Unsupervised Classification

Shruti Patil
Hof University of Applied Science, Hof, Germany
shruti.patil@hof-university.de

ABSTRACT

Analysing mobility data has become vital to understand mobility patterns, high demand areas and smart transportation planning. This study focuses on analysing geographical data in two different datasets and tries to identify high demand areas and hot-spots in the region. Supervised classification is used to predict predefined labels or class for a data point based on features. But availability of labeled data is rare and labeling data is time-consuming, labor-intensive, and costly, making it a challenge. Hence, this research tries to resolve this challenge by implementing an approach to assign labels to unlabeled data. This paper focuses on applying two phases from the C1AMP methodology suggested by Bobek in [5] and implementing this methodology on spatial-temporal data to understand high demand areas in the city. Through the application of these phases to spatial-temporal data, the effectiveness of classification on pseudo-labeled data derived from clustering is evaluated. The clustering is perfomed using ST-DBSCan algorithm, which is best suitable for Spatial-temporal data and XG-Boost classifier is used for classification. Further trained model is validated using various evaluation metrics. Additionally, this approach offers a potential for cluster analysis and following these preliminary steps to enable cluster interpretation through Explainable AI in future.

Keywords: Clustering, Classification, Spatial- Temporal

1 INTRODUCTION

The availability of large amounts of unlabeled data brings both opportunities and challenges for modern machine learning and data analysis. Traditional supervised learning methods depend a lot on labeled datasets, where each data point has a specific label or category. However, in many real-world scenarios, acquiring labeled data is costly, time-consuming, and often infeasible [17]. This limitation has driven significant interest in methodologies that can leverage unlabeled data effectively. One such approach involves using clustering as a preliminary step to label the unlabeled data, followed by the application of classification techniques [17] .

Clustering, an unsupervised learning technique, identifies natural groupings or patterns within data based solely on the inherent structure of the dataset [9][3]. By assigning labels to these clusters, the data transitions from an unlabeled state to a pseudo-labeled state, enabling its use in supervised classification models. This two-step process bridges the gap between unsupervised and supervised learning paradigms, facilitating the extraction of actionable insights from datasets where traditional labeling is impractical. For instance, in domains like mobility analysis, healthcare, and marketing, clustering-based labeling helps identify meaningful patterns or customer segments, which can then be analyzed further through classification [17] [3].

The integration of clustering and classification not only enhances the utility of unlabeled data, but also introduces challenges, such as ensuring that the initial clustering aligns well with the eventual classification objectives [14][6]. The accuracy and robustness of the clustering phase directly impact the performance of the classification model. In addition, the choice of clustering algorithms, hyperparameter tuning, and evaluation metrics play a critical role in optimizing this hybrid approach. In light of these considerations, this research aims to explore and refine the methodologies for leveraging clustering as the first step in transforming unlabeled data

into valuable labeled datasets for subsequent classification tasks. This process has significant implications for practical applications in fields that generate massive amounts of raw, unlabeled data, such as mobility planning and smart city initiatives. This mode of classification can hence be called as Un-supervised mode of classification [6] [14].

Effective transportation planning and management requires a comprehensive understanding of historical and current mobility patterns, traditionally obtained through travel surveys and traffic counts [6]. However, these conventional methods are often expensive, limited in scope, and hindered by decreasing response rates. Traditional rule-based or geometric methods often fail to differentiate between similar travel modes, while supervised techniques are constrained by the lack of labeled data. Semi-supervised learning, which leverages patterns in unlabeled data, presents a promising solution to these challenges [6].

This research is an implementation study of a methodology proposed by Bobek in [5], referred to as ClAMP (Cluster Analysis with Multidimensional Prototypes). The ClAMP approach is designed to assist in cluster analysis by generating human-readable, rule-based explanations, enabling more interpretable results compared to traditional clustering methods. This study specifically focuses on applying two phases of the ClAMP methodology to analyze large datasets characterized by spatial-temporal features, with the goal of addressing the difficulties in using unlabeled data for classification. A key aspect of this research is the exploration of unsupervised methods for analyzing and classifying spatial-temporal data. By employing clustering techniques and pseudo-labeling, the study aims to identify high-demand urban areas and popular hotspots, offering valuable insights into mobility patterns. The approach demonstrates how unsupervised methods can outperform traditional classification techniques, particularly in cases where labeled data is scarce. In the future, Explainable AI (XAI) can be integrated into this framework to further enhance the interpretability and transparency of clustering and classification results. XAI techniques can help in providing detailed insights into the rationale behind the formation of clusters and classification decisions, making it easier to understand and trust the outcomes of such analysis. This advancement could significantly contribute to smarter urban planning and mobility management.

The structure of this paper is organized as follows: Section 2 provides a comprehensive review of the relevant literature, discussing key studies in the fields of cluster analysis, unsupervised learning, and spatial-temporal data classification. It highlights existing methodologies and identifies critical research gaps, such as the challenges associated with utilizing unlabeled data for classification tasks. Section 3 outlines the methodology adopted for the study. Section 4 introduces the datasets used and describes the steps involved in data preparation. Sections 5 focus on the implementation of the study on two distinct datasets, presenting the applied techniques and analyzing their respective results in depth. Finally, Sections 6 and 7 summarize the insights gained from the analysis, discuss their implications, and present the conclusions drawn, along with recommendations for future work.

2 LITERATURE REVIEW

2.1 Information Sources and Eligibility Criteria

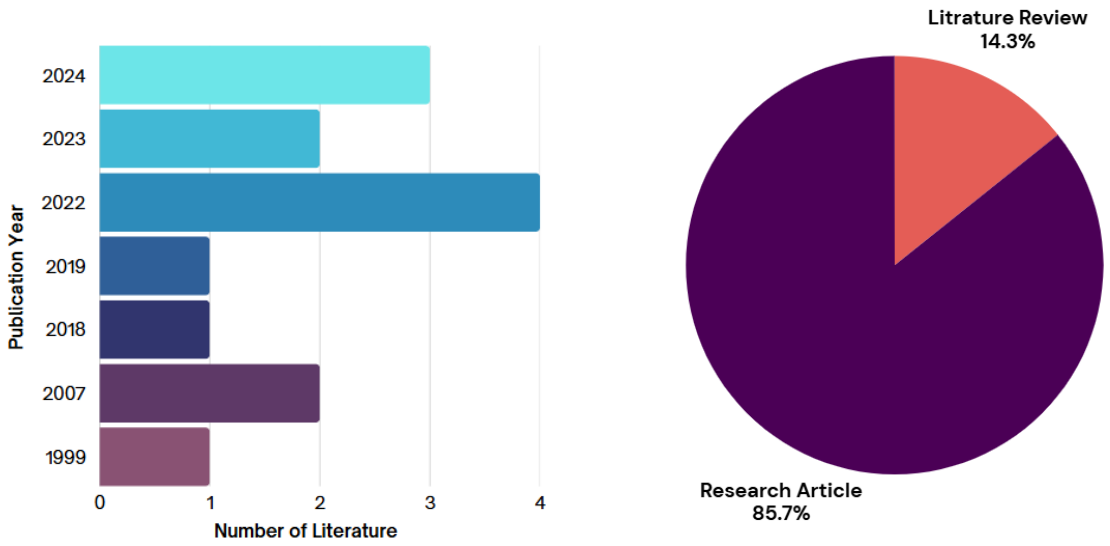
The literature search for this study was conducted using several academic databases and online resources to ensure a comprehensive review of relevant works. The primary sources included widely recognized platforms such as Google Scholar, IEEE Xplore, and ScienceDirect, which are well-suited for accessing peer-reviewed articles and conference papers. Additionally, SpringerLink and ACM Digital Library were utilized to explore domain-specific publications in artificial intelligence, clustering, and mobility analysis.

The search process involved using specific keywords and phrases, including "spatio-temporal clustering", "pseudo-labeling", "mobility data analysis", "unsupervised learning", and "explainable AI". The inclusion criteria focused on studies that addressed clustering methods, particularly those applicable to spatial-temporal datasets and mobility analysis. Preference was given to research published within the last 10 years to ensure the inclusion

of recent advancements and methodologies. Exclusion criteria included papers that lacked a clear application to clustering or spatio-temporal data and those that did not provide sufficient methodological details. Additional restrictions were applied to include only peer-reviewed studies published in English to maintain the quality and relevance of the reviewed literature.

2.2 Overview of Literature Characteristics

This section gives an overview of literature used for this research. The bar chart in Fig.1(a) shows the number of references categorized by their publication year. It highlights that most literature used in the review is from recent years, indicating a focus on current and relevant research. Older references, such as those from 1999 and 2007, are fewer but likely foundational works. The pie chart in Fig.1(b) illustrates the proportions of different types of literature used in the review. The majority of the references are research articles, while a smaller portion consists of literature reviews. This reflects a primary reliance on detailed research studies, with literature reviews providing additional background or context.



(a) Distribution of references by publication year

(b) Proportion of literature types

Fig. 1. Distribution of references by year(a) and literature type(b)

Table 1 provides a summary of the key aspects of the cited literature, highlighting the research focus and methodologies. It outlines the fields addressed by the studies, offering a concise overview of their contributions and areas of application. This helps in better understanding the scope and relevance of the reviewed works.

Table 1. Overview of literature used for the study

Reference	Research Focus (Methodology and Application)
[2] Álvarez-García et al. (2024)	Explainable cluster analysis framework for structured datasets
[10] Kumari et al. (2024)	Clustering algorithms for smart city traffic data analysis
[14] Li & Zhan (2024)	Improved semi-supervised learning algorithm for classification prediction
[8] Hussain et al. (2023)	Trajectory clustering using spatiotemporal models for transportation planning
[7] Bauer & Augenstein (2023)	Self-supervised learning for AI applications in computer vision
[5] Bobek et al. (2022)	ClAMP methodology for explainable clustering using multidimensional prototypes
[6] Breyer et al. (2022)	Semi-supervised mode classification for inter-city mobility patterns
[11] Kaminska et al. (2022)	Impact of clustering on classification in bipolar disorder data
[15] Ma & Zhang (2022)	Individual mobility prediction review for personalized mobility services
[12] Kauffmann et al. (2019)	Explaining clustering via neural networks for diverse datasets
[17] Peikari et al. (2018)	Semi-supervised learning for pathology image classification
[4] Birant & Kut (2007)	ST-DBSCAN for spatiotemporal clustering
[13] Lee et al. (2007)	Trajectory clustering with partition-and-group framework for movement pattern analysis
[9] Jain et al. (1999)	Data clustering: A review for clustering theory

2.3 Data Insights from Reviewed Studies

The reviewed literature presents diverse insights into data characteristics and recorded data items, focusing on the amount of labeled data, data types, and use cases. As depicted in Fig. 2, a significant portion of the studies utilize sparsely labeled datasets, leveraging semi-supervised approaches to enhance classification outcomes, as seen in [14], [6], and [17]. Fully unlabeled datasets are employed in 5 studies, such as [5] and [4], which rely on unsupervised techniques for clustering. In contrast, a smaller number of studies use largely labeled ([2]) or fully labeled ([10]) datasets to support methodologies like explainable clustering and traffic data analysis. The characteristics of the datasets also vary widely. Spatial and temporal data dominate studies like [8] and [4], where geographical and time-based features are essential for clustering tasks. Other studies, such as [14], highlight the use of multimodal and categorical data formats, while unstructured datasets are addressed through self-supervised methods in [7]. The use cases in the reviewed studies reflect a broad spectrum of applications. Urban and smart city interventions dominate works like [10] and [8], focusing on mobility planning and traffic optimization. Rural and inter-city mobility scenarios are explored in studies like [6] and [15], while healthcare and personalized mobility services are central to works like [17] and [11].

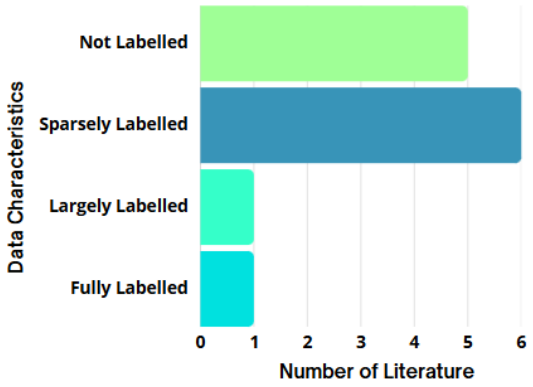


Fig. 2. Distribution of data characteristics across the reviewed literature

2.4 Research Gaps and Challenges

The reviewed studies demonstrate significant advancements in clustering and explainable artificial intelligence (AI), but several critical gaps and challenges remain. A major issue is the dependency on labeled data for supervised methods, which is both costly and time-consuming to generate. Unsupervised techniques, such as [5] and [4], address this limitation by leveraging unlabeled datasets, but they often face challenges in producing high-quality clusters and ensuring interpretability. Semi-supervised approaches ([14], [6]) attempt to bridge this gap using sparsely labeled datasets, but achieving an optimal balance between labeled and unlabeled data remains a complex task. The generalizability of clustering methods across diverse datasets is another key challenge. The ClAMP methodology proposed by Bobek ([5]) offers flexibility in clustering algorithms and enhances interpretability through human-readable rules. However, it presents several challenges, such as complexity of implementation compared to centroid-based approaches, dependence on data quality for generating meaningful clusters, and scalability issues, as its performance on large datasets is not thoroughly addressed. Furthermore, the method's evaluation is limited to benchmark datasets, raising concerns about its applicability to real-world, diverse datasets. Similarly, frameworks like [2] are effective for structured datasets but struggle with adaptability to unstructured or dynamic scenarios. The integration of clustering and classification workflows also poses challenges. While studies like [11] demonstrate that clustering can enhance classification accuracy, the scalability and optimization of these workflows for large-scale, dynamic environments remain underexplored. Applications in urban traffic analysis ([10]) and inter-city mobility ([6]) often focus on domain-specific scenarios, leaving gaps in their adaptability to areas like healthcare ([17]) or personalized services ([15]).

The interpretability of clustering results continues to be a pressing concern. Methods such as [12] introduce neuralized models for explainability, but they often become computationally intensive or lose clarity with highly dynamic datasets. Moving forward, research should focus on overcoming these barriers to develop explainable clustering methods that are robust, scalable, and versatile enough for real-world applications across diverse settings.

2.5 Contribution Beyond Existing Research

This study builds upon the existing methodologies discussed in the reviewed literature while addressing critical gaps that remain unresolved. Unlike the works reviewed, this research does not propose a new methodology or develop a novel clustering approach. Instead, it focuses on leveraging and validating existing frameworks to address specific challenges associated with clustering, explainable AI, and the use of unlabeled spatial-temporal data.

The methodology presented by Bobek ([5]) demonstrates a strong foundation for clustering completely unlabeled data and introducing interpretability through human-readable rules. However, as discussed in the previous section, it has notable challenges, including complexity of implementation, dependency on data quality, scalability issues, and limited evaluation to benchmark datasets. No existing literature directly addresses the specific challenge of using unlabeled spatial-temporal data for supervised classification, particularly in real-world scenarios. This gap provides the basis for the objectives of this study.

The primary contribution of this research is validating whether pseudo-labeled data generated from completely unlabeled spatial-temporal data can effectively be used for supervised classification tasks. This is a critical advancement, as it bridges the gap between unsupervised clustering and supervised classification by using pseudo-labeled data as an intermediary. Furthermore, the application of these methodologies to real-world mobility data provides a practical context that goes beyond the benchmark datasets typically used in previous studies. This study focuses on clustering techniques combined with explainable AI to analyze spatial-temporal mobility data, ensuring that the results are not only easy to understand but also practical for real-world applications.

3 METHODOLOGY

The methodology adopted for this study follows a systematic two-step approach to classify unlabeled data by first applying clustering techniques to assign pseudo-labels, followed by classification. The process begins with data preprocessing to ensure that the data set is clean and ready for analysis. This includes handling missing values through imputation, removing or treating outliers, and scaling numerical features to maintain uniformity across the dataset.

The clustering step is the core of the methodology, where data points are grouped into clusters based on inherent patterns. Depending on the dataset's characteristics and research, algorithms such as ST-DBSCAN, DBSCAN and TRACLUS are selected. Hyperparameters, such as the density threshold and time threshold values, are optimized using KNN and trial and tested method. KNN is a method that computes k-nearest neighbors for each datapoint to understand density distribution of data for different values of 'k', where k refers to the number of nearest neighbors. MinPts which is the minimum number of data points required to form a dense region, is set to 'k'. The Trial-and-Test method involves making multiple guesses for the unknown value, each followed by an evaluation of their effectiveness to determine the optimal values for thresholds in our case. The clustering process assigns pseudo-labels to the data, effectively transforming it into a labeled dataset. The pseudo-labeled data is then used as input for supervised classification models using XGBoost classifier. The classifier is trained on a subset of the data and evaluated using metrics like Accuracy, Precision, Recall, F1-Score, while a confusion matrix provides insights into class-wise performance and misclassification trends. To ensure robustness, K-Fold Cross-Validation is performed. This methodology provides a structured framework for leveraging clustering as a foundation for effectively classifying unlabeled data. Refer Fig.3 for visualized methodology.

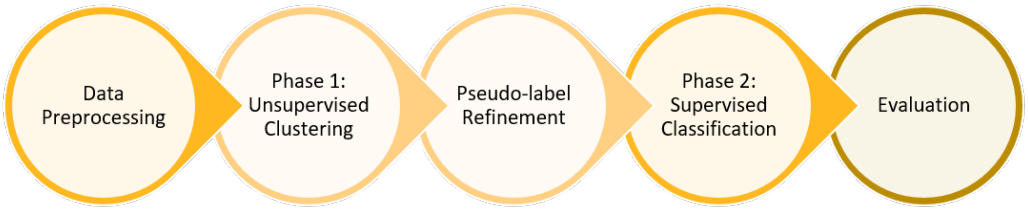


Fig. 3. Methodology used in the study

4 DATASET OVERVIEW

This section provides an overview of the datasets utilized for the analysis. It includes a detailed description of the data sources, structure, and key attributes relevant to the study. Additionally, it outlines the necessary data preparation steps that must be undertaken during the analysis process.

4.1 Dataset 1 - Mobility Uber Peru Data [16]

The Mobility Uber Peru dataset is a publicly available dataset hosted on Kaggle, providing detailed and extensive information about Uber ride trips conducted in the urban Peru city region for the year 2010. This dataset serves as a valuable resource for analyzing ride activity, offering insights into both temporal and spatial dynamics. This dataset includes key features such as start and end timestamps of each ride and the corresponding start and end

locations. The start and end timestamps for each trip, enable the analysis of ride durations, peak usage times, and temporal distribution of rides. The corresponding start and end locations, given in geographical coordinates, facilitate the examination of spatial trends, popular pick-up and drop-off points, and travel distances. With a total of 23,100 records, this dataset provides an overview of uber ride activity in Peru during this period. It offers insights for analyzing ride patterns and temporal and spatial dynamics, making it a good resource for research and exploratory data analysis.

4.2 Dataset 2 - Optimodal Hofer LandBus

The Optimodal Hofer LandBus dataset retrieved from a PostgreSQL database, contains detailed information about rides provided by the Hofer LandBus service, covering the rural area around Hof district region in Germany. This data was accessed and processed for the purpose of this analysis through pgAdmin. This dataset includes essential features such as the start and end locations of rides, timestamps for ride initiation and completion, and various other ride-related details. Spanning an entire year, from April 2023 to May 2024, the dataset captures a snapshot of ride activity in the region. With a total of 62,952 records, this dataset provides a foundation for analyzing temporal and spatial ride patterns of the LandBus service. The features pf the data like spatial and temporal details of ride makes it a good resource aiming for an in-depth analysis of mobility behavior, enabling to identify travel trends, optimize transportation services, and develop strategies tailored to regional needs.

4.3 Data Preparation

Data preparation for the source datasets begins with understanding the data and identifying the necessary steps. Initial preprocessing involved removing missing or incomplete records to prepare the data. These steps were necessary to prepare the dataset for analysis. Following the cleaning process, additional time-related features were generated to enhance the utility of the dataset for temporal analysis. New columns were created for analyzing the hourly distribution of rides, with 'start_hour' and 'end_hour' columns generated using a pseudo date and timestamps fetched from the original start and end times, respectively. The 'start_hour' and 'end_hour' columns were converted into a standard Date/Time format (YYYY-MM-DD HH:MM:SS) to represent start and end times clearly, as the time field does not work with the ST-DBSCAN algorithm in QGIS because the algorithm specifically requires the Date/Time column to be in a valid Date/Time format. Without proper formatting, such as YYYY-MM-DD HH:MM:SS, the algorithm cannot interpret the temporal values for clustering. These features added valuable temporal detail to the dataset. Another important step was transforming the spatial coordinates to the required coordinate reference system (CRS) specific to the region, enabling accurate spatial calculations. This transformation ensured that distances were expressed in meters, which was critical for clustering and analysis. All these steps were applied to both datasets, making them suitable for analysis and spatiotemporal clustering while ensuring consistency and reliability. After data preparation, the columns used for further analysis are start_hour / end_hour column in Date/Time format and start and end coordinates in Decimal format. The data preparation steps explained above is visualised in Fig 2.

Fig. 4. Data Preparation Steps

5 IMPLEMENTATION

This implementation is based on foundational paper by Bobek [5]. The author suggested a four phase approach referred as ClAMP methodology (Refer Fig.3) for enhancing cluster analysis on benchmark dataset. This section focuses on implementing two phases of the ClAMP methodology using a geographical dataset highlighted in Fig.3 , addressing challenges associated with unlabeled data.

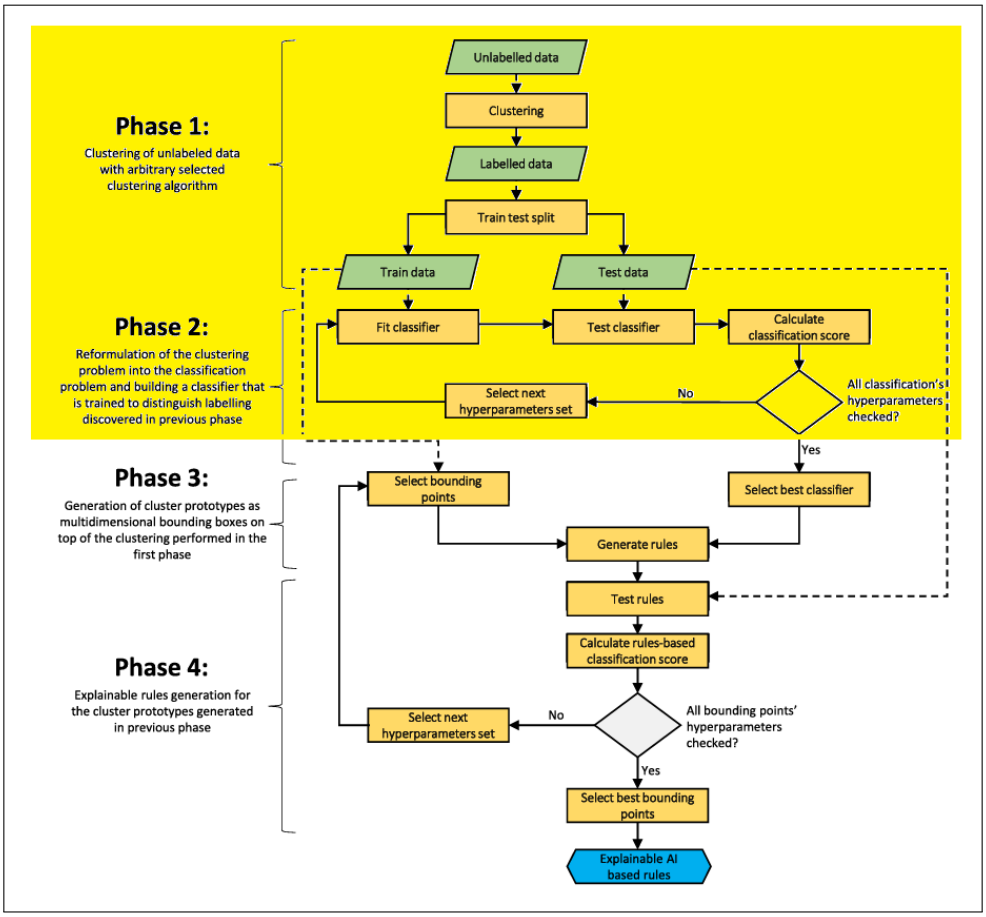


Fig. 5. ClAMP methodology reprinted by courtesy of [5]

5.1 Clustering

To perform clustering on the preprocessed data, it is essential to first select a suitable clustering algorithm and then determine the appropriate threshold values. This research focuses on identifying clustering methods suitable for spatiotemporal data with geographic and temporal components. ST-DBSCAN was chosen for its ability to handle both spatial and temporal dimensions, making it ideal for clustering points like ride start and end locations based on proximity and density while detecting outliers. TRACLUS, on the other hand, specializes in clustering trajectories, identifying common movement patterns by analyzing the shape, direction, and segmentation of

paths. Together, these algorithms address the unique challenges of geographic datasets, offering complementary approaches for analyzing both individual points and complete ride trajectories. However, the datasets used for this study consisted solely of start and end coordinates for each ride, lacking the intermediate points necessary to represent the complete ride trajectories. TRACLUS, being a trajectory-based clustering algorithm, relies heavily on detailed trajectory information to segment and cluster movement patterns effectively. Without intermediate points, the segmentation process and trajectory similarity calculations cannot be performed accurately. This limitation rendered TRACLUS unsuitable for the dataset. Consequently, ST-DBSCAN, which clusters data based on spatiotemporal proximity without requiring complete trajectory information, was taken as more appropriate approach, focusing on identifying high-demand hotspots for ride pickups and drop-offs across different times and locations.

QGIS, an open-source Geographic Information System renowned for its capabilities in spatial analysis and data visualization, was utilized for the implementation of the clustering methodology. To apply the ST-DBSCAN algorithm to the preprocessed dataset, four key parameters were required: a Date/time field, minimum points (MinPts), a distance threshold, and a time threshold.

For analyzing the hourly distribution of the start and end points of ride, the "start_hour" or "end_hour" column was selected as the Date/Time field. The MinPts parameter, which specifies the minimum number of points required to form a cluster, was set to 5. This value is the default minimum cluster size in QGIS and serves as a practical starting point for density-based clustering. By using this default, the clustering process identifies regions with at least five points as clusters, while smaller groupings are treated as noise. This choice aligns with the QGIS standard and simplifies parameter selection without requiring additional tuning. The distance threshold, determining the spatial proximity necessary for points to be considered part of the same cluster, was varied across multiple values: 50 meters, 100 meters, 200 meters, 500 meters, 1000 meters, and 2000 meters. Similarly, the time threshold, which defines the temporal proximity required for clustering, was tested using several predefined values: 5 minutes, 10 minutes, 30 minutes, 1 hour, and 5 hours. These thresholds were adjusted between separate clustering iterations, rather than during a single clustering run, to evaluate their impact on the results. The goal was to evaluate and identify the parameter values that produced the most meaningful clustering results for the input dataset which has moderate number of clusters while minimizing number of outliers. In this context, "most meaningful" refers to clustering results that clearly capture spatial-temporal patterns, align with real-world mobility trends, and reduce noise points. By running the algorithm multiple times with different parameter combinations, this process explored spatiotemporal clustering behavior under set conditions.

Following the clustering process, detailed analysis were conducted to assess the results. Box plots were created to visualize spatial and temporal variations within clusters. For distance variations, metrics include cluster size, max distance, min distance, spread, and compactness. Cluster size indicates the number of data points in each cluster, while max and min distances reflect the spatial extent and proximity of points. Spread shows the range of distances within a cluster, and compactness measures how tightly points are grouped. These metrics help compare clusters by their spatial distribution. For time variations, metrics include time span, mean time, min time, max time, and time spread. Time span captures the range of time points in a cluster, mean time represents the average time, and time spread indicates the variability in temporal activity. These metrics highlight differences in cluster activity durations and patterns over time. Additionally, a heatmap was generated to depict the number of outliers identified in each clustering scenario, offering a clear visual representation of how parameter changes influenced the clustering outcomes. Fig. 6 to 9 depicts box plots for distance and time variation for optimal paramaters selected. Box plots for other threshold values and heatmaps are provided in the Appendix section for reference. Based on these analysis, the optimal parameter values were determined to be a distance threshold of 200 meters and a time threshold of 1 hour for start points of Dataset 1. These values provided the best balance between identifying meaningful clusters and minimizing the number of outliers, ensuring interpretable clustering results for the dataset. Similarly, this method was applied to analyze the end points of Dataset 1 as well as the start and

end points of Dataset 2. For the end points of Dataset 1, the optimal parameters were determined to be a distance threshold of 500 meters and a time threshold of 5 hours. In the case of Dataset 2, the start points yielded the best results with a distance threshold of 500 meters and a time threshold of 1 hour, while the end points of Dataset 2 also performed optimally with the same parameter values of 500 meters and 1 hour. This consistent methodology enabled an evaluation of spatio-temporal clustering using optimal parameter values. Optimal paramater values mentioned above can be seen in Table 2. With these optimal parameters, the clustering results are shown in Fig. 10 and Fig. 11 for Dataset 1, and Fig. 12 and Fig. 13 for Dataset 2. The clusters shown are sorted from highest to lowest based on cluster size, with legends provided for the top 20 clusters. Fig. 14 to 17 depict the same clustering result using different symbology to represent the demand at start and end points based on the formed clusters.

Dataset	Start Points (Spatial)	Start Points (Time)	End Points (Spatial)	End Points (Time)
Dataset 1	200m	1 hr	500m	5 hr
Dataset 2	500m	1 hr	500m	1 hr

Table 2. Optimal spatial and temporal thresholds for clustering in Dataset 1 and Dataset 2

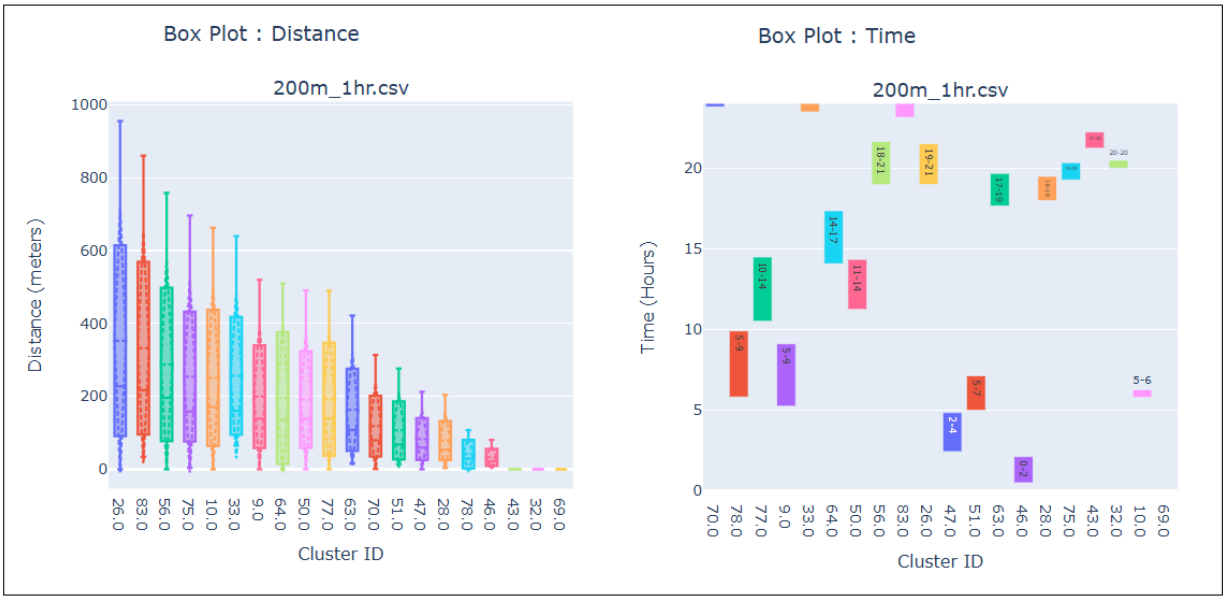


Fig. 6. Box plots showing the variation in spatial and temporal metrics for Dataset 1 start points

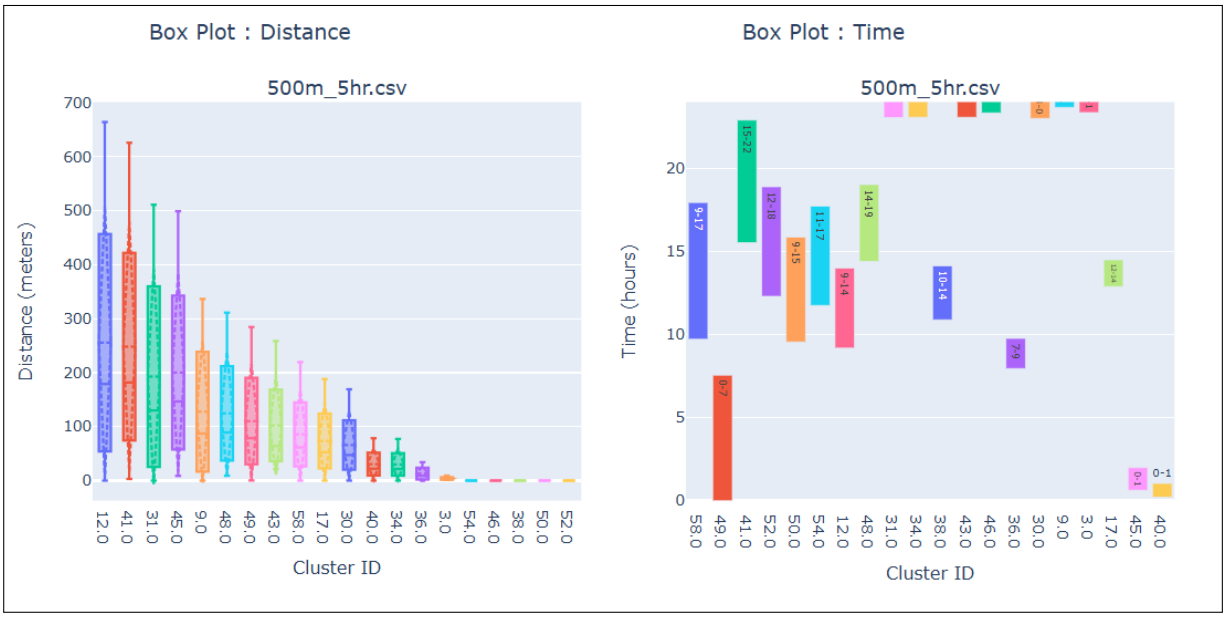


Fig. 7. Box plots showing the variation in spatial and temporal metrics for Dataset 1 end points

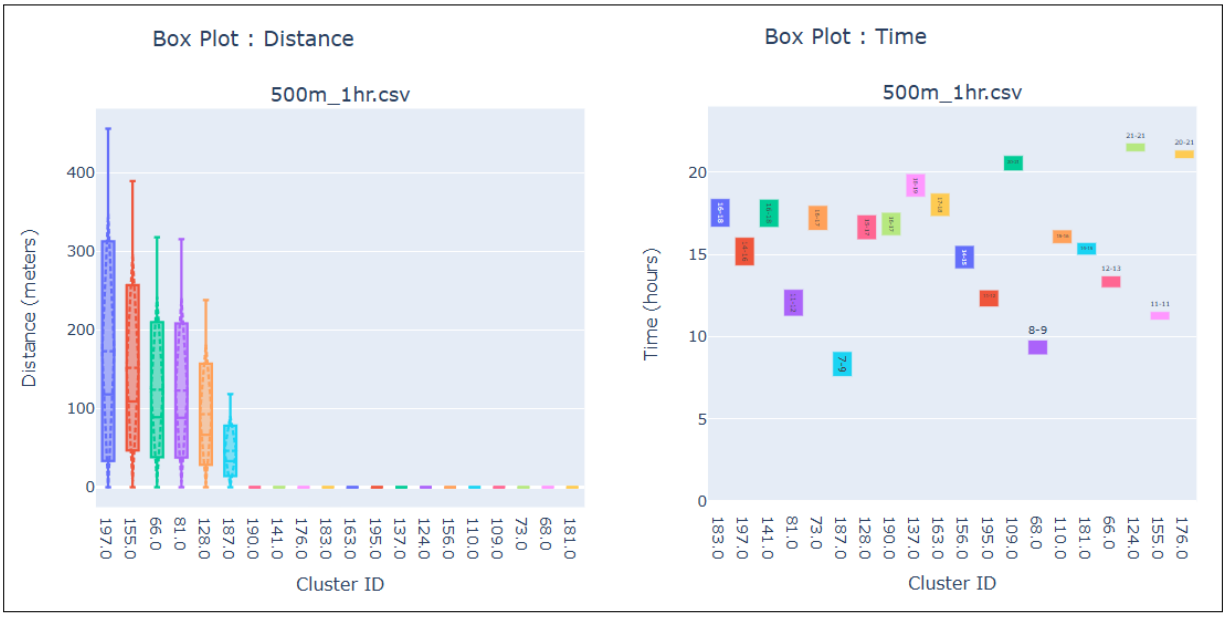


Fig. 8. Box plots showing the variation in spatial and temporal metrics for Dataset 2 start points

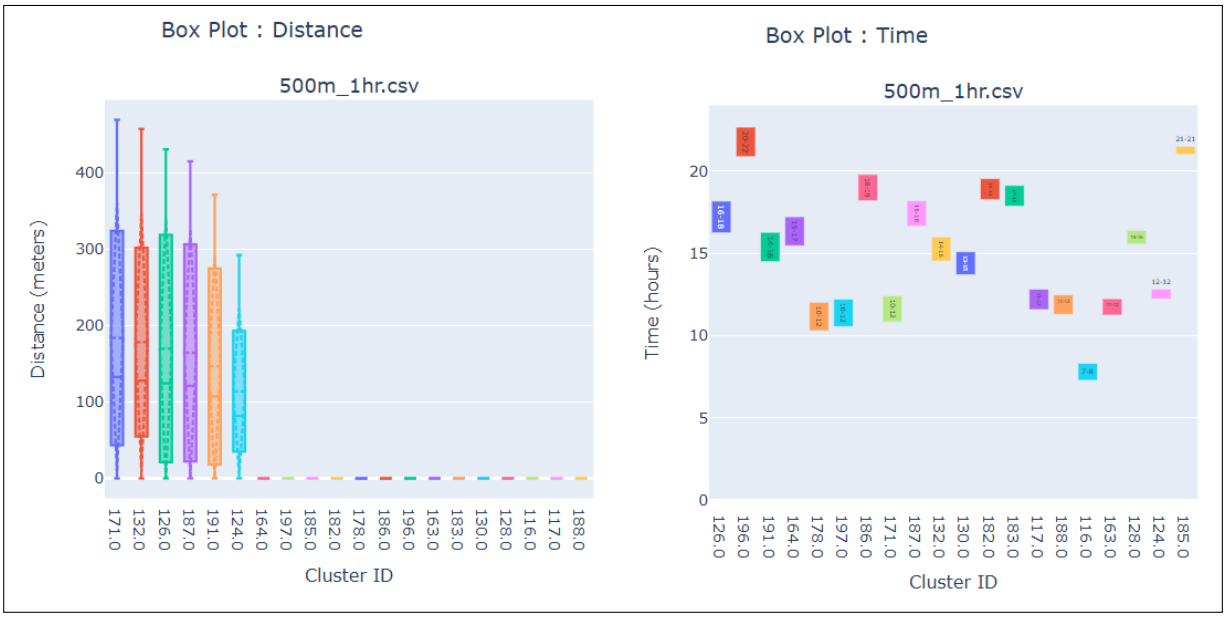


Fig. 9. Box plots showing the variation in spatial and temporal metrics for Dataset 2 end points

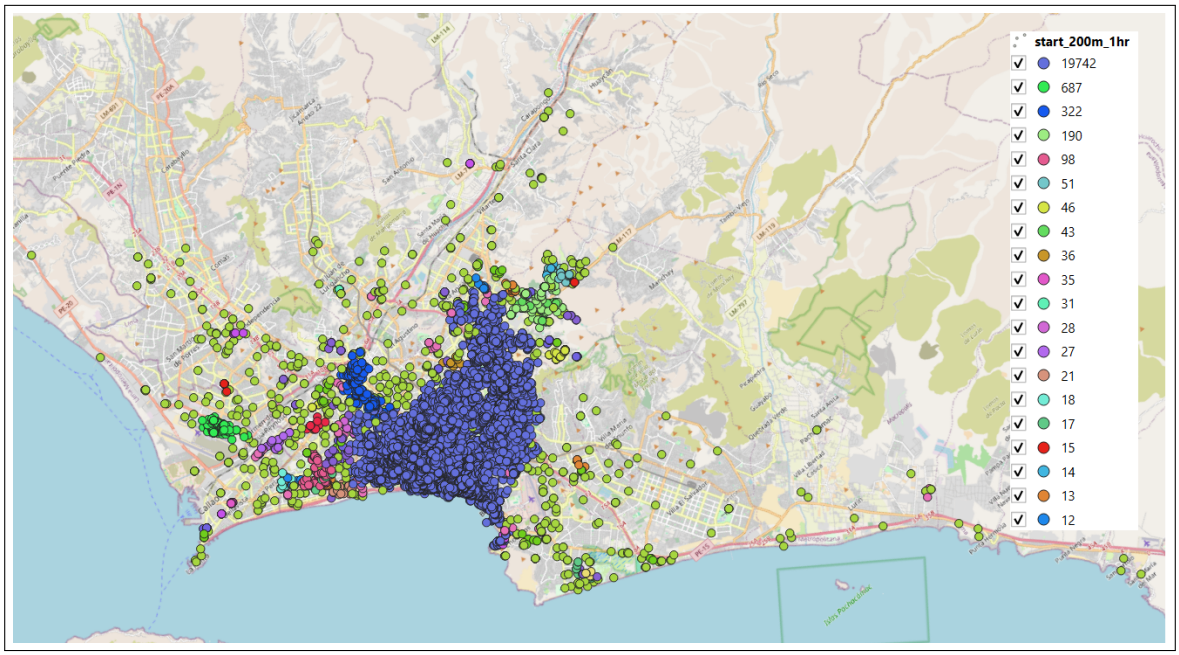


Fig. 10. Spatial Distribution of Clusters for Dataset 1 Start Points: Map data from OpenStreetMap [1]

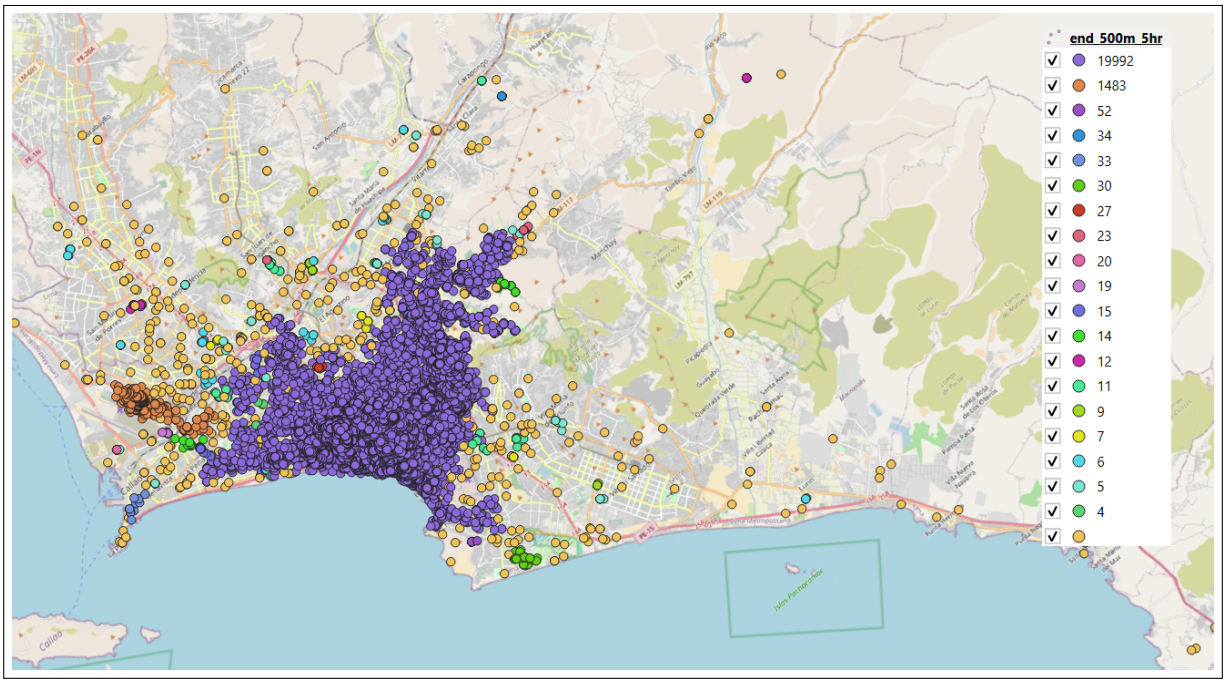


Fig. 11. Spatial Distribution of Clusters for Dataset 1 End Points: Map data from OpenStreetMap [1]

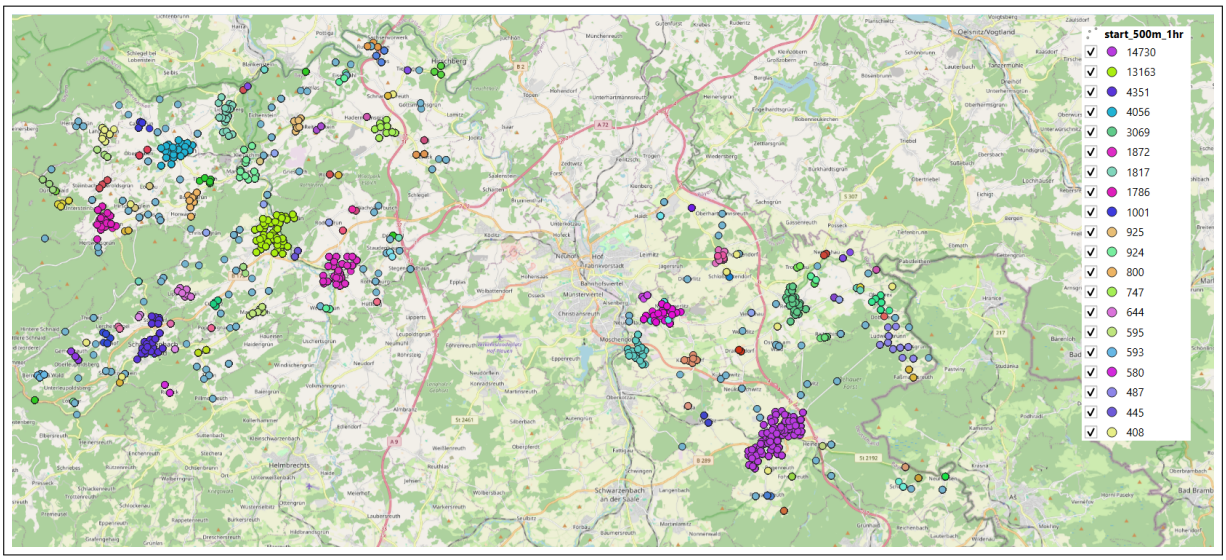


Fig. 12. Spatial Distribution of Clusters for Dataset 2 Start Points: Map data from OpenStreetMap [1]

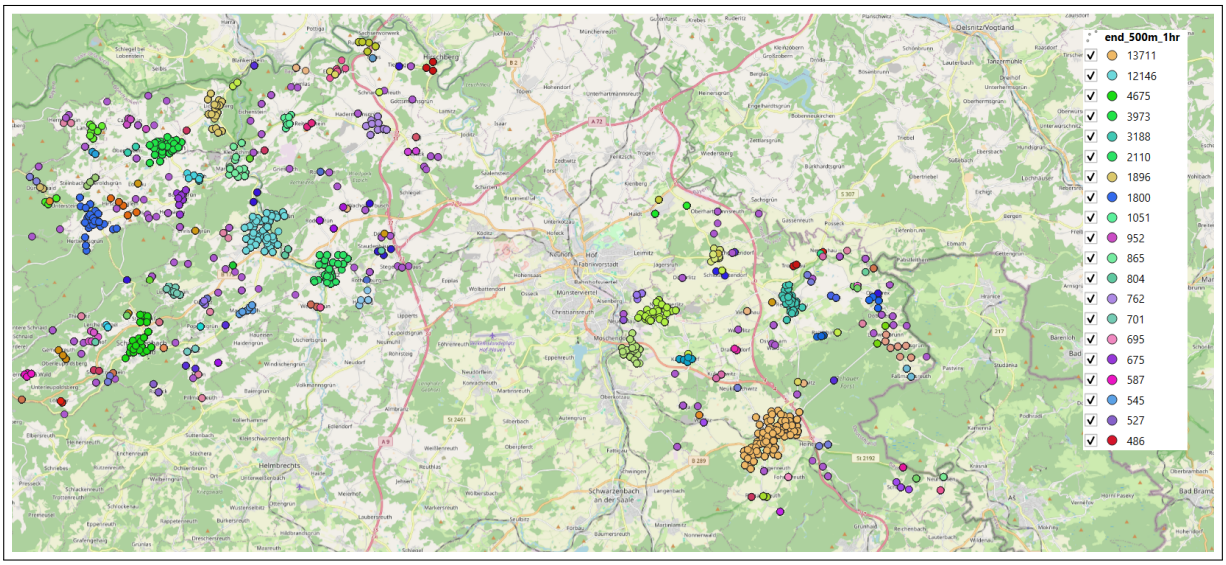


Fig. 13. Spatial Distribution of Clusters for Dataset 2 End Points: Map data from OpenStreetMap [1]

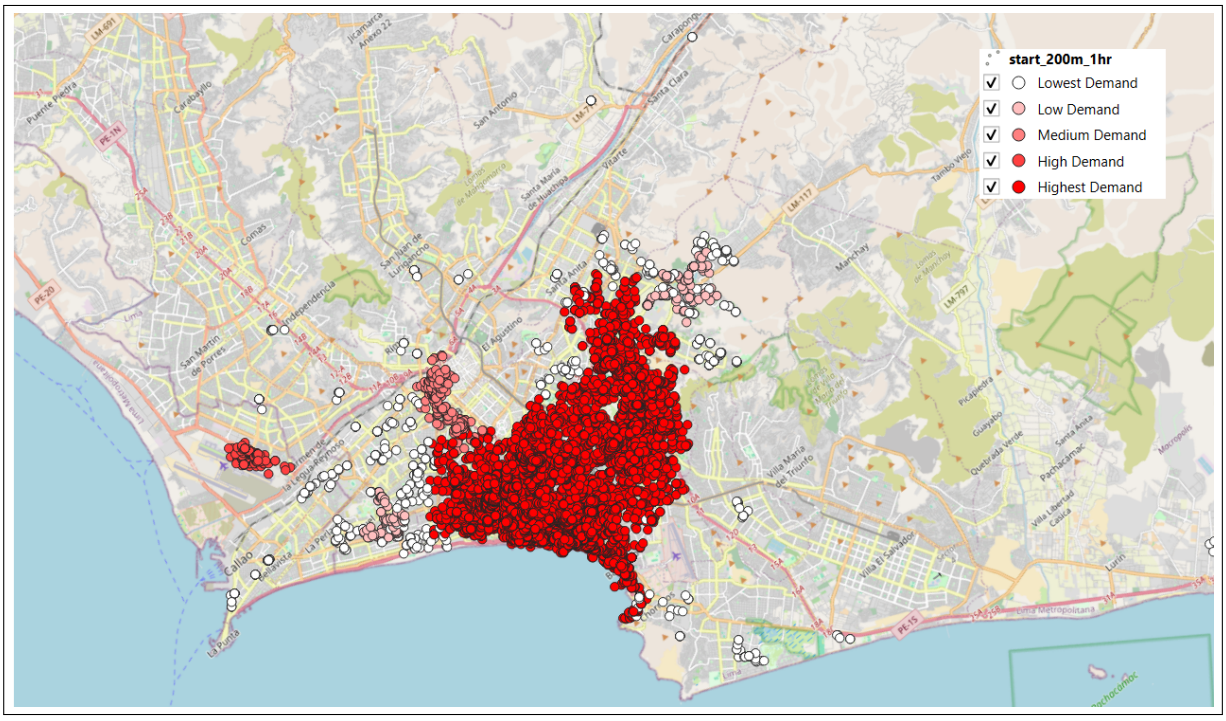


Fig. 14. Spatial Distribution of Clusters for Dataset 1 Start Points based on demand: Map data from OpenStreetMap [1]

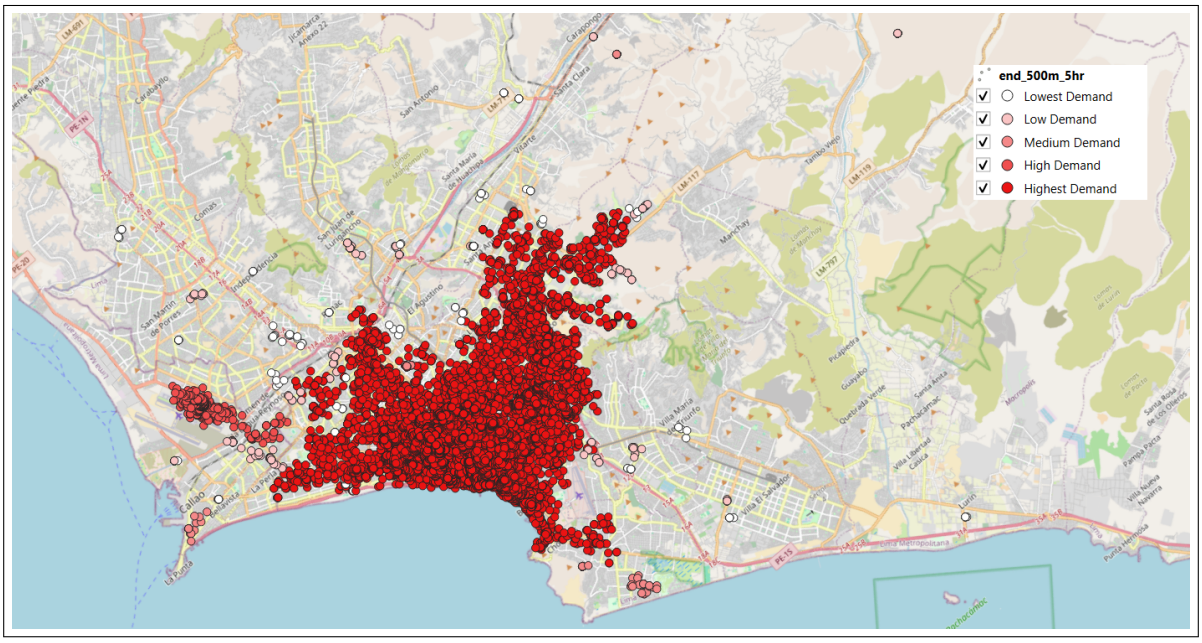


Fig. 15. Spatial Distribution of Clusters for Dataset 1 End Points based on demand: Map data from OpenStreetMap [1]

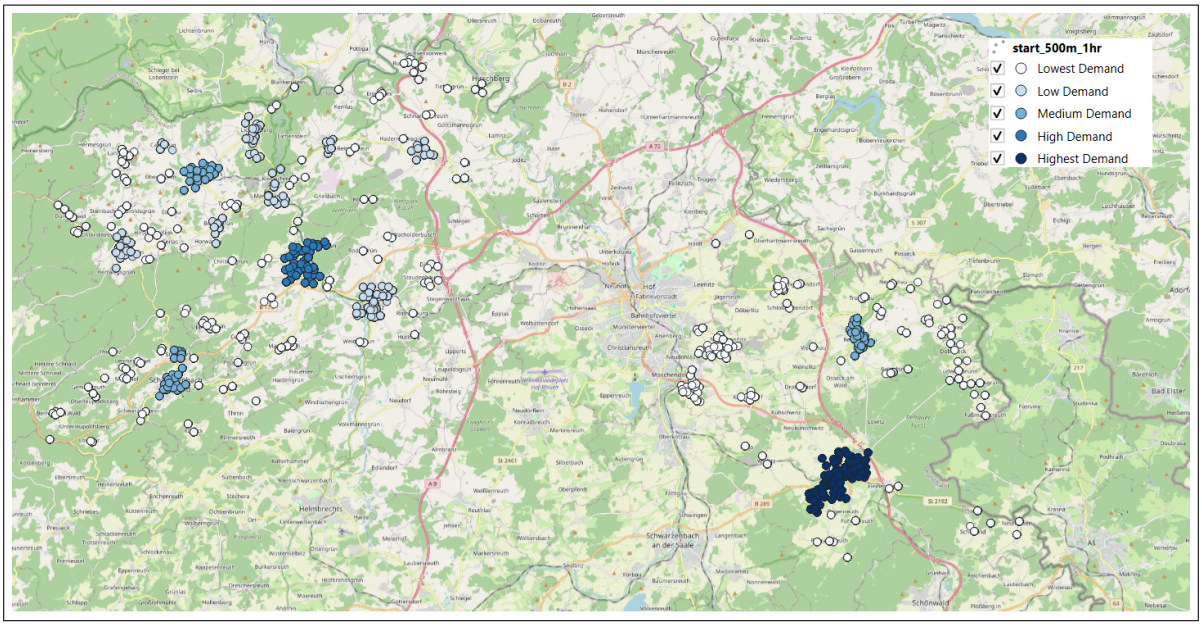


Fig. 16. Spatial Distribution of Clusters for Dataset 2 Start Points based on demand: Map data from OpenStreetMap [1]

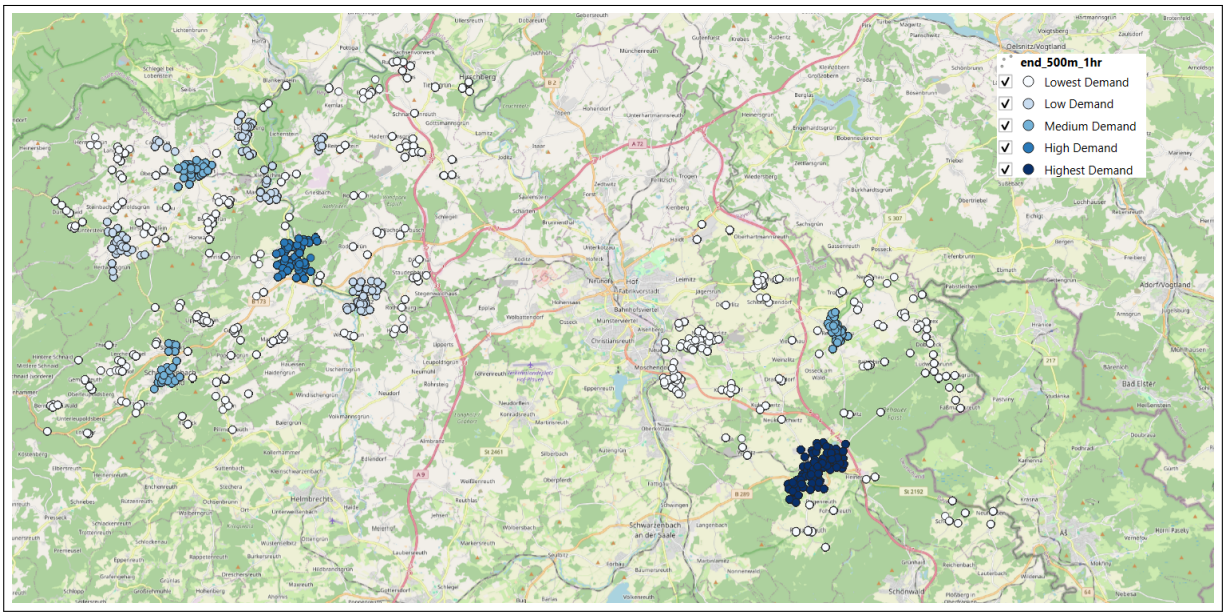


Fig. 17. Spatial Distribution of Clusters for Dataset 2 End Points based on demand: Map data from OpenStreetMap [1]

5.2 Classification

While clustering successfully identifies patterns in the data and groups similar points, it does not inherently provide a mechanism for making predictions or assigning labels to new, unseen data. Hence, the next phase of the study involves reformulating the clustering problem into a classification task by building a classifier trained to distinguish the pseudo-labels discovered in the clustering phase. This step is essential to evaluate how effectively the pseudo-labeled data can be used for supervised learning. Supervised classification is a machine learning approach where a model learns to assign labels to data points based on input features, using a labeled dataset during training to understand the relationships between the features and the target labels. In this case, the goal is to test whether the classifier can accurately distinguish the labels assigned during clustering. This process demonstrates the ability of the classifier to accurately learn and replicate the clustering labels, thereby validating the quality and consistency of the pseudo-labels generated during clustering, while also showcasing the potential of leveraging pseudo-labeled data for real-world predictive applications.

To perform classification on the large dataset, the XGBoost (Extreme Gradient Boosting) classifier was chosen. XGBoost is a powerful machine learning algorithm that excels in handling large datasets with many features [7]. It is well-known for its high accuracy and efficiency, making it a popular choice for solving complex problems. One of its key strengths is its ability to handle imbalanced datasets effectively, which is important when working with clustering results where some clusters may dominate over others[7]. Additionally, XGBoost is designed for speed and scalability. Its support for parallel processing and sparse data handling ensures it can manage large datasets efficiently without compromising performance[7]. Another advantage of XGBoost is its built-in regularization, which helps prevent overfitting and ensures the model generalizes well to new data[7]. Overfitting is not a major concern in this study because the pseudo-labels derived from clustering are inherently approximate, and the primary goal is to evaluate broader patterns rather than perfect accuracy on the training data. By applying XGBoost to the pseudo-labeled dataset, this phase aims to evaluate the reliability and predictive value of the labels

generated through clustering. The reliability of the pseudo-labels lies in their ability to reflect meaningful patterns in the data, as determined by the clustering process, while the predictive value of these labels is assessed by the model's performance in distinguishing between different clusters when presented with unseen data. This process bridges the gap between unsupervised clustering, which groups data without labels, and supervised classification, which assigns definitive labels to data points for prediction. Bridging this gap allows the pseudo-labeled data to be leveraged for practical usecases, demonstrating its potential to enhance decision-making in real-world scenarios. To perform classification on the sample dataset, the decision was made to use the top 20 largest clusters identified through clustering. This sample was chosen because these clusters account for the majority of the data points, representing approximately 80% of the dataset. By focusing on the most data-rich clusters, the classification process prioritizes areas where predictions are most meaningful and actionable, rather than allocating resources to smaller, less representative clusters. While the largest clusters are not necessarily the most significant in all contexts, they provide a practical starting point for testing the methodology. Narrowing the scope to these clusters was necessary due to the large number of clusters generated during clustering. Performing classification on hundreds of clusters would have been impractical and computationally expensive, especially since 80% of the data is already represented in the top 20 clusters. This optimization ensures faster processing and focuses on the most impactful portions of the dataset without sacrificing representativeness. The labeled dataset is prepared for training by selecting relevant features and labels. The features include latitude, longitude, and time in hours, while the target variable is the cluster ID (Label). The data is split into training and testing sets using an 80:20 ratio through the train_test_split function, ensuring that the model can be evaluated on unseen data to measure its performance accurately. An XGBoost classifier is employed for the classification task due to its efficiency and ability to handle large datasets effectively. The classifier is initialized with use_label_encoder=False to suppress a deprecation warning and eval_metric='mlogloss' for multi-class classification. The model is then trained on the training set using the fit method and used to predict labels for the test set with the predict method.

The evaluation of the XGBoost classifier began with calculating its accuracy on the test dataset using the "accuracy_score" function, which measures the proportion of correctly predicted labels. To gain deeper insights into the model's performance, a detailed classification report was generated using the "classification_report" function. This report provides key metrics, including precision, recall, F1-score, and support for each class, offering a comprehensive view of how well the classifier performs across all categories. Next, a confusion matrix was created using the "confusion_matrix" function to further assess the model's predictions. The confusion matrix highlights the number of true positives, true negatives, false positives, and false negatives for each class, enabling a detailed evaluation of the classifier's strengths and weaknesses. To make the results more interpretable, a heatmap of the confusion matrix was plotted using Seaborn, providing a visual representation of the classification results. This visualization annotates the matrix values and includes clear labels for predicted and true classes, making it easier to identify patterns or discrepancies in the predictions. The classification report and confusion matrix results for both datasets are in Fig. 18 to 21. Additionally, a 5-fold cross-validation procedure was performed using the "cross_val_score" function to ensure the robustness and generalizability of the classifier. In this process, the dataset was split into five subsets, and the model was iteratively trained and validated on different combinations of these subsets. The accuracy scores from each fold were calculated, and the mean accuracy and standard deviation were computed to summarize the classifier's performance across the folds. This step ensures that the model is not overly reliant on specific data subsets and provides a reliable estimate of its predictive capabilities on unseen data. Overall, these steps ensure a thorough and multi-faceted evaluation of the XGBoost classifier, combining direct accuracy measures, detailed metrics, visual analysis, and cross-validation to validate its performance comprehensively.

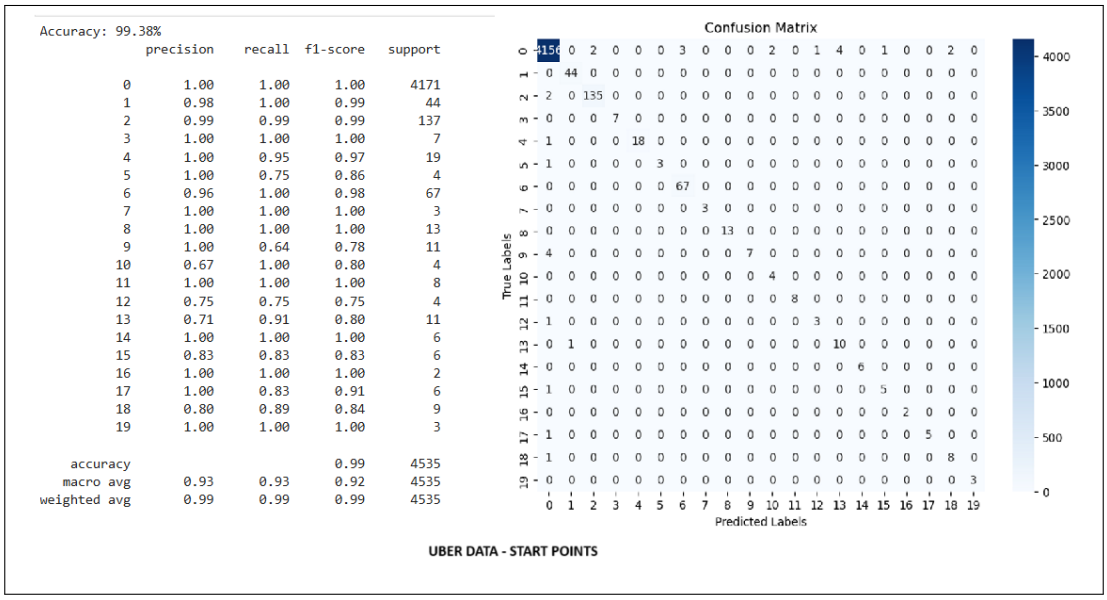


Fig. 18. Model Evaluation Metrics and Confusion Matrix for Start Point Classification in Dataset 1

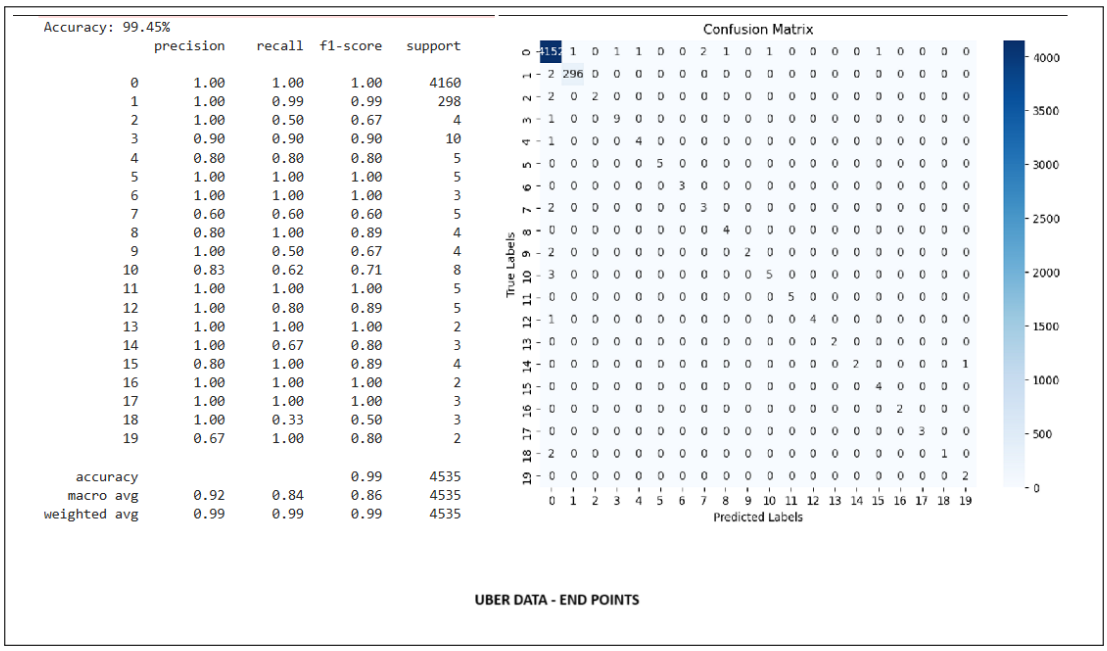


Fig. 19. Model Evaluation Metrics and Confusion Matrix for End Point Classification in Dataset 1

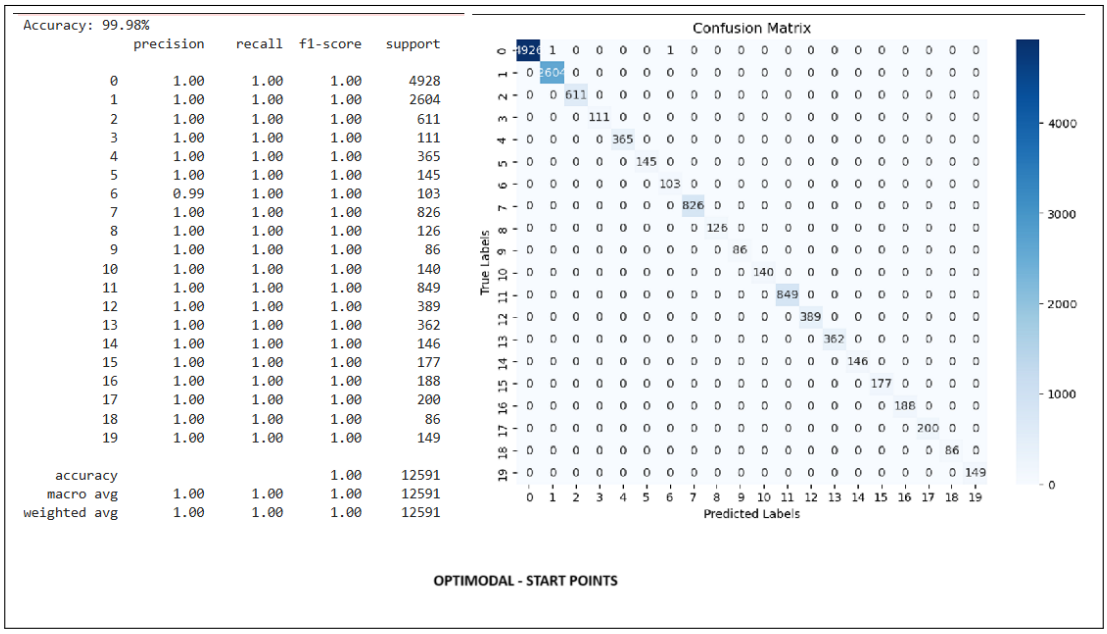


Fig. 20. Model Evaluation Metrics and Confusion Matrix for Start Point Classification in Dataset 2

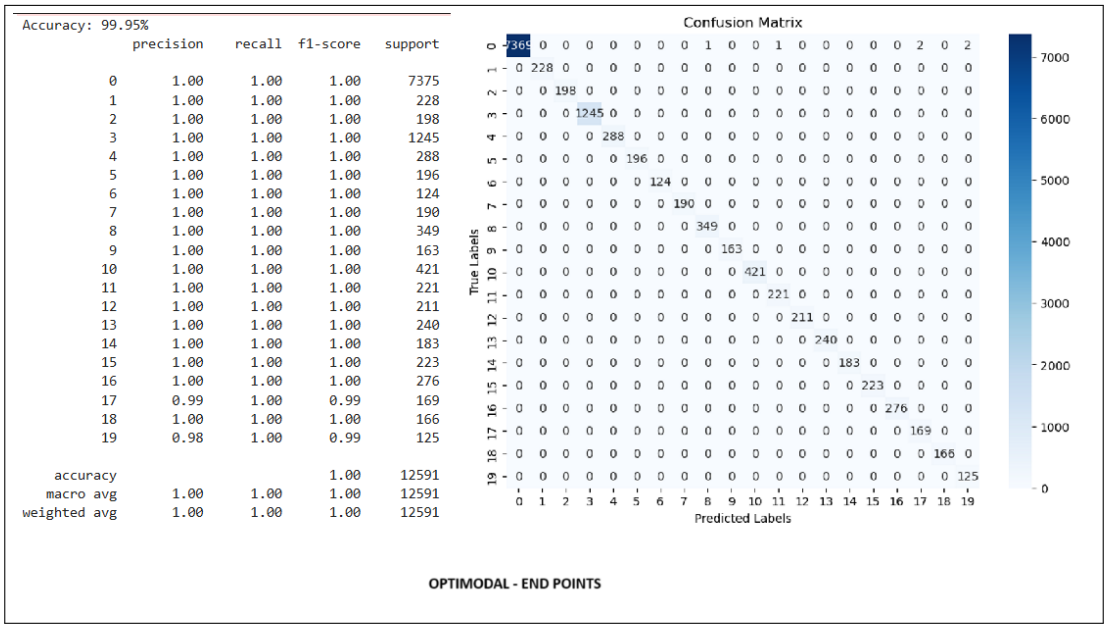


Fig. 21. Model Evaluation Metrics and Confusion Matrix for End Point Classification in Dataset 2

6 RESULT

This section discusses about the results achieved from above implementation on spatial-temporal mobility data. Using ST-DBSCAN for the start and end points in both datasets, we identified distinct mobility patterns. The clustering maps for Dataset 1 revealed concentrated areas of high demand, highlighting urban regions where transportation planning efforts could be focused. In Dataset 2, the clustering maps showed a broader geographic distribution, reflecting the inclusion of intercity mobility patterns in the data set.

In Dataset 1, the clustering process revealed highly concentrated urban hotspots. These clusters identified areas with high ride demand, particularly in central urban regions, where mobility patterns are more compact. The optimal clustering parameters were determined using the Trial-and-Test method, with box plots visualizing distance and time variations across clustering results, and a heatmap displaying the number of outliers. For start points, the ideal parameters were a spatial threshold of 200 meters and a temporal threshold of 1 hour. For end points, slightly larger thresholds of 500 meters spatially and 5 hours temporally yielded the most meaningful clusters while reducing noise. The classification phase for Dataset 1 demonstrated robust performance, with XGBoost achieving accuracies of 99.38% for start points and 99.45% for end points. These results indicate the effectiveness of the pseudo-labeling process, as the high classification accuracy suggests that the pseudo-labeled data provided meaningful input for supervised learning. Precision, recall, and F1-scores were consistently high across all major clusters, confirming that the pseudo-labels accurately captured the underlying patterns in the data. The confusion matrices indicated that most misclassifications occurred in smaller clusters, such as Cluster 13 in start points and Cluster 9 in end points, likely due to sparse data in those clusters. Despite these minor issues, precision, recall, and F1-scores were consistently high across all major clusters, indicating that the pseudo-labeling process was effective. The clustering maps for Dataset 1 highlight the potential for this approach to guide urban transportation planning. By identifying high-demand areas, policymakers can better allocate resources such as ride-sharing vehicles or public transport services to match demand patterns.

Dataset 2 represents a broader geographic range, with more dispersed clusters that reflect intercity travel patterns. The clustering maps for start and end points showed clusters spread across multiple towns and rural areas. The optimal clustering parameters for this dataset were a spatial threshold of 500 meters and a temporal threshold of 1 hour for both start and end points. These settings effectively captured the broader distribution of intercity trips while maintaining meaningful cluster definitions. The classification results for Dataset 2 were even stronger, with XGBoost achieving near-perfect accuracies of 99.98% for start points and 99.95% for end points. The confusion matrices showed almost no misclassifications, even for smaller clusters. Precision, recall, and F1-scores for all clusters were consistently at or near 1.0, underscoring the high quality of the pseudo-labels generated during clustering. The intercity distribution in Dataset 2 highlights its applicability for regional transportation planning. By understanding intercity travel patterns, planners can optimize resources such as long-distance buses or rail services to meet demand effectively.

Overall, this approach proved highly effective for both datasets. The combination of meaningful feature engineering including the transformation of raw spatial and temporal data into suitable clustering features, suitable clustering algorithms, and a robust classifier like XGBoost contributed to the success of the methodology. Dataset 1 highlighted the utility of this approach for dense urban environments, while Dataset 2 showcased its applicability for broader, intercity mobility patterns. However, the presence of smaller clusters and outliers in both datasets suggests room for improvement. Including more detailed trajectory points or intermediate points of a ride could further be more interesting to analyse.

7 DISCUSSION

In the literature review section, we discussed several studies and the challenges identified in them. While some research follows a similar approach to this paper, they often either lack applicability to real-world scenarios or fail to generalize effectively across diverse datasets. The primary objective of this study was to validate the effectiveness of pseudo-labeled data, derived from completely unlabeled spatial-temporal data, for supervised classification tasks. Furthermore, the broader objective was to explore the potential of utilizing unlabeled data to enable meaningful supervised learning, addressing a critical gap in the field. The points outlined in the "Contribution Beyond Existing Research" section served as the foundation for these objectives, and this discussion evaluates the extent to which these goals were achieved.

The results discussed above demonstrated that pseudo-labeling can effectively bridge the gap between unsupervised clustering and supervised classification. This was evident in the high classification accuracy achieved by XGBoost, along with consistently strong precision, recall, and F1-scores. These metrics confirm that the pseudo-labels generated through clustering successfully represented meaningful patterns in the data, allowing the classifier to differentiate between clusters effectively. Furthermore, the low misclassification rates in major clusters further validated the robustness of the pseudo-labeling process. The application of clustering techniques to real-world mobility data also proved successful. By analyzing spatial-temporal data, the study identified urban hotspots and areas based on start and end points of the ride. The clustering maps revealed areas of concentrated high ride demand, providing actionable insights for urban transportation planning. By moving beyond benchmark datasets, this study demonstrated the practical applicability of clustering techniques to real-world scenarios, addressing a critical gap highlighted in the literature. While the study achieved its primary objectives, it also identified some limitations. Misclassification was more prevalent in smaller clusters with sparse data, reflecting a broader challenge in handling low-density data. Another observation was related to the nature of the datasets. Dataset 1, representing urban city data, resulted in clusters where one cluster was highly dominant compared to others. This is expected in urban areas, as city centers tend to be more crowded due to the concentration of facilities and activities. However, this dominance made it challenging to pinpoint specific hotspots within the largest cluster, as it provided only a broader overview of popular demand areas rather than detailed insights. In contrast, Dataset 2, which comprised intercity ride data from a rural region, resulted in clusters that were smaller and more evenly distributed. This contrast highlights the influence of data context on clustering outcomes.This may indicate that clustering techniques, as applied here, may not work as effectively in densely populated urban areas. In such areas, the overwhelming dominance of certain clusters (like city centers) can obscure finer patterns and make it difficult to extract meaningful insights about smaller, less dominant clusters. Another limitation was the absence of intermediate ride points, as the data included only start and end points. Analyzing trajectories with intermediate points would be more interesting to analysis and could provide a more detailed understanding of mobility patterns and enhance the insights gained from clustering.

Hence, this research successfully validated the use of pseudo-labeling for supervised classification and applied clustering to real-world mobility data. These contributions address critical literature gaps and provide a strong foundation for future work.

8 CONCLUSION AND FUTUREWORK

This research addressed critical challenges in clustering and supervised classification by validating the use of pseudo-labeled data generated from completely unlabeled spatial-temporal data. It demonstrated how clustering techniques could bridge the gap between unsupervised and supervised learning, providing a practical and scalable solution to the lack of labeled data. By applying these techniques to real-world urban and rural mobility datasets, the study introduced a structured methodology to analyze mobility patterns, identify high-demand areas, and

derive meaningful insights. The approach showcased the utility of pseudo-labeling in creating reliable datasets for supervised learning, even when no labeled data is initially available.

The findings of this research have practical implications for mobility analysis and similar domains where labeled data is either unavailable or costly to obtain. By validating the effectiveness of pseudo-labeled data for supervised classification, the study highlights how clustering techniques can uncover meaningful spatial-temporal patterns in both urban and rural contexts. For instance, the clustering results provided a broad understanding of urban hotspots and identified demand clusters in rural intercity rides. These insights, while not yet fully integrated into decision-making processes, demonstrate the potential for clustering techniques to support further refinement of mobility planning and resource optimization.

The next phase of this research will focus on integrating explainable AI into the clustering and classification processes, specifically by implementing the next stages of the ClAMP methodology. This involves introducing human-readable rule-based explanations that can provide insights into how clusters are formed and how classifications are made. By enhancing the interpretability of both clustering and classification outcomes, the integration of explainable AI will make the results more transparent and accessible to stakeholders. This will not only help identify key mobility patterns but also enable decision-makers to understand and trust the results, which is crucial for applying these insights to real-world problems such as transportation planning and resource allocation. Future work will also address challenges identified in this study, such as the dominance of certain clusters in urban areas, by exploring ways to refine cluster analysis and provide more granular insights. Additionally, expanding the scope to include trajectory data, rather than limiting the analysis to start and end points, will allow for a more comprehensive understanding of mobility patterns. This extension will provide deeper insights into route choices, travel behaviors, and temporal dynamics, making the methodology more adaptable to diverse datasets and applications.

REFERENCES

[1] [n. d.]. Map data from OpenStreetMap, https://www.openstreetmap.org/copyright.

[2] Miguel Alvarez-Garcia, Raquel Ibar-Alonso, and Mar Arenas-Parra. 2024. A comprehensive framework for explainable cluster analysis. *Information Sciences* 663 (2024), 120282. https://doi.org/10.1016/j.ins.2024.120282

[3] E. Bair. 2013. Semi-supervised clustering methods. *Wiley Interdiscip Rev Comput Stat* 5, 5 (2013), 349–361. https://doi.org/10.1002/wics.1270

[4] Derya Birant and Alp Kut. 2007. ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data Knowledge Engineering* 60, 1 (2007), 208–221. https://doi.org/10.1016/j.datak.2006.01.013 Intelligent Data Mining.

[5] Szymon Bobek, Michał Kuk, Maciej Szelążek, and Grzegorz Nalepa. 2022. Enhancing Cluster Analysis With Explainable AI and Multidimensional Cluster Prototypes. *IEEE Access* PP (01 2022), 1–1. https://doi.org/10.1109/ACCESS.2022.3208957

[6] Nils Breyer, Clas Rydergren, and David Gundlegård. 2022. Semi-supervised Mode Classification of Inter-city Trips from Cellular Network Data. *Journal of Big Data Analytics in Transportation* 4, 1 (April 2022), 23–39. https://doi.org/10.1007/s42421-022-00052-9

[7] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016). https://api.semanticscholar.org/CorpusID:4650265

[8] Syed Adil Hussain, Muhammad Umair Hassan, Wajeeha Nasar, Sara Abdelwahab Abdelghani Ghorashi, Mona Jamjoom, Abdel-Haleem Abdel-Aty, Amna Parveen, and Ibrahim A. Hameed. 2023. Efficient Trajectory Clustering with Road Network Constraints Based on Spatiotemporal Buffering. 12, 3 (2023), 117–117. https://doi.org/10.3390/ijgi12030117

[9] A. K. Jain, M. N. Murty, and P. J. Flynn. 1999. Data clustering: a review. *ACM Comput. Surv.* 31, 3 (sep 1999), 264–323. https://doi.org/10.1145/331499.331504

[10] Praveena Kumari M K, D. H. Manjaiah, and K Ashwini. 2024. Clustering Algorithms to Analyse Smart City Traffic Data. *International Journal of Advanced Computer Science and Applications* 15, 8 (2024). https://doi.org/10.14569/ijacsa.2024.0150811

[11] O. A. Kaminska, Katarzyna Kaczmarek-Majer, and Olgierd Hryniewicz. 2022. Impact of clustering of unlabeled data on classification: case study in bipolar disorder. 931–934. https://doi.org/10.15439/2022F210

[12] Jacob R. Kauffmann, Malte Esders, Lukas Ruff, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2019. From Clustering to Cluster Explanations via Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems* 35 (2019), 1926–1940. https://api.semanticscholar.org/CorpusID:189998790

[13] Jae-Gil Lee, Jiawei Han, and Kyu-Young Whang. 2007. Trajectory clustering: a partition-and-group framework. In *ACM SIGMOD Conference*. https://api.semanticscholar.org/CorpusID:18004950

[14] Shucong Li and Yujiao Zhan. 2024. An Improved Semi-Supervised Learning Algorithm Used for Classification Prediction. 000 (2024), 353–358. https://doi.org/10.1109/icmiii62623.2024.00071

[15] Zhenliang Ma and Pengfei Zhang. 2022. Individual mobility prediction review: Data, problem, method and application. *Multimodal transportation* 1, 1 (2022), 100002–100002. https://doi.org/10.1016/j.multra.2022.100002

[16] marcusRB. 2019. Mobility Uber Perú dataset. data retrieved from kaggle, https://www.kaggle.com/datasets/marcusrb/uber-peru-dataset/data.

[17] Mohammad Peikari, Sherine Salama, Sharon Nofech-Mozes, and Anne L. Martel. 2018. A Cluster-then-label Semi-supervised Learning Approach for Pathology Image Classification. *Scientific Reports* 8, 1 (May 2018), 7193. https://doi.org/10.1038/s41598-018-24876-0

A APPENDIX

Fig. 22. Box Plots: Uber Ride Start Points Across Distance

Fig. 23. Box Plots: Uber Ride Start Points Across Time

Fig. 24. Box Plots: Uber Ride End Points Across Distance

Fig. 25. Box Plots: Uber Ride End Points Across Time

Fig. 26. Box Plots: Optimodal Start Points Across Distance

Fig. 27. Box Plots: Optimodal Start Points Across Time

Fig. 28. Box Plots: Optimodal End Points Across Distance

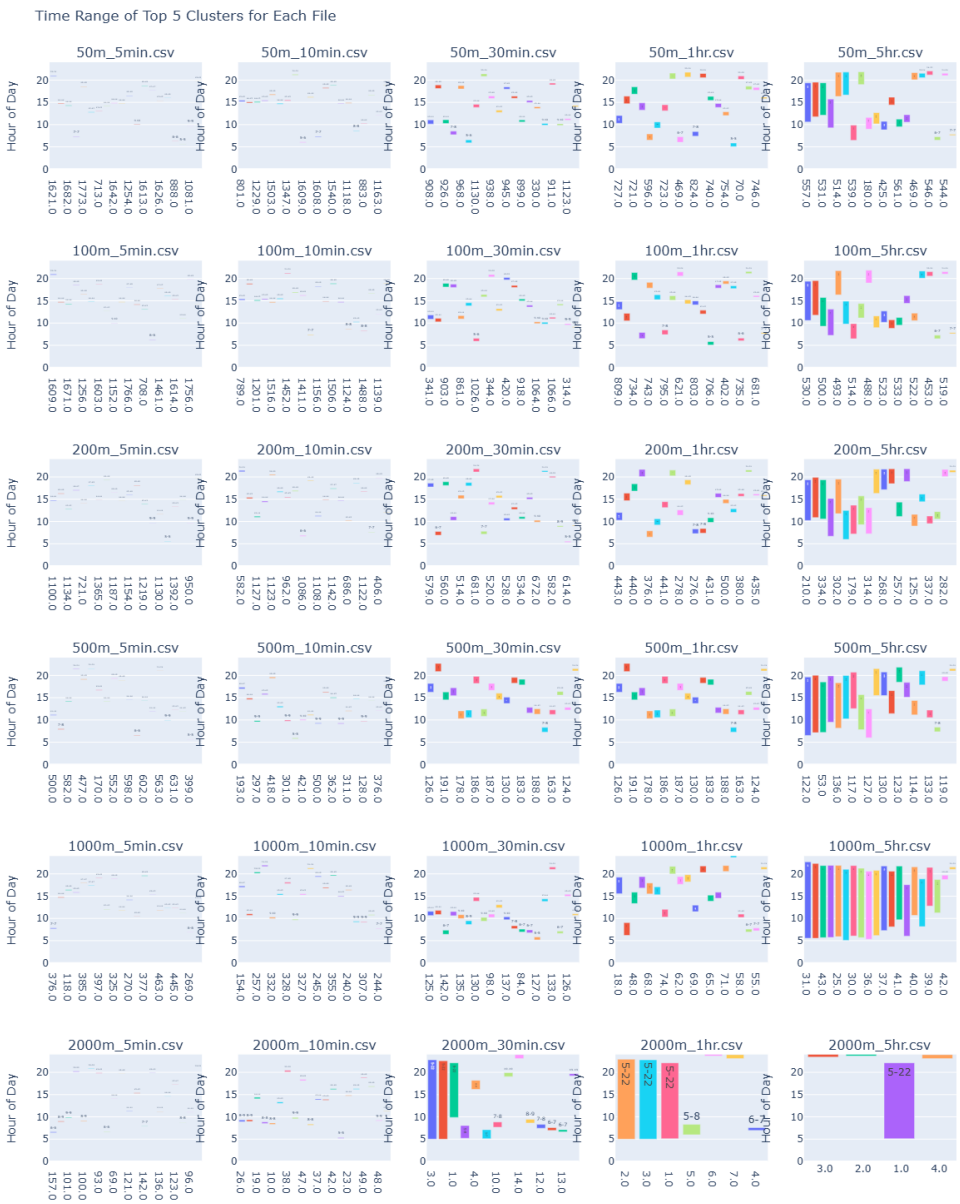


Fig. 29. Box Plots: Optimodal End Points Across Time

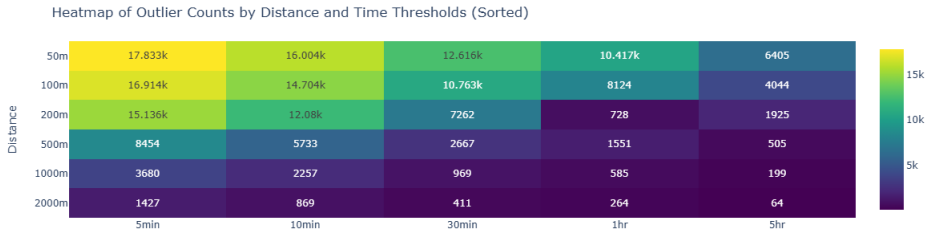


Fig. 30. Heatmap: Outliers for Uber Ride Start Points

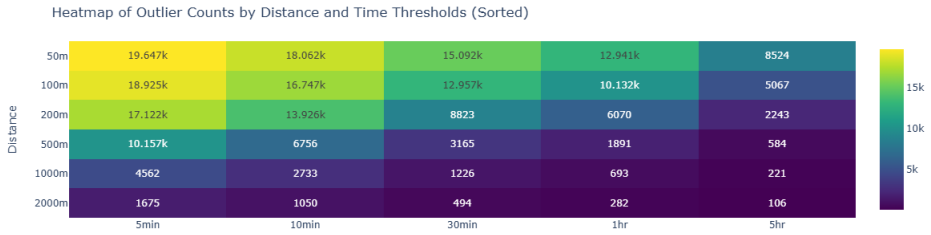


Fig. 31. Heatmap: Outliers for Uber Ride End Points

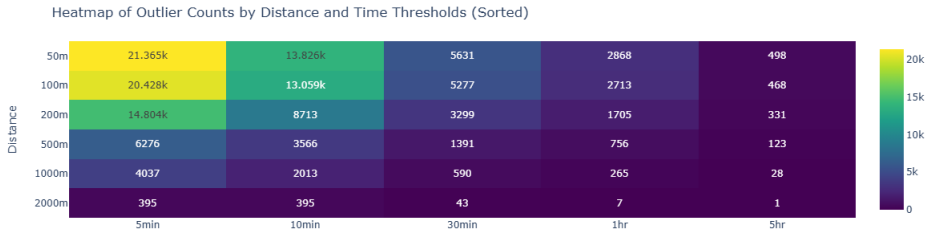


Fig. 32. Heatmap: Outliers for Optimodal Start Points

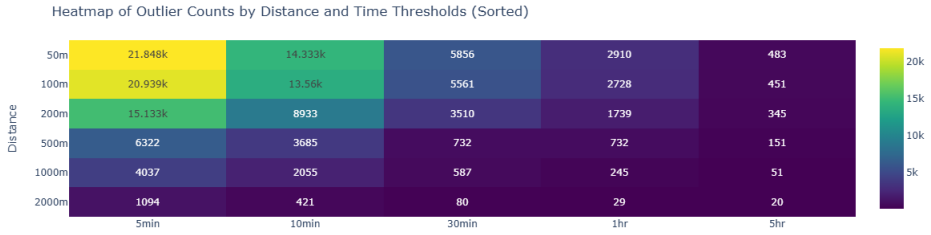


Fig. 33. Heatmap: Outliers for Optimodal End Points