Master's Thesis

Extracting mobility patterns from trip data / trajectories to feed transport models demand

Bisas Dimitrios

Supervised by:

Univ.- Prof. Dr Panayiotis Bozanis

Dr Leonidas Akritidis

March 31, 2022

Abstract

With the enormous development of floating car data, researchers have the ability to get high detail level information. Similarly, the wide expansion of these data gives the opportunity for more comprehensive travel management measures that are the result of spatial and temporal analysis.

In this thesis, a one-week dataset of floating taxi data is used to extract mobility patterns. K-means clustering is used in order to spatially cluster the pick-up and drop-off geographical points, in a way that each cluster represents an origin or destination zone. Therefore, OD matrices have been created that are based on data driven zones and capture the spatial and temporal characteristics of the trips. The next step was the calculation of the similarity of the produced for one week OD matrices with the help of cosine simulation. The results shows the OD matrices from a specific period that are similar to matrices from other periods. This helps in extracting specific and high reliable mobility patterns which capture the travel behavior of taxi users and may help the taxi providers to implement specific operational strategies in order to improve the taxi services.

Keywords: Floating Car Data, Origin- Destination (OD) Matrix, Spatial Clustering, Mobility Patterns, Travel Behavior, Similarity Measurement

Acknowledgements

I would like to express my sincere graditude to Prof Dr. Panayiotis Bozanis and Dr Leonidas Akritidis of the Department of Science and Technology at the International Hellenic University for offering me the topic and for their constant supervision.

Special thanks to Dr Josep Maria Salanova Grau head and researcher at Hellenic Institute of Transport for providing the data and his admirable support and guidance throughout my thesis. He motivated me to implement my proposed method but always steered me in the right direction.

I would also like to thank my friend M.Sc Rachit Khandelwal for his support in this thesis.

Last but not least, I have to express my very profound graditude to my parents, my fiancee, my sister and all my friends for providing me with unfailing support and continuous encouragement throughout my year of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you!

Declaration

I hereby declare that this thesis is an outcome of my own efforts and has not been published anywhere else before and not used in any other examination. Also, to mention that the materials and methods used and quoted in this thesis has been properly referenced and acknowledged.

Metsovo, 31 March 2022

_____

Contents

List of Figures

1 Introduction

1.1 Background and problem statement

Huang, Li, and Xu (2016) stated that the movements of citizens in the cities and more general their travel behavior and the reasons behind it on different time periods formulate the human mobility and travel patterns. Lian et al. (2018) supported that researching and understanding the travel patterns and behavior play a significant role in travel management, traffic forecasting, public transport management and even in parking management. The development of technology has provided a big variety of data sources such as social media, floating data, data from sensors. All these sources have been utilized from the researchers since they have been proven to be very reliable since they are based on real data. This aspect provides the ability to authorities to get high quality data in order to extract mobility patterns for an area.

However, some challenges occur from these technologies due to the enormous massive data and their complexity as far as the spatial and temporal dimension. Kumar et al. (2018) proposed the spatial clustering as a solution to this challenge since this method can identify distinct groups traffic groups based on their geographic characteristics in a way that the similarity is increased between elements of each group.

One of the most important contributions of these data collecting technologies is the extraction of spatial and temporal information of the citizens' trips. This information represents the travel demand and if it is summarized in matrices then we can have clear information about the origin and destination demand in an urban area. To date the research about origin destination matrices (OD matrix) was based on fixed zones which were formulated based on geographical criteria. OD matrix is the most common way to represent the travel demand of an area. Each origin and each destination represents a node or a zone of a network. In the matrix for each origin there is a row and for each destination there is a column, while each cell stands for the demand for the specific origin-destination pair. This thesis proposes a new method where the formulated time dependent origin and destination zones of the OD matrices are formulated based on real data, thus they can represent in much higher accuracy the travel demand of an area.

1.2 Objective and research question

This thesis utilizes floating taxi data in order to extract travel patterns and the mobility behavior of taxi users. Since from the data itself cannot be extracted important conclusions about the

travel patterns it is necessary to group these data by using clustering techniques. The produced clusters represent the origin and destination zones which with the traffic flows are the basis to formulate time and spatial dependent OD matrices. Then the similarity of these matrices is calculated in order to extract specific and accurate travel patterns.

1.3 Report structure

This thesis includes four main parts: a detailed literature review, data description and analysis, explanation of the methodology, results and conclusion.

- Literature Review: In this chapter a detailed literature review for all the aspects of this project is done. Each followed process is explained and stated by a plethora of other studies.
- Data Analysis: The floating car data are analyzed at high level in order to get a clear view of the taxi movements and their characteristics such as travel time and distance. With the help of some graphs the researchers can understand and receive significant conclusions.
- Methodology and Results: In this chapter, the developed methodology is explained in details with the help of graphs. Every step of the process is explained. Then the results are presented in a detailed manner.
- Conclusion: This chapter discusses the results of the methodology, the limitations and suggests potential future research.

2 Literature Review

2.1 Floating Car Data

The wide spread of mobility data thanks to the many technologies which have been developed the last years has given the opportunity to researchers and the city authorities to get a clear insight of the urban mobility. The main purpose of the Floating Car Data (FCD) is to collect real-time data based on the collection resource.

Especially the FCD has become the core of mobility data collection way and has a continuously increasing usage due to its very low cost and high coverage despite the reliability problems (Altintasi et al. 2017). Huber, Ladke, and Ogger (1999) confirm that FCD can be collected at low cost and can provide real-time and high quality data in total contrast to traditional methods that were not so effective and much more expensive in terms of infrastructure installation. FCD cannot provide direct information regarding the traffic flow or jams but due to their real-time information they can be used in many cases such as the OD matrices estimation, traffic queue detection, route choice guidance. FCD include two main categories the GPS based and the GSM cell-tower systems.

FCD has found a wide widespread use in traffic state estimation process. With their high detail level provide all the information that are necessary to estimate the relationship between the traffic flow parameters (speed, flow and density). Zhao et al. (2009) used FCD to estimate traffic flow parameters in order to derive the flow-speed relationship in the fundamental diagram. Reinthaler (2007) tried to estimate the traffic states for the City of Dusseldorf in Germany by using FCD from taxi fleet and the public transport. The main purpose of this study was to develop a model that could use data from different sources in order to recognize the most congested city points. Li et al. (2012) used FCD from 3 month period in order to examine the variations of the average speed. Based on the variations of the average speed the researchers managed to determine the congested points.

2.2 GPS based FCD

GPS based FCD is the most used and widely expanded data source since the higher percentage of vehicles have been equipped with a GPS system. This system utilizes the GPS technology and transmits the position of the vehicle to the service provider. This technology provides information about the exact location at a specific time point and gives the ability to extract many significant

conclusions about the travel behavior of the vehicle. Lin and Hsu (2014) explain the limitations of GPS that include the blocking of signals close to tall buildings or underground places, the limited devices' energy due to continuous transmission. These limitations lead to lower quality of data for long periods.

Xu et. al (2013) developed a method based on numerical statistics of FCD which were obtained from GPS based FCD of taxi fleets. They tried to evaluate and locate the congested points with a spatial temporal progress. Chase et al. (2012) tried to evaluate the speed data from 3 different sources such as FCD, microwave radar sensors and radar sensors. Then the obtained speed data were compared in order to create the travel patterns and investigate the differences among them. Anuar, Habtemichael, and Cetin (2015) tried to describe the relationship between speed and flow based on the Van Aerde's diagram with FCD and data loop detectors.

Jones et al. (2013) presented a method that used Support Vector Machine (SVM) in order to predict the travel time for some links of a traffic network using GPS based FCD. The study included the predicting for congested and uncongested conditions. Chu, Oh, and Recker (2005) used also FCD data to predict travel times by using Adaptive Kalman Filter (AKF) aiming at estimating noise statistics of the model by adapting to the real-time data. Miwa et. al (2004) managed to develop a travel time and route prediction model by using FCD. The proposed method included a Link Cost Table method that tries to utilize the link costs of each link in order to identify the best route.

2.3 The GSM cell-tower systems

With GSM cell-tower systems the data are transmitted usually from the mobile phones to the cell towers that provide the connecting services for a specific location. The main limitation of this technology is the constrained resolution for the location since the ability of a tower varies from hundreds square meters to square kilometers. The paths also in many cases are defined by a series of discontinuous sudden jumps which in many cases are hardly observable in fine details between the destinations (Lin and Hsu 2014). Another problem is the complexity of algorithms which are required to extract high-quality information (Abu-Aisha 2018).

Haghani et al. (2010) compared the average speeds from GPS based and GMS cell-tower FCD. He performed a statistical analysis for multiple speed categories and compared the results between the two different data sources for every speed limit.

2.4 Spatial-temporal clustering

Hair (2009) indicates that cluster analysis tries to divide data components into clusters so that the homogeneity of components within the clusters and the heterogeneity between clusters is maximized. Cluster analysis belongs to the wide field of machine learning techniques and is characterized as unsupervised learning because it learns from unlabeled and unclassified data. Generally the purpose of cluster analysis is the maximization of resemblance between the intracluster components and the dissimilarity between the intercluster components (Fraley and Raftery 2002).

Kisilevich et al. (2009) define as spatio-temporal clustering the process of grouping objects based on their spatial and temporal similarity and are categorized into four methods: hierarchical, density based, grid based and partitioning.

Hierarchical clustering builds binary merge tree starting from leaves that contain data elements on the top that contain all the data (Nielsen 2016). The produced result of this type of clustering that embeds the nodes on the plane is called dendrogram. In order to implement hierarchical clustering the user has to choose the appropriate linkage function, which defines the distance between two different elements of the dendrogram. Hierarchical clustering presents some advantages as far as visualization due to the dendrogram and the simplicity of the whole method. On the other hand, there are some important disadvantages such as the complexity of calculations that make the hierarchical clustering expensive in terms of time and memory. Another significant disadvantage is its weakness to deal with large datasets due to high time and space complexity.

Doğru and SUBAŞI (2015) stated that density-based clustering generates clusters that are connected points where the density of points is equal to or more than a threshold and in the case that density is less than the threshold then the data is considered as noise. The threshold indicates the minimum number of points in a cluster, which includes core and boundary points. The core points are these that meet the above condition while the boundaries that do not, are treated as noise.

Madhulatha (2012) defined that grid-based approach divides the space into smaller finite number of cells that formulate a grid structure on which the clustering action take place.

The partitioning methods are iterative processes that after setting an initial pre-defined number of groups the objects are iteratively relocated among groups till the convergence. Partitioning methods are presented through two methods the k-mean and the k-medoids algorithm.

The clustering methods have been applied extensively in transportation field the last decades and have replaced other methods that could not provide reliable results. Anand et al. (2018) stated that clustering and especially spatial clustering is the most common big data technique used in transportation data analysis. Their extensive use of course is due to the widespread use of FCD and is applied in cases such as traffic flow forecast, traffic accident detection, mobility patterns detection, urban planning, congestion management.

Abu-Aisha (2018) developed an hierarchical clustering analysis in order to extract urban mobility patterns for the city of Munich with the help of FCD. Based on the GPS location of the vehicles tried to investigate the origins and destinations of the drivers. He used also an innovative tool the Relative Deviation Area (RDA) in order to study the travel behavior of drivers in the clusters. This tool calculates the relative area by which a given trip deviates from the least-cost route. Kim and Mahmassani (2015) used density-based clustering algorithm on traffic flow data to investigate the temporal traffic pattern of each spatial cluster. The researchers applied the clustering method in combination with Longest Common Subsequence (LCS) as similarity tool for trajectories. This tool considers the level of closeness, relatedness and potential interaction between two vehicle trajectories. Altintasi, Tuydes-Yaman, and Tuncay (2017) tried to detect mobility patterns from FCD. They tried to estimate the level of service (LOS) for the urban network of Ankara and proved that the proposed method appeared effective in detecting congestion and bottlenecks.

Doğru and SUBAŞI (2015) compared a group of clustering methods in order to detect traffic accidents. The DBSCAN method outperformed the other methods and managed to detect 100% of the accidents. Ding et al. (2016) developed spatial clustering techniques to study the traffic flows in and out of transport hubs in order to get a clear insight of mobility patterns. They divided the study area into smaller OD zones and then studied the travel behavior of passenger within each OD zone. Chen et al. (2014) proposed a DBSCAN clustering method where the GPS trajectories are divided into line segments that are used to find the clusters with similar behavior.

2.5 GPS trajectories from taxis

An important source of FCD are taxis since they belong to the public transportation system of cities and represent a massive part of urban mobility. The GPS-equipped taxis provide tools to discover the spatio-temporal patterns of people movements and even origins and destinations distributions (Tang et al. 2015). There have been many studies conducted that use FCD technology in order to extract significant results.

Normal flow mapping approaches use predefined geographical units (e.g. states) in order to estimate the traffic flows. For this reason Zhu and Guo (2014) developed an hierarchical clustering method that considers both origins and destinations for the definition of traffic flow, where each cluster stands for the flow between origin and destination pair. Due to the hierarchical nature of clustering the method managed to generalize demand to different levels and thus minimize the potential information loss and has the potential to support multi-resolution flow mapping.

Wan et al. (2013) developed a prediction of origin destination method with the development of an improved DBSCAN algorithm. They proved that the proposed method with the clustered origins and destinations of taxis managed to predict with high accuracy the travel behavior of taxis for different regions and in different time periods. Kumar et al. (2016) tried to extract urban mobility patterns of taxi passengers for the city of Singapore using data from a large trip dataset by implementing a clustering algorithm that creates cluster of the trip data points and provide meaningful conclusions about the city mobility. Liu et al. (2015) used taxi trajectory data not only to reveal travel patterns for the city of Shanghai but also the interpretability of urban and suburban areas to transportation related issues. The researchers developed spatially embedded networks to model intra-city spatial interactions and revealed two level hierarchical polycentric city structures and proved that suburban regions are more adaptable to local traffic.

Based on taxi data Kumar et al. (2018) proposed a Dijkstra-based dynamic time warping distance measure between two trajectories in order to provide information about the road traffic and then propose better route to users. The study proposed also a clusi-VAT algorithm to suggest the optimal number of clusters and then the results were compared the produced clusters with other typical clustering algorithms. The results indicated that the proposed method appeared more effective in terms of the used measures of evaluation and the loss of cluster quality.

Mao et al. (2016) tried to get an insight of the household travel behavior and the daily urban jobs housing spatial structure by extracting spatio-temporal patterns from taxi trajectory data from the city of Shanghai. Their method included a three-step method that included the spatial clustering of taxis origins-destinations, the estimation of threshold values of OD clusters and the visualization of the results in order to understand the travel behavior of taxi users. Yue et al. (2009) developed a clustering approach that obtains spatio-temporal patterns from taxi trajectory dataset in order to investigate the points of interest of users. Similarly, they created OD matrices in order to understand better the movement patterns and the attractiveness among clusters. XueJin Wan et al. (2015) developed an hierarchical clustering method based on agglomerative clustering analysis (DBH-CLUS) in order to identify pick-up/drop-off points. The applied method appeared to be very effective in identifying important points from the trajectory dataset.

Shen et al. (2017) also tried to extract hotspots of taxi passengers by discovering behavioral patterns. After converting the studied area to polygons they tried with the help of Moran's Index to estimate the spatial autocorrelation of spatial distribution of the produced spatial clusters. Guo et al. (2012) tried to extract spatial clusters from GPS dataset in order to recognize important hotspots and map the flow measures of clusters for understanding the spatial distribution and temporal trends of movements. The process was applied in taxi trajectories in Shenzhen, China and the approach led to satisfying results regards to patterns extraction for large datasets. Xintao Liu and Ban (2013) analyzed over 85 million taxicab GPS points collected in Wuhan, China for generating spatio-temporal clusters that indicated the real-world traffic congestion.

2.6 OD matrix comparison and similarity measures

The comparison of matrices is divided into two categories: the similarity measures and the structural similarity measures. As far as the similarity measures Ashok (1996) used root mean square error in order to estimate similarity of OD matrices which were obtained from different sources. Tavassoli et al. (2016) used R squared for the comparison of OD matrices obtained from public transport. Ros-Roca et al. (2018) utilized entropy measure for the comparison of OD matrices.

On the other hand, Ruiz de Villa, Casas, and Breen (2014) proposed a new method for OD similarity measure which was based on Wasserstein´s distance but this method proved to be very

costly in computation time. Mungthanya et al. (2019) used KMEANS clustering method followed by a DBSCAN for the outliers to cluster the origins and destination points of taxi floating data. They used the cluster to formulate time and space dependent origin and destination matrices. In order to create patterns within the matrices the cosine similarity is calculated.

3 Dataset Analysis

3.1 Dataset Description

The available data for this thesis were collected from a taxi service company in Thessaloniki, Greece. The dataset includes information about 170.000 trips from taxis over a period of a month (May 2019). The metadata that the data frame contains is trip id, pick-up and drop-off coordinates, trip start timestamp, trip duration and trip length. Thessaloniki is the second largest city in Greece with a population of 800,000. Thessaloniki is a seaside city and presents high population density and all historic, commercial and recreational points are appeared if not in the city center then very close to it.

3.2 Data Limitations

Even though the dataset is very descriptive and provides many information about the taxi movements and usage of the citizens, some limitations occur which limit

the level to which this thesis can be extended. First of all, even though the dataset contains a huge number of trips it does not contain all taxi trips but only those which are recorded but this specific service provider. The wide majority of the trips is viewed in the city center so this can lead to some difficulties as far as the clustering process. The provided data include information only for the start, the end the total travel time and distance of a trip without any information about the possible intermediate stops that may increase the total travel time and distance.

3.3 Data Analysis

Before presenting the main part of this thesis an extended analysis of the data is implemented in order to get a clear picture of the taxi trips and the citizens movements. This process provides important information about the travel behavior of citizens who do not use the public transport of the city and prefer to move with in more privacy and comfort. So, some very specific results can be extracted since the exact time and space of every trip is provided.

An extensive data analysis has been applied in order to extract some results that could help the researchers get a clear view about the travel behavior of citizens who use taxi for the movements in the city of Thessaloniki. First, it is important to get an overview of the travel behaviour of taxi users as far as the travel time and travel distance. Figure 1 and Figure 2 show in detail the distributions of the travel time in min and travel distance in m. It is obvious that the most of trips

is between 5 and 12 min regards to time and 1000 to 3000m regards to trip duration. These two metrics agree totally to the observation that most of the trip are observed in the city center and more specifically start and end in the center.

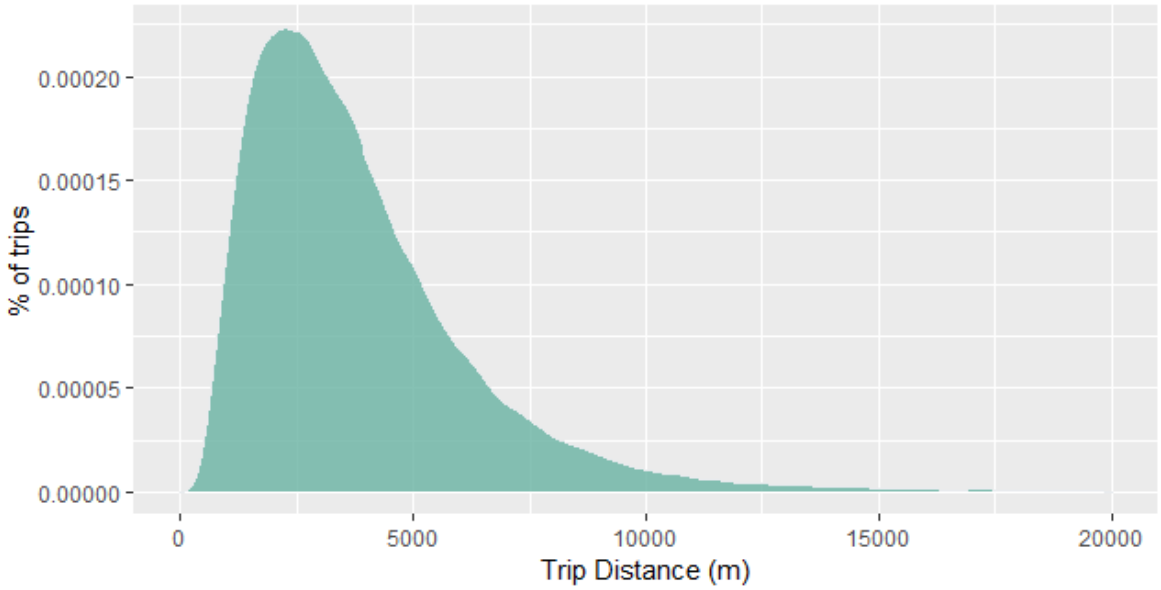

Figure 1 Distribution of trip distance

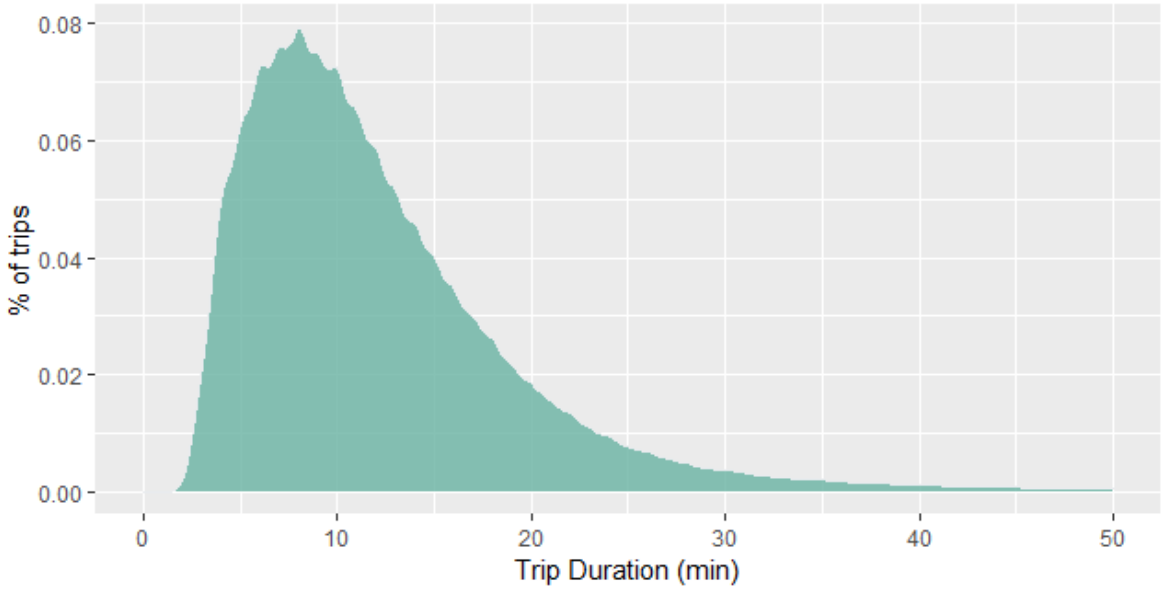

Figure 2 Distribution of trip time

Figure 3 shows the hourly average number of trips for every day. The x-axis is the time horizon of 24h and y-axis is the number of trips. Saturday and Sunday it is normal to have similar numbers since the trips these days present different characteristics and purpose such as leisure from other weekdays. For this reason, these specific two days the number of trips is higher early in the morning since more people are staying out till late and due to the lack of public transport they prefer taxis to return home. On the other hand, an important point is the increased number of trips between 10.00 and 13.00 on weekdays. This behavior indicates that people who travel early in the morning between 6.00 and 9.00 do not use the taxis, but another means of transport. Thus, the citizens who travel for other purposes tend to increase the numbers of taxi trips.

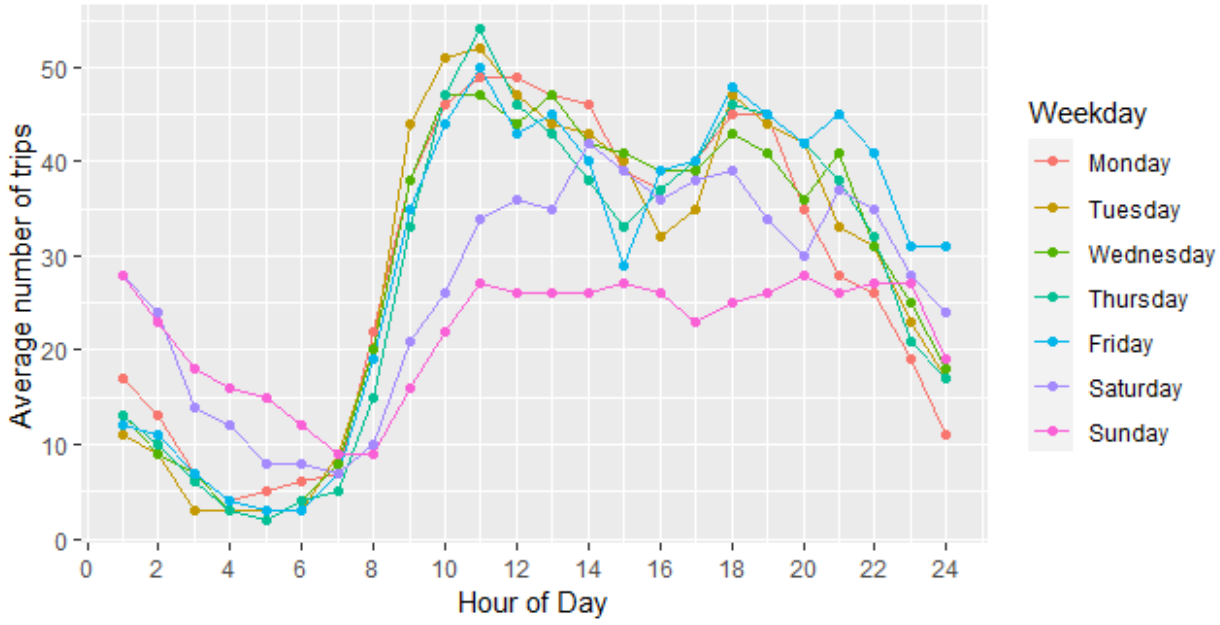

Figure 3 Hourly average number of trips

Figure 4 and Figure 5 show the hourly distribution of travel time and distance for weekdays and weekends. As far as travel time, in the early hours it is very low since there is low traffic load in the city network and the taxis can drive faster to the destinations. Similarly, the travel distance for these early in the morning hours is small indicating that citizens use the taxis for moving close the city center and not in the suburbs. It is obvious in both graphs that the numbers are increased as we move to the rush hours since both lines for weekdays and weekends are following an upward trend.

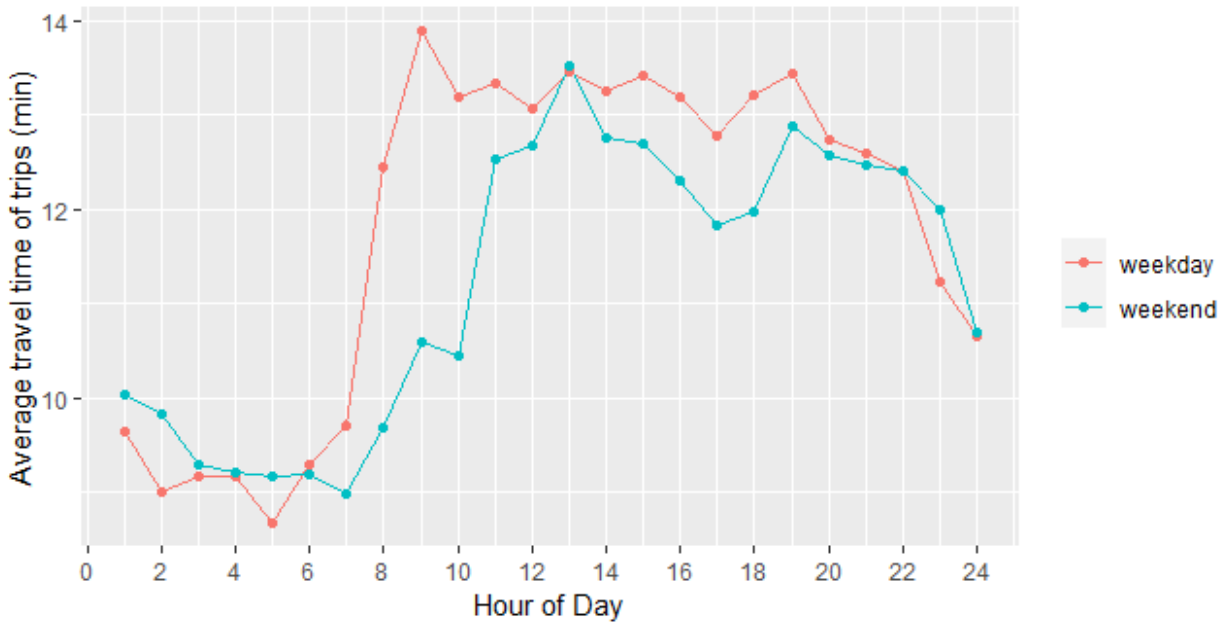

Figure 4 Hourly average travel time

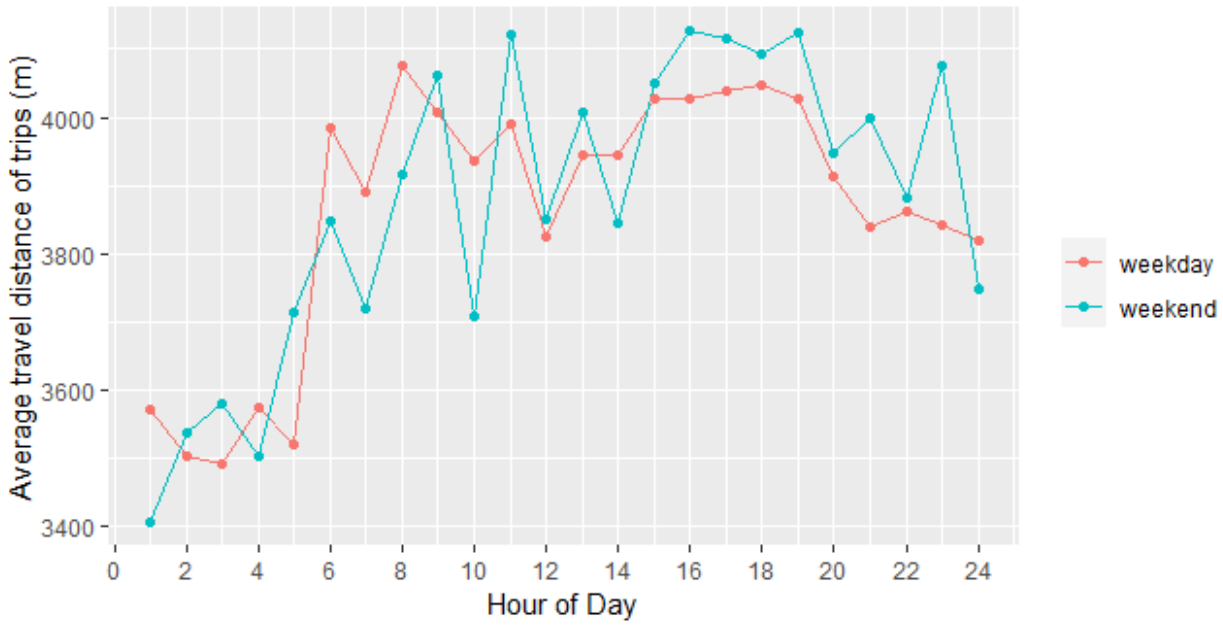

Figure 5 Hourly average travel distance

4 Methodology

The methodology of this thesis includes the construction of time-dependent OD matrices based on adaptive zoning scheme. This scheme is dependent on matrices of different size that capture the time and space variances of an area, so a method is required that tries to estimate the similarity between the available matrices. The developed process is presented in the flowchart below (Figure 6). The first step was the data preparation and cleaning. The next step is the clustering of origins and destination points with the usage of k-means clustering method. The next step includes the geographical formulation of the origin and destination zones with the representation as an enclosed convex hull polygon (Graham 1972). The final part is the similarity measure of the produced time-dependent OD matrices with the calculation of cosine similarity across the matrices.

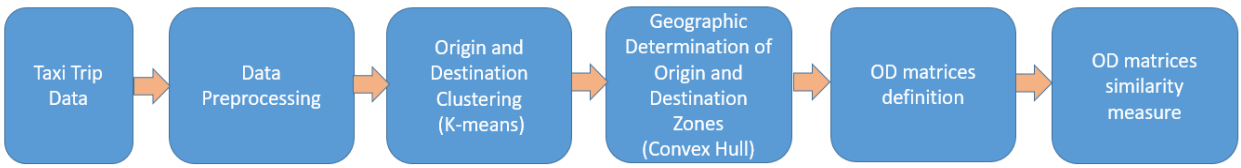

Figure 6 Method flowchart

4.1 Clustering Algorithm

K-means is an unsupervised iterative clustering method. The algorithm starts with a randomly selected number of centroids and based on a selected criterion tries to calculate the optimum number and position of centroids. The iteration process halts either when the centroids have been stabilized or the predefined number of iterations has been achieved. In this thesis the k-means clustering is used to separate into clusters the pick-up and drop-off taxi points. At this point it is important to mention that due to the high concentration most of points to a small area, (close to the city center) the predefined number of seven clusters is used in order to avoid the formation of a huge cluster that will capture the majority of the points and will miss information about the rest points, that give us interesting results about the movements for areas close to the city center. Figure 7 shows the pick-up (left) and the drop-off points (right) for a three-hour period from a weekday.

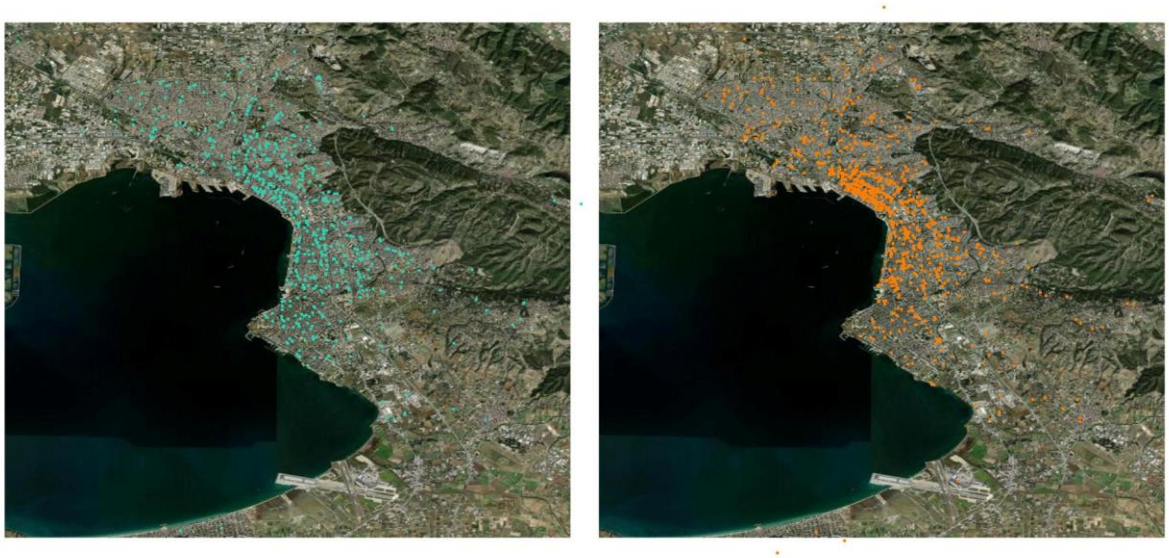

Figure 7 Left: Pick-up points. Right: Drop-off points

4.2 Convex Hull

In order to obtain the travel demand within origin-destination areas the zones have to be formulated and defined geographically based on the clustered points that came up after the k-means clustering process. Each cluster represents either an origin or destination zone and also each cluster is represented with a convex hull polygon where the most outer points were connected and created the polygon's edges. Figure 8 presents the origin and destination zones, that were formulated with the process of convex hull polygon.

Figure 8 Left: Origin clusters (zones). Right: Destination clusters (zones)

4.3 OD Matrix Construction

The OD matrix is one of the most important element of the transportation systems since it can represent in a clear way the travel demand within zones. There are many ways to create these matrices such as surveys, questionnaires and data from detectors which tend to be very costly in terms of time and money. In this project the matrices are created based on the GPS data obtained from taxis. All the vehicles that took part in this study, their starting and ending trip points define respectively their origin and destination points and these points are used to formulate the origin and destination zones. The main advantage of this method is the definition of the exact position of the origin and destination of each trip, which can lead the transport authorities to capture with high detail the travel behavior of the citizens.

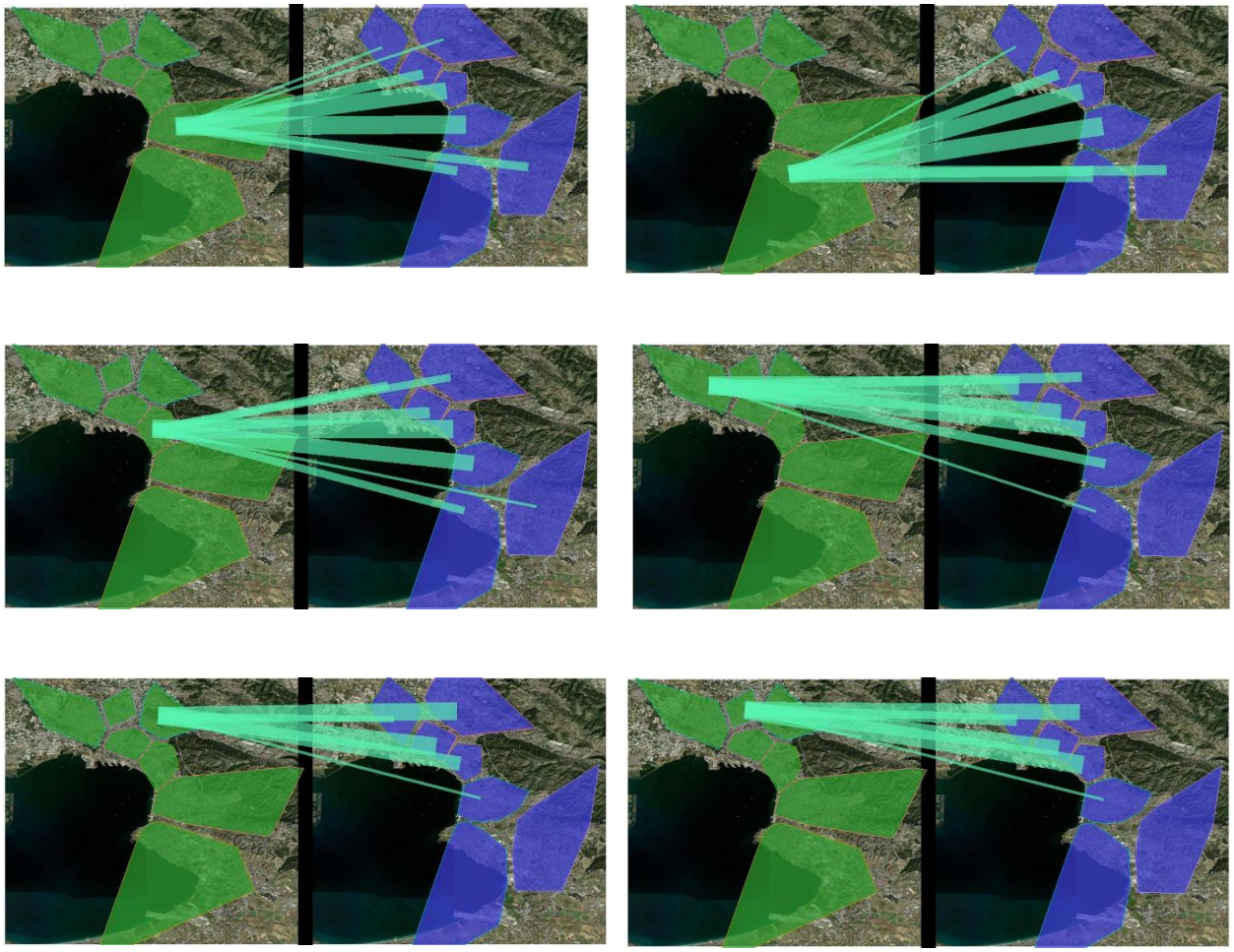

Figure 9 Trip flows based on time dependent OD matrix developed by using the proposed methodology for 6AM- 9AM

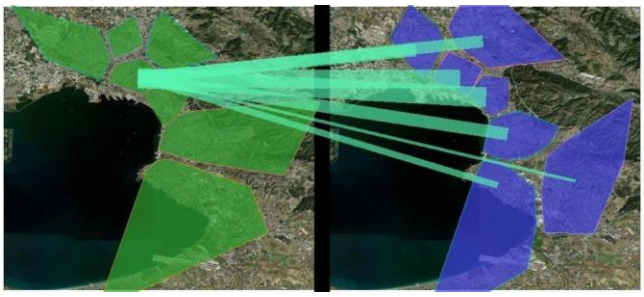

Figure 10 Trip flows based on time dependent OD matrix developed by using the proposed methodology for 6AM- 9AM of one origin zone

4.4 OD Matrix Similarity Measure

The OD matrices of taxis manage to capture the travel demand and user behaviour of an area with high accuracy. Since the OD matrices capture the demand for a specific time period it is crucial for the taxi provider to understand the changes as far as time and space. This advantage provides the taxi service the ability to implement an effective in time and space oriented management, thus reducing the cost, and improving the quality of services.

Given the high variability of OD matrices in temporal and spatial dimension it is difficult to extract patterns of the traffic analysis zones since the existing approaches implement simple comparison methods such as root mean squared error, entropy etc. In this thesis an approach (Mungthanya et al. 2019) is used that tries to measure the similarity between dynamic spatiotemporal-variant OD matrices. The most challenging part is the transformation of the matrices and all the relevant information into a comparable format such as vectors whose similarity is measured with the cosine similarity as it is appeared in the Equation 1.

$$S(X,Y)=\cos(\theta)=\frac{X\,Y}{\|X\|\|Y\|}=\frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2}\sqrt{\sum_{i=1}^{n} Y_i^2}} \quad (1)$$

Where X_i and Y_i are the components of the vector X and Y. The cosine similarity is a measurement of cosine of two vectors and shows the angle between them. As long the similarity increases the angle between the vectors is decreased and a similarity equal to 1 means two vectors with same orientation. Similarly, two vectors with similarity of 0 means that they are perpendicular to each other. This process allows the researchers to receive a clear view of the similarity measurement between the matrices, which have been transformed into vectors.

If we have two OD matrices, matrix A with N origins and M destinations and matrix B with U origins and Z destinations. We define O^A the set of origins where O^A=$\{O_1^A, O_2^A, \dots, O_N^A\}$ with N the number of origins and each dataset O_i^A= $\{o_i^A(lat), o_i^A(lon)\}$ contains the latitude and longitude

of its centroids. Accordingly, the respective vectors are formulated for the destinations D with a total of M destinations. Matrix B uses the same mathematical formulations. From each origin i of the A matrix there is travel demand that flows to each of the M destinations. So a vector is created for each that gathers all the important information and characteristics such as the travel flow, the coordinates of origin and destination centroids. The resultant vector has the following form:

$$R_i^A = [T_i^A, O_i^A, D_i^A] \quad (2)$$

Where:

O_i^A : coordinates of centroid of origin zone

D_i^A: coordinates of centroid of destination zone

T_i^A: traffic flow from an origin zone to the destination

In order to measure the similarity of two matrices A and B their resultant vectors are used, since they include all the important information. The cosine similarity for matrix A and B with a total of N and U number of origin zones respectively is calculated between the resultant vectors R_i^A and R_j^B with i = 1,2, … N and j=1,2, … U. The similarity measure is extracted from the average value for each i and j and is denoted by C_i^A and C_j^B respectively. The final similarity is calculated from the equation below:

$$Sim=\frac{\sum_{i=1}^{N} C_i^A+\sum_{j=1}^{U} C_j^B}{N+U} \quad (3)$$

The overall approach is presented in Algorithm 1 (Mungthanya et al. 2019) where the steps for the calculation of the similarity measurement between two matrices A and B are presented in detail.

Input: Resultant vectors from matrices A and B ($\{R_i^A\}$ and $\{R_j^B\}$) Output: Similarity value (Sim)

For $i \leftarrow 1$ to N (the number of origins of A) do
For $j \leftarrow 1$ to U (the number of origins of B) do
Compute cosine similarity S(R_i^A, R_j^B)
End

End

Determine the average value $C_i^A = mean_{j \epsilon \{1,2,\ldots,U\}}(S(R_i^A, R_j^B))$ For $i \in \{1,2,\ldots, N\}$

Determine the average value $C_j^B = mean_{i \epsilon \{1,2,\ldots,N\}}(S(R_i^A, R_j^B))$ For $j \in \{1,2,\ldots, U\}$

Compute $Sim = \frac{\sum_{i=1}^{N} C_i^A + \sum_{j=1}^{U} C_j^B}{N+U}$

Algorithm 1 Time-dependent OD matrix similarity measure

In conclusion, the similarity of two matrices is measured with the cosine similarity where higher values lead to higher similarity of two vectors. The similarity also is measured based on three elements the traffic flow across the zones, the geographic location (coordinates) of the centroids of origin and destination zones.

5 Results

By implementing the methodology described in Section 4, the cross similarity of OD matrices is measured based on the data from the taxi travel demand in Thessaloniki. The data used for the constructing of OD matrices has captured the demand for a three-hour period of each day. In total 56 OD matrices were developed for a period of one week. Thus, the cross similarity calculation started from Monday 0AM-3Am to Sunday 9PM-0AM was based on 56*56=3.136 values since every OD is compared with all the others. The results are presented in Figure 11 and it is obvious that the similarity in any case is very high and ranges from 0.9 to 1. It is important to mention that very high values do not mean absolutely same OD matrices but travel patterns that are very similar over these time periods. This behavior is absolutely normal since the human mobility follows some patterns and is repeated.

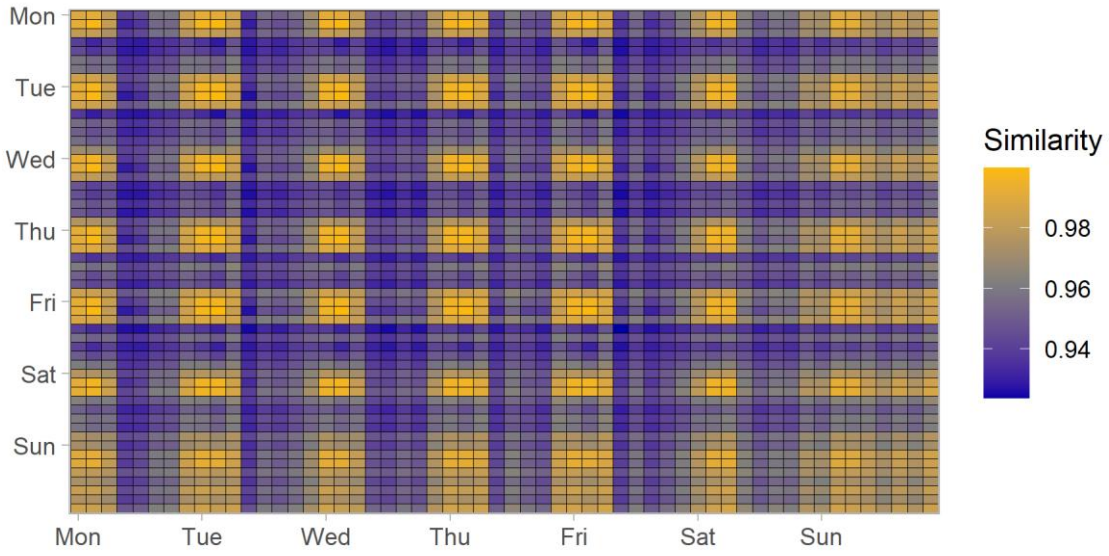

Figure 11 Cross-similarity of all matrices from Monday 0AM-3AM to Sunday 9PM-0AM

Another important result is the high similarity between time periods 0AM-6AM for all days, which means that all these time periods where the traffic load is very low there are some specific travel demand patterns. On the other hand, the similarity values are lower for periods 6AM-9PM where the travel demand is higher. Table 1 shows five time periods with maximum and minimum similarity values. The periods Sat 6AM-9AM, Fri 6AM-9AM, Sun 3AM-6AM, Thu 0AM-3AM and Thu 6AM-9AM show high cross similarity values which means that these time slots present similar travel patterns and allows the authorities to apply uniform taxi operation management

principles. On the other hand, the time period 9AM-12PM for Tuesday, Friday and Monday shows the lowest similarity values indicating that these time periods it is more difficult to extract precise travel patterns.

Table 1 Top maximum and minimum similarity values with the respective time periods

Top five time periods with highest average cross similarity values		
Ranking	Time period	Cross-similarity value
1	Sat 6AM-9AM	0.97288
2	Fri 6AM-9AM	0.97270
3	Sun 3AM-6AM	0.97139
4	Thu 0AM-3AM	0.97136
5	Thu 6AM-9AM	0.97134
Bottom five time periods with lowest average cross similarity values		
Ranking	Time period	Cross-similarity value
56	Tue 9AM-12PM	0.93524
55	Fri 9AM-12AM	0.93622
54	Mon 9AM-12PM	0.93709
53	Fri 3PM-6PM	0.93916
52	Mon 12AM-3PM	0.93936

6 Conclusion

In this chapter a summary of the whole thesis is described along with the limitations and possible further recommendations that can be applied for future research.

6.1 Summary

The floating car data allows taxi service providers to obtain all the important information on each taxi's movements, including not just the travel time and distance but also about the precise place to which each user wishes to travel. All of this high-level data is critical, and it may assist the provider in planning the essential operating measurements for improved service quality. The location of the origin and destination is one of the most essential pieces of information from taxi floating data.

In this research, we used information regarding taxi vehicle origin and destination data to create OD matrices for the city of Thessaloniki. The resulting OD matrices have the advantage of being dynamic, and the origin and destination zones are formed using the k-means clustering approach. In contrast to other techniques that use standard and consequently more static comparison measurements such as R-squared and root mean square error, the cosine similarity between the OD matrices was assessed in this thesis. The elements for the calculation of cosine similarity are the traffic flow and the coordinates of centroids which are formulated from clustering origin and destination zones. This process gives the opportunity to include all the important information of OD matrices in comparison.

The results from the comparison allow the researchers to extract travel patterns for the travel behavior of citizens when they use taxis. Many OD matrices show high cross similarity values indicated the periods with similar travel patterns.

6.2 Limitations and Recommendations

One important problem with this approach is the lack of additional data in order to validate the results and demonstrate that the proposed process of construction data driven OD matrices is reliable. Another limitation is the used clustering method for which the construction of origin and destination zones uses a specific number of clusters and the formulation of clusters was based on the Euclidean distance.

Therefore, this might be the focus of future research, as well as the selection of a different clustering approach with acceptable parameters, which could contribute to the formation of data-driven zones that more accurately replicate travel demand. In the future, more research could be conducted to validate the data.

7 Bibliography

Abu-Aisha, Abdallah. 2018. “Using Floating Car Data (FCD) to Infer Mobility Patterns.”

Altintasi, Oruc, Hediye Tuydes-Yaman, and Kagan Tuncay. 2017. “Detection of Urban Traffic Patterns from Floating Car Data (FCD).” *Transportation Research Procedia* 22:382–91.

Anand, Sesham, P. Padmanabham, A. Govardhan, and Rajesh H. Kulkarni. 2018. “An Extensive Review on Data Mining Methods and Clustering Models for Intelligent Transportation System.” *Journal of Intelligent Systems* 27(2):263–73.

Anuar, Khairul, Filmon Habtemichael, and Mecit Cetin. 2015. “Estimating Traffic Flow Rate on Freeways from Probe Vehicle Data and Fundamental Diagram.” Pp. 2921–26 in *2015 IEEE 18th international conference on intelligent transportation systems*. IEEE.

Ashok. 1996. “Estimation and Prediction of Time-Dependent Origin-Destination Flows.” *MIT Thesis*.

Chase, R. Thomas, Billy M. Williams, Nagui M. Rouphail, and SangKey Kim. 2012. “Comparative Evaluation of Reported Speeds from Corresponding Fixed-Point and Probe-Based Detection Systems.” *Transportation Research Record* 2308(1):110–19.

Chen, Jiangping, Ting Hu, Pengling Zhang, and Wenzhong Shi. 2014. “Trajectory Clustering for People’s Movement Pattern Based on Crowd Souring Data.” *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* 40(2):55.

Chu, Lianyu, S. Oh, and Will Recker. 2005. “Adaptive Kalman Filter Based Freeway Travel Time Estimation.” in *84th TRB Annual Meeting, Washington DC*. Citeseer.

Ding, Linfang, Mathias Jahnke, Shirui Wang, and Katre Karja. 2016. “Understanding Spatiotemporal Mobility Patterns Related to Transport Hubs from Floating Car Data.” Pp. 175–85 in *Proc. Int. Conf. Location-Based Services*.

Doğru, Nejdet and Abdülhamit SUBAŞI. 2015. “Comparison of Clustering Techniques for Traffic Accident Detection.” *Turkish Journal of Electrical Engineering & Computer Sciences* 23(Sup 1):2124–37.

Fraley, Chris and Adrian E. Raftery. 2002. “Model-Based Clustering, Discriminant Analysis, and Density Estimation.” *Journal of the American Statistical Association* 97(458):611–31.

Graham, Ronald L. 1972. “An Efficient Algorithm for Determining the Convex Hull of a Finite Planar Set.” *Info. Pro. Lett.* 1:132–33.

Guo, Diansheng, Xi Zhu, Hai Jin, Peng Gao, and Clio Andris. 2012. “Discovering Spatial Patterns in Origin-destination Mobility Data.” *Transactions in GIS* 16(3):411–29.

Haghani, Ali, Masoud Hamedi, Kaveh Farokhi Sadabadi, Stanley Young, and Philip Tarnoff. 2010. “Data Collection of Freeway Travel Time Ground Truth with Bluetooth Sensors.” *Transportation Research Record* 2160(1):60–68.

Hair, Joseph F. 2009. “Multivariate Data Analysis.”

Huang, W., S. Li, and S. Xu. 2016. “A Three-Step Spatial-Temporal-Semantic Clustering Method for Human Activity Pattern Analysis.” *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* 41:549.

Huber, Werner, Michael Ladke, and Rainer Ogger. 1999. “Extended Floating-Car Data for the Acquisition of Traffic Information.” in *PROCEEDINGS OF 6TH WORLD CONGRESS ON INTELLIGENT TRANSPORT SYSTEMS (ITS), HELD TORONTO, CANADA, NOVEMBER 8-12, 1999*.

Jones, Michael, Yanfeng Geng, Daniel Nikovski, and Takahisa Hirata. 2013. “Predicting Link

Travel Times from Floating Car Data." Pp. 1756–63 in *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*. IEEE.

Kim, Jiwon and Hani S. Mahmassani. 2015. "Spatial and Temporal Characterization of Travel Patterns in a Traffic Network Using Vehicle Trajectories." *Transportation Research Procedia* 9:164–84.

Kisilevich, Slava, Florian Mansmann, Mirco Nanni, and Salvatore Rinzivillo. 2009. "Spatio-Temporal Clustering." Pp. 855–74 in *Data mining and knowledge discovery handbook*. Springer.

Kumar, Dheeraj, Huayu Wu, Yu Lu, Shonali Krishnaswamy, and Marimuthu Palaniswami. 2016. "Understanding Urban Mobility via Taxi Trip Clustering." Pp. 318–24 in *2016 17th IEEE International Conference on Mobile Data Management (MDM)*. Vol. 1. IEEE.

Kumar, Dheeraj, Huayu Wu, Sutharshan Rajasegarar, Christopher Leckie, Shonali Krishnaswamy, and Marimuthu Palaniswami. 2018. "Fast and Scalable Big Data Trajectory Clustering for Understanding Urban Mobility." *IEEE Transactions on Intelligent Transportation Systems* 19(11):3709–22.

Li, Qiang, Qian Ge, Lixin Miao, and Mingyao Qi. 2012. "Measuring Variability of Arterial Road Traffic Condition Using Archived Probe Data." *Journal of Transportation Systems Engineering and Information Technology* 12(2).

Lian, Jing, Yang Li, Weixi Gu, Shao-Lun Huang, and Lin Zhang. 2018. "Joint Mobility Pattern Mining with Urban Region Partitions." Pp. 362–71 in *Proceedings of the 15th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*.

Lin, Miao and Wen-Jing Hsu. 2014. "Mining GPS Data for Mobility Patterns: A Survey." *Pervasive and Mobile Computing* 12:1–16.

Liu, Xi, Li Gong, Yongxi Gong, and Yu Liu. 2015. "Revealing Travel Patterns and City Structure with Taxi Trip Data." *Journal of Transport Geography* 43:78–90.

Liu, Xintao and Yifang Ban. 2013. "Uncovering Spatio-Temporal Cluster Patterns Using Massive Floating Car Data." *ISPRS International Journal of Geo-Information* 2(2):371–84.

Madhulatha, T. Soni. 2012. "An Overview on Clustering Methods." *ArXiv Preprint ArXiv:1205.1117*.

Mao, Feng, Minhe Ji, and Ting Liu. 2016. "Mining Spatiotemporal Patterns of Urban Dwellers from Taxi Trajectory Data." *Frontiers of Earth Science* 10(2):205–21.

Miwa, Tomio, Takaaki Sakai, and Taka Morikawa. 2004. "Route Identification and Travel Time Prediction Using Probe-Car Data." *International Journal of ITS Research* 2(1):21–28.

Mungthanya, W., S. Phithakkitnukoon, M. G. Demissie, L. Kattan, M. Veloso, C. Bento, and C. Ratti. 2019. "Constructing Time-Dependent Origin-Destination Matrices With Adaptive Zoning Scheme and Measuring Their Similarities With Taxi Trajectory Data." *IEEE Access* 7:77723–37.

Nielsen, Frank. 2016. "Hierarchical Clustering." Pp. 195–211 in *Introduction to HPC with MPI for Data Science*. Springer.

Reinthaler, Martin. 2007. "Evaluation of Speed Estimation by Floating Car Data within the Research Project Dmotion."

Ros-Roca, Xavier, Lídia Montero, Arne Schneck, and Jaume Barceló. 2018. "Investigating the Performance of SPSA in Simulation-Optimization Approaches to Transportation Problems." *Transportation Research Procedia* 34:83–90.

Ruiz de Villa, Aleix, Jordi Casas, and Martijn Breen. 2014. *OD Matrix Structural Similarity: Wasserstein Metric*.

Shen, Jingwei, Xintao Liu, and Min Chen. 2017. "Discovering Spatial and Temporal Patterns from Taxi-Based Floating Car Data: A Case Study from Nanjing." *GIScience and Remote Sensing* 54(5):617–38.

Tang, Jinjun, Fang Liu, Yinhai Wang, and Hua Wang. 2015. "Uncovering Urban Human Mobility from Large Scale Taxi GPS Data." *Physica A: Statistical Mechanics and Its Applications* 438:140–53.

Tavassoli, Ahmad, Azalden Alsger, Mark Hickman, and Mahmoud Mesbah. 2016. "How Close the Models Are to the Reality? Comparison of Transit Origin-Destination Estimates with Automatic Fare Collection Data." Pp. 1–15 in *Australasian Transport Research Forum*.

Wan, Xuejin, Jianchu Kang, Mengdan Gao, and Jianling Zhao. 2013. "Taxi Origin-Destination Areas of Interest Discovering Based on Functional Region Division." Pp. 365–70 in *Third International Conference on Innovative Computing Technology (INTECH 2013)*. IEEE.

Wan, XueJin, Jiong Wang, Yong Du, and Yuan Zhong. 2015. "DBH-CLUS: A Hierarchal Clustering Method to Identify Pick-up/Drop-off Hotspots." Pp. 890–97 in *2015 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*. IEEE.

Xu, Lin, Yang Yue, and Qingquan Li. 2013. "Identifying Urban Traffic Congestion Pattern from Historical Floating Car Data." *Procedia-Social and Behavioral Sciences* 96:2084–95.

Yue, Yang, Yan Zhuang, Qingquan Li, and Qingzhou Mao. 2009. "Mining Time-Dependent Attractive Areas and Movement Patterns from Taxi Trajectory Data." Pp. 1–6 in *2009 17th international conference on geoinformatics*. IEEE.

Zhao, Nale, Lei Yu, Hui Zhao, Jifu Guo, and Huimin Wen. 2009. "Analysis of Traffic Flow Characteristics on Ring Road Expressways in Beijing: Using Floating Car Data and Remote Traffic Microwave Sensor Data." *Transportation Research Record* 2124(1):178–85.

Zhu, Xi and Diansheng Guo. 2014. "Mapping Large Spatial Flow Data with Hierarchical Clustering." *Transactions in GIS* 18(3):421–35.