Archetypes of urban travellers: Clustering of mobile phone users in Singapore

Conference Paper

Author(s):
Anda, Cuauhtémoc; Ordoñez Medina, Sergio A.

Publication date:
2018-06

Permanent link:
https://doi.org/10.3929/ethz-b-000374607

Rights / license:
In Copyright - Non-Commercial Use Permitted
Archetypes of urban travellers:
Clustering of mobile phone users in Singapore

Cuauhtemoc Anda
ETH Zürich
Singapore-ETH Centre
anda@arch.ethz.ch

Sergio Arturo Ordonez Medina
ETH Zürich
Singapore-ETH Centre
ordonez@ivt.baug.ethz.ch

Abstract

An important task in transportation planning is to segment the travelling demand into groups of homogenous individuals for a better analysis. Traditionally travel demand is segmented through socio-demographic information. However, this information is often times not available in new streams of Big Data. We propose then a methodology to uncover the different archetypes of urban travellers based only on the mobility traces left by their mobile phones. We defined 5 variables that explain traits of travel behaviour within these digital traces, followed by an evaluation of two clustering algorithms on a dimensional-reduced space. We finally present 16 archetypes of urban travellers for the case of Singapore. Results from this study constitute one of the first steps towards the development of new Big Data travel demand models when privacy is a concern.

1 Introduction

An important task in transportation planning is to segment the travelling demand into groups of homogeneous individuals for a better analysis. Traditionally this process is done through the socio-demographics of the population, for instance, grouping them into workers, students and non-workers. However, socio-demographic information is often times not available in new streams of data such as mobile phone location data or Call Detail Records (CDR). These new streams of data available, despite lacking personal information, they provide a proxy for the trajectory of each user through long observation periods and for large extents of the urban population.

The challenge is then to characterise urban travellers based only on the observed mobile phone network trajectories. This means that we seek to uncover different types of travellers based only on their observed mobility patterns. For this end, we propose a set of features that characterises the travel behaviour of mobile phone users. We then select the appropriate clustering approach and reveal the archetypes of urban travellers for the case of Singapore.

The rest of the paper is organised as follows. Section 2 covers the related work, section 3 describes the mobile phone location dataset used, section 4 the feature engineering to design the variables to cluster, section 5 the clustering strategies considered, section 6 the low dimensional representation for visualising and evaluating the clustering results, section 7 the results of the clustering algorithms and the final clusters selected, and section 8 the conclusions.

2 Related Work

Not much work has been done in regards to characterising different types of travellers based only on their digital trace left. One recent and relevant work is [7] in which daily patterns of human activities are clustered. The original information used to find the clusters came from a large survey

Preprint. Work in progress.
in which different types of activities throughout two days were reported by each of the respondents. Hence, the main focus was to find clusters that represent different types of people depending on the types of activities performed during the two days. In contrast, in this study, we are interested in finding different types of people based on their observed trip behaviour. Another difference is the choice of clustering algorithm and the general strategy to implement it. In [7] they first used Principal Component Analysis (PCA) to reduce their original feature space and then find the clusters via $K$-means clustering. In this study, we tested two more robust clustering algorithms than $K$-means, namely Gaussian mixture clustering (a generalised version of $K$-means clustering) and HDBSCAN (hierarchical density-based clustering for applications with noise) directly on our original space, and then, we used a non-linear dimensionality reduction algorithm to visualise both clustering results and select the one that performed better. Another relevant work is [10]. In there, the urban population is characterised into a set of mobility networks, called *motifs*. This mobility networks are directed graphs were each node represents a unique activity location for a person for a given day. And the edges connect how the individual was visiting each of these locations. It is similar to the concept of *tour* [9] in transport planning but without the label of the performed activity at each node. The authors found that only 17 unique *motifs* can capture up to 90% of the population in surveys and mobile phone datasets for different countries. This gives us a base on the number of expected clusters to obtain in this study. However, we argue that these motifs are not sufficient to segment different types of travellers in a robust way. For instance, a worker that performs his/her commute with three extra activities after work (e.g. workout, dinner, shopping) has a different mobility *motif* than the worker with two extra activities instead (e.g. workout and dinner). In many cases these two workers might represent very similar type of person despite having different *motifs*.

3 Data

The series of geo-coded time-stamps collected by mobile phone network operators constitute what has been commonly referred as Call Detail Records (CDR) dataset [1]. However, since the user location updates in the mobile phone network occur not only as a consequence of calls or SMS messages, but also by means of the internet usage and other telecommunication network mechanisms, we can simply call them mobile phone location datasets. These datasets however need to be firstly processed to account for signal jumps between cell towers [5] and also need to be segmented into stay episodes and travel episodes. Those problems have been tackled before in the literature [8] and for this study the data provider has done these preprocessing steps. This study departs then from mobile phone stay-point data of one of the mobile phone network operators in Singapore. In this preprocessed dataset, a stay-point happens when the user’s mobile phone connects to a unique network antenna for a period of time larger than 15 minutes, meaning that the user was engaged in a non-travel related activity. For privacy reasons the data provider has fully anonymised the dataset and spatially aggregated the locations into *subzone* planning boundaries as defined by the Urban Redevelopment Authority in Singapore[1]. The period of study corresponds principally to the 18th of April 2017, a regular working Tuesday, however, some stay-points were captured from the end of 17th of April 2017 and the beginning of the 19th of April to have the complete observations in the transition boundaries of the day. Other important information about the dataset is that it corresponds to 50% of the population in Singapore with a total of 3,036,073 active mobile phones for the study period.

3.1 Data Processing

The first step was to calculate a measurement of *observability* for each of the stay-point series. This was done because in several cases the time distance between two detected stay-points was more than 2 hours. In Singapore the longest public transport trip is no more than 2 hours, hence we defined the observed travel times of more than 2 hours as periods of *unobservability*. Those unobservable periods can be caused by several factors, for instance, a mobile phone running out of battery or the user switching off his/her mobile phone. In other instances, it can also be caused by errors in the mobile phone network. For the case of mobility and transportation studies, this can be a strong bias when analysing travel times and number of trips/activities. Hence, we filtered out unobservable users

---

by keeping the fully observable ones. A fully observable user is defined as a user with no more than 120 minutes gap between each of his/her pairs of consecutive stay-points. By doing so we ended up with 1,053,369 users which represent 34.7% of the total number of active phones for the day of interest.

4 Feature Engineering

In order to uncover the different archetypes of urban travellers without any contextual information (e.g. no activity labels, no socio-demographic information, no mode of transport) and based only on the stay-point locations, one of the challenges was to design a set of features that could describe some traits of travel behaviour. These traits of travel behaviour need to be extracted by only the information encoded in the series of stay-point locations. We restricted the design of these variables to take into account only the temporal dimension of the trips, and we kept the same units (minutes) for all our 5 variables to avoid any normalisation issues.

Mean of duration of stay-points: This variable calculates the average duration across all stay-locations of a user. It also gives a proxy of the number of trips done by a person. For instance, for non-travellers, the mean of duration of stay-points is 1440 min (i.e. 24hr x 60min/hr), while a frequent daily traveller would have a much lower mean.

Standard deviation of duration of stay-points This variable calculates the standard deviation of the duration across all stay locations of a user. It differentiates users with homogeneous activity durations from users with non-homogeneous activity durations. For instance, someone that stayed almost all day at home and went out for a short activity will have a high standard deviation, while a person with multiple activities with similar durations will have most likely a lower standard deviation.

Bias morning-night This variable is calculated as the difference between the average of stay-points durations before noon (12pm) and the average of stay-point durations after noon. It describes whether a person is an early traveller or a late traveller. For instance, someone that performs several trips during the morning will have a large negative morning-night bias, while someone that stays at home during the morning and later performs several activities in the evening will have a large positive morning-night bias. For the case of users equally active before and after noon, then bias morning-night is close to 0 min.

First departure This variable refers to the time of the day in which the user performs his/her first trip departure. Users with no departures during the day are given a dummy value of 1440 min that represents the right side boundary of the day.

Last arrival This variable refers to the time of the day in which the user made his/her last arrival. Users with no departures during the day are given a dummy value of 0 min which represents the left side boundary of the day. For the case of users that were travelling during midnight of the 18th to the 19th are given the dummy value of 1440 min.

5 Clustering

Our strategy was to test two of the most prominent clustering algorithms in the recent years: Gaussian mixture clustering and DBSCAN (density-based spatial clustering of applications with noise), and compare the results in a 2 dimensional reduced space by T-SNE (t-distributed stochastic neighbor embedding). This allowed us to select the algorithm that shows more consistent clusters in the reduced space.

5.1 Gaussian Mixture clustering

A Gaussian mixture model is a probabilistic model that attempts to find a mixture of multi-dimensional Gaussian probability distributions that best model the input dataset. It assumes then that all data points are generated from this mixture which has a finite number of Gaussian distributions and unknown parameters. In the clustering case it is required to define the total
number of components (i.e. Gaussian distributions) which is equivalent to the total number of expected clusters. The unknown parameters are estimated through Expectation Maximisation (EM) algorithm, which is a method to find the parameters that maximise the likelihood function. For a comprehensive tutorial on EM algorithm and Gaussian mixture models the reader can refer to [2].

In order to find the number of components the Bayesian information criterion (BIC) is commonly used. BIC tests a finite set of models and then calculates a measurement related to the maximum likelihood and a penalty function. This measurement constitutes a trade-off between the goodness of fit of the model and the simplicity of the model. For our case, the BIC resulted in 23 as the optimal number of components for the mixture model.

5.2 DBSCAN

DBSCAN is a density based algorithm that assumes the formation of clusters in regions were datapoints are more dense [4]. Hence, it extracts these dense regions and leaves sparse datapoints unclustered. The way it works is that it first labels core points to all the points that within their neighborhood have at least a minimum number of points, then searches through the core points neighbours for other dense areas and assigned the new dense points to the original core point cluster. This is repeated recursively for all unexplored neighbour points and for all unassigned core points. Hence, the two important parameters of DBSCAN are the neighbourhood parameter or $\epsilon$ which is the radius distance of the neighborhood, and the minimum number of points parameter, which controls the minimum requirement of points to form a dense region. In contrast with Gaussian mixture clustering, it is not required to know a priori the expected number of clusters.

For this study we employed a recent variation on DBSCAN called HDBSCAN [3], which has the benefit that allows varying density clusters and it is implemented with higher performance than its predecessor. In addition, HDBSCAN trades the parameter $\epsilon$ for a more intuitive one: the minimum size of a cluster.

6 Low dimensional representation

Since we have proposed 5 dimensions in the data, it becomes difficult to visualise the clustering results. For this reason, we proposed using a dimensionality reduction algorithm to transform the original data into a low dimensional representation. This allows not only to visualise the clustering results but also to evaluate and select the method that performs best.

We choose then T-SNE (T-Stochastic Neighbor Embedding) [11] as our dimensionality reduction algorithm. In contrast with PCA (Principal Component Analysis), T-SNE offers a non-linear dimensionality reduction technique which instead of focusing on creating a mathematical mapping function from the high dimensional space to the low dimensional space, it is designed specifically to visualise high-dimensional datasets through a probabilistic model that preserves the local distances of the high dimensional data. The way it works is that T-SNE models the probability distribution of neighbour around each point in the original high dimensional space as a Gaussian distribution, and in the 2-dimensional output space as a t-distribution. The goal is then to find a mapping onto the 2-dimensional space that minimises the difference between these two distributions over all points. This difference is computed using the Kullback-Leibler divergence. The heavier tails of the t-distribution in comparison to the Gaussian distribution help to spread the points more evenly in the 2-dimensional space.

T-SNE important parameter is perplexity, which intuitively lets you select between a focus of local variations between closest neighbour points (i.e. low perplexity), and a focus on global variations where more neighbour points are considered (i.e. high perplexity).
7 Results

7.1 T-SNE and clustering results comparison

Fig. 1(a) shows the dimensional reduced representation of the 5 proposed variables using T-SNE. Since T-SNE has a computational and memory complexity that is quadratic in the number of data points [11], we randomly selected a subset of 1/8 of our dataset. Another important point is that since the data points in the original 5 dimensional space are compact, it is required a high perplexity parameter in T-SNE to separate the data points into *islands of similarity*. Hence, we have set perplexity equal to 70.

The Gaussian mixture clustering was estimated with 23 components and full covariance matrices, meaning that each of the components has its own covariance matrix. The results are presented on the T-SNE representation in Fig. 1(b). For the case of HDBSCAN we have set up the minimum cluster size equal to 1500 and the minimum number of samples to 250. With these settings we are able to get 20 clusters which is close to the BIC of 23, and to the 17 different types of human mobility networks or *motifs* that represent around 90% of of the population in surveys and mobile phone datasets as reported in [10]. These results can be observed in Fig. 1(c) along with the datapoints (in white colour) that were considered noise, and in consequence, no cluster was assigned to them. After
inspecting the resulted clusters of HDBSCAN, we identified 3 pathological clusters, namely, cluster 1, 6 and 14. The reason is that these clusters captured the mobile phone users that were travelling during midnight. For these cases, we have artificially imputed their last arrival variable as the last minute of the day (i.e. 1440 min). This characteristic originated the formation of these clusters that are biased on the way we have defined particular variable last arrival. Hence, we redefined the clusters by setting these datapoints as noise, and then later reclassified them (along with the original noise datapoints) into the remaining valid clusters.

For the task of reclassifying the noise, we used K-nearest neighbours classifier, which selects the K closest clustered datapoints and casts a vote among them. The cluster of which the majority of these neighbours belongs is assigned to the noise datapoint. We set up k equal to 500 with a weighting function inversely proportional to the distance between the noise datapoint and the neighbour. This means that votes of closer neighbours count more than votes of further neighbours. We present in Fig. 1[d] the reclassification results of the noise and the final HDBSCAN clusters represented on the T-SNE vectors.

By visual comparison between Fig. 1[b] and Fig. 1[d] we can see how HDBSCAN clusters match better with the islands created by T-SNE, meaning that the clustering results of HDBSCAN are more consistent than the results by Gaussian mixture. Concluding that for the variables that we have defined in the mobile phone stay-location dataset, HDBSCAN with noise reclassification performs better than Gaussian mixture clustering.

7.2 Traveller archetypes in Singapore

To better understand the meaning of each of the resulted HDBSCAN clusters, we present in Fig. 2 the histograms of the departure times for each of the trips done by the mobile phone users of each clusters. In this way, each subfigure represents a different cluster with the following information: the percentage of total mobile phone users that belong to that cluster, the departing time histograms of the original clustered users (without considering noise), where each colour in the plot represents each of the trips (e.g. blue histogram represents the departure time histogram for the first trip during the day, green histogram represents the departure time histogram for the second trip during the day, and so on), and the distribution of number of trips for the original clustered mobile phone users on the right side as a vertical grey-scale coded plot.

Commute archetypes Clusters 11, 14, 6, 13, 15, 16 and 7 relate to commute patterns with at least a trip in the morning around 7am - 8am and a trip in the after work hours, around 6pm - 7pm. Cluster 11 is the largest cluster (23.48%) and it shows the regular commute trips without any further trip during the day. Cluster 14 (8.32%) has an additional trip in the afternoon, possibly meaning that the person before going back home, decided to do a stop somewhere else (e.g. shopping or dinner). Cluster 6 (5.51%) has also an additional trip but in this case during the morning. This cluster depicts the frequent observed behaviour in Singapore in which commuters stop for breakfast on the way to work whether it is some place around their home or around their work place. Another remark is that users in this cluster have a first departure time mean at 7:18am which is around 30 minutes earlier than the first departure time mean of cluster 11 at 7:47am. This means that people that stop for breakfast (or any other activity) before work leave their houses 30 minutes before on average. Cluster 13 (5.26%) is similar to cluster 14 with the difference that there is more than 1 extra trip after work. This is the case for people that do several activities after work (e.g. going to the gym and then going out for dinner before going back home). Cluster 15 (5.26%) has two additional trips around noon, meaning that most probably users travelled for lunch and then went back to work. It is important to note that the lunch trip is the important identifier for this cluster, so users belonging to this cluster may have in addition some extra trip(s) after work, or may not. Cluster 16 (2.87%) is characterised by having an extra trip in the morning and one after work which is the joint version between cluster 14 and 6. Finally, cluster 7 (1.84%) explains users that were out the night before and travelled back home on the early hours of the day and then performed their commute. A remark here is that users in this cluster left their house 28 minutes later in average than users in cluster 11.

Non-commute archetypes Clusters 17, 10, 12, 3 and 8 relate to non-commute travel patterns. Cluster 17 (9.36%) corresponds to users performing several trips and starting their activities around
7am. These clusters can have a broader array of examples, from the traditional housewife doing multi-activities during the day, to a taxi driver or a service delivery driver. Also, the example of a university student fits here, especially if the university has an open campus (e.g., National University of Singapore) in which the students need to travel from one lecture hall to the next lecture hall. Cluster 10 (5.15%) relates to users that have at least two activities in the afternoon and that their first departure time is around noon. Cluster 12 are users that have an afternoon or evening activity. Some of these users have their first departure time around noon with a return trip around 5pm, while others have their first departure time around 6:30pm with their return trip around 9pm. Cluster 3 relates to users that have early activities and later in the day just remain at home. Cluster 8 corresponds to users that perform only one short activity in the morning and return back home around noon.

**Special case archetypes**  Cluster 1, 9, 5 and 2 are special case archetypes. Cluster 1 (16.85%) is the second largest cluster and represents non-travellers. This means that for these mobile phone users no trip was detected during the study day. Cluster 9 (3.67%) are mobile phone users which travelling activity was detected throughout the whole day. Cluster 5 (3.15%) and 2 (1.5%) on the other hand, represents users that only one trip was detected, in the morning for the case of cluster 2, and in the afternoon/evening for cluster 5.

### 7.3 Discussion

It is interesting to analyse the case of special clusters 9, 5 and 2 because they are clusters that cannot be related very easily to common examples. Cluster 9 on one hand can be explained for instance by the trace left by mobile phones owned to logistics and delivering companies. It could be the case that two drivers from the same company use the same company mobile phone, one in the late shift and the other one in the morning shift handing it over at the change of shifts. At the same time, we should also be careful, because signals from this cluster might also come from systematic errors in the connection of mobile phone devices with the mobile phone network. The cluster percentage of 3.67% might possibly indicate the presence as well of these type of systematic errors, where the device keeps connecting to different network antennas even though the user remains at the same spot. This relates to the common problem in mobile phone data known as signal jumps [8], that despite being addressed by the mobile phone data provider with preprocessing algorithms, there could have been some instances that were not correctly filtered. For the case of clusters 5 and 2, we can think of the example of a person staying overnight at someone else’s place, and during the day travelling back to his/her place, hence performing just one trip during the day. Conversely to the counter example in cluster 9, we can also have some cases in this cluster that reflect some other situations. For instance, it could be that for some part of the day the mobile phone network was not able to detect the device, or the user just switched off his/her phone.

For those reasons it is important to have a deeper analysis of the clusters obtained, especially when the data has been preprocessed by the mobile phone operator or by any other third party. Hence, the approach presented in this paper can serve not only as a tool to uncover the different archetypes of travellers in a city, but also as a tool to detect possible systematic errors in the ways the mobile phone data is processed and interpreted. This errors being possibly originated by the mobile phone network, some human behaviour attitudes (e.g., switching off the phone) or other user-related circumstances (e.g., phone out of battery).

### 8 Conclusion

In this paper we presented a methodology to uncover the different archetypes of urban travellers in a city based only on their mobile phone location data. In specific, we tested our methodology with Singapore data and uncovered 16 different archetypes of urban travellers. The methodology consists on defining 5 variables that relate to travel behaviour and that can be extracted from mobile phone stay-location data. Then we tested two of the most relevant clustering algorithms on the dimensional reduced space of T-SNE. For the case of HDBSCAN we reclassified the unclustered points using K-nearest neighbours. After comparing both clustering results on the T-SNE plane, we concluded that HDBSCAN outperformed Gaussian mixture clustering for the case of Singapore data.
Other important insight that we have gained, was that from the resulted clusters some of them might possibly resemble biases in the mobile phone data due to systematical network errors or user-related situations. Often times, these biases are difficult to detect, specially when the data provider has processed the data \textit{a priori}. However, with a further inspection at the resulted clusters, the proposed methodology can also served as a screening tool to detect these type of major biases in mobile phone data.
On a final note, the relevance of being able to segment the different types of travellers without any socio-demographic information is that it constitutes one of the first steps to being able to improve travel demand models based on Big Data when privacy is a concern. On the next step, we will be designing a generative model for each of the clusters found in order to construct synthetic populations for large-scale microsimulations (e.g. MATSim [6]).

Acknowledgments

This research has been conducted at the Singapore-ETH Centre for Global Environmental Sustainability (SEC), co-funded by the Singapore National Research Foundation (NRF) and ETH Zurich.

References


