A Spatial-temporal Comparison of EV Charging Station Clusters Leveraging Multiple Validity Indices

René Richard[1 [0000−0002−1342−6225]], Hung Cao[1 [0000−0002−0788−4377]], and Monica Wachowicz[1,2 [0000−0002−4659−0101]]

[1] University of New Brunswick, Canada
{rene.richard, hcao3}@unb.ca
[2] RMIT University, Australia
monica.wachowicz@rmit.edu.au

Abstract. Decoupling vehicles from the immediate consumption of fossil fuels introduces new opportunities in supporting sustainable mobility. Fostering a shift from vehicles with internal combustion engines to Electric Vehicles (EV) often involves using publicly funded subsidies. Given early EV adoption challenges, some charging stations may be under-utilized, others will serve a disproportionate number of users. An understanding of EV charging patterns is crucial for optimizing charging infrastructure placement and managing costs. Clustering has been used in the energy domain to ensure service continuity and consistency. However, clustering presents challenges in terms of algorithm and hyperparameter selection in addition to pattern discovery validation. The lack of ground truth information, which could objectively validate results, is not present in clustering problems. Therefore, it is difficult to judge the effectiveness of different modelling decisions since there is no external validity measure available for comparison. This work proposes a clustering process that allows for the creation of relative rankings of similar clustering results that will assist practitioners in the smart grid sector. The approach supports practitioners by allowing them to compare a clustering result of interest against other similar groupings over multiple temporal granularities. The efficacy of this analytical process is demonstrated with a case study using real-world EV charging event data from charging station operators in Atlantic Canada.

Keywords: Agglomerative Hierarchical Clustering · EV Adoption · Charging Infrastructure Usage Patterns · Clustering Process · Cluster Validity Indices

1 INTRODUCTION

The trend of vehicle electrification is happening rapidly in many countries around the world. In spite of the pandemic-related worldwide downturn in car sales, new electric car registrations increased by 41% alongside $120 billion in consumer expenditures on electric vehicles (EV) in 2020 [9]. The International Energy Agency predicts that global EV stock will reach 145 million vehicles by 2030 in the Stated Policies Scenario and global EV fleet will reach 230 million vehicles by 2030 in the Sustainable Development Scenario [9]. The futuristic vision of advanced and modern urbanization is a core concept of a smart city, in which cutting-edge infrastructure is able to offer a high quality

of life for citizens and the sustainable management of natural resources. Adopting the usage of EVs is expected to improve air quality, provide sustainable mobility, mitigate greenhouse gas emissions, reduce urban noise pollution, and therefore contributes to this vision.

Building public charging infrastructure brings about high capital costs in addition to the usage of public funds to accelerate the transition to EVs. This necessitates smart decision-making at all stages of the adoption life-cycle. Given the challenges of early EV adoption, some charging stations may be under-utilized, others will serve a disproportionate number of users. Moreover, uncontrolled EV charging behaviors may cause numerous problems for existing power grids such as high load peaks, increased operational costs, degraded power quality, increased energy consumption, and the potential risk of power outages [31,3]. Therefore, reliable control of the EV charging behavior will be paramount for a successful mass market penetration. Clustering stations together based on usage patterns is an important and useful planning tool for operators. In addition, as the number of EVs increases, so does the demand for electricity and the possible strain on electrical grids. Utilities and other power generators need to prepare for increased demand. Accurate load forecasting is a tool that can help operators ensure service continuity and consistency.

Clustering is an unsupervised machine learning technique that assists practitioners in revealing hidden patterns and insights from a given dataset. In smart grid applications, this method has been used by practitioners to group similar consumers, categorize related energy consumption reports, forecast future demand, and grow EV adoption. Statistical and probabilistic models, built with data from EV charging stations having similar charging patterns, will reportedly have increased accuracy [29]. As a result, energy load projecting methods might perform better when applied to homogeneous clusters of EV stations as opposed to all stations. Hidden patterns in energy usage behavior are the key to improving services provided by utility companies, which are responsible for managing peaks and imbalances in EV charging infrastructure usage patterns [12].

Although clustering algorithms have been applied in many knowledge domains and applications, practitioners face the challenge of selecting the proper clustering algorithm with hyperparameter combination for their specific application. An additional concern includes evaluating the quality of clustering results. Moreover, it tends to require specialists to be able to assess and make sense of the clustering results due to the subjectivity found in deep expert knowledge. This is one of the main reasons why existing automated machine learning frameworks tend to focus on supervised learning tasks that require labeled data as input rather than unsupervised learning tasks that deal with unlabeled data [22]. Because the identification of the most similar clusters can be subjective, it usually requires different approaches to automate this process [23]. In addition, one of the challenges in clustering is finding the results that align with a practitioner's needs. In practice, there are several plausible clusters in complex datasets. What's more, practitioners may have different priorities and preferences. An unsupervised clustering algorithm has no way to intrinsically infer which clusters exhibit these desired priorities and preferences [5].

In spatial-temporal datasets (e.g. EV charging event datasets), evaluating the structure consistency of discovered clusters over different temporal granularities is normally

an arduous, manual and time-consuming activity. Several examples of metrics can be utilized to determine the structure consistency of the clusters such as inter-cluster homogeneity, inter-cluster separation, density, and uniform cluster sizes. Nevertheless, the question of how to select a particular clustering result that is more meaningful than another based on practitioner priorities and preferences, still heavily depends on the practitioner's expert knowledge. 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. Towards this challenge, this study explores whether, given the prospect of a clustering result of interest, a process of objectively highlighting and recommending similar clustering results can be automated in order to support practitioners in evaluating how clustering patterns persist over multiple temporal granularities, allowing practitioners to find meaningful clusters according to their preferences and priorities. This work aims to assist practitioners in identifying multiple clustering results of interest for different temporal partitions of the same data. Providing the practitioner with an initial ranked list of clustering results and a mechanism to determine clustering similarities can assist practitioners in downstream analytical tasks such as improving regression or classification model performance.

Consequently, a clustering process in which internal cluster validity indices are utilized to enable the identification of similar clustering results across various temporal slices of data is proposed. The main focus of this study is to support practitioners in identifying similar clustering results by using a reference result of interest and comparing this reference result with other results where all results are obtained from a-priori selected temporal partitions of the input data (i.e. weekly, monthly and seasonal paritions). To demonstrate the proposed approach, a case study using real-world charging event data from EV station operators in Atlantic Canada is utilized to evaluate our clustering process in identifying similar clusters of charging stations according to their usage patterns (e.g. high vs low utilization). This work is part of a larger ongoing research project. It continues the activities documented in [25] which examined charging events from EV charging stations exclusively. This paper extends this work by providing additional spatial context to the interpretation of the weekly clustering results. In addition to these enhanced results, supplementary background and clustering process details have been added.

The rest of the paper is organized as follows. In section 2, previous research work is described. Section 3 describes the background of this work. Section 4 describes the proposed clustering process underpinning our work. Section 5 provides a detailed description of the real-world EV charging event data and the end-to-end automated implementation of our proposed clustering process. In section 6, we discuss the results. Finally, section 7 concludes and indicates future research work.

2 RELATED WORK

Clustering techniques have played a significant role in finding new value and insights in many smart grid applications [26]. They are an essential tool in the pattern analysis process to discover energy usage behavior (i.e. the EV charging demand) in the energy domain [3]. For example, Straka and Buzna [29] carried out a comparison of the clus-

tering results from the k-means, hierarchical, and DBScan algorithms aiming to explore usage patterns related to segments of charging stations. This experiment is based on a dataset of 1700 charging stations distributed across the Netherlands with about 1 million charging transactions collected during a 4-year period. The clustering algorithms successfully identified four groups of EV charging stations characterized by distinct usage patterns. In [8], the authors analyzed a dataset of seven public smart charging stations located across the City of Rochester, US. These stations recorded the charging activities of vehicles during a period of 2-years. By applying the k-means clustering algorithm, they were able to identify different clustering patterns and their behaviors with respect to charging activity, parking without charging activity, and parking durations. Another example of using clustering algorithms to reveal the charging patterns from EV stations can be found in [31]. In this study, the k-means clustering technique is used to categorized EV user behavior into different groups and label them for further prediction purposes. This work is developed based on a dataset collected from more than 200 EV charging stations installed in public parking structures in many locations in Los Angeles, US.

The aforementioned research [29,8,31] had a common point that the clustering results are mainly analysed based on the time series and the temporal characteristics of the datasets. Indeed, Xiong et al. [31] mainly used the arrival and departure schedule that are fixed at certain timestamps with little variance to label the groups, while [29,8] mainly used timestamps of charging events and utilization or energy consumed (kWh) to compute the clusters.

Very few research works exploit the spatial component of EV-related datasets to enhance knowledge discovery [20]. Recent work by Kang et al. [14] used location-based service data to identify spatial-patterns of EV usage behavior in urban areas to characterize the distribution of home and charging station clusters as well as user charging preferences. From the literature, few attempts [11,13] have conducted spatial-temporal clustering to improve the integration of an EV fleet with power management and operations.

A common issue in clustering is how to objectively and quantitatively assess and analyse the results. From this, some important research questions emerge such as (Q1) How to use the spatial-temporal information from a given dataset to assist practitioners in understanding hidden patterns revealed in the clustering results? (Q2) How to interpret and make sense of the clustering results yielded from a large quantity of grouped data points? and (Q3) How to automatically identify similar patterns across multiple temporal granularities without manually inspecting the results one by one?

Cluster validation is an essential task in the clustering process since it aims is to compare clustering results and solve the question of optimal cluster count. Many internal validity indices have been proposed in the literature to evaluate the "success" level that a clustering algorithm can achieve in discovering the natural groupings in data without any class label information [24,18]. Currently, the majority of studies validating cluster results have been focused on the computation of individual cluster validity indices (CVI), which are normally selected to specify the relative performance of clustering results. For example, Arbelaitz et al. [4] perform a comparison of 30 CVIs using an experimental setup on multiple datasets with ground truth information to propose

the "best" partitioning. The optimal suggested number of partitions is defined as the one that is the most similar to the correct one measured by partition similarity measures. The authors found that noise and cluster overlap had the greatest impact on CVI performance. Some indices performed well with high dimensionality data sets and in cases where homogeneity of the cluster densities disappeared. The conclusion in this work suggests using several CVI to obtain robust results.

Sun et al. [30] proposed a time series clustering method using a modified Euclidean distance to group the similar charging tails from ACN-Data collected from smart EV charging stations. In this work, they evaluated their clustering results with Dynamic Time Warping distance (DTW) and Euclidean distance method using the silhouette coefficient. In [32], the Davies-Bouldin index is used to determine the best value for the cluster count parameter using the k-means algorithm.

All in all, the CVIs have been traditionally used for validation purposes. However, utilizing multiple CVIs together in combination with a proximity measure such as Euclidean distance has a strong potential to offer a new pairwise similarity measure that can enhance the comparison of clustering results by practitioners. This is also the key to answering the research questions (Q1), (Q2), (Q3) that we mention above. Certainly, this is not a common practice in data science as well as in the energy domain.

3 BACKGROUND

Partitioning data into groups based on internal and a-priori unknown schemes inherit in the data is a main concern of clustering. In this unsupervised learning approach, algorithms are presented with data instances having features describing each object but no information, or label, is given as to how instances should be grouped in terms of their similarity. Clustering plays an important role in discovering hidden patterns in a dataset. It has been utilized in the energy domain to group similar consumers and help predict future demand. Clustering can serve as a pre-processing step for other algorithms. For example, statistical models built with data from charging stations having similar charging patterns will reportedly have superior accuracy [29].

Many clustering algorithms have been developed. These have been broadly categorized into a handful of groupings in the literature based on aspects of the approach such as the partitioning criteria, clustering space, procedures used for measuring the similarity and whether samples belong strictly to one cluster or can belong to more clusters in differing degrees. A common grouping of clustering algorithms is *partitioning*, *density*, *grid* and *hierarchical* methods [10,19,2,21].

3.1 Partitioning Methods

Partitioning-based methods split data points into k partitions, where each partition represents a cluster. The data is split to optimize a certain, often distance-based, criterion function. Examples of commonly known partition-based methods include k-means and k-medoids [19]. The k-means clustering algorithm is easy to implement and is appropriate for large datasets. However, it has the disadvantage of being inappropriate for

clusters of different densities and being dependent on initial centroid values. Additionally, noisy data and outliers are problematic for this algorithm. Centroids can be tugged by outliers, or outliers might get their own cluster instead of being ignored. Similarly, the k-Medoids algorithm is easy to implement. The advantage of this algorithm is that it converges quickly and is less sensitive to outliers. However, it is also dependent on the initial set of medoids and can produce different clusterings on iterative runs [2]. Partitioning methods have the drawback of whenever a point is close to the center of another cluster; poor results are obtained due to overlapping.

3.2 Density Methods

Density-based methods group neighboring objects into clusters based on local density conditions instead of distance-based criterion. Groups are formed either according to the density of neighborhood objects or a density function. This class of methods interprets clusters as dense regions that are separated by low density noisy regions. Examples of commonly known density-based methods include DBSCAN, and OPTICS. This class of methods can handle noisy data and can discover arbitrarily shaped clusters. Outliers are not problematic with this class of methods. However, density-based techniques have difficulty with data of varying densities. Together with hierarchical and partitioning-based methods, density-based methods have difficulties working with high dimensional data. As dimensionality increases, the feature space increases and objects appear to be sparse and dissimilar which affects clustering tendency [28].

3.3 Grid Methods

Grid-based methods form a grid structure from a finite number of cells quantized using the original data space. This class of methods denotes a fast processing time. Density-based methods require the practitioner to specify a grid size and a density threshold. However, this can be done automatically by using adaptive grids. Examples of commonly known grid-based methods include STING and CLIQUE. These methods are typically not effective for working with high dimensional data [19].

3.4 Hierarchical Methods

Hierarchical-based methods create a hierarchical decomposition for a given set of data points (i.e. divide similar instances by constructing a hierarchy of clusters). The family of methods can take an agglomerative (bottom-up) or divisive (top-down) approach. This class of methods can easily work with many forms of similarity or distance measures and are applicable to many attribute types. These methods suffer from a vagueness in termination criteria and also have difficulties in handling outliers or noisy data [19]. However, hierarchical clustering has the added advantage in that clustering results can be easily visualized and interpreted using a tree-based representation called a dendrogram (See Fig. 1).

One example of a hierarchical clustering method is the Hierarchical Agglomerative Clustering (HAC) algorithm. The HAC algorithm needs to determine the distance

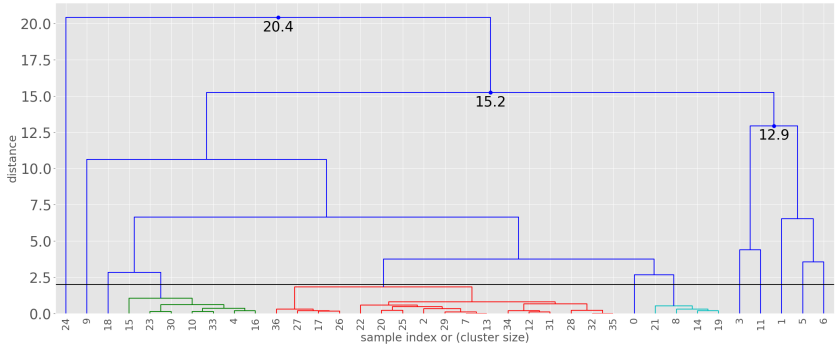


Fig. 1: Example Dendrogram

between samples in order to form similar groupings of data points. There are many options available to practitioners when selecting a distance measure. Among popular metrics are the Euclidean and Manhattan distance metrics. Proximity measures can affect the shape of clusters. Different similarity measures can produce valid clusterings but they will have different meanings. Often, the importance of the clustering depends on whether the clustering criterion is associated with the phenomenon under study. Euclidean distance is a preferred distance measure by researchers in the field of clustering. This distance metric measures the root of square differences between co-ordinates of pairs of objects [27] and is defined as [7]:

$$D(x,y) = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2} \tag{1}$$

The Manhattan distance computes the absolute differences between coordinate of pairs of objects and is defined as [27]:

$$Dist_{XY} = |X_{ik} - X_{jk}| \tag{2}$$

Kapil and Chawla [15] found that clustering using Euclidean distance outperformed clustering using Manhattan distance in terms of the number of iteration, sum squared errors and time taken to build the model. Manhattan distance is usually preferred over the more common Euclidean distance when there is high dimensionality in the data [1].

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.

4 METHODOLOGY

4.1 Clustering Algorithm Selection

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 HAC algorithm is usually limited to small datasets because of it's 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 [33,17].

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 have become interested in recent years in combining several algorithms (e.g. clustering ensemble methods) to process datasets [16].

The clustering method selected for use in this work is the HAC algorithm. The input data is of low dimensionality and the number of instances is small. A single and simple algorithm was selected in order to simplify the workflow execution and experimental setup. The case study is focused on how the proposed solution facilitates the comparison of clustering results and reduces the cognitive demand on practitioners in identifying, understanding and comparing similar clustering results.

4.2 The Proposed Analytical Workflow

Fig. 2 provides a conceptual overview of the main tasks of our proposed workflow. The numbered items in the figure link back to individual Python scripts described in detail in the implementation section. At the end of the process, a database is used to persist all clustering results and a RESTful Application Programming Interface (API) facilitates querying these results by different practitioners.

Data Preprocessing and Fusion The data preprocessing and fusion task uses raw data from the public EV charging stations. Preprocessing consists of data cleaning and consolidation steps. Data cleaning, ensures good data quality and produces a set of cleaned files by eliminating errors, inconsistencies, duplicated and redundant data rows, and handling missing data. Data consolidation combines data from various data files into a single dataset. A variety of files from the cleaned dataset are used as the input for this operation. The output of these steps is a unique file that merges all attributes into one big table.

Moreover, data fusion consists of combining multiple data sources followed by a reduction or replacement for the purpose of better inference. In our proposed clustering process, consolidated station location information and charging event data files are combined to produce more consistent, accurate, and useful data files.

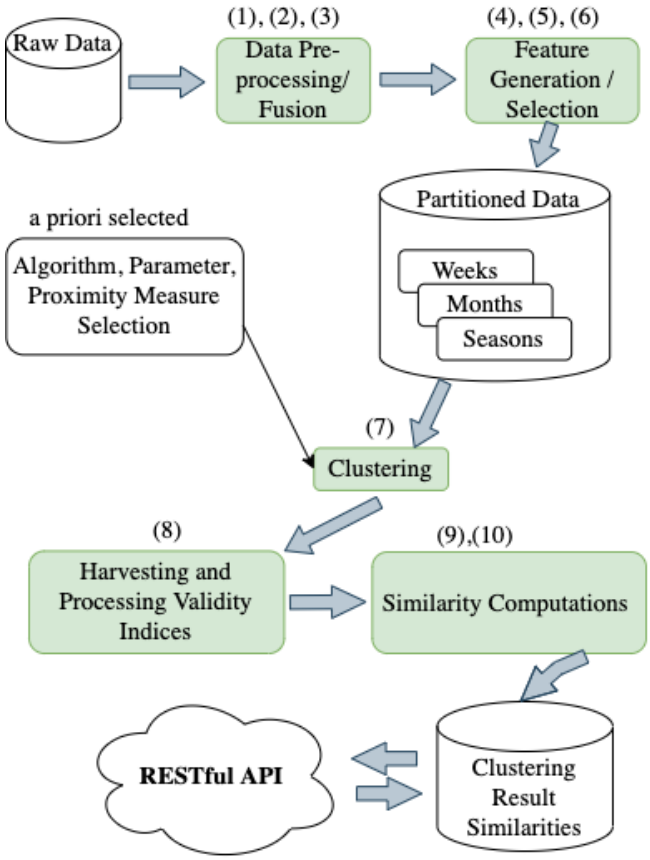


Fig. 2: Our Proposed Analytical Workflow [25]

Feature Generation and Selection The aim of the feature generation and selection task is to enrich pre-processed and fused data files by adding new attributes to each data row according to a specific context. This task is defined by a contextualization function that can produce a set of new data rows using contextualization parameters to add new attributes to the fused data rows. Transformed data is then partitioned using multiple temporal granularities (e.g. weekly, monthly or seasonally).

Clustering The aim of the clustering task is to find the patterns from transformed input data using a hierarchical agglomerative clustering algorithm. 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 the input data, producing multiple

clustering results. Internal cluster validity indices are recorded during each application of the clustering algorithm.

Harvesting and Processing Validity Indices 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 task utilizes the combination of eight cluster validity indices which are thoroughly described in [25] and listed in Table 3.

Similarity Computations Our work uses a proximity measure in the clustering task and in the computation of the results similarity matrix. Selecting a measure to determine how similar or dissimilar two data points is an important step in any clustering process. Proximity measures can affect the shape of clusters as some data points may be relatively close to one another according to one measure and relatively far from each other according to another.

In addition to the clustering task, the similarity computation task uses Euclidean distance as the proximity measure between clustering results. All index values (e.g. multidimensional points in Euclidean space) of each clustering result are used in the distance computations. The pair-wise similarity comparisons (e.g. the similarity matrix) are then persisted in a database for down-stream results exploration via a RESTful API.

The similarity matrix is stored in the database using two tables. The first table summarizes clustering results with rows consisting of a unique clustering result ID (*result_id*) and meta-data about running the algorithm (e.g. input file name, clustering execution time, all validity index values, etc.). The second table, which is linked to the first table, contains rows consisting of a source result ID (*from_result_id*), a target result ID (*to_result_id*) and a Euclidean distance. Links between result IDs are not duplicated as directionality is not considered.

4.3 Clustering and Results Exploration

The proposed analytical workflow enables the basic identification and interactive querying of potentially interesting clustering results. Additionally, the resulting assembly enables drilling down into relative rankings of comparable results for diagnostic and downstream analytical tasks. This process leverages the aforementioned RESTful API in order to facilitate this capability. The workflow facilitates the comparison of clustering results by practitioners with different priorities and preferences.

Selecting the appropriate algorithm and hyperparameters in clustering is critical. However, interpreting the level of "success" achieved once modeling results are available can be cognitively demanding. Their may exist several viable combinations of algorithms and hyperparameters that result in plausible clusters. Comparing and contrasting multiple clustering results can help uncover interesting structure in data. Nevertheless, this comes at a cost since practitioners 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 mental demand in comparing and contrasting results.

The traditional usage of CVI has been for validation purposes. However, utilizing multiple CVI together in combination with a proximity measure such as Euclidean distance has a strong potential to offer a new pairwise similarity measure that can enhance the comparison of clustering results. Supporting the practitioner by automating clustering workflows and presenting meaningful analytical results in a way that increases the opportunity to understand and compare similar groupings can assist in recognizing patterns and identifying meaningful results for downstream analysis.

5 IMPLEMENTATION

This work makes use of real operational data from public EV charging stations provided by the New Brunswick Power Corporation. 9,505 EV charging events that occurred between the dates of April 2019 and April 2020 at Level-2 (L2) and Level-3 (L3) public charging stations were included in the analysis. Table 1 describes the raw EV charging dataset features. Our practitioners are utility company managers and planners that are responsible for coordinating various projects including EV charging station condition assessments, operating and capital budget forecasting, and maintenance and operation practices development. Fig. 3 describes the overall end-to-end implementation of our EV case study.

Table 1: Raw Data [25]

Column Name	**Description**
Connection ID	Unique identifier for a connection
Recharge start time (local)	Timestamp denoting start of charging event
Recharge end time (local)	Timestamp denoting end of charging event
Account name	Unused (all null)
Card identifier	Unique identifier for a charging plan member
Recharge duration (hours:minutes)	Duration of charge event
Connector used	Connection used during charge event
Start state of charge (%)	State of charge % at beginning of charging event
End state of charge (%)	State of charge % after charging event is complete
End reason	Charge event end reason
Total amount	Unused (all null)
Currency	Unused (all null)
Total kWh	Energy transferred to vehicle during charging event
Station	Unique identifier for charging station

Custom-written Python code and a scientific Python stack were leveraged to implement the proposed clustering process. Task elements were executed in sequence from

a centralized management script. The software programs used in this work were packaged using a Docker [6] container in order to ensure a reproducible and consistent computational environment.

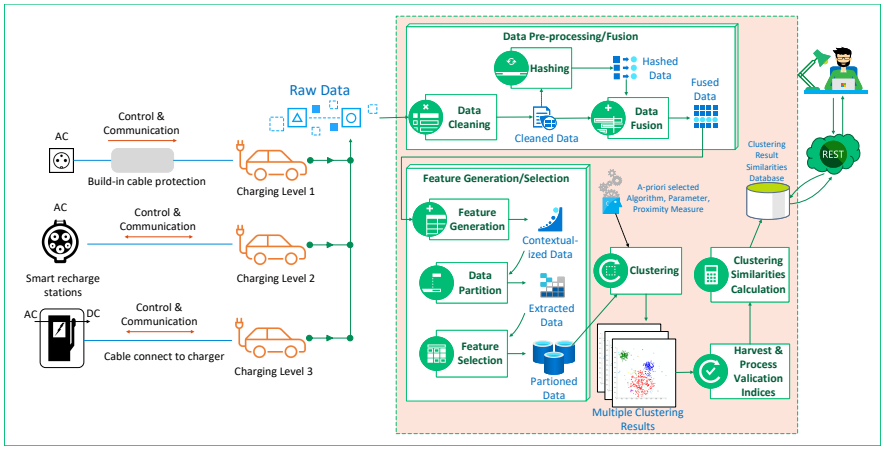


Fig. 3: Overview of Our Implemented EV Case Study [25]

Fig. 4 highlights noteworthy aspects of the implementation. The numbered boxes represent individual parameterized Python scripts. The data flow is such that the output of one script is the input for the next script. Input and output file names contain parameter values that were used when calling the workflow's scripts. The grey elements represent a job's input file(s). The blue elements represent a job's output file(s). The detailed implementation of each script is described as follows:

- **Script (1)**: The ***one_way_hash.py*** script imports raw event data and casts column elements to appropriate types. Additionally, a one-way hash function is applied to the *Card identifier* column.
- **Script (2)**: The ***locations_to_parquet.py*** script imports raw station location data and integrates multiple input files into one.
- **Script (3)**: The ***fuse_location_w_events.py*** script fuses event data with charging station location information.
- **Script (4)**: This work focuses on recharge report event data in the downstream analysis. The ***feat_eng_rech_report.py*** script creates new features (contextualized) based on calculations involving existing data attributes and removes events with a duration of 5 minutes or less (eliminating 11% of the raw records).
- **Script (5)**: The ***create_batch_ranges.py*** script creates temporal partitions of the data. These partitions facilitate the cluster analysis based on charging events occurring during a particular week, month or season of the year.
- **Script (6)**: The ***generate_ev_station_features.py*** prepares the input data for clustering 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.

- **Script (7)**: The ***cluster_data.py*** script applies the agglomerative clustering algorithm to all temporal slices of the data produced in the previous task. This is done for a cluster count hyperparameter that varies from 2 to 7. Other hyperparameter settings are kept constant to simplify the experimental setup. Internal cluster validity indices are recorded during each application of the clustering algorithm (See Table 2 for the list of indices).
- **Script (8)**: The ***scale_indices.py*** script normalizes the internal cluster validity indices in preparation for the downstream Euclidean distance computations.
- **Script (9)**: The ***similarity_matrix.py*** script performs pairwise Euclidean distance computations for each clustering result. All index values (i.e. multidimensional points in Euclidean space) of each clustering result are used in the distance computations.
- **Script (10)**: The ***load_data.py*** script persists the similarity matrix data produced in the previous task in a relational database to enable querying of clustering results and corresponding similarities across months, weeks and seasons. The database query functionality is made available via a RESTful API.

After results are generated and persisted (i.e. Script (10) in Fig. 4 is complete), the practitioner can navigate these results via a RESTful interface. Fig. 5 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 request 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, mapped station cluster memberships and silhouette plots.

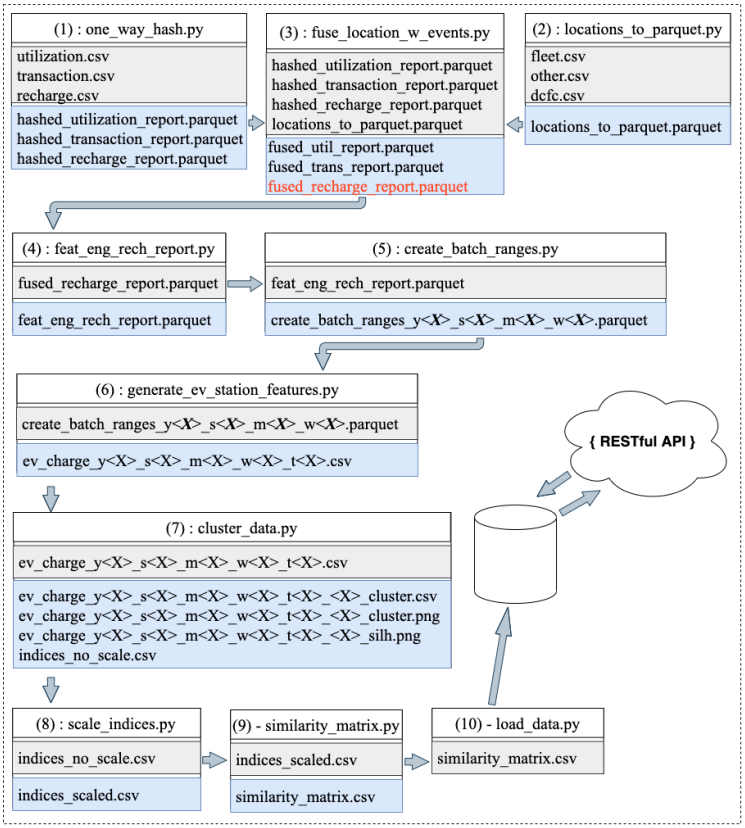


Fig. 4: Implemented Clustering Process Data Flow [25]

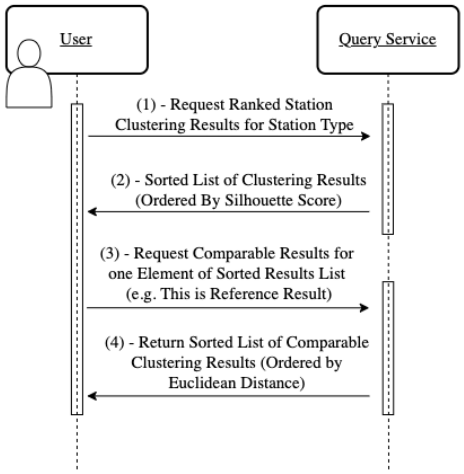


Fig. 5: Results Query Sequence [25]

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 our approach. The case study is presented in the next section.

6 DISCUSSION OF THE RESULTS

This section highlights the results of our proposed approach in identifying similar station clusterings over multiple weeks with a case study. Table 3 highlights similar clustering results relative to station clusterings for a target week starting on May 27th, 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 target week. According to the multi-dimensional pairwise distance calculations obtained using the indices described in Table 2, the most similar clustering result to the week starting on May 27th, 2019 is the result for the week starting on February the 17th 2020. The least similar clustering result is the result for the week starting on December 2nd, 2019.

Table 2: Clustering Validity Index Data [25]

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

A corresponding visual presentation of the clustering results found in Table 3 can be seen in Figures 6 through 10. Each figure contains a silhouette plot, scatter plot and a map describing the clustered data. In the silhouette plots, an observation with a silhouette width near 1, means that the 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.

Table 3: Clustering Similarities - L3 - May 27th, 2019 [25]

WEEK	Sil	CH	DB	C	S	RMS	RS	XB	*Dist*
MAY-27-2019	**0.60**	**51.37**	**0.51**	**1.12**	**2.40**	**0.15**	**0.68**	**0.09**	***N/A***
FEB-17-2020	0.60	49.35	0.57	0.19	2.44	0.16	0.67	0.10	*0.081*
MAR-02-2020	0.65	55.51	0.52	1.14	2.63	0.15	0.70	0.07	*0.101*
JUL-29-2019	0.60	55.82	0.53	0.99	2.30	0.14	0.70	0.11	*0.105*
...	...	...	...	...	...	...	...	...	...
DEC-02-2019	0.63	56.55	0.58	1.26	2.97	0.16	0.70	0.09	*0.177*

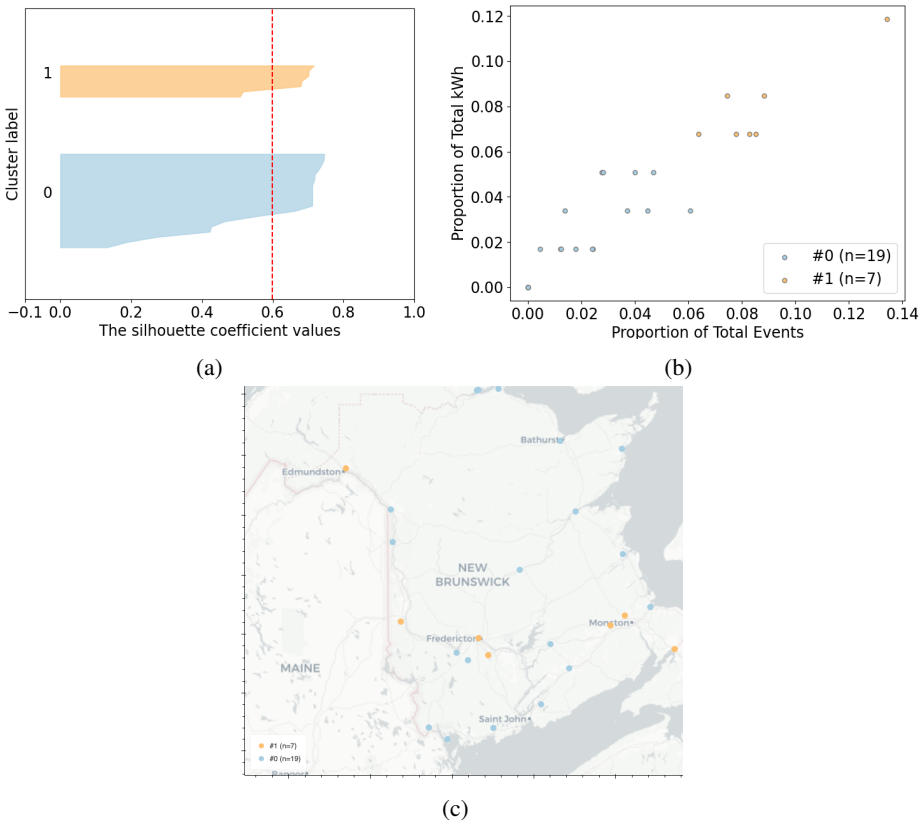
Column Name Abbreviations :

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 Week of May 27th, 2019 - (Reference Week)

We can see from Fig. 6 that a reasonable structure in the data has been found for our reference week, which starts on May 27th, 2019. 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 (See Fig. 6a).

In Fig. 6b, cluster 0, the cluster with relatively lower utilization rates, has more station members than cluster 1. Cluster 1 is the grouping of stations with relatively higher utilization rates. In the scatter plot, crisp clusters identified by the HAC algorithm can be observed. However, cluster 1 has an observation that is comparatively far from its other station members. The map in Fig. 6c, indicates that cluster 1 member stations are mostly located in the lower half of the province.



(a) (b)

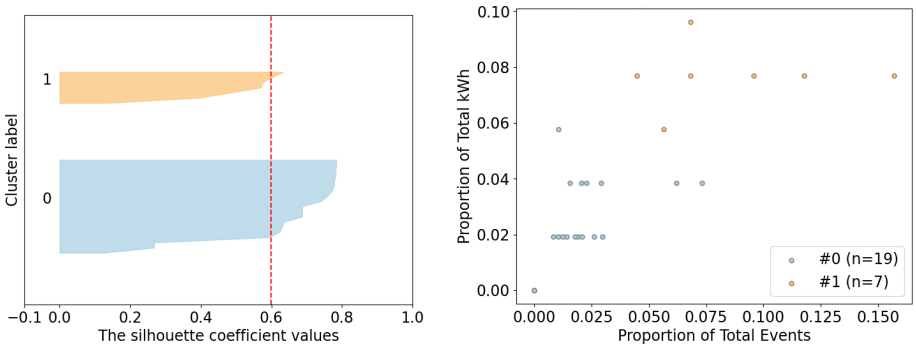
(c)

Fig. 6: L3 Station Clusters - MAY-27-2019

6.2 Week of February 17th, 2020

We now focus on the closest clustering result relative to our reference week. This grouping is for the week starting on the 27th of February, 2020. The average silhouette score for this result is also 0.60 (See Fig. 7a). The scatter plot of Fig. 7b denotes relatively well separated clusters similar to our reference week. The clusters can also be thought of as groupings of high vs. low station utilization rates with this result. Additionally, the number of observations in each cluster is the same as the reference week. Results for the week of May 27th, 2019 are slightly better when considering all cluster validation indices. This can also be observed visually. Data points seem to be closer together in the scatter plot of Fig. 6b than in Fig. 7b. The in-between cluster separation in both results are similar.

The map in Fig. 7c reveals that cluster 1 - the higher utilization stations cluster - member stations are mostly located in the right half of the province with this clustering result.



(a) (b)

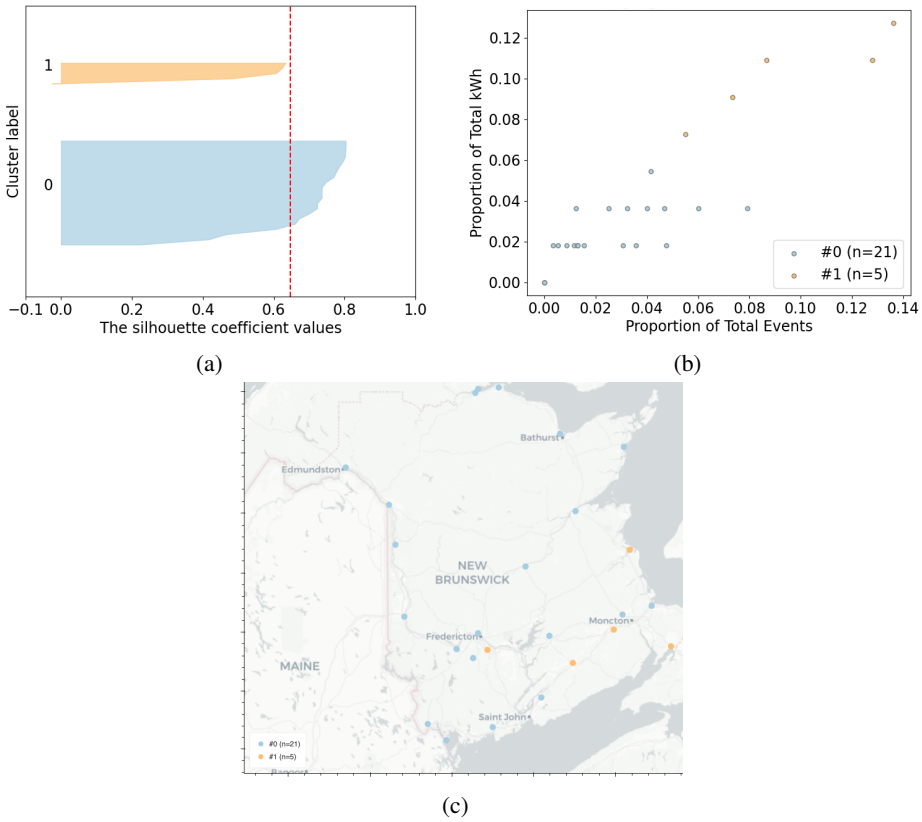
(c)

Fig. 7: L3 Station Clusters - FEB-17-2020

6.3 Week of March 02nd, 2020

The next closest clustering result relative to our reference week is the grouping for the week starting on March 02nd, 2020. The average silhouette score for this result is 0.65. The silhouette plot in Fig. 8a 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 less than optimal clustering is confirmed by observing the scatter plot of Fig. 8b. Some observations in cluster 1 could be outliers. Additionally, the cluster's cohesion is not as prevalent as cluster 0's. Perhaps a cluster count of 3 would be more appropriate with this result.

Fig. 8c, indicates that cluster 1 - the higher utilization stations cluster - member stations are mostly located in the lower-right half of the province with this clustering result.



(a) (b)

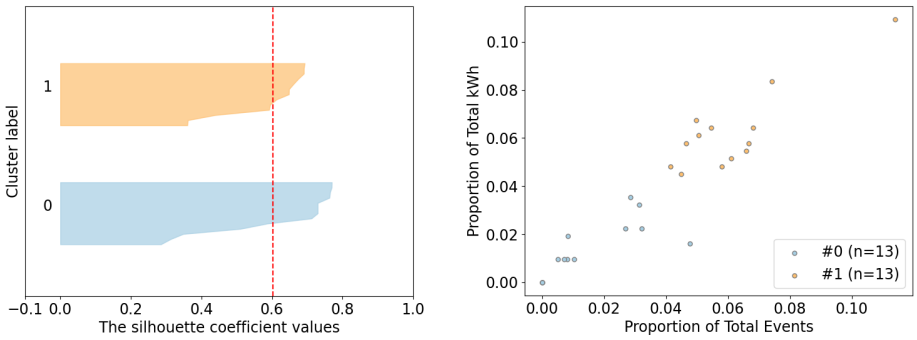
(c)

Fig. 8: L3 Station Clusters - MAR-02-2020

6.4 Week of July 29th, 2019

The silhouette plot in Fig. 9a and the average silhouette score of 0.60 suggest a reasonable structure in the data has also been found in this week. Fig. 9b denotes relatively well separated clusters. Cluster 1 has an observation that is comparatively far from its other station members. The number of observations in each cluster for both the reference clustering result and this result are different. Based on the various indices, clustering results for July 29th, 2020 are better in some aspects and inferior in others to results for the week of May 27th, 2019. This result was identified as being the 3rd most similar result for our target week.

Fig. 9c, indicates that cluster 1 - the higher utilization stations cluster - member stations are mostly located along a major freeway in the province, mostly covering the left and the bottom sections of the province.



(a) (b)

(c)

Fig. 9: L3 Station Clusters - JUL-29-2019

6.5 Week of December 02nd, 2019

The decreasing relative similarity of results is especially visible when comparing the results for the week of May 27th, 2019 with results having the least similarity (i.e, results for the week of December 2nd, 2019). In Fig. 10a we can see that all cluster 1's members have below average silhouette scores and the clustering of stations is much less similar than the other clusterings. Additionally, as can be observed in Fig. 10b, perhaps a cluster count of 3 would be more appropriate with this result.

Fig. 10c, indicates that cluster 1 - the higher utilization stations cluster - member stations are mostly located in the lower-right half of the province with this clustering result.

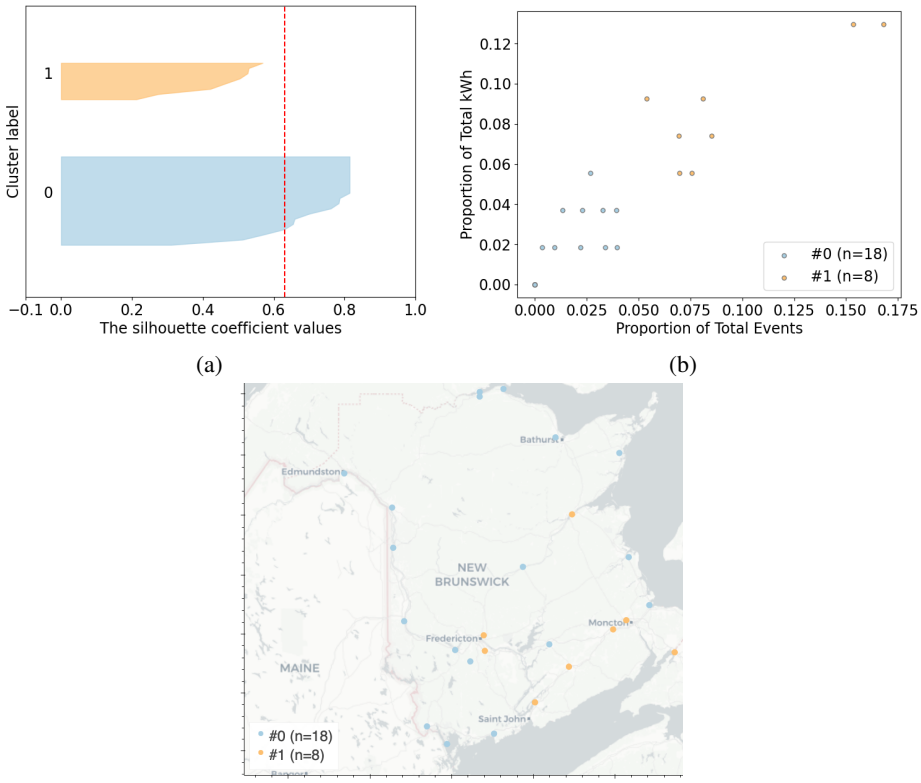


Fig. 10: L3 Station Clusters - DEC-02-2019

6.6 Overall Results

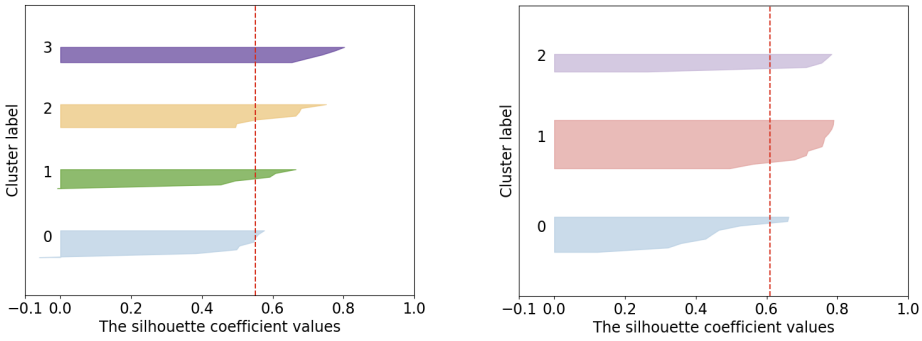
As can be observed in Figures 6 to 10 of the previous sections, the decreasing relative similarity of clustering results is especially noticeable when visually comparing the silhouette and scatter plots for the week of May 27th with the same visualizations in other weeks and doing so in a step-wise fashion down the ranked list of results.

Individual index calculations embed implicit trade-offs on what is prioritized when expressing inter-cluster separation, inter-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. This is illustrated in Table 3 where for example, the silhouette, Caliński-Harabasz, separation and R-squared index values for December 02nd suggest a better clustering than on the week starting on May 27th. However, the Davies-Bouldin, cohesion and RMS index values inverse this assessment.

Capital investments in public charging infrastructure involves the use of public funds and necessitates robust informed decision making. Identifying similar station uti-

lization patterns over multiple weeks can be useful planning information for station operators. The cluster analysis presented in our case study provides useful insights by identifying similar groupings of EV charging stations according to their usage patterns in time.

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 dataset with temporal properties. The relative ranking of similar clustering results that our approach affords makes it easy to objectively identify similar station groupings over multiple weeks based on a reference week. 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. For example silhouette plots representing a reference month (where K=4) and season (where K=3), see Fig. 11.



(a) (b)

Fig. 11: L3 Station Clustering References - August and Spring [25]

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. To optimize operations, EV station operators often seek market-related insights. EV charging station clustering can reveal useful segmentations in service consumption patterns.

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 struc-

tures 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.

The contributions of this work include an end-to-end analytical workflow that enables the analysis of energy utilization patterns at public charging infrastructure using real charging data from station operators in Atlantic Canada. This workflow facilitates the comparison of clustering results by practitioners with different priorities and preferences. Utilizing the combination of eight internal cluster validity indices to compute a proximity measure of clustering results in a priori selected temporal partitions of the data reduces the cognitive demand on users in identifying, understanding and comparing of similar clustering results over time. A case study demonstrates that given a clustering result of interest, the process of objectively highlighting and recommending similar clustering results can be automated in order to support the practitioner in evaluating how structure in data persists over multiple time slices and reduce effort in identifying multiple meaningful clustering results from a large number of modeling artifacts.

Currently, the initial ranked list of clustering results described in Step 1 of Fig. 5 is created using silhouette scores only. Framing the creation of the initial ranked list of results as a Multiple Criteria Decision Making (MCDM) problem may improve the initial results exploration experience. This will be included in future work. EV charging patterns can be more effectively analyzed by referencing the social and economic contexts in which they occur. Once clusters are obtained, it may be useful to explain the clusters with features other than the original features used to obtain the clusters. The use of real-world EV charging event data combined with nearby traffic volumes and nearby amenities may help to further contextualize the clustering results. This will also be included in future work. Lastly, other avenues will explore if utilizing the Euclidean distances and clusters obtained in this work can improve predictive performance of a baseline classifier such as improving the predictive performance of classifiers for predicting peak day of week kWh.

Acknowledgements. The authors of this paper like to thank the New Brunswick Power Corporation for providing access to station operator users and the EV charging data referenced in this research. This work was partially supported by the NSERC/Cisco Industrial Research Chair, Grant IRCPJ 488403-1.

References

1. Aggarwal, C.C., Hinneburg, A., Keim, D.A.: On the surprising behavior of distance metrics in high dimensional space. In: International conference on database theory. pp. 420–434. Springer (2001)
2. Ahmed, S.T., Kumar, S.S., Anusha, B., Bhumika, P., Gunashree, M., Ishwarya, B.: A generalized study on data mining and clustering algorithms. In: International Conference On Computational Vision and Bio Inspired Computing. pp. 1121–1129. Springer (2018)
3. Al-Ogaili, A.S., Hashim, T.J.T., Rahmat, N.A., Ramasamy, A.K., Marsadek, M.B., Faisal, M., Hannan, M.A.: Review on scheduling, clustering, and forecasting strategies for controlling electric vehicle charging: challenges and recommendations. Ieee Access **7**, 128353–128371 (2019)
4. Arbelaitz, O., Gurrutxaga, I., Muguerza, J., PéRez, J.M., Perona, I.: An extensive comparative study of cluster validity indices. Pattern Recognition **46**(1), 243–256 (2013)
5. Bae, J., Helldin, T., Riveiro, M., Nowaczyk, S., Bouguelia, M.R., Falkman, G.: Interactive Clustering: A Comprehensive Review. ACM Computing Surveys **53**(1), 1–39 (May 2020). https://doi.org/10.1145/3340960, https://dl.acm.org/doi/10.1145/3340960
6. Boettiger, C.: An introduction to docker for reproducible research. ACM SIGOPS Operating Systems Review **49**(1), 71–79 (2015)
7. Chakrabarty, A.: An investigation of clustering algorithms and soft computing approaches for pattern recognition. Ph.D. thesis, Assam University (2010)
8. Desai, R.R., Chen, R.B., Armington, W.: A pattern analysis of daily electric vehicle charging profiles: operational efficiency and environmental impacts. Journal of Advanced Transportation **2018** (2018)
9. Ekta Meena, B., Elizabeth, C., Marine, G., Christopher, L., Leonardo, P., Jacopo, T., Jacob, T., Chase, L., Owen, M., Dan, W., Ralph, P., Disha, S., Chengwu, X.: Global ev outlook 2021: Accelerating ambitions despite the pandemic (2021)
10. Han, J., Pei, J., Kamber, M.: Data mining: concepts and techniques. Elsevier (2011)
11. Heuberger, C.F., Bains, P.K., Mac Dowell, N.: The ev-olution of the power system: A spatio-temporal optimisation model to investigate the impact of electric vehicle deployment. Applied Energy **257**, 113715 (2020)
12. Iglesias, F., Kastner, W.: Analysis of similarity measures in times series clustering for the discovery of building energy patterns. Energies **6**(2), 579–597 (2013)
13. Ji, D., Zhao, Y., Dong, X., Zhao, M., Yang, L., Lv, M., Chen, G.: A spatial-temporal model for locating electric vehicle charging stations. In: National Conference on Embedded System Technology. pp. 89–102. Springer (2017)
14. Kang, J., Kan, C., Lin, Z.: Are electric vehicles reshaping the city? an investigation of the clustering of electric vehicle owners' dwellings and their interaction with urban spaces. ISPRS International Journal of Geo-Information **10**(5), 320 (2021)
15. Kapil, S., Chawla, M.: Performance evaluation of k-means clustering algorithm with various distance metrics. In: 2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES). pp. 1–4. IEEE (2016)
16. Khedairia, S., Khadir, M.T.: A multiple clustering combination approach based on iterative voting process. Journal of King Saud University-Computer and Information Sciences (2019)
17. Kuwil, F.H., Atila, Ü., Abu-Issa, R., Murtagh, F.: A novel data clustering algorithm based on gravity center methodology. Expert Systems with Applications **156**, 113435 (2020)
18. Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J.: Understanding of internal clustering validation measures. In: 2010 IEEE International Conference on Data Mining. pp. 911–916. IEEE (2010)

19. Mann, A.K., Kaur, N.: Review paper on clustering techniques. Global Journal of Computer Science and Technology (2013)
20. Morton, C., Anable, J., Yeboah, G., Cottrill, C.: The spatial pattern of demand in the early market for electric vehicles: Evidence from the united kingdom. Journal of Transport Geography **72**, 119–130 (2018)
21. Ofetotse, E.L., Essah, E.A., Yao, R.: Evaluating the determinants of household electricity consumption using cluster analysis. Journal of Building Engineering **43**, 102487 (2021)
22. Oliveira, M.: 3 Reasons Why AutoML Won't Replace Data Scientists Yet (2019), https://www.kdnuggets.com/3-reasons-why-automl-wont-replace-data-scientists-yet.html/, (March 2019)
23. Poulakis, G.: Unsupervised AutoML: a study on automated machine learning in the context of clustering. Master's thesis, Πανεπιστήμιο Πειραιώς (2020)
24. Rendón, E., Abundez, I., Arizmendi, A., Quiroz, E.M.: Internal versus external cluster validation indexes. International Journal of computers and communications **5**(1), 27–34 (2011)
25. Richard, R., Cao, H., Wachowicz, M.: 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,. pp. 67–77. INSTICC, SciTePress (2021). https://doi.org/10.5220/0010485000670077
26. Si, C., Xu, S., Wan, C., Chen, D., Cui, W., Zhao, J.: Electric load clustering in smart grid: Methodologies, applications, and future trends. Journal of Modern Power Systems and Clean Energy **9**(2), 237–252 (2021)
27. Singh, A., Yadav, A., Rana, A.: K-means with three different distance metrics. International Journal of Computer Applications **67**(10) (2013)
28. Sisodia, D., Singh, L., Sisodia, S., Saxena, K.: Clustering techniques: a brief survey of different clustering algorithms. International Journal of Latest Trends in Engineering and Technology (IJLTET) **1**(3), 82–87 (2012)
29. Straka, M., Buzna, L.: Clustering algorithms applied to usage related segments of electric vehicle charging stations. Transportation Research Procedia **40**, 1576–1582 (2019)
30. Sun, C., Li, T., Low, S.H., Li, V.O.: Classification of electric vehicle charging time series with selective clustering. Electric Power Systems Research **189**, 106695 (2020)
31. Xiong, Y., Wang, B., Chu, C.C., Gadh, R.: Electric vehicle driver clustering using statistical model and machine learning. In: 2018 IEEE Power & Energy Society General Meeting (PESGM). pp. 1–5. IEEE (2018)
32. Xydas, E., Marmaras, C., Cipcigan, L.M., Jenkins, N., Carroll, S., Barker, M.: A data-driven approach for characterising the charging demand of electric vehicles: A uk case study. Applied energy **162**, 763–771 (2016)
33. Zolhavarieh, S., Aghabozorgi, S., Teh, Y.W.: A review of subsequence time series clustering. The Scientific World Journal **2014** (2014)