A Work Project, presented as part of the requirements for the Award of a Master's degree in Business Analytics from the Nova School of Business and Economics.

CLUSTERING OF TRANSACTIONAL TRAFFIC DATA TO ANALYSE MOBILITY PATTERNS

FELIX PAGEL

Work project carried out under the supervision of:

Patrícia Xufre

17-12-2021

Abstract:

This work develops a framework to analyse the development of mobility patterns over time. Based on flow data of individual vehicles on Portuguese motorways the proposed methodology defines a set of features that describe each individual's movement characteristics. K-means was identified as most suitable algorithm to cluster the set of features allowing for a meaningful interpretation of mobility patterns. The analysis showed a conflict of objectives between cluster quality, and the interpretability of the defined features. Therefore, for an optimal outcome of the analysis the number of clusters should be manually aligned with the goal of the analysis.

Keywords:

Business Analytics, Mobility pattern analysis, Clustering traffic data, Cluster evaluation

This work used resources provided by Brisa – Autoestradas de Portugal, S.A. which are protected under a non-disclosure agreement and infrastructures of the Nova SBE Data Science Knowledge Centre.

Table of Contents

1. Introduction

The Covid-19 pandemic made ways of analysing people's mobility behaviour become a pressing need and gave the entire section a push. Frameworks such as the Covid-19 mobility data aggregator, a scraper of Google, Apple, Waze, and TomTom emerged. However, the means of analysing this kind of data and deriving meaningful conclusions from it could not keep up with this pace. The thesis taps into this field, proposing a framework for analysing a certain type of mobility data and draws conclusions on the effect of the pandemic on people's mobility behaviour.

This thesis was conducted within a project in cooperation with the Portuguese motorway company Brisa. Brisa operates dozens of motorways throughout Portugal, of which 11 are forming the ViaVerde toll collection system. Traffic data for those motorways, aggregated on a quarterly level, was provided by Brisa with the objective of developing a medium-term traffic prediction model. Alongside this high-level traffic prediction model, this thesis investigates traffic on a more granular basis. For this purpose, another dataset was provided by Brisa, covering all automatically recorded toll transactions on all 11 ViaVerde motorways from January 2019 until September 2020.

The provided database records single transactions, each of which denotes a vehicle passing through one or maximum two (closed or open system) toll stations. Thus, each line item in the database defines a trip, meaning a trajectory of a vehicle from point A to point B, both points are indicated as entry and exit points. Consequently, a round trip would be a reverse trajectory from point B to point A within a reasonable amount of time.

There has been considerable amount of work on the topic of analysing mobility. However, these are mostly focused on making sense of an unordered mix of GPS references. Instead, this work focuses on large scale mobility analysis based on transactional data.

2. Literature review

2.1 Mobility Data Structure

After clarifying the basic terms used within the following work and the broad objective of this research, the following section will provide an overview of the existing literature on the mining of traffic data. Our perspective will not only be limited on traffic, but also to look into work trying to derive insight or value from mobility data in the broader sense to check for possible transfers of methodologies.

There are two major frameworks to structure and treat mobility data, called scikti-mobility (L. Pappalardo et al. 2019) and MovingPandas (Anita Graser 2019). While both are based on the common Python library Pandas and are focused on the handling of trajectories, scikit-mobility is providing additional modelling and measurement functionalities, whereas MovingPandas is focused on geospatial analysis with an integration of the extensive open-source project GeoPandas. The measurement and modelling functionalities of scikit-mobility make it a naturally more interesting target for the further analysis, as the trajectories of the available dataset are fixed along the limited number of motorways, and thus are not useful for a more detailed geospatial analysis.

Scikit-mobility distinguishes between two forms of mobility data, trajectory data which is handled by a TrajDataFrame and flow data which is handled by a FlowDataFrame. Trajectories are the trace of the movement of an object, meaning every row is a record of the location of the object at a specific point in time. Flows on the other hand side are characterized by an origin-destination matrix, of which both points have already been discretized from an original set of coordinates. Trajectories describe movements on an individual level and are naturally more applicable in the case of GPS data where there is a

proper trace of the individual object. Flow data on the other hand is already aggregated and as such not preserving characteristics on an individual level.

Comparing these data structures to the available dataset there is a natural fit into the flow category of mobility data. After all, the dataset is not providing any real trajectories but mere entry and exit points of a vehicle's trip. Even so, the dataset is still providing information on an individual level, combined with a temporal reference of each flow, opposed to the FlowDataFrame structure.

We can therefore conclude that the data can be organized as aggregated flow data between entry and exit points, or as constructed trajectory data where entry and exit information will be brought from a wide to a long format. However, these trajectories would not resemble a classical trace of a movement but rather a jumping around between the motorways' entry and exit points. To fit these unique characteristics of the dataset, it shall be organized in a mix between the two proposed data structures.

L. Pappalardo et al. (2019) review in their library scikit-mobility a number of mobility measures at both the individual and collective levels of human mobility. These mobility measures will be discussed in greater length later, but it is important to mention that all measures chosen by the authors for their library implementation are based on trajectory data only. Whether a simple conversion of the available dataset to trajectory data makes it eligible for those measures needs to be checked.

2.2 Mobility Data Visualization

The difficulty in the visualization of mobility data lies in the combination of its time variant component and its spatial component. Landesberger et al. (2016) state that simultaneous spatial and temporal data simplification or visualization is still problematic. In their work on mobility graphs, a mean for visual analysis of mass mobility data, they identify the spatial

situations and the temporal variations in the data as main characteristics to define mobility dynamics.

To visually analyse those dynamics effectively, the authors introduce a procedure to simplify both dimensions. Spatial data can be simplified by filtering a number of locations that appear most relevant to preserve the spatial structure of the data. Alternatively, only flows above a certain magnitude threshold might be considered. As the latter approach does significantly reduce the number of nodes in an evenly connected graph of flows, the first approach requires an automatic procedure to avoid a manual selection of location, which possibly hides important information.

For those reasons the authors continue with aggregating the spatial locations to reduce the number of nodes in the graph. They follow Kisilevich et al. (2010) who have described clustering approaches for geospatial data. Accordingly, after an initial reduction of locations into regions, the flows between locations are aggregated to flows between regions. 'Density-based Spatial clustering of applications with noise' (DBScan) is identified as the most suitable algorithm for the reduction in the number of locations into regions (Martin Ester et al. 1996).

The main reasons for the authors' choice are the computational speed, the adaptability of the algorithm to unconventional cluster shapes and the non-requirement of pre-setting the number of clusters. The authors slightly modify the DBScan algorithm to also account for flow magnitudes between locations as a deciding factor for the clustering. Hereby, closely located data with very little flow magnitude are treated as outliers.

There is the question whether this algorithm should be applied to the entire dataset, spanning all time variants, or should be processing the different spatial situations at different time variants separately. The authors state that region correspondence is important for the comparison in the second step (clustering of time variants), and therefore the algorithm is

applied on the entire dataset, a so called supergraph spanning all time steps. The described approach results in a reduction of 606 locations to 42 regions, resulting in a reduction of 21 thousand flow links down to 396.

In a second step the authors reconstruct based on the reduced graph the spatial situations at different time steps, which are then clustered for similarity. The similarity is defined by a feature vector describing the magnitude of all flow links. K-means is identified as an appropriate clustering algorithm as the similarity between feature vectors can be characterized by the Euclidean distance (Han, Kamber, and Pei 2012). At this point the initial reduction of locations plays an important role as the feature vector on the initial dataset would have an extremely high dimensionality (21 thousand), the problems of which are explored by H. Kriegel, P. Kröger, and A. Zimek (2009) and Beyer et al. (1999), namely an unstable and unreliable set of clusters.

In the end the authors identify seven clusters that represent different times of the day and their corresponding spatial situations. Visualising these along with a grid showing the different hours per day for a multitude of days provides great insight into the shift of the clusters across different days of the week.

Overall, there is a strong parallel between the proposed spatial reduction and the basic setup of the available dataset. As already discussed above, the available dataset for this thesis does not provide real trajectories, meaning pathways of individuals, but rather a discrete set of entry and exit locations. Since the subject of the analysis are motorways, these pre-defined locations are not scattered densely but often span several dozen kilometres. Hence, the dataset is already providing a naturally reduced graph of 94 nodes. This does still seem like a large number of connections compared to the 42 locations derived by the authors; thus attention would need to be paid to the dimensionality of the feature vectors. However, comparing the

size of the analysed area, greater London, and Portugal we conclude that the toll system is an opportunity to avoid the need for initial spatial clustering.

Nevertheless, the question remains whether the proposed analysis would help with the goal of identifying mobility patterns as described in the introduction. The authors have shown an example timeframe of one week, while the available dataset's time period to examine spans one and a half years. There is the question if the visual analysis could be done on a higher level, clustering spatial situations at different days, opposed to the hourly level introduced by the authors. There is no similar literature work that transfers the authors visual analysis on such a more granular level.

The other concern is about the data size, as the mobility data of an entire country for an entire one and a half years is naturally more extensive than for a city. Ultimately, the utility of the described work for the available dataset comes down to the definition of what constitutes a mobility pattern.

2.3 Clustering temporal spatial data

Asadi and Regan (2019) work with traffic surveillance data to develop an approach identifying temporal and spatial clusters in such data. The developed approach uses deep embedded clustering models and is found to identify meaningful patterns in traffic data. The traffic surveillance data that the authors use to develop their models is a matrix of the locations where the sensors are placed, the time stamps of the recordings and the traffic features that the sensors are able to record, such as the speed of the vehicles.

The major difference of the authors dataset and the available dataset in this thesis are the vehicle identifiers that allow to match records belonging to the same vehicle. Asadi and Regan (2019) do not have this identifying feature available and as such their dataset

resembles neither the trajectory dataset structure nor the flow dataset structure introduced in chapter 2.1.

Similar to Landesberger et al. (2016) the authors identify the need for data reduction of the time component, and divide their time series with a sliding window into different traffic states at different points in time. While Landesberger et al. (2016) restricted their clustering of temporal structures to a latent feature representation, Asadi and Regan (2019) also use a non-linear distance function as a similarity measure and compare both approaches. In the actual clustering Asadi and Regan (2019) proceed very different, as they identify a lack of analysis on deep embedded clustering models on spatio-temporal data they set up a neural network to assign each time series window of one hour to a set of initial clusters predefined with k-mean.

The authors' used dataset is significantly different and also their work's objective is more focused on determining the state of a street at a certain point in time, as opposed to analysing the vehicles' behaviour. Therefore, the authors' developed model is of limited utility for this thesis. However, we note the authors' procedure on validating the latent feature representation of a single sensor's traffic states with the technique of t-distributed stochastic neighbourhood embedding (t-SNE) introduced by Maaten and Hinton (2008). The approach is to compare the two-dimensional feature representation at different times per day, where a positive outcome would assign traffic states at similar times of the day into similar regions in the two-dimensional pane.

This approach could be transferred as an additional way of validating the produced clusters on the available dataset. Also, the authors analyse in their work the size of the temporal clusters built with the deep embedded approach. The frequency plot shows a large number of clusters with very few assigned data points. This might be attributable to the large number of clusters (70), but also shows that the deep embedded approach is not immune to such shortfalls. Instead, the data has to be analysed for outliers beforehand.

Asma Belhadi et al. (2020) provide a more general review of available space-time series clustering methods. After examining the theory of the different algorithms, the authors apply each category to two urban traffic datasets. The authors state that space-time series are one of the most powerful representations in many domains and highlight the importance of data mining techniques in this field.

Broadly, the authors distinguish between hierarchical clustering, pure partitioning and overlapping partitioning (Asma Belhadi et al. 2020). According to the authors, pure partitioning algorithms are most suitable for large scale space-time series data as they are much faster than the other categories. The downside of this category of algorithms is the need for parameter setting. Similar to the dataset in the work of Asadi and Regan (2019) the authors' data used in their case study is a collection of sensor recordings measuring the passing vehicles at specific locations. Therefore, it is also missing the unique identifier, what is limiting its comparability with the available dataset for this work. Nevertheless, the procedure is still revealing interesting aspects.

The case study is meant to compare the clusters quality build by different categories of algorithms. The authors evaluate the quality of the clusters in two ways, one is the error sum of squares and the other a set of artificially created labels, that serve as a reference cluster to be recognized. Specifically, the labels are weekday, Saturday, and Sunday. These are meant to be distinguished by the clustering algorithms setting the parameters to three clusters. The usage of labels is a promising way of validating the clusters in a more sophisticated way and might be applied to the available dataset, for example in the form of heavy and light vehicles, or in the same format as the authors.

Overall, Asma Belhadi et al. (2020) conclude that the domain of space-time series clustering still requires further exploration and progress, meaning the existing approaches did not yielding the expected results. The authors identify a number of challenges which are

connected to the poor results. Runtime improvement of the clustering algorithms is one of these challenges and presumably highly relevant for the work in this paper, as the mere size of the data increases the requirement of a quick and efficient processing. Correlation between space-time series data is another issue the authors suggest developing further. Finally, the authors conclude that one of the central challenges of yielding high cluster quality is to adapt the existing advanced clustering methods to the specific scenarios of space-time series data. Finding common ground between those special scenarios to develop specialized methods that work for all, has yet to be examined.

W. Weijermars and E. van Berkum (2005) apply a hierarchical clustering model with the goal of identifying historical traffic patterns. Like most of the examined research, the data analysed by the authors is sensor data capturing the speed and the flow of traffic in a 15-minute interval. Given that the data is collected from only one sensor location, it is not spatial temporal but rather a time series of traffic features at one location. This reduction of dimension eases some of the problems previously examined in other works.

In their pre-processing procedure of the dataset, the authors also remove days that had shown strong road congestion, following the logic that these traffics flows do not represent travel demand properly but rather add noise. This is an important aspect to consider in the feature engineering for the clustering in this thesis. Ideally, the traffic situations across time should not be reflected in the derived features, following the same logic as the authors that road congestions are not to be considered. As such, the duration of a trip, for example, would not be an ideal feature as it not only captures the distance travelled but also the traffic situation. Moreover, the time needed for a trip is dependent on the vehicle class, as larger vehicles typically travel slower than smaller vehicle classes.

W. Weijermars and E. van Berkum (2005) see the biggest advantage of hierarchical clustering methods, a category already discussed in previous works, in the self-adaption of the model's

number of clusters. For this reason, they only implement the analysis with Ward's clustering procedure (J. H. Ward 1963). Since the authors are clustering classical time series, the results and the algorithm choice is not transferable to the own dataset.

However, it is noteworthy that the authors determine a substantial difference in cluster quality by pre-classifying the dataset into working days and non-working days. Non-working days are Saturdays, Sundays, and holidays. Due to the large variation in flow profiles between those two groups of days, the clusters without pre-classification also show high variation. Filtering out the largest differences in the data in advance might be an important method to improve cluster quality on the available dataset. After all, the algorithm shall detect smaller differences and patterns in the data and not capture distinctions that are already known upfront.

As this effect made a difference already in the relatively small dataset of the authors, it should become even more apparent when performing clustering on a large dataset. The ways in which such a pre-classification could take place needs to be determined but separating the data by vehicle classes could be a natural division for the start. Moreover, there might be a need to divide the light vehicle class further, as removing heavy vehicles only affects a small fraction of the overall traffic transactions.

The authors' procedure of dividing the data could be a potential fit for classifying the light vehicle data further, before proceeding with the clustering. With respect to the working days the authors conclude that there are four distinguishable types: Mondays, core-weekdays, Fridays and days within vacation period.

Heber Hernández et al. (2021) examine in their work ways of estimating traffic data with a linear based radial basis function. As part of the estimation process, the authors also cluster their data into partitions that are then further used for the basis function. For this clustering, the authors use the already discussed k-means algorithm, as they see a good fit with the

multidimensional continuous data while being highly scalable. The elbow plot serves as a validation method to find the best number of clusters. In a third analysis step, key statistics of the clusters are compared, but most important is the plot of the geolocation of each data point and the associated cluster represented by colour. The authors conclude that the three identified clusters are a meaningful representation of the real importance of each road segment. They argue that this clustering based on real traffic is yielding a more objective and reliable way of assessment than a manual assessment based on factors such as adjacency and connectivity to other regions.

Although the application of k-means to identify different levels of traffic increase confidence in the algorithms applicability, the results' transferability to our own dataset is limited. After all, there is no time component in the authors' dataset, but the features of each road intersection are an already aggregated summary statistic on an annual level. The work does also not contribute to the spere of spatial analysis, as the coordinates of the road intersections are not considered in the clustering, and neither are their connectivity or the flow between them.

Gerhard Münz, Sa Li, and G. Carle (2007) apply k-means clustering to detect anomalies in web traffic flows. Despite the fundamental differences in the examined domains, there are similarities to conventional traffic flow data, namely the source IP address as the origin and the destination IP address as the destination. The derived features include for example the number of packets sent from and to the given port, or the number of bytes sent from and to the given port in the considered time interval. The authors conclude that the proposed procedure work well and that the k-means centroids can be used to detect anomalous data points in live streamed data.

2.4 Conclusion of literature

We conclude that there have been numerous attempts to cluster traffic data and that the fields of application of such clustering approaches are growing. However, we determine a lack of research on spatial temporal datasets that allow to draw conclusions on an individual level. This might be due to the difficulty of obtaining large scale traffic data with a unique identifier, connected to both the required infrastructure for the data collection but also the issue of privacy.

Most research analysed in this literature review is based on sensor data that capture traffic states at certain points in time across a number of locations. The other major part of research is devoted to analysing trajectory data based on GPS samples. Due to the relevance of privacy but also limitations in scale, these datasets usually come in much smaller sizes, making the results hard to roll out in a large scale context. Also, these kinds of works are more focused on spatial analysis and the finding of a meaningful representation of a large amount of trajectories, opposed to a motorway structure that is a fixed given and cannot be aggregated into pathways.

We conclude that this thesis has a unique dataset available that provides a mix of flow and trajectory data in a large scale. The application of k-means to cluster traffic data stood out in the analysis and will be transferred to the available dataset. Finally, this work will try to preserve the unique attributes of the dataset, meaning its possibility of identifying each vehicle.

3. Approach

3.1 General procedure

As described in the introduction, the goal of the analysis is to develop a framework that helps with identifying changes of mobility patterns of motorway users based on the Covid-19 pandemic. Depending on the definition of mobility patterns, the analysis can be conducted in different ways. Generally, there are two perspectives that can guide the analysis, a macro perspective looking at the road network and traffic flows overall, and a micro perspective looking at individuals using the motorway network across time. Under the first perspective, a mobility pattern is defined as a certain amount of traffic that is flowing between a defined set of network nodes at a certain time interval. Under the latter perspective, a mobility pattern is defined as a typical road usage pattern that an individual vehicle shows within a certain time interval. Such a road usage pattern is naturally related to a metric of road usage such as travel time or kilometres travelled, measured along common temporal characteristics such as weekdays, time of the day or week of the month.

The literature review revealed that most research is focused on the macro perspective, presumably due to a lack of large-scale mobility datasets on an individual level. Flow maps were identified as underlying data structure to handle this kind of macro analysis, whereas the micro analysis typically deals with trajectory data. It was concluded that the available dataset is a unique opportunity to perform a large-scale mobility analysis on an individual, micro perspective, based on the vehicle identifiers.

There are different hypotheses on behavioural changes also affecting the mobility patterns of individuals, based on common rational. For example, the possibility of remote working should be a driving factor to reduce the number of people commuting to work. These hypotheses can be used to make sense of the identified cluster changes to check for potential

pitfalls. A mobility pattern in the context of this thesis is equivalent to a typical road usage behaviour, that is applicable to a group of individuals. Finding those groups with similar patterns is a classical unsupervised learning problem that tries to structure a set of data into a set of clusters showing maximum similarity within each cluster.

The general approach will be to deploy a clustering algorithm to an initial subset of data to calibrate the model, to then predict the cluster belonging for every month of the remaining dataset. Finally, the resulting clusters' sizes can be compared, ideally there should also be centroids that allow for an easy explanation of a persona representing each cluster. The first step of calibrating the clustering model on a subset of data is an integral part of this analysis and is a cyclic approach of trying different algorithms and feature setups to derive the best cluster structure.

The clustering models shall be evaluated along three quality measures, the inertia or similar error measure, the silhouette score, and a visual inspection. The latter is based on a t-SNE dimensionality reduction into the two-dimensional space, and a subsequent comparison of the plotted reduced dataset with the labels obtained by the respective clustering algorithm. This procedure is similar to the evaluation procedure applied by Maaten and Hinton (2008). t-SNE is a good proxy in detecting cluster structure due to its flexibility to fit non-linear shapes. A detailed overview of the final pipeline setup for the analysis is provided in the chapter 3.3.

Eventually we note that there is a conflict of objectives of high cluster quality measured by the described metrics, and high interpretability of the cluster centroids. The latter is determined by the set of chosen features, which ultimately strongly affects the cluster quality.

3.2 Feature engineering

As outlined in chapter 3.1, interpretability of the features is important for the analysis to be insightful. After all, the identified mobility patterns need to be explainable to be useful in the end. As there are no quantitative measures of the features' interpretability, the feature engineering follows a subjective approach by defining features that appear to characterise road usage behaviour by common sense. In the feature engineering there is a trade-off between the interpretability and the amount of detail captured from the data. A small number of features is generally easier to interpret, while a large number of features is able to represent the underlying data even better.

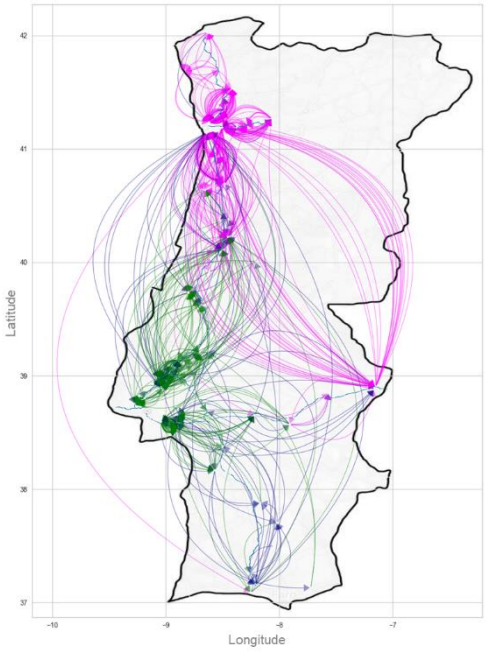

Figure 1: Sample transactions' trajectories clustered based on coordinates

Due to the size of the available dataset (>200M transactions) the initial feature considerations are performed on a sample of one hundred thousand transactions from March 2019. Figure 1 provides an example overview of a subset of transactions clustered into three regions indicated by colour, to demonstrate the geospatial characteristics of the entry and exit id attributes. As we can see in the chart, the single trips can span several motorways in the network, which spans the entire country of Portugal.

A fundamental decision in the feature engineering is how to represent the coordinates of the trajectories as well as the time attributes to be able to cluster the dataset in a meaningful way. As discussed earlier, the analysis shall be conducted on an individual level, based on the vehicle identifiers. Naturally, this suggests aggregating the data in a first step to a dataset where each instance corresponds to an individual vehicle with the attributes describing the spatial-temporal characteristics of the transactions belonging to that vehicle. However, this

approach of pre-aggregating the transactions comes with a loss of information. After all, summary statistics such as the sum, mean, median or standard deviation might preserve the central information, but cover up details of the underlying distribution of each attribute. Even so, the pre-processing approach is considered appropriate, because working with the raw transactions is dominated by the attributes time and space.

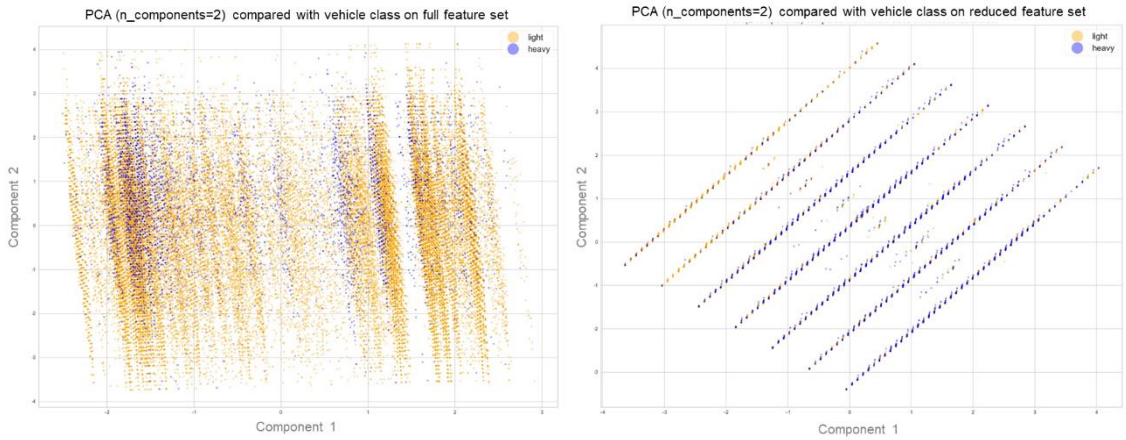

Figure 2: Left - PCA results including coordinates, Right - PCA results excluding coordinates

Visible in figure 2, showing the results of a principal component analysis performed on the raw transactions (time stamps represented by temporal features), filtering the spatial information of the entry and exit points provides a much clearer cluster structure in the principal components. The structure resembles the seven days of the week with 24 hours each, whereas on the left chart no such clear structure is visible. It seems that the information of the 94 toll station coordinates is adding noise that cannot be properly processed by the principal component analysis. After all, it is already known that the transactions' coordinates are bound by certain locations and there is no need to define spatial clusters since it is not possible to change the motorway structure.

A number of clustering algorithms have been applied on the dataset, though several ones resulted in computations too heavy for the scale of the analysis. For example, both the Spectral Clustering and the AffinityPropagation require memory allocation above 20 GiB

already at a reduced dataset size of 60 thousand vehicles. For this reason, k-means and DBScan were the chosen algorithms to be applied in large scale and compared with each other for suitability. Despite its advantage of automatically setting the n_cluster parameter, DBScan was dropped from the analysis because of very poor cluster quality, both visually and per metric. In particular, the algorithm did not deliver evenly sized clusters but only managed to identify very small micro clusters, presumably characterized by extreme values, while often assigning a large number of vehicles as outliers. For this reason, k-means algorithm was used for the final pipeline setup.

For each month, data of the beginning 15 days was sampled for the analysis. It was decided to exclude vehicles that have less than ten trips within this 15-day time period. Even though this distorts the results slightly as trips are being excluded on a rather arbitrary basis, it helps to reduce the size of the dataset to run the algorithm and to evaluate performance. One possible way of mitigating the loss of information would be to include the entire month, this way there would be more vehicles with ten or more transactions in the dataset. However, this would counteract the goal of data reduction.

After removing features that are highly correlated, such as the number of trips on a time of the day, and the sum of trip times on the time of the day, results appeared to show clear cluster structure. Figure 3 shows a comparison of two feature sets, out of many tested combinations of features, to illustrate the feature engineering process. The left side of figure 3 shows the evaluation of cluster quality of a feature set based on the sum of trip durations at different times of the day (night, morning, midday, afternoon, evening) without differentiation between weekdays and weekends.

In the t-SNE two-dimensional representation of the data, there are clearly distinguishable clusters visible, that evolve around a central, more jammed area. Of these suspected clusters,

k-means algorithm manages to identify three main groups that are nearby each other. Those results seem to be reflected in the metrics below with a silhouette score above 0.7, indicating a strong cluster structure. Partially, this high silhouette score is driven by extreme clusters that only have a few vehicles allocated but are very separable from the rest. This is reflected in the number of clusters of 16, required to match the t-SNE cluster structure shown on the left side of figure 3. Additionally, the feature set on the left side is not very interpretable, because the sum of trip times misses important information on how long a typical trip takes, or context on the difference between weekdays and weekends.

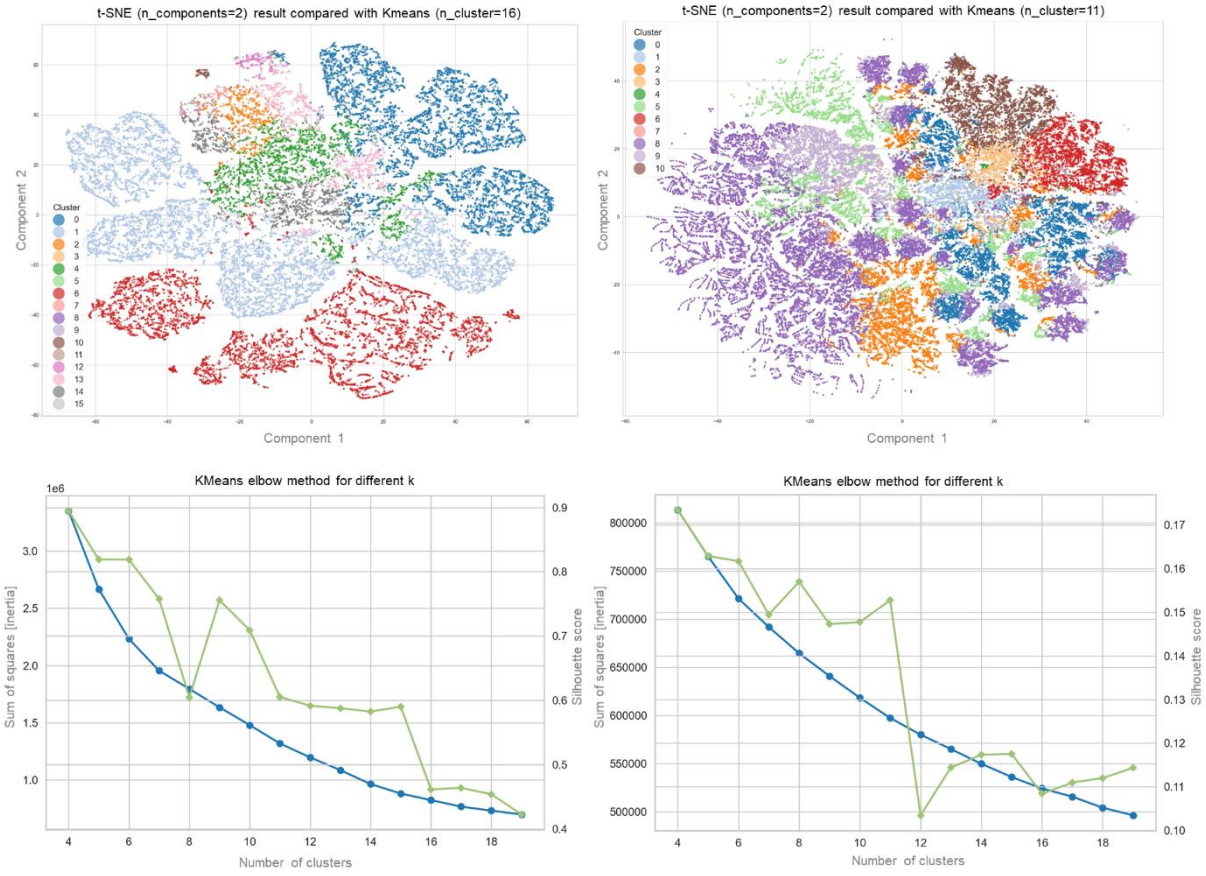

Figure 3: Left – t-SNE reduction based on sum of trip time features without weekday/weekend differentiation and corresponding elbow plot, Right – t-SNE result based on number of trips features with weekday/weekend differentiation and corresponding elbow plot

The right side of figure 3 shows the evaluation of a feature set providing more detail and focusing more on interpretability. Specifically, instead of the sum of trip durations, it uses the number of trip per week at different times of the day (night, morning, midday, afternoon,

evening) and also distinguishes between weekdays and weekend. This differentiation gives important context for the interpretation of the clusters, but as we can see in the chart, deteriorates the cluster quality in all categories. With a very low silhouette score and no clearly distinguishable cluster structures this feature set provides more insight into mobility behaviour but lacks clearly distinguishable patterns.

This is one of the central conflict lines revealed in the feature engineering process, the conflict of objectives between interpretability, the capturing of detail, and the quality of the clusters. In this context the findings also contradict what W. Weijermars and E. van Berkum (2005) state about the difference in cluster quality between working days and non-working days. Similar to the authors, distinguishing between weekends and weekday has a considerable effect on cluster quality, even though in the case of the thesis they are not treated separately but merely passed on as an additional feature.

After discussions with the client Brisa the interpretability of the features was prioritised over the quality of clusters. Therefore, the feature set corresponding to the right side of figure 3 was selected for the final analysis, the detailed set of features is visible in the appendix. Although the interpretability was prioritized for the feature selection, choosing the number of clusters still requires thoughtful balance between the cluster quality, the capturing of detail and the distinguishability between the individual clusters.

3.3 Model pipeline

Figure 4 shows the way the overall analysis is conducted. The entire dataset spans from January 2019 to September 2020. The Covid-19 pandemic had its first effects on public life in the second half of March 2020. Thus, the first three months of 2019 will serve as a baseline to calibrate the clustering model. The first quarter of the year appears to be a good baseline, since it is not distorted by major public holidays and holiday seasons.

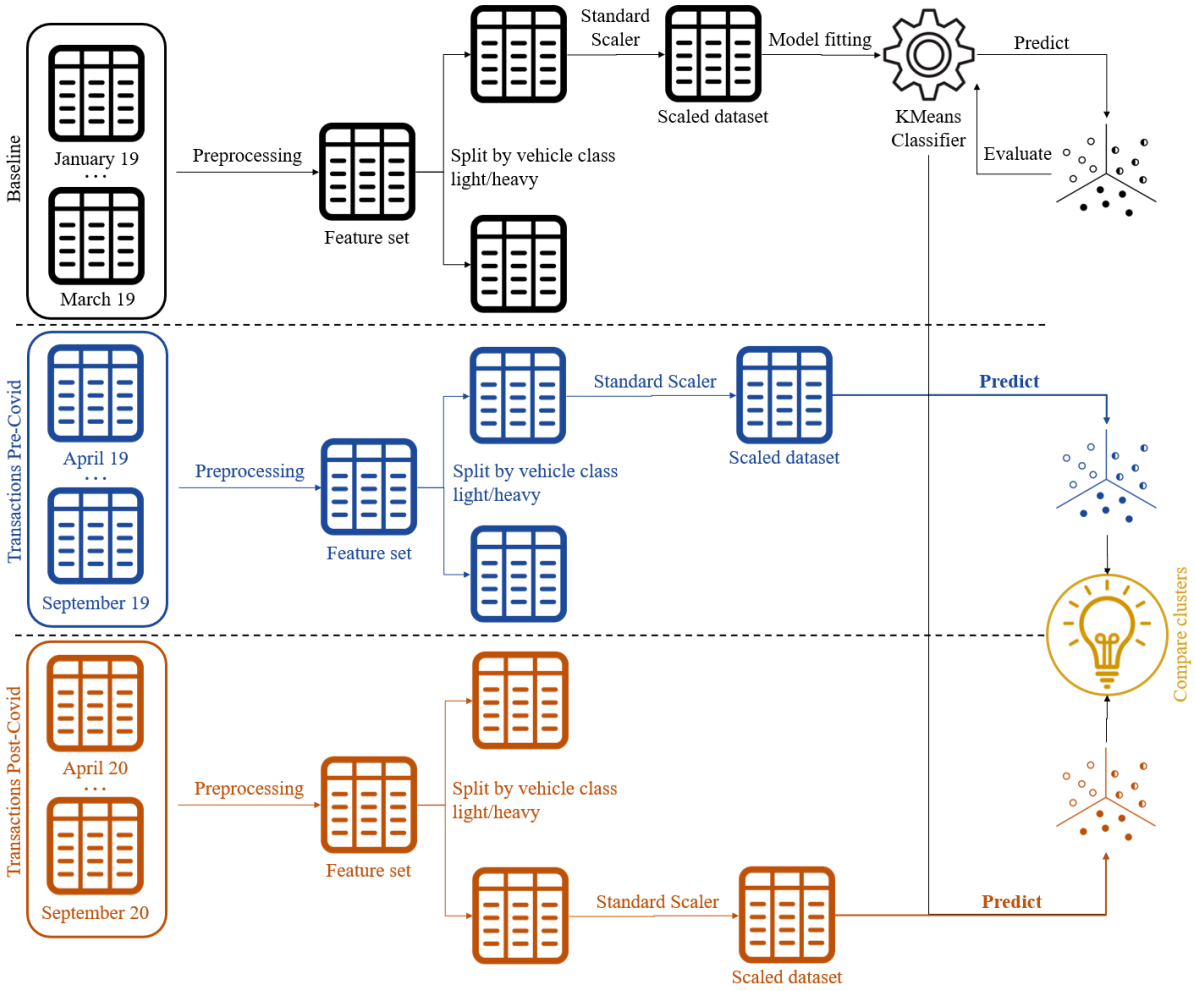

Figure 4: Pipeline setup for the analysis

In a second step, the calibrated model is used to predict the cluster allocation of the remaining data. This excludes the time of October 2019 to March 2020, which is not required for the analysis. The third step is then to compare the cluster sizes across the different months. Particular attention will be paid to the year-on-year changes, as they compare the pre and post Covid-19 behaviour. Lastly, the obtained clusters' centroids must be interpreted, to make a final statement on the change of mobility patterns. This procedure will be conducted for both light and heavy vehicles separately.

4. Modelling results

4.1 Cluster interpretation

This chapter briefly discusses the centroids obtained in the model training. Each centroid represents a persona of road user (in the following also referred to as vehicle), that shows a characteristic behaviour along the defined attributes. Consequently, each centroid represents a mobility pattern, of which the development over time will be examined in the subsequent chapter. Both chapters will focus on the results of the light vehicle traffic analysis, due to page count constraints. However, a similar analysis has been conducted for the heavy vehicle traffic results. The appendix shows the resulting clusters' centroids based on the k-means algorithm with the number of clusters set to 11, based on the right elbow plot in figure 3 with a spike of the silhouette score at 11.

All but one clusters though seem to be meaningful with more than a thousand assigned vehicles. To sum up the main characteristics of each cluster:

Cluster	Interpretation
1	Occasional users only on weekdays
2	Commuting in mornings and evenings during the week
3	Occasional user on mornings/afternoons/evenings on weeka and weekends
4	Night trips both on weekdays and weekends of which several round trips
5	Heavy users commuting but at different times, not strictly same times of day
7	Strictly commuting in the morning and afternoon but no weekend usage
8	Commuting mornings and middays but no weekend usage
9	Occasional users on both weekends and weekdays
10	Occasional users only on weekdays
11	Occasional users on both weekends and weekdays

Table 1

Comparing the initial size of the clusters we determine a relatively even distribution with no single cluster accounting for the great majority of vehicles. Only cluster stands out with only

40 vehicles assigned. As the cluster size is so small relative to the others, its informative value is limited. This cluster is characterized by very large values across all attributes, particularly the night and morning trips. Hence cluster 6 seems to capture mostly commercial light vehicle traffic, such as Uber drivers for example.

As we can see comparing the interpretations in table 1, optimising the k parameter for cluster quality does not guarantee good distinguishability between clusters. Instead, there are several clusters with little distinction, especially in the group of occasional users. Specifically, it is hard to find a meaningful difference between cluster 9 and 11, as well as 1 and 10. The differences between those clusters are only nuances in some features' values and appear more meaningful to be clustered together. Reducing the number of clusters has shown to improve the meaningfulness of the clusters, although the extreme cluster 6 persists. Therefore, it is recommended to align the number of clusters with the intended level of detail for the analysis. The following chapter concludes the analysis based on the 11 clusters. Due to the similarity in the mentioned clusters, one would expect the same clusters to merge when lowering the parameter k. However, the result for nine clusters is not that clear, and slight redundancies in the interpretations persist.

4.2 Mobility pattern development

Figure 5 shows the size development of the identified clusters both in absolute terms and relative to the total number of vehicles. In absolute terms there is a large contraction in the number of vehicles in April 2020, which logically follows the Covid-19 related restrictions of public and private life. Since these restrictions were mostly mobility related, the development is not surprising. Looking at relative terms it becomes clear that the identified clusters remain relatively stable in their size, compared to the absolute terms. There are contrary developments, for example cluster 10 seems to grow in relative size during the pandemic months, while cluster 11 seems to contract.

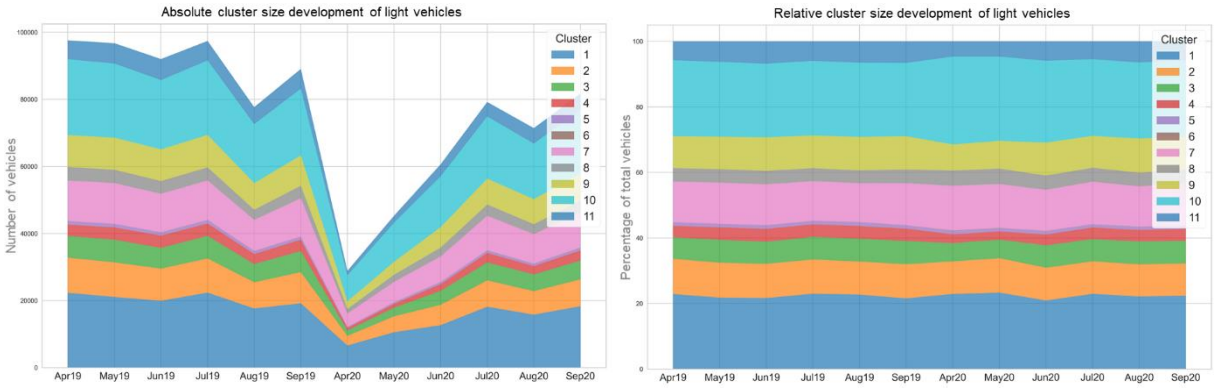

Figure 5: Left - cluster size development absolute, right - cluster size development relative to total vehicles

Figure 5 adds more detail to the overall development with two heatmaps visualising the change of the clusters. The month-on-month difference between the relative cluster sizes facilitates the comparison between the clusters and their response to the pandemic situation.

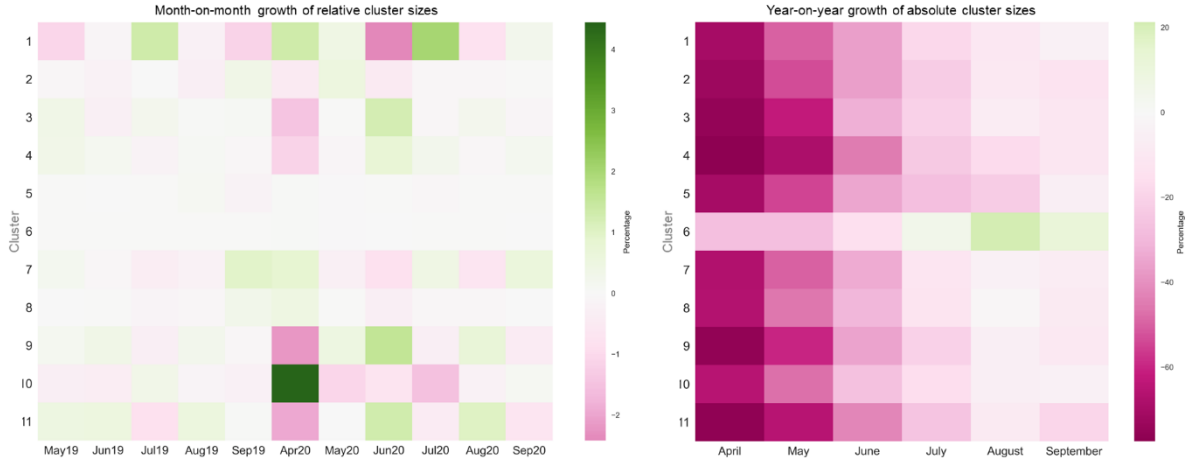

Figure 6: Left - monthly changes of relative cluster sizes, right - year on year cluster size growth

We determine that the cluster of occasional users on weekdays (10) grew strongly, but with subsequent declines in the middle of 2020. A similar picture can be determined for the other cluster of occasional users only on weekdays (1), although the development over time appears to be more volatile. This makes sense, because clusters with similar characteristics and similar persona descriptions are expected to develop similarly. Both cluster 3 and 4 contracted strongly during the peak of the pandemic but rebounded again in June with little change afterwards. This development is also explainable, as the imposed lockdowns restricted traffic

at night-time and social activities that typically take place in the evenings and at night, the corresponding cluster of strong night traffic is expected to decrease as well. Cluster 3 accounts for occasional users on both weekday and weekends, what was also affected by weekend restrictions.

There appears to be little change in cluster 5, which are heavy users commuting at different times of the day. This group would have been expected to contract; however, it remains stable even during the pandemic months. There is no noticeable change in cluster 6, as there are always a number of vehicles with those extreme usage patterns. Clusters 9 and 11 behave similarly to the night and occasional weekend users, with a strong decline in April 2020 and a rebound in June 2020, but with also fallbacks in the subsequent months. This behaviour can be explained logically as occasional users on both weekends and weekdays are more affected by weekend restrictions. Overall, the almost identical development of several clusters reaffirm that the number of clusters should be reduced for an optimal analysis outcome.

4.3 Discussion of results

The feature engineering was done in a subjective way and thus lacks evidence of an optimal outcome. Even though the objective of interpretability cannot be measured in any target metric a more systematic approach could be developed in which different feature combinations are evaluated across the entire pipeline. Such a bottom-up approach would be expected to provide more clarity on the trade-off between interpretability and cluster quality, and hence inform a better decision on feature selection.

As the pre-processing is computationally intensive, this would require additional computational resources. Another point addressing the limitations of computational power is the sampling of only 15 days of data per month. Especially for the weekend related features, two weekends are arguably not enough days to determine a pattern of repetitive behaviour.

The sampling of 15 days was required for data size reduction purposes and could be overcome with additional computational power and a distributed setup.

One drawback in the proposed analysis' framework is a lack of full year pattern visibility. It would be better for the analysis to determine annual cyclic effects in a first step, and then adjusting the identified changes in the post-Covid time for the seasonal change. Such an analysis spanning more than two years before the Covid pandemic would require additional data. A similar argument can be made when it comes to the upper time limit, which was September 2020 in the proposed analysis. Arguably the Covid-19 pandemic had not ended by that time. Indeed, case number in Portugal were rising again at that time, thus it would be advisable to expand the analysis to the year 2021 to capture a truly normal post-pandemic (endemic) situation.

Additional analysis should be conducted at the flow between clusters. Currently the proposed analysis is only looking at net flows, combining inflows of vehicles from other clusters and outflows of vehicles to other clusters. It would be interesting to investigate further on these flows between clusters at different times of the year.

Overall, the proposed analysis provided the intended insights into the change of mobility patterns. Nevertheless, there are overlaps between clusters and defining a distinct behavioural persona for each cluster is not always possible. One way of solving this dilemma would be to optimize the number of clusters with respect to the distinguishability, meaning one increases k as long as there can be found distinct persona interpretations for each cluster.

Another option is to describe the change with summary statistics, such as the average number of trips per user, the average time travelled per user etc. While this approach might be suitable to identify changes in traffic behaviour overall, it falls short of building subgroups based on several features/summary statistics, and thus provides limited depth of insight.

Overall, the presented methodology reveals several patterns that appear meaningful, albeit symmetrical in the case of several clusters which raises some questions. The matching growth in one cluster and the decline in another cluster over several months, always opposite of each other, indicates that the cluster boundaries are not separating the data well. There is a need to analyse the flow between clusters from one month to the other, on an individual level. This way the identity of the vehicles changing clusters from one month to another can be compared. If most vehicles between two clusters are always the same, this would indicate the growth values to be less meaningful and rather a cause of the mismatch between the cluster boundaries and the data's cluster structure.

References

Anita Graser. 2019. “MovingPandas: Efficient Structures for Movement Data in Python.” *GI_Forum*, 54–68.

Asadi, Reza, and Amelia Regan. 2019. “Spatio-Temporal Clustering of Traffic Data with Deep Embedded Clustering.” In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Prediction of Human Mobility - PredictGIS'19*. New York, New York, USA: ACM Press.

Asma Belhadi, Youcef Djenouri, Kjetil Nørvåg, Heri Ramampiaro, and Jerry Chun-Wei Lin. 2020. “Space Time Series Clustering: Algorithms, Taxonomy, and Case Study on Urban Smart Cities.” *Engineering Applications of Artificial Intelligence* 95.

Beyer, Kevin, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. 1999. “When Is “Nearest Neighbor” Meaningful?” In *Database Theory — ICDT’99: 7th International Conference Jerusalem, Israel, January 10–12, 1999 Proceedings*, edited by Catriel Beeri and Peter Buneman, 217–35. Springer eBook Collection Computer Science 1540. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg.

Gerhard Münz, Sa Li, and G. Carle. 2007. “Traffic Anomaly Detection Using K-Means Clustering.” *International Conference on Recent Advancements in Information Technology*.

H. Kriegel, P. Kröger, and A. Zimek. 2009. “Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering.” *ACM Transactions on Knowlegde Discovery from Data* 3 (1): 1–58.

Han, Jiawei, Micheline Kamber, and Jian Pei. 2012. “10 - Cluster Analysis: Basic Concepts and Methods.” In *Data Mining: Concepts and Techniques*, edited by Jiawei Han. 3rd ed.,

443–95. Morgan Kaufmann series in data management systems. Waltham, MA: Morgan Kaufmann/Elsevier.

Heber Hernández, Elisabete Alberdi, Heriberto Pérez-Acebo, Irantzu Álvarez, María García, Isabel Eguia, and Kevin Fernández. 2021. “Managing Traffic Data Through Clustering and Radial Basis Functions.” *Sustainability* 13 (5): 2846.

J. H. Ward. 1963. “Hierarchical Grouping to Optimize an Objective Function.” *Journal of the American Statistical Association* 58: 236–44.

Kisilevich, Slava, Florian Mansmann, Mirco Nanni, and Salvatore Rinzivillo. 2010. “Spatio-Temporal Clustering.” In *Data Mining and Knowledge Discovery Handbook*, edited by Oded Z. Maimon. 2. ed., 855–74. New York, Heidelberg: Springer.

L. Pappalardo, F. Simini, Gianni Barlacchi, and Roberto Pellungrini. 2019. “Scikit-Mobility: A Python Library for the Analysis, Generation and Risk Assessment of Mobility Data.” *undefined*.

Landesberger, Tatiana von, Felix Brodkorb, Philipp Roskosch, Natalia Andrienko, Gennady Andrienko, and Andreas Kerren. 2016. “MobilityGraphs: Visual Analysis of Mass Mobility Dynamics via Spatio-Temporal Graphs and Clustering.” *IEEE Trans. Visual. Comput. Graphics* 22 (1): 11–20.

Maaten, L. V. D., and G. Hinton. 2008. “Visualizing Data Using T-SNE.” *Journal of Machine Learning Research* 9: 2579–2605.

Martin Ester, H. Kriegel, J. Sander, and Xiaowei Xu. 1996. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.” *Knowlegde Discovery and Data Mining*, 226–31.

W. Weijermars, and E. van Berkum. 2005. “Analyzing Highway Flow Patterns Using Cluster Analysis.” In *2005 IEEE Intelligent Transportation Systems Conference (ITSC): Vienna, Austria, 13-16 September 2005*, 308–13 2005. Piscataway, N.J: IEEE Operations Center.

Appendix 1: Light vehicle cluster centroids

		Night trips		Morning trips		Midday trips		Afternoon trips		Evening trips		Round trips	
Cluster	**Size**	**Workweek**	**Weekend**	**Workweek**	**Weekend**	**Workweek**	**Weekend**	**Workweek**	**Weekend**	**Workweek**	**Weekend**	**Workweek**	**Weekend**
1	22,343	0.2	0.1	1.4	0.1	1.2	0.1	1.9	0.1	0.8	0.1	3.7	0.3
2	10,481	0.2	0.1	3.9	0.2	0.3	0.1	0.9	0.2	3.5	0.1	11.9	0.7
3	6,440	0.3	0.2	1.6	0.6	0.6	0.2	1.2	0.4	1.4	1.2	5.2	2.7
4	3,397	2.4	1.1	1.4	0.4	0.8	0.2	1.2	0.3	1.0	0.3	7.3	2.5
5	1,033	1.5	0.8	5.2	2.0	2.5	1.1	3.6	1.5	3.4	1.2	20.6	8.2
6	40	30.5	12.4	26.0	10.2	11.5	4.4	17.8	7.2	16.7	6.9	54.1	19.7
7	12,076	0.1	0.0	4.5	0.1	0.4	0.1	4.0	0.1	0.6	0.1	13.9	0.4
8	4,018	0.2	0.1	3.8	0.2	3.9	0.1	2.7	0.1	1.0	0.1	11.6	0.6
9	9,517	0.2	0.1	1.7	0.7	0.6	0.2	1.4	1.0	0.9	0.2	4.6	2.1
10	22,624	0.1	0.0	3.3	0.1	0.4	0.0	1.4	0.1	0.7	0.0	4.0	0.2
11	5,633	0.3	0.2	1.8	0.5	0.9	1.2	1.4	0.5	1.1	0.3	5.0	2.4

All values are calculated per workweek/weekend, except round trips which are calculated as overall aggregates. Also, round trips are double counted, meaning one full round trip from A to B to A is counted as two. The code to generate the results of this thesis is available under https://github.com/felixpagel/thesis-clustering-traffic. Contact the author to get access to the repository: felix.julian.pagel@gmail.com

Appendix 2: Definition of daytime for feature engineering

Time interval	Discrete time of day
23:00 – 04:59	Night
05:00 – 11:59	Morning
12:00 – 13:59	Midday
14:00 – 18:59	Afternoon
19:00 – 22:59	Evening