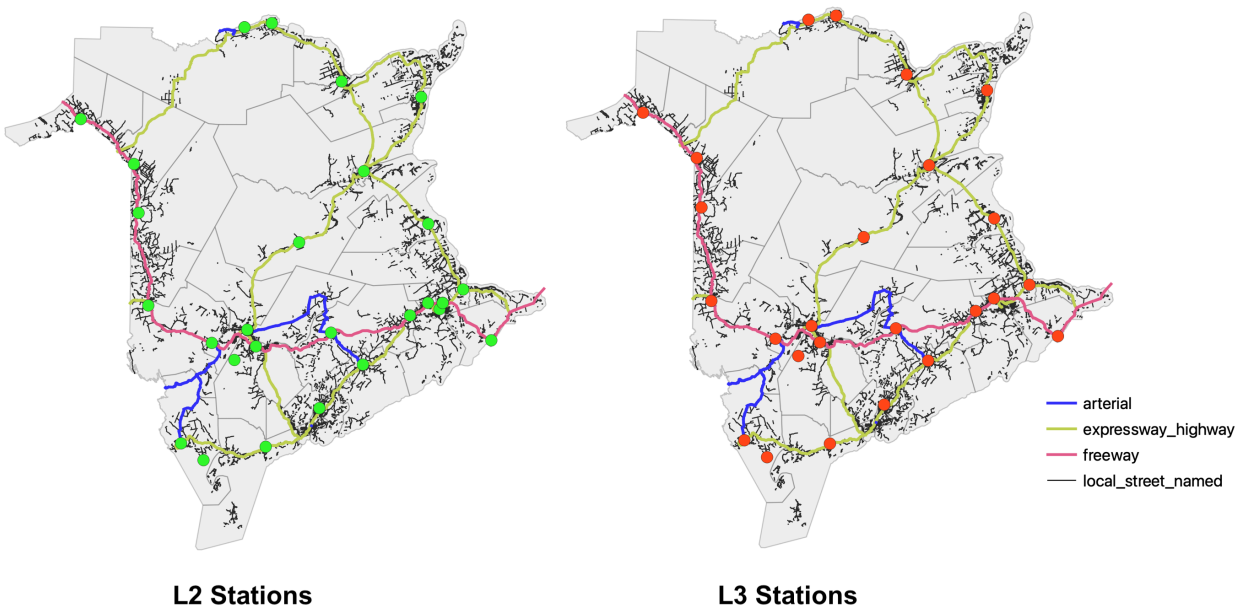


Figure 4: New Brunswick Major Highway Classes and EV Charging Stations

3. Related Works

3.1. Clustering research in the EV applications

Clustering is a popular method applied in many smart grid applications to group similar consumers, predict future energy consumption or detect outliers. By grouping data points

where knowledge about classes is unavailable, clustering algorithms play an essential role in data summarization, discovering hidden patterns for energy usage behavior, such as the EV charging demand [14], and improving downstream modeling performance. Straka et al. [4] compared the k-means, hierarchical, and DBScan clustering algorithms aiming to explore the usage patterns related to charging stations segments. The authors demonstrated the potential of clustering to reveal new insights by successfully identifying four groups of EV charging stations characterized by distinct usage patterns. With available EV charging data, exploring energy usage behavior may improve emerging applications and services. For example, energy load forecasting methods might perform better when applied to homogeneous clusters of stations as opposed to all stations [15]. The improvements in predictive algorithm performance could provide meaningful support for smoothing frequent peaks and imbalances in power grids [5].

However, dealing with clustering in general is a non-trivial work. Various steps must be taken by a practitioner/developer such as the selection of an appropriate algorithm and its hyper-parameters, the right choice of an adequate proximity measure, and how to properly validate the modeling results. A common problem in clustering is how to objectively and quantitatively evaluate the results. Cluster validation is an important task in the clustering process because it aims to compare clustering results and solve the question of optimal cluster count. From the literature, many internal validity indices have been proposed to evaluate performance of a clustering algorithm in finding the natural clusters in a random data set without any class label information [16, 17]. Numerous studies on validating clustering results use individual CVI to determine relative partitioning performance. Arbelaitz [18] et al. has selected 30 CVI to compare and evaluate over multiple data sets with the ground truth to find the best partitioning results. The optimal recommended number of partitions is defined as the one that is the most similar to the correct one measured by partition similarity measures. From this study, authors indicated that cluster overlap and noise had the greatest impact on CVI performance. Some indices achieved good performance with high dimensionality data sets. They also performed well in some cases where homogeneity of the cluster densities did not exist. This study recommended using several CVI to obtain robust results.

Few publications have explored using CVI for evaluating clustering results in the energy domain. Recently, the Davies-Bouldin index [19] was used to determine the best value for the cluster count parameter applied to the k-means algorithm. Sun et al. [20] evaluated the clustering results obtained with Dynamic Time Warping distance (DTW) and Euclidean distance based on the silhouette coefficient using charging tails from ACN-Data collected from smart EV charging stations. Euclidean distance, which measures the root of square differences between coordinates of pairs of objects, is a popular distance measure favored by many researchers in the field of clustering [21]. CVI have traditionally been used for validation purposes. However, combining multiple CVI together with a proximity measure such as Euclidean distance may lead to new similarity comparisons for EV charging event data clustering results, improving the ability to compare results over multiple time slices. By automating clustering pipelines and organizing meaningful analytical results in a way that intensifies the opportunity to compare and understand similar clustering results, the practitioner/developer can reduce cognitive load and demand in discovering EV charging patterns and detecting meaningful results for downstream analysis. The central approach of utilizing multiple CVI together in this article represents an innovation in the energy domain

and in the field of data science.

3.2. Challenges in EV Data Clustering and Results Exploration

Selecting an appropriate algorithm in clustering is critical since its performance may vary according to the distribution and encoding of data. For instance, the application of the Hierarchical Agglomerative Clustering (HAC) algorithm is usually limited to small data sets because of its quadratic computational complexity. Additionally, hierarchical methods are not always successful in separating overlapping clusters and the clusters are static in the sense that a point previously assigned to a cluster cannot be moved to another cluster once allocated [22, 23]. Essential to the practice of clustering is that different clustering techniques will work best for different types of data. There is no clustering algorithm that can be universally used to solve all problems. In fact, practitioners/developers have become interested in recent years in combining several algorithms (e.g. clustering ensemble methods) to process data sets [24].

Selecting the appropriate algorithm and hyper-parameters in clustering is critical. However, it could be cognitively demanding to successfully interpret the clustering results. There may exist several viable combinations of algorithms and hyper-parameters that result in plausible clusters. Comparing and contrasting multiple clustering results can help uncover interesting structure in data. Nevertheless, this comes at a cognitive cost since users will have to expend effort to cognitively encode and interpret these results. Additionally, in data with a temporal component such as EV charging events for example, assessing the structure consistency of discovered clusters over different temporal granularities adds additional demands. Supporting the practitioner in analytical results exploration helps reduce cognitive demand in comparing and contrasting results.

3.3. The state-of-the-art of platforms for exploring EV charging events

Research activities aimed at assessing the impact of widespread EV adoption on distribution networks are taking place with great enthusiasm in academia and industry [25, 26]. However, a large proportion of studies use data from simulated sources rather than real-world EV charging events [27]. Many research works have applied mathematical modeling and and computer-based simulations to gain knowledge regarding patterns of infrastructure usage and deployment [28]. There are few studies which perform location analysis based on data-driven analysis and modeling [29, 30]. This is evident in a recent review article by Hardman et al. [31] on consumer preferences with regards to EV charging infrastructure, which lists studies that employed surveys, interviews, modeling and vehicle GPS data in addition to a small number of studies using EV charging equipment information. Recently, Ashkrof et al. [2] explored EV driver travel behaviors in the Netherlands and point out the main limitations of their work was related to the low number of Battery Electric Vehicle (BEV) driver participants. Hybrid Electric Vehicle (HEV) and Plug-in Hybrid Electric Vehicle (PHEV) drivers were added to the study to compensate for this limitation. In spite of the hindrance in targeting BEV owners uniquely, this study pointed out that route attributes (e.g. travel time, charging infrastructure characteristics en route to and at the destination of travel), recharging wait times, and State-of-Charge (SoC) considerably influences EV charging behavior and route selection. In this research, the authors also found that the two main concerns of EV users are lack of charging opportunities and SoC.

To recognize and analyze patterns in data, especially when dealing with a new data set, descriptive analytics provide a good framework to guide the introductory stages of information processing. This process is meant to acquire an initial understanding of historical data and organize it for advance analysis. Advanced stages of descriptive analytics can include grouping sets of similar and dissimilar objects in clusters. These base analytical functions form the foundation for increasingly complex downstream analytical tasks. Using real-world, contextualized data and supporting the practitioner in analytical results exploration can help reduce cognitive demand in results interpretation and improve downstream modeling performance.

By using real-world charging event data, related demand characteristics can provide a better assessment of increased EV adoption and the potential impact on energy demand. However, EV charging patterns cannot be effectively analyzed in isolation from the social and economic contexts in which they occur. A more complete picture of EV infrastructure usage patterns can be formulated by combining data from different sources. As noted previously, scarcity of foundational EV charging data sources is a concern. This is also true for complementary data sources. Some research works based on real-world EV charging data with contextual information (e.g. traffic volumes, land use, neighboring amenities and driving distances between charging sites) have started to appear in the literature. For example, a study by [30] suggests that enriching EV charging data sets with contextual information provides useful infrastructure-related insights. However, this study's authors note that low utilization rates of some charging infrastructure in an early adoption context introduces some limitations in the obtained results.

Few studies focused on developing platforms to support practitioners in analyzing EV charging data are found in the literature. Recent works describe web platforms which display EV charging information and enable monitoring of charging infrastructure. For example, Maase et al. [32] developed an web-based assessment platform to show key performance indicators such as to kWh, connection time, charging time. The platform utilizes data from four of the largest cities in the Netherlands (Amsterdam, Utrecht, The Hague, Rotterdam) and their surrounding areas. Another example is a web platform for sharing the information about privately owned charging stations [33]. This platform is operated as an interactive map, based on the Google Maps service. No work in the literature attempts to build a platform that is able to automate the clustering process and proposes a solution to facilitate the comparison of clustering results, reduce the cognitive demand on users in identifying, understanding, and comparing similar clustering results.

4. The Proposed End-to-End EVStationSIM Platform

This section describes the proposed analytical platform that was developed to cluster and explore patterns that emerge around EV charging infrastructure usage. The system information flow is utilized to diagnose the cause of phenomena observed at charging stations based on clustering of real-world charging events. The information flow runs through 7 stages as shown in Figure 5, which are described as follows:

- **Data Collection - (Section 4.1)** : The gathering of EV charging event data and supplementary information related to charging stations.

- **Data Preprocessing and Fusion - (Section 4.2)** : The cleaning, transforming and combining of data collected from multiple sources to produce consistent, accurate, and useful data files.
- **Feature Generation and Selection - (Section 4.3)** : The creation of contextualized, semantically enriched data for downstream analytical tasks.
- **Clustering - (Section 4.4)** : The discovery of internal structure in data using the hierarchical agglomerative clustering algorithm.
- **Processing and Harvesting Validity Indices - (Section 4.5)** : The computation and normalization of eight cluster validity indices for each clustering result.
- **Similarity Computation - (Section 4.6)** : The computation of relative proximity measures for all clustering results and the creation of relative rankings of comparable station groupings.
- **Facilitate Pattern Exploration - (Section 4.7)** : The suggestion of initial relevant clustering results to explore and the ability to query relative rankings of comparable results for diagnostic and down-stream analytical tasks.

4.1. Data Collection

This stage consists of retrieving EV charging event logs and station location information for Level-2 and Level-3 charging stations. The raw data were made available by the New Brunswick Power Corporation (NB Power) in a Microsoft Excel file format. This stage also involves retrieving data from other sources in various file formats. For example, relevant traffic count totals were provided by the New Brunswick Department of Transportation and Infrastructure (NBDTI) in a tabular, PDF file format. Supplementary information related to charging station locations such as information regarding nearby amenities, average property assessment values and closest metro area populations were retrieved from the GEONB [34] data catalogue. These were available in the ESRI shape file format. Once all data were retrieved, the raw data was transformed and enriched in subsequent stages.

4.2. Data Preprocessing and Fusion

This stage uses raw charging event data from public EV charging stations. Preprocessing consists of data cleaning, consolidation and applying a one-way hash function to mask sensitive customer data. Data cleaning ensures superior information quality and produces a set of cleaned files by eliminating errors, inconsistencies, duplicated and redundant rows, and handling missing data. Consolidation combines data from various files into a single data set. Multiple files from the cleaned data set are used as the input for this operation. The output of this activity is a single file that merges all attributes into one large table.

Additionally, data fusion consists of combining multiple data sources which can be followed by a reduction or replacement for the purpose of maximizing performance. In the proposed system information flow, charging event data files are combined with station location information, neighbouring amenities and permanent traffic count totals to produce more consistent, accurate, and useful data files.

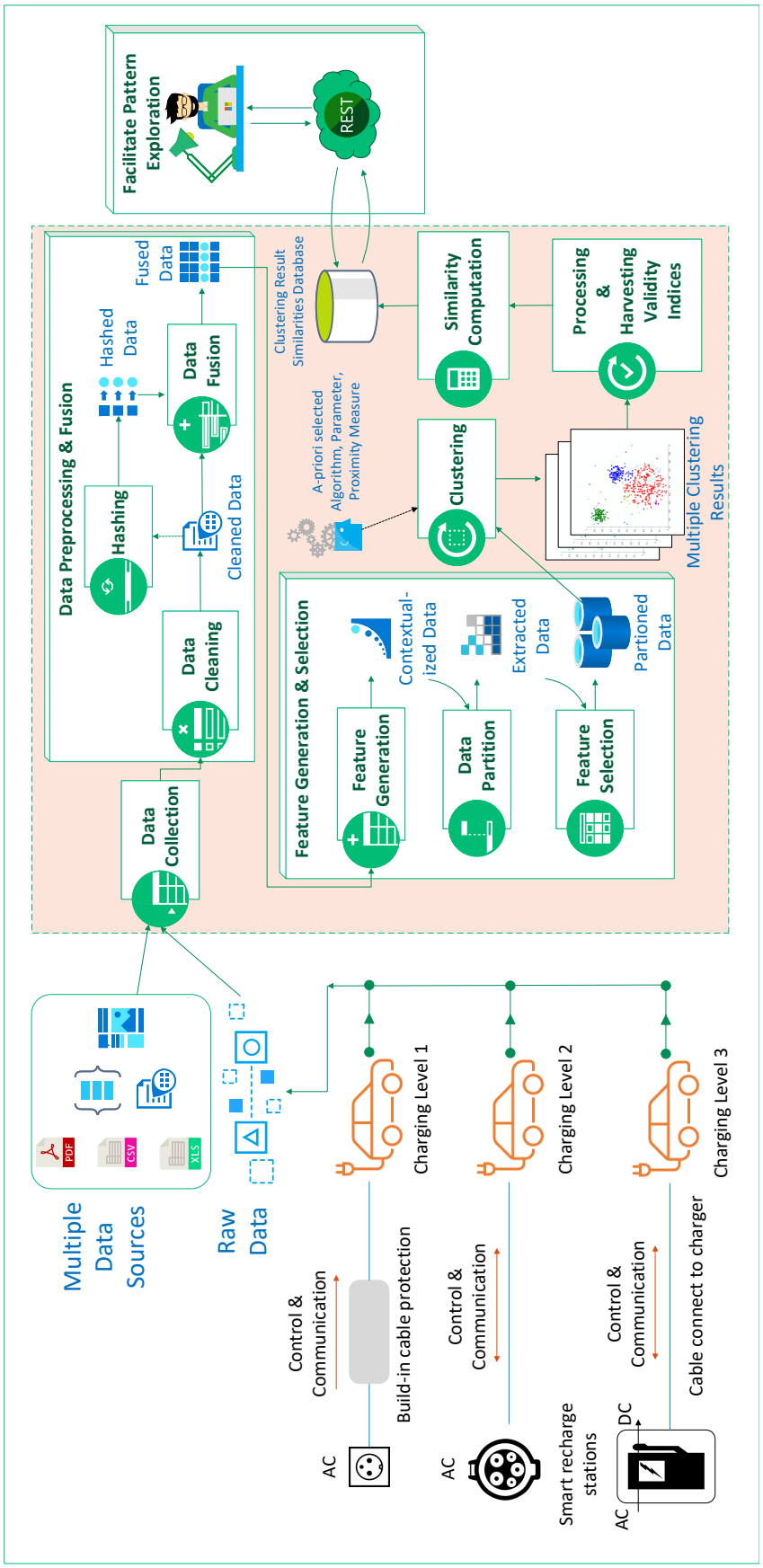


Figure 5: Proposed Analytical System Information Flow

4.3. Feature Generation and Selection

The aim of the feature generation and selection stage is to enrich preprocessed data files and to select relevant information for further analysis [35, 36]. Feature generation incorporates the construction of new attributes from raw data, which typically involves creating a mapping that converts original attributes into new attributes. Feature selection consists of identifying relevant variables that will improve the performance of downstream analytical tasks such as finding meaningful clusters or better inference. Once feature generation and selection are complete, the transformed data files are partitioned using a priori selected temporal granularities (e.g. weekly, monthly or seasonal) as a final data preparation activity. This facilitates the ensuing analysis over various time slices.

4.4. Clustering

The objective of the clustering stage is to find the patterns from transformed input data using the Hierarchical Agglomerative Clustering (HAC) algorithm [37]. The algorithm seeks to build a hierarchy of clusters by merging current pairs of mutual closest input data points until all the data points have been used in the computation. The measure of inter-cluster similarity is updated after each step using Ward linkage. This a priori selected algorithm is utilized to fit the various temporal granularities of input data, producing multiple clustering results. Internal clustering validity indices are recorded during each application of the clustering algorithm.

HAC is an unsupervised learning method which uses an iterative, bottom-up approach when determining cluster memberships. The algorithm does not require pre-specifying the number of clusters prior to its usage. The procedure starts with each individual point in the data set forming its own cluster; then the two closest clusters are merged together at each iteration until all points are merged into one large cluster. This process produces a tree structure called a dendrogram (See Figure 6), which contains all possible clusterings of the data set. The number of clusters is selected from the method's output by deciding where to cut the dendrogram in order to get the best possible partitioning of the data. Given that the dendrogram embeds all possible clusterings, cutting the tree in order to get different partitioning of the data is performed in constant time [38].

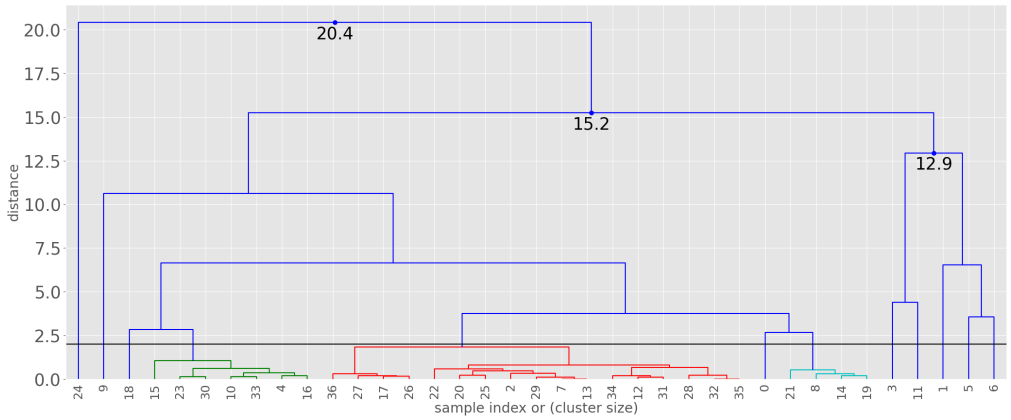


Figure 6: Example Dendrogram

HAC also requires a measure of distance between the clusters when deciding how to group the data at each iteration. This measure of cluster distances is done with a linkage function that captures the distance between clusters. Common measures of distance in this context include Ward and complete. Ward minimizes the variance of the clusters being merged. When making a merge decision with the Ward approach, two clusters will be merged if the new partitioning minimizes the increase in the overall intra-cluster variance. Complete uses the maximum distances between all observations of the two sets. When making a merge decision with the complete approach, two clusters will be merged if the new partitioning maximizes the distance between their two most remote elements. Even though the algorithm does not require pre-specifying the number of clusters prior to its usage, in order to get the best possible partitioning of the data, a decision on exactly where to cut the tree must be made. The objective of the clustering is to provide the most compact and well-separated clusters. As a final remark in the discussion of clustering, it should be noted that the HAC method is but one of many possible clustering algorithms that could be leveraged by the proposed platform.

4.5. Harvesting and Processing Validity Indices

The purpose of harvesting and processing the validity indices is to prepare the clustering operation output for the subsequent creation of a clustering result similarity matrix (described in Section 4.6). This similarity matrix is utilized to identify clustering result similarities across various temporal slices in the data. Each application of the clustering algorithm generates a record consisting of the cluster count parameter value, the various cluster validity index values and the input data used to generate the clusters. Processing the validity indices involves selecting and normalizing the index values in preparation for Euclidean distance computations. This stage utilizes a combination of eight cluster validity indices, which are listed in Tables 2 and 3 and thoroughly described in [39].

4.6. Similarity Computation

The objective of the similarity computation stage is to perform pairwise Euclidean distance calculations for each clustering result and create similarity rankings of station groupings. The similarity computation stage uses Euclidean distance as the proximity measure between clustering results. A combination of eight cluster validity indices, individually described in [39], are utilized in the distance computations. The pairwise similarity comparisons (e.g. the similarity matrix) are then persisted in a database to facilitate exploring clustering result similarities across various temporal slices in the data.

4.7. Facilitate Pattern Exploration

The purpose of the pattern exploration facilitation stage is to enable the basic identification and interactive querying of potentially interesting clustering results. Additionally, this stage enables drilling down into relative rankings of comparable results for diagnostic and downstream analytical tasks. This stage leverages a RESTful API in order to facilitate this capability.

5. EVStationSIM Platform Implementation

In this section, we provide key implementation details of the EVStationSIM platform proposed in the previous section. Custom-written Python code and a scientific Python stack were leveraged to implement the platform. The software programs used in this work were packaged using Docker [40] in order to ensure a reproducible and consistent computational environment. An architecture composition diagram is provided in Figure 7. The software used for the platform implementation is summarized in Table B.5 of Appendix B.

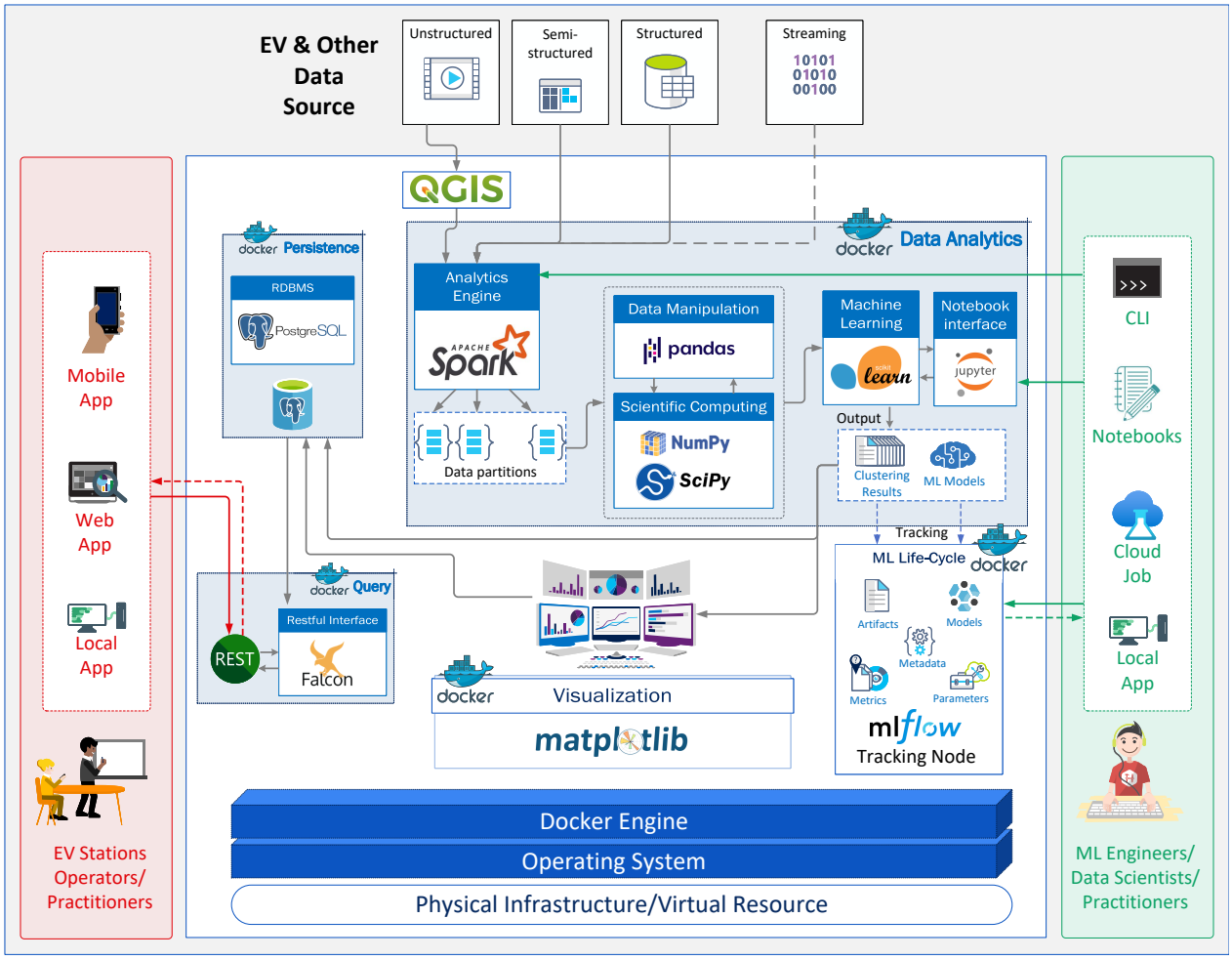


Figure 7: Architecture Overview

In Table 1, we can observe how system information flow elements introduced in Figure 5 are linked to concrete platform implementation components in Figure 7.

Table 1: From Platform Proposal to Concrete Implementation Components

Stage	**System Information Flow** - *(Figure 5)*	**Architecture Overview** - *(Figure 7)*
Data Collection	Data Collection Raw Data from Multiple Data Sources	EV Data Sources Structured and Other. QGIS Transformations
Data Preprocessing and Fusion	Hashing, Data Cleaning, Data Fusion	Data Analytics & ML Life-Cycle Containers
Feature Generation and Selection	Feature Generation, Data Partition, Feature Selection	Data Analytics & ML Life-Cycle Containers
Clustering	Clustering - Multiple Clustering Results	Data Analytics, ML Life-Cycle, Persistence & Visualization Containers
Harvesting and Processing Validity Indices	Processing & Harvesting Validity Indices	ML Life-Cycle and Persistence Containers
Similarity Computations	Similarity Computation	ML Life-Cycle & Persistence Containers
Facilitate Pattern Exploration	Facilitate Pattern Exploration - RESTful Interface	Query and Persistence Containers

The platform is meant to be of assistance to EV station operators, which are primarily concerned with accessing analytical results. Station operators access analytical result artefacts via custom applications written for mobile, web or local client access. All clients will leverage the same RESTful web interface. The platform also encapsulates and facilitates the use of common data analytics tools; to be leveraged by machine learning engineers and data scientists. These practitioners can interact with the analytical tools using custom-written code launched from a command line or Jupyter interface. The primary concern for this class of user is to make interesting analytical results available to EV station operators. Table A.4 in Appendix A outlines individual Docker container deployment details.

5.1. System Information Flow

Figure 8 provides a visual overview of the implemented platform's information flow; where weekly clustering results are the primary focus of the exploration. Key implementation details are listed below :

(1) The Raw EV charging data contains around 9500 recharging events.

(2) The overlap between EV charging events and traffic counter counts is 9 months which spans from April 2019 to December 2019.

(3) The traffic counter and GeoNB data are available to provide additional context to clustering results.

(4) Once feature generation and selection are complete, the EV charging data is split into weekly partitions. The selected weekly partitions spans the 9 months for which there is an overlap between EV and traffic counter events.

(5) The clustering process is performed on each weekly partition for a K parameter - also referred to as the cluster count parameter - that ranges from 2 to 7. All other hyperparameters for the HAC algorithm are kept as default.

(6), (7), (8) This produces numerous clustering results (i.e. one result for each K value and week). In the case study presented in the discussion section, results for a K parameter of 2 are explored.

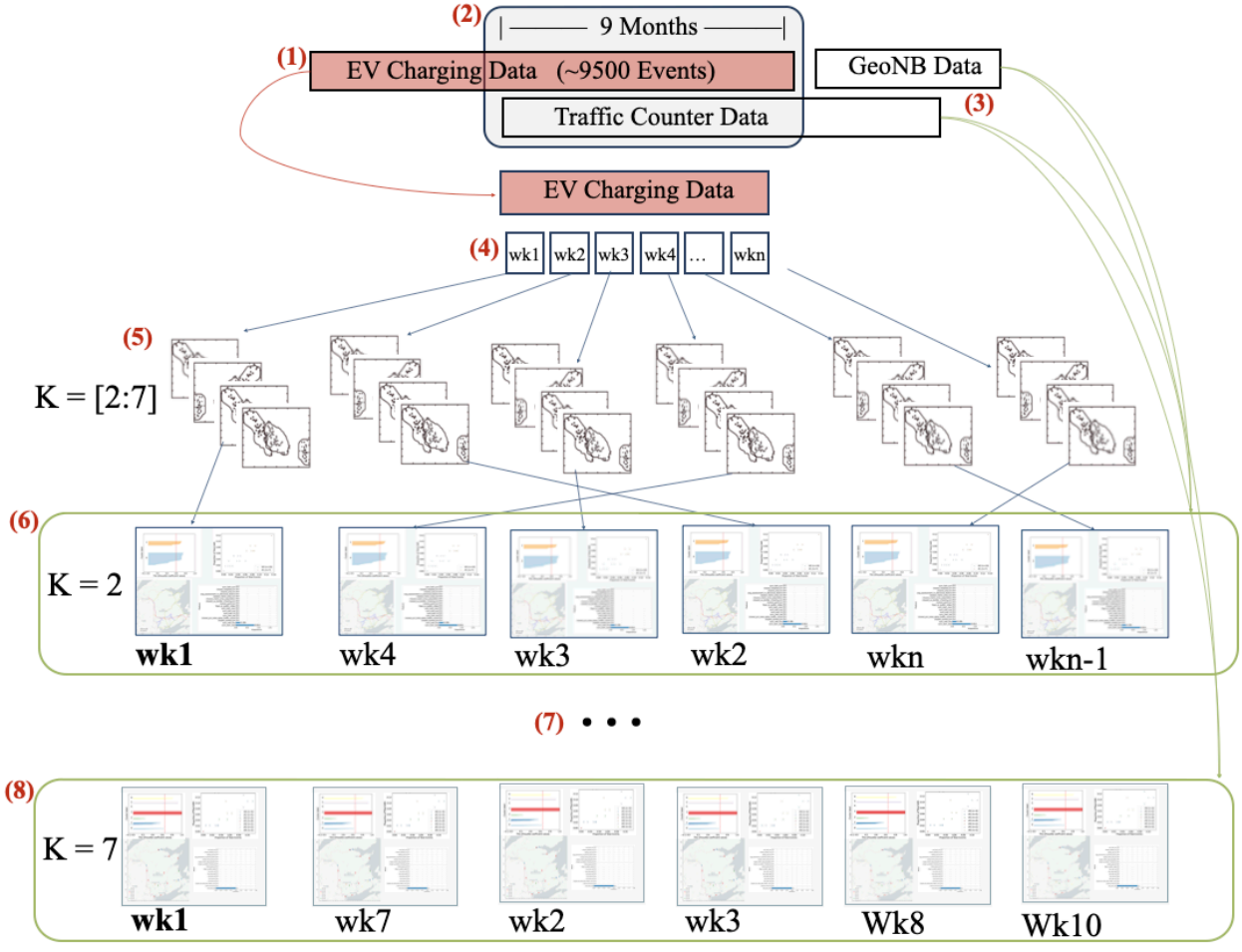


Figure 8: Platform Information Flow Overview

The next few sections describe the implementation details supporting the system information flow described in Figures 5 and 8.

5.2. Data Collection

This section describes the process of gathering the EV charging event data that is foundational to this work. In addition, it describes how this data was combined with supplementary data to formulate a more complete picture of EV infrastructure usage patterns.

Real operational data from public electric vehicle charging stations was provided by the New Brunswick Power Corporation. The raw data consisted of recharging reports and charging station location information. The data set contained 9,505 EV recharging events occurring between the dates of April 2019 and April 2020; representing 9,194 hours of charging infrastructure usage and a 97,148.65 kWh energy transfer.

Additional sources of data were collected for integration at the final stage of the EV station clustering process in order to enhance result interpretation capabilities. As an example of a supplementary data source, the New Brunswick Department of Transportation and Infrastructure (NBDTI) provided access to the total daily traffic counts for 36 permanent traffic counter installations in the province of New Brunswick. For each traffic counter installation, hourly and daily traffic count totals were provided. Traffic counter totals were provided for a date range that spans from January 2019 to December 2019.

As a final step in the data collection phase of the implementation, geospatial data regarding available amenities in the province of New Brunswick were harvested from the GeoNB data catalogue [34]. Examples of such amenities include schools, provincial parks, hospitals, provincial government service locations and border crossings. Important information such as amenity names and corresponding GPS locations were harvested.

5.3. Data Preprocessing and Fusion

Here, details of the cleaning, transforming and linking of data harvested from the various sources described above are discussed. The Apache Spark distributed analytics engine and the QGIS Open Source Geographic Information System were primarily used to perform the Extract Transform and Load (ETL) tasks.

Initial pre-processing involved loading raw EV charging event CSV data files into a local, thread-based Spark cluster. Data elements were cast to appropriate types when required (i.e. if they were initially imported as strings). For example, charging event times, initially imported as strings into Spark, were cast as timestamps. Cleaning the data consisted of removing charging events that were less than 5 minutes in duration (eliminating 11% of the raw records). These events were interpreted as connection disturbances and discarded as part of the normal data cleaning process to avoid skewing results with less relevant information. Consolidation consisted of integrating the separate charging event CSV files into one and persisting these on disk in the Apache Parquet file format.

A similar operation was performed on the EV charging station location CSV files. QGIS and the MMQGIS [41] plug-in were also used to generate hub-line distances between charging stations and closest permanent traffic counter. After charging stations were linked to traffic counters based on proximity, the resulting QGIS layer's data table was exported for downstream usage. The fusion of EV charging station information with the nearest amenities such as hospitals, national parks, government services, largest metropolitan areas, etc. was

done in QGIS. In total, 8 additional GIS features from the GEONB [34] data catalogue were linked to EV charging station location records using this approach.

The Parquet format is a popular file format with Big Data cloud service providers. Using this file format can save computation time and money because it is optimized for query performance and I/O minimization. The file format also supports very efficient compression and encoding schemes. Using the Parquet file format whenever possible in the platform translates into overall performance gains due to the advantages in I/O and data compression over the CSV file format.

5.4. Feature Generation and Selection

The feature generation process creates new features (contextualized) based on calculations involving existing data attributes. It prepares data in order to be compatible with a machine learning algorithm's requirement and can lead to performance improvement of a machine learning algorithm. In order to simplify the experimental setup, two features were utilized to cluster stations in the system information flow and case study.

Weekly, monthly and seasonal temporal partitions of the charging event data were created prior to applying feature transformations. These partitions facilitate the comparison of clustering results based on charging events occurring at stations during a particular week, month or season of the year. Actors in the energy domain typically define a medium-term duration lasting from one week to one month, while a long-term duration can span one month to years. Once the data partitions were established, the charging events were prepared for the clustering stage by calculating, for each charging station, station type and temporal granularity, the proportion of total charging events and the proportion of total power used to charge vehicles relative to all stations. Additionally, the total daily kWh used to charge vehicles for each station was calculated for all weekly, monthly and seasonal time slices of the charging event data. These daily aggregate kWh values in addition to the fused data from the data fusion stage are used in the pattern exploration stage to explain the clusters.

5.5. Clustering

In this stage of the system information flow, the agglomerative clustering algorithm is applied to all temporal slices of the data produced in the previous stage. The clustering algorithm's input features are, for each station, the proportion of total charging events and the proportion of total power used to charge vehicles relative to all stations. This is done for a cluster count hyperparameter that varies from 2 to 7. Other hyperparameter settings are kept to the algorithm's defaults to simplify the experimental setup. Internal clustering validity indices are recorded during each application of the clustering algorithm (See Table 2 for the list of recorded indices). Additionally, the clustering process also generates maps and various charts describing the clusterings. These modeling artefacts are also recorded to facilitate the clustering pattern exploration.

Table 2: Recorded Clustering Validity Index Data

Column Name	Description
file_name	File name for clustering results for station type and time granularity
n_cluster	K parameter value used in applying the clustering algorithm
silhouette_score	Silhouette index value for clustering result
calinski_harabasz	Caliński-Harabasz index for clustering result
davies_bouldin	Davies-Bouldin index for clustering result
cohesion	Cohesion index for clustering result
separation	Separation index for clustering result
RMSSTD	Root mean square standard deviation index for clustering result
RS	R-squared index for clustering result
XB	Xie-Beni index for clustering results

5.6. Harvesting and Processing Validity Indices

Every application of the clustering algorithm on weekly, monthly and seasonal partitions of the data generates a record consisting of the cluster count parameter value, the 8 cluster validity index values and the input data used to generate the clusters. Additionally, visual representations of the clustering results such as maps, scatter plots, dendrograms etc. are also generated.

Harvesting and processing the validity indices involves normalizing the recorded index values of the clustering process in preparation for Euclidean distance computations. This prepares the clustering validity index data for the creation of a clustering result similarity matrix.

5.7. Similarity Computation

Pairwise Euclidean distance computations are performed for each clustering result. All validity index values (e.g. multidimensional points in Euclidean space) of each clustering result are used in the distance computations. The similarity matrix created in this step of the system information flow is utilized to identify clustering result similarities across various temporal slices in the data.

Finally, the similarity matrix is persisted in a relational database to enable the querying of clustering results and corresponding similarities across months, weeks and seasons. The database query functionality is made available via a RESTful API. Clustering result visualizations are stored on the file system in such a manner as to support the database query functionality enabling the retrieval of available clustering result visualizations over this RESTful interface.

5.8. Facilitate Pattern Exploration

At this point in the system information flow, for each station type, temporal granularity and for a cluster count hyperparameter that varies from 2 to 7, there is a clustering result

data set that consists of rows containing a **Station** and a **cluster_number**. Additionally, pairwise Euclidean distances between all clustering results have been recorded.

The distance, in Kilometers, from the closest school, park, hospital etc. for each charging station has also been noted. In addition, for each temporal partition in the data, daily kWh totals have been recorded for each charging station. Given that charging stations were linked to traffic counters based on proximity in the data fusion stage of the information flow, traffic counter counts can be aggregated to match the weekly, monthly and temporal partitions in the EV charging data. This auxiliary information can be used to enhance clustering result interpretation capabilities by providing additional contextual information at the charging station level.

After clustering results are processed, contextualized and persisted, the practitioner can navigate these results via a RESTful interface (See Appendix D and Appendix E for request/response examples). Figure 9 illustrates how the practitioner interacts with the results system. First, the practitioner requests ranked station clustering results for either L2 or L3 station types (Step 1). The system then returns a sorted list of clustering results ordered by silhouette score (Step 2). From this list, the practitioner selects one result as the reference result for which comparable results are desired and then requests these comparable results from the system (Step 3). Finally, the system returns a sorted list of comparable clustering results that is ordered by Euclidean distance (Step 4). This sorted list contains result-specific artefacts such as scatter plots, station cluster membership maps, silhouette plots and CART feature importance plots.

The clustering process implementation and RESTful API facilitate the comparison of clustering result similarities across various temporal granularities. This process is useful in identifying avenues for further analysis. One Level 3 station clustering result for the week of May 27th, 2019 has been selected as a case study to demonstrate the proposed approach and platform implementation. The case study is presented in the next chapter.

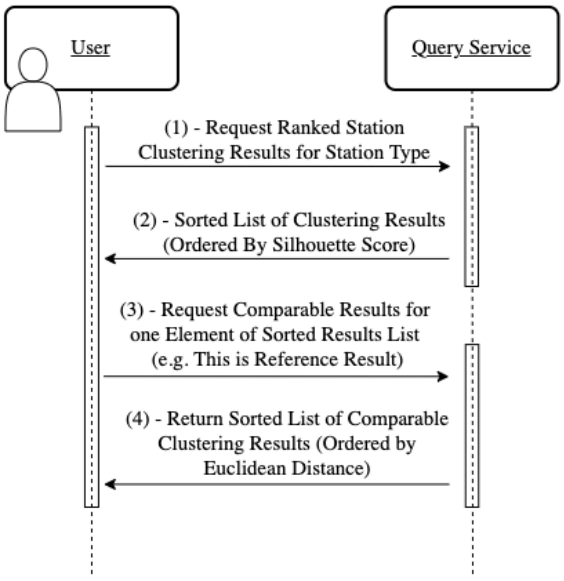


Figure 9: Results Query Sequence

6. Case Study: Applying the EVStationSIM Platform to Analyze New Brunswick EV Charging Data

This section highlights analytical results of our proposed approach and system implementation with a case study. The results presented here identify similar station clusterings over multiple weeks. Table 3 summarizes the clustering similarities relative to station clusterings for a reference week starting on May 27$^{\text{th}}$, 2019. In all results, the number of clusters is 2 and the station type is L3. The table is sorted in ascending order by Euclidean distance relative to the reference week. According to the multi-dimensional pairwise distance calculations obtained, the most similar clustering result to the week starting on May 27$^{\text{th}}$, 2019 is the result for the week starting on August 26$^{\text{th}}$ 2019. The least similar clustering result is the result for the week starting on December 23$^{\text{rd}}$, 2019.

Table 3: Clustering Similarities - L3 - May 27$^{\text{th}}$, 2019

WEEK	Sil	CH	DB	C	S	RMS	RS	XB	*Dist*
27-MAY-2019	**0.600**	**51.370**	**0.512**	**1.123**	**2.403**	**0.153**	**0.682**	**0.0919**	***N/A***
26-AUG-2019	0.623	60.355	0.518	1.135	2.854	0.154	0.715	0.078	*0.138*
29-AUG-2019	0.600	54.727	0.594	1.163	2.652	0.156	0.695	0.107	*0.156*
06-MAY-2019	0.630	57.190	0.600	1.153	2.749	0.155	0.704	0.089	*0.169*
17-JUN-2019	0.624	46.222	0.618	1.145	2.206	0.154	0.658	0.111	*0.181*
07-OCT-2019	0.586	60.513	0.540	1.234	3.111	0.160	0.716	0.099	*0.191*
02-DEC-2019	0.630	56.551	0.582	1.261	2.972	0.162	0.702	0.090	*0.197*
...	...	...	...	...	...	...	...	...	...
04-NOV-2019	0.648	72.433	0.503	1.505	4.542	0.177	0.751	0.081	*0.550*
01-APR-2019	0.575	39.765	0.620	2.028	3.361	0.206	0.624	0.107	*0.613*
28-OCT-2019	0.825	92.435	0.215	0.656	2.525	0.117	0.794	0.0184	*0.784*
23-DEC-2019	0.801	50.970	0.110	0.701	1.489	0.121	0.680	0.0174	*0.825*

Column Name Abbreviations for Table 3

Sil :	Silhouette index
CH :	Caliński-Harabasz index
DB :	Davies-Bouldin index
C :	Cohesion
S :	Separation
RMS :	Root mean square standard deviation
RS :	R-squared
XB :	Xie-Beni index
Dist :	Euclidean distance between current and previous row

6.1. Weekly Grouping Observations

A few weekly groupings were observed based on spatial and temporal patterns in the results. These groupings appear across the ordered list of results in Table 3 in the following arrangement :

- High Utilization in Lower Half of Province
  - Weeks Starting : **27-MAY-2019**, 26-AUG-2019, 29-AUG-2019
- High Utilization at Provincial Boundary Edges
  - Weeks Starting : **06-MAY-2019**, 17-JUN-2019, 07-OCT-2019
- High Utilization in South-East of Province
  - Weeks Starting : **02-DEC-2019**
- Lowering Utilization Rates
  - Weeks Starting : 04-NOV-2019, **01-APR-2019**
- Low Utilization Rates Across The Province
  - Weeks Starting : **28-OCT-2019**, 23-DEC-2019

This work utilizes the groupings mentioned above to organize the ranked list of weekly clustering results into bite-sized portions. Grouping the weeks together in this fashion helps the reader navigate a long list of results in small increments. A representative visual presentation of the clustering results for each grouping can be seen in Figures 10 through 14. Each figure contains a silhouette plot [42], a scatter plot and a map describing the clustered data. Additionally, a feature importance bar chart is included; highlighting which features are the most important, if predicting the cluster labels were attempted using a decision tree algorithm. CART predictive models, are used as an exercise in explaining the clusters using the harvested, transformed and fused features such as nearby traffic counts and amenities.

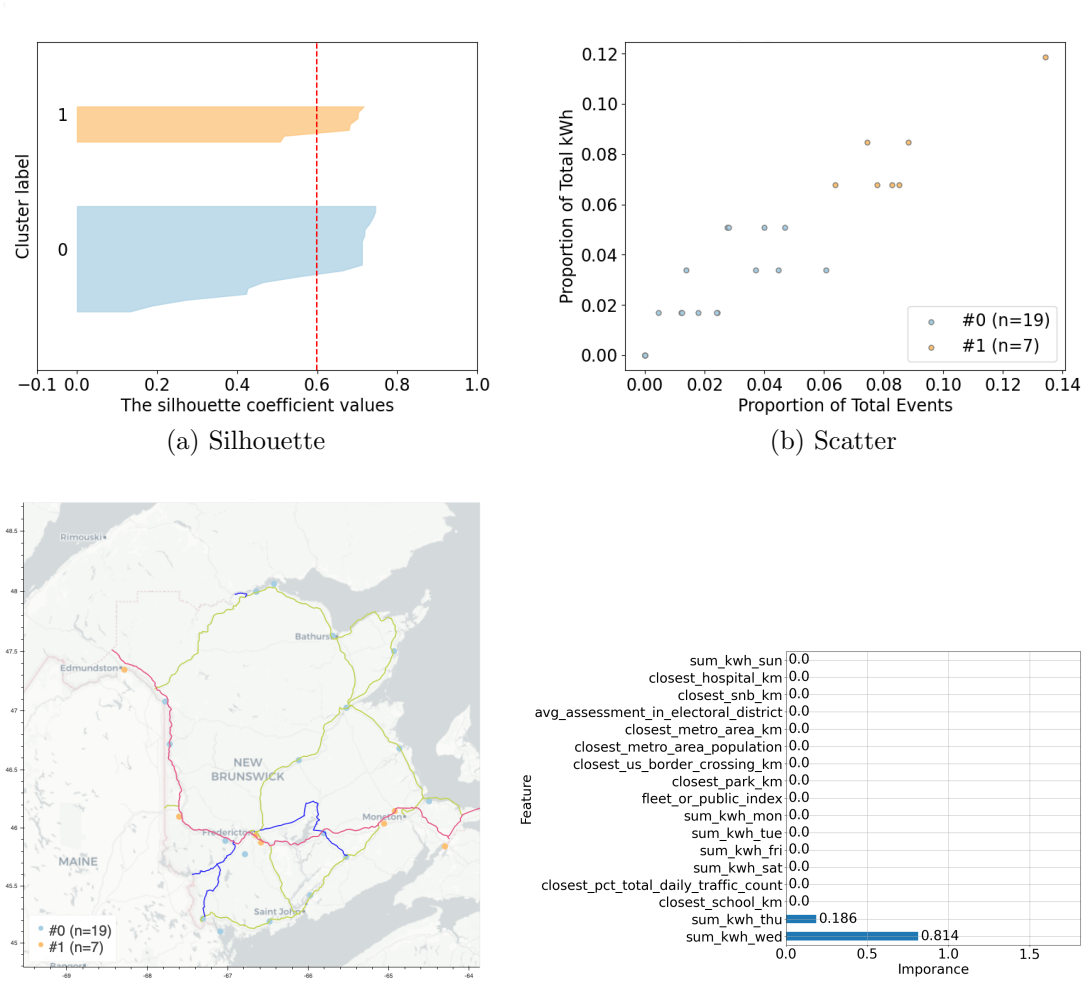
In all silhouette plots found in Figures 10a through 14a, an observation with a silhouette width near 1 means that a data point is well placed in its cluster; an observation with a silhouette width closer to negative 1 indicates the likelihood that this observation might really belong in some other cluster. We describe weekly clustering results individually while contrasting these with the results for our reference week of May 27^{th}, 2019 in the following sections.

6.2. High Utilization in Lower Half of Province

The clustering results for the reference week of May 27^{th}, and the weeks starting on August 26^{th} and 29^{th} are very similar. The clustering result similarities for these three individual weeks are noticeable both in spatial and temporal terms. Spatially, in the three clustering results, cluster 1 - the higher station utilization cluster - member stations are mostly located in the lower half of the province. More specifically, cluster 1 station members are mostly located along a section of the province's road network classified as a highway (i.e. the Trans-Canada Highway in red on the maps). Temporally, the sum of kWh transferred to vehicles on Wednesdays is in the most top important features list in the three results. The clustering result for the week of May 27^{th} is described in detail next. This clustering result also serves as the representative week for the first 3 clustering results found in Table 3.

6.2.1. Reference - Week Starting May 27th, 2019

We can see from Figure 10a that a reasonable structure in the data has been found for our reference week, which starts on May the 27th. In this clustering, stations are grouped in terms of relatively higher and lower utilization rates. The average silhouette score is 0.600 in this clustering result. In Figure 10b, cluster 0, the cluster with relatively lower utilization rates, has more station members than cluster 1, which is the cluster with relatively higher utilization rates.



(a) Silhouette (b) Scatter

(c) Map (d) CART Feature Importance

Figure 10: L3 Station Clusters - (27-MAY-2019)

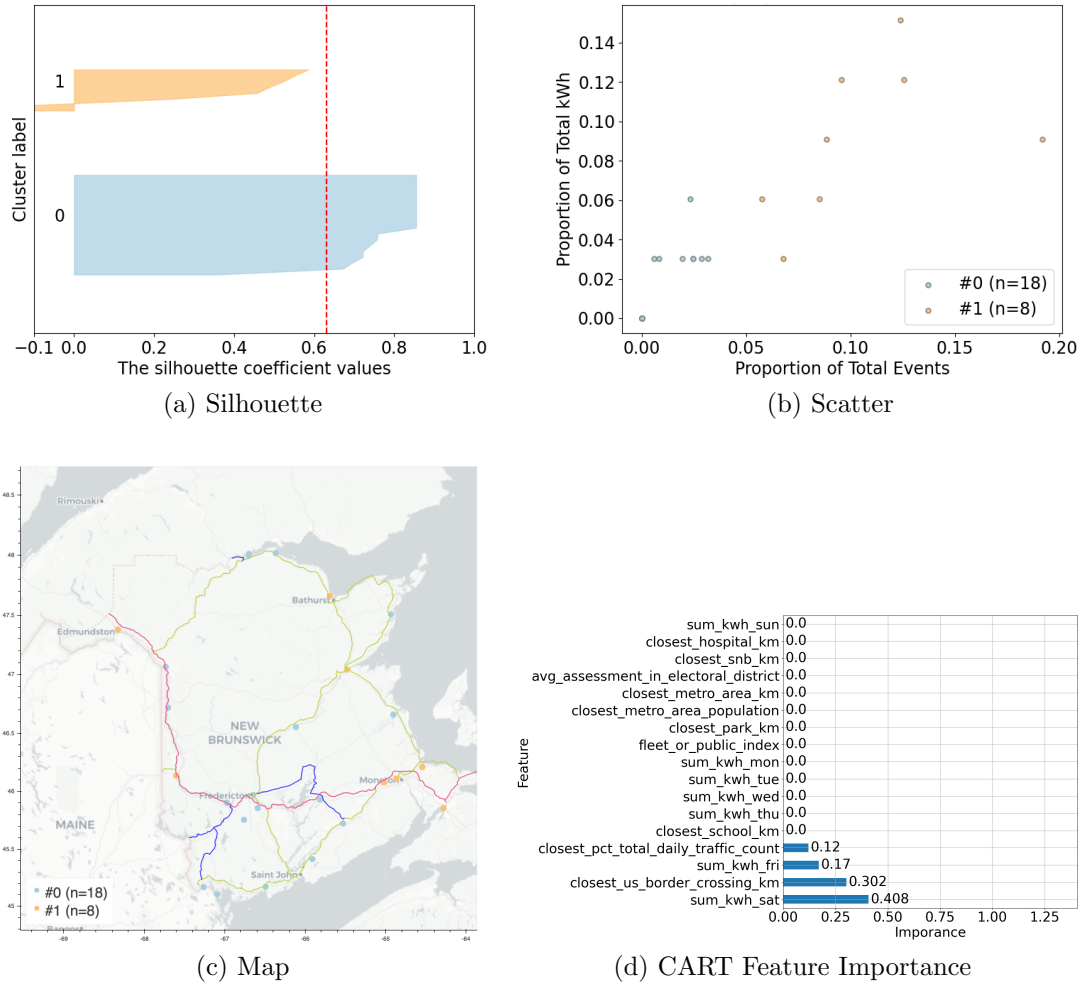
In the scatter plot, we can clearly see the crisp clusters identified by the HAC algorithm. The map in Figure 10c, indicates that cluster 1 member stations are mostly located in the lower half of the province. Finally, in the CART feature importance chart of Figure 10d, we can observe the most significant feature, if we were to use a classification and regression trees algorithm to predict the station member clusters for the purpose of explaining the clusters using supplementary data, is the sum of kWh transferred to vehicles on the Wednesday.

6.3. High Utilization Provincial Boundary Edges

The clustering results for the next 3 weeks in the ordered list of clustering results of Table 3 are also very similar. Spatially, in the three clustering results, cluster 1 - the higher station utilization cluster - member stations are mostly located at the edge of the province. Temporally, the sum of kWh transferred to vehicles on Fridays and Saturdays are top important features in the three results. The clustering result for this group's representative week is described in detail next.

6.3.1. Week Starting May 6th, 2019

The next clustering result we describe is for the week starting on the 6th of May, 2019 (See Figure 11). The average silhouette score for this result is 0.630. The silhouette plot in Figure 11a suggests a less optimal clustering. This plot indicates that some observations would seemingly belong to clusters other than the one they are in; these observations have a negative silhouette width value.



(a) Silhouette (b) Scatter

(c) Map (d) CART Feature Importance

Figure 11: L3 Station Clusters - (06-MAY-2019)

A less than optimal clustering is confirmed by observing the scatter plot of Figure 11b. Some observations in cluster 1 seem to be outliers. The cluster cohesion is not as prevalent as cluster 0's. The cluster station members of cluster 1 in the map of Figure 11c are no longer generally located in the lower half of the province. Cluster 1 station members are mostly located at the province's boundary edges with this result. The important features outlined in the bar chart of Figure 11d no longer includes Wednesday kWh totals. The most important feature is now kWh power transfers to vehicles on Saturday.

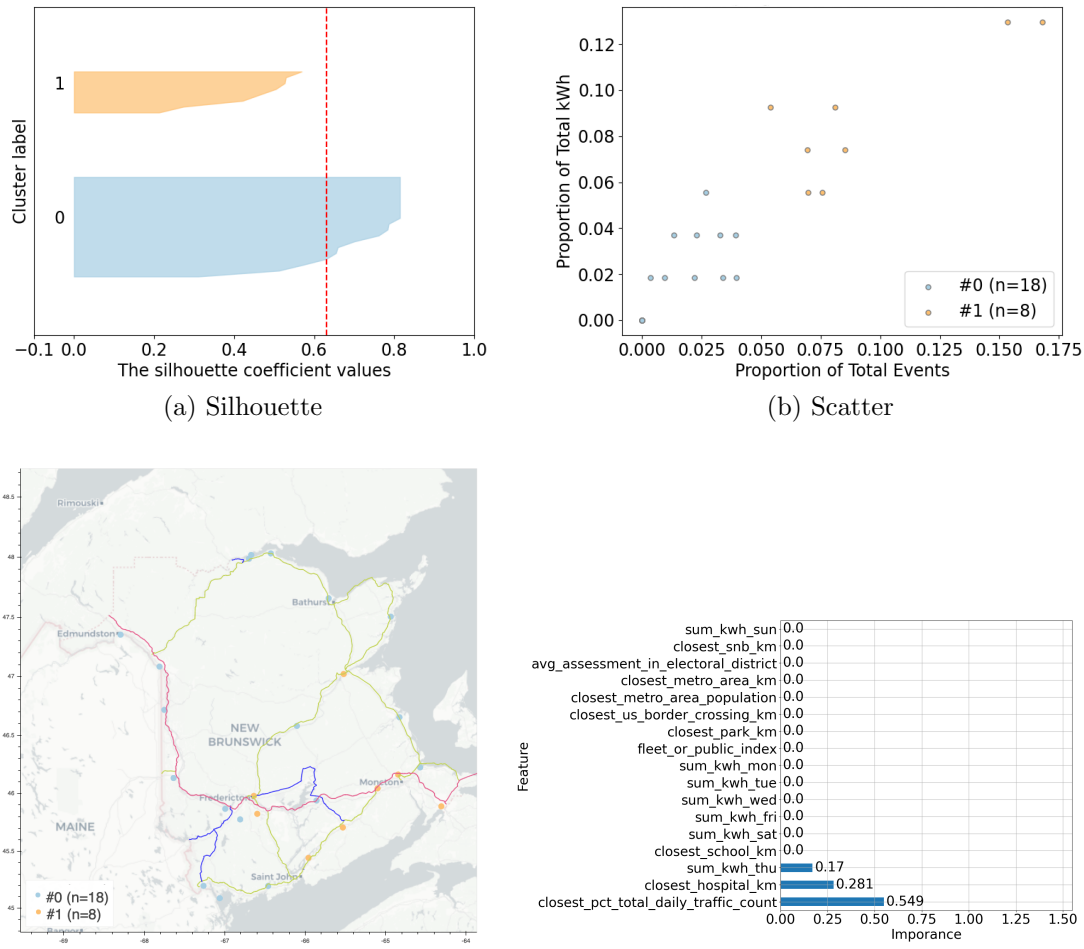
6.4. High Utilization in South-East of Province

The clustering results for the next week in the ordered list of results is unique in the sense that neighbouring results have less similarities when considering spatial and feature importance aspects to explain the clusters. This clustering result is described in detail next.

6.4.1. Week Starting December 2nd, 2019

For the clustering result of the week starting on December 2nd, 2019, the average silhouette score is 0.630. The silhouette plot in Figure 12a and the scatter plot in Figure 12b suggest that perhaps utilizing 3 clusters would result in a better partitioning of the stations. We can observe in the scatter plot of Figure 12b that cluster 1 - the higher station utilization cluster - seems to have large "within" dissimilarities, which leads to the lower silhouette widths that are observed in the silhouette plot for this cluster. In this result, it is apparent from the scatter plot (Figure 12b) that cluster 1 could be split in two smaller but well separated clusters.

The cluster station members of cluster 1 in the map of Figure 12c are generally located in the South-East section of the province. The important features outlined in the bar chart of Figure 12d indicate that the traffic counts for the closest traffic counter is the most important variable for predicting station cluster membership for the purpose of explaining the clusters using supplementary data.



(a) Silhouette

(b) Scatter

(c) Map

(d) CART Feature Importance

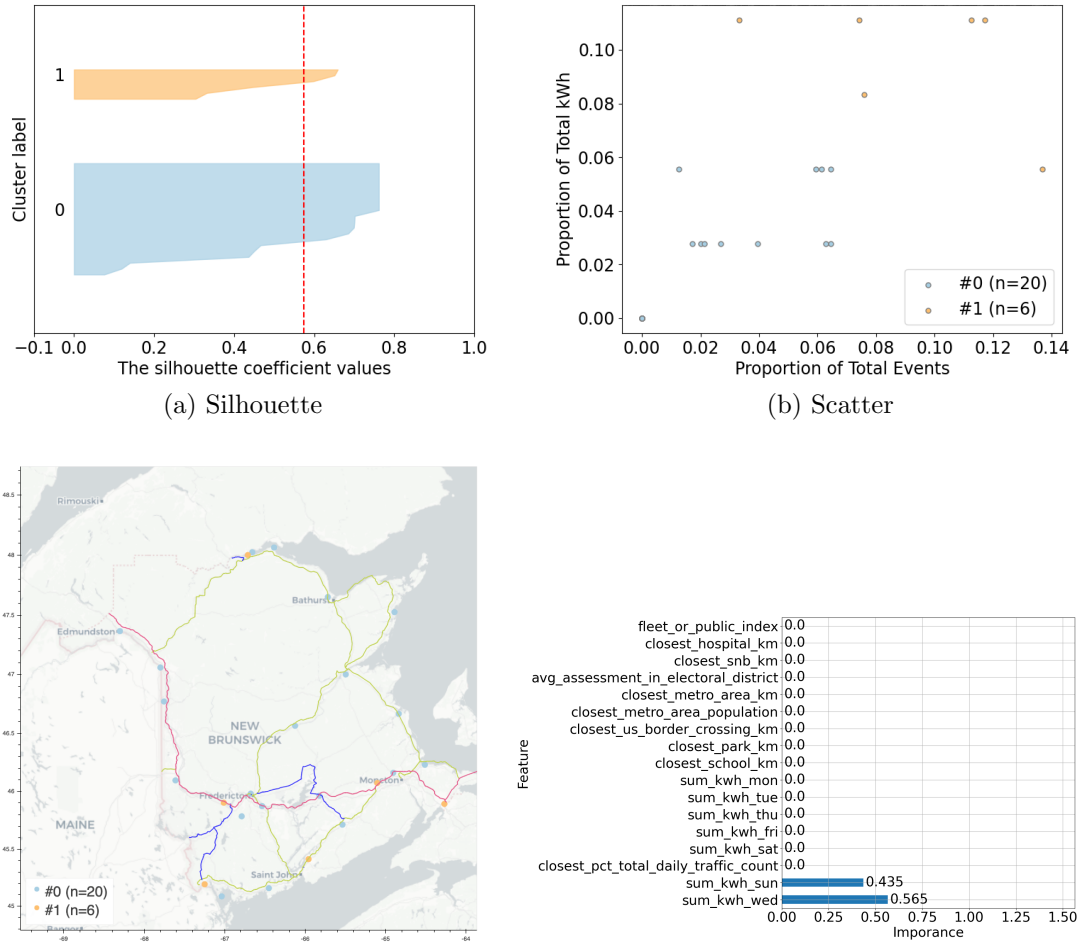
Figure 12: L3 Station Clusters - (02-DEC-2019)

6.5. Lowering Utilization Rates

The clustering results for the next 2 weeks in the ordered list of results have some interesting properties. There does not seem to be an obvious and common spatial pattern in these two clustering results. The number of cluster members for the high utilization cluster (cluster 1) decreases when comparing the week starting on November 04$^{\text{th}}$, 2019 to the week starting April 01$^{\text{st}}$, 2019. Temporally, the sum of kWh transferred to vehicles on Wednesday is in the list of important features for both results. The clustering result for this group's representative week is described in detail next.

6.5.1. Week Starting April 1st, 2019

In the result for the week starting on April 1$^{\text{st}}$, 2019, the average silhouette score is 0.575. The silhouette plot in Figure 13a and the average silhouette score value suggest a reasonable structure in the data has been found. Observing the scatter plot of Figure 13b, we can see that the observations in each cluster are more dispersed than in the reference result.



(a) Silhouette (b) Scatter

(c) Map (d) CART Feature Importance

Figure 13: L3 Station Clusters - (01-APR-2019)

The cluster station members of cluster 1 in the map of Figure 13c are generally located in the lower half of the province. The important features outlined in the bar chart of Figure 13d highlight the total kWh transferred to vehicles on Wednesday as the most important feature.

6.6. Low Utilization Rates Across The Province

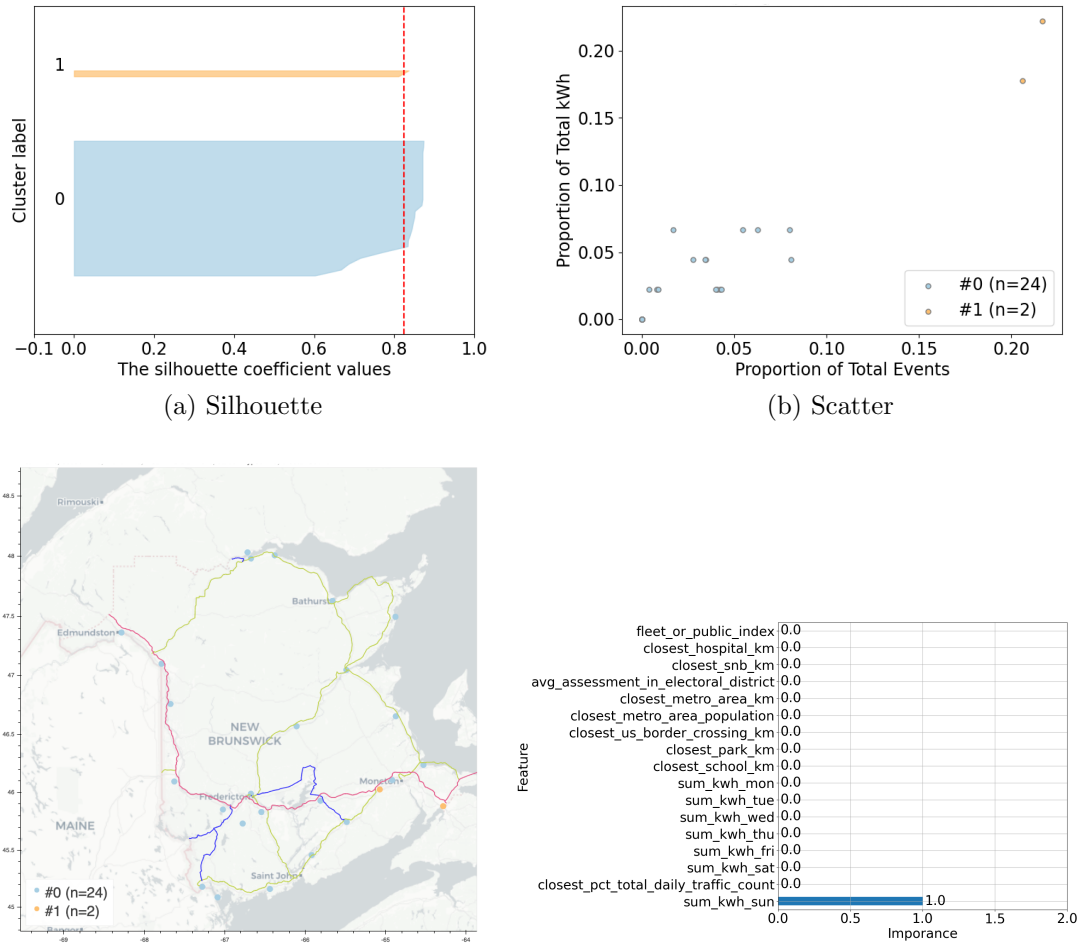
The clustering results for the last 2 weeks in the ordered list of clustering results both depict a majority grouping of low utilization EV stations across the province. The sum of kWh transferred to vehicles at the end of the week (i.e. Friday and Sunday) are top important features in these clustering results. The clustering result for this group's representative week is described in detail next.

6.6.1. Week Starting October 28th, 2019

We can see from Figure 14a and the average silhouette score of 0.825 that a strong structure in the data has been found for the week starting on October 28$^{\text{th}}$, 2019. Like all clustering results, stations are grouped in terms of relatively higher and lower utilization

rates. In the scatter plot of Figure 14b, we visually observe high intra-cluster similarity and inter-cluster separation. We can also observe that cluster 1, the cluster with relatively higher utilization rates, has 2 station members only and are located in the lower-right of the province according to the map in Figure 14c. The most important feature outlined in the bar chart of Figure 14d is total kWh transferred to vehicles on Sunday.

Every cluster validity index depicted in Table 3 for this result indicates a superior grouping of stations relative to our reference week.



(a) Silhouette (b) Scatter

(c) Map (d) CART Feature Importance

Figure 14: L3 Station Clusters - (28-OCT-2019)

6.7. Summary of Observations

As can be observed in Figures 10 to 14 of the previous sections, the decreasing relative similarity of clustering results is especially noticeable when visually comparing the silhouette and scatter plots for the week of May 27th with the same visualizations in other weeks and doing so in a step-wise fashion down the ranked list of results. As we move further away from the reference week, the spatial and feature importance aspects of the clustering results change. The grouping of weeks at the beginning and end of the ordered results list are

quite apparent. However, in the middle of the list (e.g. weeks starting 04-NOV-2019 and 01-APR-2019) the spatial-temporal characteristics of the grouping of weeks are less obvious. Although helpful in presenting and thinking about the results, there is a level of subjectivity in grouping the weeks in this manner. A subjectivity that, as mentioned previously, is also present in assessing the quality of individual clustering results in general.

Individual cluster validity index calculations embed implicit trade-offs on what is prioritized when expressing inter-cluster separation, intra-cluster homogeneity, density, and compactness as one numeric value. One can view the various indices as averages where a certain precision is lost in the summary. This can lead to situations where one index will suggest a better clustering relative to another grouping and another index will inverse this assessment.

The results highlighted in the case study provided in this section demonstrate that given a clustering result of interest, a process of objectively highlighting and recommending similar clustering results can indeed be automated in order to support the practitioner in evaluating how structure in data persists over multiple time slices in a data set with temporal properties. The relative ranking of similar clustering results this approach affords makes it easy to objectively identify similar station groupings over multiple weeks based on a reference week. Using real-world, contextualized data and supporting the practitioner in analytical results exploration can help reduce cognitive demand in results interpretation and improve downstream modeling performance. Not highlighted in the case study, are the clustering results for other a priori selected temporal partitions in the data, which are also available as reference points for exploring monthly or seasonal clustering similarities.

As a final remark on the results, Appendix F demonstrates how the weekly groupings obtained in the case study can provide useful insights in planning and expanding infrastructure allocation. As outlined in the appendix, utilizing the top 5 stations with the most kWh transferred to vehicles per station grouping can inform long-term planning and investment decisions.

7. Conclusions and Future Work

A broad EV adoption scenario will require adequate public charging infrastructure. An understanding of EV charging patterns at public charging stations is crucial to foster adoption while managing costs and optimizing placement of charging infrastructure. The outcomes of this research is believed to provide useful insights in planning and expanding infrastructure allocation (See Appendix F for this discussion). To optimize operations, EV station operators often seek market-related insights. EV charging station clustering can reveal useful segmentations in service consumption patterns. The insights offered by this framework can assist station operators and policy makers in the development of consumer engagement strategies, demand forecasting tools and in the design of more sophisticated tariff systems.

Capital investments in public charging infrastructure involves the use of public funds and necessitates informed decision making. Identifying similar station utilization patterns over multiple weeks can be useful planning information for station operators. As demonstrated in our case study, the results produced with our platform can help identify and explain over or under-utilized EV charging stations in addition to other patterns of use in the context of time. These additional insights can assist with capital investment decisions. Descriptive analytics provide a good framework to guide the introductory stages of information processing.

Advanced stages of descriptive analytics can include grouping sets of similar and dissimilar objects into clusters. This process is meant to acquire an initial understanding of historical data and organize it for advance analysis.

Although clustering has become a routine analytical task in many research domains, it remains arduous for practitioners to select a good algorithm with adequate hyperparameters and to assess the quality of clustering and the consistency of identified structures over various temporal slices of data. The process of clustering data is often an iterative, lengthy, manual and cognitively demanding task. The subjectivity in determining the level of "success" that unsupervised learning approaches are able to achieve and the required expert knowledge during the modeling phase suggest that a human-in-the-loop process of supporting the practitioner during this activity would be beneficial. Ascertaining whether a particular clustering of data is meaningful or not requires expertise and effort. Doing this for multiple results on data that has been sliced by weekly, monthly or seasonal partitions prior to applying the clustering algorithm would be very time consuming. Manually identifying one meaningful result of interest and then having an automated mechanism to select similar results is extremely useful in reducing the amount of effort required to identify avenues that merit further analysis and assist in downstream analytical tasks such as improving regression or classification model performance.

Future work will examine to what extent the weekly clustering results obtained with our platform improves downstream regression modeling results. Specifically, we will explore whether the Euclidean distances and clusters obtained in the case study can improve the predictive performance of a baseline regression model for predicting peak day of week kWh. Additionally, framing the creation of the initial ranked list of results as a multiple-criteria decision-making (MCDM) problem will be included in this work.

Acknowledgements

This work summarizes, extends and further disseminates core concepts originating from the first author's master's program thesis. The authors of this paper would like to thank the New Brunswick Power Corporation in addition to the New Brunswick Department of Transportation and Infrastructure for providing the data leveraged in this work.

Appendix A. Computation Environment Deployment Details

Table A.4: Computation Environment - Docker Container Configurations

Docker Host

Category	Value
OS	Fedora Core 32
CPU Model	Intel(R) Xeon(R) CPU E5-2660 0 @ 2.20GHz
Core Count	32
RAM	28GB

Data Analytics Container

Category	Value
OS	Fedora Core 31
Core Count	20
RAM	20GB
Spark Master	local[15]
Spark Driver Mem.	2GB
Spark Executor Mem.	15GB

Tracking Container

Category	Value
OS	Fedora Core 31
Core Count	2
RAM	2GB

Persistence Container

Category	Value
OS	Fedora Core 31
Core Count	4
RAM	4GB

Query Container

Category	Value
OS	Fedora Core 31
Core Count	2
RAM	2GB

Appendix B. EVStationSIM Platform Software Components

Table B.5: Software Used to Implement the EVStationSIM Platform

Data Analytics	*Analytics Engine*	**Apache Spark** is an open source analytics engine for big data and machine learning. The engine offers implicit data parallelism, fault tolerance, supports multiple programming languages and enables executing data engineering, data science, and machine learning tasks on single or multiple node clusters. The **PySpark** interface for Spark in Python enables the programmer to write Spark applications using Python language APIs.
	Data Manipulation	**Pandas** is an open source data analysis and manipulation library written in Python. Pandas is a widely used in data data analysis and machine learning tasks and works well with many other data science modules inside the Python ecosystem.
	Scientific Computing	**NumPy** is an open source library for scientific computing in Python providing support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these structures. This fundamental library also enables basic statistical operations, random simulation in addition to other functionality for working with numerical data. **Scipy**: is an open source scientific computation library that extends NumPy and provides additional utility functions for optimization, statistics and signal processing.
	Machine Learning	**Scikit-Learn** is an open source library for machine learning in Python. The library provides machine learning and statistical modeling functionality such as classification, regression, clustering and dimensionality reduction.
	Notebook Interface	**Jupyter** is an open source web-based interactive development environment for notebooks, code, and data.
Data Visualization	*Visualization Package*	**Matplotlib** is an open source plotting library for the Python programming language. The library enables the creation of static, animated, or interactive visualizations and provides an API for embedding plots into applications.
Tracking	*ML Life Cycle*	**MLflow** is an open source software package which streamlines the machine learning development lifecycle and addresses the common challenges of experimentation, reproducibility and deployment in machine learning projects.
Persistence	*RDBMS*	**PostgreSQL** is an open source relational database management system which is highly stable and backed by more than 30 years of development in the open-source community.
Query	*RESTful Interface*	**Falcon** is a minimalist Web Server Gateway Interface (WSGI) library for building fast web APIs and application backends.
Containerization	*Packaging*	**Docker Engine** is a software framework that uses OS-level virtualization to run applications on servers and the cloud. Containerization is a form of virtualization where a packaged application runs in an isolated user space called a container and everything required to run the application is encapsulated and isolated in the container.

Appendix C. Example Stage Element Implementation

Figure C.15 provides an example feature generation and selection stage element which was described in Section 4.3 and in Item (4) of Figure 8. In this stage, once feature generation and selection are complete, the transformed data files are partitioned using a priori selected temporal granularities (e.g. weekly, monthlyor seasonal). In this example, the parameterized ***create_batch_ranges.py*** Python script is executed from a Bash script. The platform workflow is comprised of multiple parameterized Python scripts like this one. The script in this example creates weekly temporal slices of an input data set. Noteworthy elements of this script are outlined below.

- **Line 8** outlines the individual weekly slices of data we are interested in creating. The list of weeks in this example covers a little over a month.
- **Lines 17 to 35** defines a function that is used to call a parameterized Python script of interest. Parameters include the start date of the week we are interested in creating a temporal slice of the data for *($WEEK)*, the input data *("$HOME"/data/nb_power/...)*, the output directory *("$HOME"/data/nb_power/work)* and whether we should track logging and artefact information inside MLflow *($TRACK_EXPERIMENTS)*.
- **Lines 40 to 51** defines a function that iterates over each week defined in **Line 8** with appropriate parameters and calls the function defined in **Lines 17 to 35**.
- **Line 53** launches the process to create weekly batch ranges. The output of this process is 6 data files. One file for each week defined in **Line 8**. These individual files are used downstream in the workflow in tasks such as clustering.

```
#!/bin/bash
# stop on script errors
# set -e

##############################################
# To manage which parts of the workflow to run
##############################################
WEEKS_2019=(1-APR-2019 8-APR-2019 15-APR-2019 22-APR-2019 29-APR-2019 6-MAY-2019)
TRACK_EXPERIMENTS=True
REDIR_OUT=True
WEEKLY_DO_CREATE_BACH_RANGES=true

##############################################
# To create temporal slices of clustering
# algorithm input data
##############################################
create_batch_ranges() {
    echo "Running create_batch_ranges on recharge report files"
    echo Year : "${YEAR}", Season : "${SEASON}", Month : "${MONTH}", Week : "${WEEK}"
    echo "..."
    python ./pre_processing/create_batch_ranges.py \
    --master local[2] \
    --driver_memory 2g \
    --executor_memory 2g \
    --track_experiment $TRACK_EXPERIMENTS  \
    --redir_out $REDIR_OUT \
    --experiment create_batch_ranges \
    --i_input "${HOME}"/data/nb_power/work/feat_eng_rech_report.parquet \
    --o_output_dir "${HOME}"/data/nb_power/work/ \
    --i_year $YEAR \
    --i_season $SEASON \
    --i_month $MONTH \
    --i_week_starting $WEEK
    echo "..."
}

##############################################
# To create weekly temporal slices of data
##############################################
run_weeky_batch_ranges(){
    if [ "${WEEKLY_DO_CREATE_BACH_RANGES}" = true ]; then
        echo "Running weekly batch range creation"
        YEAR=2019
        SEASON=na
        MONTH=na
        for WEEK in "${WEEKS_2019[@]}"
        do
            create_batch_ranges
        done
    fi
}

run_weeky_batch_ranges
```

Figure C.15: Parameterized ***create_batch_ranges.py*** Python Script Execution

Appendix D. Query Request - Station Groupings

(1) First, the practitioner requests ranked station clustering results for either L2 or L3 station types. Here the query parameters in the URL indicate that L3 station types are of interest with a silhouette score ranging from 0.4 to 0.6.

(2) The system then returns a sorted list of clustering results ordered by silhouette score. We see that the top result, according to this silhouette score range, is the clustering results for the week of May 27th, 2019 where the cluster count is 2. This particular record contains all cluster validity index values and a link to similar clustering results.

(3) Other clustering results, ranked in descending order by silhouette score, are also available for exploration.

Figure D.16: Search for Interesting Result

Appendix E. Query Response - Similar Weekly Groupings

(1) From the sorted list of clustering results ordered by silhouette score list, the practitioner selects one result as the reference result for which comparable results are desired.

(2) Details of the reference clustering result are outlined with corresponding analytical artefacts.

(3) & (4) A sorted list of comparable clustering results, ordered by Euclidean distance, is also available to the user. This list contains result-specific artefacts such as scatter plots, station cluster membership maps, silhouette plots and CART feature importance plots.

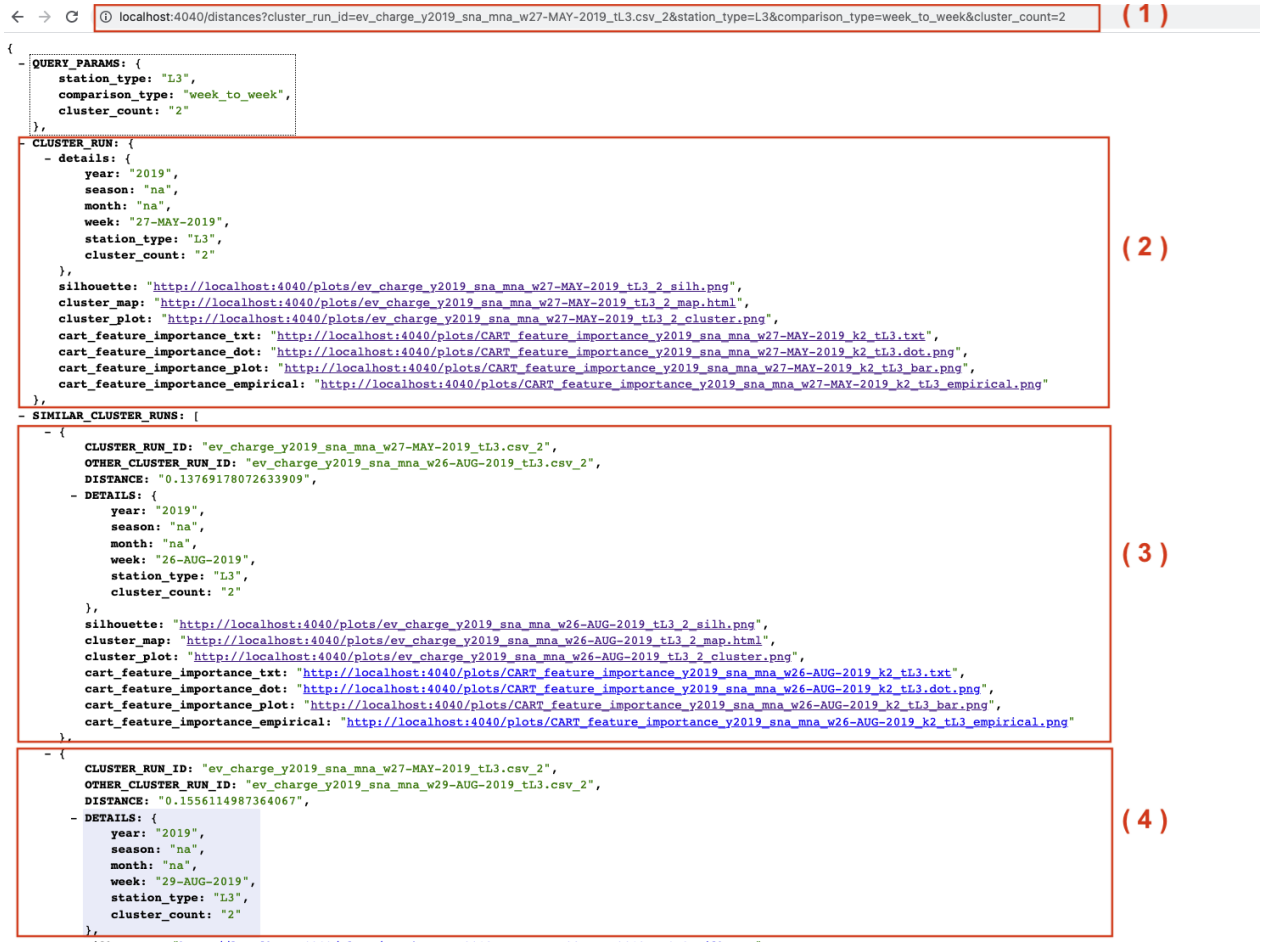


Figure E.17: Similar Results to Week Starting May 27, 2019

Appendix F. Weekly Groupings - Top 5 Station Members

Tables F.6 to F.16 list the top 5 stations with the highest total kWh transferred to vehicles per cluster during the weeks included in our case study. Ranking changes relative to the reference week are illustrated with the following symbols : (✢) indicates a new station not seen in the reference week's top 5 list, (↗) indicates a move up in ranking, (→) no changes in ranking and finally, (↘) indicates a drop in raking.

In the relatively higher station utilization rate cluster of cluster 1, Irving, Salisbury is often in the top 5 stations with the most kWh transferred to vehicles overall followed by Irving, Aulac and Grey Rock, Edmundston. As EV adoption increases in the province, station grouping trend observations such as this can be identified and applied to long-term planning and investment decisions.

Table F.6: Top 5 Stations per Cluster - (27-MAY-2019)

Station Name	Business Location	Total Sum kWh	Cluster
NBC-10004	Irving, Aulac	152.879	1
NBC-10001	Irving, Salisbury	100.63	1
NBC-10007	Mountain Road	97.092	1
NBC-10008	Grey Rock, Edmundston	94.347	1
NBC-10009	Murray's, Beardsley	88.679	1

Station Name	Business Location	Total Sum kWh	Cluster
NBC-10012	Atlantic Host, Bathurst	69.077	0
NBC-10003	Irving, Grand Falls	53.361	0
NBC-10006	Acorn, Lake George	51	0
NBC-10002	Irving, Youngs Cove	45.48	0
NBC-10020	Garcelon Civic Center	42.173	0

Table F.7: Top 5 Stations per Cluster - (26-AUG-2019)

Station Name	Business Location	Total Sum kWh	Cluster
NBC-10001	↗ Irving, Salisbury	261.121	1
NBC-10006	✢ Acorn, Lake George	249.519	1
NBC-10009	↗ Murray's, Beardsley	245.675	1
NBC-10008	→ Grey Rock, Edmundston	206.073	1
NBC-10026	✢ Town of Shediac	197.387	1

Station Name	Business Location	Total Sum kWh	Cluster
NBC-10004	✢ Irving, Aulac	128.865	0
NBC-10012	↘ Atlantic Host, Bathurst	116.36	0
NBC-10021	✢ Shoppers Drug Mart, Sussex	100.068	0
NBC-10003	↘ Irving, Grand Falls	97.377	0
NBC-10013	✢ Saint Quentin	93.521	0

Table F.8: Top 5 Stations per Cluster - (29-AUG-2019)

Station Name	Business Location	Total Sum kWh	Cluster
NBC-10009	↗ Murray's, Beardsley	186.57	1
NBC-10006	✢ Acorn, Lake George	160.428	1
NBC-10001	↘ Irving, Salisbury	158.952	1
NBC-10004	↘ Irving, Aulac	129.71	1
NBC-10021	✢ Shoppers Drug Mart, Sussex	119.355	1

Station Name	Business Location	Total Sum kWh	Cluster
NBC-10008	✢ Grey Rock, Edmundston	74.164	0
NBC-10002	↗ Irving, Youngs Cove	66.513	0
NBC-10020	↗Garcelon Civic Center	65.854	0
NBC-10026	✢ Town of Shediac	65.645	0
NBC-10015	✢ Northumberland Square Mall	57.653	0

Table F.9: Top 5 Stations per Cluster - (06-MAY-2019)

Station Name	Business Location	Total Sum kWh	Cluster
NBC-10015	✢ Northumberland Square Mall	105.749	1
NBC-10008	↗ Grey Rock, Edmundston	69.111	1
NBC-10007	→ Mountain Road	68.182	1
NBC-10001	↘ Irving, Salisbury	52.772	1
NBC-10004	↘ Irving, Aulac	48.745	1

Station Name	Business Location	Total Sum kWh	Cluster
NBC-10016	✢ Osprey Truck Stop	17.337	0
NBC-10005	✢ Irving, Lincoln	15.807	0
NBC-10002	↗ Irving, Youngs Cove	13.5	0
NBC-10010	✢ Johnson's Pharmacy	13.442	0
NBC-10021	✢ Shoppers Drug Mart, Sussex	12.746	0

Table F.10: Top 5 Stations per Cluster - (17-JUN-2019)

Station Name	Business Location	Total Sum kWh	Cluster
NBC-10008	↗ Grey Rock, Edmundston	171.968	1
NBC-10009	↗ Murray's, Beardsley	143.419	1
NBC-10005	✢ Irving, Lincoln	97.156	1
NBC-10003	✢ Irving, Grand Falls	64.398	1
NBC-10010	✢ Johnson's Pharmacy	61.931	1

Station Name	Business Location	Total Sum kWh	Cluster
NBC-10020	↗ Garcelon Civic Center	54.24	0
NBC-10019	✢ Irving, Quispamsis	44.94	0
NBC-10004	✢ Irving, Aulac	34.569	0
NBC-10025	✢ Visitor Information Center, Caraquet	21.554	0
NBC-10016	✢ Osprey Truck Stop	17.756	0

Table F.11: Top 5 Stations per Cluster - (07-OCT-2019)

Station Name	Business Location	Total Sum kWh	Cluster
NBC-10006	Acorn Lake George	165.795	1
NBC-10008	Grey Rock, Edmundston	122.943	1
NBC-10009	Murray's, Beardsley	122.616	1
NBC-10007	Mountain Road	118.928	1
NBC-10004	Irving, Aulac	115.62	1

Station Name	Business Location	Total Sum kWh	Cluster
NBC-10020	Garcelon Civic Center	51.529	0
NBC-10013	Saint Quentin	34.286	0
NBC-10023	O'Neill Arena	30.571	0
NBC-10022	Quality Inn, Campbellton	29.558	0
NBC-10002	Irving, Youngs Cove	28.557	0

Table F.12: Top 5 Stations per Cluster - (02-DEC-2019)

Station Name	Business Location	Total Sum kWh	Cluster
NBC-10024	NB Power, Fredericton	151.696	1
NBC-10015	Northumberland Square Mall	138.531	1
NBC-10007	Mountain Road	76.959	1
NBC-10001	Irving, Salisbury	73.232	1
NBC-10004	Irving, Aulac	68.446	1

Station Name	Business Location	Total Sum kWh	Cluster
NBC-10025	Visitor Information Center, Caraquet	35.658	0
NBC-10012	Atlantic Host, Bathurst	35.386	0
NBC-10008	Grey Rock, Edmundston	30.647	0
NBC-10017	Richibucto	29.424	0
NBC-10011	Tracadie	24.252	0

Table F.13: Top 5 Stations per Cluster - (04-NOV-2019)

Station Name	Business Location	Total Sum kWh	Cluster
NBC-10009	Murray's, Beardsley	96.329	1
NBC-10001	Irving, Salisbury	95.546	1
NBC-10008	Grey Rock, Edmundston	87.194	1
NBC-10021	Shoppers Drug Mart, Sussex	62.15	1
NBC-10019	Irving, Quispamsis	61.582	1

Station Name	Business Location	Total Sum kWh	Cluster
NBC-10017	Richibucto	16.705	0
NBC-10012	Atlantic Host, Bathurst	16.086	0
NBC-10007	Mountain Road	12.513	0
NBC-10002	Irving, Youngs Cove	7.057	0
NBC-10010	Johnson's Pharmacy	6.979	0

Table F.14: Top 5 Stations per Cluster - (01-APR-2019)

Station Name	Business Location	Total Sum kWh	Cluster
NBC-10025	Visitor Information Center, Caraquet	77.505	1
NBC-10001	Irving, Salisbury	66.328	1
NBC-10004	Irving, Aulac	63.805	1
NBC-10020	Garcelon Civic Center	42.991	1
NBC-10019	Irving, Quispamsis	42.066	1

Station Name	Business Location	Total Sum kWh	Cluster
NBC-10007	Mountain Road, Moncton	36.524	0
NBC-10026	Town of Shediac	35.608	0
NBC-10009	Murray's, Beardsley	34.709	0
NBC-10015	Northumberland Square Mall	33.706	0
NBC-10013	Saint Quentin	22.457	0

Table F.15: Top 5 Stations per Cluster - (28-OCT-2019)

Station Name	Business Location	Total Sum kWh	Cluster
NBC-10004	Irving, Aulac	191.299	1
NBC-10001	Irving, Salisbury	181.971	1

Station Name	Business Location	Total Sum kWh	Cluster
NBC-10025	Visitor Information Center, Caraquet	71.36	0
NBC-10015	Northumberland Square Mall	70.484	0
NBC-10003	Irving, Grand Falls	55.306	0
NBC-10021	Shoppers Drug Mart, Sussex	48.136	0
NBC-10022	Quality Inn, Campbellton	38.062	0

Table F.16: Top 5 Stations per Cluster - (23-DEC-2019)

Station Name	Business Location	Total Sum kWh	Cluster
NBC-10015	Northumberland Square Mall	439.331	1

Station Name	Business Location	Total Sum kWh	Cluster
NBC-10012	Atlantic Host, Bathurst	164.358	0
NBC-10003	Irving, Grand Falls	149.745	0
NBC-10013	Saint Quentin	89.708	0
NBC-10008	Grey Rock, Edmundston	67.787	0
NBC-10017	Richibucto	62.896	0

References

[1] New Brunswick Electric Vehicle Advisory Group, An electric vehicle roadmap for new brunswick a discussion document for public and stakeholder engagement (2016).

[2] P. Ashkrof, G. Homem de Almeida Correia, B. van Arem, Analysis of the effect of charging needs on battery electric vehicle drivers' route choice behaviour: A case study in the Netherlands, Transportation Research Part D: Transport and Environment 78 (2020) 102206. doi:10.1016/j.trd.2019.102206.
URL https://linkinghub.elsevier.com/retrieve/pii/S1361920919309757

[3] S. L. Monaca, L. Ryan, The State of Play in Electric Vehicle Charging Services: Global Trends with Insight for Ireland (2018).
URL https://esipp.ie/images/upload/UCDpolicyreport_EVChargingInfrastructureNov2018.pdf

[4] M. Straka, L. Buzna, Clustering algorithms applied to usage related segments of electric vehicle charging stations, Transportation Research Procedia 40 (2019) 1576–1582.

[5] F. Iglesias, W. Kastner, Analysis of similarity measures in times series clustering for the discovery of building energy patterns, Energies 6 (2) (2013) 579–597.

[6] M. Oliveira, 3 Reasons Why AutoML Won't Replace Data Scientists Yet, (Accessed 2021-07-23) (2019).
URL https://www.kdnuggets.com/3-reasons-why-automl-wont-replace-data-scientists-yet.html/

[7] J. Bae, T. Helldin, M. Riveiro, S. Nowaczyk, M.-R. Bouguelia, G. Falkman, Interactive Clustering: A Comprehensive Review, ACM Computing Surveys 53 (1) (2020) 1–39. doi:10.1145/3340960.
URL https://dl.acm.org/doi/10.1145/3340960

[8] Pollution Probe and The Delphi Group, Guide to electric vehicle charging in multi-unit residential buildings (2020).

[9] E. Abotalebi, D. M. Scott, M. R. Ferguson, Why is electric vehicle uptake low in atlantic canada? a comparison to leading adoption provinces, Journal of Transport Geography 74 (2019) 289–298.

[10] Statistics Canada, Population and dwelling count highlight tables, 2016 census, (Accessed 2021-06-16) (n.d.).
URL https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Tableau.cfm?Lang=Fra

[11] Statistics Canada, Vehicle registrations, by type of vehicle, (Accessed 2021-06-16) (n.d.).
URL https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=2310006701&pickMembers%5B0%5D=1.5&cubeTimeFrame.startYear=2015&cubeTimeFrame.endYear=2019&referencePeriods=20150101%2C20190101

[12] E. Abotalebi, D. M. Scott, M. R. Ferguson, Can canadian households benefit economically from purchasing battery electric vehicles?, Transportation Research Part D: Transport and Environment 77 (2019) 292–302.

[13] Transport Canada, Road transportation, (Accessed 2021-07-14) (n.d.).
URL https://tc.canada.ca/en/corporate-services/policies/road-transportation-0

[14] A. S. Al-Ogaili, T. J. T. Hashim, N. A. Rahmat, A. K. Ramasamy, M. B. Marsadek, M. Faisal, M. A. Hannan, Review on scheduling, clustering, and forecasting strategies for controlling electric vehicle charging: challenges and recommendations, Ieee Access 7 (2019) 128353–128371.

[15] E. Y. Shchetinin, Cluster-based energy consumption forecasting in smart grids, in: International Conference on Distributed Computer and Communication Networks, Springer, 2018, pp. 445–456.

[16] E. Rendón, I. Abundez, A. Arizmendi, E. M. Quiroz, Internal versus external cluster validation indexes, International Journal of computers and communications 5 (1) (2011) 27–34.

[17] Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu, Understanding of internal clustering validation measures, in: 2010 IEEE International Conference on Data Mining, IEEE, 2010, pp. 911–916.

[18] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. PéRez, I. Perona, An extensive comparative study of cluster validity indices, Pattern Recognition 46 (1) (2013) 243–256.

[19] E. Xydas, C. Marmaras, L. M. Cipcigan, N. Jenkins, S. Carroll, M. Barker, A data-driven approach for characterising the charging demand of electric vehicles: A uk case study, Applied energy 162 (2016) 763–771.

[20] C. Sun, T. Li, S. H. Low, V. O. Li, Classification of electric vehicle charging time series with selective clustering, Electric Power Systems Research 189 (2020) 106695.

[21] A. Singh, A. Yadav, A. Rana, K-means with three different distance metrics, International Journal of Computer Applications 67 (10) (2013).

[22] S. Zolhavarieh, S. Aghabozorgi, Y. W. Teh, A review of subsequence time series clustering, The Scientific World Journal 2014 (2014).

[23] F. H. Kuwil, Ü. Atila, R. Abu-Issa, F. Murtagh, A novel data clustering algorithm based on gravity center methodology, Expert Systems with Applications 156 (2020) 113435.

[24] S. Khedairia, M. T. Khadir, A multiple clustering combination approach based on iterative voting process, Journal of King Saud University-Computer and Information Sciences (2019).

[25] J. Zhang, J. Yan, Y. Liu, H. Zhang, G. Lv, Daily electric vehicle charging load profiles considering demographics of vehicle users, Applied Energy 274 (2020) 115063.

[26] I. Vermeulen, J. R. Helmus, M. Lees, R. Van Den Hoed, Simulation of future electric vehicle charging behavior—effects of transition from phev to fev, World Electric Vehicle Journal 10 (2) (2019) 42.

[27] Y. Xiang, Z. Jiang, C. Gu, F. Teng, X. Wei, Y. Wang, Electric vehicle charging in smart grid: A spatial-temporal simulation method, Energy 189 (2019) 116221.

[28] Z. Wang, P. Jochem, W. Fichtner, A scenario-based stochastic optimization model for charging scheduling of electric vehicles under uncertainties of vehicle availability and charging demand, Journal of Cleaner Production 254 (2020) 119886.

[29] J. Yan, J. Zhang, Y. Liu, G. Lv, S. Han, I. E. G. Alfonzo, Ev charging load simulation and forecasting considering traffic jam and weather to support the integration of renewables and evs, Renewable energy 159 (2020) 623–641.

[30] M. Straka, P. De Falco, G. Ferruzzi, D. Proto, G. Van Der Poel, S. Khormali, L. Buzna, Predicting Popularity of Electric Vehicle Charging Infrastructure in Urban Context, IEEE Access 8 (2020) 11315–11327. doi:10.1109/ACCESS.2020.2965621.
URL https://ieeexplore.ieee.org/document/8955852/

[31] S. Hardman, A. Jenn, G. Tal, J. Axsen, G. Beard, N. Daina, E. Figenbaum, N. Jakobsson, P. Jochem, N. Kinnear, P. Plötz, J. Pontes, N. Refa, F. Sprei, T. Turrentine, B. Witkamp, A review of consumer preferences of and interactions with electric vehicle charging infrastructure, Transportation Research Part D: Transport and Environment 62 (2018) 508–523. doi:10.1016/j.trd.2018.04.002.
URL https://linkinghub.elsevier.com/retrieve/pii/S1361920918301330

[32] S. Maase, X. Dilrosun, M. Kooi, R. Van den Hoed, Performance of electric vehicle charging infrastructure: Development of an assessment platform based on charging data, World Electric Vehicle Journal 9 (2) (2018) 25.

[33] V. Komasilovs, A. Zacepins, A. Kviesis, C. Marinescu, I. Serban, Development of the web platform for management of smart charging stations for electric vehicles., in: VEHITS, 2018, pp. 595–599.

[34] GeoNB, Geonb web site, (Accessed 2021-04-17) (n.d.).
URL http://www.snb.ca/geonb1/e/index-E.asp

[35] H. Motoda, H. Liu, Feature selection, extraction and construction, Communication of IICM (Institute of Information and Computing Machinery, Taiwan) Vol 5 (67-72) (2002) 2.

[36] E. Hancer, B. Xue, M. Zhang, A survey on feature selection approaches for clustering, Artificial Intelligence Review 53 (6) (2020) 4519–4545.

[37] F. Nielsen, Hierarchical clustering, in: Introduction to HPC with MPI for Data Science, Springer, 2016, pp. 195–211.

[38] D. Kartsaklis, M. Sadrzadeh, S. Pulman, Separating disambiguation from composition in distributional semantics, in: Proceedings of the Seventeenth Conference on Computational Natural Language Learning, 2013, pp. 114–123.

[39] R. Richard, H. Cao, M. Wachowicz, An automated clustering process for helping practitioners to identify similar ev charging patterns across multiple temporal granularities, in: Proceedings of the 10th International Conference on Smart Cities and Green ICT Systems - SMARTGREENS,, INSTICC, SciTePress, 2021, pp. 67–77. doi:10.5220/0010485000670077.

[40] C. Boettiger, An introduction to docker for reproducible research, ACM SIGOPS Operating Systems Review 49 (1) (2015) 71–79.

[41] MMQGIS, Mmqgis plug-in, (Accessed 2021-10-29) (2021).
URL https://michaelminn.com/linux/mmqgis/

[42] P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of computational and applied mathematics 20 (1987) 53–65.