Eur. Transp. Res. Rev. (2014) 6:333–342
DOI 10.1007/s12544-014-0133-5

ORIGINAL PAPER

A pattern matching approach to speed forecasting of traffic networks

V. Fabrizi · R. Ragona

Received: 10 September 2012 / Accepted: 20 January 2014 / Published online: 25 February 2014

Abstract
Objective Unlike other existing traffic data collection techniques, probe vehicles, or floating cars traveling on a road network, have appeared as a complementary solution for increasing coverage areas without requiring expensive infrastructure investments. When organized in a fleet with communication capabilities and exchange of information with a central data system, they give rise to a Floating-Car Data (FCD) system. The purpose of this paper is to present a model for short-term traffic speed forecasting based on an operating FCD system, developed and operated in Italy, delivering real-time traffic speed information throughout the Italian motorway network and along some important arterial streets located in major Italian metropolitan areas.
Design Specifically, a database covering the whole period ranging from April 2008 to October 2011 is available for Rome Ring Road, a toll-free motorway that encircles Rome (Italy), and the developed case study pertains to a portion of its available speed data.
Method A Pattern Matching method of prediction will be detailed, which reports interesting properties in terms of forecast accuracy; the method tries to identify, in the past data history, patterns neighboring (in a proper sense) the current one, which describes the actual traffic load, and produces forecasts supposing that the current situation will evolve in a similar way.

Keywords Floating car data (FCD) · Speed time series · Short-term speed forecasting · Pattern matching

1 Introduction

The successful wide-scale deployment of Advanced Traveler Information Systems (ATIS) and Advanced Traffic Management Systems (ATMS) relies significantly on the capability to perform accurate short-term predictions of traffic parameters over the entire road network [1]. Most of the studies and applications conducted so far used statistical or heuristic techniques to predict traffic conditions starting from historical and real-time traffic data collected by fixed sensor networks. Among these, we can mention loop detectors, which provide flows and speeds through a road section, and automatic vehicle identification systems, which keep track of the vehicles passing through a particular installed device or checkpoint and compute travel times from the difference between arrival times at consecutive checkpoints.

However these traffic information collection techniques have limitations in terms of spatial coverage due to the high cost of installation and maintenance.

Recently, mobile sensors or probe-vehicles have appeared as a complementary solution to fixed sensors for increasing coverage areas and prediction accuracy without requiring expensive infrastructure investments. Unlike other existing traffic data collection techniques, probe-vehicles act as moving sensors traveling in a traffic stream and do not require instrumentation to be set up on the roadway. Key and innovative aspects of probe-vehicle technology are that vehicle travel time is measured directly and the quality of data is fixed by the percent of vehicles monitored.

In this view the use of real-time FCD [2], based on traces of GPS positions, is emerging as a reliable and cost-effective method of collecting accurate traffic data for a wide-area road network and improving short-term predictions of travel conditions. This new method of urban traffic data collection has become a new frontier, and there are a number of theories about and applications of FCD [3].

The FCD technique is based on the exchange of information between a fleet of floating cars traveling on a road network and a central data system. The floating cars periodically send the recent accumulated data on their positions (latitude, longitude and altitude) and, optionally, instantaneous speed, whereas the central data system tracks the received floating car data along the traveled routes by matching the

V. Fabrizi · R. Ragona (✉)
ENEA—UTTEI/VEBIM, Via Anguillarese, 301, 00123 Rome, Italy
e-mail: roberto.ragona@enea.it

related trajectories data to the road network. The reliability of travel time estimates based on FCD is highly dependent on the percentage of floating cars participating in the traffic flow; other factors affecting the reliability of travel time estimates, mainly for lower penetration of floating cars, are traffic conditions and road link capacities.

As an intuitive rule, a lower percentage of floating cars is required in more congested traffic condition while a higher percentage of floating cars is needed in low/free flow conditions. In fact, following the usual assumption that drivers aim to minimize travel time, a few vehicles constrained to low speeds suffice to reliably denote a congested traffic condition (low data variability); on the contrary, in the case of low/free flow, driving styles affect individual speed choices, causing a larger data variability.

The purpose of this paper is to present a model for short-term traffic speed forecasting based on an operating FCD system, developed and operated in Italy by OCTO Telematics [4] [5], delivering real-time traffic speed information throughout the Italian motorway network and along some important arterial streets located in major Italian metropolitan areas.

Traffic speed values are deduced at an interval of 3 min from GPS traces transmitted in real-time from a large (and still growing) number of privately owned cars (more than 1,000,000) equipped with a specific device covering a range of insurance-related applications.

The paper is organized as follows: Sections 2 and 3 will present the FCD data that supported this research, Sections 4, 5 and 6 detail the proposed method, and finally Section 7 reports on its performance in terms of accuracy.

2 Outline of the OCTO Telematics probe-vehicle system

The aim of the Probe-Vehicle System (PVS) operated by OCTO Telematics is basically insurance profiling. This PVS is based on the wireless exchange of information between a large fleet of privately-owned probe-vehicles traveling across Italy and a Data Processing Center (DPC), which performs a variety of functions such as statistical computation on driver behavior, total mileage, accident detection and reconstruction, real-time traffic condition estimation, anti-theft satellite tracking and fleet management.

Each vehicle is equipped with an on-board unit (OBU) that integrates the following components: a GPS receiver, a GPRS transmitter, a three-axis accelerometer sensor, a battery pack, a mass memory, a processor and a RAM. The OBU stores GPS measurements (position, heading, speed, quality) and periodically transmits (on request or automatically) the recent accumulated measurements to the DPC. As aforementioned, a primary function performed by the DPC is the collection and processing of the received location and time information from the probe-vehicles in order to provide real-time estimates of traffic conditions, in terms of average link speeds, every 3 min, 24 h a day, 7 days a week.

The method of inferring traffic conditions from probe-vehicle position and time information involves four sequential steps, as follows: 1. map matching; 2. path identification; 3. travel time allocation; 4. travel speed aggregation.

Field tests were performed by comparing speed values obtained by OCTO Telematics PVS with measured values coming from the automatic vehicle identification systems installed on the Italian motorway network. The results [5] indicated that the speed estimates furnished by this PVS are within an error bound of +/− 10 % on the entire road network.

3 The Rome ring road case study

The case study selected for the development and evaluation of the proposed speed prediction method is Rome Ring Road (GRA—Grande Raccordo Anulare). The GRA is a toll-free motorway (68.2 km in circumference) that encircles Rome (Fig. 1). It is a dual carriageway with three lanes in both the clockwise (inner carriageway) and the anticlockwise (outer carriageway) direction.

The GRA has 33 numbered entry/exit junctions (starting from Aurelia Junction and proceeding in a clockwise direction) and is a major city traffic artery distributing traffic on radial routes and handling circumferential traffic in the city.

The GRA is a very challenging test bed because it is located in one of the most congested metropolises in the country. Traffic is heavy for most of the day and frequent delays and traffic-jams are experienced due to accidents or queue spillbacks from the exit ramps or the adjacent radial arterial streets leading into the city center.

The most severely congested links are in the east quadrant, specifically from Junction 10 (A1—Roma Firenze) to Junction 23 (Appia) (from km 21.0 to km 44.6, clockwise direction) and in the south quadrant from Junction 30 (Fiumicino) to Junction 18 (Casilina) (from km 7.6 to km 30.2, anticlockwise direction).

In a working day, about 15,000 probe-vehicles monitored by OCTO Telematics PVS pass through the GRA. The average distances traveled by these vehicles on the GRA is about 10 km. During the peak period, on average, more than 2,000 floating cars per hour travel on the GRA.

Currently, the penetration level of equipped private cars in Rome (about 2.5 %) is significantly higher than the national average. The link travel speed time series used in this study are organized in 3-min periods (480 values per link per day) and cover the whole period ranging from April 2008 to October 2011, for a total of 1,079 days effectively present in the database (apart from the incompletely measured days) and 517,920 speed values for each link.

Fig. 1 The Rome Ring Road (GRA)

Our analysis considers the link enclosed between Junction 11 (Nomentana) and 12 (Centrale del Latte), in the clockwise direction (see Fig. 1), in one of the more congested quadrants; it will be denoted as L_{11-12}.

The left panel of Fig. 2 shows the superimposition of all daily histories of this link, reporting a high variability during the available period and a complex structure; some daily profiles also denote congestion during night-time hours.

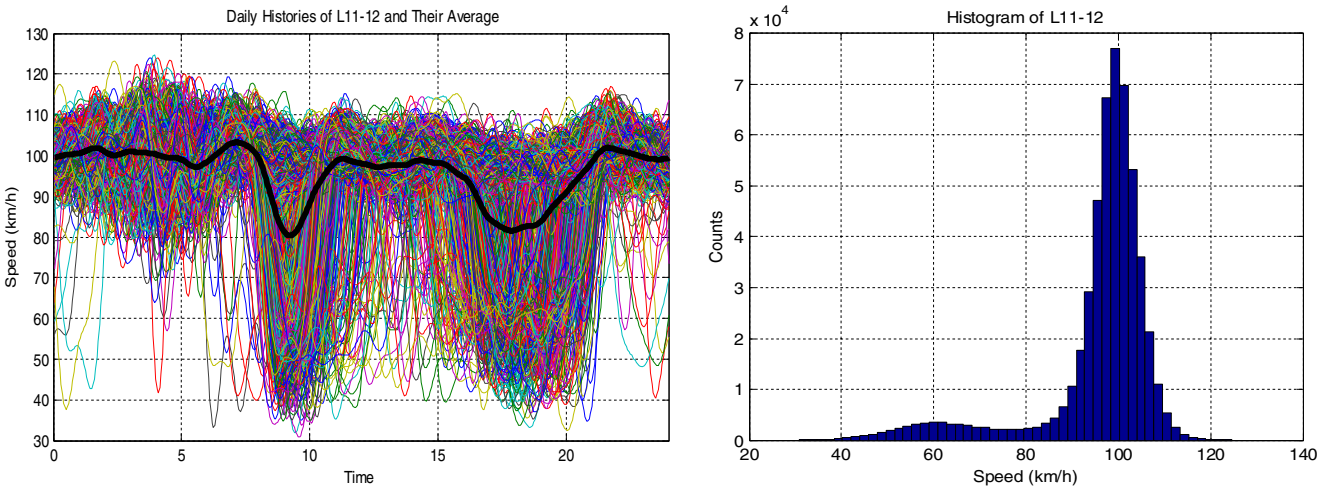


Fig. 2 Time/frequency characteristics of the link L_{11-12}

The bolded black curve presents the daily averaged speed; it is worth noting that it summarizes the erratic time evolution of this link poorly. This behavior suggested the use of an approach of *local modeling*, in contrast with global modeling, which does not benefit from the reported variability.

The right panel of Fig. 2 shows the histogram of the L_{11-12} time history over the whole period and, with its bimodal profile, justifies its complex evolution. There we can clearly observe two different traffic states centered at about 60 km/h and 100 km/h, denoting situations of near congestion and free flow, respectively; the congested states involve about 10 % of the overall area.

4 Local models and pattern matching

The problem of modeling a process from observed data has been the object of several disciplines. In the literature dealing with this problem, many different approaches have emerged.

A possible classification of the approaches to modeling is based on structural considerations, that view the approaches as *local* versus *global* methods [6, 7].

Global modeling consists in describing the behavior of the system at hand by means of a single model that covers all the space of possible operating regimes, for example a global linear model or a neural network. On the other hand, local modeling provides a description of the system by combining several models pertaining to different operating regimes, and is suitable for problems where one cannot assume that a unique statistical distribution underlies the system. Each of the local models is obtained by giving full attention to a reduced portion of the space of the possible behaviors, yielding a more accurate description even when simple approximations are used: therefore this approach refers to the concept of breaking up the domain into several small neighboring regions and analysing these separately.

Generally local methods are memory-based algorithms, in the sense that they defer data processing until a prediction is required. A database of observed data is always kept and the estimation is derived from a neighborhood of the query point/time.

Identifying the examples neighboring the query point/time may require a possible large amount of memory and high computational costs. For example, given an evolving database with several thousands or millions of examples, the continuous choice of examples can be burdensome for the system. Nevertheless the evolution of computer hardware may help to overcome these problems.

The method of *Pattern Matching (PM)* or *Pattern Imitation* [8] presented in this paper can be ascribed to local modeling; it avoids the assumptions normally required in the context of time-series modeling (e.g. stationarity and/or Gaussian condition) and is based on the imitation of the past patterns in the data history.

The method tries to identify patterns (intended in a sense that will be defined later) in the past data history, neighboring the current time, and assumes that the current situation will evolve in a similar way.

Clearly the method does not enable forecasts for unusual events or accidents, because generally they lead to few past similar situations.

We now present the element which plays a leading role in our forecasting method, *the reference pattern*.

5 The reference pattern

Let us suppose we have a multivariate time-series $\boldsymbol{L}(k) = \{L_1(k), L_2(k), \ldots, L_p(k)\}$ of p speed time-series defined on a discrete time history indexed by k = 1, 2, …., N.

In our context $\boldsymbol{L}(k)$ is a vector composed of speed data of all the links L_i of Rome GRA at time k; L_i denotes equivalently, for the sake of brevity, the link $L_{i-(i+1)}$ in the clockwise direction.

The pattern is defined as a collection of a subset $s = \{s_1, s_2, \ldots, s_h\}$ of $h \leq p$ links taken from the original vector $\boldsymbol{L}(k)$ and composed of time fragments of observations of length $b_1, b_2, \ldots, b_h$:

$$\boldsymbol{P}(s,b_1, b_2, .., b_h)=\{L_{s1}(1)\ L_{s1}(2)\ ..\ L_{s1}(b_1)\ \vdots\ L_{s2}(1)\ L_{s2}(2)\ ..\ L_{s2}(b_2)\ \vdots\ .....\ \vdots\ L_{sh}(1)\ L_{sh}(2)\ ....\ L_{sh}(b_h)\} \quad (1)$$

The association in (1) of each pattern with the set of structural parameters $\{s, b_1, b_2, .., b_h\}$ is pointed out, but it will be omitted later on. In particular we denote a pattern $\boldsymbol{P}_t(k)$ for our *target link* L_{11-12} at a generic time k as follows:

$$\boldsymbol{P}_t(k) = \{L_{10\text{-}11}(k\text{-}5)\ \ L_{10\text{-}11}(k\text{-}4)\ \ldots\ L_{10\text{-}11}(k)\ \vdots\ L_{12\text{-}13}(k\text{-}5)\ L_{12\text{-}13}(k\text{-}4)\ \ldots\ L_{12\text{-}13}(k)\ \vdots\ L_{13\text{-}14}(k\text{-}5)\ L_{13\text{-}14}(k\text{-}4)\ \ldots\ L_{13\text{-}14}(k)\ \vdots\ L_{14\text{-}15}(k\text{-}5)\ L_{14\text{-}15}(k\text{-}4)\ \ldots\ L_{14\text{-}15}(k)\ \vdots\ L_{11\text{-}12}(k\text{-}7)\ L_{11\text{-}12}(k\text{-}6)\ \ldots\ L_{11\text{-}12}(k)\}, \quad (2)$$

where:

$L_{10–11}(j)$ is the j-th speed value of the immediately preceding upstream link, enclosed between Junction 10 and 11 (see Fig. 1)

$L_{12–13}(j)$ is the j-th speed value of the immediately following downstream link (see again Fig. 1); similar considerations hold for $L_{13–14}(j)$ and $L_{14–15}(j)$.

All the link fragments are lined up to the same final time value k (see e.g. Fig. 3, where L_t represents a target link fragment and L_{t+1} the fragment of the immediately following downstream link, lined up to the same final time and date, etc..); one upstream and 3 downstream links are involved in the pattern, which it is built by considering the actual and the nearest past speed values of the target and of the spatial correlated upstream/downstream links.

Clearly the pattern structure is link-dependent, in the sense that it is to be tailored and calibrated to the target link every time.

A dynamical sequence results, represented on both temporal and spatial scales, whose time depths depend on the time-forecasting horizon.

The structure described in (2), with the time depths reported there (five time steps for associated links and seven time steps for the target link), was obtained by following a trial and error procedure, driven by the tradeoff of pursuing data reduction and obtaining acceptable results in our forecast procedures.

When the time k assumes the meaning of *forecast origin*, the associated pattern will be denoted as the *reference pattern*; our aim is to achieve a forecast for $L_{11–12}(k + n)$, which is n time-steps ahead, and results will be presented for $n=5$ (corresponding to 5×3 min=15 min ahead), $n=10$ (30 min), $n=15$ (45 min), $n=20$ (1 h).

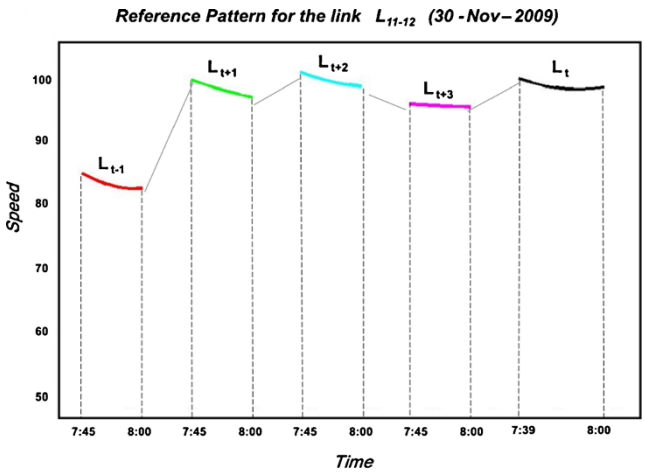


Fig. 3 A reference pattern for the target link $L_t \equiv L_{11–12}$ (forecast origin 30-Nov-2009, 8:00)

Figure 3 presents a true example of a reference pattern relative to the target link $L_{11–12}$, and to an arbitrarily chosen date and time.

6 Detailing the proposed pattern matching method

Three aspects are fundamental in our approach, and we now detail them in sequence.

a) *The bandwidth*

Following the definition of a reference pattern related to a forecast origin k, we have to find all possible past patterns present in the database and compliant with the structure described in (2). In effect our procedure is computationally reduced, because the scanning of the database proceeds in the past within a time frame of ±30 min from the selected time origin k, to simplify the search and reduce the computing time; when the scanning is complete, the set of *candidate patterns* is thus obtained.

The frame value of ±30 min constitutes a *bandwidth* parameter. This parameter decides how wide the pattern neighborhood should be, and clearly influences the results.

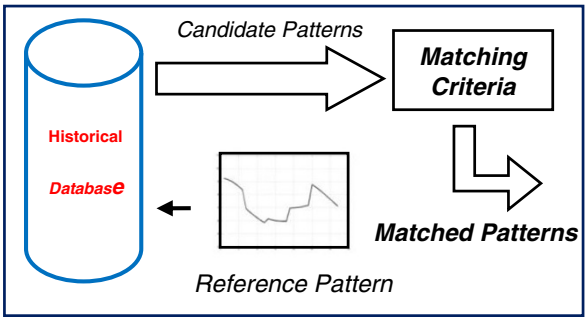
b) *The matching criteria of selection*

The next action consists in an oriented selection among the candidate patterns: patterns are to be near to the reference pattern in the sense of usual Euclidean distance, *and similar in shape*, so as to have fully comparable trends among the reference and the matched profiles.

In our procedure, we used a sequence of two criteria:

1) a classification of the Euclidean point-to-point distances between the reference and each of the candidate patterns is preliminarily made, and only those showing a distance below the *fourth percentile* of the empirical distance distribution enter the next step;
2) to evaluate shape similarity, the *Spearman coefficient* [9] between the reference pattern and each of the patterns surviving the previous selection step is calculated. The Spearman coefficient is a measure of rank correlation, and obeys the property that when two variables X and Y are perfectly monotonically related, its value becomes +1, the largest possible value. This last criterion of selection was based on the decision of choosing only the past patterns showing a Spearman value larger than 0.95.

These two last parameters, the fourth percentile of the distance distribution and a Spearman value of 0.95, constitute the other free parameters of our procedure.



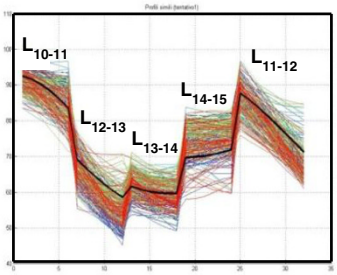


Fig. 4 The selection process for a reference pattern of L_{11-12} with its 231 matched patterns

The final effect is the selection of patterns that present marked resemblance in time evolution; the current situation is then assumed to develop in a way that resembles its matched patterns.

Figure 4 (left panel) summarizes the selection procedure. Figure 4 also presents (right panel) a result pertaining to the reference pattern at an assigned date and time (5 Feb 2009, 08:36); 231 matched patterns that were capable of passing the two-phase selection among all candidates were found in the past and are shown. The black bolded curve represents the reference pattern, the colored lines refer to all the matched patterns.

Therefore the selection procedure helps us to find past analogous situations.

The next developments of our method will discard the contributions of the neighboring links L_{10-11}, L_{12-13} and so on, and our considerations will be concentrated only on the target link L_{11-12}.

c) *The local model*

Looking at the right panel of Fig. 4, a question immediately arises: how can the information conveyed by matched patterns be used to furnish a forecast n steps ahead? Figure 5 (right panel), where the effective time evolutions of all 231 matched patterns of L_{11-12} into their true future are presented, suggests different possible solutions: a simple averaging, referred to the time of forecast (i.e. over the line +15 min., +30 min., etc..), a weighted averaging, and so on.

Any choice based on averaging constrains the forecast speed to lie within the corresponding line ranges, constituting a suitable procedure; but this procedure may generate limitations.

In order to assign more flexibility to the method, that is to assure better *generalization* capabilities, we prefer to define a local model involving the L_{11-12} fragments of all the matched patterns.

Referring for example to the situation presented in Fig. 5, we build up an *autoregressive* model using the usual least squares (LS) procedure, where we properly employ all these fragments (denoted $L_{11-12,\,m}$, left and right panel, with m = 1, 2, …, 231). In detail, in matrix form the solving system for a time horizon of 15 min (that is, five steps ahead) can be expressed as:

$$\underset{(Terms\ in\ left\ panel\ of\ Fig.\ 5)}{\begin{vmatrix} L_{11-12,1}(k_1-7) & L_{11-12,1}(k_1-6) & \ldots\ldots & L_{11-12,1}(k_1) \\ L_{11-12,2}(k_2-7) & L_{11-12,2}(k_2-6) & \ldots\ldots & L_{11-12,2}(k_2) \\ \ldots\ldots & & & \\ L_{11-12,231}(k_{231}-7) & \ldots\ldots & & L_{11-12,231}(k_{231}) \end{vmatrix}} \bullet \underset{}{\begin{vmatrix} a_{1,15} \\ a_{2,15} \\ \cdot \\ a_{8,15} \end{vmatrix}} = \underset{(Terms\ on\ the\ line+15\,min)}{\begin{vmatrix} L_{11-12,1}(k_1+5) \\ L_{11-12,2}(k_2+5) \\ \cdot \\ L_{11-12,231}(k_{231}+5) \end{vmatrix}} \tag{3}$$

Once solved, a set $\{a^*_{i,15}\}$, i = 1, 2, .., 8, of optimal coefficients is obtained that can be applied to the *current* L_{11-12} time fragment (the bolded curve of the left side of Fig. 5, denoted $L_{11-12,\,c}$), to obtain the desired five-step-ahead forecast (that is, 15 min ahead) $L^*_{11-12,\,c}(k+5)$:

$$L^*_{11-12,c}(k+5) = a^*_{1,15}\ L_{11-12,c}(k-7) + a^*_{2,15}\ L_{11-12,c}(k-6) + .. + a^*_{8,15}\ L_{11-12,c}(k) \tag{4}$$

Other time horizons forecasts can be found similarly; only the right hand term of (3) has to be lined up to the times (k + 10) for 30 min ahead, (k + 15) for 45 min ahead, and so on.

Possible variations can be considered in (3), for example, by taking into account the subtraction of a constant value from speed values (mean detrending), or of an average profile, and so on.

The system matrix in (3) is prone to rank deficiency; in fact rows are similar by choice, therefore linear (or quasi-linear) combinations among rows may exist, in this case leading the LS procedure to numerical problems or instabilities.

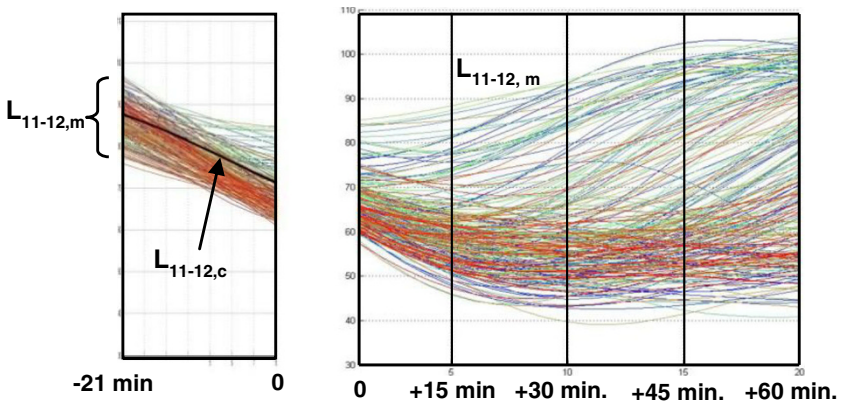


Fig. 5 Time history of all matched patterns (left panel: the 231 time fragments of $L_{11–12}$ and the current one—right panel: evolutions of the 231 profiles into their future until 1 h ahead)

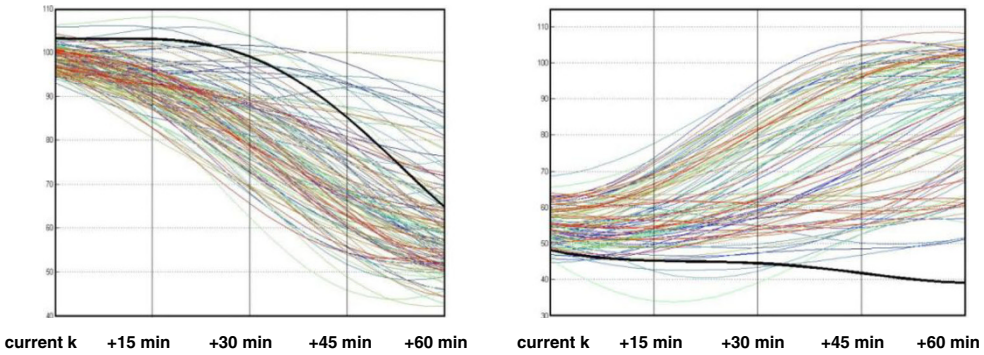


Fig. 6 Two real cases of time evolution of $L_{11–12}$ from a time k

In any case, the LS procedure can take advantage of Ridge Regression [10], which successfully treats situations of regression with collinearity or near-collinearity.

Moreover, complex and misleading situations can occur when the future time evolutions of the reference patterns disagree with those of the matched patterns.

An example is presented in Fig. 6.

It refers to two real cases of an effective time evolution of the target link $L_{11–12}$ (black bolded curve) from a *current time k* into the subsequent 60 min: the left panel shows a time evolution always included within those of its matched patterns (the colored lines), whereas the right panel shows the opposite case.

In the former case, the usual forecasting techniques can furnish suitable values; in the latter (configuring an unusual event) there is a component of unpredictability which affects the results. In this case the PM method produced the better estimation, when compared with other methods like averaging and global linear regression (see below).

Table 1 Forecasting errors for PM modeling of $L_{11–12}$

	Max absolute error (km/h)	Mean absolute error (km/h)	Mean absolute percentage error (%)
15 min	0.01720	0.00086	0.001
30 min	0.63602	0.03237	0.037
45 min	5.37979	0.28758	0.337
60 min	23.31256	1.18310	1.395

7 Forecast results for the link $L_{11–12}$

An extensive analysis of the link $L_{11–12}$ was done, considering samples referred to daytime situations (from 6 a.m. to 11 p.m.). In other words, we randomly extracted about 20,000 reference patterns (4 % of those available) from the recorded traces of $L_{11–12}$, $L_{10–11}$, $L_{12–13}$ and so on and submitted them to a process of forecast estimation for $L_{11–12}$ relative to 15, 30, 45, 60 min ahead.

Having fixed (as a measure of quality) the requirement of finding a minimum amount of 12 matched patterns to continue

Table 2 Forecasting errors for global regression modeling of $L_{11–12}$

	Max absolute error (km/h)	Mean absolute error (km/h)	Mean absolute percentage error (%)
15 min	0.54989	0.14489	0.16292
45 min	18.28674	6.01728	6.57748

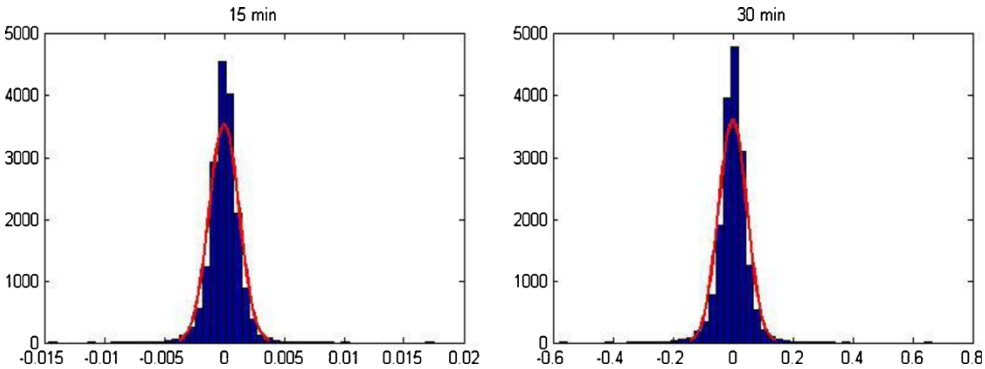


Fig. 7 Histograms of the forecasting errors of PM modeling (15 and 30 min ahead)

every estimation process, only 17,359 of them (about 88 %) reached the end of the procedure.

In effect we enabled a three-steep procedure that progressively loosened the selection limits (the fourth percentile on Euclidean distances and a Spearman coefficient of 0.95) in order to obtain at least 12 matched patterns: if the first trial on the current reference pattern failed to furnish 12 matched patterns at the highest limits, we loosened the fixed values slightly and tried again to obtain at least 12 matched patterns, and so on until the third trial.

In this way only 12 % of the available cases did not support the forecasting procedure; 13,991 cases enabled a first-trial treatment, 3,035 required a second e 333 a third with a progressive loosening of the criteria.

We will see the immediate consequences of this procedure in terms of the precision. Table 1 summarizes the results in terms of the following errors:

$$\begin{aligned}
&\text{Max Absolute Error (km/h)} = \max \left| SP_{11-12,i} - SP_{11-12,\ i} \right| \\
&\text{Mean Absolute Error (km/h)} = \sum \left| SP_{11-12,\ i} - SP_{11-12,\ i} \right| / 17359 \\
&\text{Mean Absolute Percentage Error}(\%) = 100 \cdot \left(\sum \left| SP_{11-12,\ i} - SP_{11-12,\ i} \right| / SP_{11-12,\ i} \right) / 17359,
\end{aligned}$$

where $SP_{11-12,\ i}$ and $\overline{SP}_{11-12,\ i}$ are, respectively, the effective and the forecast speed, with the index i ranging between 1 and 17359, the sample size.

Table 1 presents the results related to PM forecasts, showing good levels of precision, including when the forecasting horizon reaches as far as 60 min; in this last case only, a maximum absolute error of 23.31 km/h is reported, while other results in the same column maintain acceptable values of error.

Both types of mean absolute errors (second and third column of Table 1), on the other hand, show encouraging results in all cases.

Table 3 Relations among max absolute error and number of matched patterns (#MP)

Max absolute error (km/h)	$12 \leq$ #MP < 20	$20 \leq$ #MP < 30	#MP ≥ 30
15 min	0.01720	0.00811	0.00759
30 min	0.63602	0.32034	0.29799
45 min	5.37979	2.98723	2.77643
60 min	23.31256	12.86972	12.01528

A comparison with a global linear regression model is presented in Table 2 with regard to the forecasts 15 and 45 min ahead. As usual the global regression modeling was performed on matched and unmatched patterns, in accordance with the structure described in (2); as unique differences, the same regression order of six was assumed for all of the links and a preliminary mean value detrending on all components of $\boldsymbol{P}_t$ (k) was conducted.

Interestingly, we can note a systematic lowering of the forecast quality when compared to PM modeling.

The histograms of the forecasting errors of PM modeling show Gaussian-like profiles, with small positive excess kurtoses (see Fig. 7. Left: 15 min. ahead; right: 30 min. ahead).

Moreover, Table 3 shows the dependence of the maximum absolute error on the cardinality of the matched pattern set (#MP), distributed into three possible ranges; for example, the last value of the first row means that among all the 15-min-ahead forecasts with a population of matched patterns larger than 30, the maximum absolute error was found to be 0.00759.

We can note that situations where #MP is higher show lower levels of maximum absolute error among all the time

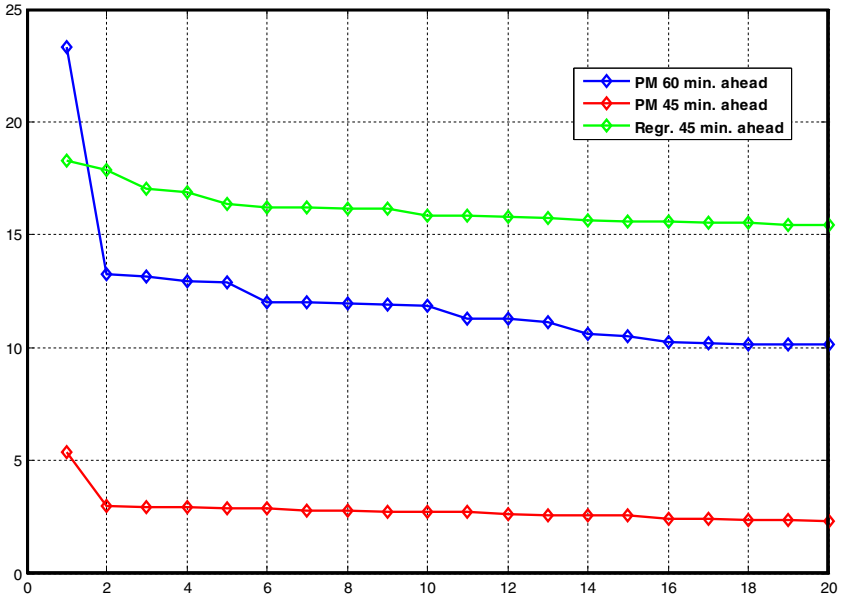


Fig. 8 The 20 maximum absolute prediction errors (km/h) for PM and global regression modeling in descending order of magnitude

horizons considered, so we are able to conclude that a cautious consideration of the parameter #MP, in terms of the assumed value during the selection procedure, is to be taken into account.

Table 3 justifies our choice of finding at least 12 matched patterns to continue the PM forecast procedure and obtain satisfying errors (apart from the value of 23.31 for the 60-min-ahead forecast (see Table 3) which probably pertains to an unusual situation); our choice is derived from a tradeoff, and thus limits the intractable situations to only 12 % of the total cases.

Finally, Fig. 8 shows the 20 maximum absolute prediction errors of two situations of PM modeling compared to global regression modeling, in descending order of magnitude.

A comparison of the green profile (pertaining to 45-min-ahead forecasts of global regression modeling) versus the red one (45-min-ahead forecasts of PM modeling) confirms on a point-to-point scale the superior performance of PM modeling. In effect even the 60-min-ahead forecasts of PM modeling (blue profile) perform better, apart from the first well known situation of an error of 23.31 km/h.

8 Conclusions

Present-day computer technology enables effective storage of large database at reduced costs, and computational techniques and tools to gain access to large amount of data are now possible; in particular procedures for finding useful patterns in available data are to hand and enable the development of PM modeling, which is based on searching similar patterns in the past data history, under the hypothesis that the current situation will evolve in an analogous way.

The application of these concepts to speed values of the link L_{11-12} (and of its neighbors) of Rome GRA, taken from a data collection of more than 3 years obtained by OCTO Telematics PVS, following the method presented in Section 6, produced superior results when compared to global regression model predictions, in both average and point-to-point comparisons, as reported in Section 7.

Possible drawbacks of the presented approach are the lack of guarantee of obtaining a forecast (because in some cases the number of past similar examples may be insufficient), or the computational costs due to pattern tracing when scanning large databases. Our experience reported a PM forecast failure of 12 % with regard to the former and an average computing times of 20 s/forecast on a desktop computer with a typical configuration (a Pentium dual-core CPU @ 3.16 GHz, 8GB of RAM) with regard to the latter. On the other hand the encouraging results obtained in terms of forecast quality suggest that this method should be applied on a wider basis for an extensive verification of its performances, and efforts in this direction have been undertaken.

References

1. Vlahogianni EI, Golias GC, Karlaftis MG (2004) Short-term traffic forecasting: overview of objectives and methods. Transp Rev 24(5): 533–577
2. Yang Z, Gong B, Lin C (2009) Travel time estimate based on floating car. Second Int Conf Intell Comput Technol Autom 3: 868–871

3. Turksma S (2000) The various uses of floating car data. Tenth Intl. Conference on Road Transport Information and Control, Publ. no. 472, London
4. Liberto C, Ragona R, Valenti G (2010) Traffic Prediction in Metropolitan Freeways, Seventh International Conference on Traffic & Transportation Studies (ICTTS 2010: 846–860), Kunming, China
5. de Fabritiis C, Ragona R, Valenti G (2008) Traffic Estimation and Prediction Based on Real Time Floating Car Data. IEEE Conf. on Intelligent Transportation Systems (ITSC 2008: 197–203), Beijing, China
6. Bontempi G (1999) Local Learning Techniques for Modeling, Prediction and Control. Universitè Libre de Bruxelles, Belgium www.ulb.ac.be/di/map/gbonte/ftp/thesis_partial.pdf
7. Loader C (1999) Local regression and likelihood. Springer, Berlin
8. Motnikar BS, Pisanski T, Cepar D (1996) Time-series forecasting by pattern imitation. OR Spectr 18:43–49
9. Maritz JS (1981) Distribution-free statistical methods. Chapman & Hall, London
10. Hansen PC (1998) Rank-deficient and discrete ill-posed problems. SIAM, Philadelphia