Transportation Research Part C 73 (2016) 183–201

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Short-term speed predictions exploiting big data on large urban road networks

Gaetano Fusco [*], Chiara Colombaroni, Natalia Isaenko

Department of Civil, Constructional and Environmental Engineering, Sapienza University of Rome, Via Eudossiana, 18, I-00184 Rome, Italy

ARTICLE INFO

Article history:
Received 25 March 2016
Received in revised form 25 October 2016
Accepted 27 October 2016
Available online 7 November 2016

Keywords:
Short-term traffic predictions
Big data
Floating car data
Measure accuracy
Bayesian networks
Neural networks
SARMA models
Supervised learning

ABSTRACT

Big data from floating cars supply a frequent, ubiquitous sampling of traffic conditions on the road network and provide great opportunities for enhanced short-term traffic predictions based on real-time information on the whole network. Two network-based machine learning models, a Bayesian network and a neural network, are formulated with a double star framework that reflects time and space correlation among traffic variables and because of its modular structure is suitable for an automatic implementation on large road networks. Among different mono-dimensional time-series models, a seasonal autoregressive moving average model (SARMA) is selected for comparison. The time-series model is also used in a hybrid modeling framework to provide the Bayesian network with an a priori estimation of the predicted speed, which is then corrected exploiting the information collected on other links. A large floating car data set on a sub-area of the road network of Rome is used for validation. To account for the variable accuracy of the speed estimated from floating car data, a new error indicator is introduced that relates accuracy of prediction to accuracy of measure. Validation results highlighted that the spatial architecture of the Bayesian network is advantageous in standard conditions, where a priori knowledge is more significant, while mono-dimensional time series revealed to be more valuable in the few cases of non-recurrent congestion conditions observed in the data set. The results obtained suggested introducing a supervisor framework that selects the most suitable prediction depending on the detected traffic regimes.

1. Introduction

1.1. Motivation

Fast and accurate predictions of future traffic conditions are a crucial requirement for reliable applications of Intelligent Transportation Systems (ITS) devoted to traffic management and traveler information, whose intelligence is related to their capability to foresee future states of the system and individuate the most appropriate actions to undertake. Advances in Information and Communication Technologies (ICT) are currently making available an unprecedented amount of measures of traffic variables from the road network that are a premise for introducing new models and methods for traffic predictions (Shi and Abdel-Aty, 2015). Traditional traffic monitoring systems are based on fixed measure stations where flows, occupancy and possibly speed are detected. Collected data are then transmitted to the traffic control center, where they are

[*] Corresponding author.
E-mail addresses: gaetano.fusco@uniroma1.it (G. Fusco), chiara.colombaroni@uniroma1.it (C. Colombaroni), natalia.isaenko@uniroma1.it (N. Isaenko).

http://dx.doi.org/10.1016/j.trc.2016.10.019
0968-090X/© 2016 Elsevier Ltd. All rights reserved.

processed to derive short-term predictions. The relatively high cost of investment and maintenance of fixed monitoring system was one of the most relevant limiting factors for a full ITS deployment although efficient algorithms for optimizing sensor locations were developed (Cipriani et al., 2006).

Availability of Floating Car Data (FCD) obtained by tracking GPS-enabled vehicles and mobile devices opens new perspectives to develop novel predicting models. In fact, they provide a pervasive tool to explore the road network and get information related to theoretically any point of the network (Fusco et al., 2015) and, in a near future, perform self-organizing monitoring techniques (Baiocchi et al., 2015). The existence of very detailed road graphs developed for on-board navigators would require equally detailed estimations of present and future traffic conditions. However, a suitable trade-off between reliability and accuracy of traffic estimates and predictions should be investigated. The main drawback of FCD is that the information is collected from only a sample of vehicles that send their current positions and speeds. Thus, they provide ubiquitous but partial information. This requires a supplementary effort to process these data and combine measures collected at different points and different instants. Moreover, while the sampling rule is usually specified, the actual sampling rate on each road link is unknown, so that the reliability of the measures is variable and difficult to estimate, except for the few links equipped with fixed traffic counting stations. Furthermore, in links not traveled by equipped vehicles data are missed at all. In the last years, several private companies have started collecting and selling real-time speed data from different sources, including floating car data. Aggregate measures supplied by private providers are usually paired with some qualitative confidence value and so preclude performing a rigorous estimation of the statistical significance of the data. Although the accuracy appeared to be improved since the earliest independent evaluation (Kim and Coifman, 2014), the reliability of traffic measures is still a crucial issue for studies dealing with short-term prediction methods that use floating car data. The huge amount of data collected in real-time on the road network requires also efficient analysis methods to catch the most useful information embedded in such time–space big data.

A large interest for machine learning methods arose in the last years in the literature on big data analysis and many network-based approaches, such as neural networks and Bayesian networks, were proposed with the aim of exploiting existing correlations among measures collected at different time intervals and on different links of the network. Specifically, Bayesian networks, which combine graph structure and Bayes approach to posterior probability from a priori estimate seem to offer a sound methodology for formulating short-term predictions from the pervasive sampling of traffic performances provided by floating car data.

In this paper, we aim at investigating the potentials of these methods to produce accurate short-term traffic predictions by exploiting floating car data collected ubiquitously on the network from a number of probe vehicles that is indeed large in absolute but is a relatively small fraction of the traffic flow on each link of the road network.

1.2. Approaches to short-term traffic predictions

Two main approaches can be individuated to perform short-term traffic predictions: either explicit or implicit traffic modeling. Explicit approach is based on mathematical models that represent the interactions between the physical variables that describe traffic phenomena. Traffic on freeways is usually modeled by macroscopic continuous models that discretize in time and space the partial differential equations that describe traffic dynamics. Traffic on urban road networks needs dynamic traffic assignment models that simulate the complex dynamic interactions between drivers' trip choices, vehicular congestion and road performances on the traffic networks.

Application of traffic models for real-time short-term predictions requires recursive methods implementable online. The rolling horizon method exploits current traffic measures to update trip demand estimation at every given short time interval and runs a new traffic simulation, which covers a longer time interval and holds until a new update is available. Relevant examples are the Dynasmart-X (Mahmassani et al., 2005) and Dynamit (Ben-Akiva et al., 2012). State-space models formulate the dynamic evolution of all traffic variables on the road network based on available real-time traffic measurements under a probabilistic environment (Muñoz et al., 2003). Typical applications for short-term real-time predictions imply the linear approximation of non-linear macroscopic traffic models that leads to the extended Kalman filter formulation (Stathopoulos and Karlaftis, 2003; Wang and Papageorgiou, 2005), although other approximation methods such as particle filter (Mihaylova et al., 2007) and Newtonian relaxation (Herrera and Bayen, 2010) were developed. The switching-mode model, which can be thought of as a combination of the hidden Markov model and the linear state-space model (Sun et al., 2003), was introduced to reproduce the possible transitions from a discrete traffic state to another, namely free-flow and congestion states that characterize the cell transmission model (Daganzo, 1994). A more complex architecture implements artificial neural networks to derive density values and determine transitions between traffic states on the linearized triangular fundamental diagram (Celikoglu, 2014).

Implicit approach derives dynamic relationships directly from time series of observed data and therefore is usually called data-driven approach. Although we acknowledge that explicit models have superior interpretation capabilities with respect to implicit models and can be applied to generate control and information strategies that prevent system over-reaction (Ben-Akiva, 1985), we recognize also that they require a huge effort to achieve an adequately accurate calibration of a large urban network. On the other hand, the enormous amount of available data on urban mobility makes implicit models a valuable alternative, easier to implement and open to possible integrations with explicit models within a hybrid rolling horizon framework that applies an explicit model to forecast traffic states over a time horizon of a few hours and an implicit model that adjusts prior model forecasts on the basis of real-time measures and supplies posterior short-term predictions. Thus, in this paper we focus on studying suitable structures of data-driven models to exploit time-space information embedded in floating car big data and testing the accuracy of the short-term predictions so obtained.

1.3. Related work

Data driven methods for short term traffic forecasting are object of a huge literature, which has been the object of a recent special issue on this journal (Zhang, 2014). We refer to the papers by Vlahogianni et al. (2014) and Oh et al. (2015) for a complete review of the state-of-the art and we focus here on the following issues: (i) the relevance of capturing the time-space correlation for short-term traffic forecasting in urban road networks through implicit models; (ii) the opportunities and concerns that arise from variable point traffic measures collected by sparsely sampled vehicles on the whole road network; and (iii) the generalization capability of probabilistic graphical models with respect to different congestion patterns.

Although the majority of previous studies conducted independent forecasting for each single monitored section of the road (Cai et al., 2016), several attempts were made in the past to catch spatial correlation between traffic variables on the road network by extending time-series models to multivariate form (Kamarianakis and Prastacos, 2005; Chandra and Al-Deek, 2009; Guo et al., 2014; Mai et al., 2015; Li et al., 2015a), through implicit prediction models that include a network structure, such as artificial neural networks (Fusco and Gori, 1996; Dougherty and Cobbett, 1997; Zhang, 2000; Zhu et al., 2014; Ma et al., 2015a, 2015b), Bayesian networks (Sun et al., 2006; Castillo et al., 2008; Hofleitner et al., 2012; Chen et al., 2015), deep architecture models (Lv et al., 2015). Several authors devised hybrid methods that combine different techniques and use multiple predictors (among others: Zhang, 2003; Zheng et al., 2006; van Hinsbergen et al., 2009; Wang et al., 2014). Chen et al. (2012) performed a systematic comparison of different methods for the short-term prediction on a single loop sensor and found that Bayesian networks and artificial neural networks be effective and efficient prediction models, although traffic breakdowns can be identified but cannot be accurately predicted. Other authors focused on the spatial-temporal correlation among traffic measures to face the complementary problem of estimating missed data, and applied either a tensor-based method (Tan et al., 2013) or a kernel probabilistic principle component analysis (Li et al., 2013). Recently, Lv et al. (2015) pointed out that traffic prediction models are still unsatisfying for many real-world applications and rethought the traffic flow prediction problem based on with big traffic data.

With reference to the second issue mentioned above, an increased interest in the opportunity of using FCD for traffic predictions arose in last years. First studies were based on data collected by special fleets like taxis (*Cfr*. Castro-Neto et al. (2009) for a review) or vehicles equipped with GPS specifically for the traffic experiment (Herrera et al., 2010; Bucknell and Herrera, 2014). Other studies on short-term traffic predictions from FCD used synthetic data to estimate the suitable penetration rate of vehicles to get accurate predictions (Deng et al., 2013). Feng et al. (2014) analyzed vehicle trajectories tracked in NGSIM experiment and developed a Bayesian method to estimate the probability distribution of travel times among different vehicles by taking into account synthetic GPS data and signal setting parameters to identify prevailing actual traffic conditions in real-time. Ye et al. (2012) studied a method to accommodate data recorded at irregular intervals, which exploits information from adjacent links. Among the studies based on-the-field data, Kim and Coifman (2014) analyzed aggregated information provided by INRIX company against loop detector measurements on 44 links and highlighted that they do not appear to reflect the latency with respect to reference measures or the occurrence of repeated reported speeds. Schneider et al. (2010) compared the effectiveness and accuracy of floating car studies with that achievable by Bluetooth technology. Patire et al. (2015) discussed the opportunities and challenges related to the use of non-aggregated point-speed GPS data and developed a data fusion method to exploit raw probe data in addition to fixed sensor counts.

As far as the generalization capability of prediction models to provide accurate predictions under different congestion patterns, almost all studies, with the exception of Guo et al. (2014), applied the short-term prediction models to a selected set of data covering a suitable time interval and assess their performances on the whole period, without inspecting the reliability of predictions in the case of heavy congestion. Many authors looked at the problem from a different perspective and tried to improve the traffic prediction by adapting the model framework to different traffic states. Two main approaches can be individuated: a clustering approach, which classifies traffic states either on the basis of the observed time-series pattern (Cai et al., 2016) or over the fundamental diagram (Celikoglu and Silgu, 2016; Antoniou et al., 2013), and a regime switching approach, again based on either time-series pattern (Cetin and Comert, 2006; Kamarianakis et al., 2012) or on the fit to the fundamental diagram (Dunne and Ghosh, 2012). Charle et al. (2010) addressed a rather different problem, which was route travel time reliability, and analyzed the historical space correlations between travel times of close links. Their perspective highlights the significance of long-term effects to individuate recurrent congestion conditions, which the short-term variation superimposes to. A reliable historical estimate is significant especially when dealing with FCD, whose sampling rate in real-time is often low other than unknown, so reducing the reliability of predictions founded on short-term series only. So far, few studies were based on large real data sets of FCD, as it would be necessary to face the question concerning the reliability of traffic forecasting methods based on FCD with respect to the reliability of the measures. Hofleitner et al. (2012) used individual FCD collected by 500 cars in a specific experiment; Cai et al. (2016) used a data set of space mean speed data collected on 30 road segments for 20 weekdays. Data were suitably preprocessed to fill missed data and eliminate abnormal values and filtered to get smoothed data. In a very recent paper (Fusco et al., 2016), we compared different network-based short-term forecasting models on a 10-month long series of aggregated measures obtained from FCD and we proposed a model structure conceived to perform forecasts on large networks exploiting speed estimates on all the links where they are available.

1.4. Paper contribution

The paper aims at providing a consistent method for short-term speed predictions on large networks based on raw floating car data and presents a modeling framework that implements some well-known network-structured prediction models. The paper also focuses on the issues that the analysis of the literature revealed to be worthy of further examination: the reliability of traffic measures collected at random points of the road network; the suitability of different prediction models with respect to different traffic conditions, such as free-flow, recurrent and non-recurrent congestion. The approach that we aim at following is that the nature of traffic congestion implicates that the computational methodologies of artificial intelligence must be transportation-inspired.

We introduce different architectures of machine learning models based on different levels of exploration of the road network in order to catch possible spatial correlations among traffic measures taken on different links of the network. In contrast to our previous study (Fusco et al., 2016), where we used the historical average speed as an a priori estimation, we are here closer to the Bayesian approach and we try to provide an as good as possible a priori estimation based on previous observations. Thus, we formulate a hybrid modeling framework where we integrate the best a priori estimation based on time correlation, which is provided by a consolidated Seasonal ARIMA model, with the spatial correlation estimated through a Bayesian network. Unlike our previous paper as well as other works in the literature, with the exception of Patire et al. (2015), we deal with issues and advantages of using raw data of individual cars. While Patire et al. focus on the question of sampling and penetration rates and present a data fusion framework to integrate floating car data and fixed point measurements, we introduce here a consistent method designed to use disaggregated raw data. We specify the model variables to exploit all the available information about traffic estimation. Specifically, variances between individual speeds and the number of measures in each time interval on each link are considered to account for the time-variable accuracy of the measures. However, no flow measure is assumed because the number of counts available is very often insufficient to get an accurate estimation in a reasonable time interval. We also enhance the validation method by introducing specific error indicators that relate the accuracy of different prediction models with the accuracy of the measures.

Unlike most studies in the literature, we assess the performances of the models under different traffic congestion conditions. Fig. 1 provides a flow-chart of the problems arising from sparse floating car data, the specific procedures implemented to face with each of them and the corresponding solutions that compose our method. It highlights also the main advantages that this method offers with respect to the state-of-the-art: the variable selection focuses on a fundamental issue of sparse floating car data, that is their variable sampling rate, and allows considering the accuracy of observed data in both model structure and prediction results; the double-star network structure of the forecasting models allows an easy modular implementation of the procedure even on very large networks as well as a parallel computation that preserves anyway the possible spatial correlation among the links; hybrid model formulation with a priori autoregressive predictions allows an easy extension of the model to integrate a supervisor mechanism that selects the best forecasting model based on estimated traffic conditions.

In contrast with other papers in the literature that aim at adapting the model framework to different traffic states, such supervisor exploits only individual point speed observations, so it does not require flow measure. Moreover, it does not seek to estimate traffic states but to individuate the occurrence of anomalous conditions and then it relaxes relationships based on recurrent observations. Finally, while only limited tests have been presented until now in the literature on traffic predictions using GPS-equipped Floating Car Data, we present a large numerical experiment conducted on a big data set composed of about 300,000 single point-speed data collected on a wide portion of an urban street network (120 links) selected on a sub-network of a large town, Rome.

The rest of the paper is organized as follows. Section 2 presents the methodology proposed for short-term forecasting, describes the state-of-art methods selected for reference and introduces the error indicators chosen for the comparison between different methods. Section 3 illustrates the experimental application on a suitable subarea of the road network

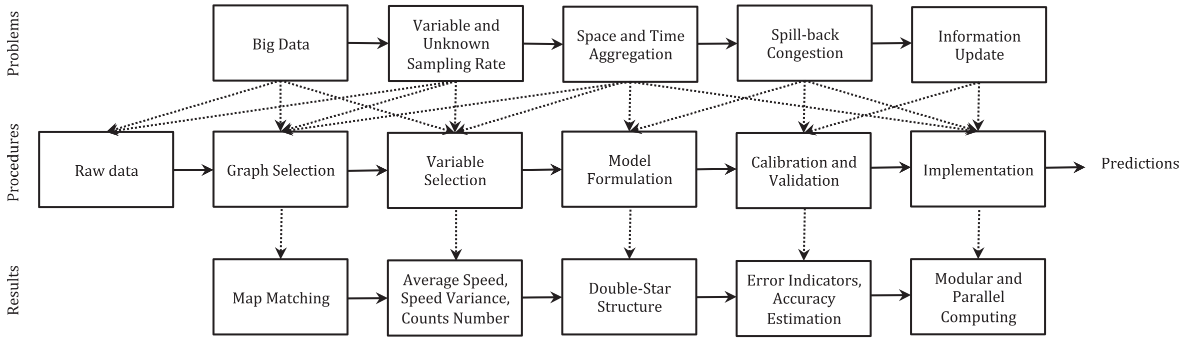
Problems
Big Data
Variable and Unknown Sampling Rate
Space and Time Aggregation
Spill-back Congestion
Information Update
Procedures
Raw data
Graph Selection
Variable Selection
Model Formulation
Calibration and Validation
Implementation
Predictions
Results
Map Matching
Average Speed, Speed Variance, Counts Number
Double-Star Structure
Error Indicators, Accuracy Estimation
Modular and Parallel Computing

Fig. 1. Flow chart of the proposed procedure.

of Rome, where the data set was available. Results of different prediction methods under different traffic conditions are illustrated and commented. Conclusions and suggestions for further research are reported in Section 4.

2. Prediction models

2.1. Time series analysis

Autoregressive Integrated Moving Average (ARIMA) is one of the most consolidated methods for time-series forecasting, used in various fields and introduced in traffic forecast on freeways since the late '70s by Ahmed and Cook (1979). In the case of stationary time series, the forecast provided by the Autoregressive Moving Average (ARMA) model is a linear combination of past observations multiplied by coefficients reflecting autoregressive (AR) and moving average (MA) nature of the process. In case the time series displays a trend the data must be differenced and an integration term (I) is usually introduced for making the time series stationary. Whether the time series presents seasonality additional terms are introduced, like in Seasonal ARIMA (SARIMA) models, applied in traffic engineering for traffic state prediction, among others, by Smith et al. (2002) and Williams and Hoel (2003). In its general form, the SARIMA model for the speed variable v_t on a generic road link is formulated as follows.

$$\Phi(B^s)\phi(B)(1-B^s)^D(1-B)^d v_t = \Theta(B^s)\theta(B)a_t \tag{1}$$

$$\Phi(B^s) = 1 - \Phi_1 B_s - \Phi_2 B^{2s} - \ldots - \Phi_P B^{P_s} \tag{2}$$

$$\Theta(B^s) = 1 - \Theta_1 B_s - \Theta_2 B^{2s} - \ldots - \Theta_Q B^{Q_s} \tag{3}$$

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p \tag{4}$$

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_p B^p \tag{5}$$

where B is the backshift operator; Θ and Φ are, respectively, the seasonal and non-seasonal polynomials representing the autoregressive component of the model; ϕ and θ are the seasonal and non-seasonal polynomials representing the moving average component of the model; s is the length of the seasonal period; a_t is a white noise series assumed to be independent and identically distributed with mean 0 and variance 1; D and d are the numbers of seasonal and non-seasonal differences needed for stationarity, respectively; (p,q) and (P,Q) define the orders of the non-seasonal and of the seasonal parts of the model, respectively. Conventionally, the specification of SARIMA family models is denoted as ARIMA(p, d, q) × SARIMA(P, D, Q)$_s$, with the abovementioned definition of the parameters p, d, q, P, D, Q, s. According to Lippi et al. (2013), the structure of the time series model can be illustrated through a directed graphical model, where black nodes represent the observed time series, grey nodes represent the unobserved noise process, and the white node represents the random variable that is to be predicted. An example is provided in Fig. 2, which reproduces a SARMA model.

In this study, different model specifications are tested. In the particular case of a linear SARMA model, the forecasted value of the speed on a generic link at time $t+1$ can be expressed in a closed-form as a linear combination of the previous values and the moving average of the random terms estimated on q and Q previous periods for the non-seasonal and seasonal components, respectively:

$$\widehat{v}_{t+1} = \mu + \sum_{t=1}^{p}\phi_i v_{t-i} + \sum_{t=1}^{P}\Phi_i v_{t-i-s} + \sum_{t=1}^{q}\theta_i a_{t-i} + \sum_{t=1}^{Q}\Theta_i a_{t-i-s} \tag{6}$$

Application to forecasting requires recursion of the observed values to estimate unobserved disturbances (Ansley, 1979):

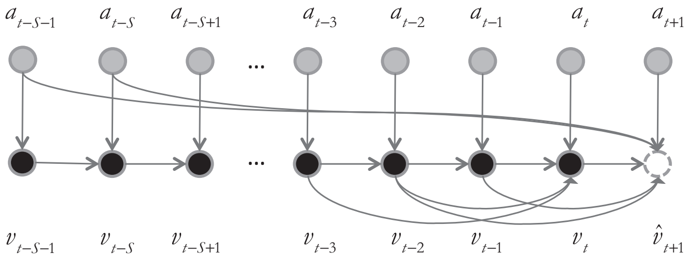

Fig. 2. Architectural graph of a SARMA $(3,0,0)(0,0,1)_S$ model.

$$\widehat{v}_{t+1} = \sum_{t=1}^{p}\phi_i v_{t-i} + \sum_{t=1}^{P}\Phi_i v_{t-i-s} + \sum_{t=1}^{q}\theta_i(v_{t-i} - \widehat{v}_{t-i}) + \sum_{t=1}^{Q}\Theta_i(v_{t-i-s} - \widehat{v}_{t-i-s}) \tag{7}$$

2.2. Bayesian network

Bayesian networks (BNs) are probabilistic graphical models. This definition outlines the two components that must be specified in a BN: a graphical component, represented by a directed acyclic graph, and a probabilistic component, expressed by probability distributions. In particular, each node of the graph represents a random variable, while the links that connect the nodes represent probabilistic dependencies between the corresponding random variables. The cause-effect relations used in BNs can be represented by considering the neighbor links in the case of traffic dynamics. The forecasted value on a generic link is formulated as the expected value of the posterior probability function of the speed, attained as the result of an a priori estimation $f^0{}_V$ and conditioned by the probability density function f_U of the variables **u** observed in the recent past on some parent nodes.

According to Fusco et al. (2016) we assume that the set L of parent nodes is composed of the forward and the backward stars of the downstream and upstream nodes of the target link, respectively. In this study, moreover, we introduce the vector of variables **u** = {**v**, σ, **n**}, consisting in the average speeds **v** = $\{v^1{}_t, v^1{}_{t-1}, \ldots, v^1{}_{t-p}, v^2{}_t, v^2{}_{t-1}, \ldots, v^2{}_{t-p}, \ldots, v^l_t, v^l_{t-1}, \ldots, v^l_{t-p}\}^{\mathrm{T}}$, the standard deviation of speeds σ = $\{\sigma^1{}_t, \sigma^1{}_{t-1}, \ldots, \sigma^1{}_{t-p}, \sigma^2{}_t, \sigma^2{}_{t-1}, \ldots, \sigma^2{}_{t-p}, \ldots, \sigma^l_t, \sigma^l_{t-1}, \ldots, \sigma^l_{t\text{-}p}\}^{\mathrm{T}}$, and the number **n** = $\{n^1{}_t, n^1{}_{t-1}, \ldots, n^1{}_{t-p}, n^2{}_t, n^2{}_{t-1}, \ldots, n^2{}_{t-p}, \ldots, n^l_t, n^l_{t-1}, \ldots, n^l_{t-p}\}^{\mathrm{T}}$ of vehicles observed in p previous time intervals on the l links belonging to the set L. By applying the chain rule, the forecasted value v_{t+1} can be formulated as follows.

$$\widehat{v}_{t+1} = E\left[f^0_v(v_{t+1})f_U(u_j)\prod_{j=1}^{N}\right] \tag{8}$$

where is $N = m{\cdot}l{\cdot}p$, as before.

The BN architecture is represented in graphical form in the Fig. 3 where the single components of the vector **u** are represented to highlight the space-time correlations assumed among the variables. Such double star architecture revealed to be the best trade-off between complexity of the architecture and accuracy of predictions. In fact, it has the great advantage to be modular, so that it can be easily implemented on large networks through simple automatic routines that explore the road graph, select the forward star of the end node and the backward star of the initial node for each prediction segment, and build a forecasting system for parallel speed prediction of each link of the road network.

2.3. Artificial neural network

The well-known Feed-Forward Neural Network (NN) is selected among the numerous specifications of neural networks, because it revealed to be more suitable than recursive networks for predicting phenomena affected by a relevant fraction of missed data (Fusco et al., 2016). NN is a static nonlinear vector multivariate function that forecasts the future value of speed v_{t+1} on an output link as a nonlinear function of the observed values **u** = $\{v_j\}$ of m traffic variables detected in p previous time intervals on l links, including the output one, being $j = 1, 2, \ldots, N = m \cdot l \cdot p$.

$$\widehat{v}_{t+1} = f_1\left[\sum_{h=1}^{H}\alpha_h f_2\left(\sum_{j=1}^{N}\beta_{j,h}\mu_j + \delta_j\right) + \delta_h\right] \tag{9}$$

In the formula, t is the current time interval, H is the number of neurons in the hidden layer, f_1 and f_2 are nonlinear activation functions, $\{\alpha_h\}$ and $\{\beta_{j,h}\}$ are coefficient matrices, δ_j and δ_h are threshold values associated with the hidden and output layers, respectively. The same double star architecture assumed for the BN is implemented for the NN. The variable vector **u** is composed of the speed values observed in the previous time interval on the prediction link and on the links belonging to its double star. A graphical representation of the architecture of the NN model is depicted in Fig. 4.

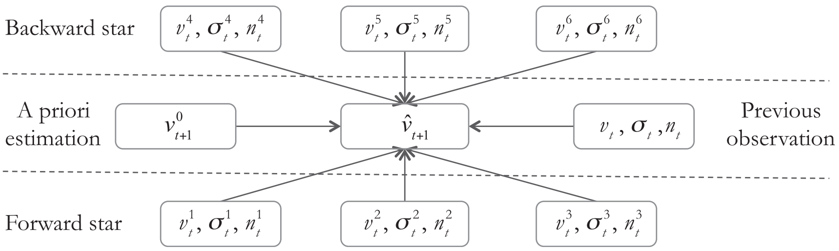
Backward star
v^4_t, σ^4_t, n^4_t
v^5_t, σ^5_t, n^5_t
v^6_t, σ^6_t, n^6_t
A priori estimation
v^0_{t+1}
$\hat{v}_{t+1}$
v_t, σ_t, n_t
Previous observation
Forward star
v^1_t, σ^1_t, n^1_t
v^2_t, σ^2_t, n^2_t
v^3_t, σ^3_t, n^3_t

Fig. 3. Double star architectural graph of the Bayesian Network in the case: $p = 1$; $l = 6$.

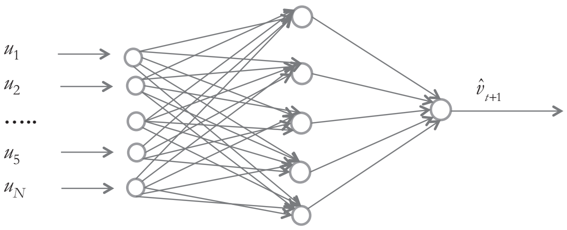

Fig. 4. Architectural graph of the Feed-Forward Neural Network.

3. Case study

3.1. Data set

The study area is composed of the primary urban road network of the EUR district in the Southern area of Rome, depicted in 0. The complete data set included one month of raw Floating Car Data obtained by a fleet of about 100,000 GPS equipped private vehicles, corresponding to about the 2.5% of the whole vehicular fleet of the town. Every data point, detected with a frequency rate of 1 reading every 2 min, reports the individual position and speed, the state of the engine (turned on, turned off, in motion), the traveled distance from the previous measurement, the direction of motion, and the quality of GPS signal (depending on the number of satellite signals received).

3.2. Time and space aggregation of raw data

Raw data were aggregated into 5-min time intervals. For each interval, the mean speed, the number of observations and the standard deviation of speed were computed. The nature of data used for this application implies that the average speed on a link is a random variable that can be estimated only on a sample of available FCD positions. Since the accuracy of the estimate depends upon the number of available FCD positions and the inter-vehicular variance of individual speeds, to achieve a suitable trade-off between accuracy and granularity of the estimation, the FCD were matched to an aggregated graph composed of 120 links having an average length of 520 m. About 300,000 high-quality individual observations were available on the selected sub-network.

A specific analysis was conducted about the effect of different time aggregation intervals on the accuracy of the average speed estimates obtained by FCD. Setting the confidence interval to a 10% deviation from the estimated value, the fractions of observations having confidence levels $P > 0.95$, $P > 0.90$ and $P > 0.80$ were computed for different time aggregation intervals, ranging from 1 to 15 min. Results are shown in Fig. 6. At 5-min aggregation level, only the 13% of observations has a 10% or better accuracy with a confidence level higher than 0.80. This percentage reduces to about 10% for 3-min aggregation level and to about 6% for 1-min aggregation level. However, it increases to 17.3% for 10-min aggregation and to 20.8% for 15-min aggregation. Similar trends but lower percentages can be observed for higher confidence levels. While it might be desirable to get a higher significance of data measures, a too long time aggregation interval would preclude a quick update of the traveler information system which the prediction system is designed to. As pointed out in Fusco and Gori (1996), a minimum update time of 5 min is required to ensure traveler information systems be beneficial on narrow grid congested urban networks. Thus, this value was chosen for temporal aggregation.

The frequency distribution of individual speeds observed during the whole reference period is illustrated in Fig. 7 for the 20 highest-frequency links. By observing the frequency distribution of raw data (grey curve in the figure), it is interesting to notice that some links show a clear Gaussian shape frequency; these are uninterrupted flow road segments (links *f*, *i*, *k*, *o*, *p*, *q*, *r*, *t* in the figure). A second group of links (denoted by *a*, *b*, *c*, *d*, *g*, *h*, *j*, *l*, *m*, *n*, *s* in the figure), other than the link *r* already included in the first group, exhibit a very high-frequency value near zero speed; these links are either signalized or ramp road segments. While analyzing the frequency distribution of the corresponding average estimated speed (blue curve in the figure), the variance reduces and more regular shapes are obtained. Specifically, the first group maintains the Gaussian shape; some links of the second group, represented by signalized road segments, lose after aggregation the peak of frequency near zero speed and take an asymmetric bell-shaped distribution (links *d*, *j*, *n*, *s*); the remaining links, composed by road ramp segments (links *c*, *b*, *e*, *g*, *h*, *l*, *m*), show a double bell shape, which reveals the existence of two different phenomena: a free-flow condition whose speed values distribute normally around the average desired speed and a congested condition that averages the low values of the observed speeds. It is worth mentioning that the two more spurious links, *e* and *r*, are both off-ramp segments. The latter has a signal in the middle, so that zero speed value has a high frequency, while the former shows an irregular frequency distribution, as other road ramp segments, with a low value at zero speed because there is no signal along it.

Most authors in the literature (see for example Cai et al., 2016) preprocess or filter the data to clear them from the noise component. Specifically, Chan et al. (2012) experienced that training an artificial neural network on filtered data instead on raw measures improved the predictions of the observed values in the validation phase with respect to a similar neural

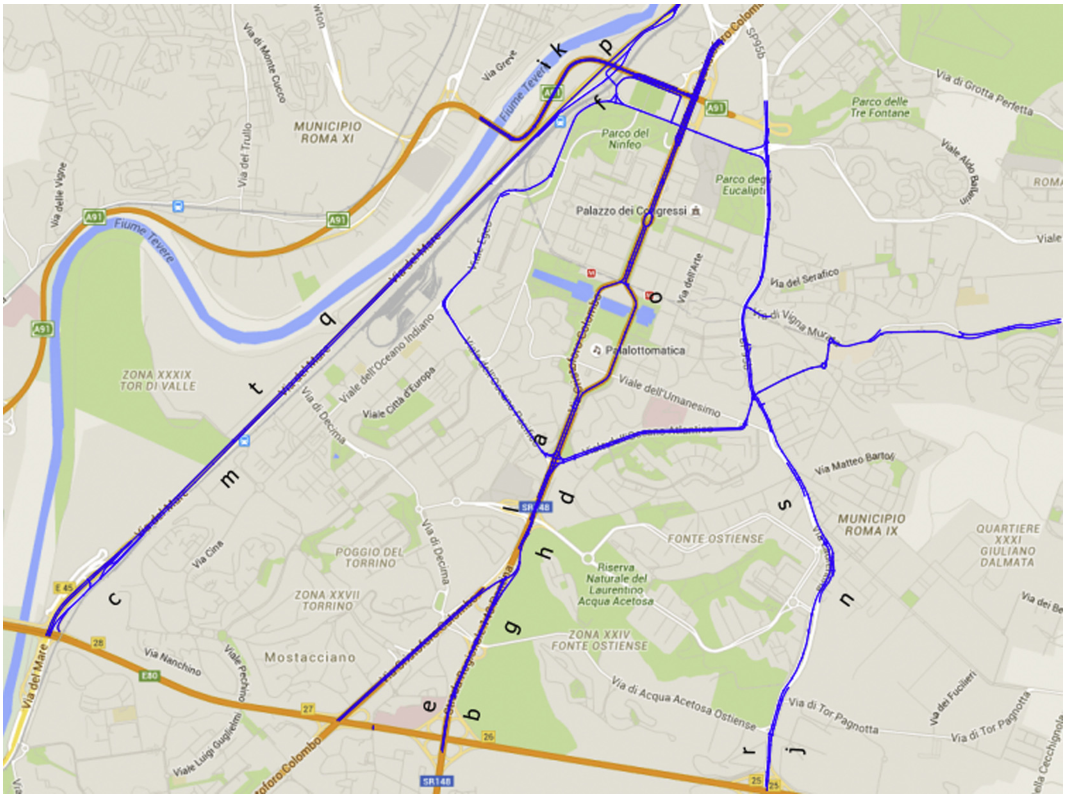

Fig. 5. Study area in Rome, Italy, with highlighted the 20 highest-frequency links.

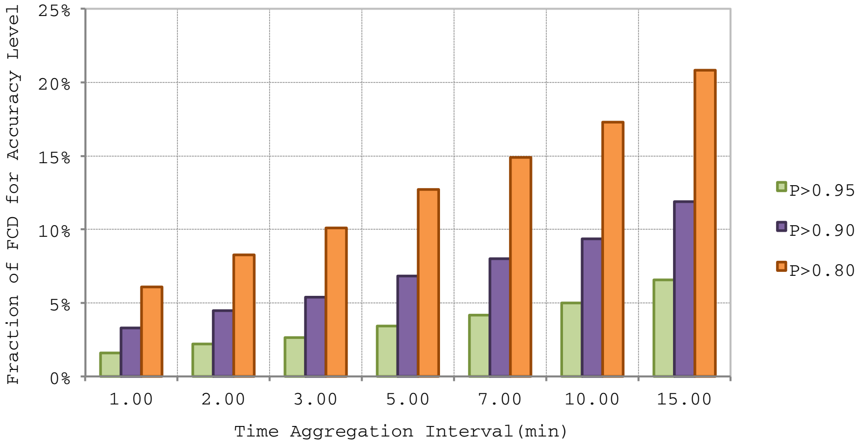

Fig. 6. Fraction of observations with different confidence levels *P* of the estimation of the average speed for different time aggregation intervals.

network trained on raw data. However, such a result depends on the specific data collected and cannot be seen as general, since we did not achieve comparable benefits in our case. More important, we are interested in taking advantage of the information contained in individual positioning data in order to get a measure of reliability to the data and exploit it to both improving the prediction capabilities of the model and deriving a measure of accuracy for the predictions. Actually, our method uses raw data to compute the inter-vehicular variance and the number of observations in each time period and each road link; these variables are then fed into the prediction model as measures of data reliability. An example of the raw data distribution for two different links, namely *f* and *a* in Fig. 5, is provided in Fig. 8. In order to discriminate the data trend from inter-vehicular speed variance, the simple average in each time 5-min time intervals, the corresponding confidence intervals for 0.80 confidence level and a moving average filter on the last 3 observations are also depicted in the figure.

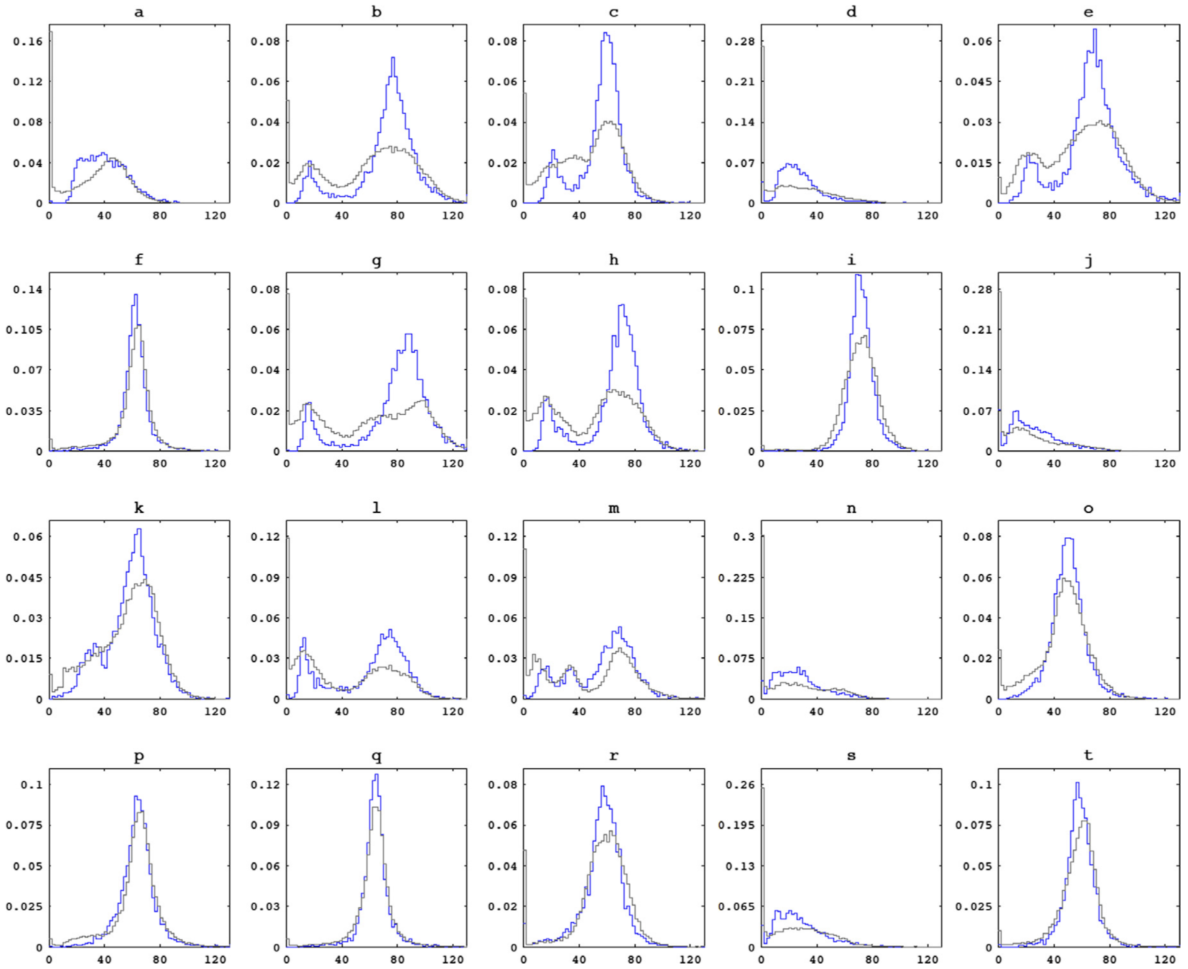

Fig. 7. Frequency distribution of individual speeds observed during one month on the 20 highest-frequency links in the study area (in grey) and the corresponding estimated averages with 5-min aggregation (in blue). Links locations labeled by *a* to *t* are individuated in Fig. 5. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

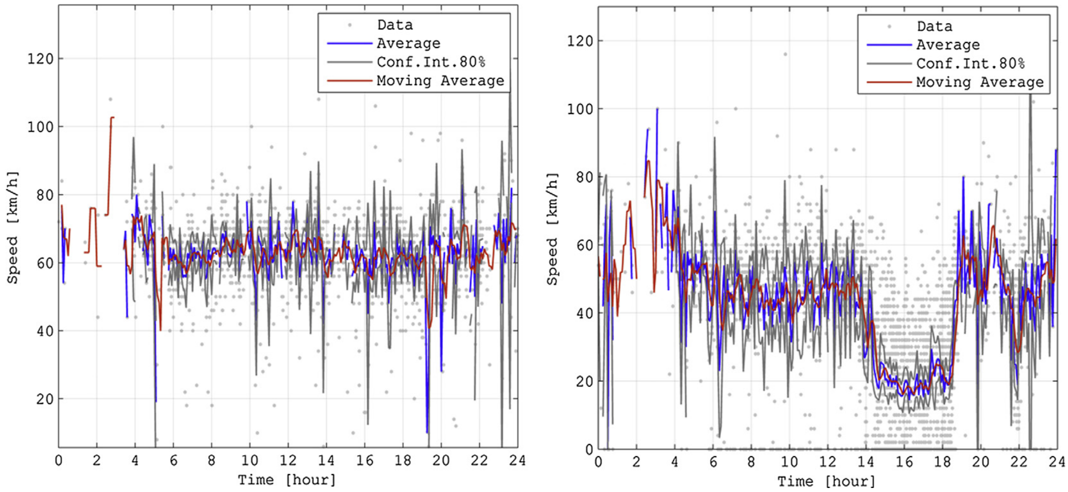

Fig. 8. Raw data, the average observed speed in one typical day, the corresponding confidence intervals for 0.80 confidence levels and the moving average on the links denoted as *f* (on the left) and *a* (on the right) in Fig. 5.

A large dispersion of individual speeds is observed both under uncongested conditions, due to the geometrical characteristics of the urban roads under study that allow even much higher speeds than the legal limit of 50 km/h, and under congested conditions, when lower speed are observed due to the presence of long queues a signal on link *a*.

3.3. Error indicators

The introduction of error indicators to compare different prediction methods requires a thorough assessment of the accuracy of the reference measure. Since the estimates of the average speed obtained by FCD are affected by variable confidence levels, in addition to traditional error indicators (MAE, MAPE; RMSE, RMSN), we introduce a new error indicator, the Relative Accuracy of Prediction (RAP), which compares the model accuracy with the accuracy of the reference measures. In other words, this indicator reduces the score of prediction inaccuracy if also the corresponding measures are inaccurate. Another indicator, $\mathrm{PAE}_{10\%}$, is related to the percentage of highly inaccurate predictions, independently of measure accuracy.

In summary, the following error indicators are introduced.

- Mean Absolute Error (MAE) : $\sum_{i=1}^{n} \frac{|\widetilde{x_i} - x_i|}{n}$
- Mean Absolute Percentage Error (MAPE): $\frac{1}{n}\sum_{i=1}^{n} \frac{|\widetilde{x_i} - x_i|}{x_i}$
- Root Mean Square Error (RMSE): $\sqrt{\sum_{i=1}^{n} \frac{(\widetilde{x_i} - x_i)^2}{n}}$
- Root Mean Square Error Normalized (RMSEN): $\frac{\mathrm{RMSE}}{\sum_{i=1}^{n} \frac{x_i}{n}}$
- Percentage of absolute errors greater than 10% ($\mathrm{PAE}_{10\%}$): $f(|\mathrm{MAPE}| > 10\%)$
- Relative Accuracy of Prediction (RAP): $\frac{1}{n}\sum_{i=1}^{n} \frac{|\widetilde{x_i} - x_i|}{\mathrm{conf.int}^{P}}$

where $\widetilde{x_i}$ is the forecast, x_i is the observed value at time *i*, *n* is the size of observation set, $\mathrm{conf.int}^P$ is the width of the confidence interval computed for *P* level of confidence.

Many other error measurements exist and some were recently applied also in BN predictions (de Oña et al., 2013; Chen et al., 2015), such as the Receiver Operating Characteristic (ROC) curve and the Area Under the ROC Curve (AUC value). These indicators provide effective measures of errors in classification problems, but are not apt for models that perform estimations of continuous variables such single models introduced in this paper, nor for combined forecasts obtained by a supervisor, as that introduced in Section 4.3, which applies different models selected according to a given traffic-dependent criterion without any automatic congestion type classifier.

3.4. Model calibration

All the models were calibrated on the data collected in the weekdays of first three weeks in order to define conditions similar to the real life operations, when the prediction model can be launched online in a rolling horizon framework and periodically validated and recalibrated, if necessary. The variables of BN model are assumed to be distributed according to Gaussian probability density functions. Coefficients of the distribution functions were calibrated by EM algorithm (Murphy, 2001), which maximizes the expected value of likelihood. It is widely used in the case of missing data; when using FCD the missing data are represented by some links that remain unobserved during some time intervals. NN model was calibrated by Levenberg-Marquardt algorithm, a well-known approximation to the Newton's method, which was said to be more efficient in comparison to other methods for convergence of the Backpropagation algorithm in the case of moderate-sized feedforward neural networks (Hagan and Menhaj, 1994; Tiwari et al., 2013). Missed data were substituted with the last observed value. The a priori estimate for the Bayesian network is provided by a time-series model with seasonality, whose benefit is to remove the trend before feeding the data into prediction models (Li et al., 2015b). In order to determine the most appropriate SARIMA model, we selected several model specifications by evaluating their performances on the 20 links with highest numbers of observations in the whole month. Since the models were calibrated only on weekdays, we chose a seasonality period of 1 day (that is, 288 intervals of 5 min). This assumption precludes catching the specific characteristics of individual weekdays and introduces a bias that can be assessed as about 4.5%, that is the average coefficient of variation of the speed for each weekday from the average day. To compare the performance of different models we used the Akaike Information Criterion: $AIC = -2\ln(L) + 2k$, where k is the number of estimated parameters and L is the maximized likelihood value, which leverages model complexity, expressed by the number of parameters that need to be estimated, and model accuracy, expressed by likelihood. Table 1 provides the numeric values of AIC for the different models calibrated; the values have been normalized to the maximum AIC value. Among SARIMA models (ones including a differencing term), ARMA(3,0,0) $\times$ SARIMA(0,1,1)288 provided the best results; while for SARMA non-differenced models, ARMA (3,0,0) $\times$ SARMA(0,0,1)288 is the best performing model. We choose both these models for the extended evaluation on the whole study area and we denote them shortly as SARIMA and SARMA in the following.

Table 1
Akaike Information Criterion for different SARIMA models based on the weekdays of the first three weeks.

Model specification	AIC
$\text{ARMA}(3,0,0) \times \text{SARMA}(0,0,1)_{288}$	0.983
$\text{ARMA}(2,0,0) \times \text{SARMA}(0,0,1)_{288}$	0.988
$\text{ARMA}(2,0,1) \times \text{SARMA}(0,0,1)_{288}$	1.000
$\text{ARIMA}(2,1,1) \times \text{SARMA}(0,0,1)_{288}$	0.978
$\text{ARIMA}(2,1,0) \times \text{SARMA}(0,0,1)_{288}$	0.989
$\text{ARMA}(3,0,0) \times \text{SARIMA}(0,1,1)_{288}$	0.925
$\text{ARMA}(2,0,0) \times \text{SARIMA}(0,1,1)_{288}$	0.926
$\text{ARMA}(2,0,1) \times \text{SARIMA}(0,1,1)_{288}$	0.948
$\text{ARMA}(1,0,1) \times \text{SARIMA}(0,1,1)_{288}$	0.950

3.5. Model validation

Model validation was performed by comparing speeds forecasted by different models with the corresponding average speeds observed during the last five weekdays of the month. Table 2 reports the error indicators for different models: the two seasonal autoregressive moving average models (SARMA and SARIMA), and the corresponding Bayesian Networks with a priori estimation obtained by time-series models (BN SARMA and BN SARIMA), Neural Network (NN) computed on the whole area for different levels of confidence *P* of the average speed estimates. To appreciate the contribution achieved by applying the prediction models to speed forecasts with respect to a simple data filtering technique, we applied as simple predictors the moving average filter (MA) computed on the last three observations and the Historical Average (HA) computed on the calibration dataset.

BN with SARMA a priori estimate provides the best results for all the confidence levels examined: as the confidence level increases from 0.80 to 0.95, MAE and RMSE indicators reduce by 6.7% and 4.5% respectively. NN shows similar indicators with MAE and RMSE reducing from 7.2% and 4.9% respectively, while SARIMA and BN SARIMA models show the highest error measures. This is an unexpected result because the first seasonal difference introduced in SARIMA, performing a stationary transformation with respect to recurrent with-day variation, is expected to improve traffic predictions, as denoted by the values of AIC indicator computed in the preselection phase performed on 20 links in the first three weeks and as often observed in the literature (Smith et al., 2002; Williams and Hoel, 2003). This issue is studied more in depth in the following, where specific analyses of results under recurrent and non-recurrent congestion conditions are presented. Since the differences between the errors of the various models are often within the accuracy interval of the measures, such specific analyses are useful to highlight when the different models exhibit significantly different performances. It is worth noticing also that

Table 2
Error indicators for the selected models and for the historical average on the validation data set for different confidence levels *P* of the average speed estimates.

Model	MAE	MAPE (%)	RMSE	RMSEN	$\text{PAE}_{10\%}$ (%)	RAP
			P > *0.80*			
BN SARMA	7.01	12	9.59	0.15	42	2.00
SARMA	7.76	13	10.54	0.17	47	2.22
BN SARIMA	7.16	12	9.77	0.15	43	2.05
SARIMA	9.43	15	12.94	0.20	54	2.73
NN	7.13	12	9.83	0.15	42	2.04
MA	8.70	14	12.17	0.19	50	2.52
HA	8.55	14	12.33	0.2	48	2.42
			P > *0.90*			
BN SARMA	6.73	11	9.28	0.15	39	1.53
SARMA	7.51	12	10.33	0.16	44	1.70
BN SARIMA	6.86	12	9.41	0.15	41	1.56
SARIMA	9.03	15	12.56	0.20	51	2.11
NN	6.82	12	9.48	0.15	40	1.54
MA	8.47	14	12.03	0.19	48	1.93
HA	8.23	14	11.83	0.19	47	1.89
			P > *0.95*			
BN SARMA	6.54	12	9.16	0.15	38	1.27
SARMA	7.37	13	10.23	0.16	43	1.43
BN SARIMA	6.64	12	9.22	0.15	39	1.30
SARIMA	8.74	15	12.20	0.19	50	1.82
NN	6.62	12	9.35	0.15	37	1.28
MA	8.36	15	11.92	0.19	47	1.63
HA	8.06	14	11.78	8.06	46	1.61

bigger errors (that is, greater than 10% of average observed speed) reduce when the confidence level of the entire set gets higher; for instance, for BN SARMA and NN the percentage of big errors reduces by 4% and 5%, passing *P* from 0.8 to 0.95 respectively. This means that bigger errors are often computed on data with lower significance. This conclusion is confirmed by the reductions of other error indicators with an increase of the level of confidence. RAP indicator, ranging from 2 to 1.27 specifically for BN SARMA, highlights that the magnitude of forecasting errors is almost in the same order of the confidence interval of the reference measure. Because the absolute error MAE decreases for higher *P* values, the observed reduction of RAP cannot be attributed only to the increase of the confidence interval.

In order to verify if the forecasting models are able to reproduce the distribution of the observed speed frequencies, we computed the analogous frequency distributions for the predicted values. To ensure the readability of the figures, here and in the following, we limit the number of curves to the models that have provided the overall best performances, namely: BN SARMA, SARMA and NN models. Results are illustrated in Fig. 9 for the same set of 20 links represented in Fig. 7 with reference to speeds observed in the 5 weekdays chosen for validation. As expected, distributions of predicted values for all models have smaller variances than the data, so that the frequencies corresponding to their medians are higher than those of the data. The distribution of predicted values fits the observations well, even when their distribution derives from the superimposition of two bell-shaped regimes, congested and uncongested. This means that correlated observations conditioned the multivariate probability distribution toward the congestion regime. This is formalized by the probabilistic structure of the BN and is empirically revealed in the case of NN. SARMA also catches the congested regimes but it often overestimates the congested speed because it lacks space correlation, which better reproduces the congestion propagation onto the road network.

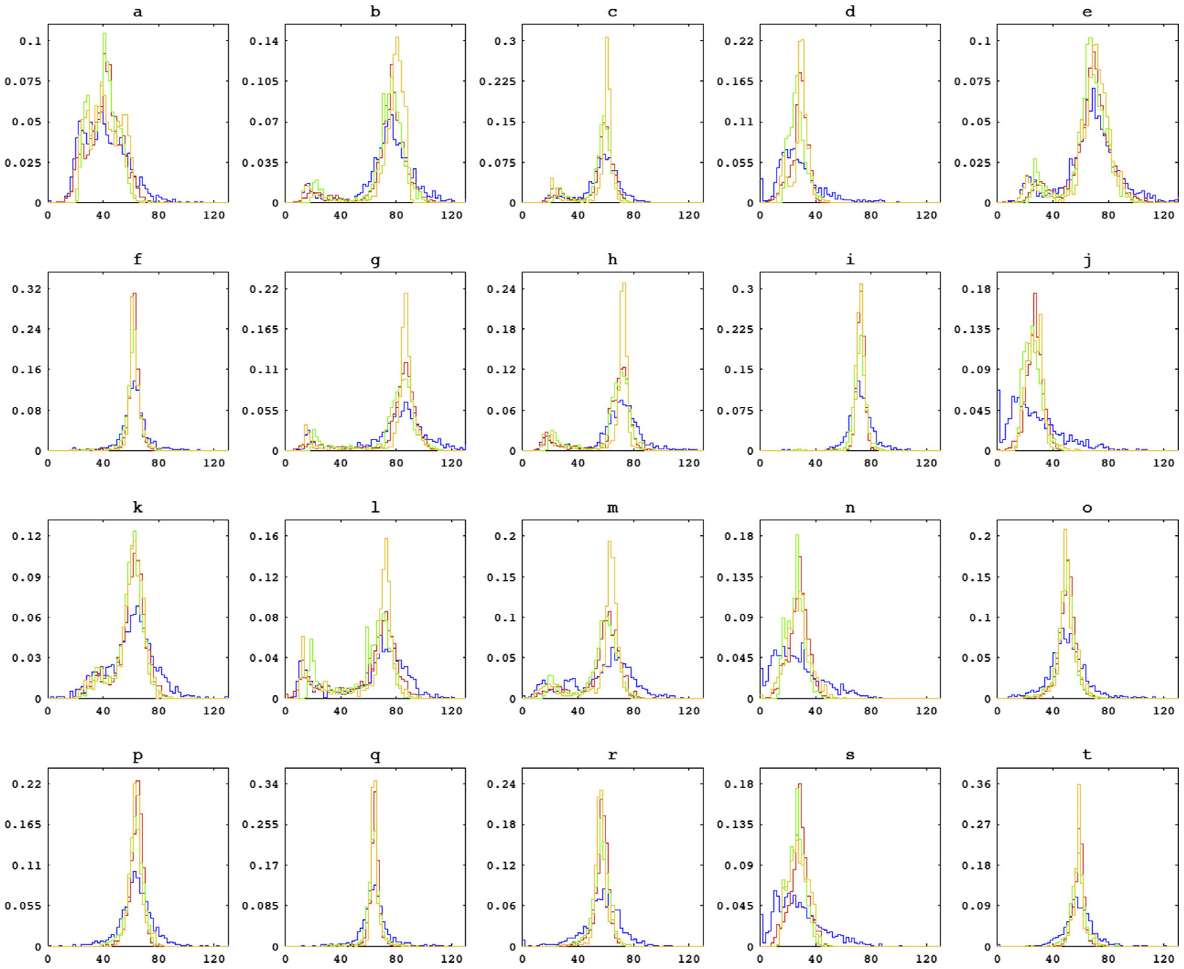

Fig. 9. Comparison of frequency distribution of individual speeds observed (blue) and predicted by Bayesian Network SARMA (red), Neural Network (orange), and SARMA (green) during the five days used for validation on the 20 highest-frequency links in the study area identified by labels in Fig. 5. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4. Performance analysis under different traffic congestion conditions

In order to assess model performances in different traffic patterns, we divided traffic condition into two groups: recurrent traffic condition, i.e. traffic pattern which is normally observed on a link, and non-recurrent traffic condition, which can be defined as a strong and sudden deviation from the standard situation.

4.1. Recurrent congestion

For practical purposes of analysis recurrent congestion time intervals are defined here as: $R = \left\{k : v^{HA}_{k+i} \leqslant \alpha v_o \cap s^{Pinf}_{k+i} \leqslant v_{k+i} \leqslant s^{Psup}_{k+i}\right\}, \forall i = 0, 1, \ldots, m$, where k is the generic time interval, v_k is the observed speed, v^{HA}_k is the historical average at time interval k, v_o is the free-flow speed, defined as usual as the 85th percentile speed on the link, α is a suitable reduction factor, s^{Pinf}_{k+i} and s^{Psup}_{k+i} are confidence interval bounds of the historical average speed for P level of confidence, m is the threshold for minimum congestion duration. This condition aims to capture recurrent speed reductions observed during the reference period; thus, it requires the historical average speed to be significantly low, while the observed speed to be within the historical confidence interval. The minimum congestion duration threshold m is set to 20 min and α factor is set to 0.5. The so defined conditions occurred in 1870 intervals. The main error indicators were computed for the different models on these intervals and the results are reported in Table 3.

Because the recurrent condition was defined concretely as a small deviation from the historical average profile, the historical average HA is the best speed estimator in this case. The proposed data-driven models are affected by random data fluctuations. With respect to the validation results obtained for the whole data set, lower MAE and RMSE values are observed, while RMSEN and the PAE are much higher because of the lower speed ($\alpha = 0.50$) imposed for congested conditions. Unlike the overall validation, in the specific case of recurrent congestion, as expected, SARIMA, which performs a stationary transformation of daily patterns, exhibits the best performances in terms of MAE = 5.18 km/h. NN and BN SARIMA are slightly worse in terms of MAE but better in terms of RMSE (6% and 7% respectively), while SARMA results the worst, although the differences are anyway quite small (13% for MAE). Differences in terms of RMSEN are negligible. This result highlights that spatial correlation included into NN model improves the forecasting accuracy with respect to the prediction based on only endogenous time-series supplied by SARMA. BN that exploits SARMA as a priori estimate provides intermediate results.

Fig. 10 illustrates 4 cases of recurrent congestion identified within the validation set. In this case, the observed data are within the historical confidence interval, which is rather narrow, due to the data nature. Both network-based models (that is BN and NN) forecasts almost always fall inside the confidence interval, while SARMA forecasts provide a larger overestimate and are often higher than the upper confidence bound.

4.2. Non-recurrent congestion

Non-recurrent congestion time intervals are defined here as: $NR = \left\{k : v_{k+i} \leqslant v^{HA}_{k+i} - s_v \cap v_{k+i} \leqslant s^{Pinf}_{k+i}\right\}, \forall i = 0, 1, \ldots, m$, where k is the generic time interval, v_k is the observed speed, s_v is an absolute deviation threshold from the historical average speed, and the other symbols are defined as before. This condition aims to capture only significant sequences of speed breakdown, and so to exclude random speed fluctuations. Introducing s^{Pinf}_{k+i} threshold, that takes into account the intrinsic variance of the phenomenon, ensures that the speed significantly deviates from standard condition while the absolute deviation s_v enables selecting only remarkable speed breakdowns. In this application we set s_v to 20 km/h, the level of confidence P to 95% and the minimum congestion duration m to 4, that is 20 min. There are 200 intervals corresponding to the described condition. Table 4 shows the performances of the different models in the selected cases of non-recurrent congestion.

As expected, error indicators are much higher than in the case of the overall data set, for several reasons: the training data set contained only rare cases of non-recurrent congestion which were too limited for the calibration of model; the abrupt change of the speed trend requires a latent time for model adapting; congestion conditions observed had a limited extent and mostly affected only a single link at the time. Relative error indicators (MAPE, PAE and RAP) are even higher than on the overall data set because of the lower average speed that is taken as reference term in heavy congestion conditions.

Table 3
Validation of the selected models compared with the historical average in cases of recurrent congestion.

Model	MAE	MAPE (%)	RMSE	RMSEN	$PAE_{10\%}$ (%)	RAP
BN SARMA	5.60	38	7.21	0.41	80	1.26
SARMA	5.86	41	7.35	0.42	82	1.39
BN SARIMA	5.24	35	7.13	0.41	76	1.12
SARIMA	5.18	33	7.60	0.43	74	1.08
NN	5.23	35	7.11	0.41	76	1.12
HA	2.63	17	3.22	0.18	63	0.57

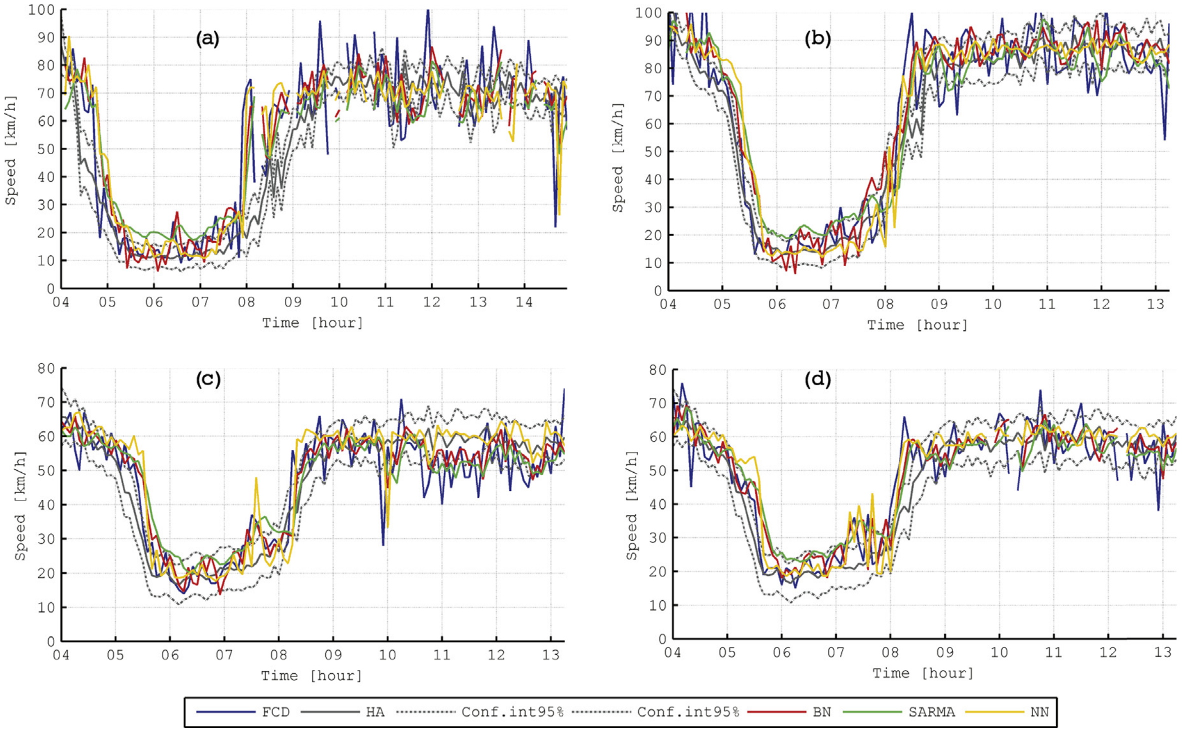

Fig. 10. Observed speed values (FCD) and forecasts provided by Bayesian Network (BN), SARMA and Neural Network (NN) models compared with the historical average and the confidence intervals for 95% confidence level in four example cases of recurrent congestion.

Table 4
Validation of the selected models compared with the historical average in cases of non-recurrent congestion.

Model	MAE	MAPE (%)	RMSE	RMSEN	$PAE_{10\%}$	RAP
BN SARMA	20.62	141	25.12	1.13	94	3.88
SARMA	13.02	85	17.65	0.80	84	2,42
BN SARIMA	22.25	151	27.99	1.32	91	3.90
SARIMA	28.06	187	32.01	1.51	97	5.50
NN	22.45	151	28.34	1.28	91	3.90
HA	34.79	230	36.35	1.64	100	7.38

Differently to the overall validation, SARMA model, which is highly affected by autoregressive terms, provides the best response to sudden changes of data trend and shows the best performances in cases of non-recurrent congestion, although error indicators are much higher than those obtained on the whole data set, RMSE = 17.65 and MAE = 13.02. It is to notice that SARMA, which is more sensitive to short-term within-day than day-by-day speed variations, performed better than SARIMA under non-recurrent congestion conditions (RMSE = 32.01 and MAE = 28.06). Such better performances of SARMA in non-recurrent congestion conditions overcompensate the weakness exhibited in recurrent congestion. This explains why also the overall performances in the application phase are slightly better than those of the first seasonal differenced SARIMA model. The same considerations hold for the Bayesian Network structured with SARMA a priori with respect to that with SARIMA a priori.

As obvious, Historical Average estimate (HA) is the worst estimator, since non-recurrent congestion is defined as a significant deviation from it. For illustrative purposes, speed forecasted by different models in non-recurrent congestion conditions in four different cases are reported in Fig. 11. It is interesting also to observe how BN SARMA almost replicates SARMA curve with a positive bias while NN model presents a different evolution. Depending on the form of congestion propagation, NN structure is highly affected by traffic patterns on conditioning links (that are backward and forward stars): in fact, if the conditioning links result noisy or uncongested, NN provides overestimated oscillating forecasts (cases b and d). Unluckily for research purposes, the data set used for validation, consisting of 5 days, included only two cases of congestion back-propagation (that is case a and case c in Fig 11). In the case a, the same non-recurrent congestion occurs only on 1 out of 3 conditioning links, while the remaining two links report very noisy data: this highly affected the NN forecast which also resulted in being noisy. In the case c, the forward-conditioning link was highly affected by abnormal trend while the

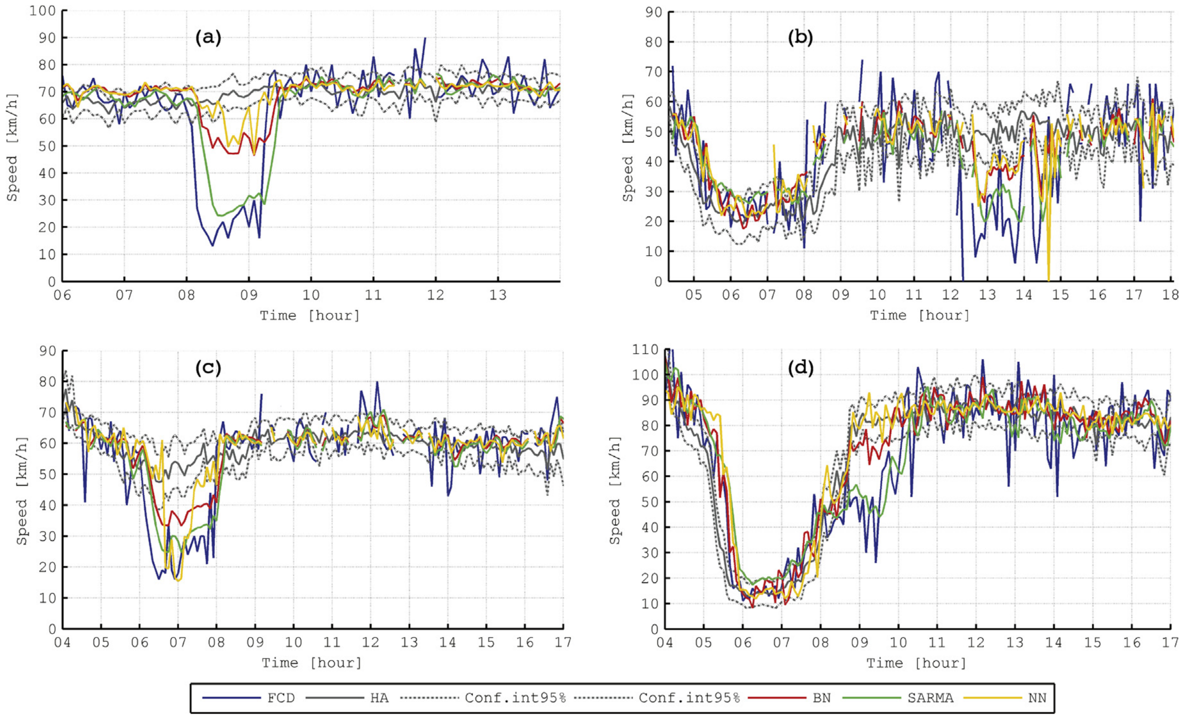

Fig. 11. Observed speed values (FCD) and forecasts provided by Bayesian Network (BN), SARMA and Neural Network (NN) models compared with the historical average and the confidence intervals for 95% confidence level in four example cases of non-recurrent congestion.

backward-conditioning link reported normal data: this case of the heaviest non-recurrent congestion is well-captured by NN model.

4.3. Supervisor mechanism

It is to notice that the best results are provided by SARMA model, which was the worst performer for recurrent congestion conditions. This counterintuitive result is due to the different structure of the two models. In fact, the Bayesian Network uses SARMA prediction, based on the 3 last speed values observed on the target link, as an a priori estimate for speed values in the next time interval. Such a priori estimate is then corrected by exploiting correlations with data collected in the last time interval on the close links. Since the models are trained on the whole dataset without discriminating between different traffic conditions, the Bayesian Network improves on average the a priori SARMA predictions, especially in recurrent conditions, which were observed repeatedly during the training phase and are well captured by the coefficients of the Bayesian Network. However, non-recurrent congested conditions were rarely observed; especially, in very few of them the congestion involved more than one link and spilled back to upstream links, which is the condition to exploit information from the links of the forward star. Thus, information from close uncongested links not only did not improve the prediction on a link affected by local congestion growth but in such cases can even worsen the prediction provided only by the simple time series of the speed observed in the last time periods on the same link.

These results suggest the opportunity of combining different models, depending on the currently observed traffic patterns. We try to obtain transportation-combined forecast, provided by a supervisor mechanism based on traffic conditions. The supervisor can be formulated as a decision-making model, which assigns the proper forecast based on estimated traffic conditions. Fig. 12 provides an example of a decision tree model for selecting the most appropriate forecast among those provided for different traffic conditions.

In this case of study, we combined the forecasts of SARMA model and BN under the following decision rule: whenever a non-recurrent congestion condition is detected the forecast for the next time interval is provided by SARMA a priori estimate, by nullifying the effect provided by the adjacent links; elsewhere the model framework is maintained by BN and the forecast is provided by BN with a priori SARMA estimate. In order to determine non-recurrent congestion we used *NR* condition as defined in Section 3.5, setting the absolute deviation from the historical average speed to 20 km/h and *m*, the number of previous intervals that satisfy *NR* condition to 1.

Fig. 13 reports an example of combined forecasts obtained by the supervisor. In the illustrated case two kinds of congestion took place: recurrent morning congestion (from 6 a.m. to 8 a.m.) evolved into abnormal congestion until 10 a.m. In this case both network-based models performed well during the recurrent early morning congestion (Fig. 10d); however, both

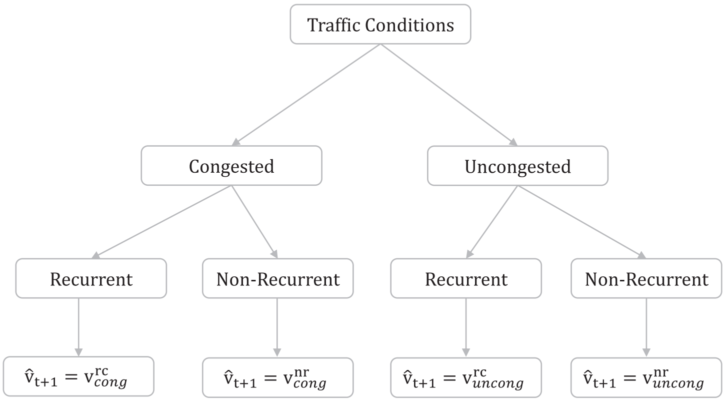
Traffic Conditions
Congested
Uncongested
Recurrent
Non-Recurrent
Recurrent
Non-Recurrent
$\hat{v}_{t+1} = v^{rc}_{cong}$
$\hat{v}_{t+1} = v^{nr}_{cong}$
$\hat{v}_{t+1} = v^{rc}_{uncong}$
$\hat{v}_{t+1} = v^{nr}_{uncong}$

Fig. 12. Decision tree supervisor model for selecting forecast based on observed traffic conditions.

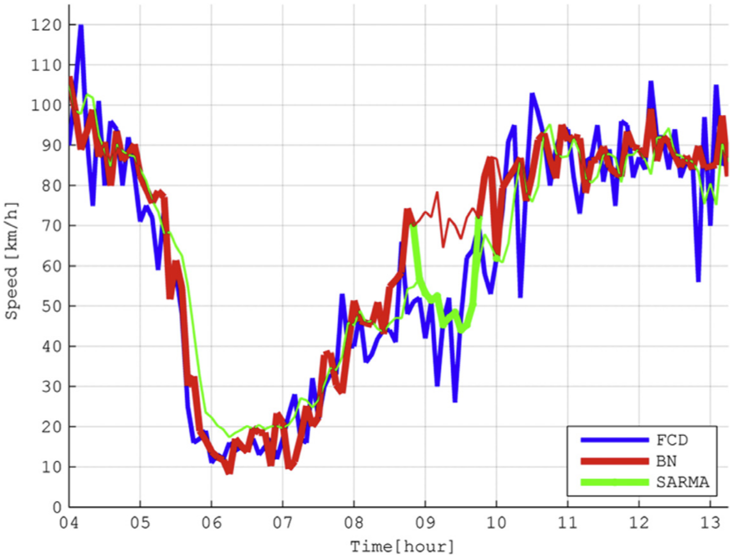

Fig. 13. Observed speed values (FCD) and combined forecasts provided by Bayesian Network (BN) and SARMA for different traffic conditions.

NN and BN failed to capture the abnormal trend. With respect to network-based models, SARMA provided worse results during the early morning but captured well the non-recurrent congestion occurred later. In the early morning, the supervisor mechanism detects normal conditions, that is a normally low speed, held within the historical confidence interval, and therefore BN model is applied. After 8 a.m. the observed speed is out of historical confidence interval and significantly lower than the historical average speed and the supervisor applies SARMA model.

Such a simple example has been introduced for illustrative purposes in order to underline the need of a focusing model analysis and development on different traffic conditions. The supervisor mechanism should be developed in a more comprehensive way, exploiting available machine-learning techniques, such as Random Forest, metamodels or Bayesian decision models. The most suitable attributes should be determined and the corresponding thresholds should be calibrated in order to obtain an unbiased combination of forecasts. In this case, due to the lack of data a simple decision rule based on observation assumptions was applied. However, even combining the forecasts by a trivial supervising rule, entailed a qualitative enhancement of the overall model performance. Indeed, providing an accurate forecast during the non-recurrent conditions is crucial in order to obtain real-time traffic management schemes.

5. Conclusions

The paper dealt with the problem of providing reliable short-term forecasts on urban road traffic networks by exploiting ubiquitous big data composed by individual point speeds from a large fleet of private cars. In order to reflect the transportation nature of the problem, the topology of the road network was taken into account by different network-based models, namely Bayesian Network (BN) and Neural Network (NN), trained to reproduce the spatial-temporal correlation between traffic variables. These models were compared with a Seasonal Auto Regressive Moving Average (SARMA) time-series model, which forecasts future traffic values from only observed time correlations. Moreover, following the Bayesian approach, BN model integrates SARMA as the a priori estimate. Space architecture of BN and NN models consists of previous measures on backward and forward stars links, other than on the prediction link, in order to reflect both forward traffic progression and congestion spillback. Such a simple regular structure makes possible to automate the engineering process of building a modular prediction system for large road networks by applying simple algorithms for graph exploration.

A large data set, containing raw floating car data was used to calibrate and validate the prediction models. Other than the average speed, inter-vehicular speed variance and the number of sampled vehicles were included into the model to account for average speed accuracy estimation. The validation phase was addressed to evaluate model performances in different traffic conditions in addition to the overall evaluation. Besides usual error indicators, a relative accuracy of prediction indicator was introduced to relate the forecasting accuracy to the accuracy of reference measures, which in sparse ubiquitous big data of floating cars varies both with time and with links. The results obtained in the overall evaluation showed that the network-based NN and BN models exhibited relative accuracy of prediction of 1.28 and 1.27, respectively. This means that their mean absolute prediction error is only about 27% greater than the measure error. However, SARMA showed a relative accuracy of prediction of 1.43, while it revealed to be the best performer for abnormal non-recurrent congestion conditions. Since SARMA model is an autoregressive endogenous model it can only capture the congestion that already took place; on the other hand, it is expected that including information from nearby links would enable capturing congestion propagation. However, during the time period taken for validation very rare conditions of congestion propagation were observed, so that very few conditions for appreciating the appropriateness of network structured forecasting models occurred.

Nevertheless, the qualitative differences in model performances suggested taking a first step towards the operational integration of BN and SARMA models under a supervisor mechanism that selects the proper forecast depending on the detected traffic conditions. Due to the lack of data only an illustrative example was derived from the supervisor implementation: under non-recurrent conditions the forecast was provided by SARMA model, and elsewhere the forecast was obtained by BN. Our ongoing research is focused on extending the validation of network-based machine learning models on a wider data set including evident cases of heavy congestion propagation observed in the whole town. Longer data sets will allow us to analyze the effect of long period seasonal variation and, if necessary, investigate ways to deal with such long period seasonality. More in general, the research is finalized to develop a supervising framework that applies machine learning techniques that integrate multiple predictions conditioned to the occurrence of different congestion regimes. Indeed, data collected from probe vehicles are growing exponentially in the present and they are expected to grow even more and more in the next future thanks to the diffusion of the new protocols of communications that will allow exploiting the data produced by a great quantity of in-vehicle sensors. Although the ultimate goal of the advances in vehicle-to-vehicle communications is the full deployment of autonomous driving on the road network, an intermediate step will be the production of updated information shared among vehicles to facilitate different tasks of human driving including congestion avoidance. These expectations are a further incentive to advance research efforts on the exploitation of big data generated by probe vehicles to predict short-term traffic conditions as well as on other forms of interaction of the vehicle with the surrounding environment.

References

Ahmed, M.S., Cook, A.R., 1979. Analysis of freeway traffic time-series data by using Box-Jenkins techniques. Transp. Res. Rec., 1–9

Ansley, C., 1979. An algorithm for the exact likelihood of a mixed autoregressive-moving average process. Biometrika 66 (1), 59–65.

Antoniou, C., Koutsopoulos, H.N., Yannis, G., 2013. Dynamic data-driven local traffic state estimation and prediction. Transp. Res. Part C Emerg. Technol. 34, 89–107. http://dx.doi.org/10.1016/j.trc.2013.05.012.

Baiocchi, A., Cuomo, F., De Felice, M., Fusco, G., 2015. Vehicular ad-hoc networks sampling protocols for traffic monitoring and incident detection in intelligent transportation systems. Transp. Res. Part C Emerg. Technol. 56, 177–194. http://dx.doi.org/10.1016/j.trc.2015.03.041.

Ben-Akiva, M., 1985. Dynamic network equilibrium research. Transp. Res. Part A: Gen. 19 (5), 429–431.

Ben-Akiva, M., Gao, S., Wei, Z., Wen, Y., 2012. A dynamic traffic assignment model for highly congested urban networks. Transp. Res. Part C 24, 62–82.

Bucknell, C., Herrera, J.C., 2014. A trade-off analysis between penetration rate and sampling frequency of mobile sensors in traffic state estimation. Transp. Res. Part C Emerg. Technol. 46, 132–150. http://dx.doi.org/10.1016/j.trc.2014.05.007.

Cai, P., Wang, Y., Lu, G., Chen, P., Ding, C., Sun, J., 2016. A spatiotemporal correlative k-nearest neighbor model for short-term traffic multistep forecasting. Transp. Res. Part C Emerg. Technol. 62, 21–34. http://dx.doi.org/10.1016/j.trc.2015.11.002.

Castillo, E., Menéndez, J.M., Sánchez-Cambronero, S., 2008. Predicting traffic flow using Bayesian networks. Transp. Res. Part B Methodol. 42, 482–509. http://dx.doi.org/10.1016/j.trb.2007.10.003.

Castro-Neto, M., Jeong, Y.-S., Jeong, M.-K., Han, L.D., 2009. Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions. Expert Syst. Appl. 36, 6164–6173. http://dx.doi.org/10.1016/j.eswa.2008.07.069.

Celikoglu, H.B., 2014. Dynamic classification of traffic flow patterns simulated by a switching multimode discrete cell transmission model. IEEE Trans. Intell. Transp. Syst. 15 (6), 2539–2550.

Celikoglu, H.B., Silgu, M.A., 2016. Extension of traffic flow pattern dynamic classification by a macroscopic model using multivariate clustering. Transp. Sci. http://dx.doi.org/10.1287/trsc.2015.0653.

Cetin, M., Comert, G., 2006. Short-term traffic flow prediction with regime switching models. Transp. Res. Rec. J. Transp. Res. Board 1965, 23–31. http://dx.doi.org/10.3141/1965-03.

Chan, K.Y., Dillon, T.S., Singh, J., Chang, E., 2012. Neural-network-based models for short-term traffic flow forecasting using a hybrid exponential smoothing and Levenberg–Marquardt algorithm. IEEE Trans. Intell. Transp. Syst. 13 (2), 644–654.

Chandra, S.R., Al-Deek, H., 2009. Predictions of freeway traffic speeds and volumes using vector autoregressive models. J. Intell. Transp. Syst.

Charle, W., Viti, F., Tampère, C., 2010. Estimating route travel time variability from link data by means of clustering. In: Proceedings of the 12th World Conference on Transport Research. Lisbon, Portugal, July 11–15, 2010.

Chen, C., Wang, Y., Li, L., Hu, J., Zhang, Z., 2012. The retrieval of intra-day trend and its influence on traffic prediction. Transp. Res. Part C Emerg. Technol. 22, 103–118. http://dx.doi.org/10.1016/j.trc.2011.12.006.

Chen, C., Zhang, G., Wang, H., Yang, J., Jin, P.J., Walton, C.M., 2015. Bayesian network-based formulation and analysis for toll road utilization supported by traffic information provision. Transp. Res. Part C: Emerg. Technol. 60, 339–359.

Cipriani, E., Fusco, G., Gori, S., Petrelli, M., 2006. Heuristic methods for the optimal location of road traffic monitoring. In: 2006 IEEE Intelligent Transportation Systems Conference. IEEE, pp. 1072–1077. http://dx.doi.org/10.1109/ITSC.2006.1707364.

Daganzo, C.F., 1994. The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. Transp. Res. Part B: Methodol. 28 (4), 269–287.

de Oña, J., López, G., Mujalli, R., Calvo, F.J., 2013. Analysis of traffic accidents on rural highways using latent class clustering and Bayesian networks. Accid. Anal. Prev. 51, 1–10.

Deng, W., Lei, H., Zhou, X., 2013. Traffic state estimation and uncertainty quantification based on heterogeneous data sources: a three detector approach. Transp. Res. Part B Methodol. 57, 132–157. http://dx.doi.org/10.1016/j.trb.2013.08.015.

Dougherty, M.S., Cobbett, M.R., 1997. Short-term inter-urban traffic forecasts using neural networks. Int. J. Forecast. 13, 21–31. http://dx.doi.org/10.1016/S0169-2070(96)00697-8.

Dunne, S., Ghosh, B., 2012. Regime-based short-term multivariate traffic condition forecasting algorithm. J. Transp. Eng. 138, 455–466. http://dx.doi.org/10.1061/(ASCE)TE.1943-5436.0000337.

Feng, Y., Hourdos, J., Davis, G.A., 2014. Probe vehicle based real-time traffic monitoring on urban roadways. Transp. Res. Part C Emerg. Technol. 40, 160–178.

Fusco, G., Colombaroni, C., Comelli, L., Isaenko, N., 2015. Short-term traffic predictions on large urban traffic networks: Applications of network-based machine learning models and dynamic traffic assignment models. In: 2015 International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS). IEEE, Budapest, Hungary, pp. 93–101. http://dx.doi.org/10.1109/MTITS.2015.7223242.

Fusco, G., Gori, S., 1996. The use of artificial neural networks in advanced traveler information and traffic management systems. In: Proceedings of the 1995 4th International Conference on Applications of Advanced Technologies in Transportation Engineering. ASCE, New York, NY, United States, Capri, Italy, pp. 341–345.

Fusco, G., Colombaroni, C., Isaenko, N., 2016. Comparative analysis of implicit models for real-time short-term traffic predictions. IET Intell. Transp. Syst. http://dx.doi.org/10.1049/iet-its.2015.0136.

Guo, J., Huang, W., Williams, B.M., 2014. Adaptive Kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification. Transp. Res. Part C Emerg. Technol. 43, 50–64. http://dx.doi.org/10.1016/j.trc.2014.02.006.

Hagan, M.T., Menhaj, M.B., 1994. Training feedforward networks with the Marquardt algorithm. IEEE Trans. Neural Networks 5 (6), 989–993. http://dx.doi.org/10.1109/72.329697.

Herrera, J.C., Bayen, A.M., 2010. Incorporation of Lagrangian measurements in freeway traffic state estimation. Transp. Res. Part B—Methodol. 44 (4), 460–481.

Herrera, J.C., Work, D.B., Herring, R., Ban, X. (Jeff), Jacobson, Q., Bayen, A.M., 2010. Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment. Transp. Res. Part C Emerg. Technol. 18, 568–583. http://dx.doi.org/10.1016/j.trc.2009.10.006.

Hofleitner, A., Herring, R., Bayen, A., 2012. Arterial travel time forecast with streaming data: a hybrid approach of flow modeling and machine learning. Transp. Res. Part B Methodol. 46, 1097–1122. http://dx.doi.org/10.1016/j.trb.2012.03.006.

Kamarianakis, Y., Prastacos, P., 2005. Space–time modeling of traffic flow. Comput. Geosci. 31, 119–133. http://dx.doi.org/10.1016/j.cageo.2004.05.012.

Kamarianakis, Y., Shen, W., Wynter, L., 2012. Real-time road traffic forecasting using regime-switching space-time models and adaptive LASSO. Appl. Stoch. Model. Bus. Ind. 28, 297–315. http://dx.doi.org/10.1002/asmb.1937.

Kim, S., Coifman, B., 2014. Comparing INRIX speed data against concurrent loop detector stations over several months. Transp. Res. Part C Emerg. Technol. 49, 59–72. http://dx.doi.org/10.1016/j.trc.2014.10.002.

Li, L., Li, Y., Li, Z., 2013. Efficient missing data imputing for traffic flow by considering temporal and spatial dependence. Transp. Res. Part C Emerg. Technol. 34, 108–120. http://dx.doi.org/10.1016/j.trc.2013.05.008.

Li, L., Su, X., Wang, Y., Lin, Y., Li, Z., Li, Y., 2015a. Robust causal dependence mining in big data network and its application to traffic flow predictions. Transp. Res. Part C Emerg. Technol. 58, 292–307. http://dx.doi.org/10.1016/j.trc.2015.03.003.

Li, L., Su, X., Zhang, Y., Lin, Y., Li, Z., 2015b. Trend modeling for traffic time series analysis: an integrated study. IEEE Trans. Intell. Trans. Syst. 16 (6), 3430–3439.

Lippi, M., Bertini, M., Frasconi, P., 2013. Short-term traffic flow forecasting: an experimental comparison of time-series analysis and supervised learning. IEEE Trans. Intell. Transp. Syst. 14, 871–882. http://dx.doi.org/10.1109/TITS.2013.2247040.

Lv, Y., Duan, Y., Kang, W., Li, Z., Wang, F., 2015. Traffic flow prediction with big data: a deep learning approach. IEEE Trans. Intell. Transp. Syst. 16, 865–873. http://dx.doi.org/10.1109/TITS.2014.2345663.

Ma, T., Zhou, Z., Abdulhai, B., 2015a. Nonlinear multivariate time–space threshold vector error correction model for short term traffic state prediction. Transp. Res. Part B Methodol. 76, 27–47. http://dx.doi.org/10.1016/j.trb.2015.02.008.

Ma, X., Tao, Z., Wang, Y., Yu, H., Wang, Y., 2015b. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. Transp. Res. Part C: Emerg. Technol. 54, 187–197.

Mahmassani, H.S., Fei, X., Eisenman, S., Zhou, X., Qin, X., 2005. DYNASMART-X Evaluation for Real-Time TMC Application: Chart Test Bed. Maryland Transportation Initiative, University of Maryland, College Park, Maryland, pp. 1–144.

Mai, T., Ghosh, B., Wilson, S., 2015. Short-term traffic-flow forecasting with auto-regressive moving average models. Proceedings of the Institution of Civil Engineers-Transport, vol. 167. Thomas Telford Ltd, pp. 232–239 (4).

Mihaylova, L., Boel, R., Hegyi, A., 2007. Freeway traffic estimation within particle filtering framework. Automatica 43 (2), 290–300.

Muñoz, L., Sun, X., Horowitz, R., Alvarez, L., 2003. Traffic density estimation with the cell transmission model. Proceedings of the 2003 IEEE American Control Conference, vol. 5, pp. 3750–3755.

Murphy, K., 2001. The Bayes net toolbox for Matlab. Comput. Sci. Statist. 33 (2), 1024–1034.

Oh, S., Byon, Y.-J., Jang, K., Yeo, H., 2015. Short-term travel-time prediction on highway: a review of the data-driven approach. Transp. Rev. 35, 4–32. http://dx.doi.org/10.1080/01441647.2014.992496.

Patire, A.D., Wright, M., Prodhomme, B., Bayen, A.M., 2015. How much GPS data do we need? Transp. Res. Part C 58, 325–342. http://dx.doi.org/10.1016/j.trc.2015.02.011.

Schneider IV, W.H., Turner, S.M., Roth, J., Wikander, J., 2010. Statistical Validation of Speeds and Travel Times Provided by a Data Service Vendor. No. FHWA/OH-2010/2. Univ. Akron 1–309.

Shi, Q., Abdel-Aty, M., 2015. Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transp. Res. Part C 58, 380–394. http://dx.doi.org/10.1016/j.trc.2015.02.022.

Smith, B.L., Williams, B.M., Keith Oswald, R., 2002. Comparison of parametric and nonparametric models for traffic flow forecasting. Transp. Res. Part C Emerg. Technol. 10, 303–321. http://dx.doi.org/10.1016/S0968-090X(02)00009-8.

Stathopoulos, A., Karlaftis, M.G., 2003. A multivariate state space approach for urban traffic flow modeling and prediction. Transport. Res. Part C: Emerg. Technol. 11 (2), 121–135.

Sun, X., Muñoz, L., Horowitz, R., 2003. Highway traffic state estimation using improved mixture Kalman filters for effective ramp metering control. Proceedings of 42nd IEEE Conference on Decision and Control, vol. 6, pp. 6333–6338.

Sun, S., Zhang, C., Yu, G., 2006. A Bayesian network approach to traffic flow forecasting. IEEE Trans. Intell. Transp. Syst. 7, 124–132. http://dx.doi.org/10.1109/TITS.2006.869623.

Tan, H., Feng, G., Feng, J., Wang, W., Zhang, Y.-J., Li, F., 2013. A tensor-based method for missing traffic data completion. Transp. Res. Part C Emerg. Technol. 28, 15–27. http://dx.doi.org/10.1016/j.trc.2012.12.007.

Tiwari, S., Naresh, R., Jha, R., 2013. Comparative study of backpropagation algorithms in neural network based identification of power system. Int. J. Comp. Sci. Inf. Technol. 5 (4), 93.

van Hinsbergen, C.P.I., van Lint, J.W.C., van Zuylen, H.J., 2009. Bayesian committee of neural networks to predict travel times with confidence intervals. Transp. Res. Part C: Emerg. Technol. 17, 498–509.

Vlahogianni, E.I., Karlaftis, M.G., Golias, J.C., 2014. Short-term traffic forecasting: where we are and where we're going. Transp. Res. Part C Emerg. Technol. 43, 3–19. http://dx.doi.org/10.1016/j.trc.2014.01.005.

Wang, Y., Papageorgiou, M., 2005. Real-time freeway traffic state estimation based on extended Kalman filter: a general approach. Transp. Res. Part B: Methodol. 39 (2), 141–167.

Wang, J., Deng, W., Guo, Y., 2014. New Bayesian combination method for short-term traffic flow forecasting. Transp. Res. Part C Emerg. Technol. 43, 79–94. http://dx.doi.org/10.1016/j.trc.2014.02.005.

Williams, B.M., Hoel, L.A., 2003. Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: theoretical basis and empirical results. J. Transp. Eng. 129, 664–672. http://dx.doi.org/10.1061/(ASCE)0733-947X(2003) 129:6(664).

Ye, Q., Szeto, W.Y., Wong, S.C., 2012. Short-term traffic speed forecasting based on data recorded at irregular intervals. IEEE Trans. Intell. Transp. Syst. 13, 1727–1737. http://dx.doi.org/10.1109/TITS.2012.2203122.

Zhang, G.P., 2003. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 50, 159–175. http://dx.doi.org/10.1016/S0925-2312(01)00702-0.

Zhang, H.M., 2000. Recursive prediction of traffic conditions with neural network models. J. Transp. Eng. 126, 472–481. http://dx.doi.org/10.1061/(ASCE)0733-947X(2000) 126:6(472).

Zhang, Y., 2014. Special issue on short-term traffic flow forecasting. Transp. Res. Part C Emerg. Technol. 43, 1–2. http://dx.doi.org/10.1016/j.trc.2014.05.009.

Zheng, W., Lee, D.-H., Shi, Q., 2006. Short-term freeway traffic flow prediction: bayesian combined neural network approach. J. Transp. Eng. 132, 114–121. http://dx.doi.org/10.1061/(ASCE)0733-947X(2006) 132:2(114).

Zhu, J.Z., Cao, J.X., Zhu, Y., 2014. Traffic volume forecasting based on radial basis function neural network with the consideration of traffic flows at the adjacent intersections. Transp. Res. Part C: Emerg. Technol. 47, 139–154.