Clustering of Boarding Patterns in Public Transport

A Case Study of Public Transport, Antalya

Fatma Erkan  
Computer Science Engineering  
Akdeniz UniversityAntalya, 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.

# 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.

# Objectıve

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 o 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.

# Lıterature 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.

# 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.

## 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.

## 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” [15]. 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:

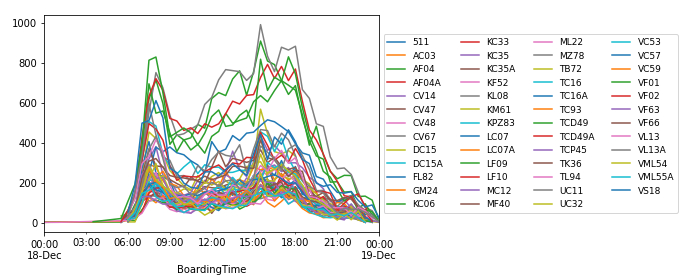
* K-means clustering is one of the simplest models in data mining. Every cluster refers to a group of data points aggregated together because of certain similarities. It begins to work with defining a target number k that refers to the number of centroids for dataset. The centroid represents the center 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.
* Gaussian mixture clustering model finds the mixture of multi-dimensional Gaussian probability distributions that best model the dataset. It can be used to find clusters in the same way with the k-means clustering.

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.

# Results and Dıscussıon

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.

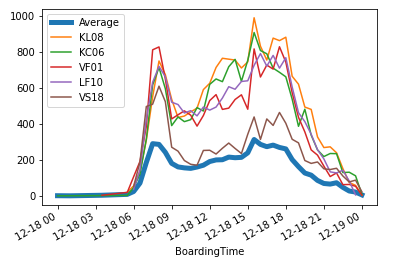
The first step of the study is to analyze the dataset that helps to improve the performance and results when defining the clusters. The dataset contains exactly 1839 data inside of it and every column contains the information about bus line, time interval and the passenger count. It provides information 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.

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. 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.



**Figure 2. Maximum Occurrences of Bus Lines**

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. 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.

**The template is designed for, but not limited to, six authors.** A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

### For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

### For papers with less than six authors: To change the default, adjust the template as follows.

#### Selection: Highlight all author and affiliation lines.

#### Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select the correct number of columns from the selection palette.

#### Deletion: Delete the author and affiliation lines for the extra authors.

## Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1”, “Heading 2”, “Heading 3”, and “Heading 4” are prescribed.

## Figures and Tables

#### Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

1. Table Type Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

1. Sample of a Table footnote. (*Table footnote*)
2. Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

##### Acknowledgment *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

##### References

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

1. <https://towardsdatascience.com/k-means-clustering-for-beginners-ea2256154109>
2. <https://www.mdpi.com/2504-3900/31/1/18/pdf>
3. <https://www.hindawi.com/journals/jat/2019/1628417/>
4. <https://hrcak.srce.hr/file/196661>
5. <https://en.wikipedia.org/wiki/Cluster_analysis>
6. <https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/#:~:text=Clustering%20is%20the%20task%20of,and%20assign%20them%20into%20clusters>
7. <https://link.springer.com/article/10.1007/s41060-015-0002-x>
8. <https://www.researchgate.net/publication/282866062_A_Mixture_Model_Clustering_Approach_for_Temporal_Passenger_Pattern_Characterization_in_Public_Transport>
9. <https://www.tandfonline.com/doi/full/10.1080/23249935.2018.1493549>
10. <https://datawisdom.ca/paper/2013-Transport.pdf>
11. <https://www.cirrelt.ca/documentstravail/cirrelt-2018-28.pdf>
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) - <http://www.mgi.polymtl.ca/agard/fr/publications/doc/2008/2008_Chapter-Agard-Morency-Trepanier.pdf>
13. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.156.2219&rep=rep1&type=pdf>
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.
15. <https://www.antalya-ulasim.com/antalya-otobus-guzergahlari.html>
16. S
17. S
18. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*
19. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
20. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
21. K. Elissa, “Title of paper if known,” unpublished.
22. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
23. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
24. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**