A Classification Approach Using Machine Learning for Predicting Traffic Flows in Areas with Missing Sensors

Martina Fazio
Department of Physics and Astronomy
University of Catania, Catania, Italy, 95125
Email: martina.fazio@phd.unict.it

Piergiorgio Vitello
Department of Engineering
University of Luxembourg, Esch-sur-Alzette, Luxembourg, 4365
Email: piergiorgio.vitello@uni.lu

Juan Pineda-Jaramillo
Department of Engineering
University of Luxembourg, Esch-sur-Alzette, Luxembourg, 4365
Email: juan.pineda@uni.lu

Richard D. Connors
Department of Engineering
University of Luxembourg, Esch-sur-Alzette, Luxembourg, 4365
Email: richard.connors@uni.lu

Francesco Viti
Department of Engineering
University of Luxembourg, Esch-sur-Alzette, Luxembourg, 4365
Email: francesco.viti@uni.lu

Word Count: 6,851 words + 2 tables = 7,351 words

Submitted 31th of July

ABSTRACT

The high rate of urbanization and the ever-increasing number of vehicles on roads require multiple efforts in order to correctly identify and evaluate interventions to improve transport systems. This phenomenon pushes cities to invest in new technologies and increasingly advanced support tools. In the era of Big Data, the exploitation of innovative data can allow the resolution of various issues related to the world of transport planning. However, data are not always available, e.g. in many rural areas, and even in the cities of developing countries. In such cases, it is possible to resort to sophisticated models able to fill in the lack of information. In this respect, this paper focuses on the prediction of traffic flows in areas where no data are available. To do this, we develop a classification approach using machine learning (ML) models involving available, but limited, datasets. After the prediction of traffic flows, a validation and similarity analysis have been performed. From the latter it is possible to identify sensor devices and regions that best represent those areas with no data. Moreover, a SHAP analysis has been carried out to identify the impact of features associated to the traffic volume. The main finding of this study is the developing of a support tool able to aid government agencies by providing an exploratory overview of what should be the traffic volume in areas with poor data. We believe that the proposed methodology has the potential of becoming a first-step analysis in the demand forecasting process.

Keywords: Traffic flow prediction, machine learning, data extraction, data analysis

INTRODUCTION

Transport systems have long been the basis of the economic development of cities. Inefficient management of transport networks reflects negative consequences on the entire community. The rapid increase in traffic volumes generates concrete repercussions in terms of pollution, accidents, and road congestion, constituting a serious risk for urban sustainability and livability. For this reason, the prediction of traffic flows plays a fundamental role for providing an idea to what extent this phenomenon negatively affects system performance (*1*). Due to the complexity of transport networks, it is crucial to resort to advanced models and tools able to provide support to the already consolidated theoretical models. In this respect, it is worthy to underline the important role of ML models, especially for the prediction of transport systems features (*2*). Furthermore, the reliability of these models for predicting and simulating the traffic states is strictly depending on the type of data involved in the modeling process. In recent years, new challenges have arisen on how to analyze and address these issues with the support of new technologies. The digitalization can help in providing a large amount of data to monitor transport networks in order to find new solutions and bringing greater efficiency to the transport system. These data can be collected by different sources such as GPS devices, sensors, social networks, smartphones, quickly providing huge datasets constituting of different types of data (structured, unstructured, semi-structured). Precisely with the adjectives velocity, volume, and variety (3Vs model) a first definition of Big Data was given (*3*).

The use of Big Data has totally transformed the world of transport thanks to the spread of Internet of Things (IoT) devices, giving greater importance to Intelligent Transport Systems (ITS) (*4*). In the transport field, the exploitation of Big Data is now widespread: OD matrix estimation (*5*, *6*), mobility pattern analysis (*7*, *8*), travel mode detection (*9*, *10*), road safety analysis (*11*, *12*). In this respect, Big Data can provide an important aid for both governments and private companies to do in-depth analyses, monitoring, make decisions, and improve transport network' performance (*13*).

The complexity of Big Data requires many efforts to manage the datasets with the aim of obtaining information that can suggest strategic decisions. For this reason, the world of Big Data analysis is now exploiting ML models, which take advantage of this data to identify patterns and predict future events (*2*). The basic concept of ML models is the capability of learning from data without prior knowledge on disparate application fields. From a transportation point of view, ML models have been involved for route optimization analysis (*14–16*), travel behavior (*17–19*), traffic flows prediction (*13*, *20–22*). Regarding this latter, it is worthy to underline the importance of conducting this type of analysis in a more automatic and reliable way, without resorting to manual counting in the fields (*23*). The purpose of this paper is framed in this context, i.e. the prediction of traffic flows through a classification approach using ML models with free available datasets. The case study is the Grand-Duchy of Luxembourg, in which sensors devices detecting traffic flows are installed and the relative data are freely accessible by the "Portail des Travaux Publics" of Luxembourg government website (*24*). Sensors are localized almost homogeneously around Luxembourg except for Luxembourg city. For this reason, we decided to develop a methodology capable of predicting traffic flows also in areas without traffic sensors using the information available from another dataset retrieved through TomTom move platform (sample dataset) (*25*) which provides information also for Luxembourg city but in a limited time period (i.e. January 2020).

The main challenge of this study is the prediction of flows in an area without counting devices. Research studies that have dealt with a similar topic usually focused attention on the forecasting of flows in short/long-term (*20–22*, *26–28*). To the best of our knowledge, there are no studies that addressed how to predict traffic flows in absence of other data sources for the study area under consideration. The methodology is capable of providing valuable information for government agencies about the most representative areas and counters useful to obtain a first-step exploratory overview of flows in poor data regions, without resorting to count fields or data purchasing.

BACKGROUND

Research on the prediction of traffic flows is rooted in several studies starting from the '70s (*29–32*). One of the first efforts to predict traffic flows has been carried out by Nicholson and Swann (*33*), who elaborated a short-term prediction of traffic flows across the Mersey tunnel (Liverpool) using spectral analysis and

collecting data from inductive loop traffic detectors. The methodology provided reasonable results, but it was not able to make predictions for an entire day. Lu (*34*) proposed an adaptive prediction system which consists of a set of iterative mathematical steps to minimize the mean squared error and then maximize the accuracy of the prediction of traffic flows. Despite the results of these studies are promising, the model seems to have no general applicability, but it is closely linked to the case study considered. An analogous study is the one conducted by Williams et al. (*35*). They integrated seasonal AutoRegressive Integrated Moving Average (ARIMA) analysis with Winters exponential smoothing models to forecast seasonal traffic flow in urban freeway. The aim was to compare the performance level of the analysis proposed compared to other similar approaches (e.g. k-Nearest Neighbor, Artificial Neural Network, and historical average models).

The accuracy of the prediction models increases with the amount of data that are engaged in the analysis process. The spread of new technologies has generated many sources of mobility data (e.g. sensors, smartphones) that make available a large amount of real-time information that enhances the quality of the analysis carried out (*36–38*). Exploiting innovative data for prediction analysis is extensively addressed in the literature. Moghaddam and Hellinga (*39*), developed a data-driven model for the prediction of travel times in arterial roadways using Bluetooth detectors data, Iqbal et al. (*40*) elaborated a combined optimization-based approach with a microscopic traffic simulation analysis for OD matrices estimation, collecting data from mobile phones and traffic counts devices. Similarly, the study of Cantelmo and Viti (*37*) also demonstrated the effectiveness of the combined use of mobile phone data and Floating Car Data (FCD) for the estimation of OD matrices. The importance of exploiting innovative data is also underlined by Di Donna et al. (*41*), who carried out a user mobility pattern analysis gathering data from Call Detail Records (CDR).

Over the years, pushed by the technologies progresses and the advent of Big Data, research moved to develop methodologies based on more sophisticated ML models. There are several studies that have dealt with the prediction of traffic flows adopting a ML approach. Chikaraishi et al. (*20*) focused on traffic prediction during disastrous events, conducting an exploratory analysis to test the applicability of a set of ML models. Lv et al. (*21*) used a Stacked Autoencoder model (SAE) for the estimation of traffic flows of roads with different levels of traffic loads. Yi and Bae (*22*) proposed a Deep Neural Network (DNN) model in order to predict and analyze link-based traffic conditions. In this study the authors focused on distinguishing the different levels of congestion using real-time traffic data. A similar study was conducted by Xu et al. (*42*) focusing on bikesharing trips dynamic forecasting. Through the use of a deep learning model, authors gained accurate predictions in terms of produced and attracted trips in different time intervals. Other studies moved in this direction using different ML models: Fuzzy-neural model, Deep Belief Networks, Long-Short-Term Memory network (LSTM), Convolutional Neural Network (CNN) , k-Nearest Neighbor, among others (*26*, *43–46*).

This paper proposes a classification approach using ML models able to predict traffic flows in areas where no data are available, and the only way to obtain it would require field surveys or data purchase. This is an aspect that has not been investigated in the literature, and this study is aiming at bridging this gap. Bauer et al. (*47*) conducted a similar study dealing with the estimation of O/D matrices based on flow count data from sensors without resorting to a prior O/D matrices. Combining least squares with maximum entropy methods authors proposed a statistical tool to calibrate path flows even in absence of prior knowledge. However, the data that the authors used have been collected by devices installed inside the study area.

METHODOLOGY

In this section we will describe the data and the methodology adopted to predict traffic flows within areas in the absence of sensor devices, through a classification approach using ML models with insufficient datasets. **Figure 1** shows the framework steps of the proposed methodology that will be thoroughly detailed below.

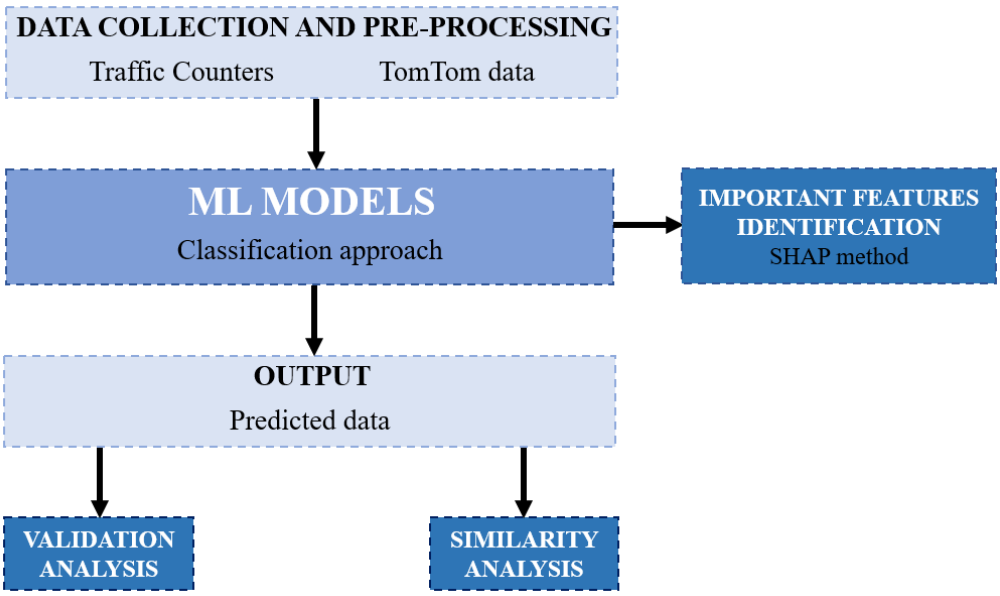


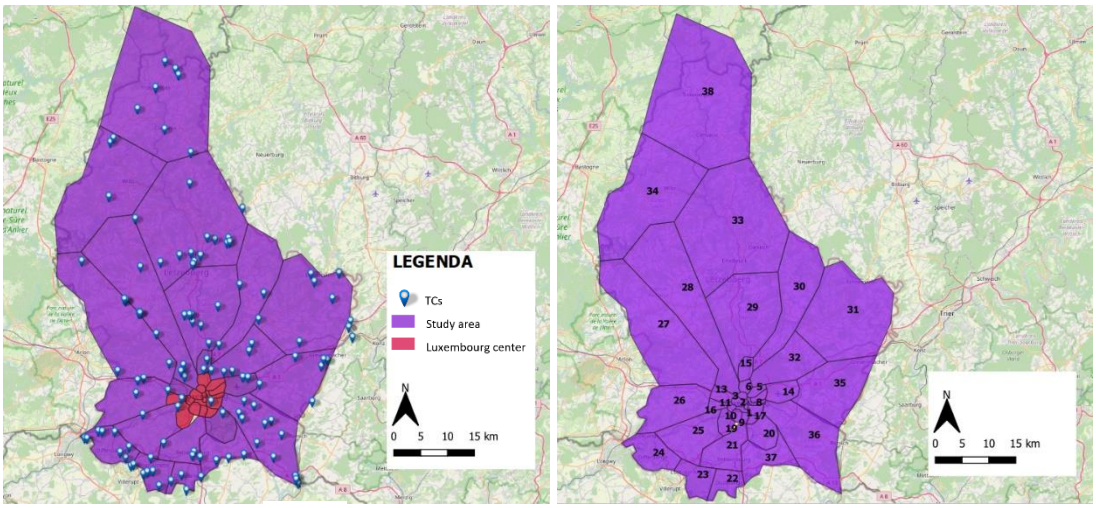
Figure 1 Flowchart of the proposed methodology

Data collection and pre-processing

Data has been collected from two data sources. The first collected data is provided by the "Portail des Travaux Publics" of the Luxembourg government website. This open data provides car flows detected by 135 traffic counters (TCs) distributed throughout the country of Luxembourg (**Figure 2a**). TCs are classified according to the name and type of road in which they are localized (e.g., Highway, National, Secondary Roads, and mobile counting). Each TC collects daily traffic flows per hour for the two road directions, which we classified as "inflow" (i.e., toward the center) and "outflow" (i.e., away from the center). Nevertheless, TCs have a different time availability depending on the year in which the sensor was created, besides this data can be only accessed with a limit of the past 5 years.

The second collected data comprises a sample of the traffic flows provided by the TomTom Move platform. TomTom collects FCD information from different devices: smartphones, car navigators, black boxes. This data includes information on flows into and out from different user-defined regions, with a time range of 3 hours. The sample, freely downloadable from the platform, consists of data for the month of January 2020. For our analysis we divided the study area (i.e. the Grand-Duchy of Luxembourg) into 38 regions (**Figure 2b**). We also collected aggregate traffic flows coming from outside, into the study area.

As can be seen from **Figure 2a** the TCs are distributed throughout the country except for the center-south (Luxembourg City), where just one TC is located. Notwithstanding, TomTom data provides information also for the center regions (**Figure 2b**), the main issue with using this information is the limited time period in which is possible to collect the data, whereas the datasets from TCs are available for any calendar date. In this respect, the purpose of this study is to develop a ML model able to combine both datasets in order to predict traffic flows in areas in which there are not TCs (i.e., Luxembourg center). **Table 1** presents the composition of the final datasets after the pre-processing analysis.



(a) (b)

Figure 2 TCs' location in the study area (a); study area zoning used for the retrieval of TomTom data (b)

The two datasets were pre-processed using common processes traditionally performed in data mining (e.g., creating dummy variables for categorical features, normalizing the numerical features by using the z-score standardization method, among others) (*48*). Moreover, some additional features were created using the features of the dataset (e.g., time_range (ranges of 3 hours) in the Traffic_counters dataset, in order to merge both datasets) and the two datasets were merged using 4 matching features existing in both datasets (i.e., region_of_origin, region_of_destination, day, time_range). This resulted in a unique dataset with 101,872 rows and 19 columns. Some features have many "null values" for different reasons (e.g., no available data for some TCs, no available volume data for some regions or some TCs, etc.). The target feature, volume in TC, had a right-skewed distribution, with a range between 0 and 3732 vehicles, a mean of 230, a standard deviation of 306 and a median of 136 vehicles. Therefore, we decided to adopt a classification approach by creating the "volume_counter_ranges" feature and selecting 12 volume ranges to the initial "volume_counter" feature based on the distribution of their values. Therefore, the target feature (i.e., volume_counter_ranges) has the following ranges (in vehicles/hour): (a) 0, (b) 1-50, (c) 51-100, (d) 101-200, (e) 201-300, (f) 301-400, (g) 401-500, (h) 501-600, (i) 601-700, (j) 701-800, (k) 801-900, (l) more than 900 (see **Table 1**). By setting specified input features, the model will predict the range of traffic volume in the counter.

TABLE 1 Composition of the final dataset. 101,872 observations.

Feature	**Description**	**type**	**null values**
TARGET: volume_ranges in Traffic_counters	Traffic flow recorded in the counter [vehicles] in 12 ranges	categorical	71824
ID_counter	ID of the counter	categorical	52744
time	time of the registered volume in the Traffic_counters dataset	categorical	52744
direction	direction of the traffic flow(Inflow, Outflow)	categorical	52744
region_localization_counter	region where is located the traffic counter	categorical	52744
route_name	name of the road in which the traffic counter is located (e.g. A7, N2, CR184, etc..)	categorical	52744

Feature	Description	type	null values
route_type	type of the road in which the traffic counter is located (Autoroute, Route nationale, Chemin repris, Comptage mobile)	categorical	52744
region_of_origin	region of origin of the traffic flows (based on the zoning shown in Figure 2b)	categorical	0
region_of_destination	region of destination of the traffic flows (based on the zoning shown in Figure 2b)	categorical	0
day	day of the week of the registered volume	categorical	0
time_range (ranges of 3 hours)	time range (range of 3 hours) of the registered volume	categorical	0
volume in TomTom_move	volume registered in TomTom_move dataset in the range of 3 hours (time_range)	numerical	0
peak_time	weight indicating if the time is peak time or not	numerical	0
night	weight indicating if the time is night or not	numerical	0
regions_OD	Feature indicating the OD flow	categorical	0
sum_volume_counter_origin_tomtom	This feature adds all the volume for the same origin region in the same day and the same time_range	numerical	0
sum_volume_counter_destination_tomtom	This feature adds all the volume for the same destination region in the same day and the same time_range	numerical	0

ML models

The data was divided in three subsets (external, middle, center) according to the regions presented in **Figure 3**. This zoning has been performed with the aim of using the external regions data to train the ML models, and the middle regions data to validate the models. Then the model's output will be used to perform a final validation using the single existing TC in Luxembourg city center (i.e., id: 400) and finally the best model will be applied to predict the traffic volume in the latter regions.

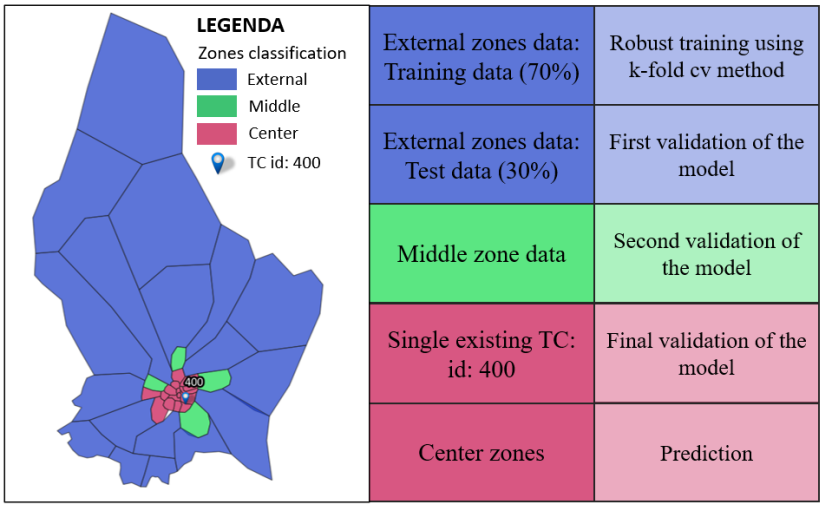


Figure 3 Area division for the validation analysis

With the aim of predicting traffic volume ranges in Luxembourg city in absence of TCs' data, we selected the subset of external regions to train the models. To achieve this goal, data from external regions was randomly split into training data and test data (70% and 30%, respectively) where each subdivision was composed by the traffic volume ranges and the input features in the same proportion for both subdivisions. The *SMOTE* method was used to handle imbalanced classes within the target feature (*49*). A set of 11 ML models were trained to predict the traffic volume ranges: Logistic Regression (LR), K

Neighbors Classifier (KNN), Naive Bayes (NB), Decision Tree (DT), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Ada Boost (ADA), Extra Trees (ET), Random Forest (RF), Gradient Boosting (GB) and Light Gradient Boosting Machine (lightGBM). These models, whose detailed descriptions can be found in numerous studies (*17, 18, 50–55*), were selected since they have been widely and successfully used across a variety of classification problems.

The stratified k-fold cross-validation method has been implemented to validate the trained models. This method is commonly used to assess the performance of classification models performed, thanks to its capability of reducing any bias produced by the models (*56*). Moreover, each ML model has a set of hyperparameters that need to be tuned in order to improve its performance. This process, commonly known as "hyperparameter tuning", is carried out by implementing the random search method, which allows assessing the values of the hyperparameter with larger impact on model performance (*56, 57*). To assess the performance of the models, the Area Under the Receiver Operating Characteristics (*ROC*) Curve, generally known as *AUC*, the Recall, Precision and F1-score evaluation metrics were employed as loss functions. Additionally, the learning curve method was implemented to identify possible *underfitting* or *overfitting* problems in the models by analyzing the model performance as more observations are used in the training process (*58*).

Validation analysis

After analyzing the ML models according to the evaluation metrics, the best model was chosen based on the confusion matrix, where it is possible to analyze the performance of each model in its prediction of each range. Then, once we obtained the predicted traffic volumes from the best ML model, we still need to analyze the accuracy of such results. As shown in **Figure 3** the validation phase involves the data of the middle regions in order to compare the predicted traffic volumes with the real ones (i.e., the data collected by TCs) and then the process is also repeated for the single existing TC in Luxembourg center. Therefore, we establish the goodness of our model in the test data of the external regions, in the middle regions, and in one TC in Luxembourg center.

Identification of important features

After identifying the best model based on the evaluation metrics and the confusion matrices, the impact of the features associated to the traffic volume range was obtained using the SHapley Additive exPlanation method (SHAP) (*59*). This method obtains the importance and the direct impact of each feature in order to better interpret the results of the model (*60*), and allows to identify the direct effect and the magnitude of the contributions of each input feature in the traffic volume ranges.

Similarity analysis

Similarity analysis is focused on identifying similarity trends between predicted values and the data collected by TCs. This analysis allows identification of the most representative existing TCs compared to predicted traffic flows in the center regions. The aim is to depict the inflow and the outflow traffic volumes from Luxembourg City center only using the TCs in the surrounding areas. The first step of this phase is to divide the real TCs data into two categories, depending on which type of road they are installed: highways or national roads. For both categories, we compute the similarity between the predicted TCs in center regions and the real TCs. To measure similarity between two temporal profiles of TCs, we exploit the symmetric index of Jensen-Shannon divergence (JSD) that outperforms the asymmetric Kullback-Leibler divergence (KLD) that always returns a finite value. The similarity of two TCs *i* and *j* is calculated through **Equation 1**:

$$J\left(D_i, D_j\right) = H\left(\frac{D_i + D_j}{2}\right) - \left(\frac{H(D_i) + H(D_j)}{2}\right) \quad (1)$$

where J represents the divergence of two temporal profiles, H is the Shannon entropy (*61*) and D is the temporal profile of a TC. The similarity can assume values in the range [0-1]. The value 0 represents the

maximum similarity (e.g., the temporal pattern of a TC with itself) and 1 represents the maximum divergence. In our dataset we have 8 TCs from highways, for each of them we compute an average value of similarity with all the TCs in center regions. The similarity for the *k*-th TC located in a highway (S_h) is computed as follows (**Equation 2**):

$$S_h = \frac{\sum_{k=1}^{N} J(D_h, D_k)}{N} \quad (2)$$

where J is the JSD, and N is the number of TCs predicted in the center. From the similarity values, we were able to detect the highway TCs that have the highest similarity with the predicted traffic flows, and consequently we can identify which TCs are the most representative for the traffic flows of center regions.

Regarding the TCs installed on National roads, since these are the majority in our dataset (89), we choose a different approach by combining TCs for a region and hence defining the similarity at a regional level. The similarity for a region *r* (S_r) is defined as follow (**Equation 3**):

$$S_r = \frac{\sum_{i=1}^{K_r} S_i}{K_r} \quad (3)$$

where S_r represents the similarity value between a TC and the center TCs described in **Equation 2**, and K_r is the number of TCs belonging to region *r*. Looking at the similarity values at the regional level we can identify which areas have traffic trends similar to center regions, this tell us where we should focus more in order to detect the general traffic flows of center regions.

RESULTS AND DISCUSSION

This section provides the results of the models to predict the traffic flows within a region in the absence of devices' data, through a classification approach using ML models. The process involved four main stages: (a) training a set of 11 ML models to identify the best in terms of the aforementioned evaluation metrics; (b) validation of the models using middle regions' data and the single TC in Luxembourg center; (c) identification of the direct effect and importance of input features in the predicted traffic volume ranges; (d) application of the best model to Luxembourg center in order to perform the similarity analysis.

ML models results

The ML models were trained with the training data and implementation of the k-fold cross-validation method (with k=10). Evaluation metrics were then obtained using the test data described. **Table 2** summarizes the results of the initial models: lightGBM, RF and ET are the best models in terms of the evaluation metrics. Therefore, the hyperparameter tuning method was applied using the random search method in order to tune the performance of these models, analyzing the behavior of their learning curves to ensure they are correct (i.e., the necessary clear convergence trend between training and cross-validation scores, allowing to foresee if adding more observations to the training of these models will probably improve their performance, decreasing the risks of *overfitting* and *underfitting*). Those results are also presented in **Table 2**.

TABLE 2 Results of the ML models assessed using the test data.

Initial results				
Model	**AUC**	**Recall**	**Precision**	**F1-score**
Light Gradient Boosting (lightGBM)	0.9478	0.5933	0.7166	0.7140

Random Forest (RF)		0.9446	0.6020	0.7446	0.7425
Extra Trees (ET)		0.9166	0.6075	0.7473	0.7458
Logistic Regression (LR)		0.9012	0.4965	0.5988	0.5584
K Neighbors Classifier (KNN)		0.9005	0.5881	0.7068	0.6933
Gradient Boosting (GB)		0.8948	0.4843	0.5888	0.5753
Decision Tree (DT)		0.8873	0.5999	0.7412	0.7391
Linear Discriminant Analysis (LDA)		0.8849	0.4511	0.5711	0.5126
Naive Bayes (NB)		0.7064	0.2683	0.4755	0.2091
Ada Boost (ADA)		0.5936	0.1791	0.3530	0.2025
Quadratic Discriminant Analysis (QDA)		0.5023	0.0881	0.2861	0.0685
Tuned versions after hyperparameter tuning					
Model	**Behavior of the Learning Curve**	**AUC**	**Recall**	**Precision**	**F1-score**
Initial lightGBM	***Good***	***0.9478***	***0.5933***	***0.7166***	***0.7140***
Tuned lightGBM	Good	0.9569	0.6066	0.7456	0.7441
Initial RF	Good	0.9446	0.6020	0.7446	0.7425
Tuned RF	Good	0.8909	0.5126	0.6132	0.5809
Initial ET	Good	0.9166	0.6075	0.7473	0.7458
Tuned ET	Bad	0.8688	0.4740	0.5690	0.5031

Validation of the models in all regions

After the ML analysis, we created confusion matrices for the six best performing models in order to select the single best model, based on its predictive performance for each range. **Figure 4** shows the confusion matrices of the best model, the Initial lightGBM model. LightGBM is an open-source gradient boosting framework initially developed by Microsoft, which has gained popularity due to its advantages in training speed, high efficiency and accuracy, and low memory usage, among others. The gradient boosting model is a tree-based learning algorithm that combines numerous weak-learner models to build a strong-learner model, where every single weak-learner model meaningfully improves the performance of the entire model (*62*, *63*). Moving to the description of the matrices, in **Figure 4** each cell contains the number of occurrences of a predicted class when testing true inputs. The legend bars colors represent the percentage of correct predicted occurrences over the total of true class values. The columns show predicted class values, while the sum of all values in each row indicates the total occurrences for each such class. The diagonal presents the occurrences of correct predictions for each class, and the accuracy is the sum of all elements on the diagonal divide by all elements of the matrix. In more detail, **Figure 4a** shows the performances of the classification algorithm for external regions; as explained earlier the model was trained on the data of these regions. In this case, the matrix depicts a satisfactory accuracy for this area (64.8%), and most of the predicted occurrences are on the diagonal, except for class 0, in which occurrences are often predicted as class 1-50. The reason is clearly connected with the length of this class that includes only one value. Analyzing the middle regions, **Figure 4b** shows that the total accuracy for this area (49.7%) is lower than for the external regions. Despite the accuracy, this result is relevant for this study since mispredictions of our algorithm are often between classes close to each other. Since the number of classes is rather high, and the focus of the analysis is to estimate the general trend of the TCs, errors between neighboring classes is acceptable. Note that the highest number of prediction errors occur for higher classes (more than 500), this is likely due to having fewer occurrences of high traffic volumes, so the model is less trained for those values. The last confusion matrix in **Figure 4c** focuses on the center regions. In this area we have only one real TC, therefore the comparison of true and predicted traffic volumes is based on very few values. The accuracy for this area (43.3%) is similar to the one obtained from middle regions with the difference that

in the matrix we have some classes with 0 true occurrences, such as the class 0 and the class >900. As we noticed for the middle regions's predictions, also for center areas it is clear how the model finds more difficulties with higher volumes. The matrix shows us how accuracy decreases as traffic volumes increase.

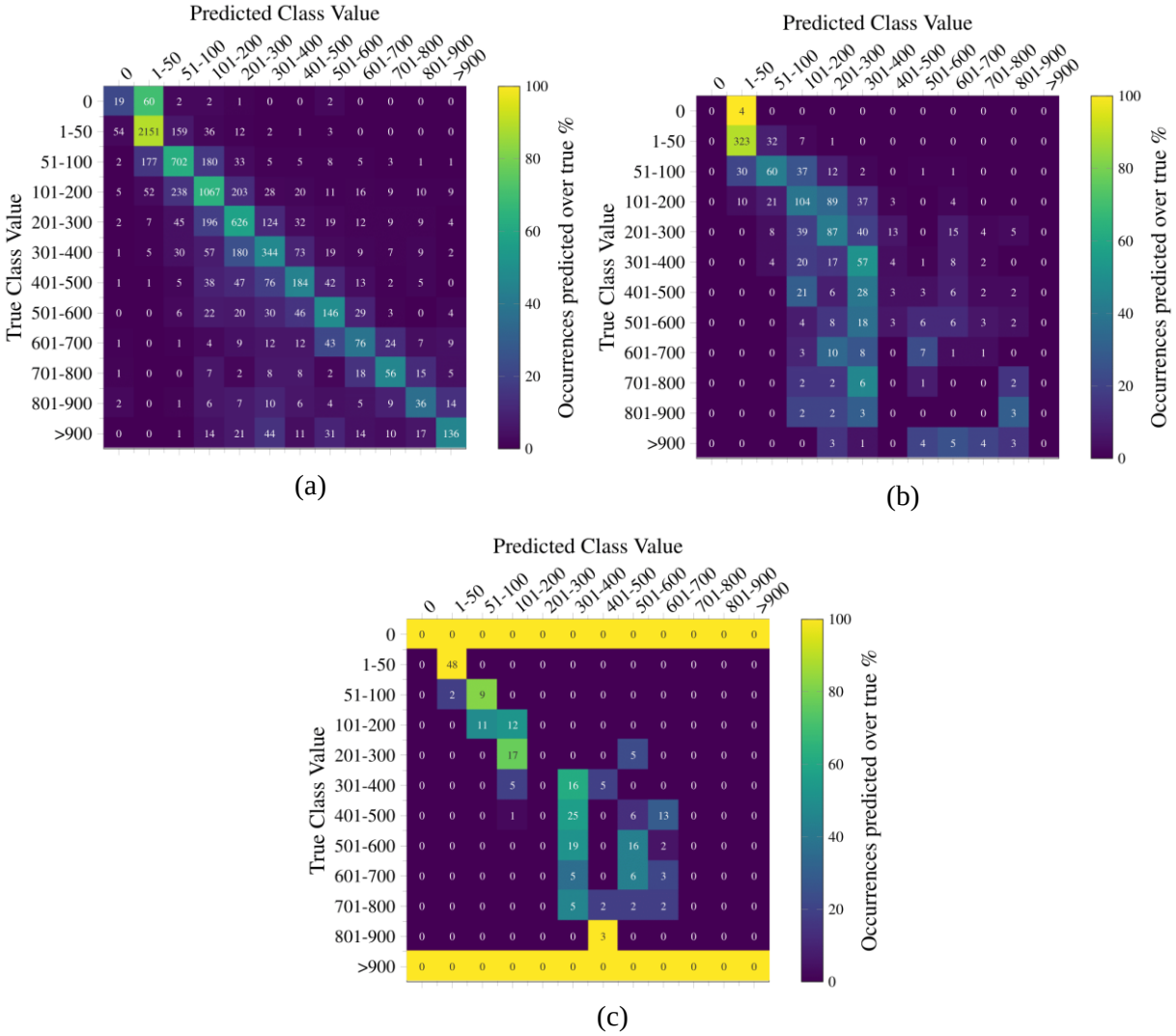
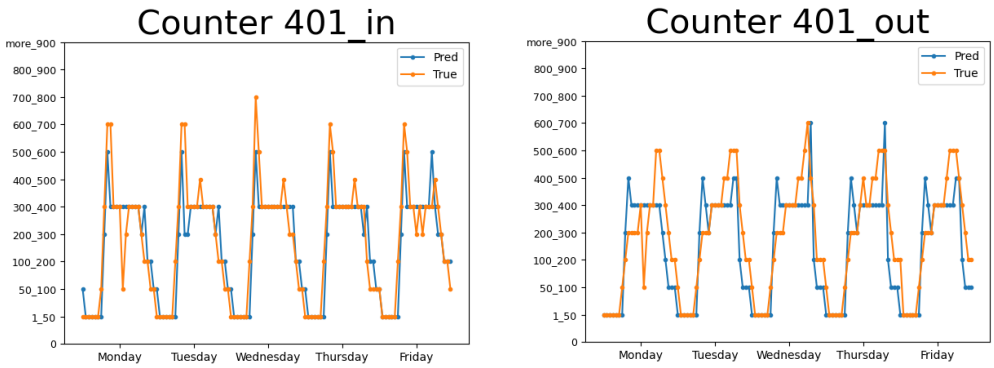


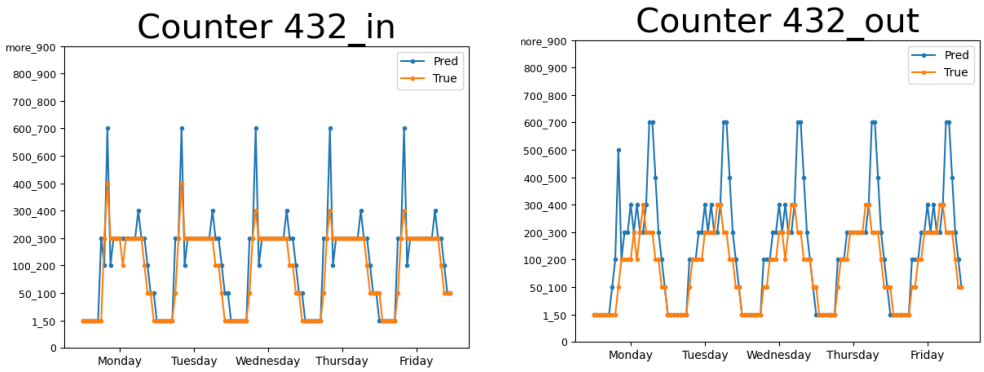
Figure 4 Confusion matrices for external (a), middle (b) and center (c) regions created using the Initial lightGBM model

To show the results in **Figure 4** more intuitively, below we present the predicted temporal profile at TC level. First, we order the prediction output by TCs, time, and direction of TCs, then we compare the predicted temporal profiles with the real ones. **Figure 5** shows the results for two TCs (ids: 401 and 432) that are located in middle region. The plots are divided by direction of the TC (inflow and outflow). From figure **Figure 5a**, we can observe how the predicted traffic volumes of TC 401 resemble the true values for both directions, especially the inflow plot shows high accuracy throughout the week. It is interesting to notice that the prediction are accurate also on different day profiles, e.g. we can see how the model is able to detect the differences between the Monday and the Friday profile. **Figure 5b** presents the temporal profile of TC 432 in which the prediction results are characterized by a lower accuracy compared to TC 401, especially for the outflow direction. Despite this, we can notice how for both directions the general trends are similar, the peaks occur in the same time intervals, while what our model is not able to estimate correctly

is the intensity of the peaks. For instance, the 432 Inflow plot shows that the maximum peaks reach the 400-500 class, while the predicted values indicate higher volumes (class 600-700).



(a)



(b)

Figure 5 Comparison between predicted and true traffic flows for TCs 401 (a), and 432 (b) in middle regions

Identification of the importance of the input features in the predicted traffic volume ranges

Figure 6 presents the importance of the most relevant input features shown in **Table 2** for the traffic volume ranges in Luxembourg and the direct impact of each one on the output magnitude for the initial lightGBM model. All features are presented in descending order according to the impact of the features on the output of the model, and it is also possible to see the magnitude of the prediction power of the feature for each traffic volume range. According to this, the results suggest that indicators for night, peak time, and road type are the three features that have a greatest impact on predicting traffic volume in Luxembourg. It is also possible to visibly identify the different impact of each feature for each range (e.g., night has a greater impact for traffic volume ranges below 50 vehicles/hour, than for the other ranges; while peak time has a greater impact on the traffic volume ranges greater than 700 vehicles/hour). Moreover, it is interesting to note that among road types (Highway, National road, Secondary road, and mobile counting), Highways ("Autoroute") and National roads ("Route nationale") seem to be the most important to identify the traffic volume range.

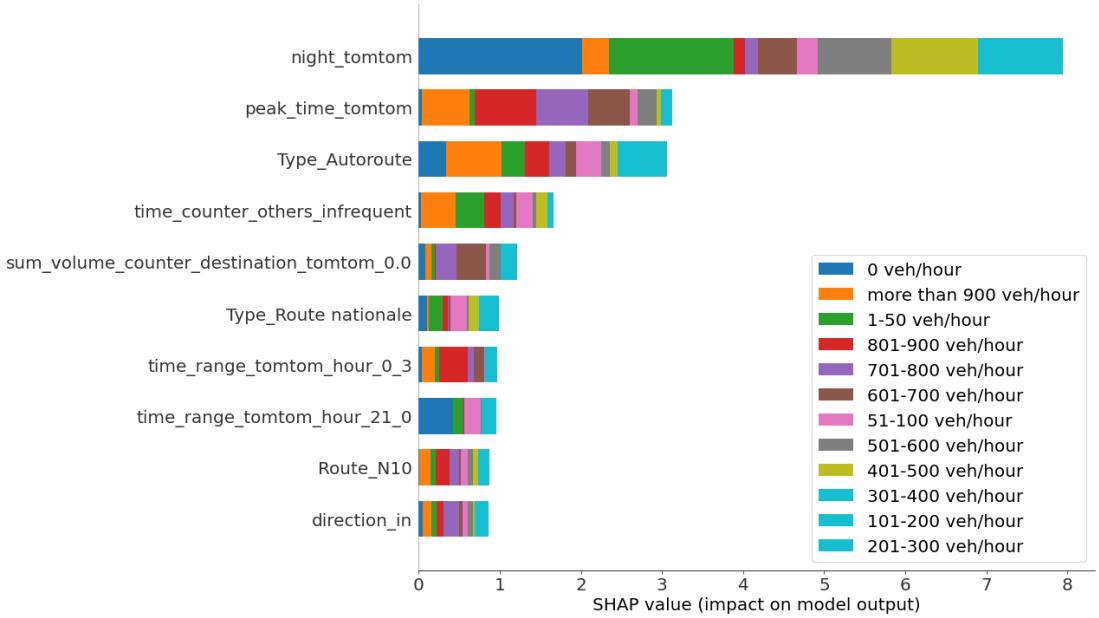


Figure 6 SHAP plot with the global feature importance in the lightGBM model

Similarity analysis

Based on the findings shown in **Figure 6** we decided to focus our attention on TCs installed on highways and national roads. In particular, a similarity analysis has been performed in order to identify TCs whose data are the most similar to those estimated for the central regions. **Figure 7** shows the results of the similarity analysis for highways TCs. Following the color bar, the lighter TCs are the higher is the similarity level. From this type of analysis, it is possible to understand which TC provides the most similar information compared with the predicted inflows (**Figure 7a**) and outflows (**Figure 7b**) in the central areas. For national road TCs, we considered the sum of flows for each TC in each region, to identify the most similar regions (external and middle) compared to center regions **Figure 8**. Similarly to the case of highway TCs, in this way it is possible to infer which are the most representative regions of the center in terms of inflow (**Figure 8a**) and outflow (**Figure 8b**). More precisely, in order to obtain a first overview of the amount of inflows and outflows in center regions, this analysis is able to select which TCs, or regions contain the most similar traffic flows value to those predicted for central areas. Such analysis can be framed as a support tool for government agencies offering two services: (i) providing information about which TCs can be considered the most representative for a fist overview of the possible traffic flows of the target area; (ii) the regional similarity values can be useful to suggest the best area where to locate new TCs that can better describe the traffic flows of the target area.

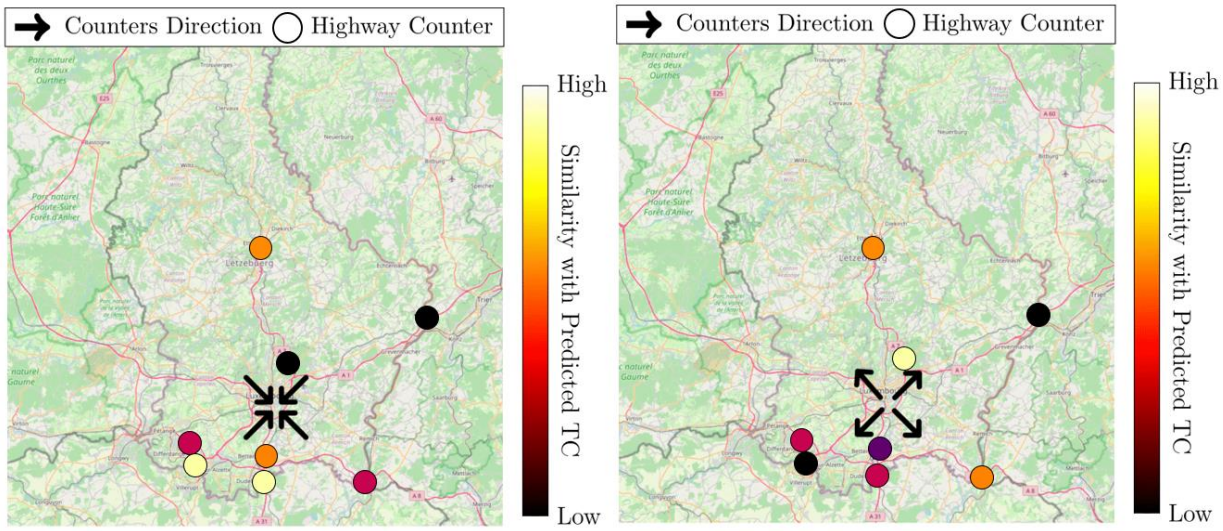


Figure 7 Similarity analysis between highway TCs and the predicted flows in center regions considering inflow (a) and outflow (b) direction

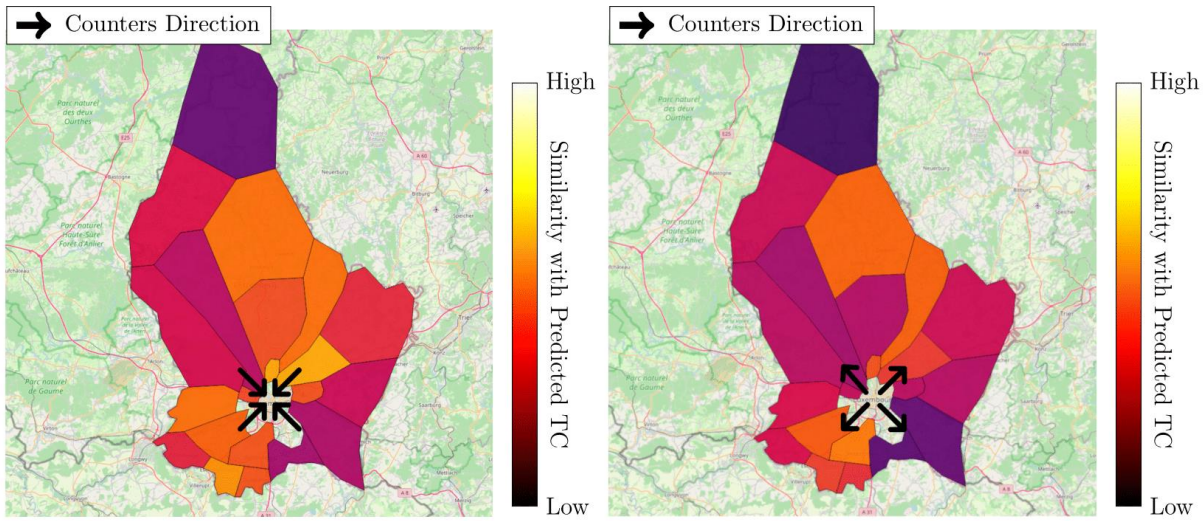


Figure 8 Similarity analysis between the sum of national roads TCs data for each region and the predicted flows in center regions considering inflow (a) and outflow (b) direction

CONCLUSIONS

The prediction of traffic flows plays a fundamental role in the management of cities and in transport planning; the increasing availability of Open Data information (*64*) is empowering rapid development of new approaches. In this respect, this paper aims at developing a methodology capable of predicting traffic flows in areas with missing sensors devices using free but limited open datasets whose extracted information has been used to train ML models for prediction, and for validation. Subsequently, SHAP and similarity analysis have been used to identify the most important features and the most representative TCs and regions, to predict traffic flows in center regions. From the analysis carried out we demonstrate that the model is able to predict traffic flows with an acceptable accuracy level. The final purpose is to provide to

government agencies a first-step exploratory approach to gain an overview of the traffic flows in areas with little or no other data sources, suggesting also the most appropriate location (i.e., road type and regions) where to install new TCs. However, the methodology has some limitations that can be investigated in further research, e.g. finding data from other data sources in order to have more heterogeneous features at a microscopic level that can improve the accuracy level of the predicted flows (e.g. travel times, path-choice).

ACKNOWLEDGMENTS

Mrs. Fazio is partially supported by EU4EU fund. Mr. Vitello is supported by the Luxembourg National Research Fund (PRIDE17/12252781/DRIVEN).

AUTHOR CONTRIBUTIONS

The authors confirm contribution to the paper as follows: study conception and design: M. Fazio, P. Vitello, J. Pineda-Jaramillo, R. D. Connors, F. Viti; data collection: M. Fazio, P. Vitello, J. Pineda-Jaramillo; analysis and interpretation of results: M. Fazio, P. Vitello, J. Pineda-Jaramillo; draft manuscript preparation: M. Fazio, P. Vitello, J. Pineda-Jaramillo. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

1. Chen, D. Research on Traffic Flow Prediction in the Big Data Environment Based on the Improved RBF Neural Network. *IEEE Transactions on Industrial Informatics*, 2017. https://doi.org/10.1109/TII.2017.2682855.
2. Bhavsar, P., I. Safro, N. Bouaynaya, R. Polikar, and D. Dera. Machine Learning in Transportation Data Analytics. In *Data Analytics for Intelligent Transportation Systems*.
3. Laney, D. 3D Data Management: Controlling Data Volume, Velocity, and Variety. *Application Delivery Strategies*, 2001.
4. Zantalis, F., G. Koulouras, S. Karabetsos, and D. Kandris. A Review of Machine Learning and IoT in Smart Transportation. *Future Internet*.
5. Alexander, L., S. Jiang, M. Murga, and M. C. González. Origin-Destination Trips by Purpose and Time of Day Inferred from Mobile Phone Data. *Transportation Research Part C: Emerging Technologies*, 2015. https://doi.org/10.1016/j.trc.2015.02.018.
6. Cantelmo, G., and F. Viti. A Big Data Demand Estimation Framework for Multimodal Modelling of Urban Congested Networks. 2019.
7. González, M. C., C. A. Hidalgo, and A. L. Barabási. Understanding Individual Human Mobility Patterns. *Nature*, 2008. https://doi.org/10.1038/nature06958.
8. Wang, F., J. Wang, J. Cao, C. Chen, and X. (Jeff) Ban. Extracting Trips from Multi-Sourced Data for Mobility Pattern Analysis: An App-Based Data Example. *Transportation Research Part C: Emerging Technologies*, 2019. https://doi.org/10.1016/j.trc.2019.05.028.
9. Gong, H., C. Chen, E. Bialostozky, and C. T. Lawson. A GPS/GIS Method for Travel Mode Detection in New York City. *Computers, Environment and Urban Systems*, 2012. https://doi.org/10.1016/j.compenvurbsys.2011.05.003.
10. Shafique, M. A., and E. Hato. Travel Mode Detection with Varying Smartphone Data Collection Frequencies. *Sensors (Switzerland)*, 2016. https://doi.org/10.3390/s16050716.
11. Strauss, J., L. F. Miranda-Moreno, and P. Morency. Mapping Cyclist Activity and Injury Risk in a Network Combining Smartphone GPS Data and Bicycle Counts. *Accident Analysis and Prevention*, 2015. https://doi.org/10.1016/j.aap.2015.07.014.
12. Fazio, M., N. Giuffrida, G. Inturri, and M. Ignaccolo. Combining GPS-Tracks and Accident Data to Improve Safety of Cycling Paths. 2020.
13. Mohammed, O., and J. Kianfar. A Machine Learning Approach to Short-Term Traffic Flow Prediction: A Case Study of Interstate 64 in Missouri. 2019.
14. Yang, J., Y. Han, Y. Wang, B. Jiang, Z. Lv, and H. Song. Optimization of Real-Time Traffic Network Assignment Based on IoT Data Using DBN and Clustering Model in Smart City. *Future Generation Computer Systems*, 2020. https://doi.org/10.1016/j.future.2017.12.012.
15. Fusco, G., C. Colombaroni, L. Comelli, and N. Isaenko. Short-Term Traffic Predictions on Large Urban Traffic Networks: Applications of Network-Based Machine Learning Models and Dynamic Traffic Assignment Models. 2015.
16. Yu, J., G. L. Chang, H. W. Ho, and Y. Liu. Variation Based Online Travel Time Prediction Using Clustered Neural Networks. 2008.
17. Hagenauer, J., and M. Helbich. A Comparative Study of Machine Learning Classifiers for Modeling Travel Mode Choice. *Expert Systems with Applications*, 2017. https://doi.org/10.1016/j.eswa.2017.01.057.
18. Zhao, X., X. Yan, A. Yu, and P. Van Hentenryck. Prediction and Behavioral Analysis of Travel Mode Choice: A Comparison of Machine Learning and Logit Models. *Travel Behaviour and Society*, 2020. https://doi.org/10.1016/j.tbs.2020.02.003.
19. Pineda-Jaramillo, J. Travel Time, Trip Frequency and Motorised-Vehicle Ownership: A Case Study of Travel Behaviour of People with Reduced Mobility in Medellín. *Journal of Transport and Health*, 2021. https://doi.org/10.1016/j.jth.2021.101110.
20. Chikaraishi, M., P. Garg, V. Varghese, K. Yoshizoe, J. Urata, Y. Shiomi, and R. Watanabe. On the

Possibility of Short-Term Traffic Prediction during Disaster with Machine Learning Approaches: An Exploratory Analysis. *Transport Policy*, 2020. https://doi.org/10.1016/j.tranpol.2020.05.023.
21. Lv, Y., Y. Duan, W. Kang, Z. Li, and F. Y. Wang. Traffic Flow Prediction with Big Data: A Deep Learning Approach. *IEEE Transactions on Intelligent Transportation Systems*, 2015. https://doi.org/10.1109/TITS.2014.2345663.
22. Yi, H., J. Heejin, and S. Bae. Deep Neural Networks for Traffic Flow Prediction. 2017.
23. Wang, S. J., and P. Moriarty. Can New Communication Technology Promote Sustainable Transport? 2017.
24. Portail Des Travaux Publics. https://travaux.public.lu/fr/infos-trafic/comptage.html.
25. TOMTOMmove. https://move.tomtom.com/login.
26. Soua, R., A. Koesdwiady, and F. Karray. Big-Data-Generated Traffic Flow Prediction Using Deep Learning and Dempster-Shafer Theory. 2016.
27. Polson, N. G., and V. O. Sokolov. Deep Learning for Short-Term Traffic Flow Prediction. *Transportation Research Part C: Emerging Technologies*, 2017. https://doi.org/10.1016/j.trc.2017.02.024.
28. Wu, Y., H. Tan, L. Qin, B. Ran, and Z. Jiang. A Hybrid Deep Learning Based Traffic Flow Prediction Method and Its Understanding. *Transportation Research Part C: Emerging Technologies*, 2018. https://doi.org/10.1016/j.trc.2018.03.001.
29. Baras, J. S., and W. S. Levine. ESTIMATION OF TRAFFIC FLOW PARAMETERS IN URBAN TRAFFIC NETWORKS. 1977.
30. Chang, M. F., and D. C. Gazis. TRAFFIC DENSITY ESTIMATION WITH CONSIDERATION OF LANE-CHANGING. *Transportation Science*, 1975. https://doi.org/10.1287/trsc.9.4.308.
31. Okutani, I., and Y. J. Stephanedes. Dynamic Prediction of Traffic Volume through Kalman Filtering Theory. *Transportation Research Part B*, 1984. https://doi.org/10.1016/0191-2615(84)90002-X.
32. Kawashima, H. Long Term Prediction of Traffic Flow. *IFAC Proceedings Volumes*, 1987. https://doi.org/10.1016/s1474-6670(17)55876-0.
33. Nicholson, H., and C. D. Swann. The Prediction of Traffic Flow Volumes Based on Spectral Analysis. *Transportation Research*, 1974. https://doi.org/10.1016/0041-1647(74)90030-6.
34. Lu, J. Prediction of Traffic Flow by an Adaptive Prediction System. *Transportation Research Record*, No. 1287, 1990.
35. Williams, B. M., P. K. Durvasula, and D. E. Brown. Urban Freeway Traffic Flow Prediction: Application of Seasonal Autoregressive Integrated Moving Average and Exponential Smoothing Models. *Transportation Research Record*, 1998. https://doi.org/10.3141/1644-14.
36. Fazio, M., M. Le Pira, G. Inturri, and M. Ignaccolo. Bus Rapid Transit vs. Metro. Monitoring on-Board Comfort of Competing Transit Services via Sensors. 2020.
37. Cantelmo, G., and F. Viti. A Big Data Demand Estimation Model for Urban Congested Networks. 2020.
38. Cantelmo, G., P. Vitello, B. Toader, Antoniou, and F. Viti. Inferring Urban Mobility and Habits from User Location History. 2020.
39. Moghaddam, S. S., and B. Hellinga. Real-Time Prediction of Arterial Roadway Travel Times Using Data Collected by Bluetooth Detectors. *Transportation Research Record*.
40. Iqbal, M. S., C. F. Choudhury, P. Wang, and M. C. González. Development of Origin-Destination Matrices Using Mobile Phone Call Data. *Transportation Research Part C: Emerging Technologies*, 2014. https://doi.org/10.1016/j.trc.2014.01.002.
41. Di Donna, S. A., G. Cantelmo, and F. Viti. A Markov Chain Dynamic Model for Trip Generation and Distribution Based on CDR. 2015.
42. Xu, C., J. Ji, and P. Liu. The Station-Free Sharing Bike Demand Forecasting with a Deep Learning Approach and Large-Scale Datasets. *Transportation Research Part C: Emerging Technologies*, 2018. https://doi.org/10.1016/j.trc.2018.07.013.
43. Yin, H., S. C. Wong, J. Xu, and C. K. Wong. Urban Traffic Flow Prediction Using a Fuzzy-Neural

Approach. *Transportation Research Part C: Emerging Technologies*, 2002. https://doi.org/10.1016/S0968-090X(01)00004-3.
44. Hochreiter, S. Long Short-Term Memory. Vol. 1780, 1997, pp. 1735–1780.
45. Jo, D., B. Yu, H. Jeon, and K. Sohn. Image-to-Image Learning to Predict Traffic Speeds by Considering Area-Wide Spatio-Temporal Dependencies. *IEEE Transactions on Vehicular Technology*, 2019. https://doi.org/10.1109/TVT.2018.2885366.
46. Munoz-Organero, M., R. Ruiz-Blaquez, and L. Sánchez-Fernández. Automatic Detection of Traffic Lights, Street Crossings and Urban Roundabouts Combining Outlier Detection and Deep Learning Classification Techniques Based on GPS Traces While Driving. *Computers, Environment and Urban Systems*, 2018. https://doi.org/10.1016/j.compenvurbsys.2017.09.005.
47. Bauer, D., G. Richter, J. Asamer, B. Heilmann, G. Lenz, and R. Kölbl. Quasi-Dynamic Estimation of OD Flows from Traffic Counts Without Prior OD Matrix. *IEEE Transactions on Intelligent Transportation Systems*, 2018. https://doi.org/10.1109/TITS.2017.2741528.
48. Pineda-Jaramillo, J. D. A Shallow Neural Network Approach for Identifying the Leading Causes Associated to Pedestrian Deaths in Medellín. *Journal of Transport and Health*, 2020. https://doi.org/10.1016/j.jth.2020.100912.
49. Xue, W., and J. Zhang. Dealing with Imbalanced Dataset: A Re-Sampling Method Based on the Improved SMOTE Algorithm. *Communications in Statistics - Simulation and Computation*, Vol. 45, No. 4, 2016, pp. 1160–1172. https://doi.org/10.1080/03610918.2012.728274.
50. Shafiq, M., Z. Tian, A. K. Bashir, A. Jolfaei, and X. Yu. Data Mining and Machine Learning Methods for Sustainable Smart Cities Traffic Classification: A Survey. *Sustainable Cities and Society*, Vol. 60, 2020, p. 102177. https://doi.org/10.1016/j.scs.2020.102177.
51. Gao, K., Y. Yang, T. Zhang, A. Li, and X. Qu. An Extrapolation-Enhanced Approach for Modeling Travel Decision Making: Integrating Ensemble Machine Learning with Knowledge-Based Decision-Making Theory. *Knowledge-Based Systems*, 2021, p. 106882. https://doi.org/10.1016/j.knosys.2021.106882.
52. Ali, M. PyCaret: An Open Source, Low-Code Machine Learning Library in Python.
53. Martínez Fernández, P., I. Villalba Sanchís, V. Yepes, and R. Insa Franco. A Review of Modelling and Optimisation Methods Applied to Railways Energy Consumption. *Journal of Cleaner Production*, Vol. 222, 2019, pp. 153–162. https://doi.org/10.1016/j.jclepro.2019.03.037.
54. Wang, F., and C. L. Ross. Machine Learning Travel Mode Choices: Comparing the Performance of an Extreme Gradient Boosting Model with a Multinomial Logit Model. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2672, No. 47, 2018, pp. 35–45. https://doi.org/10.1177/0361198118773556.
55. Hillel, T., M. Bierlaire, and Y. Jin. *A Systematic Review of Machine Learning Methodologies for Modelling Passenger Mode Choice*. Lausanne, 2019.
56. Witten, I. H., E. Frank, M. A. Hall, and C. J. Pal. *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier, 2016.
57. Bergstra, J., and Y. Bengio. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, Vol. 13, No. 10, 2012, pp. 281–305.
58. Meek, C., B. Thiesson, and D. Heckerman. The Learning-Curve Sampling Method Applied to Model-Based Clustering. *Journal of Machine Learning Research*, 2002. https://doi.org/10.1162/153244302760200678.
59. Xu, J., A. Wang, N. Schmidt, M. Adams, and M. Hatzopoulou. A Gradient Boost Approach for Predicting Near-Road Ultrafine Particle Concentrations Using Detailed Traffic Characterization. *Environmental Pollution*, Vol. 265, 2020, p. 114777. https://doi.org/10.1016/j.envpol.2020.114777.
60. Lundberg, S. M., and S. I. Lee. A Unified Approach to Interpreting Model Predictions. 2017.
61. Lin, J. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory*, 1991. https://doi.org/10.1109/18.61115.
62. Wang, J., B. Liu, T. Fu, S. Liu, and J. Stipancic. Modeling When and Where a Secondary

Accident Occurs. *Accident Analysis and Prevention*, Vol. 130, 2019, pp. 160–166. https://doi.org/10.1016/j.aap.2018.01.024.

63. Nti, I. K., A. F. Adekoya, and B. A. Weyori. A Comprehensive Evaluation of Ensemble Learning for Stock-Market Prediction. *Journal of Big Data*, 2020. https://doi.org/10.1186/s40537-020-00299-5.
64. Yadav, P., S. Hasan, A. Ojo, and E. Curry. The Role of Open Data in Driving Sustainable Mobility in Nine Smart Cities. 2017.