




# Using Classification for Traffic Prediction in Smart Cities

Konstantinos Christantonis<sup>1</sup>, Christos Tjortjis<sup>1</sup>✉ , Anastassios Manos<sup>2</sup>,  
Despina Elizabeth Filippidou<sup>2</sup>, Eleni Mougiakou<sup>3</sup>, and Evangelos Christelis<sup>2</sup>

<sup>1</sup> International Hellenic University, 14th km Thessaloniki–Moudania, 57001 Thermi, Greece  
c.tjortjis@ihu.edu.gr

<sup>2</sup> DOTSOFT SA, 3 Kountouriotou, 546 25 Thessaloniki, Greece

<sup>3</sup> Commonsense, 1-3 Akakiou & 60 Ipirou Street, 10439 Athens, Greece

**Abstract.** Smart cities emerge as highly sophisticated bionetworks, providing smart services and ground-breaking solutions. This paper relates classification with Smart City projects, particularly focusing on traffic prediction. A systematic literature review identifies the main topics and methods used, emphasizing on various Smart Cities components, such as data harvesting and data mining. It addresses the research question whether we can forecast traffic load based on past data, as well as meteorological conditions. Results have shown that various models can be developed based on weather data with varying level of success.

**Keywords:** Smart cities · Data mining · Prediction · Classification

## 1 Introduction

The deployment of modern and smart cities increasingly gains attention, as large urban centers over time present numerous challenges for citizens. Traffic is a stressful and time-consuming factor affecting citizens. Lately, many local authorities attempt to design and create smart infrastructures and tools in order to collect data and utilize models for better decision making and citizen support. Such data are often derived from sensors collecting information about Points Of Interest (POIs) in real time. Data mining techniques and algorithms can then support getting useful insights into the problem, whilst forming appropriate strategies to counter it.

This work focuses on analyzing different approaches regarding data manipulation in order to predict day-ahead traffic loads at random places around cities, based on weather conditions. Prediction efforts regard classification tasks aiming at highlighting factors that affect traffic prediction. This study utilizes weather data collected from sensor devices located in Athens and Thessaloniki, Greece. Three different day zones are introduced and compared while subsets of trimesters are also tested.

The remaining of the paper is structured as follows: Sect. 2 provides background information. Section 3 presents our approach for traffic prediction, including data selection as well as experimental results. Section 4 discusses and evaluates our findings while Sect. 5 concludes the paper with suggestions for future work.

## 2 Background

Traffic prediction is not a new subject; there are numerous scientific efforts that perform both classification and regression tasks in this domain [1]. However, few efforts attempted to predict traffic volumes based on low level data, such as weather data.

Nejad et al. examined the power of decision trees for classifying traffic loads on three levels, based only on time and temperature [2]. Results were positive and motivated our work. Xu et al. compared CART with k-NN and a direct Kalmann Filter [3]. Moreover, Wang et al. proposed the use of volume and occupancy data [4]. Such data, as well as speed, can be obtained by loop detectors. Loop detectors are sensors buried underneath highways which estimate traffic by observing information related to vehicles passing above them. Loop detectors along with rain predictions were also used in order to predict crashes [5].

Tree-based algorithms are widely used and justified, however even more sophisticated algorithms were used such as Support Vector Machines (SVMs) [6] and Neural Networks (NNs) [7]. A novel hybrid method for short-term traffic prediction under both typical and atypical traffic conditions. Theodorou et al. introduced an SVM based model that identifies atypical conditions [1]. We used ARIMA or k-NN regression to identify typical or atypical conditions, respectively. In addition to [4], Liu et al. introduced three binary variables, other than weather conditions, for holiday, special conditions and quality of roads [8].

However, in this work we chose not to include such features for several reasons. Firstly, holidays do not have the same effect on each season. For example, sunny holidays might result in lower levels of traffic in large urban centers whilst for winter holidays this is not the case. Special conditions, such as a major social events or protests are not easy to add to models since most of the times these occur unexpectedly. Abnormal conditions have a negative impact on such models, for instance when attempting to estimate the traffic flow based on Expressway Operating Vehicle [9].

Another significant decision on such predictions relates to the actual day of the week. As highlighted in [10, 11], weekdays tend to have different characteristics from weekends and should be carefully tested. Finally, most scenarios regard short-time prediction, meaning minutes to h ahead. Dunne et al. utilized neurowavelet models along with the rain expectation in the next hour [12].

## 3 Traffic Prediction

Weather data can be exploited in different smart cities problems and scenarios. Collecting accurate weather data can be beneficial for numerous daily problems which associate weather with human decisions. Traffic is affected by numerous factors, only some of which are predictable. As mentioned in Sect. 2, traffic is directly affected by weather conditions, however the scale differs across cities and cultures.

### 3.1 Problem Definition and Approach

This section tests the above intuition on traffic prediction. Traffic is a multi-dimensional problem; researchers focus on either predicting traffic loads on a certain time interval

or selecting the optimal route based on real-time adaptations in order to minimize travel time. Accidents and social events can shortly disrupt normal traffic in specific areas, while seasonality and weather conditions can affect traffic in a larger scale. Based on that, we used weather data collected from sensors installed around carefully chosen specific city spots for predicting the day-ahead traffic volume.

To select the most appropriate locations to install sensors that either measure traffic loads or collect weather data, it is crucial to define their objective in advance. Our efforts focus on the question ‘How can one exploit sensor data that are not personalized and create meaningful conclusions for the general public?’ Deployment of smart city infrastructure requires a deep understanding of the traffic problem. Roads that are busier than others do not always provide more information in comparison to more isolated ones. The number of alternative roads or the location of busy buildings can affect the necessity of measuring traffic for a specific road.

Our approach, besides examining traffic predictability based on weather data, also aims to clarify differences among locations. Moreover, we implement a series of tests regarding different monthly periods under the objective to understand which months contribute positively on the deployment of such models. The approach followed is explained in Sect. 3.2.

### 3.2 Dataset Description and Processing

For this task, our data source was the newly deployed ppcity.io, which is a set of platforms providing useful information derived from sensors to citizens and visitors of Athens and Thessaloniki, Greece. These two large urban centers host almost half the Greek population. The selected platform collects and analyses environmental, traffic and geospatial data. Those data are collected through several sensors located at central points around the two cities. In general, there are numerous city spots on both cities where the environmental and traffic conditions are measured. Initially, we chose locations about which we had information regarding both traffic and weather conditions. We focused on the ten most reliable spots that produce data, i.e. sensors with the most compact flow of data and wider range of recording.

Weather related attributes used for the research analysis involved the following: Humidity, Pressure, Temperature, Wind Direction, Wind Speed and Ultraviolet Radiation (UV). The platform provides additional weather data, but these were not included in the modeling process. Indicatively, such variables include measurements for ozone concentration, nitrogen dioxide etc.

The target variable (traffic load), named *Jam Factor*, represents the quality of travel. It ranges from 0 to 10, where 10 describes stopped traffic flow and 0 a completely empty road. Data have been collected for almost six months, a rather short period of time, but at least including August and December, two months when people tend to take holidays. August is widely considered in Greece as the month with the lowest traffic volume, since it is a period that most people go on holidays, away from large urban centers. Similarly, in December both cities face abnormalities in traffic loads, since many citizens move from and to the two large urban centers. The dataset comprises data collected between 29/7/19 and 3/2/20. Eight sensors are located in Athens and two in Thessaloniki. In

addition, two sensors are located near school areas. More information about the sensors is available in Sect. 4.

The way data are processed is conceptually simple. Regarding the data structure, we defined three distinct time intervals within a day and examined their predictability. The first interval, named *Morning*, includes signals from sensors between 07:00 and 10:00. The second interval, named *Afternoon*, includes signals between 15:00 and 18:00 and the last one, named *Evening*, includes signals between 19:00 and 22:00.

Therefore, the average value for the given signals was computed. For example, if a sensor captures information every  $h$  (e.g. at 07:10, 08:10, 09:10 etc.), we computed and assigned the average value for each weather metric and the traffic load for that specific day period. The above strategy resulted in creating three different datasets consisting of 185 instances on average. Fortunately, the quality of the dataset was high, thus missing values were minimal. It is worth noting that our averaging strategy does not replace missing values. If there is a missing value in a weather averaging is performed using the remaining two available values, instead of averaging and labeling three values.

An important decision was to transform the problem into a classification task by categorizing the target variable into two and later into three classes. Initially, the split point was the median value for each dataset, aiming at a fully balanced task which would allow for safer conclusions. Therefore, the two classes named *High* and *Low* indicate if the traffic load on that specific day was higher than usual. In the second stage the target variable was split into three categories of equal size (High, Medium and Low). The metric used for classification evaluation was accuracy, since classes are balanced and of equal interest.

Finally, we split the dataset into three subsets consisting of data regarding different trimesters. The first one contains data for the period between 29-July and 25-October, the second for the period between 15-September and 20-December and the third one, between 25-October and 3-February. We aimed at distinguishing if any periods (seasonality) affect the models negatively.

### 3.3 Experiments

Different cities demonstrated different traffic patterns; however, in almost every case we observed a few common patterns. First, on weekends citizens tended to use vehicles much less than weekdays, while on holidays people tended to leave the urban centers. To further understand the case of Greece, Fig. 1 visualizes the mean value for all available sensors in order to explain phenomena beyond seasonality. The colors on all figures share the same palette and indicate the same subsets. More precisely, for Fig. 1, the blue line indicates the *Afternoon* values while red and green indicates *Morning* and *Evening* respectively.

For Figs. 2, 3, 4, 5, 6 and 7 blue indicates the subset of 29 July–3 February, red for 15 September–20 December, green for 29 July–25 October and purple for 25 October–3 February. As expected, a rapid introductory search on the data surface highlight that the volume of traffic was consistently higher during the *Afternoon* period, while both *Morning* and *Evening* periods seem to present similar amounts of traffic.

Regarding the predictability for each of these combinations between sensors and day periods, for all the experiments on this task, we used the Random Forest Classifier. It is an

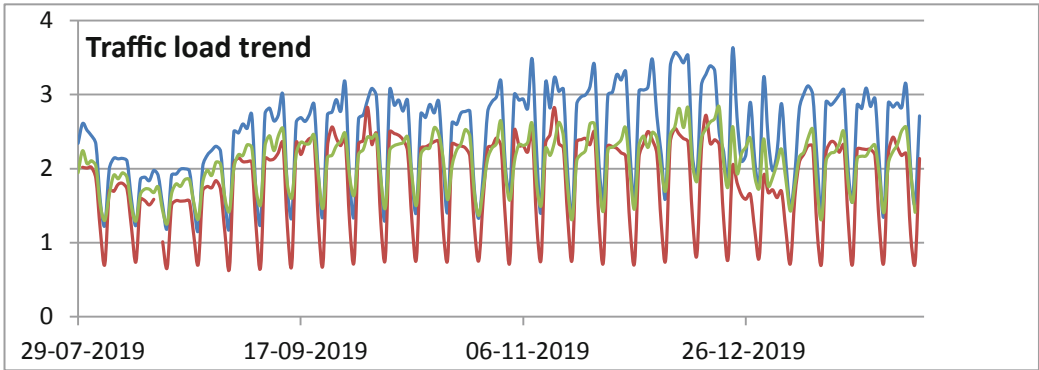


Fig. 1. Mean value for all sensors for each day period (Color figure online)

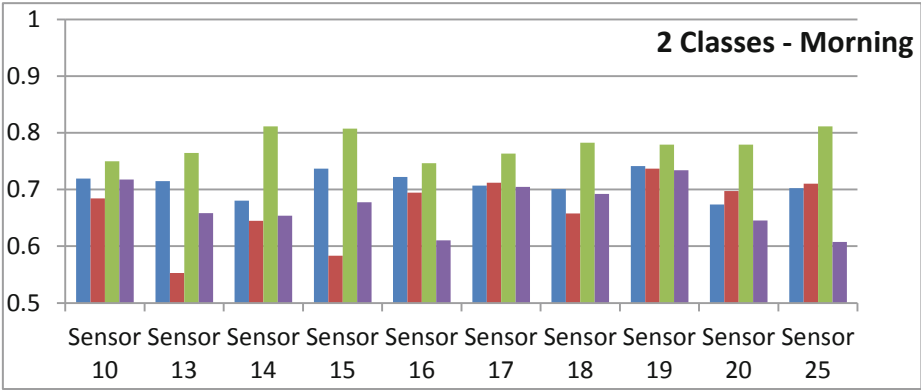


Fig. 2. Accuracy with 2-classes – Morning (Color figure online)

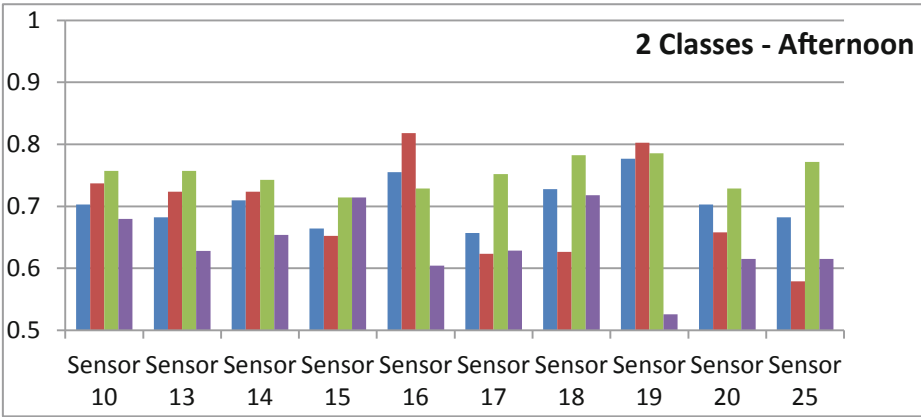
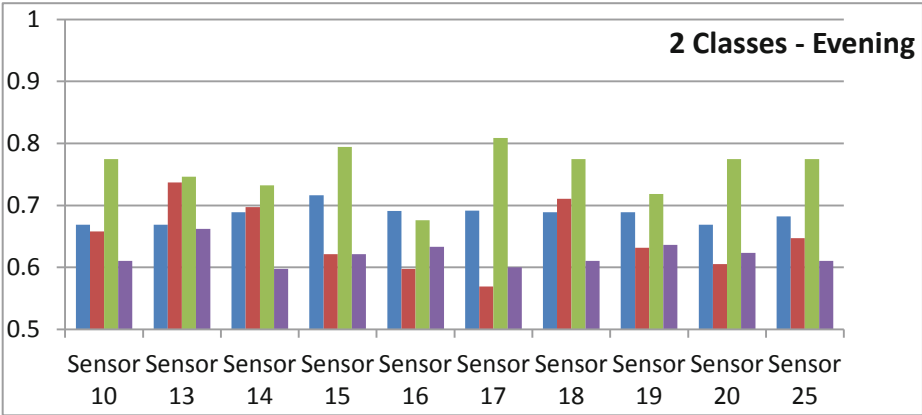
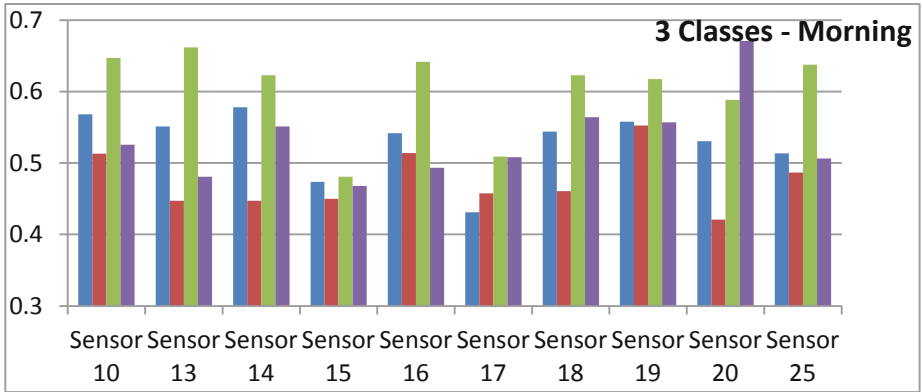


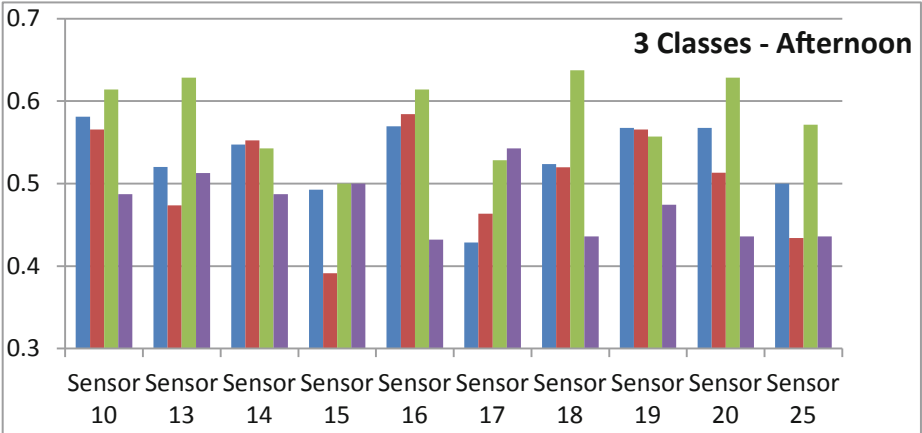
Fig. 3. Accuracy with 2-classes – Afternoon (Color figure online)



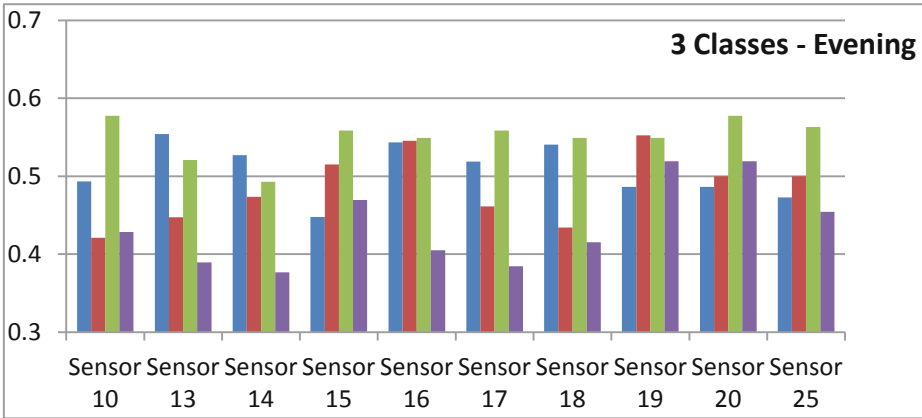
**Fig. 4.** Accuracy with 2-classes – Evening (Color figure online)



**Fig. 5.** Accuracy with 3-classes – Morning (Color figure online)



**Fig. 6.** Accuracy with 3-classes – Afternoon (Color figure online)



**Fig. 7.** Accuracy with 3-classes – Evening (Color figure online)

ensemble learning method that operates by constructing a multitude of decision trees. It is considered as a powerful method and stands as a top-notch solution for various problems [13]. On top of that, it handles both scaled and not scaled data. In order to prevent over-fitting, we used 10-fold cross-validation and for the optimal hyper-parameter tuning we used an extended grid search.

It is important to clarify that all experiments were also conducted using the Logistic Regression algorithm, however the results were not included since they were almost identical. That is to justify that our algorithm does not over-fit since our datasets are small. Further, Figs. 2, 3, 4 show classification results for each of the described time intervals. The Y-axis indicates accuracy with a start-point at 0.5 to highlight the margin of a baseline model on a perfectly balanced classification task.

Since the purpose of this work was to analyze and highlight differences between time and day periods on traffic prediction we did not focus on exact values. In most cases feeding machine learning algorithms with more data is considered a crucial step to achieve generalization and expand the margin for higher accuracy. However, in all cases above, results with the initial dataset clearly did not outperform the ones with subsets. Figures 5, 6 and 7 show results obtained by dividing the target variables into three classes of equal instances.

Figures 5, 6 and 7 illustrate the same intuition as for the case of two classes (i.e. Higher accuracy for the 29-Jul/25-Oct period), but also highlight significant changes on the individual performance for each sensor. In Sect. 4 we evaluate and discuss results, whilst explaining abnormalities and unexpected behavior.

## 4 Evaluation and Discussion

The evaluation of this work examines all the essential steps on a data mining project. Firstly, data acquisition, which is a critical component on every project, revealed the importance of sufficient data. The data collection process should be constant and clearly defined in advance. We dealt with a well-structured database that was recording sensor

signals in a nearly perfect synchronization between traffic and weather data. Regarding the pre-processing step, we did not use all the available features, instead the models contained only weather metrics tested and introduced by the literature. For weather data exploitation, it is crucial to fully understand the correlation between variables, because the same level of information might be repeatedly captured. The rest of this section further discusses our findings per case.

The general problem in terms of real-time adjustment of routes in order to achieve traffic “congestions” is still most important for many cities. However, the day-ahead prediction of the volume was based on a major assumption that there would not be any accident or abnormal conditions in general. Starting from this point, the factor of environmental conditions was essential since many employees, students and tourists decide in advance the way they travel around the city on the upcoming day. The results of the approach presented in Sect. 3.3 are encouraging and justify what was stated above. All three-day periods on the binary classification achieved quite satisfying accuracy levels.

More precisely, the initial model, resulted in accuracy higher than 0.7 for most sensors on Morning and Afternoon. On the other hand, for the 3-classes experiments, results do not allow to reach safe conclusions. However, the initial model demonstrated robust performance achieving on average accuracy higher than 0.5. For the Morning period we observed predictability similar with the binary case. For Afternoon and Evening periods there were not significant gaps in accuracy even though the 29-Jul/25-Oct (light green line) period shows better performance than the rest. In addition, we observed that the 3-class experiments do not show similar patterns with the previous results.

Surprisingly, the middle period 15-Sep/20-Dec (red line) was expected to achieve much better results especially for the sensors 13 and 19 which are installed close to school areas. Indeed, the range of this period was selected to cover and adjust for the periods that schools operate, undisrupted. Results were contradicting, while Sensor 19 outperforms the rest on the Morning period for the binary case, Sensor 13 resulted in the lowest accuracy.

Figure 1 highlights the special case of August in Greece and the fact that the traffic load is highly reduced. In addition, as expected, on weekends traffic was heavily reduced while we observed that peaks on Afternoon and Evening periods for weekdays, happen mostly on Thursdays and Fridays.

Finally, it is worth noting that Sensors 15 and 17 are in Thessaloniki. For both, results were lower than the average for Athens. Based on these results, we observed that the range of traffic values for those two is amongst the highest.

## 5 Conclusions and Future work

Experiments for both stages produced some clear and informative results. The main conclusion is that the systematic recording and use of weather data can support decision making. The list below summarizes the conclusions.

- Standalone weather metrics can assist in building reliable prediction models regarding traffic volumes.



- Roads are busier in the afternoon and for most of the sensors even evenings have higher traffic volumes in comparison to mornings.
- Mornings do not return steady results for all sensors. As discussed, the results are stable only for sensors located near schools.
- For morning periods, the peaks of traffic happen in the beginning of the week. No earlier than the middle of September the load gets similar to that of Evenings and gets clearly lower again by the start of Christmas holidays.
- The first data subset regarding the period 29-Jul/25-Oct outperforms both other subsets and the initial data set consisting of all available data. That indicates that winter months introduce uncertainty and volatility, thus related models underperform significantly.
- Transforming the target variable into three classes resulted in admittedly good results, firmly better than the baseline model.

The above conclusions emerge from the detailed analysis of the models, weather metrics and traffic volumes. However, threats to their validity exist. We briefly summarize them in Sect. 5.1.

### 5.1 Threats to Validity

The biggest threat on approaches as such, is the fact that those day-ahead models rely on weather data which also are predicted. Thus, it is crucial that we have accurate predictions of weather conditions. Moreover, including the weekends that admittedly have different loads of traffic in the same models with weekdays may affect the validity on a negative way. Another threat could be sufficient deseasonalising; the factor of time could be possible analyzed into more explanatory variables. Not having available data for at least one year of recordings may lead to questionable conclusions about seasonal effects; however, this is not a rule. Finally, the sensors regard roads of different volume of traffic and even though traffic usually fluctuates uniformly that may conclude to misleading results.

**Acknowledgements.** The work is implemented within the co-funded project Public Participation City (ppCity - T1EΔK-02901 and MIS 5029727) by Action Aid “Research-Create-Innovate” implemented by General Secretariat of Research and Technology, Ministry of Development and Investments. The project ([www.ppcity.eu](http://www.ppcity.eu)) provides a set of tools and platforms to collect city environmental data and use these in an intelligent way in order to support informed urban planning. Key element in this process is the opinions and views of citizens which are collected in a crowd sourcing manner. The research depicted under this paper is based on Platform 3 (run from <https://panel.ppcity.eu/platform3/>) which provides an open data city portal from environmental data in Athens and Thessaloniki, Greece.

## References

1. Theodorou, T.I., Salamanis, A., Kehagias, D., Tzovaras, D., Tjortjis, C.: Short-term traffic prediction under both typical and atypical traffic conditions using a pattern transition model. In: 3rd International Conference on Vehicle Technology & Intelligent Transport Systems, pp. 79–89 (2017)

2. Nejad, S.K., Seifi, F., Ahmadi, H., Seifi, N.: Applying data mining in prediction and classification of urban traffic. In: 2009 WRI World Congress on Computer Science and Information Engineering, pp. 674–678 (2009)
3. Xu, Y., Kong, Q., Liu, Y.: Short-term traffic volume prediction using classification and regression trees. In: 2013 IEEE Intelligent Vehicles Symposium (IV), pp. 493–498 (2013)
4. Wang, Y., Chen, Y., Qin, M., Zhu, Y.: Dynamic traffic prediction based on traffic flow mining. In: 2006 6th World Congress on Intelligent Control and Automation, Dalian, pp. 6078–6081 (2006)
5. Abdel-Aty, M.A., Pemmanaboina, R.: Calibrating a real-time traffic crash-prediction model using archived weather and ITS traffic data. *IEEE Trans. Intell. Transp. Syst.* **7**(2), 167–174 (2006)
6. Yan, H., Yu, D.: Short-term traffic condition prediction of urban road network based on improved SVM. In: 2017 International Smart Cities Conference, pp. 1–2 (2017)
7. Tang, J., Li, L., Hu, Z., Liu, F.: Short-term traffic flow prediction considering spatio-temporal correlation: a hybrid model combining type-2 fuzzy c-means and artificial neural network. *IEEE Access* **7**, 101009–101018 (2019)
8. Liu, Y., Wu, H.: Prediction of road traffic congestion based on random forest. In: 2017 10th International Symposium on Computational Intelligence and Design, pp. 361–364 (2017)
9. Ai, Y., Bai, Z., Su, H., Zhong, N., Sun, Y., Zhao, J.: Traffic flow prediction based on expressway operating vehicle data. In: 2018 11th International Conference on Intelligent Computation Technology and Automation, pp. 322–326 (2018)
10. Clark, S.: Traffic prediction using multivariate nonparametric regression. *J. Transp. Eng.* **129**(2), 161–168 (2003)
11. Christantonis, K., Tjortjis, C.: Data mining for smart cities: predicting electricity consumption by classification. In: IEEE 10th International Conference on Information, Intelligence, Systems and Applications, pp. 67–73 (2019)
12. Dunne, S., Ghosh, B.: Weather adaptive traffic prediction using neurowavelet models. *IEEE Trans. Intell. Transp. Sys* **14**(1), 370–379 (2013)
13. Tzirakis, P., Tjortjis, C.: T3C: improving a decision tree classification algorithm's interval splits on continuous attributes. *Adv. Data Anal. Classif.* **11**(2), 353–370 (2017). <https://doi.org/10.1007/s11634-016-0246-x>