Clustering of Boarding Patterns in Public Transport

A Case Study of Public Transport, Antalya

Fatma Erkan
Computer Science Engineering
Akdeniz University
Antalya, Turkey
20150807029@ogr.akdeniz.edu.tr

***Abstract*— This assignment is about public transport data for Antalya city. It is a resource for studying the travel habits of the Antalya city inhabitants. This experiment contains an approach to the data mining the temporal behavior of the passengers in a public transportation to extract relevant and easily interpretable clusters. The study can help to improve the demand of customers and propose targeted services and tools accordingly. To get these results dataset analyzed, and clustering results contextualized.**

Keywords—clustering, public transport.

I. INTRODUCTION

Public transportation is travel that gives an opportunity to travel more people together along the designated routes. Sample types of public transportation include buses, trains, and trams. In large cities, public transportation has an important position. Passengers have smart cards that contain credit and the public transportation system based on these cards. Passengers use these cards when boarding buses or other transportation. The main purpose of the system is to manage the revenue and study the daily travel habits of passengers with the transport network. When a smart transportation card is used, it contains a detail information of passenger such as time to use of buses, boarding location and so on. These kinds of information give an opportunity to check the passengers' usage of transportation time and analyze travel behavior and regularity, or turnover rates.

Passengers' travel behaviors can be used in different applications. For example, this information can be used to improve the performance of the public transportation network, identify the demand with accuracy, evolve the service and many more. Data of the public transport can be used to identify whether some areas are underserved, and then public transport system can be improved with this way.

II. OBJECTIVE

In the assignment, dataset indicates that the behavior of passengers in daily public transport depends on different time periods. Contributions of the experiments are followings:

- The experiment contains the extraction of travel patterns from the dataset. To this effect, the temporal passenger count based on boarding time is constructed. Then some model-based clustering approaches applied to find out clusters of passengers whose behavior similar depend on their boarding times.
- The study includes that how passenger travel habits relate to the bus line. To do this, first step is to cluster of bus line data of the city. The socioeconomic characteristics are also can be related to the dataset by researching the route of the bus lines in Antalya. Then we can estimate the passengers travel behaviors depend on socioeconomic characteristics.
- Approaches are applied the dataset that covering the usage of public transport boarding times in day of 18th of December 2019 in Antalya city.
- Different clustering models are used to get specific clusters such as k-means clustering, mean shift clustering, hierarchical clustering.

The main aim is to define an approach to clustering dataset to study temporal behavior of passengers.

III. LITERATURE REVIEW

Anne-Sarah Briand et al. [7,8] were study about mixture model clustering approach for temporal passenger pattern characterization in public transport system. Researchers present an approach to passenger clustering based on travel hours. The main approach of the experiment considers a continuous representation of time. They use the Gaussian mixture model to estimate using the passenger information. The study contains a large experimental subject on the real dataset and visualize the approach to discover behaviors of passengers. Experiment contains the socioeconomic information with crossing the results with spatial information on the city and the users' information. Contiguous weekly activity profile was obtained for each group as a result. Authors uses these groups to differentiate passengers with regular travel hours and those with diffuse travel hours.

Mahnaz Moradi and Martin Trepanier [11] analyzed temporal clusters of public transit passengers using smart card data. The study includes smart public transportation card data behavioral pattern of users monthly. Main goal of the experiment is the discovery of the cluster membership's stability of passengers. It includes a method to identify the stability of temporal behaviors of passengers depend on smart card data. Results are separated in different parts. First part shows the weekdays grouping because working days contain greatest portion of unstable users, the second sections include the information about first day after long weekends analyzing every week provides best results. And the final section shows the adult passengers who are stable bus users. The result is found that more than nearly 70% of passengers are present on the public transportation more than 2 weeks.

Bruno Agard et al. [12] study about smart transport cards, onboard readers, and centralized information system. With data mining methods, authors research on the automatic extraction of passenger patterns on the public transportation system. The dataset was used in different ways in this project to evaluate actual usage of the public transportation system. It

makes it possible to identify passengers and measures their travel habits. It analyzes the different usage of the system daily, weekly, and monthly. With these results, researchers' purpose is to improve the public transport system by adjusting balance between passenger habits and the level of service provided on each route. Another aim is to enhance the public transport routes and performance. Their result also shows the difference between smart card data and usual travel data that smart card data provides more information about passengers.

Catherina Morench et al. [13] analyses the performance of the public transport system in Canada using smart card data. Main purpose of the study is to evaluate the usability of smart card data to estimate some of basic indicators. Then the methodology results some passenger movements and measures the performance. It estimates the indicators using dataset which is generated using smart card data. Result also demonstrate the ability of specific indicators depend on the critical issues.

Neal Lathia and Licia Capra research the habit of hidden passengers' information from smart card data and study of their response to travel encouragement. Firstly, authors collect data from online survey and smart card system data from public transport service from London's Oyster. These two different data compared. Comparison results shows the behaviors of passengers by different aspects. For instance, frequency of trips, regularity, modality of travel etc. Then they provide an approach that how data from smart card can improve the extent to which incentives such as prices by time.

All these studies working with smart card data of public transport system. They use some basic clustering approaches. Some of these researches contain machine learning methods to improve the development. These methods used to conduct multi-scale clustering. Some of them also consider the socioeconomic characterization cluster of passengers and information of passengers.

The difference between the approach of this study is the clustering algorithms. Most of the approaches in other studies clusters data using the similar clustering algorithms. These different approaches give different solutions to researchers. The approach of the study of public transport in Antalya is based on different clustering. Various clustering models used to identify differences and show results in more clear way. In the approach of this project, timespan is considered while developing.

IV. METHODOLOGY

Clustering public transport system is based on the temporal activities of passengers extraction from dataset. Thus, passenger groups can show the most frequent travel patterns in public transport system in Antalya city. It helps to improve transportation system. The approach in the assignment is to passenger count clustering depend on time. The generative model of this study integrates a continuous representation of time.

A. Dataset Analyse

Dataset of the study contains the collected passenger information depend on time intervals by bus line. It contains approximately 1900 different rows. The data shows that Antalya's public transport system's daily traffic. Different bus lines have different routes. For instance, the most occurrence bus line which is KL08 has o route from Güzeloba to Sarısu. The dataset also shows the daily information of passenger actions depend on time interval on date 18-12-2019.

B. Methodology

The main aim of this study is to analyze groups of passengers count depend on time intervals. So, research contain the information about passengers taking public transport system at same time intervals without route of the bus line. But in that case, for the more comprehensive research, the route of the lines of buses can be searched on the website named "Antalya Ulasim". Clustering of these passenger groups identify frequent patterns of usage the public transport system and it helps to improve the demand according to the results.

The first step of the study is to analyze the dataset according to timespan. Looking at the boarding time information it is seen that the dataset contains the hourly knowledge of the bus lines and there are different bus lines which have different routes in the city. It is also estimated that bus lines with maximum passengers for first 6-time units and bus lines with maximum occurrence in the total version of the dataset. The minimum passenger bus lines and minimum occurrence in the total dataset can be commented. In particular time units various bus lines have different number of passengers. From discovering a result that which bus line contains maximum number of passengers can improve the performance of the public transport system in Antalya city. Many more analyze results can help to raise the productivity. However, in some cases dataset can contain more information about bus lines, number of passengers, boarding times and some other features, when studying with this kind of data methodology is important to help public transport system.

Then, the study examines in detail correlation between bus lines and passenger data. From this view, all the bus lines have very similar characteristic which implies that for rasping the general pattern of the date, focusing on time is effective.

Next step is to cluster the data depend on time intervals. Estimation is done by using different clustering models to compare results to find best groups. From this perspective, clusters of the bus lines can be evaluated in different aspects. Various clustering models draws independent graphs as a result. In every model, each graph is first generated by selecting a cluster, then graphs are drawn from the conditional distribution relative to that specific cluster. Each cluster is thus described concisely as a distribution over the passenger count or time interval. In this context, clusters are description of when trips are more or less probable to be made.

In this assignment, generative models are:

- K-means clustering is one of the simplest models in data mining. Every cluster refers to a group of data points aggregated together because of certain similarities. It begins to work with defining a target number k that refers to the number of centroids for dataset. The centroid represents the center of that cluster. The k-means algorithm defines k number of centroids and it allocate each data point to the nearest group and it keeps the centroid small. Means keyword in the k-means model refers to the average of the data which finds the centroid. In the study the 6 cluster is more effective, and it shows the affect of time clearly.
- Partitional clustering method that is mean shift used in the study. Mean shift clustering aims to find the

high point of the distribution. It is a process to find the peek point of the clusters. The method is simple and operational. In the k-means there is a risk of dividing cluster to multiple clusters. However, mean shift clustering method the risk is does not exist. Result of the mean-shift method is show clusters that affect of time even more clear.

- Agglomerative clustering is a kind of hierarchical clustering model that groups data in clusters based on their similarities. The method starts clustering with treating each data as a single cluster. Then, it pairs groups until every group compound into bi group which contains other objects. So in this study the agglomerative clustering model is used to group.

When estimating the model to find the best clustering for the study, various kind of clustering used. In order to retrieve the model which best fits the dataset, clustering models evaluated one-by-one. To select an appropriate number of clusters, different cluster numbers examined and evaluated.

V. Results and Discussion

With using various clustering models, the study obtained different outputs that each of the clustering models describe temporal mobility pattern. To analyze the clustering results, each of the models are evaluated with its pros and cons.

The first step of the study is to analyze the dataset that helps to improve the performance and results when defining the clusters. The dataset contains exactly 1839 data inside of it and every column contains the information about bus line, time interval and the passenger count. It provides information about traffic of public transport system in Antalya in the date of 18-12-2019 and it shows the daily usage of various bus lines.

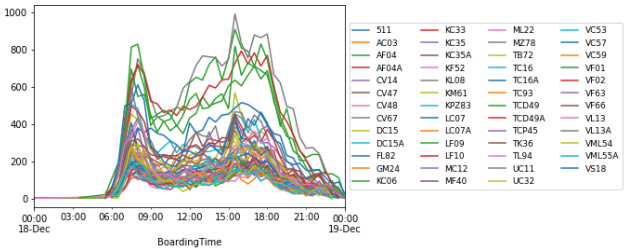

Figure 1. Boarding Time vs Passenger Count of Bus Lines

In the Figure 1, each line shows the different routes. It is seen that some lines are visibly have higher values than the other lines. There is also some noticeable trends, for instance, there are noticeable spikes between from 06:00 AM to 09:00 AM and from 03:00 PM to 06:00 PM for all lines that are rush hours.

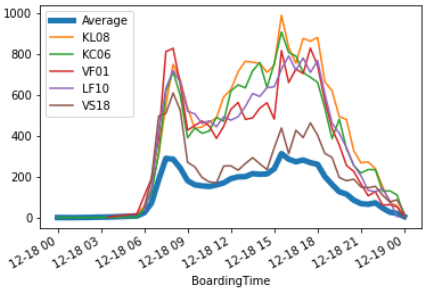

Figure 2. Maximum Occurrences of Bus Lines in Maximum Recorded Passengers

In the first 6 time points, which is from beginning hour to 06:00 AM, VF01 has the maximum number of passengers with the number 107 at the time 06:00 AM. However, in the total version of the dataset the most occurrence bus line is the KL08 with the number of occurrences 22. And the KC06 follows it with the number of occurrences 6.

In the first 6-time units, KC06, TL94 AND TCP45 has minimum recorded passengers. The Figure 3 shows the bus lines with maximum occurrences in the minimum recorded passengers in total.

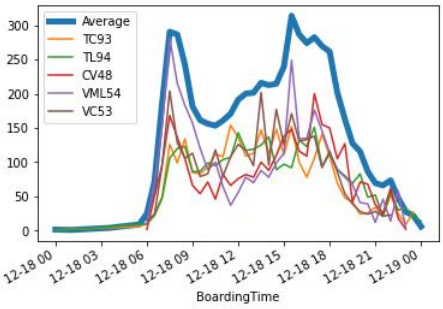

Figure 3. Maximum Occurrences of Bus Lines in Minimum Recorded Passengers

In the total dataset, most used bus lines are KL08 with 19400 passengers, KC06 with 17457 passengers and LF10, VF01, LF09 follows. Lines with least total passengers are KC35 with 3023 passengers, TC93 with 3028 passengers and CV48, TCD49 AND KC35A follows the list. It is seen that the busiest line is KL08. For the least busy, there are multiple lines with similarly low amount of passengers.

Correlation analysis is used to study about relationship between variables such as degree that the variables associated with each other. In this study the average of all the correlation coefficients between bus lines calculated nearly 0.8367 and the standard deviation of all correlation coefficients between bus lines is nearly 0.07748. From this analysis, all of the bus lines have very similar characteristic. This result is in line with previous results and implies that for rasping the general pattern of the date.

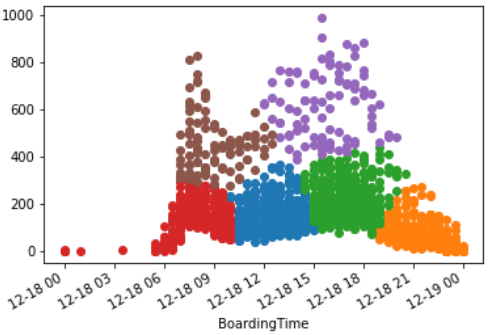

Figure 4. Clustering with k-means method

In k-means clustering which is one of the most popular clustering model in data mining, time is taken in units of minutes. This decision is important for clustering, as the unit of time affects the distance calculations. Figure 4 represents the different clusters in different colors. The number of clusters is chosen as 6 because it shows the affect of time clearly. Top two clusters are mainly differentiated from others based on their number of passengers. However, the rest and the top two between themselves are separated are mainly by time.

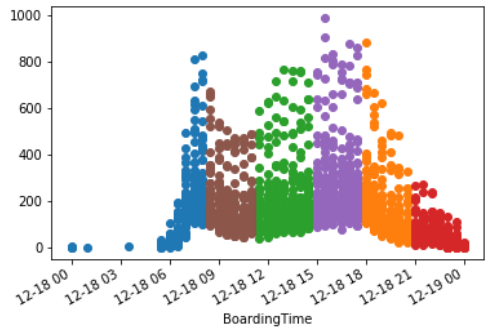

Figure 5. Clustering with k-means method with seconds

In the k-means method, other time units are investigated. In the Figure 5, the time unit is selected as seconds, which created clusters entirely on time with equal time difference for each cluster which is nearly 3 hours.

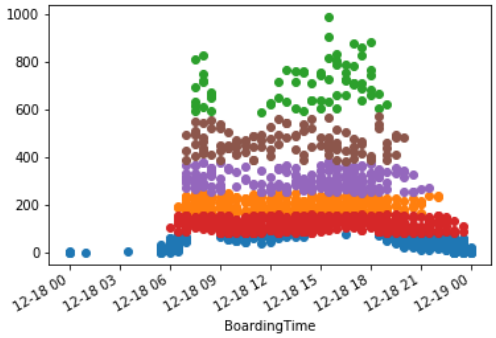

Figure 6. Clustering with k-means method with hours

In the Figure 6, the time unit is selected as hour. This time the clusters are observed entirely horizontally that is only based on the number of passengers data because of these two clusterings are simplifying the difference to only one feature, it can be argued that selecting minutes as the time unit is sensible. However, further investigation on time units that are multiples of 1 minute are not investigated as the obtained results seemed sufficient.

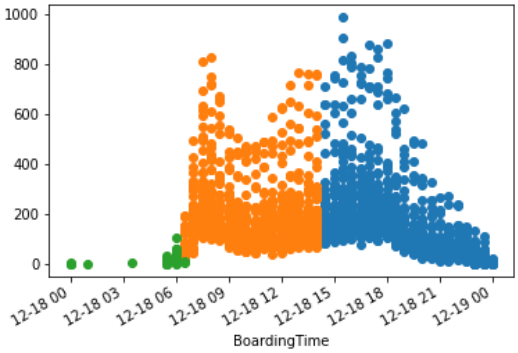

Figure 7. Clustering with mean shift method

Figure 7 represents the mean shift clustering method's result. The mean shift is a partitional clustering method that aims to find the high point of the distribution. In k-means method, there is a risk of dividing a cluster to multiple clusters. However, in the mean shift method the risk is too low. Obtained result from this method shows the affect of time even more clearly.

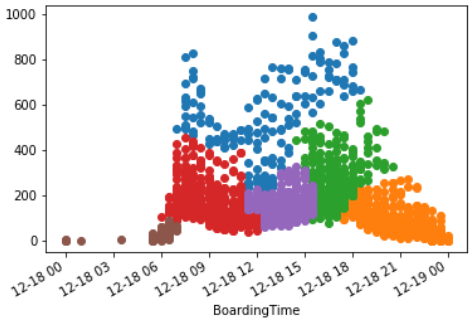

Figure 8. Agglomerative hierarchical clustering

Hierarchical methods are different kind of clustering method which uses an iterative method to either combine smaller clusters or divide a cluster to a given number. Because of their method, they are not especially inclined to spherical clusters, such as circles, which can be seen as an advantage compared to k-means method. Here it is seen that sharper dividers between clusters and a new cluster at the smaller values for the earlier hours.

If the number of clustering changed to 7, then the other divider between the top groups is seen as shown in the Figure 9.

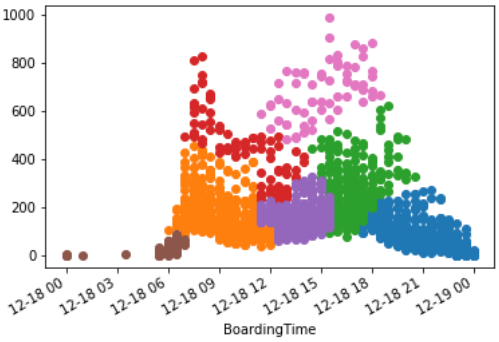

Figure 9. Agglomerative hierarchical clustering with 7 clusters

In this study the number of passengers of the bus lines in the date of 18-12-2019 is shown the dataset analyzed by using various clustering algorithms. Firstly, to understand the structure of the data, the dataset was visualized. With these graphs a pattern is appeared for bus lines. This pattern shows the higher passenger count of the bus lines are seen in the time of working. The least and the most used lines are identified, and they compared with the average and their intensity during the day was visualized. Then all the data was visualized to show different bus lines over a diagram. To analyze the similarities between bus lines, the data converted into time series according to the bus lines and correlation analysis performed. The result of the correlation coefficients shows the mean is 83.7% and the standard deviation is 7.7%. The values show most of the lines fit the daily pattern. After that, the time converted to minutes because clustering methods using the Euclidean distance. Scikit library used to implement clustering methods. K-means clustering, mean shift clustering and agglomerative hierarchical clustering methods selected because those methods generate meaningful results. Results of this study shows that the data clustered depend on time. When confirming the result with correlation analysis, it is noticed that certain pattern of the bus lines corresponds with the analyzation. Lastly, k-means clustering method is used with changing the time unit as seconds and hours as an alternative approach to the minute. Observation shows that the clustering

with using the second symbolized the clustering based on the difference of the seconds. And while using the hour, the clustering done according to the passenger data. These results did not give much information about the data. It was concluded that using minutes was reasonable.

ACKNOWLEDGMENT

This study is done to get information about clustering models under the data mining course with the public transport system in Antalya city. The author would like to thank the instructor of the data mining course who provide the data for the study and share experiences with his students.

REFERENCES

[1] A. Haussmann: K-Means clustering for beginners, Toward Data Science, 2020.

[2] T. Cristobal, G. Padron, A. Quesada-Arencibia, F. Alayon, G. Blasio and C. R. Garcia: A study on the behavior of clustering travel time in road-based mass transit system, Institude for Cybernetics, University of Las Palmas de Gran Canaria, Spain, 2019.

[3] M. Tahnin Tariq, M. Hadi, Y. Xiao: Pattern recognition using clustering analysis to support transportation system management, operation, and modeling, Politecnica de Madrid University, 2019.

[4] T. Galba, Z. Balkic, G. Martinociv: Public Transportation BigData Clustering, JJ. Strossmayer University of Osijek, 2013.

[5] Wikipedia: Cluster analysis, 2021.

[6] S. Kaushik: An introduction to clustering and different methods of clustering, 2016.

[7] A. Briand, E. Come, M. K. El Mahrsi, L. Oukhellou: A mixture model clustering approach for temporal passenger pattern characherization on public transport, International Journal of Data Science and Analytics, 2016.

[8] E. Come, M. Khalil El Mahrsi, L. Oukhellou: A mixture model clustering approach for temporal passenger pattern characterization in public transport, International Journal of Data Science and Analytics, 2016.

[9] Y. Liu, T. Cheng: Understanding public transit patterns with open geodemographics to facilitate public transport planning, Spatiotemporal big data analytics for transportation application, 2020.

[10] B. Agard, V. Partovi Nia, M. Trepanier: Assessing public transport travel behaviour from smart card data with advanced data mining techniques, Rio de Janerio, Brazil, 2013.

[11] M. Moradi, M. Trepanier: Temporal clusters analysis of public transit passengers using smart card data, Interuniversity Research Centre on Enterprice Networks Logistics and Transportation, Canada, 2018.

[12] Morency, C., Trépanier, M., Agard, B.: Mining smart card data from an urban transit network. In: The International IEEE Conference on Intelligent Transportation Systems, Ecole Polytechnique de Montreal, Canada, September (2009).

[13] C. Morency, M. Trépanier, B. Agard, Measuring Transit Performance usingSmart Card Data, World Conference on Transport Research, San Francisco,USA, June 24-28, 2007.

[14] N. Lathia and L. Capra. How smart is your smartcard measuring travel behaviours, perceptions, and incentives. In Proceedings of the 13th International Conference on Ubiquitous Computing, UbiComp ’11, pages 291–300, New York, NY, USA, 2011. ACM.