Data Recovery Algorithms Based on the Nearest Neighbor Method for Predicting Traffic Flows

Arthur Zhigalov [a], Irina Bolodurina [a] and Lubov Zabrodina [a]

[a] *Orenburg State University, Prospekt Pobedy, 13, Orenburg, 460018, Russia*

Abstract

This research examines the problem of short-term forecasting of traffic flows in conditions of incomplete information. For the task of forecasting, a mathematical model has been compiled, which makes it possible to compare the main characteristics of the traffic flow with the value of the forecast - the speed of the flow. Within the framework of this work, various forecasting algorithms have been investigated, which make it possible to analyze the influence of the number of recorded observations and similar records, as well as allowing modifications to increase the forecast accuracy. Also, a data recovery approach has been developed to eliminate noisy and missing data. Experiments of short-term forecasting were carried out on an open dataset from Yandex and confirmed the importance of the stage of preprocessing and analysis of the effectiveness of recovery algorithms. After applying the proposed approach to data processing and recovery, the predicted value turned out to be more accurate on average by 2 km / h. The results of the study showed that the combined model with weighing showed a more accurate result compared to simple models, the average error was 9.81 km / h. Among the possible directions for further research, one can single out the search for several combined methods that have equivalent algorithms using neural networks.

Keywords [1]

data recovery algorithm, short-term forecasting, modeling of transport networks, data preprocessing, nearest neighbors method

1. Introduction

With the rapid growth of the number of vehicles and the progress of urbanization, the number of traffic congestions in large cities increases every year, resulting in low efficiency of transport networks and wasted time, wasted fuel, and excessive air pollution. Also, intelligent traffic management is a key issue, which is also an important tool to guide scientific decision making in traffic management. Traffic forecasting is a challenging task in intelligent traffic management.

The goal is to anticipate changes over time in the number of vehicles at observation points (eg intersections or stations). The time interval is usually set to 5, 15, or 30 minutes. The forecasting task is associated with the processing of current data and the construction of a forecasting model for traffic flows.

Because traffic data has sharp nonlinearities resulting from transitions from free flow to failure, and then to congested traffic, predicting traffic flow is challenging due to rapidly changing traffic conditions. Also, processing the data before building the model can significantly affect the result of the prediction algorithm. This is due to many factors, such as speed noise, missed travel times, traffic accidents (data outliers), etc. Such phenomena affect the accuracy of the constructed forecast. In this regard, the purpose of this work is to develop a data recovery algorithm for solving the problem of short-term forecasting of traffic flow.

YRID-2020: International Workshop on Data Mining and Knowledge Engineering, October 15-16, 2020, Stavropol, Russia
EMAIL: leroy137.artur@gmail.com (Arthur Zhigalov); prmat@mail.osu.ru (Irina Bolodurina); zabrodina97@inbox.ru (Lubov Zabrodina);
ORCID: 0000-0003-3208-1629 (Arthur Zhigalov); 0000-0003-0096-2587 (Irina Bolodurina); 0000-0003-2752-7198 (Lubov Zabrodina);

2. Related works

The problem of predicting traffic flows has been studied in several studies both from the point of view of small road sections and for large-scale road networks.

In the framework of the research conducted by Y. Han and F. Moutarde [1], the change in global congestion configurations in networks (spatial congestion configurations of the entire network) is considered. As part of the development of a unified data mining system, the authors presented a description of the most typical time patterns of network traffic status and a long-term forecast of large-scale traffic dynamics. The experiments have shown the effectiveness of clustering and forecasting based on the results of tensor factorization.

In [2], A. Ladino et al. proposed a real-time forecasting technique for dynamic travel time based on merging individual forecasts obtained from clustered time series. The resulting prediction error along the horizon of future values within the cross-validation algorithm has the smallest value if there is a priori knowledge about the cluster.

The authors of the study [3] developed a deep learning model for predicting traffic flows to solve the problem of non-linearity, recovery, and congestion. In some cases, when the traffic mode changes abruptly, deep learning provides fairly accurate short-term traffic forecasts. However, the results obtained are not generalized to global large-scale networks.

The application of deep learning to short-term real-time traffic speed prediction is also discussed in the study [4]. The experiments have shown that the performance of the speed prediction model increases when the missing values are taken into account and restored in the multi-view approach.

Statistical problems in processing a data stream with different quality are presented in more detail in [5]. The authors of the study analyzed algorithms for converting missing or fuzzy data, estimated speed, and made a forecast of driving time on motorway networks. The results obtained can be used in conjunction with other short-term forecasting algorithms.

The study [6] presents a system for predicting traffic congestion using the K-nearest neighbor algorithm. However, traffic congestion modeling is based on the example of a specific road, based on its spatial and temporal data, and requires generalization to large-scale networks.

In the framework of [7], an improved K-nearest neighbor approach is presented in conjunction with the method of ranking local minima of distances, which in most cases exceeds competing algorithms in terms of forecast accuracy.

A modification of the k-nearest neighbor method is also presented in the study [8] and is based on a three-level nonparametric regression algorithm. According to experiments, the proposed algorithm has improved predictive ability, as well as significantly increased the accuracy and performance of short-term forecasting of traffic flows in real time.

The authors of the study [9] conducted a comparative analysis of the average and weighted nonparametric regression model of k-means, and also evaluated the reliability of the predicted result. The experiments confirmed the feasibility of using the methods in short-term traffic flow forecasting and showed a fairly high accuracy.

In [10], attention is paid to the problem of predicting the state of short-term motion based on spatio-temporal characteristics in conditions with damaged or missing data. Critical sections of the road had the strongest impact on the accuracy of the forecast, and therefore, the authors determined the need to use a spatial-temporal correlation algorithm.

Thus, the research has shown that data processing before building the model can significantly affect the accuracy of the prediction algorithm. In this regard, the task of research and analysis of data recovery algorithms for solving the problem of short-term forecasting is relevant.

3. Generalized statement of the problem of forecasting traffic flows

The transport network is represented as a graph $G = (W, U)$, $arcs\ u \in U$ representing road segments and a set of vertexes representing intersections. The following values can be used as the predicted traffic flows value on the edges of the road graph: flow density, average time of vehicle passing along the road section. In this paper, the average speed of a vehicle passing a road section will be used for conducting a numerical experiment. Let’s assume that data processing has been completed and there is an observation database in the form of sets of traffic flows values for all sections of the road network in the form:

$$X = X_i , X_i = \left(v_{k_i}\left(t_{k_i}\right), t_{k_i}\right), \tag{1}$$

where k_i is number of observations on the i-th edge; t_{k_i} are observed time points; $v_{k_i} \in R^+$ is speed at a time t_{k_i}. Taking into account the entered designations, a formal statement of the problem of getting a forecast for a given parameter of the transport network's traffic flows: Using the road graph $G = (W, U)$ and current data on the traffic flows value in the form (1), we need to calculate the estimation of the traffic flows parameters for all edges of the transport network:

$$\begin{cases} v_i(t^* + d) = f(X) \\ v_i(t^* + d) \in R^+ \\ \forall\, t^* \geq t_{k_i} \end{cases} , \tag{2}$$

where d is forecast horizon; t^* is current time; $f(X)$ is function for predicting the traffic flows value vi on edge ui at time $t^* + d$. When building a model for short-term forecasting of traffic flow, it is necessary to:

- define a model that displays the values of known observations of the traffic flows value in the forecast value;
- determine the parameters of the algorithm of the mathematical model of traffic flows based on real data from the observation database.

4. A prediction model based on information about nearest neighbors using linear regression

When building a model for predicting traffic flows, appropriate databases are needed, which often have data gaps and noise in practice. The missing information can be filled in when performing local forecasting for the road network graph in sections with gaps. The idea is to use information about traffic flow on neighboring edges to predict this value based on similar incidents in the past. The function of restoring the value of the transport flow, depending on the values on neighboring edges, can be represented as:

$$\mathrm{v}^{\mathrm{r}}_{\mathrm{d,t}} = f(\mathrm{v}^{\mathrm{r_1}}_{\mathrm{d,t}}, \mathrm{v}^{\mathrm{r_2}}_{\mathrm{d,t}}, \ldots, \mathrm{v}^{\mathrm{r_N}}_{\mathrm{d,t}}), \tag{3}$$

where $v^r_{d,t}$ is speed on the road r in time f of the day d; f is some recovery function. We will use linear regression as a function. We will look for the type of function f as a linear dependence on the speeds on adjacent roads:

$$f\left(\mathrm{v}^{\mathrm{r_1}}_{\mathrm{d,t}}, \mathrm{v}^{\mathrm{r_2}}_{\mathrm{d,t}}, \ldots, \mathrm{v}^{\mathrm{r_N}}_{\mathrm{d,t}}\right) = \mathrm{b_1 v}^{\mathrm{r_1}}_{\mathrm{d,t}} + \mathrm{b_2 v}^{\mathrm{r_2}}_{\mathrm{d,t}} + \ldots + \mathrm{b_N\, v}^{\mathrm{r_N}}_{\mathrm{d,t}} + \varepsilon = \mathrm{X\,N\,i} = \sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{b_i v}^{\mathrm{r_i}}_{\mathrm{d,t}} + \varepsilon, \tag{4}$$

where b_i are linear regression coefficients; ε is the approximation error function f.

The goal is to find the coefficients b_i. These coefficients can be found using the least squares method. Let there are a history of observing the road, as well as a vector of points $T = (t_1, t_2, \ldots, t_N)$ where t_i indicates a point in time in the past. Vector of transport flow values at time points ti on edge for which the forecast will be built we denote by V^r. For each edge there is a certain number of adjacent edges N. Let there are a matrix of the form:

$$W = \begin{pmatrix} v^{r_1}_{t_1} & \cdots & v^{r_N}_{t_1} \\ \cdots & \cdots & \cdots \\ v^{r_1}_{t_N} & \cdots & v^{r_N}_{t_N} \end{pmatrix}, \tag{5}$$

where $v^{r_i}_{t_j}$ is the amount of traffic flow on the road r_i at time t_j . Then in matrix form the dependency looks like this:

$$V^r = bW + E, \tag{6}$$

where b is vector of linear regression coefficients $b = (b_1, \ldots, b_N)$; E is the vector of forecast errors $E = (\varepsilon_1, \ldots, \varepsilon_N)$.

It is necessary to select the coefficients $b = (b_1, \ldots, b_N)$, so that the sum of the squares of deviations is minimal:

$$\sum_{i=1}^{N} \epsilon_i^2 = \sum_{i=1}^{N} \left(v_i^r - f_i(x)\right)^2 \rightarrow max \tag{7}$$

where $f_i(x)$ is linear combination of traffic flow values and coefficient vector at a time i.

To find the coefficient vector we transform the matrix form (7) to a system of equations, which in matrix form has the form

$$(W^T W)^b = W^T V^r. \tag{8}$$

The solution of this system of equations allows us to obtain coefficients for the linear dependence (4).

To restore using the least squares method, select sections of the road graph that meet the following conditions:

1. training takes place at those time intervals $T = t_1, t_2, \ldots, t_N$ for which all values of traffic flows values on adjacent edges are known at given time points;

2. trecovery occurs at those points where the traffic flows value is unknown at the time t_{i+1} but on neighboring edges all traffic flows values at the predicted time are present in the observation history.

5. Forecasting models

When predicting the amount of traffic flows on the road r at some point in time t we can use observations of this section. Then the base model I is represented in following form:

$$v_{d,t}^r = average\left(v_{d-1,t-\delta}^r, \ldots, v_{d-1,t-1}^r, \ldots, v_{d-1,t+1}^r, \ldots, v_{d-1,t+\delta}^r, v_{d-2,t-\delta}^r, \ldots\right), \tag{9}$$

where $v_{d,t}^r$ is speed on the road r at time t of day d.

To take into account the situation with a small number of observations we will take the average value of the traffic flows value for all observations on the transport network and shift it in its direction. Then the base model II is represented as follows:

$$v_{d,t}^r = \frac{avgValue * avgCount + globalAverage * K}{avgCount + K}, \tag{10}$$

where $avgV$ alue are traffic flows values obtained from the formula (9); $avgCount$ is number of observations used in the base model I; $globalAverage$ is the value of the traffic flows at a given time t for all roads and all previous days; K is the free parameter of the model.

To account for differences in traffic flow characteristics, it is possible take into account observations that are most similar to the current behavior of the traffic flows at a given time. For this purpose we enter the similarity metric expressed by the formula:

$$w(d_i, d) = \left(\frac{1}{|R|} \sum_{r \in R} \sum_{t \in T} \frac{\left|v_{d,t}^r - v_{d_i,t}^r\right|}{|T|}\right)^{-1}, \tag{11}$$

where $v_{d,t}^r$ is speed on the road r at time t of day d; d_i - day number i in the history of observations; R - many roads in the transport network; T is a set of moments in time for which both dimensions are known $v_{d,t}^r$ and $v_{d_i,t}^r$.

Then the base model II is modified so that it takes into account the found similarity values. This model is designated as the basic model III.

There are examples of constructing an ensemble of traffic flows forecasting models. Let's build combined models. Let there be a vector of prediction models $F = (f_1, f_2, \ldots, f_k)$ consisting of k weak models. Then the constructed model for predicting traffic flows will have the form (combined model I):

$$v_{d,t}^r = \frac{1}{k}\left(f_1(X_1) + f_2(X_2) + \cdots + f_k(X_k)\right). \tag{12}$$

For unequal models, we will take into account their coefficients. In this method, the ensemble of models A is defined as the weighted sum of k weak models (combined model II):

$$A = \alpha_1 f_1(X) + \alpha_2 f_2(X) + \cdots + \alpha_k f_k(X). \tag{13}$$

where α_i is coefficient with which the weak model is included in the ensemble.

To find linear regression coefficients $\alpha_1, \alpha_2, \dots, \alpha_k$ predictions $Y_1, Y_2, \dots, Y_k$, are made for one of the previous days, V^* is the forecast vector of the combined model. Then the Least Absolute Deviation problem is solved:

$$min\|\alpha_1 Y_1 + \alpha_2 Y_2 + \cdots + \alpha_k Y_k\|. \qquad (14)$$

Model (14) is designated as a combined model II.

6. Simulation results

To build the model, we used an open source of Internet data from Yandex. The source data consists of two parts: a description of the city's street network and observation data for these streets. The observations cover 31 days for the Moscow road graph. Intersections in Moscow correspond to the vertices of the graph, and street segments correspond to arcs (a two-way street corresponds to two multidirectional arcs).

The data are heterogeneous, for some roads a large number of speed observations are known, for others, on the contrary, less. Figure 1 shows the distribution of the number of observations in sections of the road network graph:

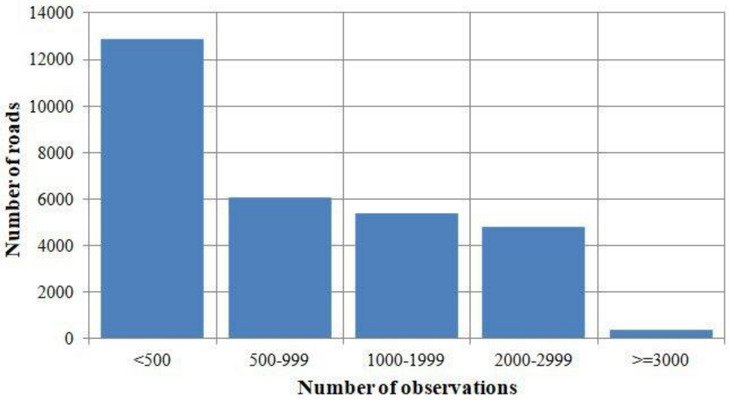

Figure 1: Distribution of observational data by road

The metric of the average deviation of the forecast values from the observed measurements can serve as an estimation measure (15).

$$Q = \frac{1}{n}\sum |v^* - \hat{v}|, \qquad (15)$$

where n is total number of plots; $\hat{v}$ is predicted speed on the section; v^* is observed speed on the site.

The results of evaluating algorithms using the metric (15) are shown in figure 2 before data recovery.

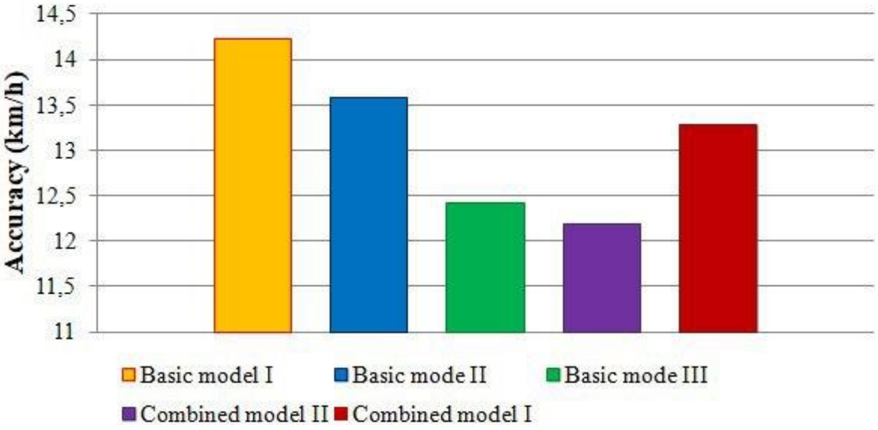

Figure 2: The results of the evaluation of the algorithms for metric (15) to recover the data

The results of evaluating algorithms using the metric (15) are shown in figure 3 after data recovery.

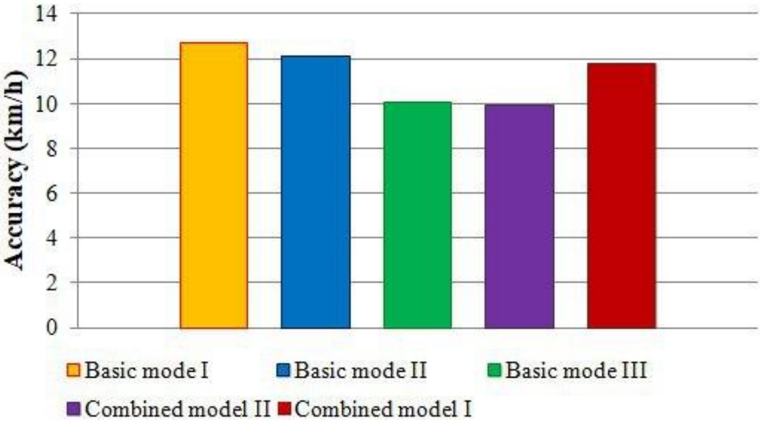

Figure 3: The results of the evaluation of the algorithms for metric (16) after data recovery

Based on the results obtained, the following conclusions can be made:

1. The approach with data recovery using information about the nearest neighbors allowed increasing the number of observations, the number of data increased by about 1.5 times. Forecasts based on the new data were more accurate.
2. The basic model III makes a significant contribution to the combined model. In this case, the coefficient α_3 = 0, 92. The other models account for only 0,08.
3. The model based on average observations of the road makes an average error of 12.31 km/h, the best model (combined model II) gives an average error of 9.81 km/h, i.e. the improvement is on average 3 km/h.
4. Combining prediction models by learning from different samples using sampling (combined model I) does not improve, since the models are not equivalent.

7. Conclusion

In this study, we consider the problem of short-term forecasting in conditions of incomplete information, the purpose of which is to develop an adaptive mathematical model for predicting traffic flows with the required accuracy. A comparative analysis and selection of the optimal TP model are carried out, as well as algorithms for predicting the flow rate of the transport network are studied.

Methods for building traffic forecasts based on previous measurements have shown that processing and clearing raw data is an important step before predicting traffic flow. So according to the processed data, the results were more accurate by an average of 2.3 km/h.

The combined model with weighing showed a more accurate result compared to simple models, with an average error of 9.81 km/h. When combining models, the average approach does not improve the model, since the presented algorithms are not equivalent.

Among the possible directions for further research, we can highlight the search for several combined methods with equivalent algorithms using neural networks.

Acknowledgments

The study was carried out with the financial support of the Russian Foundation for Basic Research in the framework of scientific projects No. 20-57-53019, 20-07-01065, as well as a grant from the President of the Russian Federation for state support of leading scientific schools of the Russian Federation (NSh-2502.2020.9).

References

[1] A. Ladino, Y. Alain, C. Canudas and H. Fourati. "A real time forecasting tool for dynamic travel time from clustered time series" Transportation research. Part C, Emerging Technologies 80 (2017): 216-238.

[2] H. Yufei, M. Fabien. "Analysis of Large-scale Traffic Dynamics using Non-negative Tensor Factorization" ITS World Congress (2012).

[3] L. Zhang, Q. Liu, W. Yang, N. Wei, D. Dong. "An Improved K-nearest Neighbor Model for Short-term Traffic Flow Prediction" Procedia - Social and Behavioral Sciences 96 (2013):653-662.

[4] N. Polson, V. Sokolov. "Deep learning for short-term traffic flow prediction" Transportation Research Part C: Emerging Technologies 79 (2017):1-17.

[5] P. Bickel, C. Chen, J. Kwon, J. Rice, Measuring Traffic. Statist Statistical Science 4 (2007):581-597.

[6] R. Rishab, M. Shreyas, P. Rajashree and T. Rohith. "Traffic Congestion Prediction System using K-Nearest Neighbour Algorithm" International Research Journal of Engineering and Technology 7 (2020):2628- 2630.

[7] S. Gutha. "A deep learning approach to real-time short-term traffic speed prediction with spatialtemporal feature prediction" Electronic Theses and Dissertations (2019).

[8] X. Pang, C. Wang, G. Huang. "A Short-Term Traffic Flow Forecasting Method Based on a Three-Layer K-Nearest Neighbor Non-Parametric Regression Algorithm" Journal of Transportation Technologies 6 (2016): 200- 206.

[9] Y. Gang, W. Yunpeng, Y. Haiyang, R. Yilong, X. Jindong "ShortTerm Traffic State Prediction Based on the Spatiotemporal Features of Critical Road" Sections Sensors 18 (2018):2287.

[10] Z. Zheng, D. Su. "Short-term traffic volume forecasting: A k-nearest neighbor approach enhanced by constrained linearly sewing principle component algorithm" Transportation Research Part C: Emerging Technologies 43 (2014): 143-157.