Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Short-term Forecast for Average Speed of Road Section based on Floating Car Data with Support Vector Regression

Short-term Forecast for Average Speed of Road Section based on Floating Car Data with Support Vector Regression

Y H Zou, Y H Feng, J C Huang, G Q Cheng, T Wang and Z Liu

College of Systems Engineering, National University of Defence Technology, Changsha, Hunan Province, China

fengyanghe@nudt.edu.cn

Abstract. The recent intelligent transportation system has yielded remarkable progress on traffic data collection, resource allocation and intelligent programming. However, development on traffic real-time data processing and prediction still remains limited. In pursuit of real-time prediction on road section average speed, we introduced a prediction method, which mines GIS floating car data with support vector regression algorithm. The result indicated our proposed method was superior in comparison with other commonly used algorithms including linear regression, artificial neural network, Bayesian regression and ridge regression. Besides, the quick convergence and well fitting confirmed the plausibility of our method in this domain.

1. Introduction

The quick development of cities has posed great demands in building intelligent systems, which can provide us with better services in living and well manage the limited resources. Among them, the Intelligent Transportation System (ITS), targeting at optimizing the urban transportation resources, has attracted more and more attention ranging from industry, academia to government agencies [1]. Integrating information, communications, abundant sensors and other plausible technologies, ITS makes attempt to guide automatic data analysis, efficient resource allocation and smart service assignment with people, road networks and kinds of vehicles involved in [2]. Though ITS has been widely employed in city planning for a few decades, a recent significant change in ITS is revealed as the explosion of volume of data in transportations. Potential causes for this phenomenon are the rapid development of sensors for data collections and various social platforms in sharing the transportation information. A significant conclusion can be drawn that the availability of massive data is leading a revolution in the domain transferring the conventional technology-driven transportation systems to data-driven ones. To exploit disciplinary data from multi-sources and advance the performance and efficiencies in transportation systems, machine learning techniques are introduced to address related problems. A significant advantage of employment of machine learning in dealing with transportation big data lies in mining task-beneficial patterns, uncovering predictable disciplines and optimizing the process of decision-making [3]. Instead of collecting expert experience, machine learning is capable of capturing plausible rules and enhancing the generalization of systems in the era of big data [4].

With respect to ITS, there are several issues to address, such as traffic status supervision, automatically path planning, transportation vehicles allocation and etc. Among them, the prediction of real-time traffic status is regarded as one of the most critical topics, which bridges the gap between data collection and ITS design. Specifically speaking, potential information after forecasting the average speed of road sections in a real-time way can well assist the traffic control and ensure the free movement in traffic if decision-makers accommodate properly. Meanwhile, the intuitive traffic information allow more plausible strategies in choosing routines for travelers. In this paper, we would focus our attention on machine learning algorithm's employment in solving the short-term prediction of the average speed of road sections. After collecting the GIS floating car data generated from urban roads and pruning the redundant information of data, we performed calculations of the average speed of floating cars on some road section. Considering the robustness and universal generality, a classical algorithm called support vector regression (SVR) was adopted in constructing the prediction model.

The remainders of this paper are arranged as follows. Some related works are summarized in Section 2, in which we discuss methodologies of average speed prediction in road networks. Then, we detail the methods and techniques in collecting and processing the GIS floating car data as well as calculations on the average speed of cars on some road in Section 3. The SVR prediction model is introduced in Section 4 and we elaborate the principles and essence of such model. In Section 5, we collect the transportation data in Shenzhen in the form of time window to verify the plausibility and effectiveness of SVR's employment in average velocity real-time prediction.

2. Related work

The field of function estimation and regression prediction has witnessed significant progress after the adaptation of SVM method which was proposed by Vapnik [5]. A novel and efficient pairing nv-support vector regression (pair-v-SVR) algorithm was introduced by Pei-Yi Hao, which successfully combines the advantages of twin support vector regression (TSVR) and classical ε-SVR algorithms [6]. In the task of identification of nonlinear systems in RKHS spaces, SVR method has proven its effectiveness and shown the excellence in fitting residue and superiority of the regularization network in reducing computation time [7]. Considering the complex characteristics of traffic system such as nonlinearity, time-varying, randomness and uncertainty, Thomas Epelbaum et al. utilized deep learning models to capture regression disciplines in time series data [8]. These algorithms are designed to address real-time average speed prediction of road section based on Floating Car Data (FCD).

The former works with respect to traffic flow characteristics focus mainly on relationships between three traffic flow characteristics as traffic flow, average speed and density. Greenshields firstly developeda linear model describing the relationship between velocity and density in 1993 [9]. The model, which was adopted by the U.S. Department of Transportation, assumed that the flow rate is linear with the velocity before it reaches to the maximum and then is illustrated in curve relation when the flow rate is between the maximum and the coordinate origin point [10]. Natalia Isaenko et al. designed an integrative framework which was capable of recognizing and selecting suitable method for traffic forecasting with individual FCD [11].

The main challenge of analyzing the FCD comes from the geographic data error. Map-matching is an important step in the information processing that can minimize errors effectively. Jia-Ching Ying et al. developed a novel modularity-based map-matching algorithm called Urban Map-Matching (UrbMatch) utilizing urban GPS trajectories [12]. The method called spatial and temporal conditional random field (ST-CRF) has better performance and robustness when facing the low-frequency trajectory data(e.g., one GPS point for every 1–2 minutes) [13]. Mahdi Hashemi put forward a weight based map-matching algorithm, which can be applied in real-time complex urban road networks [14]. In order to calculate the average speed of road sections, Yanace J. L. et al. performed least square method on the instantaneous speed of the floating car [15]. However, from the perspective of statistics, it is difficult to control the estimation error.

3. Floating vehicle data analysis and processing

3.1. Floating vehicle sample collection

Floating car, also called probe vehicle, refers to a vehicle equipped with a GPS positioning system and a wireless communication device. They can collect their own traffic data, such as speed, transmission time, latitude and longitude, direction, passenger status, the distance between the last point and other information on roads. The collection of FCD is a sampling survey process on the road traffic network.

The number of floating car samples should be determined before the collection of floating car traffic information. The quantity of floating cars in the road network should be big enough to ensure the accuracy of traffic flow parameter estimation. At the same time, however, the relationships between road coverage, the information update cycle and the number of floating vehicles should be furthr discussed. The former works [16, 17] provide ways to determine the number of floating car samples, with multiple factors considered. The relationship between the number of floating car samples, traffic parameters, road coverage, and information refresh cycle can be approximated as follows:

$$\beta = \left(1 - \mathrm{e}^{-Nk}\right)^2 + Nke^{-2Nk} \tag{1}$$

Where N is the number of floating car samples; β is the road coverage; K is the floating car density:

$$K = \frac{\overline{v}t_0}{l} \tag{2}$$

Where, $\overline{v}$ is the average speed of traffic flow; t is the information update cycle and l is the length of the link.

Given the length of the link, the information refresh cycle, and the traffic flow rate, the relationship between coverage and sample size can be obtained.

3.2. Floating car data processing

After accessing the data of the floating car, it is generally the first step to match the map and prune some anomaly data. Map-matching is quite crucial for floating car information processing. The intrinsic ideology is to compare the vehicle locating trajectory obtained by the data acquisition system with the road information in the electronic map database, and then map the vehicle to the most probable position on the map by some effective algorithm [18]. In this paper, we made use of the ST-Matching algorithm to embed the information of spatial connection in the road network. The Figure 1 illustrates the variation after matching crossroads data point with ST-matching algorithm.

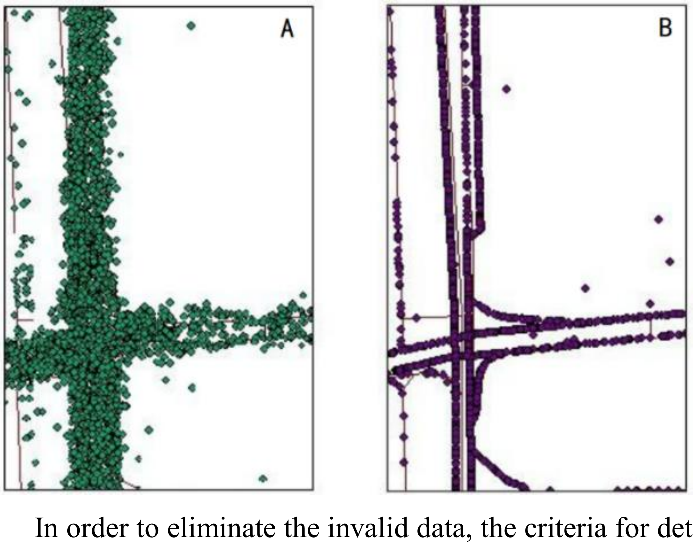


Figure 1. Map-matching of floating car data points. (A) The raw location data on the map. (B) Map matched data points are on the most probable position of the road line.

In order to eliminate the invalid data, the criteria for determining invalid data should be determined. Basing on the requirements of predictive modeling and the actual situation, this paper identifies the following screening criteria:

- The wrong data.

Repeated records. The data in one car ID which have different positions but the same time. The data of two neighbor points with the same car ID which have different positions but the distance is 0.

- The invalid data

Considering the taxis without passengers tend to find the business at a low speed or parking, their data cannot reflect the real situation. If a same ID car in the same latitude and longitude for a long time (2 *min* above), we regard these vehicles in abnormal driving conditions, which data should be removed. If two neighbor floating car points have a too-long (120 seconds above) time interval, it cannot truly reflect the real average speed. Due to the speed limit on urban roads, the calculated car speed should satisfy the constraint. The data whose average speed is more than 80km/h should be removed [19].

4. Support vector regression model

Two typical SVR algorithms, ε-SVR and ν-SVR, are commonly used in regression. ν-SVR is superior to ε-SVR in terms of regression accuracy and maneuverability [20]. In order to avoid the artificial error in the choice of coefficients, we chooses ν-SVR as the prediction model. One of advantages in SVR is the powerful feature transformation trick which maps the input into high dimensional Hillbert space. The nonlinearity of the mapping empowers better representational capability of characterizing the potential decision boundary. Besides, the model is embedded with the tolerant mechanism which makes the model more stable even in the presence of response noise. We provide brief description of ν-SVR formulation following Vapnik as follows [21]:

Given the training dataset $T = \{(x_i, y_i)|i = 1,2,\ldots,N\}$ where the x is the attribute vector while the y is the response variable, the objective of SVR is to capture the decision function in the form as $f(x) = w^T * \phi(x) + b$. It can be noticed the decision function is accompanied with the potential feature transformation which can be learned in a parameter tuning process.

$$\min_{w,b,\epsilon^{(*)},\epsilon} \frac{1}{2}\|w\|^2 + C * (\nu\varepsilon + \frac{1}{N}\,\Sigma_{i=1}^{N}(\xi_i^* + \xi_i)) \tag{3}$$

s.t.

$$|(w^T * \phi(x) + b) - y_i| \leq \xi_i^* + \xi_i, i = 1,2,..,N$$
$$\xi_i^{(*)} \geq 0, i = 1,2,..,N$$
$$\xi_i \geq 0, i = 1,2,..,N$$

The hyper parameter C controls the extent of tolerance to the noise and the norm of parameter, w can be viewed as the penalty term which generally leads better generalization [22]. The count of support vector is controlled by the parameter ν, in range (0, 1]. Note that training samples falling inside the ε-tube have zero loss, and samples outside the ε-insensitive zone are linearly penalized using the slack variables $\xi_i^*, \xi_i \geq$ 0, i = 1,…,N. Usually, problem(3) is usually solved in its Lagrange's dual form, (see [23] for details):

$$\max_{a^*} \frac{1}{2}\sum_{i,j=1}^{l}(a_i^* - a_i)(a_j^* - a_j)K(x_i, x_j) + \sum_{i=1}^{l} y_i(a_i^* - a_i) \tag{4}$$

s.t.

$$\sum_{i=1}^{l}(a_i^* - a_i) = 0$$
$$0 \leq a_i^* \leq \frac{C}{N}, \quad i = 1, \ldots, N$$
$$\sum_{i=1}^{l}(a_i^* - a_i) \leq C\nu$$

Where the kernel $K(x_i, x_j) = (\Phi(x) \cdot \Phi(x'))$.

Solution of the dual formulation (4) yields optimal values of parameters $\overline{a}^{(*)} = (\overline{a}_1, \overline{a}_1^*, \ldots, \overline{a}_N, \overline{a}_N^*)$ that can be used to construct the optimal SVR function: $f(x) = \sum_{i=1}^{N}(a_i^* - a_i)(x_i, x_j) + \overline{b}$ In this optimal solution, training samples with non-zero coefficients are the support vectors (SVs), corresponding to data points at the boundary or outside ε-insensitive zone.

The non-linear kernel $K(x_i, x_j)$can be computed with the dot product (x_i, x_j) in (7) to extend linear SVR to a non-linear setting. This kernel $K(x_i, x_j) = (\Phi(x) \cdot \Phi(x'))$ implicitly captures the non-linear mapping of the data x → φ(x).

Some commonly used kernel functions includePolynomial kernel, Radial Basis Function (RBF), and Sigmoid kernel [24].

The prediction index of the model is the average speed of the target road section, as well as its upstream and downstream sections, over the past period of time. To construct the input matrix, we assume the average speed of road l on time t is $\overline{V}_l(t)$ and the input data format is:

$$\begin{bmatrix} \overline{V}_l(t+1-p) & \overline{V}_l(t+2-p) & \cdots & \overline{V}_l(t) \\ \overline{V}_{l+1}(t+1-p) & \overline{V}_l(t+2-p) & \cdots & \overline{V}_{l+1}(t) \\ \vdots & \vdots & & \vdots \\ \overline{V}_{l+k}(t+1-p) & \overline{V}_{l+k}(t+2-p) & \cdots & \overline{V}_{l+k}(t) \end{bmatrix}$$

Where p is the number of periods (1 minute) that needs to be traced back when predicting the average speed. The data is divided into training set and testing set to carry out related experiments.

5. Experiment

In this section, we will employ the taxi float car data of Futian District, Shenzhen on October 1st, 2017 from 17:00 to 19:00 in the experiment. After the data preprocessing, we compared SVR predicted results with results using other commonly used regression prediction algorithms.

5.1. Data Processing

We select the average speed data of the unidirectional section of Hongli Road from east to west (from Caitian Road to Xintian Road). The parameters of the target section and its related sections of the attributes shown in Table 1.

Table 1. Parameters of road sections.

	Name	Num.	Range of Rd.	Length(km)
Upstream	Hongli Rd.	5	Jintian Rd.-Caitian Rd.	0.490
sections	Jintian Rd.	4	Fuzhong Rd.- Hongli Rd.	0.428
Target section	Hongli Rd.	1	Yitian Rd.- Jintian Rd.	0.702
Downstream	Hongli Rd.	2	Yitian Rd.-Xinzhou Rd.	0.705
sections	Yitian Rd.	3	Fuzhong Rd.- Hongli Rd.	0.434

The target road section has two upstream sections (No.4, No.5) and two downstream sections (No.2, No.3), which have direct influence to the target downstream. Cars on the target section all come from upstream sections, and drive to downstream sections.

We first run a map-matching process and then the scattered points are projected onto the corresponding line segments of the road based on the driving direction of the vehicle. The speed and driving direction of the vehicle can be calculated by the change of the latitude and longitude.

According to the calculation formula mentioned in Section 3.1, the road coverage rate is calculated as follows:

Table 2. Road Coverage Rate Calculation.

Rd. Number	1	2	3	4	5
Length(km)	0.702	5	0.434	0.428	0.490
Lowest speed(km/h)	14.27973	14.27973	4.083369	9.015247	6.585276

k	1.695125	1.687911	0.784057	1.755305	1.119945
The minimum of samples N	6	10	4	7	5
road coverage β	0.999923	1	0.920915	0.999991	0.992693

We define the length of a time interval as 5 minutes and the road coverage rate of each road segment are calculated under the minimum sample quantity and the lowest average speed. The minimum number of samples is collected from the time interval when there are least FCD points on the road. As can be seen in the Table 2, the road coverage rate can reach more than 90% with the minimum of average speed and the minimum of samples. In this case, the sample size in any time interval can satisfy the minimal accuracy requirement.

5.2. Prediction model

According to the requirements of SVR model training and the characteristics of road sections, the average speed of the road sections needs to be prepared in the following format.

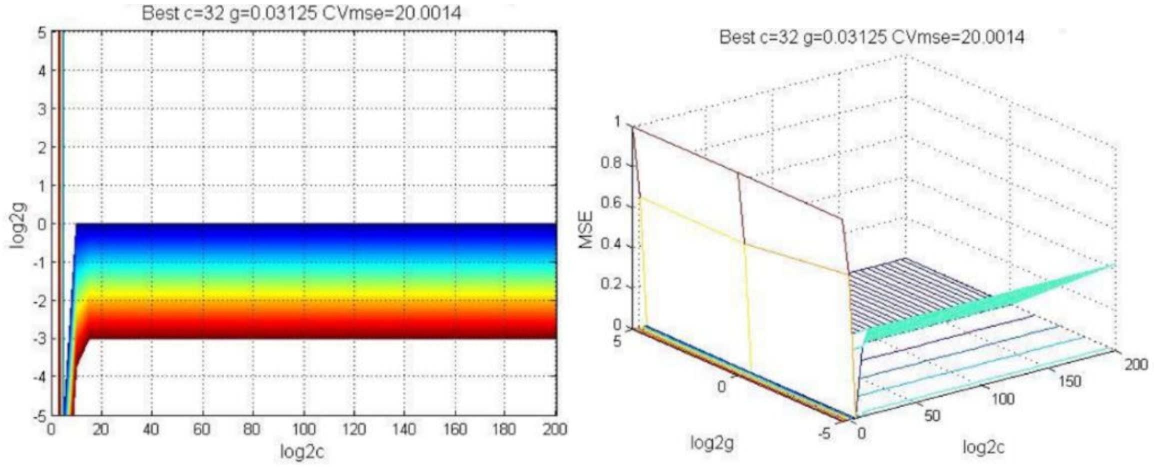
Table 3. Data Format of the Input and Output.

							Prediction Index					
Rd. Number	1			1		2	3	4			5	
Sample	V_{t+1}	V_t	V_{t-1}	…	V_{t-p+1}	…	…	…	V_t	V_{t-1}	…	V_{t-p+1}
	Output						Input					

Where p is the backtracking coefficient.

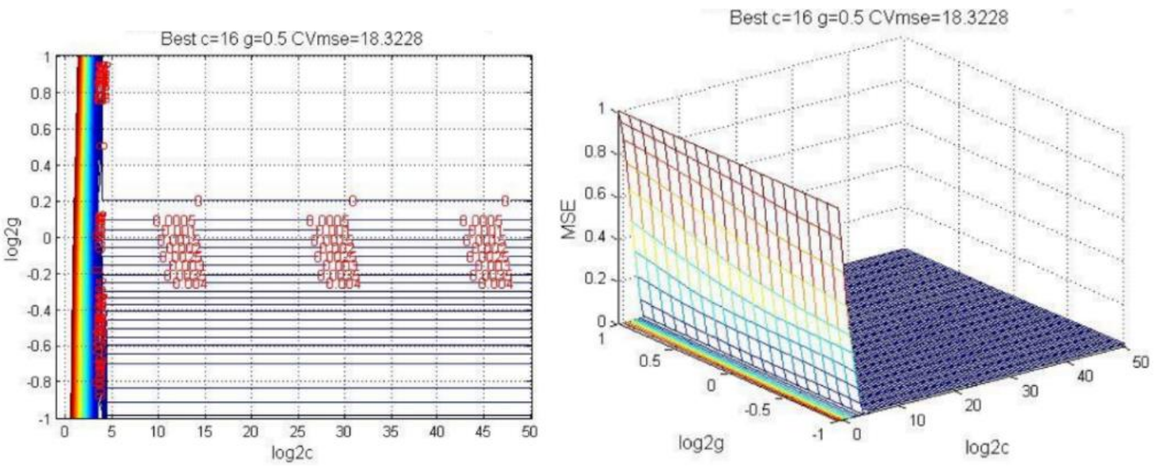
The calculated data of the various road sections are represented in a matrix according to the format in Table 4, and the data is processed by setting p = 3 as an example first, a matrix of [111 × 16] is formed. Then, we perform normalization on these data. Finally, we partitioned the first 80 lines as training set and the last 31 lines as testing set.

In the process of model training, the parameter C of the RBF kernel ranges in the interval [-5,200], with a step of 5, and the parameter g of the kernel ranges in the interval [-5,5], step size is 1. We performed grid search on these parameters to obtain the optimal parameters. The result is on Figure 2:



(a) (b)

Figure 2. Rough Optimum Calculation. (a) Result in contour line overhead view. (b) Result in 3D view plot



(a) (b)

Figure 3. Accurate Optimum Calculation. (a) Result in contour line overhead view. (b) Result in 3D view plot.

Basing on the result of the rough selection of the parameters, we narrow the search range. We set the parameter C, in the interval [-1, 50], with a step of 0.5, and the parameter g, in the interval [-1,1], with a step length of 0.1, take an accurate optimum calculation. The result is on Figure 3.

According to the optimization results, the best *C* parameter is 16, *g* is 0.5. The mean square error at this time is 18.3228. The forecast results are shown in Figure 4

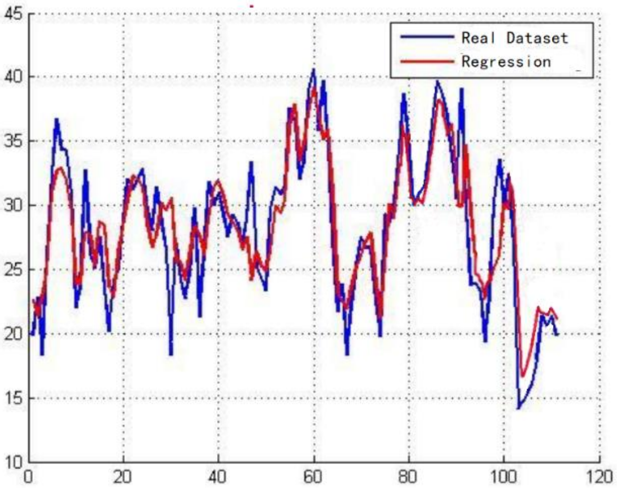


Figure 4. SVM regression curve on the data set. Training set (before 80) and the testing set (after 81).

To better evaluate the performance of the model, we compared the predicting results of SVR, ANN, linear regression, Bayesian ridge and ridge regression on the same dataset. The results shows on Figure 5, The MSE of each algorithm results on the testing set shows on Table 4.

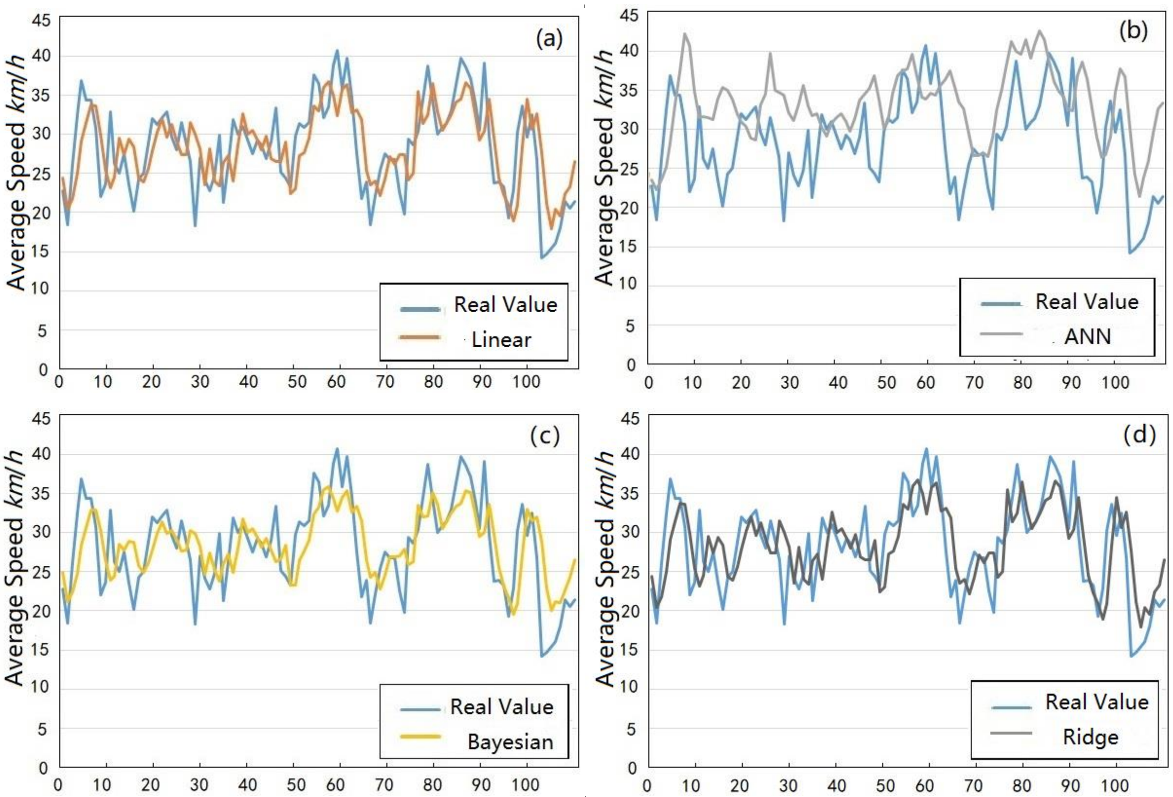


Figure 5. Comparison of the fitting degree on the dataset. (a) Linear regression. (b)ANN. (c) Bayesian ridge. (d) Ridge regression.

Table 4. The MSE of each algorithm results on the testing set.

Algorithm	SVR	Linear model	ANN	Bayesian ridge	Ridge
MSE	18.3228	22.1685	71.2724	26.0315	22.1801

6. Discussion and conclusion

In this work we demonstrated the superior performance of SVR on the average speed of the road section regression forecast. Instead of considering the data from a single road section, we aggregate the data of related road sections in the high dimensional space, aiming at achieving quick convergence. Our results suggest that SVR can handle this type of input well, with a smaller mean square error than the other algorithms.

The condition considered in this work is a typical scene on the urban road network. SVR can deal with the forecast problems under longer backtracking time and the complex road network conditions. We can find more regular patterns in the daily long-term data of each road section in our forecast index system, when accessing more data. In future work, we would address these limitations, adapting our forecast method to accommodate the daily long-term and road network domains.

Acknowledgment

Thanks to Dr. Qi Wang's help in academic wrting. Yehan Zou acknowledges Support by National Natural Science Foundation of China (Grant No: 71701205,71701206).

References

[1] Shivendra, A. and S. Kumarswamy, Intelligent Transportation System. Computer-Aided Civil and Infrastructure Engineering, 2016. 18(3): p. 173-183.

[2] Souza, A.M.D., et al. Real-time path planning to prevent traffic jam through an intelligent transportation system. in Computers and Communication. 2016.
[3] Jha, V., Study of Machine Learning Methods in Intelligent Transportation Systems. 2015.
[4] Wang, Q., et al., Addressing Complexities of Machine Learning in Big Data: Principles, Trends and Challenges from Systematical Perspectives. 2017.
[5] Vapnik, V., S.E. Golowich, and A. Smola, Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing. Advances in Neural Information Processing Systems, 1996. 9: p. 281--287.
[6] Hao, P.Y., M.S. Lin, and L.B. Tsai. A New Support Vector Machine with Fuzzy Hyper-Plane and Its Application to Evaluate Credit Risk. in Eighth International Conference on Intelligent Systems Design and Applications. 2008.
[7] Sayehi, I., et al. A comparative study of two kernel methods: Support Vector Regression (SVR) and Regularization Network (RN) and application to a thermal process PT326. in International Conference on Sciences and Techniques of Automatic Control and Computer Engineering. 2016.
[8] Epelbaum, T., et al., Deep Learning applied to Road Traffic Speed forecasting. 2017.
[9] Arnott, R., E. Inci, and J. Rowse, Downtown curbside parking capacity. Journal of Urban Economics, 2015. 86: p. 83-97.
[10] Transportation, U.S.D.o., Chapter 3. Numerical Modeling. 2008.
[11] Isaenko, N., C. Colombaroni, and G. Fusco. Traffic dynamics estimation by using raw floating car data. in IEEE International Conference on MODELS and Technologies for Intelligent Transportation Systems. 2017.
[12] Ying, J.C., et al., Spatial-temporal Mining for Urban Map-Matching.
[13] Liu, X., et al., A ST-CRF Map-Matching Method for Low-Frequency Floating Car Data. IEEE Transactions on Intelligent Transportation Systems, 2017. 18(5): p. 1241-1254.
[14] Hashemi, M. and H.A. Karimi, A weight-based map-matching algorithm for vehicle navigation in complex urban networks. Journal of Intelligent Transportation Systems, 2016: p. 00-00.
[15] 15.Ygnace, J.L., et al., Travel Time Estimation on the San Francisco Bay Area Network Using Cellular Phones as Probes. 2000.
[16] Zhi, T.U., et al., Study on the Route Coverage and the Update Cycle of Transportation Information Based on the Minimum Samples of Floating Car. China Railway Science, 2006. 27(5): p. 127-131.
[17] Chen, Y., Study on the density of floating car based on the route coverage. Journal of Shandong University of Technology, 2006.
[18] Jensen, C.S. and N. Tradišauskas, Map Matching. Encyclopedia of Database Systems, 2016: p. 1692-1696.
[19] Jahnke, M., et al., Identifying Origin/Destination Hotspots in Floating Car Data for Visual Analysis of Traveling Behavior. 2017: Springer International Publishing.
[20] Miguel, M. and A.T. Flor, Spectroscopic Determination of Aboveground Biomass in Grasslands Using Spectral Transformations, Support Vector Machine and Partial Least Squares Regression. Sensors, 2013. 13(8): p. 10027.
[21] Vapnik, V., Estimation of Dependences Based on Empirical Data. 2006: Springer New York.
[22] Wang, Q., et al., A Novel Ensemble Method for Imbalanced Data Learning: Bagging of Extrapolation-SMOTE SVM. Computational Intelligence & Neuroscience, 2017. 2017(3): p. 1827016.
[23] Cherkassky, V., Predictive Learning, Knowledge Discovery and Philosophy of Science. 2012: Springer Berlin Heidelberg. 209-233.
[24] Smola, A. and V. Vapnik, Support vector regression machines. 1997.