Real-time System for Short- and Long-Term Prediction of Vehicle Flow

Stefano Bilotta, Paolo Nesi, Irene Paoli
Department of Information Engineering, DISIT Lab
University of Florence, Florence, Italy
Https://www.disit.org, https://www.snap4city.org, <name>.<surname>@unifi.it

***Abstract*— Nowadays, traffic management and sustainable mobility are becoming one of the central topics for intelligent transportation systems (ITS). Thanks to the today's technologies, it is possible to collect real-time data to monitor the traffic situation in some specific areas. An important challenge in ITS is the ability to predict road traffic variables. The short-term predictions of traffic aspects are a complex nonlinear task that has been the subject of many research efforts in the past few decades. Accessing to precise traffic flow data is mandatory for a large number of applications which have to guarantee high level of services such as: traffic flow reconstruction, which in turn is used to perform w*hat-if analysis, conditioned routing, etc.* They have to be reliable and precise for sending rescue teams and fire brigades. This paper proposes a solution for a short- and long-term traffic flow prediction estimation by using and comparing a number of machine learning approaches. The solution has been developed in the context of Sii-Mobility smart city mobility and transport national project and it is in use in other EC projects and solution such as Snap4City PCP EC and TRAFAIR CEF, but also for REPLICATE H2020 SCC1 and control room in Florence area.**

Keywords- smart city, traffic sensors, short-term predictions, prediction models, machine learning, reconstruction algorithm, traffic flow.

I. INTRODUCTION

Traffic measuring is a central topic for intelligent transportation systems (ITS). Thanks to the today's technologies real-time data can be collected and used to monitor the traffic. The knowledge of real-time traffic information enables the development of a relevant number of services and improvements in many areas: congestion detection and reduction; computing of O-D origin-destination matrices; incident management; optimization of existing infrastructures of public transport; dynamic network traffic control; improved information services (e.g., traffic information, dynamic route guidance, road digital signage); plan for future investments on mobility solutions; reducing fuel consumption and emissions for CO2, NO2 that strongly depend on traffic. See on this fact, the normative of the European Commission regarding the conformant of the environmental value with respect to the reference values (2008/50/EC Directive on Ambient Air Quality and Cleaner Air for Europe and 2004/107/EC Directive on heavy metals and polycyclic aromatic hydrocarbons in ambient air).

Traditional methods for traffic flow measuring via spire sensors are very expensive for installation and maintenance and thus can be only arranged on a limited number of points in the city. Surrogated of traffic flow data can be obtained from data collected by mobile applications such as navigators, as well as from on board units, as performed by tracking systems of fleets and insurances. For example, in [15] a smartphone-based crowd sensing system for traffic regulator detection and measure has been proposed, where the data are gathered from the handheld devices located within the running vehicles. On the other hand, the data coming from navigator Apps (e.g., TomTom, Google map, Waze) could be very expensive. The usage of TV Cameras located in specific critical points and not at the crossing reduces the costs thus allowing the installation of a higher number of traffic sensors, and thus made possible their exploitation for the above-mentioned applications.

Traffic sensors provide continuous measuring of the traffic on selected roads at fine grain, and in most cases also providing information about the vehicle's kind. On traffic flow sensors, a variety of dysfunctions can be experienced which provoke the lack of data. Therefore, when data measures are missing, it is important to be capable to provide services to citizens in any case. There are relevant services among the above-mentioned that need to guarantee a significant service level and continuity thus providing a needed level of quality for the real-time services. Among them, (i) the *traffic flow reconstruction*, that allow to compute a traffic flow in each segment of the road network [1]; (ii) *conditioned routing* or the *what-if analysis* for rescue teams, fire brigade, etc. The use of stationary scattered sensors data the monitoring of traffic status can be combined with the short-term predictions of traffic flow to reduce discontinuities in the above mentioned services, thus accepting a certain level of error that could maintain the needed service level, and prevent the infringement of relevant constraints of the service level agreements.

Thus, on traffic flow sensors, different types of dysfunctions may prevent the reception of data: network failure, broken device, wrong data production, and byzantine errors. When the fault is temporary, an accurate short-term traffic prediction model could solve the problem producing the missed real-time data. In the case of long-term dysfunction, a long-term predictive model (e.g., one week) could be adopted to use predicted values in place of broken device ones. On the other hands, long-term traffic predictions are very challenging due to the dynamic nature of traffic flow data. And, an anomaly detection algorithm has to be adopted to alert the municipality about the device's failure to start fixing

the problem. The anomaly detection can be the technique to activate the predictions in place of the lack of data.

In literature, short-term and long terms traffic flow predictions have attracted extensive research efforts in the past decades and nowadays remained a growingly active research topic. In [10, 11, 12, 13, 14], the traffic state analysis is related to the monitored areas in terms of short-term traffic flow prediction on fixed points and no information is given where sensors are not located. In [10, 11], the theoretical basis for modelling univariate traffic condition data streams as seasonal autoregressive integrated moving average processes are considered. In [12, 14], the problem of short-term prediction has been assessed in freeways thorough deep learning models exploiting historical information only. In [19], the authors discussed the random forest model for the prediction of short-term traffic flow achieving an accuracy about 94%. In [20], neural networks, random forest, a gradient boosting machine, and a generalized linear model have been investigated in order to short-term predict traffic volume, speed, and occupancy of a single roadway segment. The authors have applied the model only on 1.3-mile section of westbound Interstate 64 (I-64) in St. Louis, Missouri, in the United States, obtaining an accuracy about 92%. Even in these cases the authors involved historical information only to predict the traffic conditions. In [24], three methods for a short terms traffic prediction of a single road have been compared (i.e., CNN, GRU, GRU+STFSA) exploiting historical information in a period of 40 working days. In [23], traffic volume is predicted on highway domain, using characteristics as: weak time continuity, structural space topology, and wider spatio-temporal correlation.

In the above presented cases, simple network areas as freeways or rings are considered and only for short-term prediction, taking into account only historical data. The present paper covers a complex urban network in a real-world road structure, and it investigates on the influence of historical traffic related features and on external information as actual weather and weather forecast features. In particular, the relevance of each variable has been evaluated to have a view of factors that are the most related with the traffic conditions in the city. In addition, a study of sensors through a clustering approach has been presented with the aim to have a topological point of view of the entire urban network reflecting the streets categorization.

This paper presents a solution for the computing both short and long-term traffic flow sensors predictions. The solution has been implemented in the context of Sii-Mobility project and infrastructure (national smart city project of Italian Ministry of Research for terrestrial mobility and transport, http://www.sii-mobility.org). Sii-Mobility is based on Km4City model and tools (https://www.km4city.org) [21]. Sii-Mobility is presently covering the whole Tuscany region, Italy, which hosts 3.5 inhabitants and 40M of tourists per year. The solution proposed in this paper is at the basis of Snap4City on traffic flow analysis and reconstruction, and Trafair CEF for computing NOX production from traffic, and it is presently exploited in the Smart City Control Room for Florence area according to REPLICATE H2020 SCC1 project and challenge. Moreover, for Florence, Pisa and Livorno municipalities in the Tuscany region the traffic flow data are used for traffic flow reconstruction and for other services [1].

The examples reported on the paper are related to the sensors traffic flow data in the Florence area, Italy, enabling the above-mentioned Smart City services. To this aim, a flexible model has been adopted to predict vehicle flow values one hour in advance with a resolution of 10 minutes. Therefore, the main contribution of this paper consists in presenting a machine learning approach for real-time short and long terms prediction of traffic flow sensor values.

The paper is structured as follows. Section II provides a description of the traffic flow data, and their characterization in terms of clustering in groups. In addition, the identification of a number of features at the basis of the predictive models are proposed. In Section 3, the machine learning approaches adopted to identify and validate the predictive models and framework are presented. The section also focusses on the comparison of the predictive models (short terms as 1 hour) exploiting the data collected within Florence area for traffic sensors, to achieve the identification of the best resulting approach in terms of prediction error and processing time. Section IV analysed the changes and the impact of them on the predictive model of Section III for long terms prediction (1 week). Section V presents one of the most critical application in which the exploitation of traffic flow data is very relevant for the production of traffic flow reconstructions. Conclusions are drawn in Section VI.

II. DATA DESCRIPTION AND FEATURE IDENTIFICATION

As mentioned in the introduction, the main goal was to find a solution to predict the traffic flow in the locations of traffic sensors. Typically, for each traffic sensor, the traffic flow is registered every 10 minutes. The data exploited refer to the 135 devices located in the municipality of Florence as depicted in Figure 1. Please note that, each device sensor location may measure the traffic flow on both sides of the road, and on multiple lanes. Therefore, in each location may corresponds to two distinct device logic sensors.

Figure 1. Map of the traffic sensors location in Florence municipality

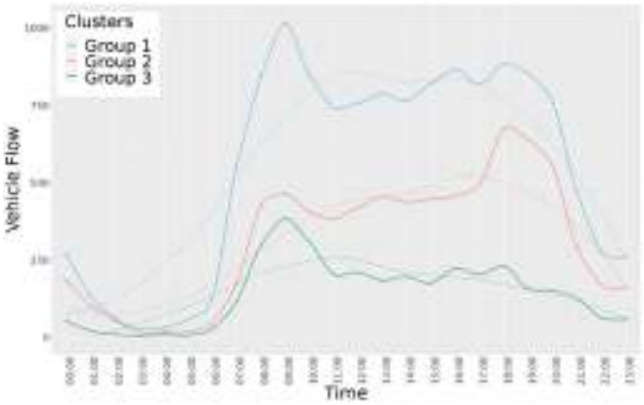
The trends of traffic flow data are strongly dependent on a number of road features: road relevance (primary, secondary, etc.), number lanes, speed limits, presence of speed meters,

distance from road crossing, etc. Moreover, a certain class of roads (e.g., the so called primary/main roads of the open street map), may provide higher capability with respect to local, single lane cases. In order to characterize the typical time trend H24 of the whole traffic flow sensors located in the city, we have performed a clustering. This approach allowed to aggregate device sensors with the same behaviour over time. The data taken into account have been those from November 2019, to February 2020.

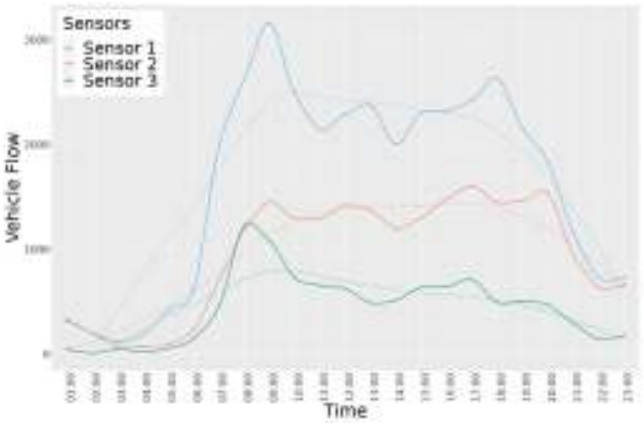
As a first step, we have tested cluster tendency by measuring the probability that a given data set has been generated by a uniform data distribution using the Hopkins statistics [8]. The value of Hopkins statistic resulted to be equal at 0.86, then the data set was proven to be significantly clusterable. As a second step, the K-means clustering method has been applied to identify clusters of traffic flow sensors. Please note that, K-means assigns each item to the cluster having the nearest centroid. In K-means clustering, there is an ideal center point that represents a cluster [3]. The clustering has been performed on the basis of the time trend H24, considering the normalized vehicle flow measures. The optimal number of clusters resulted to be equal 3, and it has been identified by using gap statistic criteria [4]. In Figure 2, the identified clusters have been represented on map, at which a different color pin for each cluster has been assigned.

Figure 2. Map of the traffic sensors location per cluster in Florence municipality (blue pins: Group 1; red pins: Group 1; green pins: Group 3)

Figure 3 (a) depicts the hourly median vehicle flow trends for each cluster and Figure 3 (b) shows the average vehicle flow trends of the three most representative traffic flow sensors for each cluster. The three trends are mainly describing situations in which: (1) a peak is registered in the morning and a second peak is also present in the evening and this cluster is characterized by a high flow of vehicles; (2) an almost stable traffic is present in whole day working hours, characterized by medium flows; (3) the peak of traffic is registered in the morning, from 7:00 to 9:00 while in the rest of the day is the flow of vehicle tends to decrease.



(a)



(b)

Figure 3. (a) Hourly median vehicle flow trends per cluster (Group 1, Group 2, Group 3) and (b) hourly average vehicle flow trends per representative sensor (Sensor 1, Sensor2, Sensor 3) in each cluster

A. Feature Identification

Since the aim of the research has been to identify a traffic flow prediction solution, a set of features have been identified and evaluated to have a more general view on those factors that are most correlated with the traffic condition in the city. Starting from a set of historical variables considered in literature, additional sets of traffic related features and weather information have been considered in the present paper. Thus, the identified large set of features has been classified and presented in Table 1.

Features belonging to the *Baseline* (time series) category refer to aspects related to the direct statistical observation of sensors data over time. Date and time when measures are taken, working day or not, number of vehicles in the street, etc. belong to this category. Typically, the values are recorded every 10 minutes and are used to consider the seasonality of data which may have different trends, e.g., working days with respect to weekends. Usually, the trend of number of vehicles is similar from one week to another for the same day (e.g., Monday to Monday); thus, two other features have been included in the model for capturing:

(i) the difference between the number of vehicles v captured at the same time i and the number of vehicles during the previous time slot of the previous week (dP):

$$dP = v_k - v_{k-1}$$

where $k = i - 7days$

(ii) the difference between the number of vehicles v captured at the same time i and the number of vehicles v in the successive time slot of the previous week (dS):

$$dS = v_k - v_{k+1}$$

where $k = i - 7days$.

The value of the number of vehicles related to the previous week respect to the observed one at the same time i has been considered as additional feature ($PwVF$):

$$PwVF = v_k$$

where $k = i - 7days$.

Category	Feature	Description
Baseline	Vehicle Flow (v)	Real number of vehicles recorded every 10 minutes
	Time	Hours and minutes
	Month	Month of the year (1-12)
	Day	Day of the month (1-31)
	Day of the week	Day of the week
	Weekend	0 for working days, 1 else
	Previous observation's difference of the previous week (dP)	$dP = v_k - v_{k-1}$ where $k = i - 7days$
	Subsequent observation's difference of the previous week (dS)	$dS = v_k - v_{k+1}$ where $k = i - 7days$
	Previous week observation ($PwVF$)	$PwVF = v_k$ where $k = i - 7days$
Weather and weather forecast	Max Temperature	City maximum expected temperature during the day
	Min Temperature	City minimum expected temperature during the day
	Temperature	City temperature one hour earlier than *Time*
	Humidity	City humidity one hour earlier than *Time*
	Rain	Presence of rain one hour earlier than *Time*
	Pressure	City pressure one hour earlier than *Time*
	Wind Speed	City wind speed one hour earlier than *Time*

Table 1. Overview of the feature used in the short-term prediction models

Features belonging to the *Weather* and *Weather forecast* are also collected every 10 minutes (i.e., temperature, humidity and rainfall). According to our analysis, the significant values are those related to the hour just before measured vehicle flow time.

III. SHORT-TERM PREDICTION MODELS

In this section the machine learning techniques considered are compared with the aim of creating a solution to predict the vehicle flow of each traffic sensor in the city. The possibility of producing both short and long terms prediction (see Section IV) with a satisfactory precision strongly reduces the errors and faults in the services where real time information are necessary to produce a result. In these cases, the new predicted values can be considered as alternative source of traffic information (see Section V for more details).

During our research study a number of techniques have been discharged since they did not produce satisfactory results (e.g., Bayesian Regularized Neural Network and Recurrent Neural Network, Support vector Regression that achieves an R-squared less than 0.7). Among the well-known considered techniques, the most effective solutions are the eXtreme Gradient Boosting (XGBoost) [5] and the Random Forest (RF). In addition, we are also reporting the results of the Auto Regressive Integrated Moving Average (ARIMA) model as an alternative forecasting to show the performance of classic statistical solutions. The model has been developed by using Box-Jenkings methodology for ARIMA modeling [6]. The choice of the presented models has been led from a study of the best solutions presented in literature.

A. Experimental Results

According to the data and considerations reported in previous sections, the identified challenge was to create a flexible model to predict the vehicle flow with a resolution of 10 minutes for the next hour (as short term prediction, while the long term prediction addressed 1 week in advance). As a training data set, we have selected a period of three months, from **November 2019, to February 2020**. The test set is made by observations every 10 minutes recorded during the weeks from January 27th (Monday) to February 9th (Sunday), i.e., 24 (hours) per 14 (days) test sets were considered to calculate the error on the one-hour prediction avoiding noise. In reality we have much longer time periods of data into Snap4City platform and service, while the limitation has been defined to be sure that the learning is addressing the recent seasonality behaviour with a learning phases that is not too computationally expensive. In fact, taking a year of data to make just a prediction depending on the data of the last 4 weeks would be completely un-useful.

Therefore, three different approaches are reported: ARIMA model, RF, XGBoost, and applied on the features presented in Table 1. The ARIMA model has been executed as multi-step forward with updated iteration technique: the forecast was computed one hour in advance. Then, the training set is updated with the observations recorded in the predicted hour and a new forecast is executed for the next hour. The RF has been set with number of trees composing the forest equal to 500 and the candidate feature set equal to 1/3 of the number of the data set variables. For XGBoost the eta value was set to 0.3. In Tables 2, 3 and 4, the assessment of the prediction

models is presented, providing the result in terms of R^2, Root-Mean-Square Error (RMSE) and Mean Absolute Scaled Error (MASE). Table 2 shows the Sensors of Group 1 models results, Table 3 shows the Sensors of Group 2 results and Table 4 reports those for Sensors of Group 3. Those sensors are representative of the clusters reported above.

ML Models	R^2	RMSE	MASE
RF	0.95	240	**1.30**
XGBoost	**0.96**	**234**	1.35
ARIMA	-	286	1.38

Table 2. Sensors of Group 1 prediction models result

ML Models	R^2	RMSE	MASE
RF	0.91	187	**0.99**
XGBoost	**0.93**	**170**	1.09
ARIMA	-	263	1.56

Table 3. Sensors of Group 2 prediction models result

ML Models	R^2	RMSE	MASE
RF	0.88	203	1.19
XGBoost	**0.90**	201	1.24
ARIMA	-	**120**	**0.90**

Table 4. Sensors of Group 3 prediction models result

Considering ARIMA models, for Group 1 Sensors the best predicting model has been an ARIMA (2,0,5), for Group 2 an ARIMA (4,0,4) and for Groups 3 an ARIMA (2,0,2) model (all of them by considering a predictive window of 1 hour). For Groups 1 and 2 the achieved results of ARIMA models are similar or worst with respect to machine learning approaches. For Group 3, the results are better respect to the ML method. On the contrary, the ARIMA approach, to achieve comparable results, has to be re-trained whenever the data is missing. These motivations have led to discarding the ARIMA model for short-term prediction. The XGBoost model turned out to be the better ranked in terms of R-squared and RMSE for most of the cases. In all representative sensors the R-square value is greater than 0.90 and reach the higher value in the prediction model for sensors of group 1 (0.96).
The RF model seems to be better in terms of MASE for the sensors of group 1 and sensors of group 2 models, even if the values are similar to those obtained with the RF model. The machine learning models allowed to predict the traffic flow (and thus the number of vehicles) 24 hours/few days in advance with the same accuracy reported in Tables 2, 3 and 4. In particular, the error measures don't deviate in a significative way when the predicted time interval takes into account reach a maximum of 48 hours.
In terms of processing time, the XGBoost model takes half of the time wrt RF model training. For this reason, XGBoost has been considered the best compromise and it has been adopted as the best solution even in long-term prediction case presented in the following section.

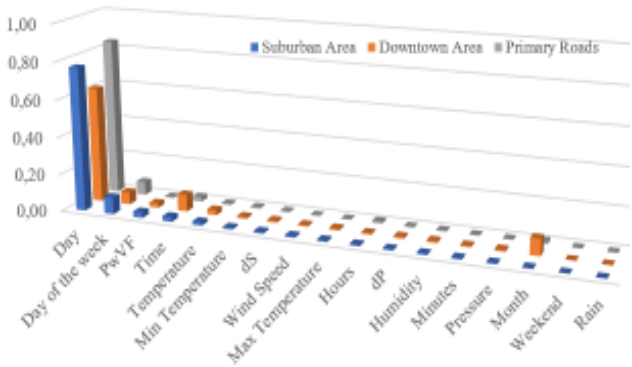
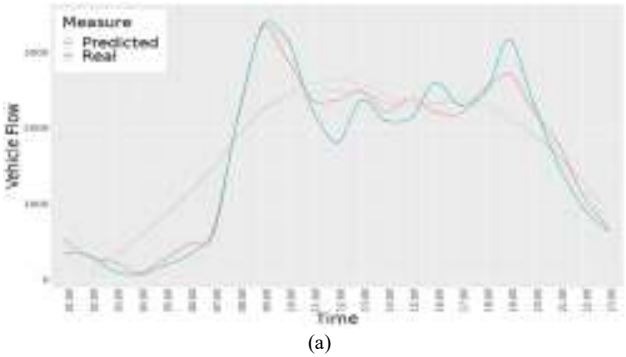
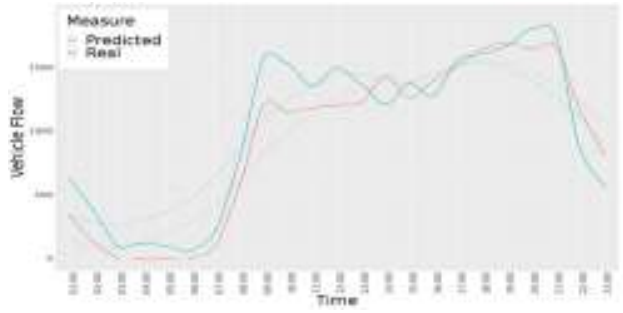


Figure 4. Variables Importance of the XGBoost model per cluster. The resulting histogram depicts that variable Day is the most relevant predictor followed by the Day of the week and the PwVF value. The weather conditions seem to be not so influent.

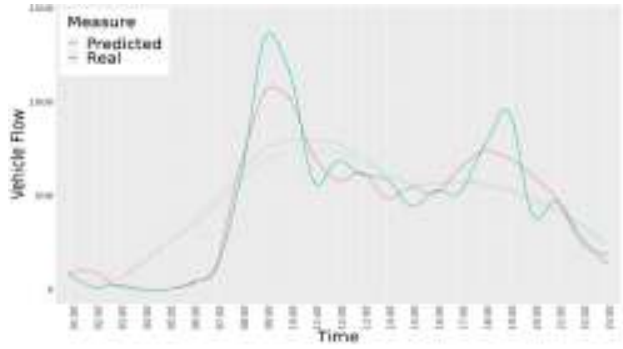
Figure 4 reports the analysis of relevance for the features presented in Table 1. The relevance of each predictor is evaluated individually: during the model training, a LOESS [18] smoother, (i.e., a nonparametric method for regression estimation) is fitted between the outcome and the predictor. To obtain a relative measure of variable importance, the R^2 statistic is computed for the model containing the considered variables against the null model (intercept only). The resulting histogram depicts that variable Day (of the baseline category) is the most relevant predictor. On the contrary, weather status and forecast seem to be not so relevant. For this reason, XGBoost model has also been tested by exploiting baseline feature only (see Table 5). In this condition, the running time of the training model turns out to be lower (from about 130 sec for the baseline + weather model to about 70 sec for the baseline model). Figure 5 shows the comparison between predicted and real values of vehicle flow for (a) Group Sensors 1, (b) Group 2 Sensors and (c) Group 3. In the example, February 17th (Monday) has been considered as a typical day.



(a)



(b)



(c)

Figure 5. Daily (Monday) vehicle flow real trend vs predicted trend for (a) Sensor 1 (b) Sensor 2 (c) Sensor 3

Note that, the possibility of a *medium*-term prediction (from 24 hours to few days in advance, using data from weather forecast) can be adopted to handle and predict the values in the case of medium-term dysfunction (e.g., a broken device).

IV. Long-term prediction models

As mentioned in the introduction, the exclusively use of fixed sensors can generate disadvantages due to device dysfunctions. In the case of long-term devices dysfunctions, long-term road traffic variables prediction (one week in advance, from Monday to Sunday) is of considerable significance. The possibility of producing long term predictions with a satisfactory precision allow the production of services that give information about the traffic condition in the future or predicting and analysing the production of pollutants such as NOx coming from traffic conditions [22].
To guarantee traffic information is useful understand and predict the traffic flow measures at most one week in advance. To this aim, the models presented in Section III, has been also adopted in a long-term perspective and results compared with the average trend for the sensor.

A. Experimental Results

The possibility of a long-term prediction (one week in advance) can be adopted to predict the values in the case of long-term dysfunction (e.g., a broken device). The XGBoost model has been applied exploiting baseline category features only presented in Table 1. Results have been reported in Table 5 and the predicted values are compared with the average for each hour slot.

XGBoost Model Results	R^2	RMSE	MASE
Sensors of Group 1	0.95	215	0.89
Sensors of Group 2	0.91	178	0.82
Sensors of Group 3	0.86	127	0.92

Table 5. XGBoost long-term prediction model results for each representative sensor group

Please note that, Table 5 results regarding long terms prediction (1 week) can be compared with respect to those obtained for short-term predictions (1 hour). So that, the long terms resulted the most accurate in terms error measures. Such results are improved by using the baseline features only. This is a double benefit: from a computational cost point of view and for the accuracy aspects. Moreover, the same training can be used for the short-medium-long term predictions achieving similar results in all cases.
On the other hand, the usage of the mean value estimated for each time slot in the week, that is the typical rule of thumb produce worst results in terms of error accuracy with respect the results presented in Table 5 (RMSE errors using average values are twice as many as errors using XGBoost model). Please note that the solution presented also outperforms the solutions proposed in the literature as mentioned in the introduction. This is due to the methods and to the fact that our approach considers a larger set of features including: metrics, and external factors such as weather condition. Moreover, the solution is developed in a complex urban network where the roads have different characteristics (e.g., primary roads, traffic limited roads etc.) and the results in terms of accuracy are satisfactory as the studies conducted in a single highway domain.

V. Predictive Model application

The aim of the present section is to provide a general idea about one the most relevant applications of the prediction model results within the city traffic flow reconstruction algorithm presented in [1]. In some occasion, sensor data may come with a variety of missing values, resulting in considerable difficulties in the analysis and maintain the service. To guarantee the service, the prediction model has been included in the traffic flow reconstruction algorithm as method to impute the missing data in real time, and thus compensate the sensors' network fault that happen in an average of 10% of time. This means that in the 10% of days some of the sensors are not producing the data thus reducing the amount of data in input at the 85%. Please note that the traffic flow reconstruction can be still feasible with higher errors until the traffic flow sensors network is working with at least the 70% of sensors, resulting in an alarm and creating a ticket for the corrective maintenance teams: https://www.snap4city.org/dashboardSmartCity/view/index.php?iddasboard=MTc2MQ==. (see Figure 6).

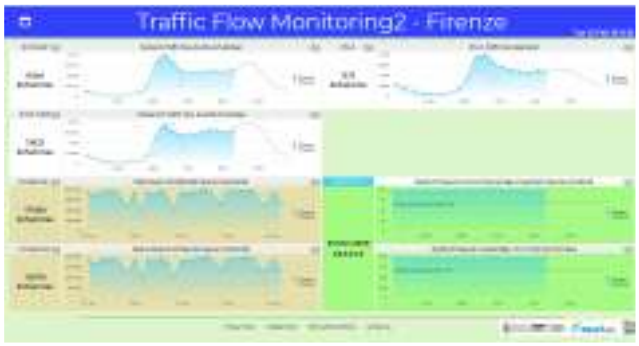


Figure 6. Traffic flow data at the border of the city. Please note the bottom right corner widgets in which the percentage of active traffic flow sensors is reported by hours.

A. City Traffic Flow Model Computational Approach

In [1], the authors consider the traffic data coming from the results produce by Sii-Mobility project where a general study of the traffic flow reconstruction model is applied in the city of Florence. More precisely, a real-time visual self-adaptive solution for traffic flow reconstruction at every location within a wide area (in terms of number of road segments) is produced, leveraging the detections from a few fixed traffic sensors deployed within the area of interest. The solution has several advantages with respect to the solutions available in the literature, since it: supports complex and real-world road structure; presents a wider applicability; does not needs of third-party engagement on providing data from installed devices on vehicles; is robust with respect to discontinuous data; declared precision rate; produces real-time visual rendering of results. A such approach is based on modelling the traffic flow in the city by means of Partial Differential Equation (PDE) based on fluid dynamics studies.

In details, the city traffic flow model proposed in [1] is a real-time visual self-adaptive solution to reconstruct the traffic density at every location of a wide urban area from a few fixed traffic sensors deployed within the area of interest. A mathematical model for fluid dynamics on networks has been applied, and the road networks have been studied as direct graph composed by arcs that meet some nodes, corresponding to road junctions. In a single road the nonlinear model is based on the **conservation of cars** described by the following scalar hyperbolic conservation law:

$$\frac{\partial \rho(t,x)}{\partial t} + \frac{\partial f(\rho(t,x))}{\partial x} = 0 \quad (1)$$

with: boundary conditions $\rho(t,a) = \rho_a(t)$, $\rho(t,b) = \rho_b(t)$ and initial values $\rho(0,x) = \rho_0(x)$. In particular, $\rho(t,x)$ denotes the vehicular density which admits values from 0 to ρ_{max}, where $\rho_{max} > 0$ is the maximal vehicular density on the road. The function $f(\rho(t,x))$ is the vehicular flux which is defined by means the product $\rho(t,x)v(t,x)$, where $v(t,x)$ is the local speed of the vehicles. In the case of first order approximation, if we assume that $v(t,x)$ is a decreasing function, only depending on the density, then the corresponding flux is a concave function. The discretization scheme in terms of *finite difference* is considered to obtain a numerical solution of the equation (1). Please note that the equation (1) and its solution are based on the hypothesis that there is a conservation of the number of cars entering and exiting in the area of observation. This means that each faults of sensors on the border of the city area causes a degeneration of the main working condition of the solution. On the contrary the usage of the predictive model allows to reduce: (i) the error estimation due to the lack of conservation, and (ii) the discontinuities provoked by the lack of data in specific points.

Figure 7 shows the traffic flow reconstruction algorithm [1] visualization in Florence municipality. Data used in the algorithm are a combination between real data and short-term prediction model results.

The prediction data can also be used to estimate the traffic reconstruction in the future considering all traffic sensors as missing. In all these cases, the predicted values can be considered as an additional source of traffic. During the real-time evaluation of the traffic flow reconstruction the predicted values can be used to improve the entire model accuracy. More precisely, to highlight the importance of the prediction model inside the traffic flow reconstruction algorithm, a comparative approach has been conducted when a given sensor data is not available, with the aim to compute an error measure in terms of mean absolute percentage error (MAPE). A simulation analysis has been conducted to calculate the average error in reconstruction assuming that the value of a given sensor is missing during a 24H time slot. The MAPE on reconstruction without considering the prediction decreases from 0.25 to 0.14.

Figure 7. Traffic flow reconstruction algorithm visualization

https://www.snap4city.org/dashboardSmartCity/view/index.php?iddasboard=MTc5NQ==

VI. CONCLUSIONS

In this paper, we have proposed a predictive approach and solution for short and long terms predictions of traffic flow data. The solution has been developed in the context of Sii-Mobility/Km4City smart city mobility and transport national project and it is in use in other EC projects and solutions such as Snap4City PCP EC and TRAFAIR CEF, but also for

REPLICATE H2020 SCC1 and control room in Florence area. The knowledge of real-time traffic information enables the development of a relevant number of services and improvements in many areas and services: congestion detection and reduction; computing of O-D matrices (commuter plans); incident management; optimization of existing infrastructures of public transport that is the improvement of efficiency of the current road network; dynamic network traffic control; improved information services (e.g. traffic information, dynamic route guidance, road message signs); plan for future investments; reducing fuel consumption and emissions for CO2, NO2, that strongly depend on traffic.

Among the applications, traffic flow reconstruction is probably the most critical for the precision needed and the reliability of the data in input to the process that are actually the traffic flow sensors data. Accessing to precise traffic flow data is mandatory to guarantee high level of services such as: traffic flow reconstruction, which in turn is used to perform what-if analysis, conditioned routing, etc. They have to be reliable and precise for sending rescue teams and fire brigades. This paper proposes a solution an approach for a short- and long-term traffic flow sensor value prediction by using XGBoost model that resulted to be the best compromise from precision and computational costs for both short and long terms prediction. The solution proposed outperformed the state-of-the-art solution also based on machine learning presented in the literature and the trivial rule that suggests the use of mean value.

ACKNOWLEDGMENT

The authors would like to thank the MIUR, the University of Florence and companies involved for co-founding Sii-Mobility national project on smart city mobility and transport. Km4City is an open technology and research of DISIT Lab. Sii-Mobility is grounded and has contributed to Km4City open solution. The present solution is also adopted in Snap4City platform and by Trafair CEF project.

REFERENCES

[1] Bellini, P., Bilotta, S., Paolucci, M., & Soderi, M. (2018, August). Real-Time Traffic Estimation of Unmonitored Roads. In 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech) (pp. 935-942). IEEE.J.

[2] A. Banerjee, R.N. Dave, Validating clusters using the Hopkins statistic, in: 2004 IEEE International Conference on Fuzzy Systems (IEEE Cat. No.04CH37542), vol. 1, 2004, pp. 149–153.

[3] Xu R, Wunsch DC. Clustering. New Jersey, USA, Wiley-IEEE Press, 2008.

[4] Tibshirani, Robert, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63.2 (2001): 411-423.

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

[6] G. E. Box, G. M. Jenkins, and G. C. Reinsel, Forecasting, Time Ser. Anal., vol. 12, pp. 137_191, Jun. 2008.

[7] S. Yang, S. Shi, X. Hu and M. Wang, Spatiotemporal Context Awareness for Urban Traffic Modelling and Prediction: Sparse Representation Based Variable Selection. PloS one10.10: e0141223, 2015.

[8] T. Q. Tang, W. S. Shi, X. B. Yang, Y. P. Wang and G. Q. Lu, A macro traffic flow model accounting for road capacity and reliability analysis. Physica A: Statistical Mechanics and its Applications 392, 6300 – 6306, 2013.

[9] Chapman, Craig H., and Oliver B. Downs. Assessing road traffic flow conditions using data obtained from mobile data sources. U.S. Patent No. 7,831,380. 9 Nov. 2010.

[10] Williams, Billy M., and Lester A. Hoel. Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results. Journal of transportation engineering 129.6, 664-672, 2003.

[11] Kamarianakis, Yiannis, and Poulicos Prastacos. Space–time modeling of traffic flow. Computers & Geosciences 31.2, 119-133, 2005.

[12] Zheng, Weizhong, Der-Horng Lee, and Qixin Shi. Short-term freeway traffic flow prediction: Bayesian combined neural network approach. Journal of transportation engineering 132.2, 114-121, 2006.

[13] Huang, Wenhao, et al. Deep architecture for traffic flow prediction: deep belief networks with multitask learning. IEEE Transactions on Intelligent Transportation Systems 15.5, 2191-2201, 2014.

[14] Lv, Yisheng, et al. Traffic flow prediction with big data: a deep learning approach. IEEE Transactions on Intelligent Transportation Systems 16.2, 865-873, 2015.

[15] Hu, Shaohan, et al. Smartroad: smartphone-based crowd sensing for traffic regulator detection and identification. ACM Transactions on Sensor Networks (TOSN) 11.4, 55, 2015.

[16] Ma, Xiaolei, et al. Large-scale transportation network congestion evolution prediction using deep learning theory. PloS one 10.3, e0119044, 2015.

[17] R. J. Hyndman and A. B. Koehler, Another look at measures of forecast accuracy, Int. J. Forecasting, vol. 22, no. 4, pp. 679_688, 2006.

[18] W. S. Cleveland, S. J. Devlin, and E. Grosse, Regression by local fitting: Methods, properties, and computational algorithms, J. Econometrics, vol. 37, no. 1, pp. 87_114, 1988.

[19] Zheng, C. H. E. N. G., & Xian-fu, C. H. E. N. (2016). The model of short-term traffic flow prediction based on the random forest. Microcomputers and Its Applications, 35(10), 46-49.

[20] Mohammed, O., & Kianfar, J. (2018, September). A Machine Learning Approach to Short-Term Traffic Flow Prediction: A Case Study of Interstate 64 in Missouri. In 2018 IEEE International Smart Cities Conference (ISC2) (pp. 1-7). IEEE.

[21] C. Badii, P. Bellini, D. Cenni, G. Martelli, P. Nesi, M. Paolucci, "Km4City Smart City API: an integrated support for mobility services", 2nd IEEE International Conference on Smart Computing (SMARTCOMP 2016), St. Louis, Missouri, USA, 18-20 May 2016.

[22] L. Po, et al. TRAFAIR: Understanding Traffic Flow to Improve Air Quality. IEEE International Smart Cities Conference, 730-737, 2019.

[23] Ding, Weilong, Xuefei Wang, and Zhuofeng Zhao. "CO-STAR: A collaborative prediction service for short-term trends on continuous spatio-temporal data." Future Generation Computer Systems 102 (2020): 481-493.

[24] Dai, Guowen, Changxi Ma, and Xuecai Xu. "Short-Term Traffic Flow Prediction Method for Urban Road Sections Based on Space–Time Analysis and GRU." IEEE Access 7 (2019): 143025-143035.