RESEARCH ARTICLE

A Kriging based spatiotemporal approach for traffic volume data imputation

Hongtai Yang[1]☯, Jianjiang Yang[2]☯, Lee D. Han[3]*, Xiaohan Liu[1], Li Pu[4], Shih-miao Chin[5], Ho-ling Hwang[5]

1 National United Engineering Laboratory of Integrated and Intelligent Transportation, School of Transportation and Logistics, Southwest Jiaotong University, Hi-Tech Industrial Development Zone, Chengdu, Sichuan, China, **2** Model Risk Management, Bank of America, Charlotte, NC, United States of America, **3** Department of Civil & Environmental Engineering, the University of Tennessee, Knoxville, TN, United States of America, **4** School of Architecture and Design, Southwest Jiaotong University, Hi-Tech Industrial Development Zone, Chengdu, Sichuan, China, **5** Center for Transportation Analysis, Oak Ridge National Laboratory, Cherahala Boulevard, Knoxville, TN, United States of America

☯ These authors contributed equally to this work.
* lhan@utk.edu

Abstract

Along with the rapid development of Intelligent Transportation Systems, traffic data collection technologies have progressed fast. The emergence of innovative data collection technologies such as remote traffic microwave sensor, Bluetooth sensor, GPS-based floating car method, and automated license plate recognition, has significantly increased the variety and volume of traffic data. Despite the development of these technologies, the missing data issue is still a problem that poses great challenge for data based applications such as traffic forecasting, real-time incident detection, dynamic route guidance, and massive evacuation optimization. A thorough literature review suggests most current imputation models either focus on the temporal nature of the traffic data and fail to consider the spatial information of neighboring locations or assume the data follow a certain distribution. These two issues reduce the imputation accuracy and limit the use of the corresponding imputation methods respectively. As a result, this paper presents a Kriging based data imputation approach that is able to fully utilize the spatiotemporal correlation in the traffic data and that does not assume the data follow any distribution. A set of scenarios with different missing rates are used to evaluate the performance of the proposed method. The performance of the proposed method was compared with that of two other widely used methods, historical average and K-nearest neighborhood. Comparison results indicate that the proposed method has the highest imputation accuracy and is more flexible compared to other methods.

OPEN ACCESS

Citation: Yang H, Yang J, Han LD, Liu X, Pu L, Chin S-m, et al. (2018) A Kriging based spatiotemporal approach for traffic volume data imputation. PLoS ONE 13(4): e0195957. https://doi.org/10.1371/journal.pone.0195957

Editor: Xiaolei Ma, Beihang University, CHINA

Received: August 17, 2017

Accepted: April 3, 2018

Published: April 17, 2018

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Along with the rapid development of Intelligent Transportation Systems (ITS), traffic data collection technologies have been evolving dramatically. [1, 2]. On the one hand, the emergence of innovative data collection technologies such as remote traffic microwave sensor (RTMS), Bluetooth sensor, and GPS-based floating car method have made traffic data collection much easier than before. [3–5].

Despite the development of technologies, the missing data problem still exists. Missing data could be due to various reasons such as malfunction of sensors and loss of communication. The Mobility Monitoring Program of the Texas Transportation Institute (TTI) reported that after screening erroneous data, the complete rate of collected data can be anywhere between 16% and 93% with a median value of 67% [6]. Williams and Hoel [7] reported that the data missing rate collected by Georgia's statewide advanced traffic management system was 10% or higher. The data missing rate of the freeway Performance Measurement System (PeMS) in Los Angeles was found to be 15% [8]. Chandra and Al-Deek [9] reported a 15% missing rate of data collected by loop detectors on I-4 in Orlando, Florida. An empirical study showed that the average missing rate of data collected by Georgia NaviGAtor system at GA 400 was between 4% and 14% [10]. In Beijing, China, the average missing rate of the daily traffic volume data was about 10% (4% due to malfunction of detectors and 6% due to other reasons) with the missing rate of data collected by some detectors as high as 25% [11].

The missing data issue has posed great challenges for data based applications such as traffic forecasting, incident detection, route guidance, and massive evacuation optimization. Therefore, a lot of efforts need to be made to impute the missing data.

Most current imputation techniques could estimate a single value for the missing data point. These techniques include heuristic imputation, prediction imputation, and statistical learning imputation etc. The heuristic imputation methods fill the missing data point by averaging data of the same time period on neighboring days or averaging data of neighboring time periods of the same day. These methods are based on the assumption that traffic characteristics are similar at the same time period of different days or the fluctuations of traffic data are low during short time period [12]. Another group of heuristic methods are called pattern-similar imputation methods which search for the most similar traffic data series from historical data and use it to estimate missing data points [13]. These heuristic methods make good use of the similarity and periodicity of traffic data. However, the local variation and unexpected changes of traffic pattern could result in high imputation inaccuracy [14, 15]. To address this issue, two advanced methods, Bayesian Principal Component Analysis (BPCA) and Probabilistic Principal Component Analysis (PPCA), were recently proposed by Qu et al. [11, 16]. Researchers first show that traffic flow data follow Gaussian distribution and that principal component analysis (PCA) can be used to retrieve the features of traffic flow. Then, a robust PCA is used to filter out the abnormal traffic flow data that disturb the imputation process. The difference between BPCA and PPCA is that BPCA is slower than PPCA but yields similar results. BPCA is usually carried out first on a small sample to determine the important parameters. Then, the imputation tasks are performed by PPCA with those parameters.

Prediction method is also an important way to impute data. Regression method is a classic example. Al-Deek et al. [9] compared the feasibility and imputation accuracy of three regression models, multiple regression, time series, and pairwise regression. They found that quadratic model performed better because of its ability to capture nonlinear relationships among variables. To impute missing traffic data during holidays, Liu et al. introduced a new procedure using non-parametric regression, the K-nearest neighborhood (KNN) method, estimate missing values for different types of highways on holidays [17]. Other regression models that have been used for imputation include ARIMA [18], support vector regression [19], exponential smoothing [18], neural network [20], hidden Markov Model [21] and so on [12]. However, these prediction methods can only use data before the missing data point and ignore the data after the missing data, which means they cannot take full advantage of the data set for imputation.

Statistical learning imputation methods assume that data are missing at random. Specifically, the missing data are considered as realizations of random variables characterized by a certain probability distribution function. Antonio et al. proposed an incremental approach theoretically motivated by the Statistical Learning Theory of Vapnik, and provided a new paradigm for missing data imputation [22]. Ma et al. employed copula theory to build a connection between the correlation function and the marginal distribution function of traffic flow, and proved effectiveness of the method to impute missing data in large-scale transportation networks [23].

Most of the methods mentioned above only use temporal information for imputation, while spatial information is not well used. As traffic flows from upstream to downstream, the traffic stream characteristics at a certain location are usually closely related to those at neighboring locations. The incorporation of surrounding traffic information has been proved to be useful to improve traffic prediction accuracy [24–26]. Literature review results show that Markov chain Monte Carlo (MCMC) [10] and PPCA [11] are two representative methods that use both temporal and spatial information. However, both MCMC and PPCA methods assume a probability distribution model of the data [12]. This assumption limits the use of these methods since when the data does not follow a specific distribution these methods may generate inaccurate imputation results. As a result, this paper proposed an alternative method, a Kriging based method, that does not assume the data follow any probability distribution and that can fully use both temporal and spatial information, to impute data.

The rest of the paper is organized as follows. Section 2 describes the proposed Kriging based imputation approach and other benchmark models that are used for comparison. Study location and data are described in Section 3. A brief description of data missing patterns and missing ratios are presented in Section 4. Section 5 compares imputation accuracy of proposed approach with benchmark models, historical average and KNN. Concluding remarks are given in section 6.

Methodology

Kriging based spatiotemporal imputation approach

Background about Kriging. Kriging originated in the mining industry in the early 1950's as a means of improving ore reserve estimation and has been used as synonym for geo-statistical interpolation for many decades. Traditionally, the Kriging method only deals with spatial variables. Consider a set of spatial data $z\ (\mu_i)$ of an attribute z at location U_i, $i = 1, 2, 3, ..,n$, where U is a vector of spatial coordinates $\mu_i = (x_i, y_i)$. The task of data imputation is to estimate missing values of z at a set of m locations. Generally speaking, Kriging is just optimal interpolation method based on regression using observed surrounding data points, weighted according to covariance values. Compared with other methods, the Kriging method has following advantages: 1) It can reduce the effect of data clustering by assigning data points within a cluster less weight; 2) It can produce a measure for possible estimation error (Kriging variance), along with the estimation of the missing values [27].

Kriging based spatiotemporal imputation. Traffic stream characteristics change over time and space. Traffic volume at a location is not only correlated with the traffic volumes at upstream and downstream locations but also correlated with volumes of the previous and next time step [28]. Thus, time dimension needs to be considered in the Kriging model to better estimate the missing data.

In the example of traffic volume, each data point is referenced by its temporal timestamp t_i, and spatial location $\mu_\alpha = (x_\alpha, y_\alpha)$. Different from the aforementioned traditional spatial Kriging models, the coordinates are simplified as $\mu_\alpha = x_\alpha$ as roads can be seen as a longitudinal system with only one spatial dimension, in which x_α is the mile marker.

In the space-time framework, traffic volume is formulated as $Q\ (\mu_\alpha, t_i); \alpha = 1, 2, \ldots, n; i = 1, 2, \ldots, m$. Similar to the spatial models, the covariance is defined as the variance of the mean squared difference between data separated by a given spatial and temporal lag (h_s, t_s):

$$C(h_s, h_t) = E[(z(u_\alpha, t_\alpha) - z(u_\alpha + h_s, t_\alpha + h_t))^2] \qquad (1)$$

To be consistent with the common practice in spatial statistics, experiment semivariogram is computed as half of covariance:

$$\hat{\gamma}_{s,t}(h_s, h_t) = \frac{1}{2}E\left[(z(u_\alpha, t_\alpha) - z(u_\alpha + h_s, t_\alpha + h_t))^2\right] \qquad (2)$$

In the ordinary space-time Kriging system, the missing value $Q^*(\mu, t)$ can be estimated as weighted average of values of surrounding locations:

$$Q^*(u, t) = \sum \lambda_{\alpha,i}(u, t) Q(u_\alpha, t_i) \ with \sum \lambda_\alpha(u, t) = 1 \qquad (3)$$

The weights $\lambda_{\alpha,\ i}\ (\mu_\alpha, t_i)$ assigned to each neighboring data point are calculated by minimizing the prediction variance:

$$\sigma^2(u, t) = Var[Q^*(u, t) - Q(u, t)], \qquad (4)$$

while maintaining unbiasedness of the estimated value $Q^*(\mu, t)$.

As the calculation of covariance is based on both spatial and temporal distance between data points, the spatial and temporal correlations of traffic volumes are well considered and utilized in the model. In this study, the Gaussian variogram method is used to approximate empirical variogram in the proposed spatiotemporal imputation method. It should be noted that the temporal and spatial properties of data are not similar, which makes it difficult for the variogram to capture the temporal and spatial variability. To address this issue, the very straightforward solution is to regard time dimension as the third orthogonal dimension and to extend traditional 2-dimensional Kriging to a 3-dimensional Kriging. In addition, the temporal dimension has to be rescaled to align with the spatial directions. All the works mentioned above are implemented using R studio and related packages.

Benchmark imputation methods for comparison

To evaluate the performance of the Kriging based spatiotemporal approach, the results were compared with those of two classical imputation models, historical average and KNN.

Historical average (same time and weekdays and same stations). The historical average model is a widely used prediction model [29]. A missing data point is estimated by averaging data points of the same location at the same time of the day on the same day of the week. To be more robust to extreme values, the historical median can be used too.

K-nearest neighborhood. Because the data were recorded every 30 seconds by sensors, 2880 data points were collected every day (*2880* = 24h * 60min/h * 60s/min / 30s). In order to implement KNN method [30], the traffic volume data needs to be reformatted as a $2880\times(s_* d)$ matrix), where *s* is the number of stations and *d* is total number of days. After the transformation, the data collected at a given station on a specific day is considered as a column of the matrix.

For the column with missing values, the Euclidean distances between this column and other columns were calculated to find *k* nearest neighbors. Finally, the weights for the *k* nearest neighbors were derived and the estimation of missing values were the weighted averages of k nearest neighbors [31].

Evaluation criteria

Mean absolute deviation (MAD) and root mean squared error (RMSE) were used to compare imputation results of proposed approach with benchmark methods. Suppose there were n missing data points in the test dataset with V_{act}^{i} as ground truth for i^{th} missing data point and V_{est}^{i} as the estimated value for the missing data point. The two measures could be calculated as follows:

$$MAD = \frac{\sum_{i}^{n} |V_{act}^{i} - V_{est}^{i}|}{n} \tag{5}$$

$$RMSE = \sqrt{\frac{\sum_{i}^{n} (V_{act}^{i} - V_{est}^{i})^{2}}{n}} \tag{6}$$

Data source and study locations

Smart Way [32], a key program of Tennessee's intelligent transportation system, uses solar-powered nonintrusive RTMS to collect real-time highway traffic information (including volume, speed and occupancy). The collected data are sent to the traffic management center. The data used in this study are collected by these RTMS radars installed along interstates across Tennessee. Vehicle presence, traffic volume, speed, and occupancy per lane are recorded every 30 seconds by these sensors [33].

To better identify data missing patterns, a long period of 33 days of data (from April 29 to May 31, 2013) are collected for six RTMS stations along Ellington Parkway in Nashville, Tennessee [34]. The detailed description of RTMS stations is presented in Table 1 and their locations are given in Fig 1. As the data are collected every 30 seconds, a total of 570,240 (2880×33×6) data points would be obtained if no data were missing.

Different from previous studies, imputation was performed on the raw data in this study instead of aggregated data to prevent information loss during the aggregation process. The data description and missing rates are shown in Table 2. Numbers in parenthesis indicate corresponding standard errors. The average count means the average number of vehicles that were recorded by sensors over 30 seconds.

Data missing rates

To understand RTMS radars' performance, a boxplot of missing rates by station and day of the week is shown in Fig 2. It shows that the performance of a station varies across days and the performance of different stations on the same day also varies significantly. Station 115 usually has the lowest data missing rates with only a few exceptions. In contrast, station 117 has the highest data missing rates across the week.

Table 1. Description of RTMS stations.

Station	Direction	Location	Lanes	Mile marker
115	Northbound	Ellington Parkway @I24	2	10.6
117	Northbound	Ellington Parkway @Cleveland	2	11.4
119	Northbound	Ellington Parkway @Granada	2	11.8
121	Northbound	Ellington Parkway @Douglas Ave	2	12.2
123	Northbound	Ellington Parkway @South of Trinity	2	12.4
124	Northbound	Ellington Parkway @Trinity	2	13.0

https://doi.org/10.1371/journal.pone.0195957.t001

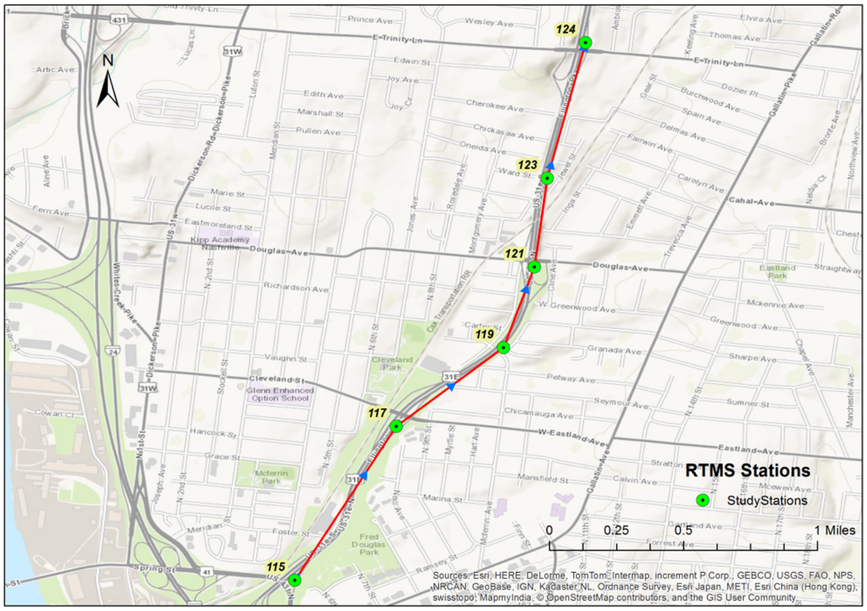

N
124
123
121
119
117
115
31E
31W
E-Trinity-Ln
Douglas-Ave
W-Eastland-Ave
Eastland-Ave
Gallatin-Ave
Dickerson-Pike
Cleveland-St
Cleveland Park
RTMS Stations
StudyStations
0
0.25
0.5
1 Miles
Sources: Esri, HERE, DeLorme, TomTom, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), swisstopo, MapmyIndia, © OpenStreetMap contributors, and the GIS User Community

Fig 1. RTMS stations for this study.

https://doi.org/10.1371/journal.pone.0195957.g001

Evaluation of imputation performance

Experiment design of data missing scenarios

A complete data set is preferred to train the proposed and benchmark models and to evaluate their performance. A close look at the data reveals that the data collected by station 119 on May 17, 2013 has a low data missing rate, 0.03% (only one data point is missing), and thus is used in this study. Another reason for choosing station 119 is that there are both upstream stations and downstream stations, which means there are both upstream and downstream information available.

To compare the imputation performance, imputation methods are tested based on simulated scenarios with different data missing rates. The missing rates are set to be different percentiles of the actual missing rates for all stations during the 33 days of the study. Also, the

Table 2. Data description.

Station number	Lane	Average speed (mi/h)	Average count	Missing rate
115	1	47.41 (22.67)	5 (4)	3.19%
	2	36.63 (31.47)	2 (3)	
117	1	38.46 (33.65)	3 (4)	33.83%
	2	25.94 (31.75)	2 (4)	
119	1	57.85 (21.71)	5 (4)	14.27%
	2	37.77 (27.38)	3 (5)	
121	1	41.95 (17.56)	5 (4)	15.28%
	2	40.47 (24.18)	4 (4)	
123	1	43.41 (26.52)	4 (5)	18.77%
	2	43.22 (26.63)	4 (5)	
124	1	40.06 (18.46)	4 (3)	17.28%
	2	35.26 (21.38)	3 (4)	

https://doi.org/10.1371/journal.pone.0195957.t002

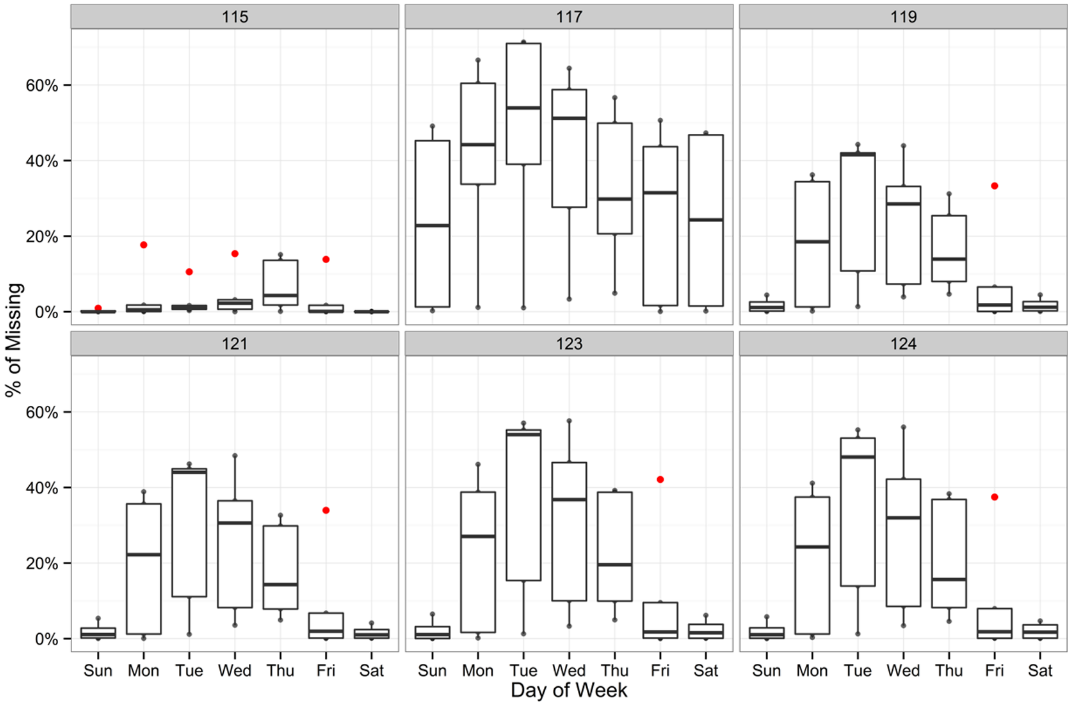


Fig 2. Boxplot of data missing rates by station and day of the week.

https://doi.org/10.1371/journal.pone.0195957.g002

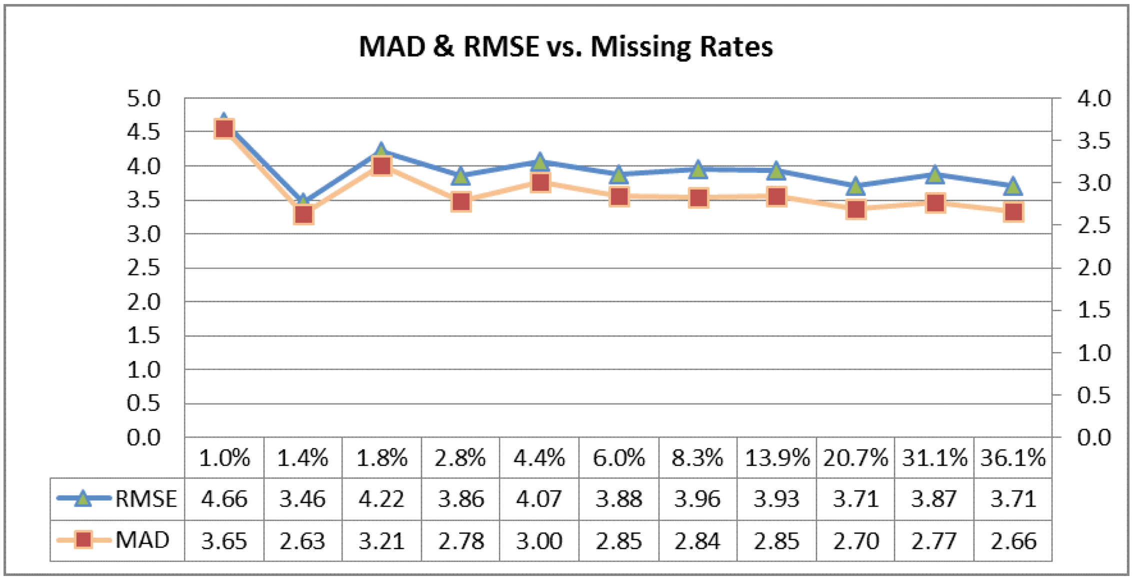
missing data points are generated randomly. The simulation process for the simulation is as follows:

1. Choose a specific data missing rate among 25%, 30%, 35%, . . ., 75% percentiles of missing rates of all stations during the study period;
2. Based on the missing rate selected above, generate the number of points to be flagged as missing in the dataset;
3. Generate missing data points randomly;
4. Repeat step 1 to 3 for different missing rates to generate corresponding scenarios;
5. Perform imputation on these generated scenarios using the proposed method and benchmark methods, and compare their results.

For the whole day of May 17, 2013, the traffic was congested during the rush hours and was in free-flow condition during the non-rush hours, just like the other days. Since the missing data points were generated randomly, with missing rate ranging from 1.0% and 36.1%, the missing data was likely to cover both free-flow conditions and congested conditions.

Imputation performance

The proposed imputation method and the benchmark methods were tested on 11 different scenarios. The semivariogram is shown in Fig 3. Imputation results are shown in Table 3 and Fig 4. It can be seen from the table that the proposed imputation method is more accurate than the other two methods in most scenarios. Only when the missing rate is lower than 1%, the performance of the historical average method is better than the proposed imputation method. KNN method usually has the lowest imputation accuracy. This may be due to that



	1.0%	1.4%	1.8%	2.8%	4.4%	6.0%	8.3%	13.9%	20.7%	31.1%	36.1%
RMSE	4.66	3.46	4.22	3.86	4.07	3.88	3.96	3.93	3.71	3.87	3.71
MAD	3.65	2.63	3.21	2.78	3.00	2.85	2.84	2.85	2.70	2.77	2.66

Fig 3. Initial variogram.

https://doi.org/10.1371/journal.pone.0195957.g003

there are only three features in this study for KNN to determine the nearest neighbors while KNN usually needs more than three features to obtain a reliable result [1].

Conclusions

The paper presents a Kriging based spatiotemporal data imputation approach that is able to fully utilize the spatiotemporal information of the traffic data and that does not assume the data follow any distribution. As traffic flows from upstream to downstream, the traffic stream characteristics at a certain location are usually related to those at neighboring locations. So the traffic stream characteristics at upstream and downstream locations can be used to impute the missing value at a specific location. Besides, the traffic characteristics of a specific location at a certain time are also related to those of previous/future days or time periods. Therefore, a Kriging based imputation method that considers both temporal and spatial information is proposed. Compared with KNN and historical average, the proposed method has higher

Table 3. Performance of proposed approach.

Quantile	% Missing	MAD(Kr)	MAD(H)	MAD(Kk)	RMSE(Kr)	RMSE(H)	RMSE(Kk)
25%	1.0%	3.65	**3.06**	5.63	4.66	**4.16**	7.26
30%	1.4%	**2.63**	3.04	3.81	**3.46**	3.89	5.25
35%	1.8%	3.21	**3.03**	5.33	**4.22**	4.44	6.89
40%	2.8%	**2.78**	3.16	4.42	**3.86**	4.46	6.25
45%	4.4%	**3.00**	3.40	4.90	**4.07**	4.72	6.37
50%	6.0%	**2.85**	3.11	4.06	**3.88**	4.52	5.78
55%	8.3%	**2.84**	3.11	4.39	**3.96**	4.52	5.99
60%	13.9%	**2.85**	2.93	3.99	**3.93**	4.13	5.55
65%	20.7%	**2.70**	3.20	4.16	**3.71**	4.44	5.75
70%	31.1%	**2.77**	3.01	4.18	**3.87**	4.26	5.87
75%	36.1%	**2.66**	3.07	4.00	**3.71**	4.30	5.68

Note: Kr represents the proposed imputation method, H represents the historical average method, and Kk represents the KNN method.

https://doi.org/10.1371/journal.pone.0195957.t003

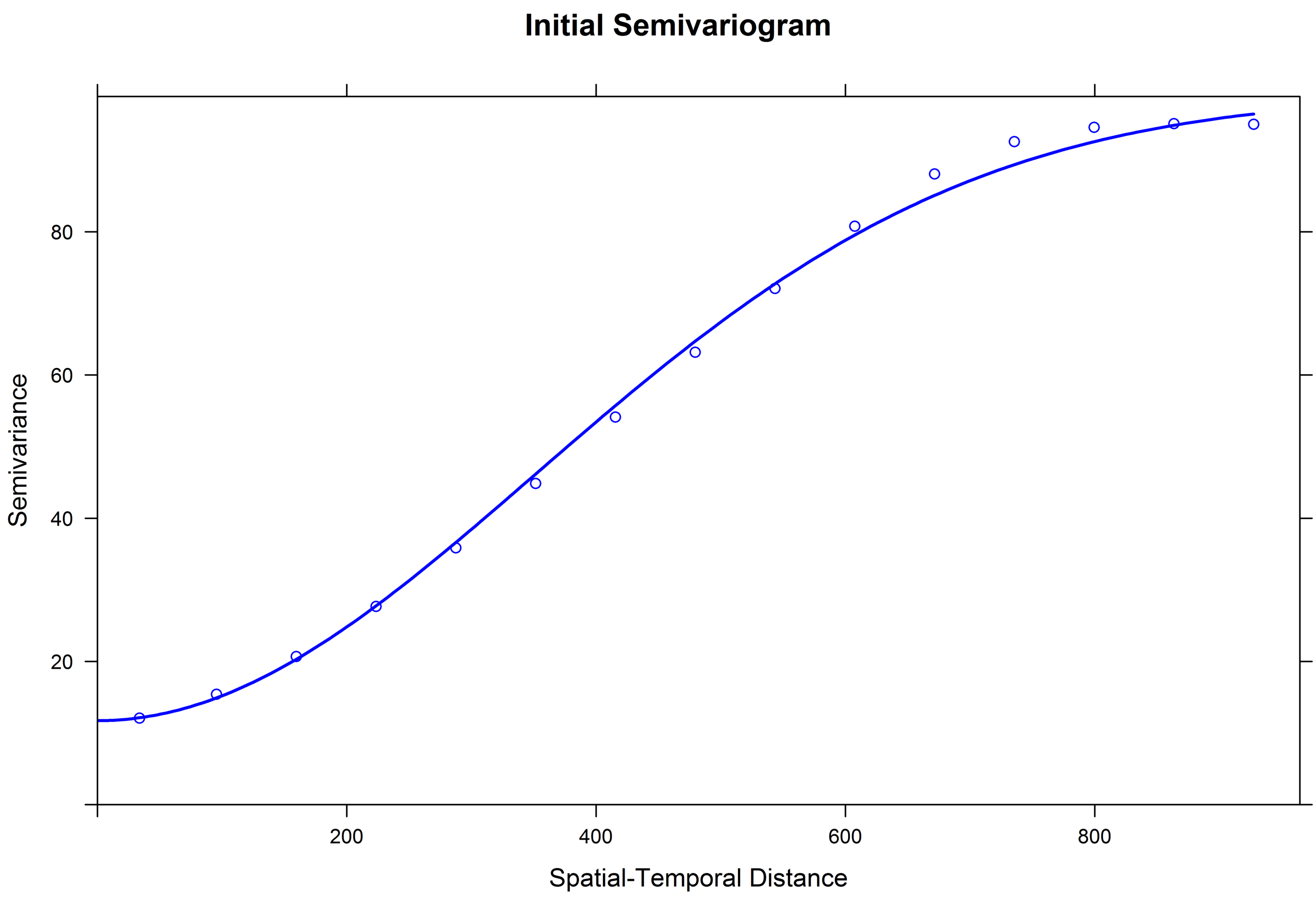


Fig 4. Imputation performance of proposed approach.

https://doi.org/10.1371/journal.pone.0195957.g004

imputation accuracy in ten out of the eleven generated scenarios. Only when the data missing rate is lower than 1%, the performance of the historical average method is better than the proposed imputation method. It suggests that the historical average method is more suitable for the scenarios in which only a few data points are missing. This study also finds that the KNN method has the lowest imputation accuracy. The result of KNN may be more reliable when there are more features to determine the nearest neighbors are available.

Supporting information

S1 Data. Data used in this study.
(RAR)

Author Contributions

Conceptualization: Hongtai Yang, Jianjiang Yang.

Data curation: Hongtai Yang, Jianjiang Yang.

Formal analysis: Hongtai Yang, Jianjiang Yang.

Funding acquisition: Lee D. Han.

Investigation: Hongtai Yang.

Methodology: Hongtai Yang, Jianjiang Yang, Xiaohan Liu, Shih-miao Chin.

Project administration: Jianjiang Yang.

Resources: Lee D. Han.

Software: Lee D. Han.

Supervision: Lee D. Han.

Validation: Xiaohan Liu, Shih-miao Chin, Ho-ling Hwang.

Visualization: Xiaohan Liu, Shih-miao Chin, Ho-ling Hwang.

Writing – original draft: Hongtai Yang, Jianjiang Yang, Lee D. Han, Xiaohan Liu, Shih-miao Chin, Ho-ling Hwang.

Writing – review & editing: Hongtai Yang, Jianjiang Yang, Lee D. Han, Xiaohan Liu, Li Pu, Shih-miao Chin, Ho-ling Hwang.

References

1. Ma X, Dai Z, He Z, Ma J, Wang Y, Wang Y. Learning Traffic as Images: A Deep Convolutional Neural Network for Large-Scale Transportation Network Speed Prediction. Sensors. 2017; 17(4).
2. Ma X, Yu H, Wang Y, Wang Y. Large-Scale Transportation Network Congestion Evolution Prediction Using Deep Learning Theory. Plos One. 2015; 10(3):e0119044. https://doi.org/10.1371/journal.pone.0119044 PMID: 25780910
3. Ding C, Duan J, Zhang Y, Wu X, Yu G. Using an ARIMA-GARCH Modeling Approach to Improve Subway Short-Term Ridership Forecasting Accounting for Dynamic Volatility. IEEE Transactions on Intelligent Transportation Systems. 2017;PP(99):1–11.
4. Ding C, Wang D, Ma X, Li H. Predicting Short-Term Subway Ridership and Prioritizing Its Influential Factors Using Gradient Boosting Decision Trees. Sustainability. 2016; 8(11):1100.
5. Ma X, Ding C, Luan S, Wang Y, Wang Y. Prioritizing Influential Factors for Freeway Incident Clearance Time Prediction Using the Gradient Boosting Decision Trees Method. IEEE Transactions on Intelligent Transportation Systems. 2017; 18(9):2303–10.
6. Turner S. Defining and Measuring Traffic Data Quality: White Paper on Recommended Approaches. Transportation Research Record: Journal of the Transportation Research Board. 2004; 1870(-1):62–9. https://doi.org/10.3141/1870-08
7. Williams BM, Hoel LA. Modeling and Forecasting Vehicular Traffic Flow as a Seasonal ARIMA Process: Theoretical Basis and Empirical Results. Journal of Transportation Engineering. 2003; 129(6):664–72.
8. Chen C, Kwon J, Rice J, Skabardonis A, Varaiya P. Detecting errors and imputing missing data for single-loop surveillance systems. Transportation Data Research. 2003; 1855(1855):160–7. WOS:000189494800020.
9. Al-Deek H, Venkata C, Ravi Chandra S. New Algorithms for Filtering and Imputation of Real-time and Archived Dual-loop Detector Data in I-4 Data Warehouse. Transportation Research Record: Journal of the Transportation Research Board. 2004;(1867):116–26.
10. Ni D, John Leonard II. Markov Chain Monte Carlo Multiple Imputation Using Bayesian Networks for Incomplete Intelligent Transportation Systems Data. Transportation Research Record Journal of the Transportation Research Board. 2005; 1935(1):57–67.
11. Qu L, Li L, Zhang Y, Hu J. PPCA-Based Missing Data Imputation for Traffic Flow Volume: A Systematical Approach. IEEE Transactions on Intelligent Transportation Systems. 2009; 10(3):512–22. https://doi.org/10.1109/TITS.2009.2026312
12. Li Y, Li Z, Li L. Missing traffic data: comparison of imputation methods. IET Intelligent Transport Systems. 2014; 8(1):51–7. https://doi.org/10.1049/iet-its.2013.0052
13. Zhong M, Sharma S, Lingras P. Matching Patterns for Updating Missing Values of Traffic Counts. Transportation Planning and Technology. 2006; 29(2):141–56. https://doi.org/10.1080/03081060600753461 WOS:000239390800004.

14. Smith B, Scherer W, Conklin J. Exploring Imputation Techniques for Missing Data in Transportation Management Systems. Transportation Research Record. 2003; 1836(1):132–42. https://doi.org/10.3141/1836-17

15. Zhong M, Sharma S, Lingras P. Genetically Designed Models for Accurate Imputation of Missing Traffic Counts. Transportation Research Record: Journal of the Transportation Research Board. 2004;(1879):71–9.

16. Qu L, Zhang Y, Hu J, Jia L, Li L, editors. A BPCA Based Missing Value Imputing Method for Traffic Flow Volume Data. Intelligent Vehicles Symposium, 2008 IEEE; 2008: IEEE.

17. Liu Z, Sharma S, Datla S. Imputation of Missing Traffic Data during Holiday Periods. Transportation Planning and Technology. 2008; 31(5):525–44. https://doi.org/10.1080/03081060802364505

18. Williams B, Durvasula P, Brown D. Urban Freeway Traffic Flow Prediction: Application of Seasonal Autoregressive Integrated Moving Average and Exponential Smoothing Models. Transportation Research Record: Journal of the Transportation Research Board. 1998;(1644):132–41.

19. Wu C-H, Ho J-M, Lee D-T. Travel-time Prediction with Support Vector Regression. IEEE transactions on intelligent transportation systems. 2004; 5(4):276–81.

20. Smith BL, Demetsky MJ, editors. Short-term Traffic Flow Prediction Models-A Comparison of Neural Network and Nonparametric Regression Approaches. Systems, Man, and Cybernetics, 1994 Humans, Information and Technology, 1994 IEEE International Conference on; 1994: IEEE.

21. Qi Y, Ishak S. A Hidden Markov Model for Short Term Prediction of Traffic Conditions on Freeways. Transportation Research Part C: Emerging Technologies. 2014; 43:95–111. https://doi.org/10.1016/j.trc.2014.02.007.

22. D'Ambrosio A. Accurate Tree-based Missing Data Imputation and Data Fusion within the Statistical Learning Paradigm. Journal of Classification. 2012; 29(2):227–58.

23. Ma X, Luan S, Du B, Yu B. Spatial Copula Model for Imputing Traffic Flow Data from Remote Microwave Sensors. Sensors. 2017; 17(10):2160.

24. Kamarianakis Y, Prastacos P. Space–time modeling of traffic flow. Computers & Geosciences. 2005; 31(2):119–33.

25. Yang J, Han LD, Freeze PB, editors. Short-Term Freeway Speed Profiling Based on Longitudinal Spatial-Temporal Dynamics. Transportation Research Board 93rd Annual Meeting; 2014.

26. Stathopoulos A, Karlaftis MG. A Multivariate State Space Approach for Urban Traffic Flow Modeling and Prediction. Transportation Research Part C: Emerging Technologies. 2003; 11(2):121–35.

27. Cressie N. The origins of kriging. Mathematical Geology. 1990; 22(3):239–52.

28. Yang H, Cherry CR, Zaretzki R, Ryerson MS, Liu X, Fu Z. A GIS-Based Method to Identify Cost-effective Routes for Rural Deviated Fixed Route Transit. Journal of Advanced Transportation. 2016.

29. Hull J, White A. Incorporating Volatility Updating into the Historical Simulation Method for Value-at-risk. Journal of risk. 1998; 1(1):5–19.

30. Altman NS. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. American Statistician. 1992; 46(3):175–85.

31. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing Value Estimation Methods for DNA Microarrays. Bioinformatics. 2001; 17(6):520–5. PMID: 11395428

32. Potts JF, Marshall MA, Crockett EC, Washington J. A Guide for Planning and Operating Flexible Public Transportation Services2010.

33. Yang H, Cherry C. Statewide Rural-Urban Bus Travel Demand and Network Evaluation: An Application in Tennessee. Journal of Public Transportation. 2012; 15(3):97–111. https://doi.org/10.5038/2375-0901.15.3.6

34. Yang H, Cherry CR. Use characteristics and demographics of rural transit riders: a case study in Tennessee. Transportation Planning and Technology. 2017; 40(2):213–27. https://doi.org/10.1080/03081060.2016.1266168