Classifying and Predicting Urban Traffic States Based on Temporal and Weather Conditions

Paraskevi Krini[1], Virginia Petraki[1*], George Yannis[1]

[1] National Technical University of Athens, Department of Transportation Planning and Engineering, 5 Heroon Polytechniou str., 15773, Athens, Greece
*vpetraki@mail.ntua.gr

Abstract. Urban traffic congestion remains one of the most pressing challenges for modern cities, affecting the overall quality of life. This study aims to identify and classify traffic states on two central roads in Athens, Greece, using a multi-stage approach that combines clustering, classification, and explainable AI techniques. Traffic speed data were collected from the Google Directions API, while hourly load data were obtained from detectors operated by the Traffic Management Centre of Attica. Covering the period from January to July 2022, the datasets were combined spatiotemporally, with a focus on peak periods during both weekdays and weekends to capture high-demand traffic conditions. K-means clustering was applied to identify two primary traffic states: (i) congested, with higher load and lower speeds, and (ii) less-congested, with lower load and higher speeds. To mitigate class imbalance, the SMOTE oversampling technique was employed. Four XGBoost classification models were trained separately for each road with two directions and evaluated using confusion matrices and standard performance metrics. SHAP values were then used to interpret model predictions and assess feature importance. Weekday, hour during the day, and temperature emerged as the most influential variables. These findings offer valuable insights to support urban traffic management and planning strategies.

Keywords: traffic congestion, K-means, classification, SHAP values.

1 Introduction and Background

Urban road traffic dynamics are a critical issue for urban planning and city sustainability. The intensive mobility between city centers, which are hubs for everyday life, creates intensified needs for transportation. These trips are serviced by various transport modes, such as private cars, public transport, taxis, and micro-mobility devices and associated infrastructures. However, a dependence upon private car usage has created significant social, environmental, and operational impacts, the predominant of which is traffic congestion. Traffic congestion is recognized to be a complex phenomenon, based on spatiotemporal circumstances, external data (e.g., road crashes, roadworks) etc. Numerous cities suffer from long-term traffic congestion, which is estimated to cost EUR 270 billion annually in Europe [1].

The literature indicates that accurate traffic analysis takes place through the combination of data and application of modern analytical methods such as clustering and deep learning [2, 3]. The number of vehicles and traffic speed are needed to determine traffic conditions [4]. Leveraging the increasing loads of available traffic data, along with the inclusion of explanatory variables (such as weather and geometric attributes), and computational techniques, provides sufficient flexibility and promotes more accurate and optimized traffic forecasting models [5]. Meanwhile, the selection of the appropriate data harvesting methods (i.e. spanning sensors, GPS, and satellite technologies) determines the quality of the output [6, 7]. Finally, data-driven workflows that integrate heterogeneous inputs with machine-learning models have been shown to enhance short-term prediction and operational decision-support in practice [5].

This study aims to characterize and predict traffic states on two central Athenian corridors. To achieve this objective, a two-step system was developed that applied unsupervised clustering with K-means followed by supervised classification with XGBoost, with SHAP providing model explanations. The analysis focuses on the influence of weather and temporal factors on the defined traffic states for each road, as well as their ability to predict those states.

The paper is organized as follows: the first section provides motivation for this study and the objectives. Section 2 presents the proposed methodological approach for identifying and predicting the traffic states. The following section describes the implementation and results, and the final section discusses the main findings and conclusions.

2 Methodology

2.1 Data Collection

Traffic data was collected from Google Maps and the Traffic Management Center (TMC) and spatiotemporally matched the segment-hour level. Specifically, the average speed, travel time, and geographical information (start, and end of each road segment) were collected for 62 road segments in Attica using Google Maps, while traffic load was recorded at 26 predefined locations by the TMC. The dataset covers the period from January to July 2022, with hourly average speed measurements at 20-minute intervals on weekdays (8:00–9:40, 12:00–18:40, and 21:00–23:40) and weekends (Saturdays: 12:00–18:40, 21:00–23:40; Sundays: 12:00–18:40). Additionally, weather conditions in terms of hourly average temperature, rainfall, snowfall, and day/night changes were obtained using the Open-Meteo API. The spatiotemporal merging of the speed and traffic load datasets enabled the mapping of sensor loops onto four main road axes in downtown Athens:

1. Alexandras Avenue (from Kifisias to Patision) [Alexandras 1]
2. Alexandras Avenue (from Patision to Kifisias) [Alexandras 2]
3. Vasilissis Sofias Av. (from Panepistimiou to Vasilissis Konstantinou) [Vas.Sofias 1]
4. Vasilissis Sofias Av. (from Vasilissis Konstantinou to Panepistimiou) [Vas.Sofias 2]

Along Alexandras 2, Vas. Sofias 1, and Vas. Sofias 2, three sensor loops are located on each direction. On Alexandras 1 one traffic loop is presented.

The final database contained hourly traffic speed, hourly traffic load per traffic lane, weather conditions, and temporal features (is_day, weekday/weekend) as presented in Table 1.

Table 1. Description of Variables.

Variable	Description	Data Source
Traffic_Speed	Traffic speed per hour (km/h)	Google Directions API
Traffic_Load	Vehicles/hour/lane	Traffic Management Center (TMC)
is_day	0=night ; 1=day	OpenMeteo
Weekday	0= weekend; 1=weekday	n/a
Month	1(=Jan) – 12(=Dec)	n/a
Day	1(= Mon)- 7(=Sun)	n/a
Hour	0-23	n/a
temperature_2m	Average hourly temperature (°C)	OpenMeteo
rain	Precipitation height (mm)	OpenMeteo
snowfall	Snowfall depth (mm)	OpenMeteo

2.2 Clustering

Clustering analysis was performed by road axis to investigate the conditions under which traffic load and vehicle speed vary and to explore their relationship. The Silhouette method was used to determine the optimal number of clusters by evaluating clustering quality, measuring how well each point fits within its cluster. The Silhouette measure is between -1 and 1, with -1 indicating poor clustering and 1 indicating effective clustering [8]. After finding the best number of clusters, the K-means algorithm was utilized due to its efficiency and widespread use in classifying data. The K-means algorithm allocates data into a given number of clusters by minimizing the square distances between individual points and their respective cluster centroids, ensuring distinct and reliable separation, which is highly beneficial when dealing with traffic data [8].

2.3 Classification

The eXtreme Gradient Boosting (XGBoost) algorithm was used for classification purposes. This algorithm uses a second-order Taylor expansion to improve loss function optimization, allowing incremental construction of decision trees via the Gradient Boosting method [9]. An important aspect of the approach is division of the dataset between training and test subsets, which allows for proper assessment of model generalizability and helps prevent overfitting. According to [10], the flexibility and comparatively low computational burden of XGBoost have made it one of the most effective tools in the field of machine learning, often proving superior to other approaches in performance.

Synthetic Minority Over-sampling Technique (SMOTE) was utilized to counteract the class imbalance problem. The technique generates new instances of the minority class through interpolation among original samples and their close neighbors, thus enlarging the minority class's decision boundary, reducing the threat of overfitting, and improving classification performance [11].

One key component of the current work is the utilization of SHAP (Shapley Additive Explanations) values for model explanation. Based on game theory concepts, SHAP estimates the contribution of each feature to making predictions, thus enabling feature importance and interpretability evaluation in both regression and classification settings regardless of the utilized model [12].

3 Results

3.1 Cluster Analysis

The traffic conditions have been clustered with respect to two basic traffic factors: hourly traffic speed and hourly traffic load per lane. By applying the Silhouette Width measure, the analysis reveals that the best number of clusters of all road segments is two. Figure 1 presents the cluster analysis results of the road segments examined.

The means of the traffic speed and traffic load for each traffic cluster, along with the number of observations and the respective average Silhouette Width, are given in Table 2. The analysis has therefore recognized two different traffic states: (a) congested, characterized by low speeds and increased traffic load, and (b) less congested, with higher traffic speed and lower traffic load. Interestingly, the analysis resulted in dividing the speed-load graph into two different areas. One can clearly identify the less-congested region and the congested area. The quality of the resulting clustering is acceptable, given that the range of the Silhouette Width measures is between 0.44 and 0.65

Table 2. Descriptive statistics of the identified traffic clusters.

Cluster	Cluster Label	Traffic Speed (km/h)	Traffic Load (vehicles/hour/lane)	Cluster Size	Av. Silhouette Width
Alexandras (from Kifisias to Patision) **[Alexandras 1]**					
1	less-congest	27.095	361.813	377	0.52
2	congested	15.527	453.579	981	0.65
Alexandras (from Patision to Kifisias) **[Alexandras 2]**					
1	less-congest	29.576	334.469	364	0.49
2	congested	17.343	406.542	991	0.62
Vas. Sofias (from Panepistimiou to Vas. Konstantinou) **[Vas. Sofias 1]**					
1	congested	16.091	524.781	821	0.51
2	less-congest	22.339	427.675	534	0.44
Vas. Sofias (from Vas. Konstantinou to Panepistimiou) **[Vas. Sofias 2]**					
1	less-congest	20.448	216.923	453	0.46
2	congested	12.659	284.094	900	0.44

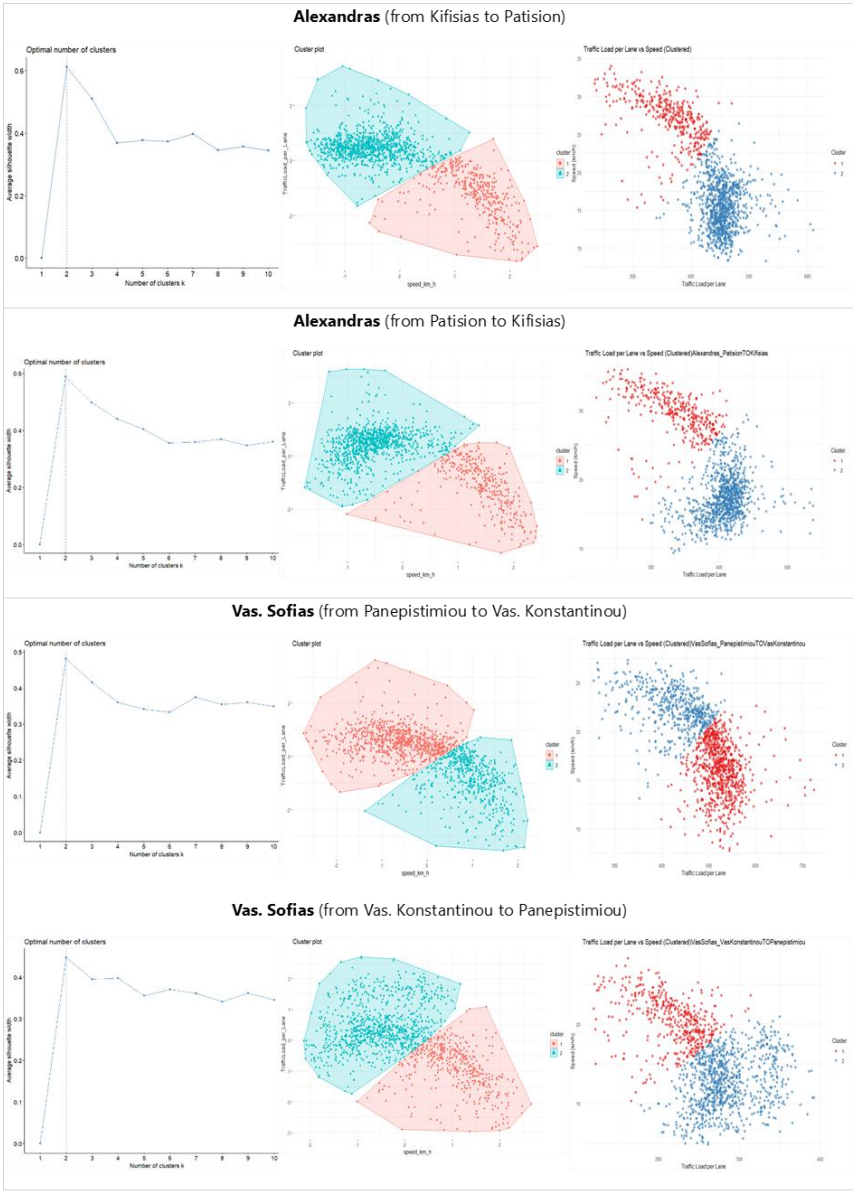

Fig. 1. Clustering results.

3.2 Classification Analysis

The XGBoost algorithm uses as input all variables described in Section 2.1. and as output the traffic states (Table 1). 80% of the sample was used for training and 20% for evaluating the model. Hyperparameter tuning with 4-fold cross-validation was carried

out to mitigate overfitting and enhance the model's performance with seven different hyperparameter combinations tested and randomly chosen from the ranges listed in Table 3. The entire process of selecting, running, and comparing these combinations was automated using R code. The objective was to determine the internal model configuration that provided the highest classification accuracy. The search was conducted over a specified parameter distribution, including learning rate, iterations, max depth, gamma, colsample by tree, minimum sum of instance weights (hessian) and subsample which was set to 1. The following best parameters were identified through the search. The final optimized hyperparameters are presented in Table 3.

Table 3. Hyperparameter tuning results for XGBoost models.

Hyperparameter	Examined Range	Optimized Values			
		Alexandras 1	Alexandras 2	Vas. Sofias 1	Vas. Sofias 2
Boosting iterations	100-300	200	300	300	300
Max depth	4-6	5	4	6	6
Learning rate	0.025-0.3	0.1	0.3	0.1	0.3
Gamma	0-2	0	1	0	2
Colsample by tree	0.5-1	1	0.5	1	0.5
Min sum weights	1-5	1	1	5	1
Subsample	fixed at 1	1	1	1	1

Overall accuracy ranges from 0.812–0.923, suggesting that the models correctly predict the traffic states 81%-92% of the time, and with high discrimination (AUC 0.867–0.959). The best results are on Alexandras Av., while Vas. Sofias is lower but still shows good F1 for the positive class.

Table 4. Confusion matrix.

Metric	Alexandras 1	Alexandras 2	Vas. Sofias 1	Vas. Sofias 2
Accuracy (Test)	0.923	0.912	0.875	0.812
F1-score (Class 1)	0.949	0.940	0.835	0.867
Balanced Accuracy	0.893	0.879	0.876	0.776
AUC	0.959	0.934	0.947	0.867
Kappa	0.791	0.771	0.735	0.550

SHAP values report the average contribution a feature has on the models' outcome, across all possible combinations of inputs. Figure 2 presents the SHAP summary plots for the four road–direction models. It should be noted a labeling nuance, for Vas. Sofias (Panepistimiou to Vas. Konstantinou) the congested state is Cluster 0, and the less-congested state is Cluster 1 (reversed relative to the other three models). Therefore, comparing the SHAP plots between travel directions shows both the ranking of influential variables and whether their effects are consistent. Day type (weekday or weekend) ("Weekday"), hourly temperature ("temperature_2m"), and time of day ("Hour") emerge as the most influential predictors. Calendar dummies (month/day) offer secondary impact, while weather extremes contribute only marginally.

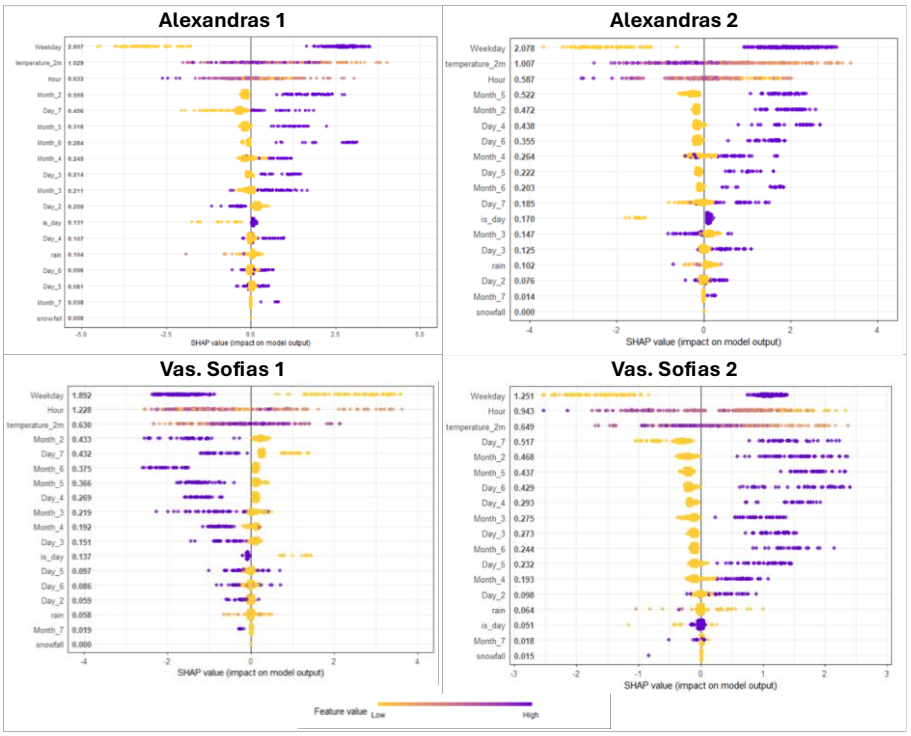

Fig. 2. Summary plots of SHAP values.

4 Discussion and Conclusions

This study aims to characterize and predict traffic states on two central Athenian corridors through a two-step workflow, unsupervised clustering followed by supervised classification with XGBoost, complemented by SHAP-based explanations.

Traffic states were defined hourly from two fundamental parameters: traffic speed and traffic load per lane. Across all four road directions, the clustering consistently yielded two traffic states: a congested state (lower speed and higher load per lane) and a less congested state (higher speed and lower load). The analysis focused on peak hours, capturing the periods of highest demand. Consequently, the findings speak primarily to peak-period management rather than all-day conditions. Similar state definitions based on speed and flow/load are common in the literature. Studies have clustered traffic using speed–load, or speed–occupancy, features to identify a small set of regimes (2 to 4), that map onto free-flow vs. congested conditions, corroborating our two-state outcome [12].

XGBoost models attained 81-92% accuracy, meaning that the developed models can reliably flag whether a road is in a congested or less-congested state. SHAP summaries show that day type, temperature, and time of day dominate model decisions across directions, with month, day-of-week dummies, and daylight providing secondary

refinements, while extreme weather playing a minor role. Those patterns are consistent with evidence that temporal regularities drive urban congestion, and weather exerts smaller but measurable effects [3]. Weekdays push predictions toward the congested state (strongest on Alexandras Av.) and higher temperatures increase congestion likelihood. Hour refines predictions with later peak hours tending toward congestion, though the magnitude is secondary and more variable across directions. On Vas. Sofias, the hour during the day impacts more than temperature compared to Alexandras Av., possibly since it acts as a primary inbound arterial to offices and public services, making the hour more influential. In summary, the models' strong test performance indicates a possible practical readiness for real-time traffic state flagging, while SHAP clarifies that temporal regularities and temperature are the primary drivers.

References

1. European Court of Auditors: Special report 06/2020: Sustainable urban mobility in the EU: No substantial improvement is possible without member states' commitment. Publications Office of the European Union, Luxembourg (2020).
2. Zhang, Y., Ye, N., Wang, R., Malekian, R.: A method for traffic congestion clustering judgment based on grey relational analysis. ISPRS International Journal of Geo-Information 5(5), 71 (2016).
3. Yin, X., Wu, G., Wei, J., Shen, Y., Qi, H., Yin, B.: Deep learning on traffic prediction: Methods, analysis, and future directions. IEEE Transactions on Intelligent Transportation Systems 23(6), 4927–4943 (2021).
4. Wang, W.X., Guo, R.J., Yu, J.: Research on road traffic congestion index based on comprehensive parameters: Taking Dalian city as an example. Advances in Mechanical Engineering 10(6), 1687814018781482 (2018).
5. Antoniou, C., Koutsopoulos, H.N., Yannis, G.: Dynamic data-driven local traffic state estimation and prediction. Transportation Research Part C: Emerging Technologies 34, 89–107 (2013).
6. Seo, T., Kusakabe, T., Asakura, Y.: Estimation of flow and density using probe vehicles with spacing measurement equipment. Transportation Research Part C: Emerging Technologies 53, 134–150 (2015).
7. Karami, Z., Kashef, R.: Smart transportation planning: Data, models, and algorithms. Transportation Engineering 2, 100013 (2020).
8. Mantouka, E.G., Barmpounakis, E.N., Vlahogianni, E.I.: Identifying driving safety profiles from smartphone data using unsupervised learning. Safety Science 119, 84–90 (2019).
9. Kostopoulos, A., Garefalakis, T., Michelaraki, E., Katrakazas, C., Yannis, G.: Modeling and sustainability implications of harsh driving events: A predictive machine learning approach. Sustainability 16(14), 6151 (2024).
10. Nielsen, D.: Tree boosting with xgboost—why does xgboost win "every" machine learning competition? Master's thesis, NTNU (2016).
11. Gu, Q., Wang, X.M., Wu, Z., Ning, B., Xin, C.S.: An improved SMOTE algorithm based on genetic algorithm for imbalanced data collection. Journal of Digital Information Management 14(2) (2016).
12. Vlahogianni, E.I., Karlaftis, M.G., Stathopoulos, A.: An extreme value based neural clustering approach for identifying traffic states. In: IEEE Intelligent Transportation Systems Conference, pp. 320–325. IEEE (2005).