An Interactive Visual Analytics Platform for Smart Intelligent Transportation Systems Management

Ilias Kalamaras, Alexandros Zamichos, Athanasios Salamanis, Anastasios Drosou, Dionysios D. Kehagias, Georgios Margaritis, Stavros Papadopoulos, and Dimitrios Tzovaras, *Senior Member, IEEE*

***Abstract*— The reduction of road congestion requires intuitive urban congestion-control platforms that can facilitate transport stakeholders in decision making. Interactive ITS visual analytics tools can be of significant assistance, through their real-time interactive visualizations, supported by advanced data analysis algorithms. In this paper, an interactive visual analytics platform is introduced that allows the exploration of historical data and the prediction of future traffic through a unified interactive interface. The platform is backed by several data analysis techniques, such as road behavioral visualization and clustering, anomaly detection, and traffic prediction, allowing the exploration of behavioral similarities between roads, the visual detection of unusual events, the testing of hypotheses, and the prediction of traffic flow after hypothetical incidents imposed by the human operator. The accuracy of the prediction algorithms is verified through benchmark comparisons, while the applicability of the proposed toolkit in facilitating decision making is demonstrated in a variety of use case scenarios, using real traffic and incident data sets.**

***Index Terms*— Traffic prediction, visual analytics, multi-objective visualization, hypothesis testing.**

I. Introduction

REDUCING road congestion implies a capability to interact either offline or in real-time with various aspects of urban road networks. Towards this goal, intuitive urban congestion-control platforms for transport stakeholders, such as policy makers and decision authorities, are deemed necessary. However, current tools only marginally or unsatisfactorily support a set of capabilities needed for advanced and evidence-based decision making. In order to fulfill this gap, interactive ITS visual analysis tools are capable to support a set of advanced features, such as (i) analysis of travel-behavioral patterns; (ii) assessment of critical quality measures; (iii) real-time visualization of traffic flow; (iv) automated recommendations of corrective actions that will lead to reduced congestion; (v) evaluation of the impact that new traffic rules have on congestion and quality of life for citizens.

Manuscript received July 29, 2016; revised January 18, 2017 and June 16, 2017; accepted July 2, 2017. This work was supported in part by the European Commission through the 7th Framework Program under Project FP7-ICT-609026-MOVESMART, and in part by Horizon 2020 under Grant H2020-RIA-653460-RESOLUTE. The Associate Editor for this paper was W. Chen. *(Corresponding author: Ilias Kalamaras.)*

The authors are with the Information Technologies Institute, Centre for Research and Technology Hellas (CERTH), 57001 Thessaloniki, Greece.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2017.2727143

This paper introduces an interactive visual analytics platform that comprises tools for recommendations about effective measurements towards the reduction of road congestion for real-time or future traffic conditions, by exploiting the principles of trial-and-error and hypothesis testing. A number of transport information services, such as road clustering, anomaly detection and traffic prediction, are integrated within the visual analytics platform allowing for measuring all critical parameters that might affect urban congestion, such as real-time vehicle occupancy. Based on the measured parameters, the goal of the visual analytics platform is to provide useful insight on those measures that would enable optimal use of available space, whilst reducing congestion in such levels to ensure both minimum energy consumption and environmental footprint, but also maximum accessibility and quality of life for citizens. Through its architecture, a major contribution of the proposed toolkit is the ability to combine several back-end data analysis algorithms into a common interface, allowing data exploration and decision making.

The rest of the paper is structured as follows. Section II reviews related work on the three technological pillars on which our visual analytics platform relies, i.e., Visualization toolkits, Traffic Prediction, and Data Clustering. Section III presents the system architecture and its structural components, whereas Section IV describes in more detail the aforementioned key technological pillars of the visual analytics platform. In Section V, we present the results obtained after comparing the traffic prediction accuracy component of our platform to similar approaches, in order to select the one that achieves the best performance. Section VI presents a set of representative use cases that showcase the advanced capabilities of our platform, and finally Section VII concludes and outlines the paper.

II. Related Work

A. Visualization Toolkits

There are several traffic-related visual analytics and exploration methods available in the literature, differing in the input data used and in the visualization techniques employed. Wang *et al.* [1] introduce an interactive tool for visual analysis and exploration of urban traffic congestion. The tool extracts traffic jam information from GPS trajectories and then allows the user to make multilevel exploration, from traffic patterns on a single road to traffic jams in a whole city. Chu *et al.* [2] present a system that discovers, analyzes and visualizes the

hidden traffic patterns by studying the movement of cars from large trajectory datasets. In more recent works, graph-based modeling of urban network data have been used to visually analyze taxi trajectory data and examine network centralities [3]. A detailed review of recent visual analytics methods and systems, targeted at traffic control and urban computing is presented in [4].

Recently, significant work has been made regarding systems combining analytical methods with visualization, in an effort to overcome the limitations of using analytical methods with data of increasing size (big data). Visual analytics methods have been used to interactively train classifiers for public transport data [5], to combine multiple traffic parameters and understand the dependencies between them [6], and to detect anomalies in the input data [7]. Visual analytics have also been used for causal reasoning, through the combination of visual elements and interactions with automatic causal reasoning algorithms [8]. Towards data clustering, Rinzivillo *et al.* [9] introduce the concept of progressive clustering as a generic tool to explore and analyze large traffic datasets, using a large number of trajectories. Andrienko *et al.* [10] combine clustering and classification with human interaction, where an analyst directs the work of the computer towards the discovery of meaningful, relevant clusters.

B. Traffic Prediction

One of the basic usages of the hereby proposed visualization toolkit is the prediction and visualization of traffic over short-term future intervals. Short-term traffic prediction algorithms can broadly be classified into parametric and non-parametric models. The parametric models are based on time series analysis. Most of them are based either on the classic Box and Jenkins Auto-Regressive Integrated Moving Average (ARIMA) model, such as the work by Williams and Hoel [11], or on Kalman Filtering (Mannini *et al.* [12]). Non-parametric models on the other hand have been inspired by Artificial Intelligence techniques. Indicative efforts in this category include the works by Myung *et al.* [13] and Zheng and Su [14], which are based on the k-Nearest Neighbor (kNN) algorithm, by Wu *et al.* [15] and Hu *et al.* [16], which apply Support Vector Regression (SVR), and the one by Zhu *et al.* [17], which makes use of Artificial Neural Networks (ANN).

A more recent approach is the exploitation of the spatiotemporal correlations between elements of the traffic networks in order to build more accurate models. Diamantopoulos *et al.* [18] enhance a Space-Time ARIMA (STARIMA) model with the traffic values of the neighboring roads for a road of interest. The correlated neighbors are discovered using a Pearson correlation-based metric called Coefficient of Determination (CoD). This work was extended in [19], where a mechanism is developed for efficiently calculating the CoD values between all pairs of roads for large-scale traffic networks.

C. Dimensionality Reduction and Visual Clustering

Mapping roads as points on a two-dimensional screen has been used as a means to allow the operator to visually distinguish clusters of roads with similar characteristics and behaviors. This task is closely related to dimensionality reduction, since usually high-dimensional attributes need to be projected on a 2D plane. Manifold learning methods for dimensionality reduction attempt to discover a low-dimensional manifold lying in the high dimensional space, based on graph structures constructed from the data. They include unsupervised methods, such as Locality Preserving Projections (LPP) [20], as well as supervised ones, such as Marginal Fisher Analysis (MFA) [21] and Local Discriminant Embedding (LDE) [22]. In the work of [21], several manifold learning methods are described as instances of a general framework for dimensionality reduction, namely Graph Embedding (GE) dimensionality reduction. Force-directed placement algorithms have also been used to embed graphs in a low dimensional space, in an intuitive and visually pleasing way [23].

Most of the existing visualization methods use a single attribute of behavior (unimodal) in order to position the entities into the graph. However, combining multiple attributes, in a *multimodal* setting, can provide better insights and visualizations regarding the data. Existing methods of multimodal visualization include methods that simultaneously combine characteristics of all modalities, e.g., through a weighted sum of graph Laplacians [24], or through Multiple Kernel Learning [25]. They also include methods that utilize information of one modality to assist learning in another modality [26]. Finally, following a different principle, the work of [27] formulates multimodal visualization as a multi-objective optimization problem, resulting in a set of Pareto-optimal solutions, instead of one. Even though such methods verify the effectiveness of multimodal fusion, to the best of our knowledge they have not been adopted for road traffic visualization, where the combination of multiple modalities may reveal important aspects for anomaly detection and decision making.

D. Contribution of Proposed Platform

The contribution of the proposed platform can be summarized to the following:

- Employment of visual analytics techniques for visually identifying clusters of roads with common characteristics, which is not supported by commercial tools for traffic planning, such as [28] and [29].
- Combination of multiple types of features, including user-defined ones, permitting the detection of more informative patterns, which is usually not provided by existing tools of the literature (Section II-A), which focus on the visualization of a single fixed type of feature.
- Unification of traffic prediction and hypothesis testing methods with visual analytics, in an interactive manner, supporting the configuration of algorithm parameters through visual interaction.

III. System Architecture

A high level architecture of the proposed platform is illustrated in Fig. 1. The system is composed of three main building blocks. The Preprocessing block involves the preprocessing of

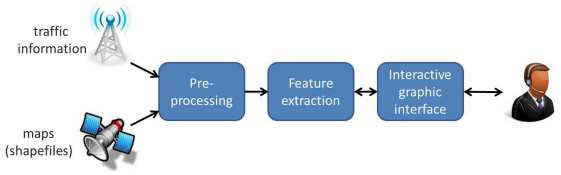

Fig. 1. The conceptual architecture of the proposed visual analytics toolkit.

the data retrieved by online notification systems (e.g., RDS), traffic monitoring data by distributed sensors (e.g., loop detectors), as well the road network structure itself (e.g., shapefiles). The Features block supports the combination of a wide variety of algorithms that are efficiently scheduled in order to produce the appropriate features, based on the preprocessed data, for each application configuration. The Interactive Graphical User Interface block involves an advanced visualization engine that allows the interactive manipulation of all entities (e.g. road segments) through various types of analyses (anomaly detection, clustering, dynamic hypothesis formulation, etc.).

In a real case scenario, the envisioned pipeline of the system's operation would start with the passive monitoring by the operator, until the anomaly detection module triggers an alarm. At that point, the operator should actively be involved, in order to find the best solution that resolves the incident. To this direction, their interaction with it would include several hypothesis scenarios and their simulation based on past data, trained models and optimization algorithms, that are expected to support the operator in decision making.

IV. Back-End Description

A. Dimensionality Reduction and Visual Clustering

As a means for visual clustering of the data, the multi-objective visualization method of [27] is used in the toolkit. Compared to state-of-the-art dimensionality reduction and visualization techniques, which only consider a single feature, the multi-objective visualization method has the significant advantage of combining multiple features extracted from the available data. This multimodal combination can be of significant value to operators, since they can combine multiple traffic-related parameters and be able to discover patterns that may not be apparent when a single parameter is considered. Compared to other multimodal dimensionality reduction methods, the multi-objective visualization method manages to uncover solutions that other fusion-based methods cannot [27]. The above advantages have led to the selection of the multi-objective visualization method in the proposed platform.

The multi-objective visualization method considers a set of objects represented by a multitude of descriptor vector types (modalities), and maps them on the 2D plane, so that points that are close to each other correspond to objects that are similar with respect to all descriptor types. For instance, in this work, the objects could be the roads and the multiple descriptor types could be the traffic volume, traffic speed, etc., characterizing each road. This positioning problem is formulated as a multi-objective optimization problem, where each objective corresponds to a descriptor type and its minimization leads to an optimal placement for the specific descriptor type. Solving for these multiple competing objectives results in a Pareto-optimal set of solutions, each corresponding to a different trade-off among the multiple objectives. The reader is referred to [27] for more details. In the proposed toolkit, the operator can select among the various trade-offs in order to put more focus on one feature or another in the visualization.

B. Anomaly Detection

For the purposes of hypothesis testing, an anomaly detection method is used by the toolkit, in order to determine if the traffic measurements for a road at a specific time instance are unusual, provided the historical measurements of the road. The Local Outlier Factor (LOF) [30] is used for this purpose. The LOF method measures the outlierness of measurements by examining their sparsity compared to other normal instances. Hereby, a slightly modified version of LOF is used, in order to also incorporate training data in the procedure, as in [31]. LOF scores higher than 1 indicate that an item is an outlier.

The LOF method was selected over other existing state-of-the-art anomaly detection methods, such as CUSUM, HMM and BRPCA, all compared in [31], as a sufficient trade-off between simplicity in the design, accuracy and efficiency, which are all considered as important factors for an interactive platform such as the proposed. Regarding simplicity and accuracy, the implementation of methods such as CUSUM is simpler, but lacking accuracy. LOF, mostly based on a k-nearest neighbor search, achieves sufficient accuracy, while at the same time having an implementation that is simpler than HMM, which requires a sophisticated training phase, or BRPCA, which is based on Singular Value Decomposition. The fact that the computational bottleneck of LOF is in the computation of the nearest neighbors, which is relatively fast, is also an advantage of LOF compared to e.g., HMM, in terms of computational efficiency. Moreover, although LOF, in the supervised version utilized herein, relies on training data, while BRPCA is unsupervised, LOF is not based on assumptions about linear dependency of the variables, as is BRPCA [31], which makes it able to handle more general anomaly detection problems. However, it should be noted that the modular design of the proposed platform allows for the use of various anomaly detection methods, so as to incorporate more accurate and more efficient methods in the future.

C. Traffic Prediction

Central to the architecture of our system is a prediction module that predicts future traffic over short-term intervals. The module takes as input historical traffic data in the form of time series in order to build a prediction model, and then applies this model to real time traffic data in order to make short-term predictions, e.g., predict the traffic flow up to 1 hour in the future. We wanted to evaluate the performance of both parametric and non-parametric models, in order to choose the one that we will use in the proposed toolkit. Therefore, we selected ARIMA [32] as a parametric model and k-NN [13] and SVR [15] as non-parametric models. ARIMA is one of the most widely used parametric models for traffic

prediction in the literature. This model has been used from various researchers and in many different forms (STARIMA, VARIMA, KARIMA, etc.) leading to high prediction accuracy results which was the main reason for selecting it. On the other hand, k-NN and SVR are non-parametric prediction models (memory-based and model-based respectively) which are characterized by both precision and simplicity. Finally, these models have been used by various researchers in the field of ITS in general, and their results (accuracy and performance) can be cross-referenced and cross-validated.

The main difference between the kNN-based [14], [33], [34] and the ARIMA-based [11], [35], [36] traffic prediction models is that the former do not include a training process, meaning that they do not require the estimation of parameters. They find the k closest neighbors, and use their values to calculate the prediction. On the other hand, the latter are parametric models, meaning that they include a training process in which a set of parameters (the ϕ parameters) needs to be estimated. This leads to an approximation problem, which can be solved using the OLS method. Since the parameters are estimated, the dot product between them and the vector of inputs is calculated and the result is the prediction of the ARIMA-based models.

V. Evaluation

A. Dataset Description

The datasets used in the current work were obtained from the Performance Measurement System (PeMS) of the California Department of Transportation (Caltrans). This system is composed of more than 39,000 Vehicle Detection Stations (VDSs) that track the flow and speed of vehicles, scattered all over the freeway system of all major metropolitan areas of the State of California. For demonstration purposes, we selected a square area of approximately 32 km by 32 km in Caltrans District 4, San Jose. After removing stations with missing or erroneous data, we ended up with a subset of 506 VDSs that cover the aforementioned area. The traffic dataset contains traffic flow and speed measurements for these VDSs, aggregated and averaged in 5-minute time intervals. The time period covered by our dataset is four months, from May 1, 2015 to August 31, 2015. For each day of the total 4-month period, time series of traffic flow and speed per 5-minute intervals were constructed, resulting in sets of 123 time series of size 288, for each VDS. We used the first 3 months as the training set for our prediction models, and the last month both as the test set to measure the prediction accuracy of the model and for the presentation of the various use cases. Fig. 2 illustrates the location of the selected VDSs over the map.

PeMS also maintains a database with incidents collected by the California Highway Patrol (CHP). Each incident is characterized by a unique ID along with a timestamp, the freeway and the exact coordinates where it occurred, its type (e.g., Traffic Hazard), its duration in minutes, and other related information. We obtained the full set of incidents that occurred in District 4 during the four month time period we examine (37,280 incidents in total, 5,841 of which happened in the last month and near the 506 VDSs that we selected).

Fig. 2. The geographic distribution of the VDSs over the map.

B. Traffic Prediction Benchmark

To measure the accuracy of the various traffic prediction models, we performed a number of experiments, using the set of time series for each VDS described in Section V-A. As presented in Section IV-C, three different models were used for traffic prediction, namely ARIMA, kNN, and SVR. The metric used to measure the prediction accuracy of the models is the *Normalized Root Mean Square Error* (*NRMSE*) given by Equation (1).

$$NRMSE = \frac{\sqrt{\sum_{i=1}^{N} \frac{(y_i - y_{i,pred})^2}{N}}}{y_{i,max} - y_{i,min}} \quad (1)$$

By observing the original traffic time series, we can see that traffic presents different patterns in different time periods of the day. Furthermore, several different traffic patterns may exist within the same time period. These patterns can be identified using an unsupervised learning algorithm, and a prediction model can be built only for a specific part of a specific time period of a day. This model will be more accurate compared to one that was constructed for the whole day. In this context, we introduce a methodology that exploits these findings in order to improve the prediction accuracy.

In the first step of this methodology, the original time series of a road are split into segments of 1-hour duration, i.e., 24 segments. Then, the partial time series of the segment are clustered using the density-based spatial clustering of applications with noise (DBSCAN) algorithm. The advantages of this algorithm is that it does not require a predefined number of clusters and that it can identify clusters of arbitrary shape (not only convex clusters). The clustering process is repeated for the partial time series sets of all segments. Finally, for each cluster of each segment, a different prediction model (in terms of different parameter values) is constructed. More details can be found in [37].

The aforementioned methodology was combined with the kNN, SVR and ARIMA base algorithms, in order to produce the Enhanced models. An Enhanced model means that the aforementioned methodology was followed and the corresponding base algorithm (kNN, SVR or ARIMA) was used for

TABLE I
PREDICTION ERROR (NRMSE) FOR DIFFERENT MODELS, WITHOUT ("BASE") AND WITH ("ENHANCED") OUR METHODOLOGY

	Base	Enhanced	Diff (%)
SVR	12.76	5.67	-55.56
ARIMA	16.52	13.42	-18.76
kNN	13.31	13.14	-2.02

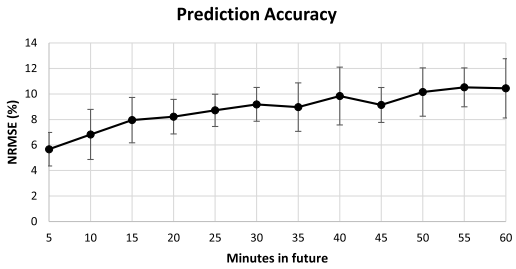

Fig. 3. Prediction accuracy of Enhanced SVR for different time intervals in the future.

prediction. All three enhanced models were tested, in terms of prediction accuracy, and the one that derived the best results was the Enhanced SVR model. This model is the one used by the proposed visualization kit.

In order to evaluate the accuracy of our methodology, we compared it with the base ARIMA, k-NN and SVR models using the NRMSE metric. Initially we trained and ran the three benchmarking models on the dataset as a whole without using the proposed methodology. Then we implemented and tested the proposed method, using each time a different prediction model for the generated clusters. The results of these experiments are shown in Table I. All three models present superior accuracy when used in conjunction with the proposed method compared to their unitary form. More specifically, the use of the proposed method reduces the prediction error of the SVR, ARIMA and k-NN models by 55.56%, 18.76% and 2.02% respectively. Also, the Enhanced SVR presents better accuracy compared to the other two enhanced models. Based on these, the choice of the Enhanced SVR for prediction model of the proposed visual analytics toolkit is justified.

Our visualization toolkit focuses on short-term traffic prediction, i.e., predicting the value of a traffic variable (speed or flow) up to one hour in the future. We create a different model for each future interval we want to make predictions (e.g., 5 minutes, 10 minutes, etc) as described in [37]. Fig. 3 shows the NRMSE, as well as its variance, when a prediction is made for $5, 10, \ldots, 60$ minutes in the future. As shown, the prediction error remains low (e.g., below 11%) for all intervals.

VI. EXAMPLES OF USE

Supported by the theoretical foundations presented in Section IV as the back-end, the proposed platform provides the following functionalities to the operator, presented in this section, covering a large number of tasks that an urban traffic operator is interested in: (a) Visual exploration of historical data, (b) Visual testing of hypotheses, (c) Traffic prediction,

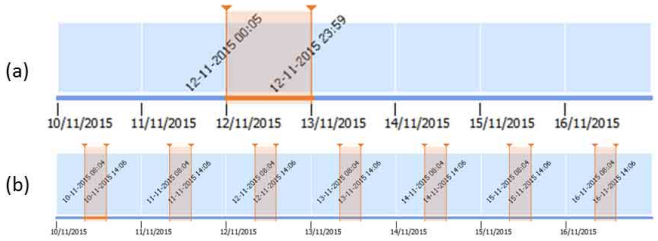

Fig. 4. Selection of time period of interest for exploratory visualization.

(d) Conditional traffic prediction. In the current work, the main network element of interest is the *link*, which is defined as the line segment between two network nodes, whether they are intersections or intermediate nodes. The data for a link consist of an aggregation of the data originating from the traffic sensors assigned at this link.

A. Visual Exploration of Historical Data

Exploration of historical data means offering a number of different ways to collectively view various aspects of the existing data. In the proposed tool, exploration is largely accomplished by transforming roads in a behavior-related domain and visualizing them as points on a two-dimensional plane. Various behavioral aspects of a road can be considered, such as the traffic volume or the average speed in predefined time intervals. The multi-objective visualization approach is used to combine multiple behavioral aspects and then present visualizations to the user. The results are linked to the map view, in order for the operator to see which roads exhibit the behavioral patterns visualized.

In an example scenario, the operator desires to detect abnormal events in the data collected from the immediate past, in order to either see how reported incidents have affected traffic, or discover incidents that have not yet been reported. As a first step, the operator selects a time period of interest, e.g., the previous day, as illustrated in Fig. 4. Any further visualizations will be based on the data of this selected time period. The time period selection can be linear, from a specific time instance up to another, as in Fig. 4(a). It can also be non-linear, e.g., repeating a selected time period over a week, as in Fig. 4(b), where the morning hours (08:00 - 14:00) are selected from each day, or using other filtering procedures.

The operator can view the collected data in the map view, which is one of the central parts of the tool. Fig. 5 depicts the traffic volume for the roads of interest. The traffic volume has been normalized using a z-score normalization. Values close to zero denote an expected traffic volume, close to the mean value, while larger values denote an increased or decreased traffic volume, compared to the standard deviation σ. The z-score values are quantized in a three-color scale: blue (values close to zero, $< 1\sigma$), orange ($< 3\sigma$) and red ($> 3\sigma$). Using this visualization, the operator can instantly see areas deviating from expected traffic, which may indicate abnormal incidents.

The traffic volume for each road can be presented either for the whole time period of interest, or as successive values covering the minimum time interval considered, hereby

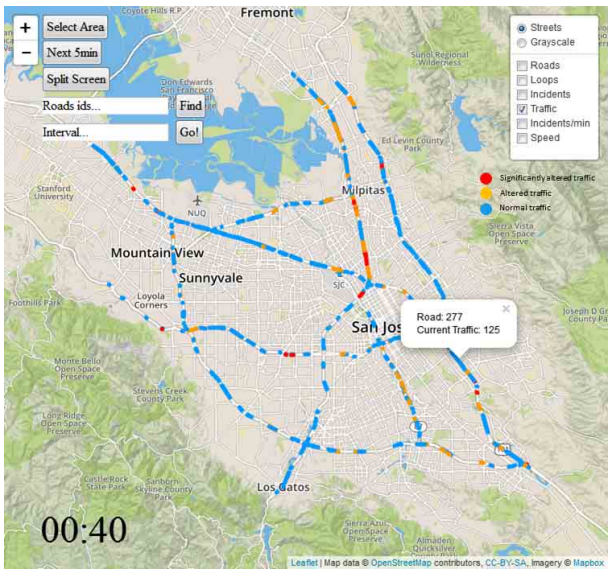

Fig. 5. Map view of the normalized traffic volume for a 5-minute interval. The blue colors denote expected traffic volume, while red colors denote a large deviation from the expected traffic volume.

5-minute intervals. In the second case, the operator can view the data in successive steps, thus having an overview of the traffic evolution through time. The map visualization supports multiple overlays, so that other attributes apart from traffic volume can be visualized as well. The available attributes include the traffic volume, the average speed, the number of reported incidents, as well as different versions of these attributes, smoothed over time or normalized with respect to their mean values and their standard deviations, as computed over larger previous historical periods. Although viewing individual attributes is itself useful for the operator, combining multiple attributes is even more powerful in discovering abnormal events. For this purpose, the operator can select a multitude of the above mentioned attributes, in order to be used in the multi-objective visualization method. In multi-objective visualization, each road is represented as a point on the 2D plane, so that nearby points correspond to roads with similar behaviors with respect to the combined attributes.

Continuing the example, the operator selects for instance the smoothed traffic and speed attributes, and provides them as input to the multi-objective visualization method. The results are presented in Fig. 6. The central part of the figure is occupied by the multi-objective visualization, combining the two features, while the right-most panes show the visualizations of the two features individually. As is apparent from the central multi-objective visualization, most of the road points are gathered in a large cluster, denoting similar behaviors with respect to the combination of attributes. However, there is a small group of roads towards the bottom of the visualization, shown in red, which deviate from the large cluster, forming a smaller one. This deviation from the behavior of the other roads indicates that these roads may be involved in an unusual event. The operator can select the roads of this smaller group, in order to perform further analysis. Due to the linking between the various views of the tool, the selected roads are also highlighted in red in the map view of Fig. 6. It appears that most of the selected roads are upon Bayshore freeway, suggesting that there may indeed be an incident on this road.

B. Visual Hypothesis Testing

The exploration methods presented in the previous section provide "horizontal" analyses, considering data of multiple roads for a relatively small time window. The proposed tool also provides means for "vertical" analyses, by considering statistical data collected for each road over a large historical time period. Such analyses are provided in the form of specific types of hypotheses, which are verified visually by the operator. The tool supports two types of hypotheses: (a) whether specific roads exhibit abnormal behavior with respect to their history and (b) whether specific geographical areas exhibit relatively increased amount of accidents.

1) Unusual Road Behavior: The multi-objective visualization view of Fig. 6 indicates that there is a group of roads with unusual behavior compared to the other roads of the network. However, such a behavior may be expected, e.g. in case the involved roads are those surrounding a stadium and there is a congestion in this area every Sunday, at the time of an athletic event. In order to distinguish such cases, the unusual road behavior hypothesis tests if the roads of the network exhibit unusual behavior with respect to their history. When the operator selects this type of hypothesis, the system retrieves historical data of the roads and provides them as training input to the LOF anomaly detection method, (Section IV-B). A LOF score is computed for each road, indicating whether this road exhibits abnormal behavior with respect to its historical traffic behavior. Continuing the example of the previous section, Fig. 7 depicts the LOF scores computed for the roads of the considered geographical area. A continuous color scale is used, from blue (low LOF scores - regular behavior) to red (high LOF scores - outlier). It can be observed that most of the roads with the highest LOF scores belong to the small group deviating from the other roads in the multi-objective visualization. This suggests that this deviation is not something that is expected and should be treated accordingly.

2) Geographical Areas With Increased Amount of Incidents: As a second type of hypothesis, the operator can see if specific areas of the map demonstrate an increased rate of incidents throughout a selected period of time. The results are presented in the form of a heatmap, layered over the map view. The number of incidents happening on a road have been used as the "temperature" of the corresponding point on the map, which is diffused to the nearby area, using Gaussian heat kernels. The visualization resulting from all the available data is an intuitive way for the operator to distinguish the "hot" areas, i.e., ones which exhibit a large number of incidents and thus require the attention of the urban planner, in order to reduce them. Fig. 8 depicts an example of this type of visualization, for incident data collected over a period of one month. It is apparent that the eastern area along the Sinclair freeway is the one with the largest amount of incidents.

For this type of hypothesis, the tool also supports a quantitative statistical assessment of whether any difference in the

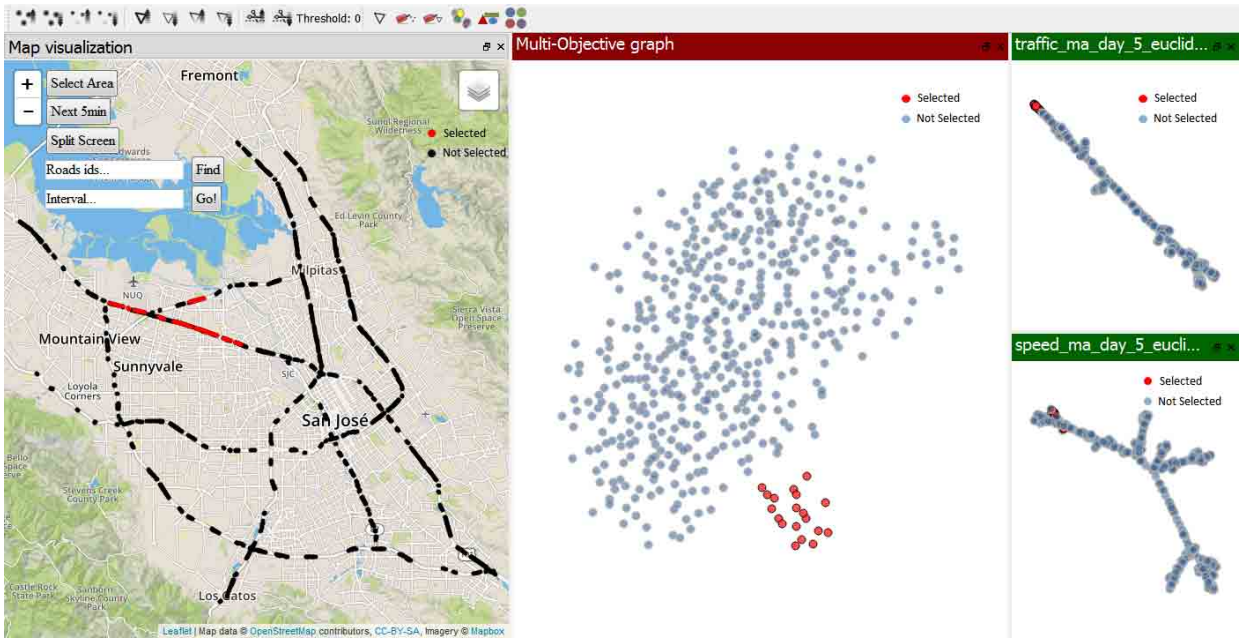

Fig. 6. Multi-objective visualization of roads from the Caltrans dataset. The rightmost panes show the individual visualizations of the two selected features (traffic volume and speed), while the middle plane depicts their combined visualization. Each point represents a road, with nearby points representing roads with similar characteristics with respect to both features. A small group of road points at the bottom of the image appears detached from the large cluster of roads, indicating that they exhibit very different behavior in the time period of interest, hereby the 5th day of the dataset. The operator has selected these points (red) and the selection is linked to the map in the left pane. The selected points correspond to the right lane of a particular road. Indeed, an accident had happened at the Bayshore-Moffet intersection, which caused a traffic jam along the lane.

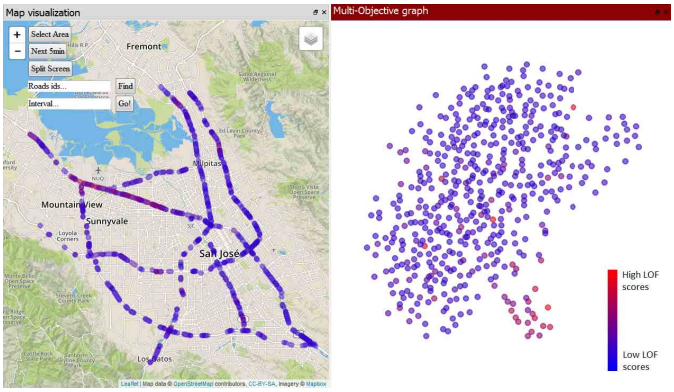

Fig. 7. Visualization of the road LOF scores. Blue color denotes low LOF scores, i.e., regular behavior, while red colors denote high LOF scores, i.e., abnormal behavior. The roads of the small group that deviates from the rest have large anomaly scores, which is another clue that there is an unusual situation in the corresponding roads.

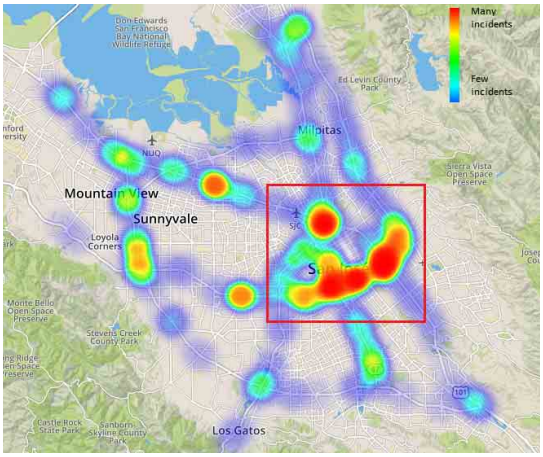

Fig. 8. Heatmap of the number of incidents happening in the roads of the considered region. The colorscale indicates the number of incidents, from blue (few incidents) to red (many incidents). The area bounded by the red box is the one with the most recorded incidents.

number of incidents between a selected area and the whole area could occur by chance. Considering that the number of incidents happening in a road during a specified time period, e.g., a month, follows a Poisson distribution, the null hypothesis is formed, stating that the mean of the distribution of the roads in the selected area, hereby the one in the red box of Fig. 8, is not larger than the mean of the roads in the whole area. The Przyborowski-Wilenski method is followed in order to test this hypothesis against the alternative hypothesis that the mean of the selected area is larger than the mean of the whole area. For the example of Fig. 8, there is statistically significant evidence ($p = 2.5 \cdot 10^{-41}$) to support that the selected area is more prone to incidents compared to the other areas. This quantitative result is presented to the operator along with the visualization, and can be of great use to the urban planner.

C. Traffic Prediction

A major functionality of the toolkit is the prediction of future traffic-related attributes, such as the volume and average vehicle speed, given the current traffic state. The SVR model presented in Section IV-C is used as the prediction model.

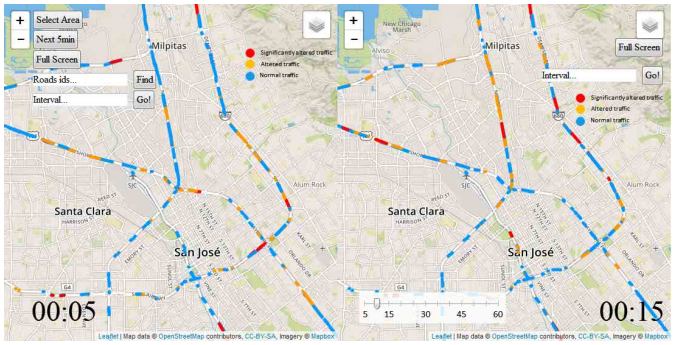

Fig. 9. Split view of the map, used for making predictions. The left side shows the current traffic, while the right side shows the predicted traffic for a future time instance selected by the operator.

Fig. 10. Example of the use of the prediction slider. The operator can select how far in the future to predict, by moving a slider. Prediction for time instances after (a) 10 minutes, (b) 20 minutes and (c) 50 minutes are depicted.

The graphical user interface of the toolkit facilitates the operator in making predictions. The map view can be split into two views, as presented in Fig. 9. The left view shows the current traffic, while the right one shows the predicted traffic for a future time instance. The operator can select how far in the future to predict, in a range from 5 minutes to 1 hour, by using a slider, as depicted in the figure. Fig. 10 illustrates an example of prediction, where the operator, by altering the slider position, views the predicted traffic for various future time instances. Of course, the further away in time one looks, the less accurate the prediction becomes (see Section V-B).

D. Conditional Traffic Prediction

A second prediction-related functionality provided by the toolkit is the ability to impose an incident on a selected road, e.g., a hypothetical accident or a road closure, and predict how the traffic will be affected in the next minutes. This allows the operator to test hypothetical scenarios and possibly prevent any future undesired situations. This is accomplished by considering, after the operator imposes an incident, a different prediction model than the one used for regular traffic. This incident-related prediction model is constructed for each road using the same theoretical methodology as the SVR model used for normal traffic, but using the time series of previous incidents on the same road as the training data. When the operator puts an incident on a road, the regular prediction model of this road, as well as of all the roads that are affected

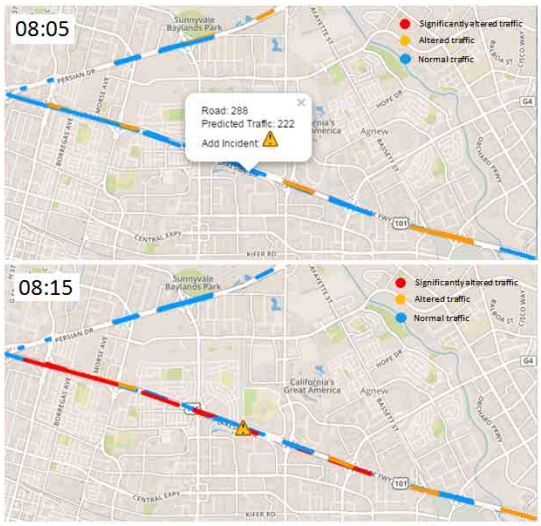

Fig. 11. Example of traffic prediction after a hypothetical incident, imposed by the operator. The operator can place an incident on a road through the graphical user interface, as illustrated in the top image, and view the predicted traffic flow for the next minutes.

by the incident, is substituted by the incident-related model, which is then followed for the next time instances.

The multi-objective visualization is utilized in order to determine which roads are affected by the incident. The affected roads are selected as the nearest neighbors of the target road in the multi-objective visualization, using the traffic volume and speed as the combined attributes, or attributes selected by the operator. Fig. 11 depicts an example of conditional prediction. At the top image, the operator places an incident on a selected road, using the graphical user interface. The next image illustrates the prediction of the traffic volume for the next few minutes. It can be observed that, for the next minutes, the traffic volume of the road passing from the point of the incident is significantly affected.

E. User Study

An online survey[1] has been conducted, participated by 6 professionals in traffic engineering and related research. The participants were presented with the proposed tool and were asked for their opinion regarding its various parts. Half of the participants work on monitoring traffic and analyzing data of large urban networks. Half of the participants consider that an interactive visualization tool would be essential for their work, although only 33% of the participants already use such a tool. The participants had a very positive first impression for the proposed tool and considered it as something they would need, although not all of them would replace the tool they already use with the proposed one.

Regarding visualization, features such as representing the roads as points on a two dimensional plane, connected to the actual geographical map, using different colors for normal and abnormal road behavior, and clustering them into groups of similar characteristics, were considered useful, as they might reveal significant traffic patterns. Most of the participants (67%) considered the heatmap as a useful and appeal-

[1] Available online at: https://goo.gl/nWHKz1

ing visualization tool. On the other hand, the participants suggested that the user interface of the platform could be improved. Traffic prediction was considered by all participants as an important feature that would significantly assist them in monitoring. Moreover, the ability to manually insert an incident in the network and examine how its impact is dissipated was considered very useful by all participants. Overall, the comments of the participants suggest that the features available in the proposed platform are useful to the traffic engineers and researchers. Based on their comments, future work on the platform will aim improve the developed technologies and the weak parts.

F. Computational Complexity

A significant part of the computational power required by the proposed platform is occupied by the multimodal visualization method. The visualization method, presented in Section IV-A, is based on force-directed placement of the points on the screen, which can have a time complexity of $O(N \log N)$. In an Intel i7 - 3.5GHz processor, with threading enabled, this makes the platform able to handle a number of 5000 VDSs in about 7 s. The Caltrans dataset considered hereby uses about 500 VDSs to represent the major roads of the city of San Jose, having a population of about a million people. Roughly, this makes the platform able to handle a large city of 10 million people, requiring about 5000 VDSs, in a relatively small amount of time (7 s).

VII. Conclusions

The main contribution of the presented work comprises a novel visual analytics tool that provides an efficient and cost-effective means for testing the impact of different spatial interventions on vehicular congestion by exploiting advanced visualization techniques. Another contribution of our platform is that it enables the exploration of traffic dependencies between roads in urban networks by appropriate visual means. This can provide useful insight about the spatial impact of a particular intervention to a particular area. It can also assist on the selection of the minimum cost interventions that can result in the maximum impact on current congestion.

Moreover, this paper contributes by providing a set of state-of-the-art underpinning algorithmic implementations integrated in the same framework. As a result, a set of sophistication capabilities are enabled through a common tool. These represent three distinguished but interoperable and orchestrated components, namely data visualization and clustering, anomaly detection, and traffic prediction. The integration of the aforementioned functionalities into the core of our visual analytics tool, enables support for a number of novel use cases that were not supported by current ITS visualization platforms, especially through its advanced visualization and hypothesis testing capabilities. Our platform goes beyond the visual exploration of historical data to the identification of unusual road behavior, as well as roads with increased occurrence of incidents, and those causing congestion. Furthermore, conditional traffic modeling is supported, for different types of occurring incidents.

Future work includes potential improvements on the provided technological contributions for further boosting the accuracy of traffic prediction, but also aesthetic improvements based on user feedback. A wide scale deployment of the current backed architecture on cloud, as well as the webification of the user interface are also foreseen as future tasks.

References

[1] Z. Wang, M. Lu, X. Yuan, J. Zhang, and H. van de Wetering, "Visual traffic jam analysis based on trajectory data," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 12, pp. 2159–2168, Dec. 2013.

[2] D. Chu *et al.*, "Visualizing hidden themes of taxi movement with semantic transformation," in *Proc. IEEE Pacific Vis. Symp.*, Mar. 2014, pp. 137–144.

[3] X. Huang, Y. Zhao, C. Ma, J. Yang, X. Ye, and C. Zhang, "TrajGraph: A graph-based visual analytics approach to studying urban network centralities using taxi trajectory data," *IEEE Trans. Vis. Comput. Graphics*, vol. 22, no. 1, pp. 160–169, Jan. 2016.

[4] Y. Zheng, W. Wu, Y. Chen, H. Qu, and L. M. Ni, "Visual analytics in urban computing: An overview," *IEEE Trans. Big Data*, vol. 2, no. 3, pp. 276–296, Sep. 2016.

[5] L. Yu *et al.*, "iVizTRANS: Interactive visual learning for home and work place detection from massive public transportation data," in *Proc. IEEE Conf. Vis. Anal. Sci. Technol. (VAST)*, Oct. 2015, pp. 49–56.

[6] S. M. Arietta, A. A. Efros, R. Ramamoorthi, and M. Agrawala, "City forensics: Using visual elements to predict non-visual city attributes," *IEEE Trans. Vis. Comput. Graphics*, vol. 20, no. 12, pp. 2624–2633, Dec. 2014.

[7] P. S. Quinan and M. Meyer, "Visually comparing weather features in forecasts," *IEEE Trans. Vis. Comput. Graphics*, vol. 22, no. 1, pp. 389–398, Jan. 2016.

[8] J. Wang and K. Mueller, "The visual causality analyst: An interactive interface for causal reasoning," *IEEE Trans. Vis. Comput. Graphics*, vol. 22, no. 1, pp. 230–239, Jan. 2016.

[9] S. Rinzivillo, D. Pedreschi, M. Nanni, F. Giannotti, N. Andrienko, and G. Andrienko, "Visually driven analysis of movement data by progressive clustering," *Inf. Vis.*, vol. 7, nos. 3–4, pp. 225–239, 2008.

[10] G. Andrienko, N. Andrienko, S. Rinzivillo, M. Nanni, D. Pedreschi, and F. Giannotti, "Interactive visual clustering of large collections of trajectories," in *Proc. IEEE Symp. Vis. Anal. Sci. Technol. (VAST)*, Oct. 2009, pp. 3–10.

[11] B. M. Williams and L. A. Hoel, "Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results," *J. Transp. Eng.*, vol. 129, no. 6, pp. 664–672, 2003.

[12] L. Mannini, S. Carrese, E. Cipriani, and U. Crisalli, "On the short-term prediction of traffic state: An application on urban freeways in ROME," *Transp. Res. Procedia*, vol. 10, pp. 176–185, Jan. 2015.

[13] J. Myung, D.-K. Kim, S.-Y. Kho, and C.-H. Park, "Travel time prediction using k nearest neighbor method with combined data from vehicle detector system and automatic toll collection system," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2256, pp. 51–59, Dec. 2011.

[14] Z. Zheng and D. Su, "Short-term traffic volume forecasting: A k-nearest neighbor approach enhanced by constrained linearly sewing principle component algorithm," *Transp. Res. C, Emerg. Technol.*, vol. 43, pp. 143–157, Jun. 2014.

[15] C.-H. Wu, J.-M. Ho, and D. T. Lee, "Travel-time prediction with support vector regression," *IEEE Trans. Intell. Transp. Syst.*, vol. 5, no. 4, pp. 276–281, Dec. 2004.

[16] W. Hu, L. Yan, K. Liu, and H. Wang, "A short-term traffic flow forecasting method based on the hybrid PSO-SVR," *Neural Process. Lett.*, vol. 43, no. 1, pp. 155–172, 2016.

[17] J. Z. Zhu, J. X. Cao, and Y. Zhu, "Traffic volume forecasting based on radial basis function neural network with the consideration of traffic flows at the adjacent intersections," *Transp. Res. C, Emerg. Technol.*, vol. 47, pp. 139–154, Oct. 2014.

[18] T. Diamantopoulos, D. Kehagias, F. G. König, and D. Tzovaras, "Investigating the effect of global metrics in travel time forecasting," in *Proc. 16th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2013, pp. 412–417.

[19] A. Salamanis, D. D. Kehagias, C. K. Filelis-Papadopoulos, D. Tzovaras, and G. A. Gravvanis, "Managing spatial graph dependencies in large volumes of traffic data for travel-time prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 6, pp. 1678–1687, Jun. 2016.

[20] X. He and P. Niyogi, "Locality preserving projections," in *Proc. Neural Inf. Process. Syst.*, vol. 16. 2004, p. 153.

[21] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.
[22] H.-T. Chen, H.-W. Chang, and T.-L. Liu, "Local discriminant embedding and its variants," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2. Jun. 2005, pp. 846–853.
[23] T. M. J. Fruchterman and E. M. Reingold, "Graph drawing by force-directed placement," *Softw., Pract. Exper.*, vol. 21, no. 11, pp. 1129–1164, 1991.
[24] H. Tong, J. He, M. Li, C. Zhang, and W.-Y. Ma, "Graph based multi-modality learning," in *Proc. 13th Annu. ACM Int. Conf. Multimedia*, 2005, pp. 862–871.
[25] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh, "Multiple kernel learning for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 6, pp. 1147–1160, Jun. 2011.
[26] K. Nigam and R. Ghani, "Analyzing the effectiveness and applicability of co-training," in *Proc. 9th Int. Conf. Inf. Knowl. Manage.*, 2000, pp. 86–93.
[27] I. Kalamaras, A. Drosou, and D. Tzovaras, "Multi-objective optimization for multimodal visualization," *IEEE Trans. Multimedia*, vol. 16, no. 5, pp. 1460–1472, Aug. 2014.
[28] PTV Group. *PTV Vissim*. Accessed 2017. [Online]. Available: http://vision-traffic.ptvgroup.com/en-us/products/ptv-vissim/
[29] TSS-Transport Simulation Systems. *Aimsun*, Accessed 2017. [Online]. Available: https://www.aimsun.com/
[30] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," *ACM SIGMOD Rec.*, vol. 29, no. 2, pp. 93–104, 2000.
[31] S. Papadopoulos, A. Drosou, N. Dimitriou, O. H. Abdelrahman, G. Gorbil, and D. Tzovaras, "A BRPCA based approach for anomaly detection in mobile networks," in *Information Sciences and Systems*. Cham, Switzerland: Springer, 2016, pp. 115–125.
[32] M. M. Hamed, H. R. Al-Masaeid, and Z. M. B. Said, "Short-term prediction of traffic volume in urban arterials," *J. Transp. Eng.*, vol. 121, no. 3, pp. 249–254, 1995.
[33] L. Zhang, Q. Liu, W. Yang, N. Wei, and D. Dong, "An improved K-nearest neighbor model for short-term traffic flow prediction," *Proc.-Social Behav. Sci.*, vol. 96, pp. 653–662, Nov. 2013.
[34] M. Bernaś, B. Płaczek, P. Porwik, and T. Pamuła, "Segmentation of vehicle detector data for improved k-nearest neighbours-based traffic flow prediction," *IET Intell. Transp. Syst.*, vol. 9, no. 3, pp. 264–274, Apr. 2015.
[35] W. Min and L. Wynter, "Real-time road traffic prediction with spatio-temporal correlations," *Transp. Res. C, Emerg. Technol.*, vol. 19, no. 4, pp. 606–616, 2011.
[36] Y. Kamarianakis and P. Prastacos, "Space–time modeling of traffic flow," *Comput. Geosci.*, vol. 31, no. 2, pp. 119–133, 2005.
[37] A. Salamanis, G. Margaritis, D. D. Kehagias, G. Matzoulas, and D. Tzovaras, "Identifying patterns under both normal and abnormal traffic conditions for short-term traffic prediction," *Transp. Res. Procedia*, vol. 22, pp. 665–674, 2017.

Ilias Kalamaras received the Diploma degree in electrical and computer engineering from the Aristotle University of Thessaloniki in 2010, and the Ph.D. degree in multimodal dimensionality reduction from the Intelligent Systems and Networks Group, Imperial College London, in 2016. He is currently a Research Assistant with the Information Technologies Institute of the Centre for Research and Technology Hellas. His main research interests include graph-based dimensionality reduction, clustering, and visualization of multimodal data.

Alexandros Zamichos received the Diploma degree in computer and communication engineering from the University of Thessaly in 2015. Since 2015, he has been a Research Assistant with the Information Technologies Institute of the Centre for Research and Technology Hellas. His main research interests include data mining and machine learning.

Athanasios Salamanis received the Diploma degree in electrical and computer engineering from the Polytechnic School of the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2013. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Polytechnic School, Democritus University of Thrace, Xanthi, Greece. Since 2013, he has been a Research Assistant with the Information Technologies Institute of the Centre for Research and Technology Hellas (CERTH/ITI). His main research interests include applied machine learning, high performance computing methods, big data analytics, traffic prediction in large-scale urban networks, pattern recognition, and time series analysis.

Anastasios Drosou received the Diploma degree in electrical and computer engineering from the Aristotle University of Thessaloniki in 2004, the M.Sc. degree in communication electronics from the Technische Universität München in 2007, and the Ph.D. degree in signal and image processing from Imperial College London in 2013. He is currently a Research Assistant with the Information Technologies Institute of the Centre for Research and Technology Hellas. His research interests are in the field of biometric security, computer vision, stereoscopic image processing and signal analysis, pattern recognition, network security, visualization, and visual analytics.

Dionysios D. Kehagias received the Diploma and Ph.D. degrees in electrical and computer engineering from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 1999 and 2006, respectively. He is currently a Researcher Grade C with the Information Technologies Institute of the Centre for Research and Technology Hellas (CERTH). His research interests include time-series analysis, big data analytics, machine learning, and intelligent transportation systems, with a focus on traffic prediction techniques under uncertainty.

Georgios Margaritis received the bachelor's, master's, and Ph.D. degrees in computer science from the University of Ioannina, Greece, in 2005, 2008, and 2014, respectively. He is currently a Research Assistant with the Information Technologies Institute of the Centre for Research and Technology Hellas (CERTH). His research interests include big data analytics, scalable datastores, search engines, and high performance computing.

Stavros Papadopoulos received the Diploma degree in electrical and computer engineering from the Aristotle University of Thessaloniki in 2010, and the Ph.D. degree in graph analytics from the Intelligent Systems and Networks Group, Imperial College London. He is currently a Research Assistant with the Information Technologies Institute of the Centre for Research and Technology Hellas. His main research interests include information visualization, visual analytics, network security, anomaly detection in mobile and IP networks, and root cause analysis.

Dimitrios Tzovaras (SM'13) received the Diploma and Ph.D. degrees in electrical and computer engineering from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 1992 and 1997, respectively. He is currently the Director with the Information Technologies Institute of the Centre for Research and Technology Hellas. His main research interests include visual analytics, 3-D object recognition, search and retrieval, behavioral biometrics, assistive technologies, information and knowledge management, multimodal interfaces, computer graphics, and virtual reality.