A Cluster Analysis of Temporal Patterns of Travel Production in the Netherlands

Dominant within-day and day-to-day patterns and their association with Urbanization Levels

Eftekhar, Zahra; Pel, Adam; van Lint, Hans

DOI
10.18757/ejtir.2023.23.3.6499

Publication date
2023

Document Version
Final published version

Published in
European Journal of Transport and Infrastructure Research

Citation (APA)
Eftekhar, Z., Pel, A., & van Lint, H. (2023). A Cluster Analysis of Temporal Patterns of Travel Production in the Netherlands: Dominant within-day and day-to-day patterns and their association with Urbanization Levels. *European Journal of Transport and Infrastructure Research*, *23*(3), 1-29. https://doi.org/10.18757/ejtir.2023.23.3.6499

TU Delft OPEN

EJTIR

ISSN: 1567-7141

http://ejtir.tudelft.nl/

Issue 23(3), 2023, pp. 1-29

https://doi.org/10.18757/ejtir.2023.23.3.6499

A Cluster Analysis of Temporal Patterns of Travel Production in the Netherlands: Dominant within-day and day-to-day patterns and their association with Urbanization Levels

Zahra Eftekhar*
z.eftekhar-1@tudelft.nl
Department of Transport & Planning, Delft University of Technology, Netherlands.

Adam Pel
A.J.Pel@tudelft.nl
Department of Transport & Planning, Delft University of Technology, Netherlands.

Hans van Lint
j.w.c.vanlint@tudelft.nl
Department of Transport & Planning, Delft University of Technology, Netherlands.

Keywords

travel production
land-use feature
demand pattern
urbanization
temporal pattern

Publishing history

Submitted: 13 June 2022
Accepted: 22 August 2023
Published: 10 October 2023

Cite as

Eftekhar, Pel, van Lint (2023). A Cluster Analysis of Temporal Patterns of Travel Production in the Netherlands: Dominant within-day and day-to-day patterns and their association with Urbanization Levels. European Journal of Transport and Infrastructure Research, 23(3), 1-29.

This paper explores temporal patterns in travel production using a full month of production data from traffic analysis zones (TAZ) in the (entire) Netherlands. The mentioned data is a processed aggregated derivative (due to privacy concerns) from GSM traces of a Dutch telecommunication company. This research thus also sheds light on whether such a processed data source is representative of both regular and non-regular patterns in travel production and how such data can be used for planning purposes. To this end, we construct normalized matrix (heatmap) representations of weekly hour-by-hour travel production patterns of over 1200 TAZs, which we cluster using K-means combined with deep convolutional neural networks (inception V3) to extract relevant features. A silhouette score shows that three dominant clusters of temporal patterns can be discerned (K=3). These three clusters have distinctly different within-day and day-to-day production patterns in terms of peak period intensity over different days of the week. Subsequently, a spatial analysis of these clusters reveals that the differences can be related to (easily observable) land-use features such as urbanization levels (i.e., Urban, Rural, and mixed-level). To substantiate this hypothesis and the usefulness of this clustering result, we apply an OVR-SMOTE-XGBoost ensemble classification model on the land-use features of the TAZs (i.e., to identify their cluster). The results of our clustering analysis show that given the land-use features, the overall production patterns are identifiable. Further analysis of the mixed-level areas shows a more complex relationship between temporal heterogeneity and spatial characteristics. Population density seems to impose additional uncertainty on the temporal patterns. All in all, feature selection and spatial and temporal discretization play essential roles in identifying the dominant trip production patterns. These findings are directly useful for data-driven estimation and prediction of demand time series. Furthermore, this study provides further insights into people's mobility, relevant for transportation analysis and policies.

* Corresponding Author

1 Introduction

Call detailed records (CDR) can show valuable insights in various contexts such as daily mobility motifs (Schneider, Rudloff, Bauer, & González, 2013; Jiang, Ferreira, & Gonzalez, 2017), population movements (Antoniou, Chaniotakis, Katrakazas, & Tirachini, 2020; Szocska et al., 2021), and disaster response (Yabe, Jones, Rao, Gonzalez, & Ukkusuri, 2022). While all the mentioned studies use raw trajectories, acquiring this type of data is difficult. Telecom operators are banned from providing ground truth or contextual information to protect people's privacy at the—inevitable—expense of data utility. Usually, what is available is processed aggregated OD matrices for which the methodology used for processing and up-scaling to the whole population is unclear. This data is a valuable source of information on people's mobility, but its suitability for transportation planning remains unclear. Are these OD matrices representative of the population demand variations and irregular patterns? In many cases, due to the limited market share of the telecom operator, only a small and possibly behaviorally biased sample of people is presented in the raw data; How well do they represent the whole population? (Mamei, Bicocchi, Lippi, Mariani, & Zambonelli, 2019) raises attention to the need for more research and experiments assessing and evaluating the OD matrices derived from these data sources.

To make these processed data useful for urban planning, we must clarify their potential biases even when facing limited data resources to test these against. In this regard, consistency of data deals with its representativeness and correspondence to the real world. Essentially, data consistency has two main aspects: uniformity across the dataset and alignment with expected patterns(Demchenko, Grosso, De Laat, & Membrey, 2013). The consistency of the data can be assessed by using multiple sources when available, checking data credibility using various quantitative tools, and evaluating the internal data structure (Rubin & Lukoianova, 2013). In this paper, we evaluate the consistency of the OD matrices resulting from CDR both by assessing their associated travel demand patterns (representing internal data structure) and by comparison with data from an independent source, in our case, land-use characteristics.

Understanding travel demand patterns is essential for transportation planning and management. Firstly, travel demand analysis plays a significant role in identifying the current problems of transportation systems and helps in modeling the future traffic state (Thakuriah, 2001; Rich & Mabit, 2012). Secondly, the demand patterns help evaluate the impact of transportation infrastructure and management policies and strategies, such as flexible-time work schedules and congestion pricing (Gärling et al., 2002). Thirdly, understanding demand patterns is useful for developing better standards for evacuation plans, and responses (Pel, Bliemer, & Hoogendoorn, 2011; Xu, Chen, & Yang, 2017).

Travel demand patterns are characterized by spatial-temporal heterogeneity, as the amount of travel differs strongly across areas as well as time-of-day and day-of-week (Shen, Zhou, Jin, & Wang, 2020). Temporal variability is especially significant when modeling motor vehicle demand in urbanized areas where morning and afternoon peaks account for almost 50% of daily travel demand (Lin & Shin, 2008). Spatial heterogeneity is derived from diverse urbanization levels, economic activities, lifestyles, transportation accessibility, and resource distribution between areas, e.g., (Fotheringham, Charlton, & Brunsdon, 1998). This spatial-temporal heterogeneity, including the identification of patterns therein, is an important part of understanding travel demand.

Many studies analyze the relationship between travel demand and land-use properties using methods based on Ordinary least squares (OLS) (Yang et al., 2018; Maat & Timmermans, 2006). OLS generally assumes homogeneous regression relationships in the data. However, spatiotemporal data has the basic features of both spatiotemporal dependence and heterogeneity (Ermagun & Levinson, 2018; Atluri, Karpatne, & Kumar, 2018). For instance, in the temporal dimension of traffic data, the traffic demand of areas is typically similar to that of recent times.

Moreover, they might show periodicity, trends, and holiday effects. In the spatial dimension, the traffic state of each road segment is often (but not always) similar to upstream and downstream traffic conditions. This is an example of spatiotemporal dependence (T. Cheng, Haworth, Anbaroglu, Tanaksaranond, & Wang, 2013). Due to complex nonlinear variations, the traffic demand also has different distributions in different geographical regions and time intervals. For instance, for some areas, the peak hours of travel production happen in the afternoon, while it is in the morning for other areas. This phenomenon represents spatiotemporal heterogeneity (Atluri et al., 2018). Overlooking spatiotemporal heterogeneity gives rise to some errors, for instance, misinterpretation of coefficients and inaccuracy in estimations (Anselin & Griffith, 1988; M. Deng et al., 2018; S. Cheng, Lu, Peng, & Wu, 2019).

To account for spatial heterogeneity, many extended OLS-like models have been developed, amongst which the geographically weighted regression (GWR) model (Brunsdon, Fotheringham, & Charlton, 1996) is widely used in transportation studies. For example, Cardozo, García Palomares, and Gutiérrez study the relationship between transit travel demand and land use mix, bus accessibility, and road density using a GWR model. Whereas the GWR model can sufficiently describe spatial heterogeneity, it does not address temporal heterogeneity. Typically, days are divided into multiple periods, in which average (proportional) values are considered for modeling. To incorporate temporal heterogeneity, a geographically and temporally weighted regression (GTWR) method to predict transit travel demand was first applied by (Ma, Zhang, Ding, & Wang, 2018).

However, little is still known about how different areas have various spatiotemporal patterns in travel production associated with their urban development. There is a need to understand the factors that mediate the interactions between urbanization and travel production in time and space. Fekih et al. proposes a framework to extract spatiotemporal travel demand patterns from large-scale GSM traces. Their analysis focuses on within-day variations of travel demand. In this paper, we build on this work and investigate both within-day and day-to-day production patterns of all the traffic analysis zones (TAZs) in the Netherlands. We tried to capture a more holistic picture of temporal production patterns through normalized heatmaps. After feeding these heatmaps to a deep convolutional neural network (DCNN) and K-means method, three main patterns were discerned. Analysis of these patterns is one of the two methods of evaluating data consistency.

The performed temporal analysis of the underlying patterns is valuable for adjusting the demand models and prediction. Later we link these temporal patterns to spatial urbanization levels through their land-use characteristics, which is beneficial for urban development strategies and policymakers. In fact, this study proposes three urbanization levels for the Netherlands: urban, rural, and other, associated with their specific land-use characteristics and travel production patterns within the day and day-to-day, and this study explores the differences in these patterns. This effort is the second method of assessing the data consistency. Furthermore, an OVR-SMOTE-XGBoost ensemble classification method is proposed to investigate the relationship between land-use characteristics and temporal production patterns. Our findings suggest that given the land-use features of each TAZ, their most probable travel production temporal pattern is detectable. The results are beneficial for dynamic demand prediction models. Furthermore, we find indications for the data's ability to represent both dominant patterns and variations correctly in spite of the data's inherent bias. This study will help planners discern and assess the representativeness and suitability of using the processed, up-scaled derivative of GSM traces in traffic planning.

The remainder of this paper is organized as follows: Section 2 describes the research data and the implemented method. In Section 3, we present and discuss the results of our analysis on the temporal patterns found in the Netherlands. Finally, Section 4 concludes the paper.

2 Methodology

To understand the temporal patterns in travel production and how these might be associated with spatial urbanization levels, we can distinguish three main research components (as shown in the overall research framework in Figure 1: the data showing the spatiotemporal travel productions, the clustering method to discern temporal patterns, and the association analysis method to discern spatial relations. These three components are explained one by one in the following subsections.

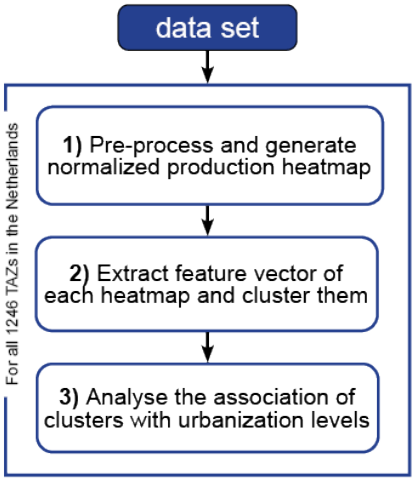


Figure 1. Overall research framework.

2.1 Travel production data:

This research uses the hourly production of the 4-digit postal code zones in the entire Netherlands during March 2017. Using 4-digit postal code zones leads to 1246 TAZs in the Netherlands (see Figure 2). Travel production of TAZ *i* is defined as the number of inter-zonal trips made by motor

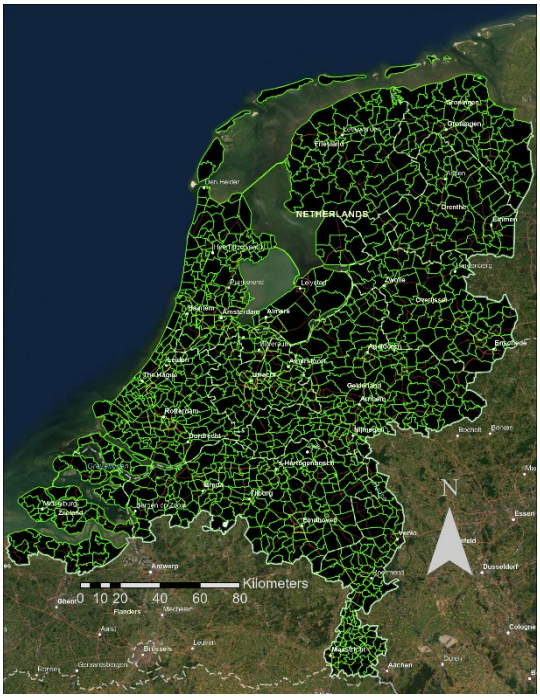


Figure 2. TAZs in the Netherlands.

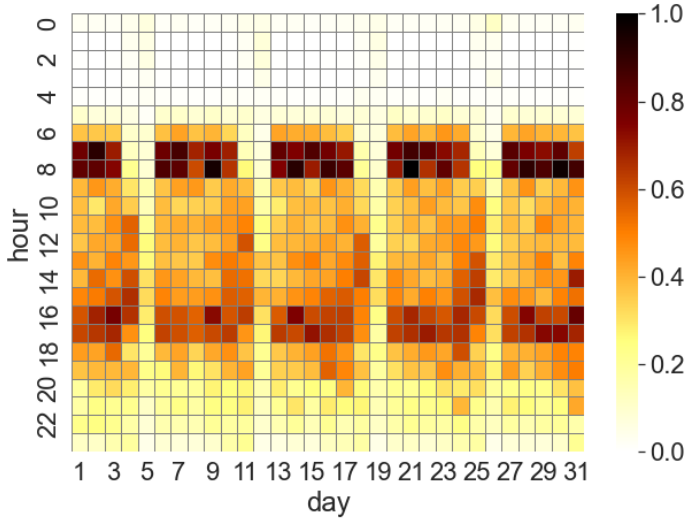


Figure 3. An example of production heatmap for one TAZ.

vehicles starting at i. The production values are derived from the GSM traces of the Dutch telecommunication company Vodafone, whose market share is about one-third of the Dutch population. Another company performed the processing due to privacy concerns regarding the raw mobile phone data. Consequently, the available data for this study, instead of the mobile phone traces, consists of origin-destination (OD) matrices of the motor vehicles based on TAZs in the Netherlands. These OD matrices have been initially scaled up to the entire Dutch population. We refer the reader to (Meppelink, Van Langen, Siebes, & Spruit, 2020) for more detail on the scaling procedure.

To begin with, we collected the (processed) data. After pre-processing and reshaping, each zone had a heatmap of normalized production values. Normalizing the production values enables fast and stable pattern comparison of various zones. The technique we applied on each production value x for normalizing is Min-Max Scaling, i.e.,

$$x_{normalized} = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad (1)$$

where $x_{normalized}$ is the normalized value, x_{min} and $x_{max}$are the minimum and maximum production values in the (month-long) time series of that particular TAZ. Thus the resulting normalized values range between 0 and 1.

The result is a production heatmap for each TAZ in which the horizontal and vertical axes represent the days of the month and the hours of a day, respectively. Figure 3 shows an example.

This representation allows us to see the temporal patterns of production both within a day (comparing across rows) and between days (comparing across columns). These 1246 heatmaps of travel production are the basis for the following analyses.

2.2 Clustering temporal production patterns:

The heatmaps of hourly travel production per TAZ are clustered based on (temporal) similarity using K-means clustering. Due to its easy application and effectiveness, the K-means clustering method is one of the most popular algorithms for clustering analysis (Potera ş, Mih ăescu, & Mocanu, 2014; Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016; Cohn & Holm, 2021; Van Gansbeke, Vandenhende, Georgoulis, Proesmans, & Van Gool, 2020). The method works by splitting the N-dimensional data set of M points (heatmaps) into K clusters such that the sum of the pairwise Euclidean distance between the points of each cluster is minimized (Hartigan & Wong, 1979). In other words, this method aims to maximize the similarity between the points in the same

cluster and maximize the dissimilarity of points from different clusters. The method's initialization is by randomly selecting K points as the cluster centroids. The clustering process has two significant steps:

- Assignment: assigning each point to its closest centroid. Mathematically this step refers to partitioning the points to the Voronoi diagram (Shamos & Hoey, 1975) generated by the centroids.
- Update: updating each cluster center to be the average of all points contained within them.

Applying the K-means method means that the number of clusters is exogenously chosen and thus needs to be justified. As we cannot determine how many patterns exist in advance, this is an unsupervised learning problem. That is, there are no "true labels" or ground truth available (i.e., there is no a-priori behavioral hypothesis on the existence of specific patterns), and therefore the appropriate number of clusters is best chosen by assessing the (dis)similarity within and between clusters, for different values of K. To calculate the goodness of clustering, we used the Silhouette index (Rousseeuw, 1987). There are several methods for clustering evaluation, such as distortion score (Camps-Valls, 2006), Rand index (Yeung, Haynor, & Ruzzo, 2001), adjusted Rand index (Hubert & Arabie, 1985), and Silhouette index. The latter is particularly useful for an unsupervised evaluation of clustering.

The Silhouette ranges between -1 and 1, where high values show a well-matched point to its own cluster and poorly matched to the neighboring clusters. If many points have a negative value, the number of clusters needs to be modified. The Silhouette of point i is calculated as $s(i) = \frac{x(i)-y(i)}{\max(x(i),y(i))}$ where, x is the average distance with points in other clusters (i.e., dissimilarity), and y is the average distance to points of the same cluster. Silhouette index of a cluster is the average s(i) of all the points inside the cluster, and the Silhouette index for all clusters (SI) is the average s(i) of all points in all clusters. The optimal number of clusters happens when the SI is maximum. That is when on average, the difference between mean intra-cluster and nearest-cluster distance is the highest. For instance, if we calculate SI for the different number of clusters (i.e., K) in the range of 2, 15 and the maximum values of SI happen when K=3, the optimal K equals 3.

To evaluate the separateness of clusters, the inter-cluster distance is commonly used (Liu et al., 2012). Visualizing the inter-cluster distances in two dimensions gives insight into the relative importance of clusters. To this end, multidimensional scaling (MDS) embed multi-dimensional cluster centers into 2-dimensional space (we refer the reader to (Bengfort et al., 2018; Kruskal, 1964)).

This method preserves the distance to other centers. For instance, if cluster centers are close to each other in the original feature space, they are also close when embedded into 2-dimensional space. In the inter-cluster distance map, the clusters are sized according to the number of instances that belong to each cluster which in a sense reflects how important each cluster is.

Another issue that needs to be addressed is that the K-means method is inherently a linear algorithm (Ning & Hongyi, 2016). Therefore, it is unsuitable for complex nonlinear data distributions. To take the non-linearity of data into account, a deep convolutional neural network (DCNN) is used for feature extraction. The DCNN transforms input heatmaps to final representations, so-called feature vectors, which are more easily separable by a linear clustering algorithm (van El teren, 2018) than the original heatmap. To this end, the DCNN identifies key (salient) features of the heatmaps for analysis and clustering, and after this transformation step, we have a set of feature vectors that are then used for K-means clustering (instead of applied directly on the heatmaps).

As DCNN, the state-of-the-art InceptionV3 based on transfer learning is used. The InceptionV3 architecture specifically improves adaptability to different scales, and overfitting is better prevented (Kaur & Gandhi, 2020). One way in which this is done is by using transfer learning.

Transfer learning lets us transfer already trained model parameters to our new model and therefore accelerates its training (Wang et al., 2019). After training a DCNN on a large dataset, e.g., ImageNet, one can adopt transfer learning because the DCNN is able to learn generic features (think of edges or other geometric shapes) that are also applicable to other images (such as heatmaps) without the need for training from scratch. Furthermore, the weights of the DCNN, which is pre-trained on a large dataset, improve its accuracy for specific tasks such as pattern recognition tasks in which the amount of available training data is limited (Iglovikov & Shvets, 2018). This also holds for our data set. In this study, we thus extract feature vectors from the demand heatmaps using the InceptionV3 deep neural network, which is pre-trained on the ImageNet dataset consisting of millions of images used for object recognition and image classification. For details about this dataset, we refer the reader to (J. Deng et al., 2009).

In summary, this step transforms the 1246 heatmaps of travel production into feature vectors that are then clustered based on (temporal) similarity, yielding K clusters with a distinct temporal production pattern, and in the process, determine how many (K) of such clusters/patterns exist.

2.3 Association with urbanization levels:

In the final step, we analyze the degree to which the K clusters are associated with specific urbanization levels. That is the degree to which the urbanization level of a TAZ can be used as a predictor for to which cluster of temporal production patterns it belongs.

Comparing the overall area of different land-use types (derived from Open Street Map (OSM) (OpenStreetMap contributors, 2017) data) in the clusters helps relate the resulting clusters (i.e., temporal production patterns) to land-use types. Assuming a non-linear relationship, we propose a tree-based ensemble machine learning method, eXtreme Gradient Boosting (XGBoost) Ensemble, to model the relationship between the land-use feature and clusters. Due to the regularization term in the loss function, one can achieve the lowest complexity with the highest accuracy. For more details on the XGBoost algorithm we refer the readers to appendix A and (Chen & Guestrin, 2016).

The framework of our method, called the OVR-SMOTEXGBoost ensemble model, for our multi-class imbalanced data is shown in Figure 4 and the three main steps in this algorithm are described as follows:

- Decompose the multi-class classification problem into multiple independent binary classification problems. Traditionally, classification methods are designed for two-class (binary) problems. Furthermore, as the generality of multi-class classification problems naturally makes learning more complex, an intuitive approach is to solve such problems by decomposing them into several binary classification problems. In One-vs.-Rest (OVR) strategy, one develops multiple classifiers (i.e., one classifier per class indicating being or not being in a specific class) (Hong & Cho, 2008). Based on the OVR decomposition method, the initial training set is decomposed into three two-class training sub-samples: $Train_1$ for class1, $Train_2$ for class1, and $Train_3$ for class3. Then using a binary logistic loss function, each classifier calculates the probability of each class. Then each instance is classified into the class of the highest probability.
- Balance the training sets. An issue regarding the classification is the imbalanced learning problem. Underrepresented data and class distribution skew affect the learning algorithm performance and cause this problem (He & Garcia, 2009). The synthetic minority oversampling technique (SMOTE) is one of the most recognized data augmentation methods to address the imbalanced data problem (Zhai, Qi, & Shen, 2022). It attempts to balance class distribution by oversampling minority class instances by randomly replicating them.

- Use the training sets to train the SMOTE-XGBoost ensemble model for a binary class of TAZs prediction (i.e., each of three OVR classifiers).

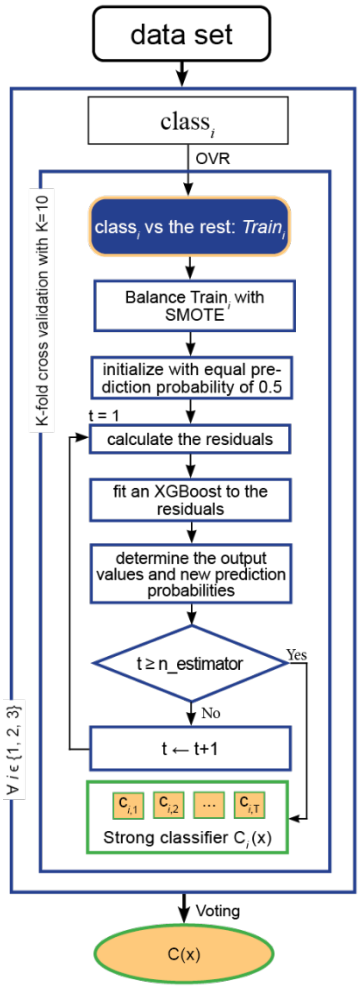


Figure 4. The framework of the OVR-SMOTE-XGBoost ensemble model.ensemble model.

To avoid overfitting, we use a K-fold strategy with k = 10 for the training and testing (i.e., cross-validation) (Note that K here is completely unrelated to K in the earlier K-means clustering method in that these are two independent parameters, but that simply both these methods happen to use the same letter). K-fold divides the data elements into K groups of samples, i.e., k folds, and the training of the model is achieved using K-1 folds, and testing is done on the left-out fold. For instance, given a dataset of 100 samples, 90 (i.e., nine folds) are used for training and validation. However, the predictive accuracy is tested on the remaining ten (i.e., one fold) samples. This process is iteratively repeated for all ten folds (with no overlap) to achieve a robust accuracy in prediction.

To further investigate the main spatial patterns inside a cluster, we apply a hierarchical clustering analysis (HCA). This step leads to several spatial sub-clusters explaining the temporal heterogeneity observed in the temporal clusters. HCA is generally a clustering analysis method that tries to build a hierarchy of clusters. In this study, we used an Agglomerative approach where each TAZ starts in its own cluster (i.e., bottom of the hierarchy). At each iteration, the similar clusters, based on their spatial characteristics, merge with others until one cluster or N clusters are formed (i.e., top of the hierarchy). In this research, we calculate the similarity between two clusters

using Ward's method, which is the sum of the square Euclidean distances between the two associated clusters.

The Supplementary data and selected hyperparameters associated with this paper will be presented online.

3 Results and Discussion

The first step in our research method was to apply the DCNN to the travel production heatmaps for feature extraction. Because we needed to use a pre-trained model (i.e., via transfer learning), we had to resize the input to the same format the network was originally trained on, that is, 224 by 224 pixels (numbers). Subsequently, the resulting feature vector had 224^2 features representing each heatmap. We clustered all these vectors of the entire 1246 TAZs in the Netherlands and calculated SI for various K values to determine the optimal number of clusters as shown in Figure 5. The scores on the y-axis are the average Silhouette index of all the features, SI, that each TAZ includes (i.e., travel production feature vector). Accordingly, the best cluster separation occurs when $K= 3$ because the maximum SI belongs to the case where $K= 3$. It is worth mentioning that SI for $K= 3$ is relatively close to that of $K= 2$. Still, this slight difference cannot be treated insignificantly. After the feature extraction step, the trip production features constitute large vectors containing many zeroes, and as the SI is the product of averaging the Euclidean distances, any difference is meaningful --- indicating non-zero elements of trip production features.

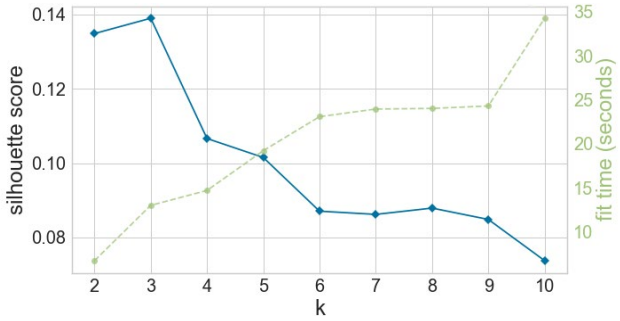


Figure 5. Silhouette index for finding the optimum number of clusters.

Also, based on the inter-cluster distance map in Figure 6, the three clusters seem to be well-separated. Clustering of the temporal patterns of production was performed using inceptionV3 with the K-mean method.

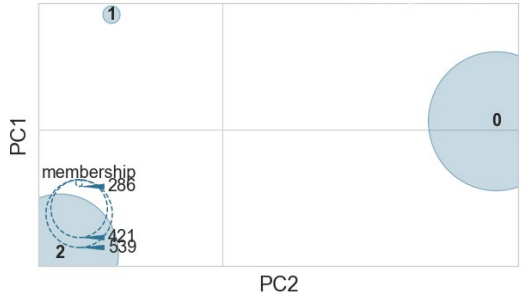
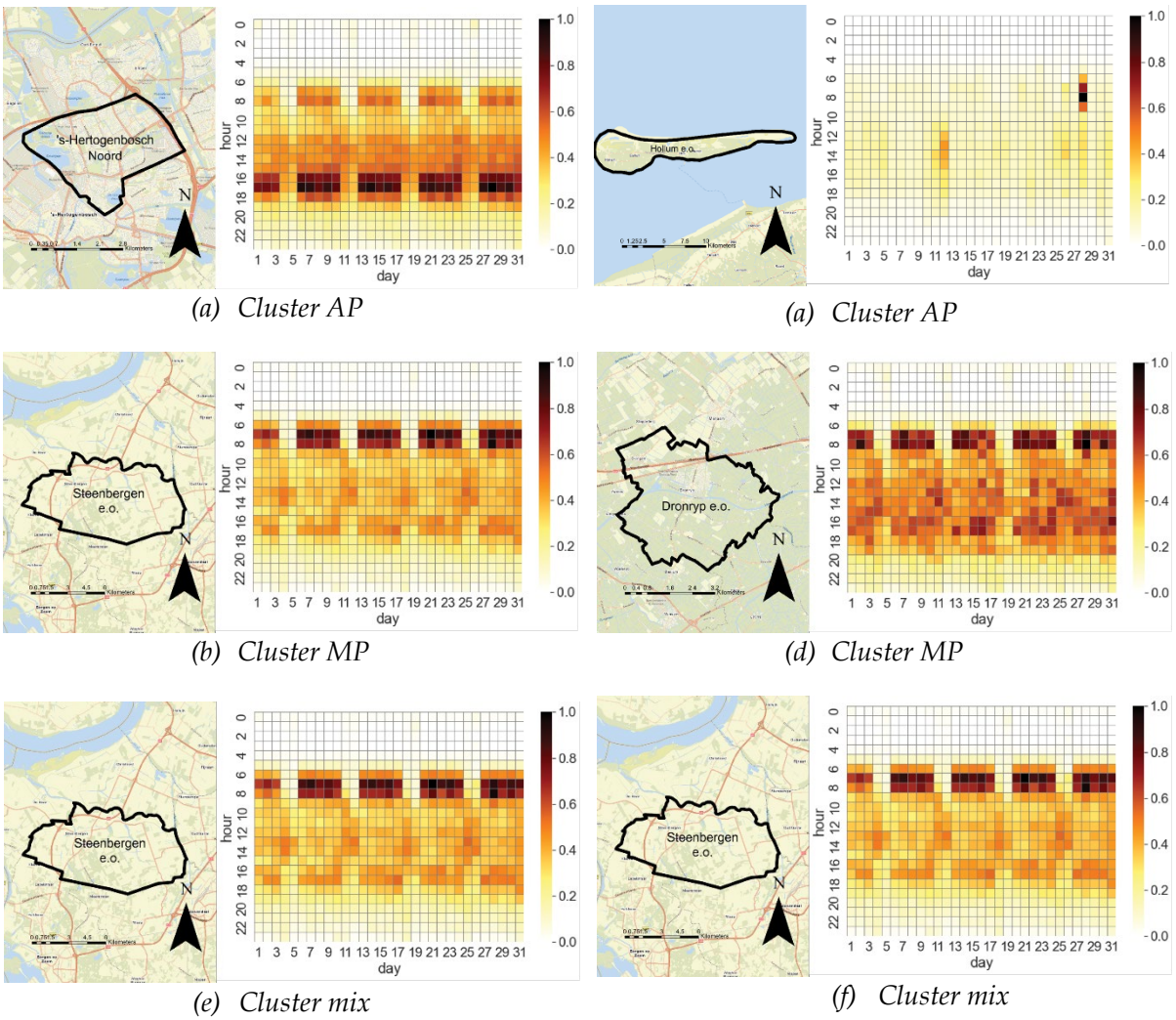


Figure 6. K-means inter-cluster distance map when K=3 and MDS is used for dimensionality reduction.

The remainder of this section is divided into two parts: Firstly, we show and analyze the results of clustering the travel production heatmaps in Section 3.1; Secondly, in Section 3.2, we discuss the association between the extracted temporal clusters and spatial characteristics.

3.1 Travel production patterns

Examples of the three clusters of temporal heatmaps of travel production are given in Figure 7. Of each cluster, two TAZ heatmaps are shown: the left images represent the closest heatmap to the cluster centroid, and the right pictures show the furthest heatmap from the cluster centroid.



(a) Cluster AP *(a) Cluster AP*

(b) Cluster MP *(d) Cluster MP*

(e) Cluster mix *(f) Cluster mix*

Figure 7. Travel production clusters.

For the first cluster, as shown in Figures 7a and 7b, containing 286 TAZs, we generally observe the presence of afternoon peaks between 15:00 and 19:00 in travel demand. Therefore, this cluster is named AP. Unlike the working days, weekends do not have significant peaks. Thus the afternoon peak could be due to more work-related areas, i.e., people tend to leave work in the afternoon, which causes a rise in travel production. Figure 7b belongs to the first cluster, although both its location and temporal production pattern seem to be indicating an outlier. Figure 8 shows the Silhouette score distribution per cluster. As negative scores reflect poor heatmap-to-cluster matches, we can see that cluster *AP* contains more outliers than the other two.

The second cluster shown by Figures 7c and 7d, contains 421 TAZs, and displays more distinct morning peaks between 06:00 and 09:00 in travel demand. Hence, this cluster is named *MP*. As this morning peak is predominantly observed on working days, this could be due to residential areas, i.e., people leaving their houses in the morning. Compared to cluster AP, the smaller peak interval reflects more scheduled activities (e.g., starting time of work) in the mornings and less obligation to leave on time (e.g., from work) in the afternoons. Additionally, less regular activities like

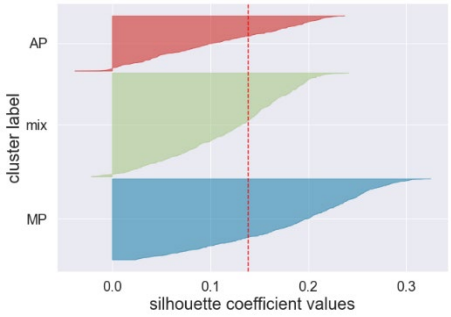
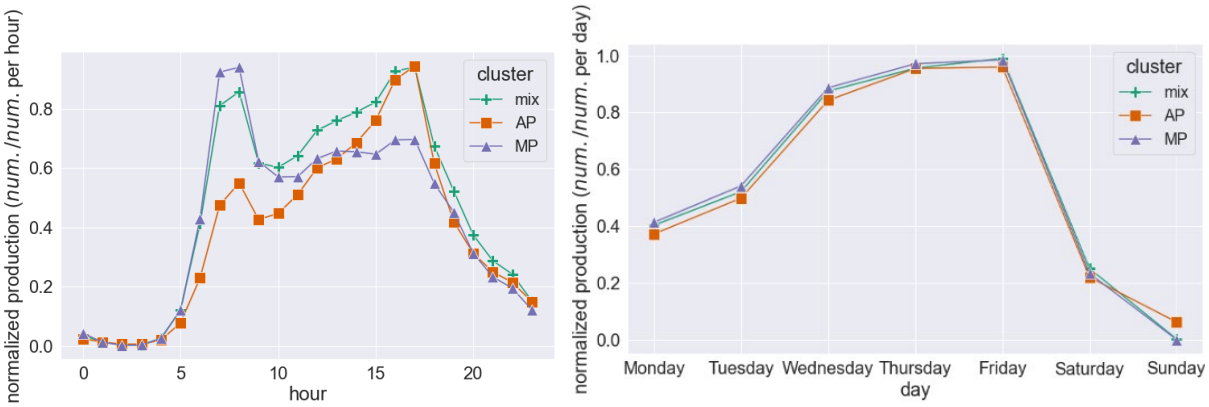


Figure 8. Silhouette score distribution per cluster.

shopping and social events in the afternoon in cluster *AP* also trigger the longer peak range. In Figure 7d, which falls further away from the cluster centroid, the afternoon peak starts to become more severe. In fact, a further increase in the afternoon values would probably cause this TAZ to be assigned to the third cluster. According to Figure 8, this cluster has higher Silhouette scores and lower outliers (i.e., negative Silhouette scores).

The third cluster, named mix, containing 539 TAZs, displays patterns other than those observed in cluster *AP* and *MP*. For instance, in Figure 7e, morning and afternoon peak seems almost equally extreme with higher values and scheduled activities in the morning, which can be a presenter of suburban areas where a mix of residential and work-related activities is established. Figure 7f, on the other hand, presents a different pattern: This TAZ seems to be a weekend trip producer as some peaks are observed during the weekends and Friday afternoons. Moreover, the afternoons seem to demonstrate higher values.

To investigate the dominant *within-day* patterns per cluster, Figure 9a shows the average normalized hourly travel production of the three clusters. To this end, for each TAZ, the hourly travel demand is averaged across all days of the month and then is min-max normalized across these 24-hour values. And then, for each cluster, these normalized 24-hour patterns are averaged across all TAZs inside that cluster. This leads to the results shown in Figure 9a, which indeed confirm the earlier observations of morning peak, afternoon peak, and mixed temporal patterns.



(a) Hourly production. *(b) Daily production.*

Figure 9. Average of normalized travel production of the three clusters.

To investigate the dominant *between-day* patterns per cluster, Figure 9b shows the average normalized daily travel production of the three clusters. To this end, for each TAZ, the hourly travel demand is summed per day, and then an average daily demand is computed for each day of the week, and then min-max normalized across these 7-day values. For each cluster, these normalized 7-day patterns are averaged across all TAZs inside each cluster. This leads to the results shown in Figure 9b. The between-day patterns in all clusters demonstrate a similar trend, although

cluster *AP* shows slightly higher values on Sundays, compensating for lower values during the rest of the week.

Figure 10 shows the *absolute within-day* travel production in each cluster. The values are computed similarly as for Figure 9a but without normalizing. The shaded area shows the *90th* and *10th* percentile. As expected, the values are widely scattered, especially in cluster *AP*, with higher average production in almost all hours of the day. In contrast, cluster *MP* displays the lowest average production throughout the day. The morning and afternoon peaks of all clusters seem to be almost in the same time range; however, their difference in the production values are more significant during the afternoon peak. Overall, the means of clusters seem to be significantly different.

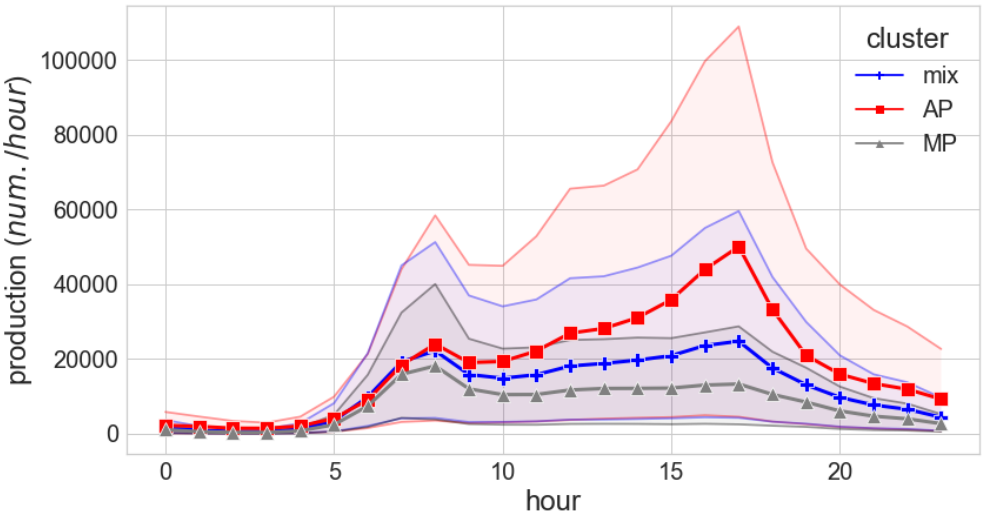


Figure 10. Average hourly production of the three clusters with 10th and 90th percentile as the shaded area.

Figure 11 shows the absolute between-day travel production in each cluster. The values are computed similarly as in Figure 10 but without normalizing. The shaded area shows the 90th and 10th percentile. In the same way as hourly patterns, weekly average values seem to be well-separated. However, the Figure shows high variance meaning a widely scattered distribution of

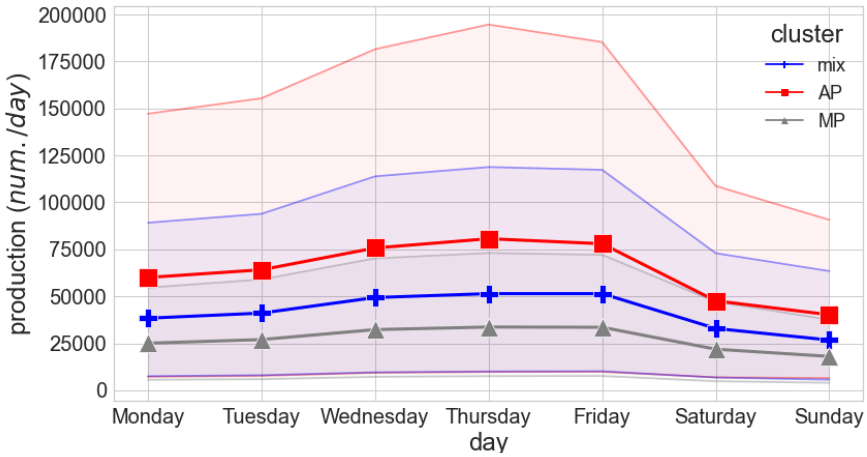
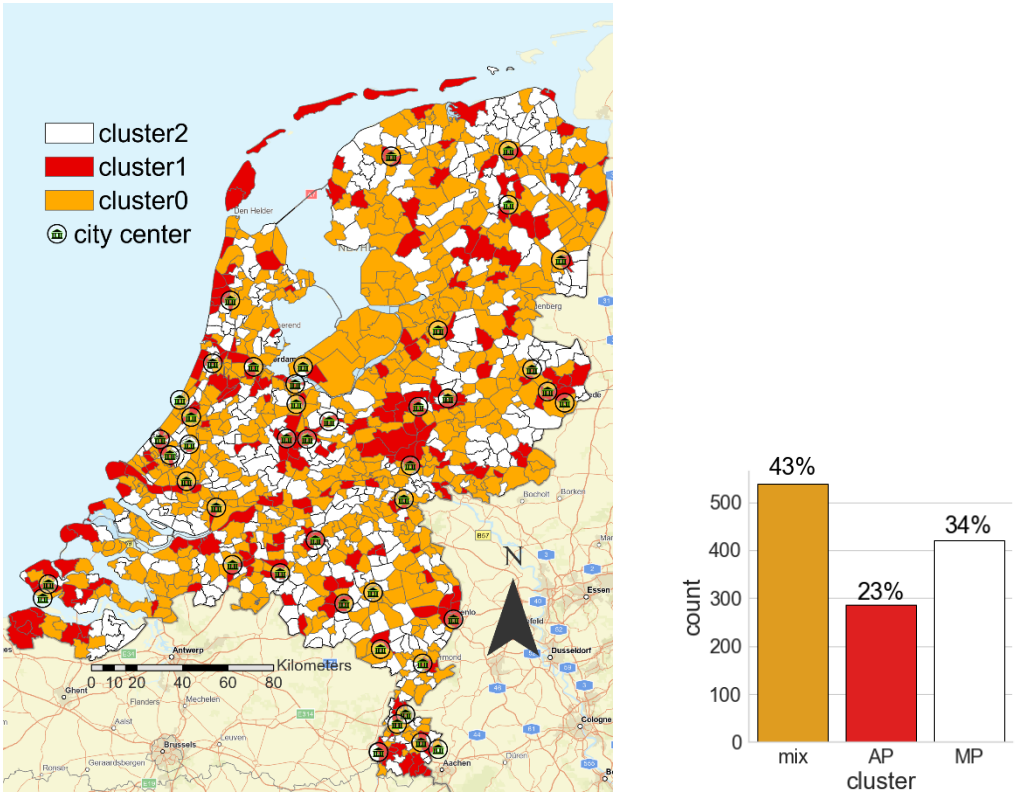


Figure 11. Average daily production of the three clusters with 10th and 90th percentile as the shaded area.

values. Moreover, the daily patterns seem very similar among all the clusters. A slight difference lies in the days with maximum averages. These are Thursday and Tuesday for cluster *AP*, which indicates working areas. The days with maximum production for cluster *mix* and cluster *MP* are Tuesday, Thursday, and Friday, which can hint that work is not as dominant as it is in cluster *AP*, i.e., as some part-time employees do not work on Fridays, having growth in the production values mainly indicates non-work-related trips. Furthermore, unlike Figure 9b in which cluster *mix* holds the maximum mean, Figure 11 display highest mean for cluster AP. Once again, it implies that productions in cluster *mix* have a smaller standard deviation (i.e., values are closer to their maximum) which produces higher values in the normalized production plot.



(a) Hourly production. *(b) Daily production.*

Figure 12. Average of normalized travel production of the three clustrs.

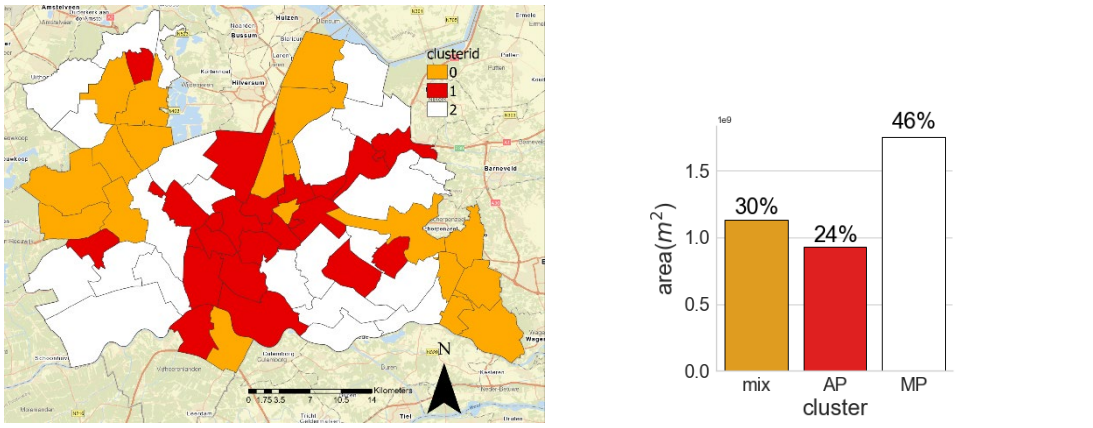
3.2 *Association between travel production patterns and spatial characteristics*

Observing the three clusters in space suggests that zones in cluster *AP* which constitute a smaller proportion of zones (as shown in Figure 12b), happen to be in more urbanized areas. In fact, as displayed in Figure 12a, out of 43, 39 city centers fall into this cluster. Moreover, the non-metropolitan (i.e., the least urbanized) areas happen to be in cluster *MP* and mix. Therefore we hypothesize that these (temporal) clusters are (spatially) associated with urbanization levels. For instance, usually, the majority of farmlands belong to the non-metropolitan (i.e., the least urbanized) areas, thus possibly also to clusters *MP* and *mix*.

To test this, the association of each cluster with the following land-use characteristics is analyzed:

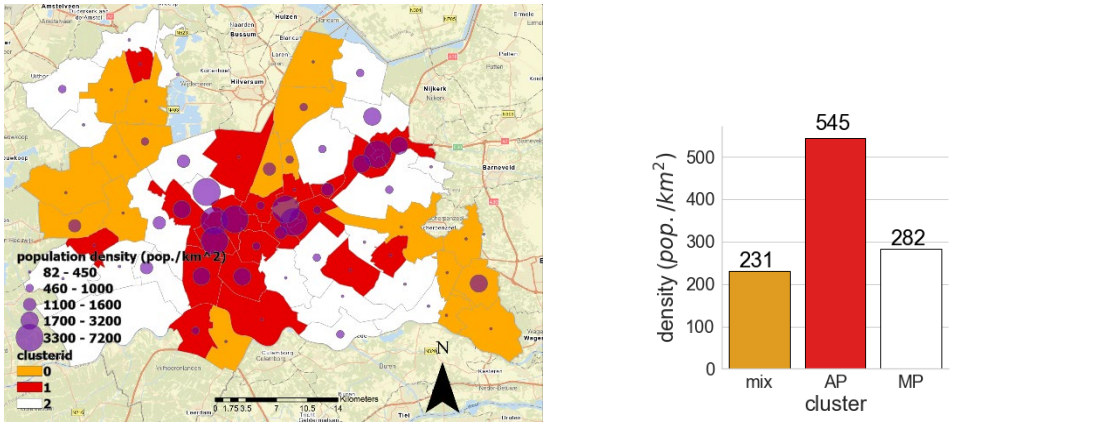
- population
- commercial and industrial buildings,
- rail and service roads,
- cycleway and footway,
- car and bicycle parking.

This is initially done for the province of Utrecht (as shown in Figure 13a), but the subsequent model will be based on all 1246 TAZs in the Netherlands. Figure 13b shows the distribution of TAZs for each cluster in the province of Utrecht.



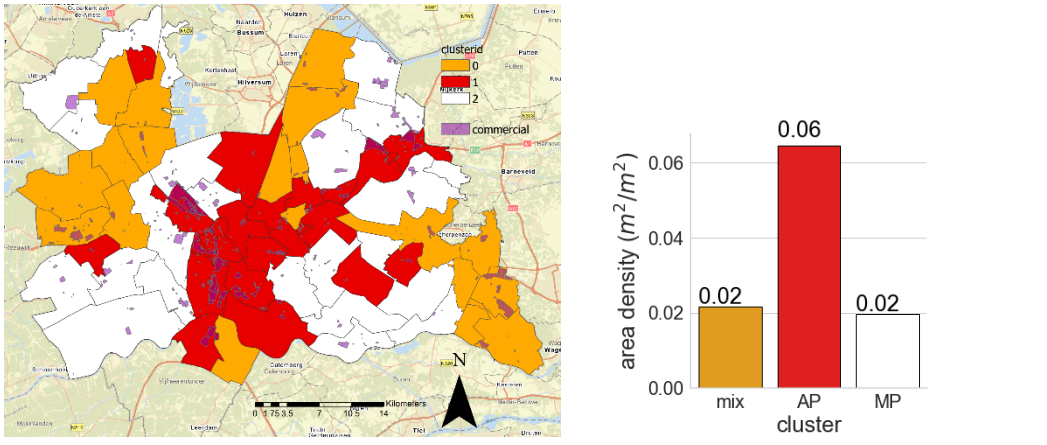
(a) *The three clusters.* (b) *Distribution of total area over the clusters.*

Figure 13. Share of clusters from the total area of Utrecht province.



(a) *Population density of the zones and the three clusters.* (b) *Population density over the clusters.*

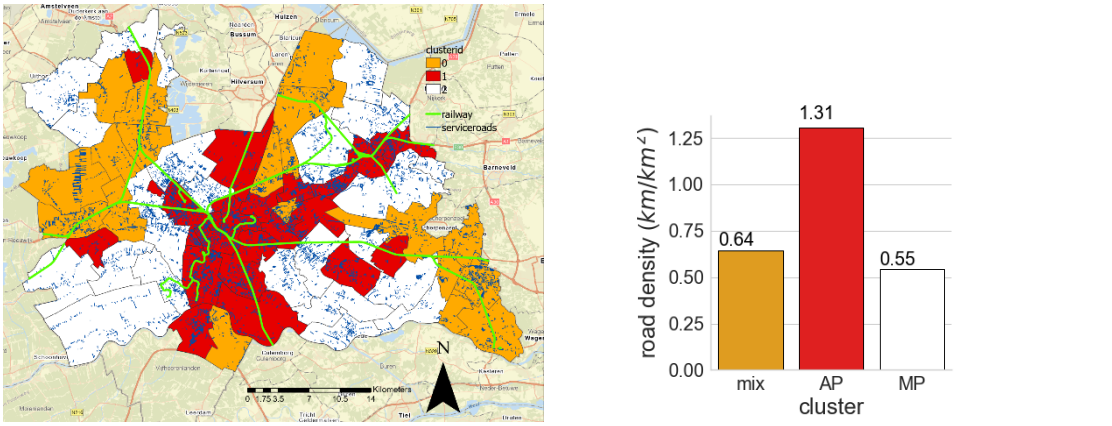
Figure 14. Share of clusters from the total population of Utrecht province.



(a) *Commercial/industrial land-use.* (b) *Density of commercial/industrial area over the clusters.*

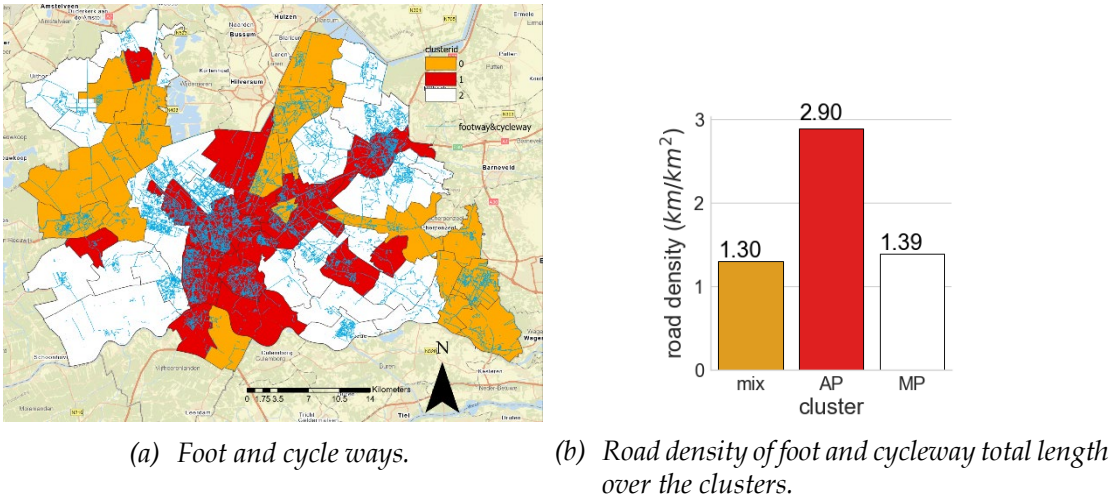
Figure 15. Share of clusters from commercial/industrial areas.

Looking specifically at the province of Utrecht, Figure 14b shows indeed a higher average population density in cluster AP. Figure 15 also shows that the majority of commercial/industrial areas, as a representation of urban areas, in the province of Utrecht belong to cluster AP. Figures 16, 17b, and 18b indicate that transportation infrastructure are densely distributed in cluster AP. However, Figure 19b shows a high density of farm and meadow lands in cluster *MP* and *mix*. Also, Figure 19 shows that 48% of farmlands which are usually interpreted as rural areas, fall into cluster *MP*. Therefore, density in the mentioned characteristics of each TAZ may be associated with their cluster.



(a) Service roads and railways. *(b) Distribution of service roads and railways total length over the clusters.*

Figure 16. Road density of clusters from the total length of service roads and railways.



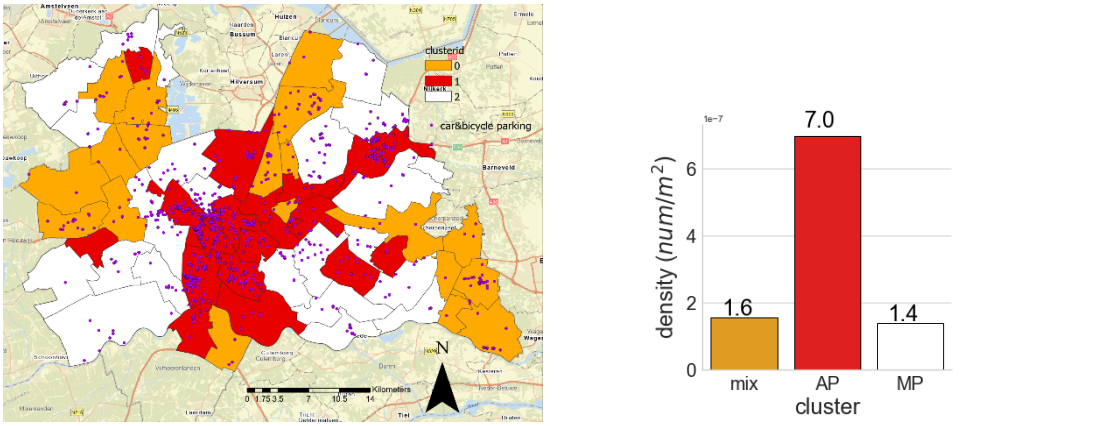
(a) Foot and cycle ways. *(b) Road density of foot and cycleway total length over the clusters.*

Figure 17. Share of clusters from the total length of foot and cycle ways.

To test this hypothesis at the level of all 1246 TAZs in the Netherlands, an OVR-SMOTE-XGBoost ensemble classification model is trained in which the inputs are the land-use characteristics (i.e., densities) and the output is the cluster, i.e., *AP, MP, or mix*. The model tries to reconstruct the clusters using the land-use characteristics and population density (as presented in Figures 14 to 19) associated with each TAZ. We used a K-fold cross-validation strategy with k = 10 to achieve robust accuracy in prediction. Accordingly, Figure 20 shows the confusion matrix of the associated classification.

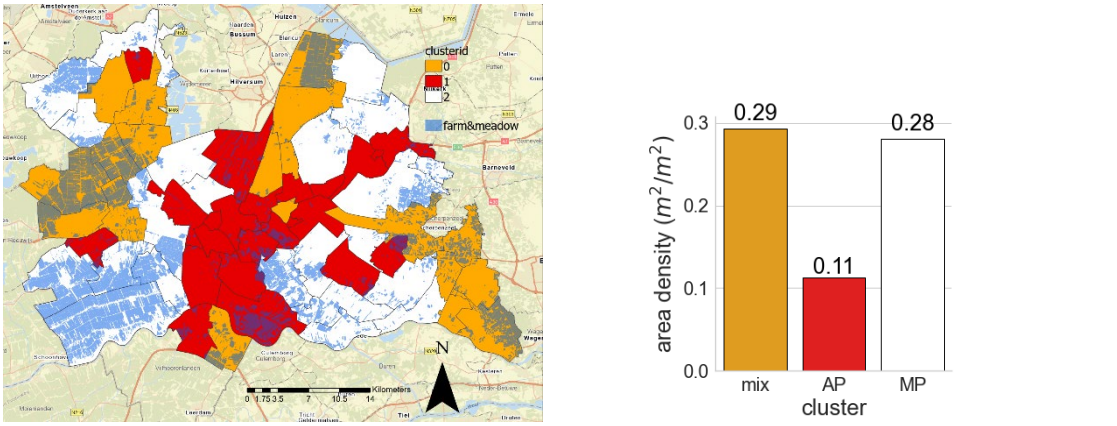
Note that due to our model structure, the possible bias caused by imbalanced training data is avoided. One can imply this by considering that classification accuracy in clusters *AP* is the highest (i.e., 64%), although the number of such TAZs is the lowest (i.e., 286 out of 1246). If the model was

biased toward the higher populated cluster, mix, the model accuracy would have been around 43%, which is derived from the proportion of TAZs belonging to mix. However, our accuracy is higher (53%). Therefore, it seems that the model is not biased due to imbalanced data. In fact, the accuracy of identifying cluster *mix* seems to be the lowest. This may be due to the similarity of patterns in cluster *mix* to both other clusters, i.e., cluster *mix* constitute mix of clusters *AP* and *MP*. Therefore, misclassification is more probable in mix. Especially, making distinction between *MP* and *mix* seems to be more difficult. This might be due to less extreme peaks in *MP* than in *AP*; thus more difference between *AP* and mix. Overall, the land-use feature we used seems to describe *AP* to a greater extent. To look closer, Figure 21 presents the confusion matrices of all 12 provinces in the Netherlands. In almost all provinces, the accuracy in *mix* is lower than in the other two. Only one province (Zeeland) shows a different result; In fact, it has the highest accuracy for mix. The cause of this behavior needs further analysis, but it can be related to the exceptional location of this province as a river delta situated at the mouth of several major rivers at the country's border(as shown in Figure 22).



(a) Car and bicycle parking garages. *(b) Density of car/bicycle parking total area over the clusters.*

Figure 18. Share of clusters from the total area of car and bicycle parking garages.



(a) Farmland land-use. *(b) Density of farmland/meadow area over the clusters.*

Figure 19. Share of clusters from farmland/meadow areas.

To take a closer look at mix, Figure 23 shows the most extreme TAZs associated with false negatives and positives of mix. That is, the TAZs with the highest probability to be false positive and false negative of the *mix* cluster. $C_{i,j}$is the most extreme TAZ whose true label is i, and the predicted label is j, for instance. $C_{mix,MP}$ shows the TAZ in the *mix* cluster, which is incorrectly classified as a

TAZ in *MP*. The normalized trip production heatmap shows the dominant pattern of *MP*. However, the irregular peak in the afternoon of Friday, March 24th, might have triggered a miss-cluster for Sleeuwijk. What stands out is that all the other three extreme zones in Figure 23 fall in Zuid-Holland, the highest-populated province in the entire Netherlands. An explanation might be that irregularities or inconsistencies between the spatial characteristics and normalized trip production patterns are most likely a consequence of the high population density and the associated uncertainties.

		predicted labels			
		mix	*AP*	*MP*	∑
true labels	*mix*	219(41%)	115(21%)	205(38%)	539
	AP	64(22%)	181(64%)	41(14%)	286
	MP	126(30%)	35(8%)	260(62%)	421
	∑	409	331	506	1246

*Figure 20. Confusion matrix of provinces for recognizing TAZs in the clusters, i.e., either*AP, *MP, or mix.*

The spatial characteristics of the extreme cases in Figure 23 are in Table 1. Comparing these values with the mean and standard deviation of the clusters (i.e., true labels) in Table 2 gives insight into the miss-classification stimuli in extreme cases. For instance, the trip production heat map of $C_{AP,mix}$, Leiden West, in Figure 23 implies the dominant behavior of cluster AP. However, despite high densities in most spatial features and thus being similar to cluster AP, high density in Farm & Meadow triggered the miss-classification to *mix*.

Table 1. Spatial characteristics of the extreme TAZ in Figure 23.

TAZ	Zwijndrecht	Sleeuwijk	Leiden West	Zoetermeer Noord
True label	mix	mix	AP	MP
Predicted label	AP	MP	mix	mix
Commercial & Industrial land-use density (m^2/m^2)	0.074	0.005	0.062	0.014
Farm & Meadow land-use density (m^2/m^2)	0.023	0.089	0.011	0.064
Footway & Cycleway density (km/km^2)	2.54	0.52	6.56	5.99
Railway & service road density (km/km^2)	3.57	0.20	2.87	1.28
Parking garage density (Num/km^2)	0.37	0.12	0.23	0.29
Population density (Num/km^2)	952	214	1521	1220

Model structure and lack of post-analysis play a vital role in analyzing and alleviating the errors in *mix*. In our probabilistic OVR classification model, we decomposed the three-class problem into three independent binary class problems. Then, using a binary logistic loss function, we calculated the probability of each class. Each instance is then classified into the class of the highest probability. It is worth mentioning that the correct class might be the second highest probability, with a slight difference from the highest class. If we aim to analyze the more severe errors, we need to flag close probabilities and evaluate the confusion matrix afterward. Then the analysis can focus on more significant errors. In this regard, Figure 24 shows that in about 43% of the miss classified TAZs, associated with *mix* (i.e., either actual or predicted cluster is *mix*), less than 20% probability

difference is observed between the first and second most probable cluster. Therefore, these edge cases need a post-analysis step to exclude them from genuine prediction errors.

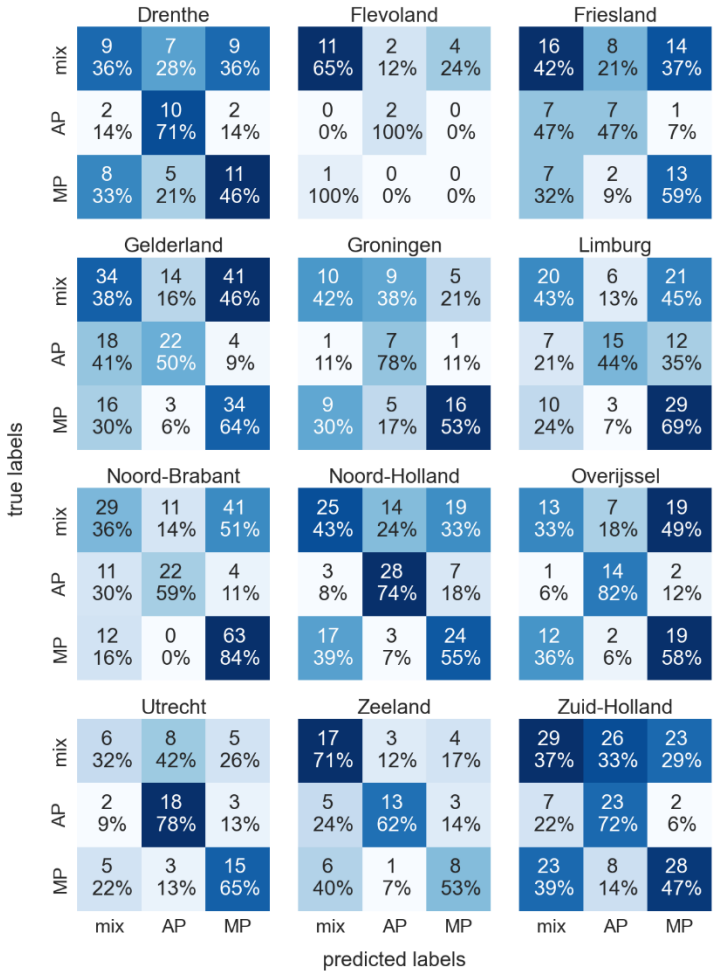


Figure 21. Confusion matrix of provinces for recognizing TAZs in the clusters, i.e., either AP, MP, or mix.

Figure 22. Province of Zeeland in the southwest border of the Netherlands.

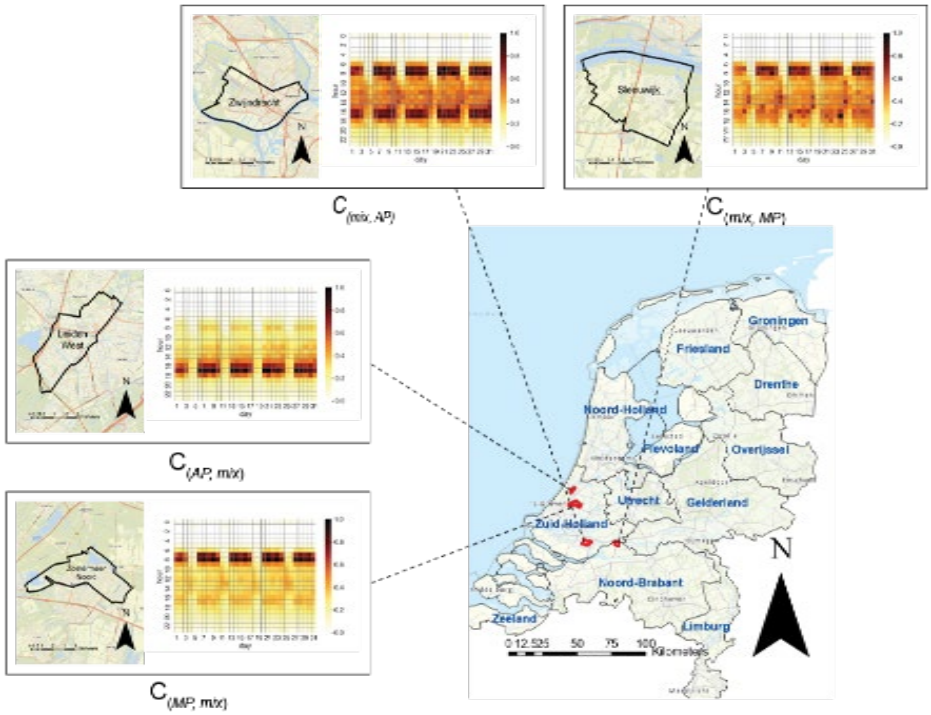


Figure 23. TAZs with the highest probability to be false positive and false negative of the mix cluster.

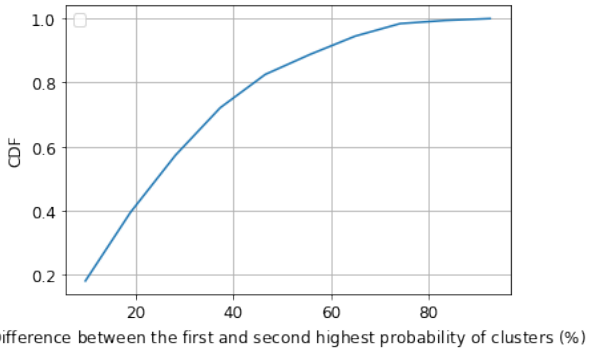


Figure 24. Cumulative density function (CDF) of the probability difference between the first and second most probable cluster. This visualization is on TAZs of the false positive and false negative of the mix cluster.

Table 2. Mean and standard deviation of Spatial characteristics of the true labels (i.e., clusters).

TAZ		*AP*	*MP*	*mix*
Commercial & Industrial land-use density (m^2/m^2)	mean	0.029	0.007	0.013
	std	0.047	0.010	0.019
Farm & Meadow land-use density (m^2/m^2)	mean	0.064	0.107	0.100
	std	0.068	0.075	0.077
Footway & Cycleway density (km/km^2)	mean	1.57	0.80	0.98
	std	1.63	0.96	1.17
Railway & service road density (km/km^2)	mean	1.18	0.35	0.5
	std	1.32	0.40	0.52
Parking garage density (Num/km^2)	mean	0.76	0.20	0.21
	std	1.67	0.49	0.68
Population density (Num/km^2)	mean	558	261	290
	std	834	497	439

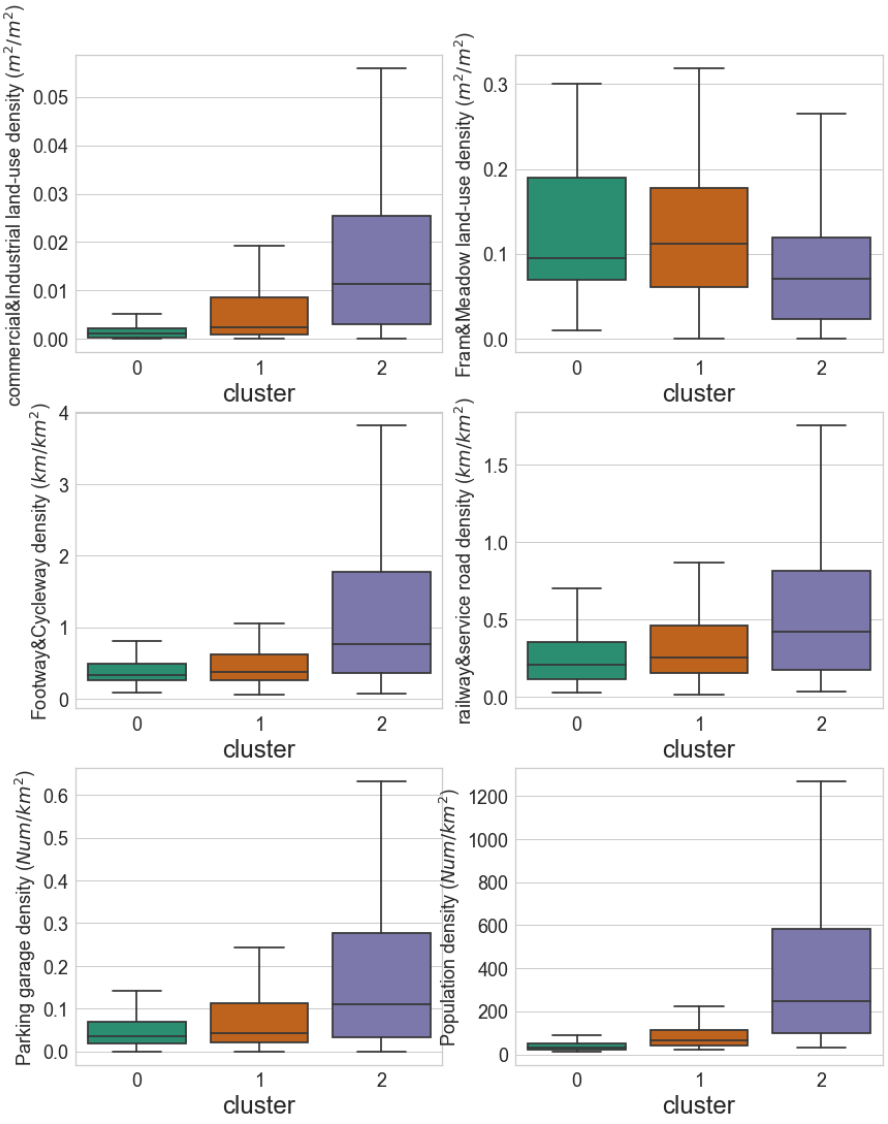


Figure 25. Comparing the spatial characteristics of mix cluster's sub-clusters.

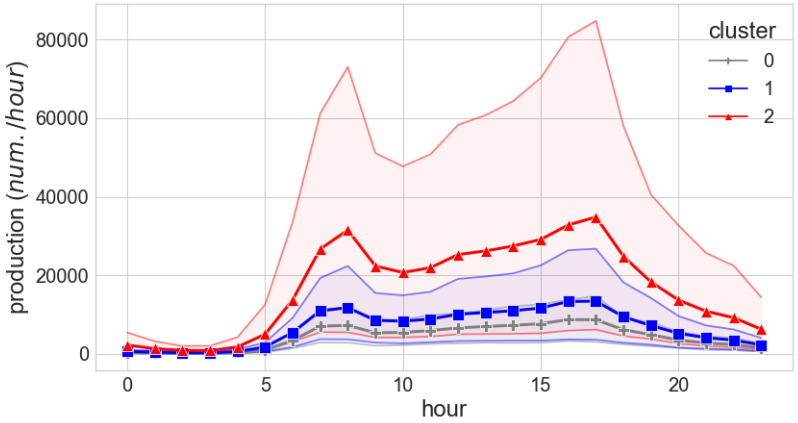


Figure 26. Average hourly production of the three sub-clusters of the mix cluster with 10th and 90th percentile as the shaded area.

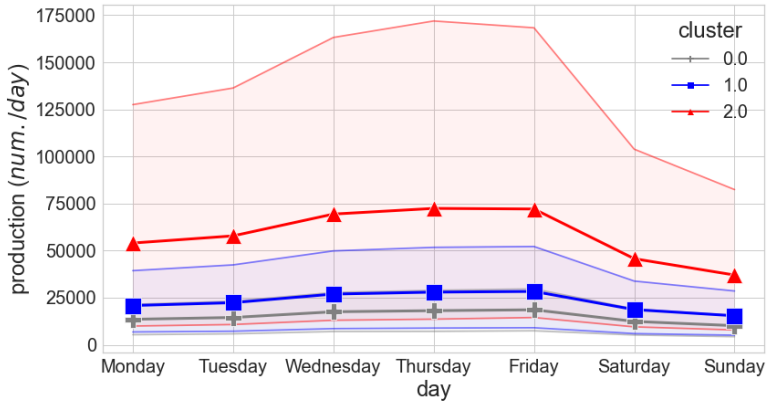


Figure 27. Average daily production of the three sub-clusters of the mix cluster with 10th and 90th percentile as the shaded area.

Having considered the above post-analysis, two main components seem to cause the identification errors in *mix*:

1. Feature selection plays a crucial task in improving the accuracy and preventing over-fitting of pattern recognition algorithms. To obtain the most informative feature subset, we select the important features and delete the unimportant features from the original subset. In this research, we only considered land-use features for our classification algorithm. Demographic information such as age, occupation, and income level is another feature group that can partially explain travel behaviors. For instance, a student city reveals a different dominant departure time pattern than a city with a higher population ratio of labour workers. As a result, the trip production pattern is also different based on the demographic profile of the TAZ. Another challenge regarding the feature selection is the dynamic relationships between the target classes(i.e., patterns) and candidate features. Therefore it is also challenging to measure the relationship between candidate features, the selected features, and target classes(i.e., patterns) in the feature selection process.
2. Spatial and temporal discretization problem is an inefficiency in the land-use data causing the inability to separate the different travel behavior patterns. The discretization step specifying the data resolution is essential for macroscopic modeling, deriving the necessary high-level insights for traffic planning and management, and reducing the computation time by decreasing the level of details. In this regard, two main factors must be addressed: Firstly, how much spatial variation is observable in each TAZ. Secondly, how often the land-use data is updated (i.e., the temporal variation)? For instance, investigating $C_{MP,mix}$ in Figure 23 reveals that in west side of *Zoetermeer Noord* a large area, called "De Nieuwe Driemanspolder", has turned into recreational meadow between 2017 and 2020. This overhaul adds up to the area density of *Farm&Meadow*, causing the trip production pattern to become *MP*. As the land-use data in this study does not seem to be updated frequently, this major change is not reflected, leading to errors in identifying trip production patterns. Information on spatial and temporal variations of land-use data helps one adjust the abstraction level for drawing relevant insights and design a traffic zoning system with a more crisp separation between different travel behaviors.

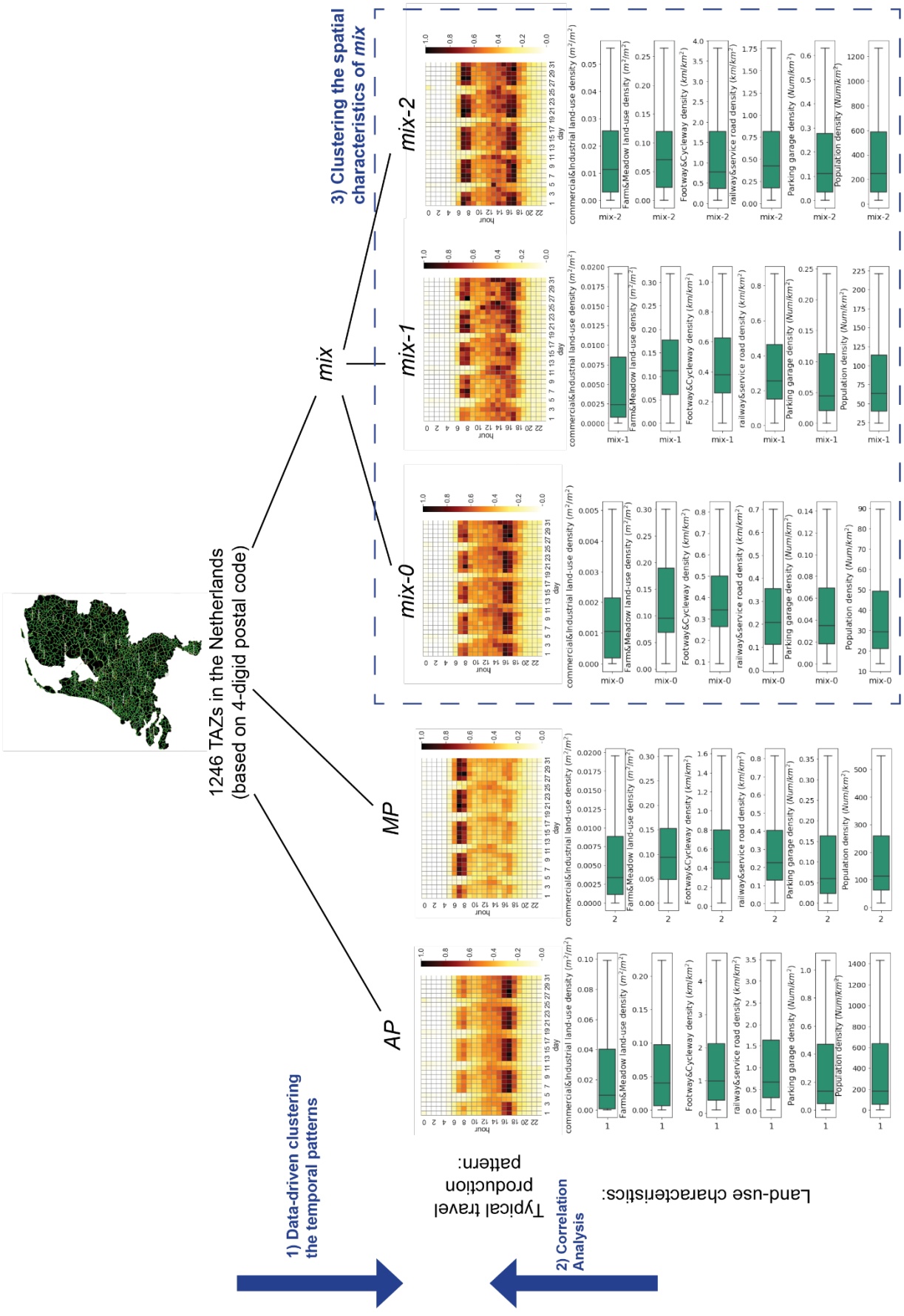


Figure 28. Overall findings of the study.

To investigate the spatial patterns of the **mix** cluster more closely, we applied hierarchical clustering on the land-use densities of all TAZs. Accordingly, we identified three main clusters with different spatial properties as shown in Figure 25. Sub-clusters {0, 1, 2} containing 68, 171, and 300 TAZs of *mix*, respectively. Sub-cluster 2 is more similar to *AP*, owing to higher densities of all the mentioned spatial characteristics except for *Farm&Meadow*. This similarity can also be observed in Figure 26 and 27, showing the average hourly and daily patterns in the three sub-clusters, respectively. On the other hand, sub-cluster 1 contains TAZs with a slightly higher median density of *Farm&Meadow* areas. Unlike cluster *MP*, which also had the highest density of *Farm&Meadow*, Sub-cluster 1 does not have the lowest densities in the other five variables. These areas might refer to the sub-urban residential TAZs connecting the metropolitan to rural areas. These areas show a morning peak indicating commuting travels and an afternoon peak due to non-commuting activities such as shopping, recreational, and social activities. Sub-cluster 0 shows fairly regular morning and afternoon peak with lowest population density and the median *Farm&Meadow* density lower than sub-cluster 1.

We recapped the overall findings of this research in Figure 28. The first clustering step resulted in three dominant temporal patterns in the trip production of all TAZs. Then we performed a correlation analysis to understand the association between spatial characteristics and the dominant patterns. More complex patterns inside *mix* induced us to perform the second clustering on the spatial characteristics of TAZs inside *mix*.

3.3 Research limitations

This paper introduces a data-driven framework to show day-to-day and within-day trip production patterns. Our method identifies the dominant temporal patterns without prior assumptions on the number of patterns. The last part of this research investigates the association of the identified patterns with spatial characteristics. Clearly, several limitations should be borne in mind when interpreting the results:

- Raw CDR can offer ample information about millions of mobile phone users. Nevertheless, their limitation concerns the privacy issue. Consequently, the data available for conducting this study was initially processed, aggregated, and transformed into origin-destination tables by another party. Inevitably, the level at which we receive the data lacks details on the methodology used for deriving the resulting dataset. For instance, the algorithms for separating motor vehicles from other modes of transport, scaling the OD data to the entire Dutch population, and mapping from the antenna Voronoi diagram to TAZs is unclear. Despite not fully understanding how the OD matrices were created, we used the data as given and focused on what we could learn from it. It's important to note that while we've done our best to explain the data, some of the finer details about how it was produced aren't clear to us. We accepted the matrices as they were, using them to guide our research and draw conclusions.
- Due to privacy reasons and ease of trip identification, the short-distance trips, mainly occurring inside each TAZ, were initially eliminated. Consequently, the present analysis only considers longer distance mobility with possibly more regular temporal patterns.
- In this research, we used origin-destination tables collected over only one month. This duration might not be sufficient to analyze the data's temporal stability. Furthermore, the resolution of the dataset is at the level of 4-digit postal code zones, which might be too coarse to evaluate the effect of special events on the trip production of TAZs.

3.4 Research implications

An implication of this study is the possibility of assessing the effect of developing/changing an urban area on trip production patterns. This can be particularly useful for decision and

policymakers. One possible direction for future research is using our framework to assess the effect of new urban areas on energy consumption and environmental footprint due to changes in temporal traffic patterns. Another benefit of the present study is that it introduces a framework for assessing the effect of special events on trip production patterns. To perform such an analysis, we need to select the spatiotemporal resolution of the data based on the extent of the event. Although this study focused on trip production patterns, one can apply a similar framework for trip attraction and OD demand.

4 Conclusion

The present study investigated the consistency and trustworthiness of the processed aggregated derivatives of GSM traces in the form of OD matrices. To this end, we analyzed the temporal patterns of travel production in TAZs of the Netherlands. The travel production of different areas reveals different temporal patterns within a day and between days. Clustering these patterns using a DCNN based on transfer learning with the K-means method introduced three distinct clusters in the Netherlands. One with a distinct afternoon peak, one with a distinct morning peak, and another with a mix of both. Further evaluation of the patterns shows robust temporal patterns in normalized production values. On the other hand, absolute values show a wider range of values for all the clusters reflecting both the regular and irregular patterns.

Another purpose of this research was to assess the association between temporal production patterns and spatial urbanization. In fact, observing these clusters in space and comparing them with the distribution (i.e., density) of land-use characteristics suggested different urbanization levels for each cluster: urban, rural, and mixed-level. An urban area mainly presents the city center with a high density of urban facilities, reveals a sharp afternoon production peak indicating more work-related activities. While a rural area shows more distinct production values in the morning, suggesting more residential land use. Mixed levels display other patterns in time and space. Further analysis of the mixed-level areas shows a more complex relationship between temporal heterogeneity and spatial characteristics. Population density seems to impose additional uncertainty on the temporal patterns. All in all, feature selection and Spatial and temporal discretization play essential roles in identifying the dominant trip production patterns.

Taken together, the results of this research support the idea that using land-use characteristics can improve the task of dynamic travel production prediction. Notwithstanding the relatively limited sample (one month of travel production at a specific spatial-temporal scale), this work establishes a quantitative framework for detecting hourly and daily temporal patterns and how spatial urbanization level helps in demand modeling. Moreover, such analysis is required before using the processed demand data for policy-making and network development. Further research could also be conducted to determine the effectiveness of using the land-use characteristics (on top of other variables) in improving the demand prediction models.

Acknowledgements

This research is sponsored by the NWO/TTW project MiRRORS under grant agreement 16270. Supplementary data associated with this article will be provided in the online version.

References

Anselin, L., & Griffith, D. A. (1988). Do spatial effects really matter in regression analysis? *Papers in Regional Science*, 65(1), 11–34.

Antoniou, C., Chaniotakis, E., Katrakazas, C., & Tirachini, A. (2020). A better tomorrow: towards human-oriented, sustainable transportation systems. *European Journal of Transport and Infrastructure Research*, 20(4), 354–361.

Atluri, G., Karpatne, A., & Kumar, V. (2018). Spatio-temporal data mining: A survey of problems and methods. *ACM Computing Surveys* (CSUR), 51(4), 1–41.

Bengfort, B., Bilbro, R., Danielsen, N., Gray, L., McIntyre, K., Roman, P., . . . others (2018). Yellowbrick. Retrieved from http://www.scikit-yb.org/en/latest/ doi: 10.5281/zenodo.1206264

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4) (No. 4). Springer.

Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical analysis*, 28(4), 281–298.

Camps-Valls, G. (2006). Kernel methods in bioengineering, signal and image processing. *Igi Global*.

Cardozo, O. D., García-Palomares, J. C., & Gutiérrez, J. (2012). Application of geographically weighted regression to the direct forecasting of transit ridership at station-level. *Applied Geography*, 34, 548–558.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Cheng, S., Lu, F., Peng, P., & Wu, S. (2019). Multi-task and multi-view learning based on particle swarm optimization for short-term traffic forecasting. *Knowledge-Based Systems*, 180, 116–132.

Cheng, T., Haworth, J., Anbaroglu, B., Tanaksaranond, G., & Wang, J. (2013). *Spatiotemporal data mining*.

Cohn, R., & Holm, E. (2021). Unsupervised machine learning via transfer learning and k-means clustering to classify materials image data. *Integrating Materials and Manufacturing Innovation*, 1–14.

Demchenko, Y., Grosso, P., De Laat, C., & Membrey, P. (2013). Addressing big data issues in scientific data infrastructure. In *2013 international conference on collaboration technologies and systems (cts)* (pp. 48–55).

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *In 2009 ieee conference on computer vision and pattern recognition* (pp. 248–255).

Deng, M., Yang, W., Liu, Q., Jin, R., Xu, F., & Zhang, Y. (2018). Heterogeneous space–time artificial neural networks for space–time series prediction. *Transactions in GIS*, 22(1), 183–201.

Ermagun, A., & Levinson, D. (2018). Spatiotemporal traffic forecasting: review and proposed directions. *Transport Reviews*, 38(6), 786–814.

Fekih, M., Bonnetain, L., Furno, A., Bonnel, P., Smoreda, Z., Galland, S., & Bellemans, T. (2021). Potential of cellular signaling data for time-of-day estimation and spatial classification of travel demand: a large-scale comparative study with travel survey and land use data. *Transportation Letters*, 1–19.

Fotheringham, A. S., Charlton, M. E., & Brunsdon, C. (1998). Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. *Environment and planning A*, 30(11), 1905–1927.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4), 367–378.

Gärling, T., Eek, D., Loukopoulos, P., Fujii, S., Johansson-Stenman, O., Kitamura, R., . . . Vilhelmson, B. (2002). A conceptual analysis of the impact of travel demand management on private car use. *Transport Policy*, 9(1), 59–70.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1), 100–108.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263-1284. doi: 10.1109/TKDE.2008 .239

Hong, J.-H., & Cho, S.-B. (2008). A probabilistic multi-class strategy of one-vs.-rest support vector machines for cancer classification. Neurocomputing, 71(16), 3275- 3281. Retrieved from https://www.sciencedirect.com/science/article/pii/S0925231208003007 (Advances in Neural Information Processing (ICONIP2006) / Brazilian Symposium on Neural Networks (SBRN 2006)) doi: https://doi.org/10.1016/j.neucom.2008.04.033

Hubert, L., & Arabie, P. (1985). Comparing partitions. Journal of classification, 2(1), 193–218.

Iglovikov, V., & Shvets, A. (2018). Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. *arXiv preprint arXiv:1801.05746*.

Jiang, S., Ferreira, J., & Gonzalez, M. C. (2017). Activity-based human mobility patterns inferred from mobile phone data: A case study of singapore. *IEEE Transactions on Big Data,* 3(2), 208–219.

Kaur, T., & Gandhi, T. K. (2020). Deep convolutional neural networks with transfer learning for automated brain image classification. *Machine Vision and Applications,* 31(3), 1–16.

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29(2), 115–129.

Lin, J.-J., & Shin, T.-Y. (2008). Does transit-oriented development affect metro ridership? evidence from taipei, taiwan. *Transportation Research Record*, 2063(1), 149–158.

Liu, X., Cheng, S., Zhang, X., Yang, X., Nguyen, T. B., & Lee, S. (2012). Unsupervised segmentation in 3d planar object maps based on fuzzy clustering. In *2012 eighth international conference on computational intelligence and security* (p. 364-368). doi: 10.1109/CIS.2012.88

Ma, X., Zhang, J., Ding, C., & Wang, Y. (2018). A geographically and temporally weighted regression model to explore the spatiotemporal influence of built environment on transit ridership. Computers, *Environment and Urban Systems*, 70, 113–124.

Maat, K., & Timmermans, H. (2006). Influence of land use on tour complexity: a Dutch case. *Transportation Research Record*, 1977(1), 234–241.

Mamei, M., Bicocchi, N., Lippi, M., Mariani, S., & Zambonelli, F. (2019). Evaluating origin–destination matrices obtained from *cdr data. Sensors*, 19(20), 4470.

Meppelink, J., Van Langen, J., Siebes, A., & Spruit, M. (2020). Beware thy bias: Scaling mobile phone data to measure traffic intensities. *Sustainability*, 12(9), 3631.

Ning, C., & Hongyi, Z. (2016). An optimizing algorithm of non-linear k-means clustering. *International Journal of Database Theory and Application*, 9(4), 97–106.

OpenStreetMap contributors, . (2017). Planet dump retrieved from https://planet.osm.org. Retrieved from https://www.openstreetmap.org

Pel, A., Bliemer, M., & Hoogendoorn, S. (2011). Modelling traveller behaviour under emergency evacuation conditions. *European Journal of Transport and Infrastructure Research*, 11(2).

Potera ̧s, C. M., Mih ̆aescu, M. C., & Mocanu, M. (2014). An optimized version of the k-means clustering algorithm. In *2014 federated conference on computer science and information systems* (pp. 695–699).

Rich, J., & Mabit, S. L. (2012). A long-distance travel demand model for europe. *European Journal of Transport and Infrastructure Research*, 12(1).

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20, 53–65.

Rubin, V., & Lukoianova, T. (2013). Veracity roadmap: Is big data objective, truthful and credible? *Advances in Classification Research Online*, 24(1), 4.

Schneider, C. M., Rudloff, C., Bauer, D., & González, M. C. (2013). Daily travel behavior: lessons from a week-long survey for the extraction of human mobility motifs related information. *In Proceedings of the 2nd acm sigkdd international workshop on urban computing* (pp. 1–7).

Shamos, M. I., & Hoey, D. (1975). Closest-point problems. *In 16th annual symposium on foundations of computer science* (sfcs 1975) (pp. 151–162).

Shen, X., Zhou, Y., Jin, S., & Wang, D. (2020). Spatiotemporal influence of land use and household properties on automobile travel demand. *Transportation Research Part D: Transport and Environment*, 84, 102359.

Srivastava, A. K., Safaei, N., Khaki, S., Lopez, G., Zeng, W., Ewert, F., . . . Rahimi, J. (2022). Winter wheat yield prediction using convolutional neural networks from environmental and phenological data. *Scientific Reports*, 12(1), 1–14.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2818–2826).

Szocska, M., Pollner, P., Schiszler, I., Joo, T., Palicz, T., McKee, M., . . . others (2021). Countrywide population movement monitoring using mobile devices generated (big) data during the covid-19 crisis. *Scientific reports*, 11(1), 1–9.

Thakuriah, P. (2001). Urban transportation planning: A decision-oriented approach. *Journal of Transportation Engineering*, 127(5), 454–454.

van Elteren, T. (2018). A comparative study of human engineered features and learned features in deep convolutional neural networks for image classification (Unpublished doctoral dissertation).

Van Gansbeke, W., Vandenhende, S., Georgoulis, S., Proesmans, M., & Van Gool, L. (2020). Scan: Learning to classify images without labels. In *European conference on computer vision* (pp. 268–285).

Wang, C., Chen, D., Hao, L., Liu, X., Zeng, Y., Chen, J., & Zhang, G. (2019). Pulmonary image classification based on inception-v3 transfer learning model. *IEEE Access*, 7, 146533–146541.

Xu, X., Chen, A., & Yang, C. (2017). An optimization approach for deriving upper and lower bounds of transportation network vulnerability under simultaneous disruptions of multiple links. *Transportation research procedia*, 23, 645–663.

Yabe, T., Jones, N. K., Rao, P. S. C., Gonzalez, M. C., & Ukkusuri, S. V. (2022). Mobile phone location data for disasters: A review from natural hazards and epidemics. *Computers, Environment and Urban Systems*, 94, 101777.

Yang, Z., Franz, M. L., Zhu, S., Mahmoudi, J., Nasri, A., & Zhang, L. (2018). Analysis of washington, dc taxi demand using gps and land-use data. *Journal of Transport Geography*, 66, 35–44.

Yeung, K. Y., Haynor, D. R., & Ruzzo, W. L. (2001). Validating clustering for gene expression data. *Bioinformatics*, 17(4), 309–318.

Zhai, J., Qi, J., & Shen, C. (2022). Binary imbalanced data classification based on diversity oversampling by generative models. Information Sciences, 585, 313- 343. Retrieved from https://www.sciencedirect.com/science/article/pii/S0020025521011804 doi: https://doi.org/10.1016/j.ins.2021.11.058

Appendix A

This part discusses more details on the XGBoost algorithm (we refer the readers to (Chen & Guestrin, 2016) for more details).

XGBoost is an ensemble of decision trees; it consists of sequentially developed decision trees where each tree works to improve the performance of the prior tree (Srivastava et al., 2022). Ensemble methods generally try to reduce the bias or variance of several weak learners by combining them into a strong learner (i.e., a learner with low bias and variance). Boosting is an ensemble method in which weak learners are fitted sequentially and aggregated to the ensemble model. In each step, the training set is updated to focus more on the weakness of the current ensemble. In other words, each model in the sequence does the fitting by giving higher weight to misclassified data points. If the weak learner of each step depends on the gradient direction of the loss function at each step, this method is also called Gradient Boosting Machines (GBM) (Friedman, 2002). The advantage of XGBoost over non-extreme gradient boosting methods is the regularization term in the loss function, which helps prevent over-fitting.

Assume our data set is $\chi = \{(x_i, y_i): i = 1, \dots, n, x_i \in \mathbb{R}^m, y_i \in \mathbb{R}\}$; hence we have n observations each having m features corresponding to their associated label y. Then $\hat{y}_i$ can be defined as a result of an ensemble, with T additive functions, represented by the generalized model as follows:

$$\hat{y}_i = \Phi(x_i) = \sum_{t=1}^{T} f_t(x_i), \quad (2)$$

where f_t is a decision tree, and $f_t(x_i)$ is the score given by the $t-th$ decision tree to the $i-th$ data point. The objective function that needs to be minimized to select the function f_t consists of two terms of training loss, $L(y_i, \hat{y}_i)$ and *regularization*, $\Omega(f_t)$:

$$obj_{\Phi} = \sum_i L(y_i, \hat{y}_i) + \sum_t \Omega(f_t) \quad (3)$$

The training loss, L, estimates the model's goodness of fit based on the training data. A common form of L for classification, which is used in this research, is the logistic loss (i.e., binary logistic) for $y \epsilon \{0, 1\}$ (Bishop & Nasrabadi, 2006):

$$L_{Logistic} = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(p_i) + (1 - y_i)\log(1 - p_i)), \quad (4)$$

where y_i is the true value, $p_i \epsilon [0, 1]$ denotes the probability prediction, and N is the number of samples. An ideal classifier has a logistic loss close to zero.

In order to prevent the model from becoming too complex, Ω applies the penalty as follows:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda||\omega||^2, \quad (5)$$

Where γ controls the penalty for the number of leaves, T and λ is the parameter for controlling the magnitude of leaf weights ω in the decision tree. The purpose of having the regularization term in the objective function is to simplify the model and prevent over-fitting.

Learning the tree structure is more difficult than the traditional optimization problem, where you can simply take the gradient. In other words, training all the trees simultaneously is not a straightforward task; therefore, XGBoost uses an additive method that optimizes the learned tree and adds a tree at each step. In the $t-th$ iteration, we need to add the following f_t, which minimizes the objective function:

$$obj^t = \sum_{i=1}^{n} L\left(y_i, \hat{y}_i{}^{t-1} + f_t(x_i)\right) + \Omega(f_t). \quad (6)$$

This function can be simplified and approximated by the Taylor expansion:

$$obj^t \simeq \sum_{i=1}^{n}\left[L\left(y_i, \hat{y}_i{}^{t-1}\right) + g_i f_i(x_i) + \frac{1}{2}h_i f_i^2(x_i)\right] + \Omega(f_t), \quad (7)$$

Where the functions g_i and h_i, the first and second order gradient of the loss function, are defined as follows:

$$g_i = \partial_{\hat{y}_i{}^{t-1}} L\left(y_i, \hat{y}_i{}^{t-1}\right), \quad (8)$$

$$h_i = \partial^2_{\hat{y}_i{}^{t-1}} L\left(y_i, \hat{y}_i{}^{t-1}\right). \quad (9)$$

We can rewrite Equation 7 by expanding Ω and find the optimal output value (i.e., weight) $\bar{\omega}_j$ for leaf j as follows:

$$\bar{\omega}_j = \frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda}, \quad (10)$$

Where I_j is the instance set of leaf j. Replacing Equation 10 and 5 in 7 gives us the following optimal value of the loss function which is used as a similarity score for measuring the quality of each tree structure:

$$\widetilde{obj}^t = -\frac{1}{2}\sum_{j=1}^{T} \frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T. \quad (11)$$

For binary logistic loss function we use for the classification, $g_i = -(y_i - p_i)$ (i.e., the residual), and $h_i = p_i(1 - p_i)$ which can be replaced in Equation 10 and 11.

To simplify evaluating tree structure when adding new branches to the tree (i.e., evaluating the split candidates), a greedy algorithm is used. This algorithm starts from one single leaf and adds new branches to the tree iteratively. Therefore, after the tree splits from a given node, the formula for loss reduction (i.e., gain) is as follows:

$$obj_{split} = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i + \lambda} + \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma, \quad (12)$$

Where IL and IR are subsets of the available observations in the left and right nodes after the split. I is the subset of the available observations in the current node so that $I = I_L \cup I_R$. Moreover, the tree structure will continue to split if obj_{split} is positive or other criteria are met, such as the maximum depth of a tree that users need in XGBoost parameters fine-tuning.

Equation 12 is used for finding the best split at any node and it only depends on g_i, and h_i (i.e., the first and second order gradient) of the training loss and the regularization parameter γ. Therefore, as long as the first and second-order gradient is provided, XGBoost can optimize any custom loss function.

XGBoost performs better than other tree boosting algorithms due to (i) having the regularization term for preventing the over-fitting, (ii) downscaling of each new tree by a constant parameter τ to reduce the impact of a single tree on the final model, i.e., it gives the future trees more space to improve the model while reducing the impact of the current tree. Moreover, (iii) XGBoost supports column sampling, which means each tree is built using a subset of the columns from the training dataset.