DUGET: Leveraging Machine Learning for Dynamic User Grouping and Evolution Tracking in Public Transit Systems

Tobias Johannesson*, Isak Rubensson†, Sina Sheikholeslami*, Ahmad Al-Shishtawy*‡, Vladimir Vlassov*
*Department of Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden
Email: tjohann@kth.se, sinash@kth.se, ahmadas@kth.se, vladv@kth.se
†Transport Administration Region Stockholm, Sweden
Email: isak.jarlebring-rubensson@regionstockholm.se
‡RISE Research Institutes of Sweden
Email: ahmad.al-shishtawy@ri.se

***Abstract*—This work aims to explore the use of machine learning techniques, particularly clustering and cluster evolution tracking, to analyze travel patterns in public transportation in a city and provide valuable insights for urban transit planning and optimization. Clustering involves identifying and grouping similar objects, such as passengers with different ticket types, and distinguishing them from dissimilar objects in other groups. Over time, groups can change, so tracking this change can provide more detailed and valuable insights than analyzing data in aggregates. Clustering and cluster evolution tracking can reveal groups of passengers that are more or less affected by changes such as seasonality or fare increases. We propose a framework called DUGET (Dynamic User Grouping and Evolution Tracking), which clusters anonymized users based on their ticket choices and temporal travel patterns using a multi-step approach. The clusters are then tracked over time using Jaccard similarity based on memberships, allowing for the analysis and visualization of changes. Our experiments using a real-world public transportation dataset collected in Stockholm, Sweden, show the feasibility of tracking change over time in public transportation by examining passenger behavior as a temporal aggregate. The framework we propose is generalizable and can be used for future projects to understand trends in groups of objects.**

***Index Terms*—Smart card data, Temporal patterns, Clustering, Customer Segmentation, Transportation, Jaccard similarity**

I. Introduction

The rapid advancement of data collection and storage technologies has unlocked new possibilities in analyzing public transportation systems [1] [2]. With automated systems capturing millions of trips daily, decision-makers can now refine services by leveraging detailed insights into user behavior [3]. These insights are crucial for optimizing transit services by improving routes, reducing wait times, and balancing accuracy with the increased computational costs of handling larger, more complex data [4]. While the availability of such vast datasets presents significant opportunities, it also brings challenges, particularly in translating this wealth of data into actionable knowledge. One such trend is the emergence of new approaches to clustering users based on varying definitions of behavior [3]. Understanding the dynamics of user groups over time is particularly challenging, as static analyses often fail to capture the temporal variations critical for effective transit planning [5] [6].

In the domain of public transportation, clustering techniques have long been employed to uncover patterns in travel behavior [7]. Static clustering methods, like k-means, are easy to implement and ideal for initial segmentation, providing a snapshot of how different user groups utilize transit services [7] [8]. However, a significant limitation of these approaches is their static nature—they typically analyze data from a single point in time and do not account for how segments evolve, especially in response to external factors such as fare changes, seasonal variations, or policy shifts [7] [9]. This limitation can lead to a superficial understanding of user behavior, missing out on critical insights into the temporal dynamics that are essential for making informed decisions.

Recent advancements have introduced dynamic clustering methods that incorporate time as a dimension in the analysis process [10]. These methods aim to capture the evolution of user groups, offering a more accurate representation of the changing landscape of public transportation usage. Clustering using more advanced data may provide deeper insights, but it often comes with increased computational complexity and resource demands [8]. The need for sophisticated algorithms and significant computational power can be a barrier to the widespread adoption of these methods, and sampling is utilized to help mitigate this [10].

This paper addresses these challenges by introducing the Dynamic User Grouping and Evolution Tracking (DUGET) framework, a novel data mining approach designed to analyze temporal travel patterns and monitor the evolution of user groups in public transportation systems. DUGET integrates traditional and advanced clustering techniques, including k-means++ clustering, Hierarchical Agglomerative Clustering (HAC), and cluster tracking, to provide a comprehensive view of user behavior over time. The evolution of clusters is tracked using Jaccard similarity based on memberships, enabling detailed analysis and visualization of changes.

While DUGET offers significant advantages in capturing temporal dynamics and providing detailed user behavior insights, its computational complexity is an important consideration. The k-means++ algorithm, with its linear scalability relative to the number of users and clusters, provides an efficient foundation for large-scale data. However, the subsequent use of Hierarchical Agglomerative Clustering (HAC), with its cubic time complexity, presents higher computational demands when scaling to datasets with millions of users. To address this, DUGET has been designed to balance accuracy and computational feasibility. The initial clustering phase effectively reduces the dataset's dimensionality, making the HAC step more manageable and allowing DUGET to provide richer, dynamic insights while remaining scalable for large urban transportation systems.

The practical utility of DUGET is demonstrated through its application to real-world data from Stockholm's public transportation system. The dataset, provided by the Transport Administration Region Stockholm, the public authority responsible for coordinating and managing public transportation services in the Stockholm region, encompasses millions of trips recorded over several years, providing a robust test case for the framework. By identifying stable user groups and tracking dynamic shifts, DUGET provides actionable insights that can be applied to optimize transit services. For example, the framework revealed significant changes in user behavior during the summer months, highlighting the sensitivity of certain segments to temporal factors. Unlike static methods, DUGET offers a more responsive understanding of user behavior, making it a valuable tool for urban transit planning.

While previous studies in public transportation analysis have explored various clustering techniques to segment user behavior, there remains a significant gap in methodologies that effectively integrate temporal dynamics into the analysis. For instance, Cats et al. utilized longitudinal smart card data to model user behavior, employing a two-step clustering approach that combined k-means with HAC [11]. Although their work successfully captured distinct user segments, it was constrained by the static nature of the clustering, which failed to account for how these segments evolved over time. This limitation restricts the ability to observe and respond to shifts in user behavior, particularly when external factors such as fare changes or service adjustments come into play. Similarly, Agard et al. introduced a novel distance metric for clustering temporal sequences of trips, highlighting the importance of considering temporal patterns. However, their approach also revealed challenges in scalability, particularly when applied to large-scale datasets typical of urban transit systems [12].

Building on previous clustering approaches, Truong et al. [13] investigated passenger flow patterns using hierarchical clustering combined with principal component analysis (PCA) to uncover latent trends across stations based on timestamps. This method allowed them to observe recurring daily patterns, such as morning and afternoon peaks and highlighted the functional distinctions between different station types [13]. However, their analysis was constrained by the aggregated and fully anonymized nature of the data, which limited the ability to dynamically track individual commuter patterns over time. In contrast, DUGET leverages pseudo-anonymized data, enabling a more granular and continuous tracking of commuter behaviors. This distinction allows DUGET not only to identify clusters but also to monitor how these user groups evolve in response to temporal and external factors, such as seasonal changes or fare adjustments, offering a more robust and responsive approach to understanding public transit usage dynamics.

In contrast, DUGET offers a more sophisticated and dynamic approach by integrating manual segmentation with dynamic clustering and cluster tracking, specifically designed for the complexities of large-scale public transportation data. Tracking of evolving clusters can showcase segments appearing, disappearing, and how clusters change with time [14]. What sets DUGET apart is its innovative use of Jaccard similarity to track clusters over time, providing a robust and quantitative measure of how user groups evolve in response to various factors. Unlike the static methods employed by Cats et al. [11], DUGET's ability to monitor temporal changes allows for a more responsive and adaptive understanding of user behavior. For example, our application of DUGET to Stockholm's public transportation data revealed significant shifts in user groups during fare changes and seasonal transitions, which static clustering methods would have missed.

In summary, this paper presents a robust and scalable framework for analyzing the temporal dynamics of user behavior in public transportation systems. By leveraging both traditional and advanced clustering techniques, DUGET provides deep insights into how user groups change over time, offering valuable guidance for improving transit services and addressing the needs of diverse user groups. The application of this framework to real data from Stockholm's public transportation system underscores its practical relevance and potential for broader applicability in urban transit planning.

II. Method

The methodology aimed to investigate changes in different user groups by leveraging machine learning, in particular clustering techniques, to categorize users based on their temporal travel patterns. The analysis focused on size variations, cluster membership shifts, and average behavioral changes in each segment, influenced by factors such as seasonal variation and the fare increase in January 2024.

The proposed framework encompassed data collection and preprocessing, data analysis, clustering, cluster tracking, and change quantification. Figure 1 outlines the key steps, which include collecting data from specified time periods/bins, transforming it into object representations, normalizing the data, and sampling users, primarily from those shared across periods, for Jaccard similarity to track clusters over time.

In short, objects were grouped and clustered in a multi-step process using k-means and hierarchical agglomerative clustering. These final clusters were matched and tracked over time using Jaccard similarity based on the unique identifiers

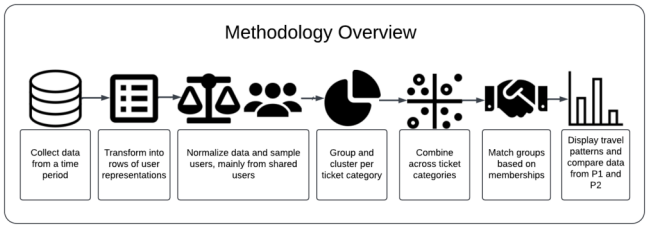

Fig. 1: Overview of the Methodological Framework

connected to cards used by travelers in Stockholm. The final step involved quantifying changes, allowing for visualization and comparison of data from different periods for the final clusters. The following sections will delve into each of these steps in greater detail, providing a comprehensive understanding of the methods used. This framework we dubbed "Dynamic User Grouping and Evolution Tracking" (DUGET). For more details on the DUGET framework refer to [15].

A. Data Description

The dataset used covers real trips made by real users in the region of Stockholm, Sweden during the period of 2022 up until (but not including) February 2024. Region Stockholm's relational database is automatically filled through their automated system which collects data in real time where each new trip initiated creates new records. When a card/ticket is used (also known as a tap-in), data such as the location, time and date, mode of transportation, and type of ticket are all recorded and linked to the CardKey, which is a unique identifier connecting trips to one entity.

The dataset is built up of three main tables along with several dimension tables. The primary tables relate to cards/tickets/users, trips, and journeys (one or multiple linked trips). These tables tracked information, such as first and last usage, and detailed all trips/journeys linked to a single CardKey, allowing for easy calculation of the number of trips/journeys a user has made in any given period. Data collected included the start and end stations, as well as the time and distance of each journey, along with various other dimensions. Example columns and their data, showing coordinates, timestamps, and locations, are shown in Figure 2.

InTime	InPointName	inCentroidEastingCoordinate	inCentroidNorthingCoordinate	OutTime	OutPointName
14:32:22	Solna	18.0095717662221	59.3665049546069	14:44:59	Stockholm City
20:14:25	Stockholm City	18.0594473186000	59.3311395346819	None	None
20:42:23	Liljeholmen	18.0230444293938	59.3106881010578	20:50:39	Slussen
23:51:12	Gullmarsplan	18.0809960892966	59.2983390154208	00:03:03	Sköndals centrum
17:12:43	Norra Sköndal	18.1169262959958	59.2610979190273	None	None

Fig. 2: Data columns showing start time, start point, start coordinates, end time, and end point of a set of trips

Trips can be linked together into journeys. A journey is simply one or multiple trips taking place close together in terms of time, where it is assumed that they are part of the same activity, such as a transfer, if the timing is right. These trips can often have the out station inferred by looking at the next trip made, but this destination is not inherently tracked within this automated system where tickets are only scanned upon entry.

The entire dataset tracks trips and journeys made between the end of 2022 and January 2024. On average, roughly 2,000,000 distinct users travel during a week, with roughly 700,000 distinct users during a single weekday. Around two million trips were made during an average day in November. Strong patterns in temporal travel can be observed, most of which are intuitively understood. For instance, one such pattern during weekdays includes two peaks corresponding to morning and afternoon rush hours, illustrating common commuter behavior as depicted in Figure 3.

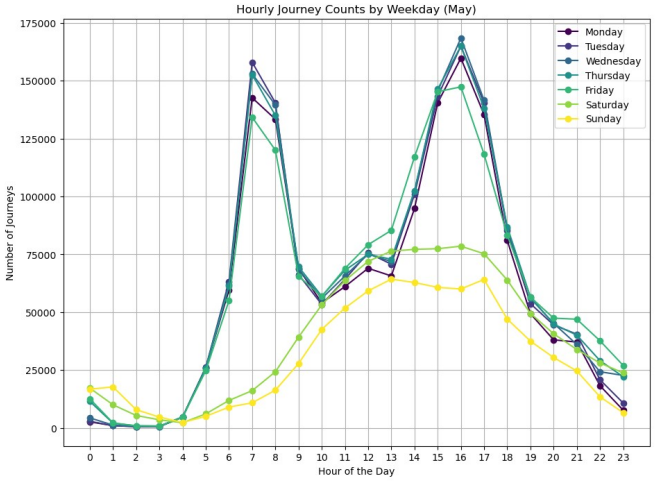

Fig. 3: Hourly patterns for all Weekdays for Journey made during the month of May 2023

B. Data Preprocessing

The smart card data from Stockholm's public transportation network, covering November 2022 to January 2024, contains timestamped records of user trips, including trip start time, mode of transportation, and station coordinates. To ensure privacy, the data is anonymized, with each user represented by a unique identifier linked solely to their travel behavior.

In the preprocessing stage, data was retrieved from an Azure data lake using PySpark and SQL queries. The dataset included four time periods (each 28 consecutive days), starting in November 2022, January 2023, November 2023, and January 2024. Each period had an equal number of weekdays to simplify comparative analysis.

To balance dataset size with analysis efficiency, temporal aggregation was applied, grouping journeys into hourly intervals. This reduced computational load while preserving key travel patterns. Hourly intervals were defined as Night (0-4), Morning (5-9), Day (10-14), Afternoon (15-18), and Evening (19-23).

Users were categorized into ticket types—Period, Tourist, School, or Single—based on the first ticket type used during

the period. This allowed for a simplified and efficient analysis of travel behavior across both time and user groups.

C. User Representation

With data from the selected periods preprocessed, journeys were aggregated so each row contained all data related to a single CardKey, assumed to represent one user. This aggregation encompassed all journeys made by a user during the period, capturing the temporal aspects of these journeys. The profiles created encoded the day of the week and time of day into features such as Monday_Morning and Tuesday_Evening, summing all journeys made by that CardKey for each time slice and day of the week as visualized in Figure 4

	CardKey	Monday_Morning	Friday_Morning	Friday_Day	Sunday_Evening
0	3732	7	0	1	4
1	4264	8	2	0	5
2	5859	1	3	4	5
3	8891	5	8	7	6

Fig. 4: Initial Columns of User Representation Data

Each user profile was assigned a single ticket category based on the first ticket type used during the observed time period. This approach simplified the analysis and mitigated potential complexities arising from users holding multiple ticket types. Attempts to determine the most frequently used ticket category were found to be computationally inefficient, increasing the processing time by over 160%, from 3 minutes to more than 8 minutes. The number of users in each of the four categories is visualized in Table I and Table II.

TABLE I: Number of Users in Each Ticket Category Across Multiple Months

Category	November 2022	January 2023	April 2023	November 2023	January 2024
Period	553,000	529,000	538,000	551,000	389,000
Tourist	55,000	42,000	61,000	55,000	22,000
School	165,000	162,000	174,000	183,000	176,000
Single	1,528,000	1,405,000	1,438,000	1,477,000	612,000

Users with fewer than four journeys during the specified time period were removed, under the assumption that representing true patterns for these users would be difficult. To accommodate the diverse requirements of clustering algorithms, particularly those sensitive to scale and distribution, normalization was applied so that each row was summed to one. This was achieved by dividing each column by the sum of all columns.

TABLE II: Number of Users in Each Ticket Category After Filtering Users with Fewer Than Four Journeys

Category	November 2022	January 2023	November 2023	January 2024
Period	520,000	513,000	534,000	235,000
Tourist	38,000	29,000	41,000	13,000
School	156,000	150,000	169,000	161,000
Single	571,000	522,000	556,000	187,000

D. Clustering Pipeline

Using temporal profiles to represent user behavior, a multi-step clustering approach was applied, integrating both domain knowledge and statistical techniques. Initially, users were divided into four manual groups based on ticket categories: Period, Tourist, School, and Single, with each user assigned to one category per period. These manual groups were then independently clustered using k-means++. The resulting clusters were consolidated into a single dataset, and further reduced in number through additional clustering using hierarchical agglomerative clustering (HAC), resulting in a final set of temporal profiles that captured behavior across all ticket categories.

The Sankey plot in Figure 5 shows the full clustering process with the initial manual grouping of users, followed by clustering within each category using the k-means algorithm. The final stage depicts how HAC combines these clusters across all ticket categories to form the final set of clusters that are randomly named.

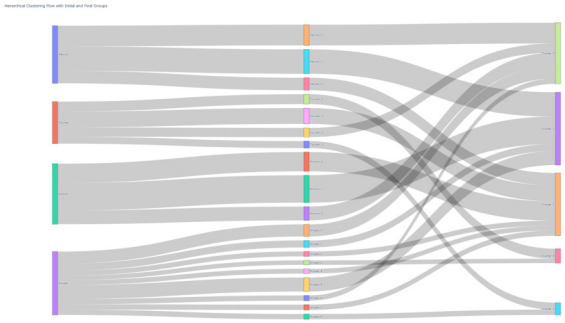

Fig. 5: Flow of User Grouping and Clustering Process

1) Manual Segmentation: The temporal profiles representing user behavior were divided into groups based on the defined ticket categories: Period, Tourist, School, and Single. These categories simplify multiple ticket types into broader groups. 'Period tickets' refer to those valid for 30 days or longer. 'School tickets' are used by elementary school students. The 'single tickets' group covers tickets intended for one-time use; multiple single tickets are still associated with the same CardKey, as is the case when switching to another ticket type. 'Tourist tickets' are primarily intended for tourists and are valid for 24 hours to seven days. These groups could be further reduced into full price and reduced tickets to separate university students and the elderly from adults.

To balance the dataset, users were sampled based on their presence in both examined time periods (November and the following January). This method reduced computational resource demands and ensured a balanced representation of different ticket groups, preventing any single group from dominating the sample. In cases where the number of users traveling in both periods was insufficient, additional users were randomly sampled from the remaining population to achieve a predetermined percentage of total users within each

ticket category. While this random sampling may reduce the Jaccard similarity between periods, it allows for the inclusion of more data, such as tourist users who may not travel consistently across both periods. This approach ensures a more comprehensive analysis while maintaining the ability to track consistent users. To preserve the integrity of the user matching process, the sampled rows were sorted by CardKeys to align with existing data."

2) Clustering Technique: The cluster analysis in this study aimed to dissect the multifaceted nature of public transportation usage by examining user data. After users were sampled from each ticket category and the manual groups were created, k-means++ clustering was applied. K-means excels at partitioning data into distinct clusters and scales efficiently compared to algorithms such as DBSCAN or hierarchical techniques. The initial seeding technique of k-means++, where initial centers are placed further apart with a probability proportional to the distance to the nearest point, was employed to reduce the risk of finding only local optima.

The number of clusters must be specified beforehand. To determine this, the within-cluster sum-of-squares (WCSS) and the elbow method were utilized. WCSS measures the similarity of objects within a cluster, offering insights into cluster cohesion. The elbow method helps identify the optimal number of clusters by indicating the point where adding more clusters does not significantly improve overall tightness. This technique balances the mathematical goodness of the number of clusters with having a manageable set of distinct and robust groups to analyze.

WCSS was applied to each manual group containing only temporal profiles for each of the ticket categories, producing elbow plots for the range of 1 to 20 clusters. From this, k-means was applied to each of the groups using a k around the elbow of these plots, with different k values for each category. These clusters were then renamed based on the ticket category, as the labels are randomly assigned by the k-means algorithm. The final groups after k-means were labeled, for example, Period_1, Tourist_1, Tourist_2, etc.

This clustering approach aimed to uncover underlying patterns and segments within the data, facilitating a deeper understanding of user behavior. By grouping different ticket types separately, it examined the variance within these groups without being influenced by other ticket categories.

3) Regrouping and Refinement: After identifying subgroups using k-means++ for all four ticket categories, the next step was to refine these clusters to capture more complex patterns and produce a final set of meaningful clusters representing user behavior. All clusters were combined into a single dataframe, relabeled from all four ticket categories.

Initially, the approach successfully captured variance and distinct temporal patterns within each ticket category. To further enhance the analysis, hierarchical agglomerative clustering (HAC) was applied. This method allowed for the exploration of similarities among different ticket types and the capture of intricate temporal profiles, using Ward linkage to ensure the formation of compact and coherent clusters. HAC was applied by calculating the centroids, pairwise distances between user profiles and current centroids, and iteratively merging clusters until a given threshold was reached.

First, centroids representing the mean of all current clusters were computed. Pairwise distances between user profiles and these centroids were calculated using Euclidean distance, the required distance metric for Ward linkage. The clustering process proceeded iteratively with Ward linkage, where at each step, the most similar clusters were merged, and distances were recalculated based on the new clusters formed. This iterative process continued until the desired number of clusters was obtained. Ward linkage minimizes the increase in the sum of squares within clusters, ensuring compactness.

To determine the optimal number of clusters, a dendrogram was constructed from an initial clustering run. This dendrogram as seen in 6 and 7 visually illustrated how clusters merged from the initial set down to a single cluster, guiding the selection of the optimal number of clusters. The goal was to maintain distinct patterns without consolidating all users into a single large cluster. Emphasis was placed on combining groups with similar temporal profiles to create a final set of clusters that effectively captured the diversity of user behaviors without losing granularity.

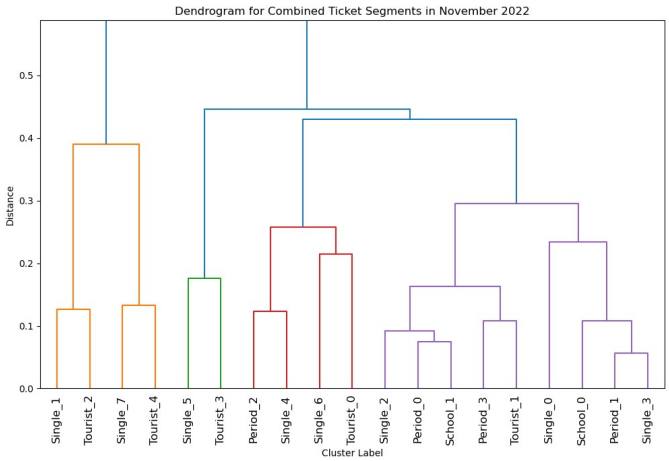

Fig. 6: Combined Dendrograms for All Ticket Categories in November 2022

The hierarchical clustering process resulted in a flat set of clusters using Ward linkage, ensuring that the identified clusters were both cohesive and representative of distinct patterns in temporal travel preferences among users. These clusters are randomly labeled between 1 and n by HAC, where n is the number of final clusters.

This approach not only refined the initial cluster assignments from k-means++ but also enhanced the interpretability and utility of the clustering results in analyzing and understanding user behaviors in public transportation.

E. Clustering Tracking

To understand how users change over time, it was important to correctly track groups of users. Since groups were labeled at

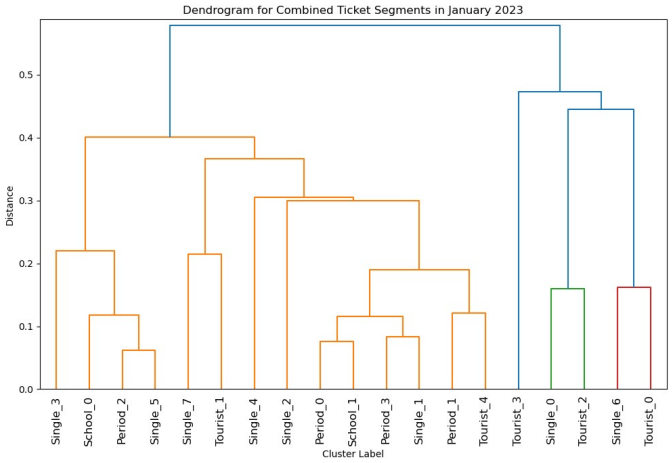

Fig. 7: Combined Dendrograms for All Ticket Categories in January 2023

random using Hierarchical Agglomerative Clustering (HAC), the final groups produced for period one and period two had varying names. All clusters were matched using Jaccard similarity on the CardKey, and clusters in period two were relabeled to match the best corresponding clusters from period one. This approach allowed for evaluating size changes, examining how CardKeys migrated between groups, and describing changes in temporal patterns between clusters. The following pipeline was proposed to understand how specific groups changed their behavior over time.

1) Label Matching: To track changes in clusters across different periods, we employed a method of relabeling based on the Jaccard similarity of CardKeys between clusters from period one (P1) and period two (P2). Temporal profiles were created and clustered using k-means and HAC, with labels randomly assigned to all clusters for November and January of the following year. These randomly assigned labels made it challenging to establish direct relationships between groups across periods. Relying solely on centroids could lead to incorrect labeling if clusters underwent significant changes between periods.

Assuming that user behaviors remained relatively stable over time and that clusters reflected meaningful distinctions among users, labels for P2 were assigned based on group membership linked to P1. Previous studies on similar temporal travel pattern data indicated that clusters typically retained consistent CardKeys over time, with most movements aligning with clusters exhibiting similar temporal profiles [16].

Each cluster was represented by the set of users that traveled in any cluster during both time periods. Utilizing Jaccard similarity and focusing on users present in both periods allowed us to establish connections between clusters based on their membership composition. This enabled analysis of transitions between groups and changes in ticket category distributions within clusters.

For each cluster, a set of CardKeys representing its membership was retained only for those present in both periods. This increased the ratios for all groups, particularly for groups comprising more transient users, where the number of CardKeys traveling in both periods was lower compared to more stable ticket categories such as period tickets. A matrix was constructed to compare the Jaccard similarity of CardKeys between clusters from different periods, resulting in an $n \times m$ matrix where the ratio of similarity ranged from zero to one. A score of zero indicated that the two clusters had no users in common, while a score of one indicated that the two clusters had identical users (aside from those traveling only in P1 or only in P2).

This matrix was used to relabel clusters in P2 to the name of the cluster in P1 with the highest Jaccard score, starting with the highest scores. To avoid dominance by a single large group and to accommodate varying cluster counts between periods, matched clusters were sequentially removed from further matching, preventing the capture of merges and splits as a trade-off.

To summarize, all clusters were represented as sets of CardKeys, with those not traveling in both P1 and P2 dropped. All clusters from P1 were compared to all clusters in P2 based on the intersection over the union of membership. Clusters in P2 were relabeled to the name of their best match in P1, starting with the highest ratio of shared CardKeys. Each cluster could only be matched once, and was then removed from potential labels.

2) Change Over Time: With users clustered and groups matched across different time periods, the focus shifts to analyzing and quantifying underlying changes. User behavior was characterized by a set of temporal profiles that consider the time of day and day of the week for their journeys. Changes within clusters are defined by alterations in these temporal profiles, such as shifts in peak times, alterations in average travel times, variations in the size of user groups, and fluctuations in the predominant types of tickets used.

Each cluster was described using parameters like average number of journeys conducted per weekday, average number of journeys per hour of the day, radius, and membership to outline basic trends in average behavior and highlight preferences within each group. These metrics provide insights into how clusters evolve over time and how certain groups may become more prominent or undergo shifts in characteristics.

To accurately assess changes, comparisons are made with previous periods to distinguish actual shifts in user behavior from seasonal variations or responses to specific events, such as changes in ticket prices. This comparative analysis helps in understanding whether observed changes are persistent trends or temporary fluctuations influenced by external factors.

F. Costs and Scalability

The proposed DUGET framework is tailored to handle the complexities of large-scale public transportation data. The initial phase employs the k-means++ algorithm, selected for its ability to efficiently handle large datasets. With a time complexity of $\mathcal{O}(n \cdot k \cdot d)$, where n is the number of data points, k is the number of clusters, and d is the dimensionality

of the data, k-means++ ensures that the initial clustering phase is both computationally feasible and scalable. This step significantly reduces the dimensionality of the dataset, providing a streamlined input for the subsequent clustering process.

Following this, Hierarchical Agglomerative Clustering (HAC) is applied to refine the clusters identified by k-means++. While HAC is known for its higher computational demands—characterized by a time complexity of $\mathcal{O}(N^3)$ and a space complexity of $\mathcal{O}(N^2)$ —its application at this stage is crucial for capturing the nuanced, nested structures within the data that are often missed by simpler methods. By operating on a reduced dataset, the resource demands of HAC are mitigated, allowing DUGET to maintain a balance between accuracy and efficiency.

The computational cost of these processes is carefully managed through implementation choices in limiting the dimensions used to represent users, and to through the domain being built on more limited groups, ensuring that the framework can scale with increasing data sizes. This balance is particularly evident in our empirical analysis, where DUGET was applied to a dataset comprising roughly 3 million users. The results demonstrated that while the HAC step required significant computational resources, these demands were within the capacity of standard computational infrastructures when processing datasets of this size. Further scalability analyses project that even with datasets ten or one hundred times larger, DUGET remains viable, provided that sufficient computational resources are available.

III. Results and Analysis

The evaluation of the DUGET framework underscores its ability to identify and track the evolution of user groups within Stockholm's public transportation system, despite the challenges posed by overlapping clusters. While the silhouette scores were lower than expected due to these overlaps, the framework consistently identified similar groupings across multiple consecutive runs and during different seasons. This consistency, particularly in identifying stable user segments such as daily commuters, highlights the robustness of the clustering approach where Figure 8 and Figure 9 shows the similarity between November 2022 and January 2023.

The Jaccard similarity for the larger, consistent clusters was observed to be upwards of 0.69, reinforcing the reliability of the framework in tracking stable user groups over time. These stable clusters, despite the inherent complexities of the dataset, provide valuable, actionable insights for transit planners, ensuring that core services remain optimized for the majority of users. In contrast, the dynamic clusters exhibited significant behavioral shifts, particularly in response to external factors such as seasonality during the summer months and the fare increase implemented in January 2024. The analysis revealed a clear migration of users between clusters, with some users adapting to the fare changes by altering their travel times or reducing the frequency of their trips. This dynamic adaptability

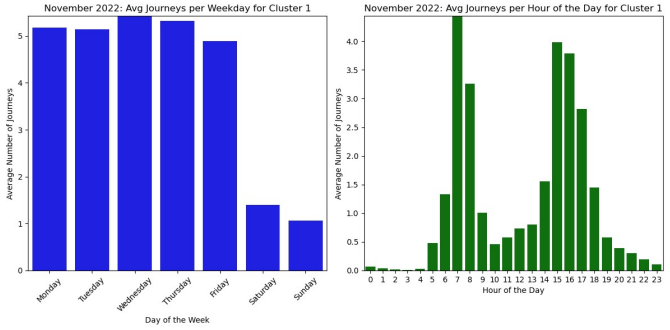

Fig. 8: Average Journeys per day of the week and per hour of the day for the commuter segment during November 2022

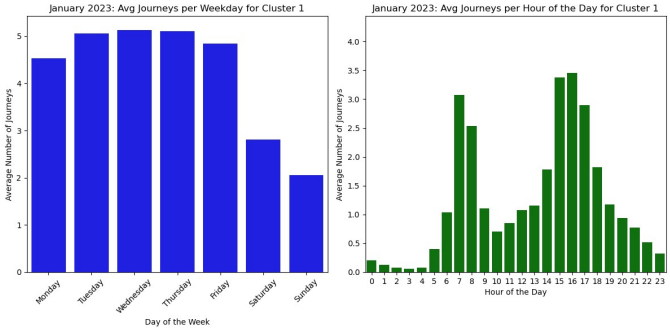

Fig. 9: Average Journeys per day of the week and per hour of the day for the commuter segment during January 2023

highlights the framework's ability to capture and reflect real-world behavioral changes, offering a comprehensive tool for understanding both stable and shifting user patterns in transit systems.

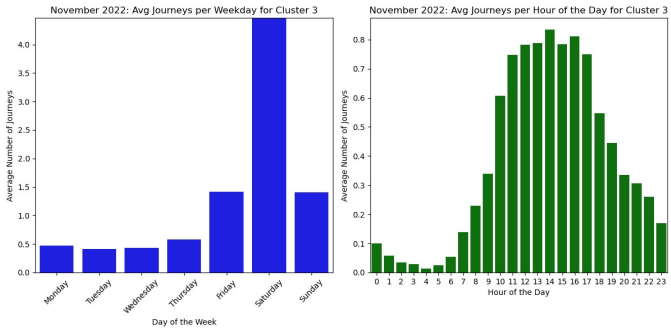

Fig. 10: Average Journeys per day of the week and per hour of the day for the Saturday segment during November 2022

The use of Jaccard similarity for cluster tracking proved instrumental in quantifying the degree of change within these user groups. For example, in the comparison of clusters between November 2022 and January 2023, and between November 2023 and January 2024, we observed a divergence in cluster composition, particularly in groups with high numbers of single ticket users. The ability to measure and visualize how clusters overlap or diverge over time provides a powerful

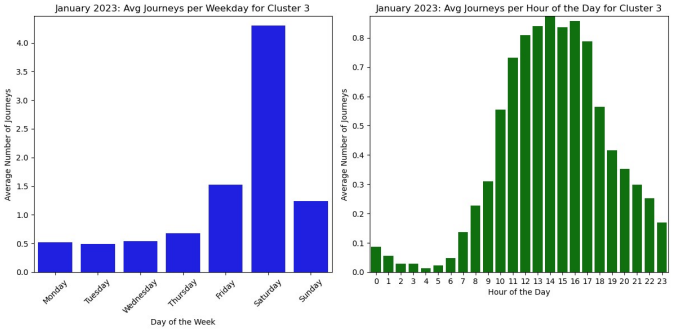

Fig. 11: Average Journeys per day of the week and per hour of the day for the Saturday segment during January 2023

tool for understanding the impact of external factors on public transportation usage—a capability that was previously not utilized at the Transport Administration Region Stockholm. This approach not only enhances our comprehension of user behavior but also offers a model that can be applied to other public transportation systems worldwide.

The seasonal analysis further enriched our understanding of user behavior. The data indicated a predictable yet significant shift in travel patterns during the holiday season, with certain user groups, such as non-daily commuters, exhibiting more sporadic travel behavior. These insights are crucial for transit authorities to anticipate demand fluctuations and adjust services accordingly. Specifically, the decrease in single ticket users within certain clusters during January 2024, as opposed to January 2023, suggests a possible reaction to the fare increase, which may highlight a sensitivity in these user segments to fare changes.

The ticket category count for each of the manual groups, as shown in Table III, illustrates the distinct clustering patterns identified using DUGET. The clusters between November 2022 and January 2023 were mapped so that cluster 1 had the most similarity in terms of membership with cluster 1, and the same mapping was applied between November 2023 and January 2024. Despite some users being dropped due to changes in group composition, the analysis showed that cluster 1 for 2022/2023 and cluster 4 for 2023/2024 consistently represented the commuter group, characterized by a high number of period, school, and single ticket users.

TABLE III: Comparison of Shared Users in Each Ticket Category Between November and January Across Years

Category	November 2022 and January 2023	November 2023 and January 2024
Period	51,000	24,000
Tourist	1,000	1,000
School	53,000	56,000
Single	52,000	19,000

The identification of these commuter groups provided valuable insights into how single tickets might be used similarly to period tickets, a behavior not captured by previously applied methods. Cluster 2 for 2022/2023 and cluster 6 for 2023/2024 represented day/activity-based travelers, showing a more even spread over the day compared to commuters. This group might include commuters with flexible schedules, as indicated by the significant portion of period and school users. The observed reduction in tourist and single ticket users within this group in January 2024, compared to January 2023, could be attributed to the overall lower number of these users during that period.

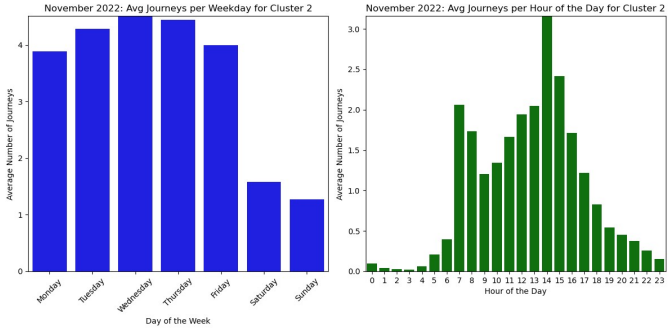

Fig. 12: Average Journeys per day of the week and per hour of the day for the activity-based segment during November 2022

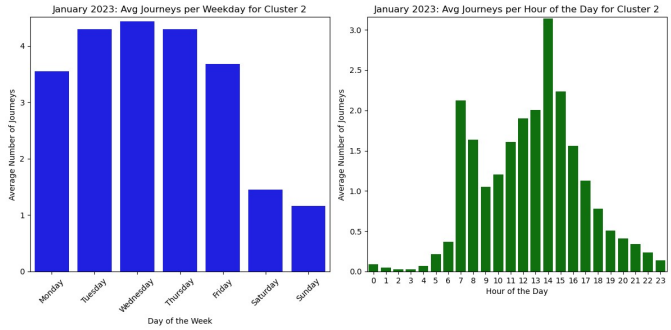

Fig. 13: Average Journeys per day of the week and per hour of the day for the activity-based segment during January 2023

Additionally, the groups with zero period and school ticket users further demonstrate how period tickets are predominantly used by individuals conducting a greater number of journeys. The grouping of users who prefer traveling on specific days of the week, such as Fridays or Saturdays, provides transit planners with insights that were previously masked by broader categorizations like weekday/weekend travel patterns.

The scalability of the DUGET framework was rigorously tested through both empirical analysis and theoretical projections. In our experiments, the initial k-means++ clustering phase demonstrated efficient performance, requiring approximately 1 minute to cluster 1 million users. The HAC step, while more computationally demanding, completed within 2 minutes using standard computational resources, underscoring the practicality of the approach for datasets of this size where the first step of clustering decreased the costs of using HAC.

The final set of groups identified and matched between November 2023 and January 2024 demonstrates the robustness of the DUGET framework, despite fluctuations in cluster

TABLE IV: Average Trips Conducted per Cluster in November 2023 and January 2024

Cluster	Average Trips (November 2023)	Average Trips (January 2024)
Cluster 1	10.07	10.43
Cluster 2	8.41	8.37
Cluster 3	9.92	8.35
Cluster 4	33.65	29.26
Cluster 5	11.66	9.21
Cluster 6	19.64	22.26

size over time. Clusters such as Cluster 1 and Cluster 2 exhibit minimal variation in behavior, with changes in average trips conducted per month remaining within 0.48% and 3.57%, respectively. This stability in behavior, despite external factors and shifts in group composition, underscores the framework's effectiveness in maintaining consistent user groups. Furthermore, the clear distinction between clusters, as evidenced by the significant differences in average trips per group—such as the contrast between Cluster 4 (29.26 trips) and Cluster 2 (8.37 trips) in January 2024—reinforces the framework's ability to accurately segment users based on meaningful behavioral patterns. Overall, these findings further validate DUGET's capability to dynamically track, adapt, and preserve the integrity of user behavior across multiple periods, providing transit planners with actionable insights for real-world applications.

IV. Discussion

This study demonstrates the effectiveness of dynamic clustering in analyzing user behavior within public transportation systems. The methodology successfully captured nuanced travel patterns across diverse user groups, offering insights into how these groups evolve in response to external factors such as fare changes. The combined use of k-means++ and Hierarchical Agglomerative Clustering (HAC) allowed for the identification of stable user segments, such as daily commuters, while also highlighting dynamic clusters sensitive to external changes.

A key strength of the DUGET framework lies in its ability to track user behavior over time using Jaccard similarity, which confirmed the stability of certain groups and captured significant shifts in others. This differentiation between stable and dynamic clusters provides a more granular understanding of user behavior, helping transit planners target interventions more effectively.

However, this study has several limitations. The reliance on temporal data, while crucial for understanding public transportation usage, does not fully capture the complexity of user behavior. Factors such as socioeconomic status, trip purpose, and multimodal travel preferences were not incorporated, potentially leading to oversimplified segmentation. Additionally, the assumption that each CardKey corresponds to a single user may introduce inaccuracies due to data anonymization.

Another key limitation is the computational complexity, particularly in the use of Hierarchical Agglomerative Clustering (HAC) and the querying and joining of large datasets. As data grows, these processes can become bottlenecks, affecting scalability and performance. While HAC effectively refines clusters, its high computational cost, combined with complex database operations, presents challenges for larger datasets.

To address these issues, sampling and dimensionality reduction are essential for improving scalability. Sampling enables processing of representative data subsets, reducing computational load without sacrificing accuracy, while dimensionality reduction retains the most relevant features. Integrating these techniques with more efficient algorithms or hybrid approaches would improve both scalability and the framework's robustness.

Despite these limitations, the DUGET framework's ability to capture dynamic user groups and temporal shifts offers significant value. Its capacity to detect granular behavioral changes that static methods miss demonstrates its utility in complex urban transportation systems. Future work could further enhance the framework by incorporating additional data sources, such as user feedback and socioeconomic indicators, and by exploring multimodal travel patterns for a more comprehensive understanding of user behavior.

V. Conclusion and Future Work

The DUGET framework effectively grouped users based on temporal patterns and tracked their evolution over time. By leveraging both k-means++ and HAC, the framework provides urban planners with detailed insights into user behavior, offering a nuanced understanding of how stable and dynamic user groups respond to external factors like fare changes and seasonality. This capability highlights the framework's practical utility for effectively capturing and analyzing evolving user behavior in public transportation systems.

One of the most notable findings is DUGET's ability to detect granular behavioral shifts, in contrast to existing methodologies that focus on aggregated metrics, such as the total number of tickets sold. This highlights the critical importance of incorporating temporal dynamics into public transportation analysis for a more nuanced understanding of user behavior. Additionally, the scalability of DUGET was validated through its ability to maintain robust performance and manageable computational requirements as the dataset size increased. The primary computational bottleneck, k-means++, scales linearly with the number of users, ensuring that clustering operations remain efficient even as data grows. This combination of scalability and detailed behavioral insights positions DUGET as a highly effective tool for modern public transportation planning.

The flexibility of DUGET lies in its ability to adjust the level of detail by varying the number of clusters and the granularity of time periods examined. This adaptability enables the tracking of seasonal changes on yearly, monthly, weekly, and daily levels. While the methodology was primarily tailored to public transportation, it is adaptable to other domains, with the critical factor being the domain knowledge required to represent the objects under examination accurately. The choice of clustering algorithms and the data size are the

primary constraints when utilizing this cluster tracking method efficiently.

While the DUGET framework has demonstrated effectiveness, several limitations may impact its broader implementation. The reliance on existing clustering algorithms, which are sensitive to parameters and data types, can lead to suboptimal results in certain contexts. The assumption of stable user group memberships may not hold for sporadic users, complicating segment relabeling. Additionally, the storage of user memberships over time can be challenging in memory-limited systems, suggesting the need for more efficient solutions. In environments with significant seasonal variations, longer initial periods may be required to establish reliable baselines, as time-binning could obscure short-term behavioral changes. While using shorter intervals, such as weekly data, could mitigate this, it would demand additional computational resources and fine-tuning.

The DUGET framework opens several exciting avenues for future research and development. Exploration of automated solutions, more advanced clustering techniques, and analysis of sporadic users and seasonal effects are a few examples. Incorporating temporal and spatial data could provide a more comprehensive framework, applicable to a wider range of data types. Further experimentation with user representation could lead to more distinct and explainable segments, improving qualitative mapping. Additionally, testing various segment mapping techniques over time, and automating the evaluation of cluster robustness using metrics like Jaccard similarity, will enhance the framework's generalizability, reliability, and scalability across diverse applications.

REFERENCES

[1] A. Oussous, F.-Z. Benjelloun, A. Ait Lahcen, and S. Belfkih, "Big data technologies: A survey," *Journal of King Saud University - Computer and Information Sciences*, vol. 30, no. 4, pp. 431–448, 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1319157817300034

[2] A. K. Sandhu, "Big data with cloud computing: Discussions and challenges," *Big Data Mining and Analytics*, vol. 5, no. 1, pp. 32–40, 2022.

[3] T. F. Welch and A. Widita, "Big data in public transportation: a review of sources and methods," *Transport Reviews*, vol. 39, no. 6, pp. 795–818, 2019. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/01441647.2019.1616849

[4] S. Kaffash, A. T. Nguyen, and J. Zhu, "Big data algorithms and applications in intelligent transportation system: A review and bibliometric analysis," *International Journal of Production Economics*, vol. 231, p. 107868, 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0925527320302279

[5] W.-L. Shang, J. Chen, H. Bi, Y. Sui, Y. Chen, and H. Yu, "Impacts of COVID-19 pandemic on user behaviors and environmental benefits of bike sharing: A big-data analysis," *Applied Energy*, vol. 285, p. 116429, 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306261920317906

[6] S. Deng, Q. Cai, Z. Zhang, and X. Wu, "User behavior analysis based on stacked autoencoder and clustering in complex power grid environment," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 12, pp. 25 521–25 535, 2022.

[7] K. Lu, J. Liu, X. Zhou, and B. Han, "A review of big data applications in urban transit systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 5, pp. 2535–2552, 2021.

[8] Z. Gan, M. Yang, T. Feng, and H. Timmermans, "Understanding urban mobility patterns from a spatiotemporal perspective: daily ridership profiles of metro stations," *Transportation*, vol. 47, no. 1, pp. 315–336, 2020. [Online]. Available: http://link.springer.com/10.1007/s11116-018-9885-4

[9] A. D. Marra, L. Sun, and F. Corman, "The impact of COVID-19 pandemic on public transport usage and route choice: Evidences from a long-term tracking study in urban area," *Transport Policy*, vol. 116, pp. 258–268, 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0967070X21003620

[10] L. He, M. Trépanier, and B. Agard, "Space–time classification of public transit smart card users' activity locations from smart card data," *Public Transport*, vol. 13, no. 3, pp. 579–595, 2021. [Online]. Available: https://link.springer.com/10.1007/s12469-021-00274-0

[11] O. Cats and F. Ferranti, "Unravelling individual mobility temporal patterns using longitudinal smart card data," *Research in Transportation Business & Management*, vol. 43, p. 100816, 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2210539522000372

[12] M. T. Bruno Agard and V. P. Nia, "Assessing public transport travel behaviour from smart card data with advanced data mining techniques," in *Proceedings of the World Conference on Transport Research (WCTR)*. WCTR Society, 2013. [Online]. Available: https://datawisdom.ca/paper/2013-Transport.pdf

[13] R. Truong, O. Gkountouna, D. Pfoser, and A. Züfle, "Towards a better understanding of public transportation traffic: A case study of the washington, DC metro," *MDPI*, vol. 2, no. 3, p. 65, 2018. [Online]. Available: https://www.mdpi.com/2413-8851/2/3/65

[14] N. Barbosa Roa, L. Travé-Massuyès, and V. H. Grisales-Palacio, ": Dynamic clustering for tracking evolving environments," *Pattern Recognition*, vol. 94, pp. 162–186, 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0031320319301992

[15] T. Johannesson, "Dynamic user grouping and evolution tracking (duget): Leveraging machine learning for public transit insights," Master's thesis, KTH Royal Institute of Technology, 2024, accessed: 2024-11-12, URN: urn:nbn:se:kth:diva-353078.

[16] A.-S. Briand, E. Côme, M. Trépanier, and L. Oukhellou, "Analyzing year-to-year changes in public transport passenger behaviour using smart card data," *Transportation Research Part C: Emerging Technologies*, vol. 79, pp. 274–289, 2017. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0968090X17301055