Traffic Speed Estimation Based on Multi-source Data Feeds:

A Case Study of the San Francisco Bay Area

Yue Lin

1 Introduction

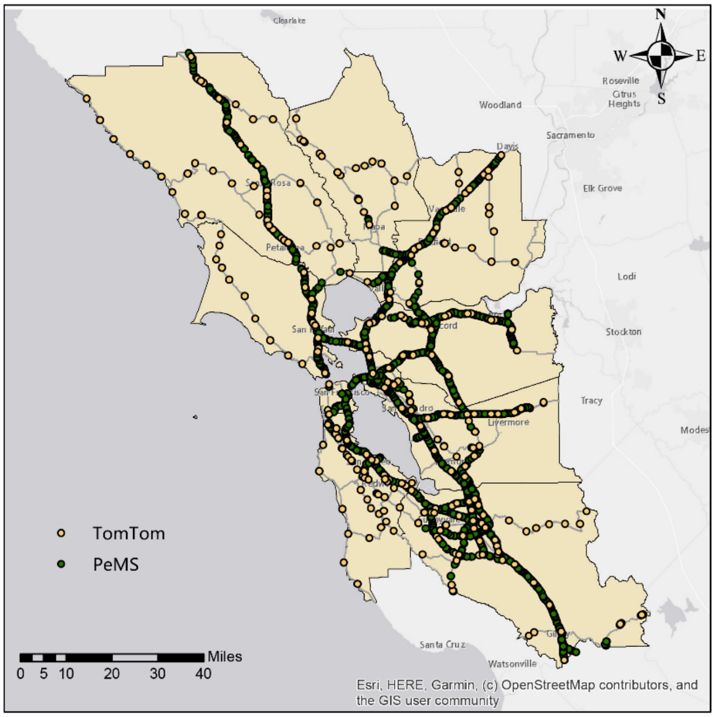
We are living in an era when geospatial big data are abundant and readily available (Miller and Goodchild, 2015). Accordingly, geolocated traffic big data have been used to assist in the control and management of transportation networks, which help to fill the gaps of simulations where empirical data are missing (Balakrishnan et al., 2020; Qiang and Xu, 2019). To date, traffic big data mainly come from two data sources: (1) traffic sensors installed by city or state governments on or along roadways, where the most prevalent ones are loop detectors and traffic cameras because they can be used for traffic detection for a long time period and are relatively cost-efficient (Lowry, 2014; Pinto et al., 2020; Yang et al., 2018), and (2) Global Positioning System (GPS) receivers equipped on vehicles or in mobile phones, where the obtained data are also known as Floating Car Data (FCD) (Santi et al., 2014; Zou et al., 2012).

However, simply using one of the above-mentioned data sources for traffic condition estimation may be problematic. For one, traffic sensors only detect vehicles at a point (e.g., loop detectors) or over a short road section (e.g., traffic cameras) and are unequally distributed in the road network systems, which results in the limited spatial coverage of traffic sensor data (Meng et al., 2017). For another, FCD is considered to be a partial and biased representation of the road traffic conditions, for it does not record the trajectories of vehicles without any GPS receivers equipped (Zhu et al., 2019).

To address the problem of data biases, in this project, we aggregate different sources of traffic data in order to provide a more accurate and reliable estimation of traffic performance across the road network systems. Our method is built based on Ordinary CoKriging (OCK), a data imputation model that can take advantage of multiple datasets. We also introduce a network distance metric in the variogram modelling for OCK, which, compared to Euclidean distance, is more suitable to describe the spatial distance between road links.

2 Materials

The San Francisco Bay Area is a populous region in Northern California that consists of nine counties: Alameda, Contra Costa, Marin, Napa, San Mateo, Santa Clara, Solano, Sonoma, and San Francisco. The Bay Area is known to have an extensive freeway and highway system with a bustling road traffic. In 2019, the Daily Miles Traveled (DMT) of the Bay Area has reached up to 23 miles per person. In this project, the primary source used for data aggregation was obtained from TomTom (https://developer.tomtom.com/), an online map service that provides real-time traffic flow information using probe vehicles. The traffic speed of 418 observations were collected at 8:35 AM on March 29, 2020 from TomTom, where these observations are sparsely distributed along the freeways and highways in the Bay Area. The secondary data used for traffic performance estimation was obtained from the Caltrans Performance Measurement System (PeMS) (http://pems.dot.ca.gov/). PeMS provides traffic information collected at a 5-minute interval from over 39,000 traffic sensors distributed across the freeway system in the State of California. In this project, we retrieved the average traffic occupancy from 3,888 PeMS traffic detectors at 8:35 AM on March 29, 2020 as our secondary data source. The study area of this project and the distributions of observations from two data sources are presented in Figure 1.



Esri, HERE, Garmin, (c) OpenStreetMap contributors, and the GIS user community

Figure 1. Study area: The San Francisco Bay Area. The Bay Area is located in Northern California that consists of nine counties. A total of 418 observations are collected from TomTom, and a total of 3,888 observations are collected from PeMS.

3 Methods

3.1 Ordinary CoKriging

OCK is a spatial interpolation technique that supports multivariate operation, which estimates the missing values of a poorly sampled variable (i.e., the primary variable) with the help of a well sampled variable (i.e., the secondary variable) (Marcotte, 1991). In this project, OCK is employed to estimate the missing values of traffic speed obtained from TomTom by taking advantage of the traffic occupancy information from PeMS. OCK can be implemented with the following steps.

Step 1: Variogram estimation. In OCK, in addition to the experimental semivariogram for both variables (i.e., the primary and secondary variables), an experimental cross-semivariogram for the combination of both variables should also be estimated. Assuming *A* as the primary variable and *B* as the secondary variable, the experimental semivariogram for variables *A* and *B*, $\widehat{\gamma_A}$ and $\widehat{\gamma_B}$, can be calculated by:

$$\widehat{\gamma_A} = \frac{1}{2N_A(\boldsymbol{h})}\sum_{i=1}^{N_A(\boldsymbol{h})}[Y_A(\boldsymbol{s}) - Y_A(\boldsymbol{s}+\boldsymbol{h})]^2 \tag{1}$$

$$\widehat{\gamma_B} = \frac{1}{2N_B(\boldsymbol{h})}\sum_{i=1}^{N_B(\boldsymbol{h})}[Y_B(\boldsymbol{s}) - Y_B(\boldsymbol{s}+\boldsymbol{h})]^2 \tag{2}$$

where $\boldsymbol{s}$ is the vector of location of an observation, $\boldsymbol{h}$ is the vector of spatial lag between $\boldsymbol{s}$ and another observed location to capture the spatial dependency, and $N_A(\boldsymbol{h})$ and $N_B(\boldsymbol{h})$ are the numbers of pairs within the spatial lag $\boldsymbol{h}$ for *A* and *B*, respectively.

Similarly, the experimental cross-semivariogram for the combination of *A* and *B* can be calculated by:

$$\widehat{\gamma_{AB}} = \frac{1}{2N_{AB}(\boldsymbol{h})}\sum_{i=1}^{N_{AB}(\boldsymbol{h})}[Y_A(\boldsymbol{s}) - Y_A(\boldsymbol{s}+\boldsymbol{h})][Y_B(\boldsymbol{s}) - Y_B(\boldsymbol{s}+\boldsymbol{h})] \tag{3}$$

where $N_{AB}(\boldsymbol{h})$ is the numbers of pairs within $\boldsymbol{h}$ for the combination of variables *A* and *B*.

Step 2: Model fitting. After estimating the semivariogram and cross-semivariogram, we fit a spherical semivariogram model (with nugget effect) by:

$$\gamma(\boldsymbol{h}) = \begin{cases} 0, \boldsymbol{h} = 0 \\ a + (\sigma^2 - a)\left(\frac{3\boldsymbol{h}}{2\emptyset} - \frac{\boldsymbol{h}^3}{2\emptyset^3}\right), 0 < \boldsymbol{h} < \emptyset \\ \sigma^2, \boldsymbol{h} > \emptyset \end{cases} \tag{4}$$

where σ^2 and $\emptyset$ refer to the still and range of the semivariogram, respectively.

Step 3: Spatial prediction using OCK. The estimated value of variable *A* can then be calculated by:

$$z_{OCK} = \sum_{i=1}^{N_{AB}(\boldsymbol{h})} \lambda_i Y_A + \sum_{j=1}^{N_{AB}(\boldsymbol{h})} \lambda_j{}' Y_B \tag{5}$$

where Y_A and Y_B are the values of variables *A* and *B*, and λ_i and $\lambda_j{}'$ are the estimated weights for variable *A* and *B*, respectively.

3.2 Network distance metric for variogram modelling

In most cases, Euclidian distance is adopted to estimate the spatial lag between pairs of observations when modelling the semivariogram and cross-semivariogram in OCK. However, for road networks, the Euclidian distance between two sampling points is not able to reflect their actual distance along the roadways and may impair the prediction accuracy of OCK. Therefore, in this project, we develop a network distance metric based on the isometric embedding theory to obtain a more accurate estimation of distance between sampling points on roads (Zou et al., 2012).

An isometric embedding is a function used to project a vector into a new measurement space where the distances between vectors are preserved and represent the actual distances between points in the network. The implementation of network distance metric using isometric embedding follows the steps below.

Step 1: Distance matrix calculation. We first construct a graph consisting of all the n observation points and the links between connected points, and then calculate the distance matrix $\boldsymbol{D} = \{d_{ij}\}_{n\times n}$, where d_{ij} denotes the shortest distance each pair of points.

Step 2: Projection. To project the observations to the new measurement space, a new matrix $\boldsymbol{B} = \{b_{ij}\}_{n\times n}$ is constructed based on the distance matrix $\boldsymbol{D}$:

$$b_{ij} = \frac{1}{2}\left(-d_{ij}^2 + \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2 + \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2\right) \qquad (6)$$

Then, the new coordinates of the points $\boldsymbol{s}'$ can be calculated by:

$$\boldsymbol{s}' = \boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2} \qquad (7)$$

where $\boldsymbol{\Gamma} = (e_1, e_2)$ is the normalized eigenvectors of the greatest two eigenvalues of matrix $\boldsymbol{B}$, λ_1 and λ_2, and $\boldsymbol{\Lambda} = diag(\lambda_1, \lambda_2)$.

Table 1. Coordinates of observations before and after isometric embedding transformation.

Coordinates in Euclidean space		**Coordinates in non-Euclidean space**	
x	*y*	*x*	*y*
587009.7	4162039	36736.92	11536.4
575504	4174618	-40713.9	-6833.54
586917.7	4173355	30831.1	18338.09
595194.9	4173270	-39922.2	-35005.3
603578.5	4173324	17396.21	1341.501
611868	4173987	-54732	68162.48

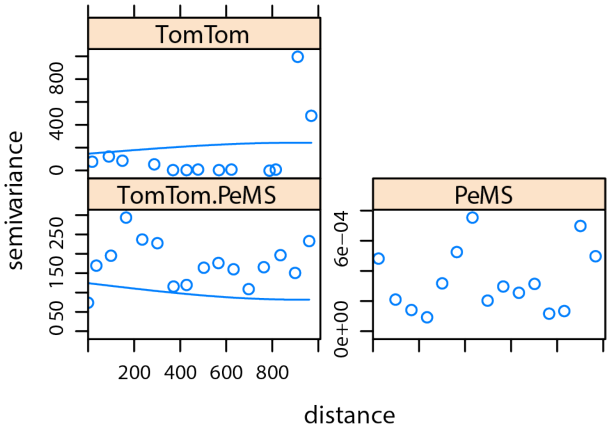


Figure 2. Semivariogram and cross-semivariogram. The semivariogram is calculated for traffic data obtained from TomTom and PeMS, respectively, and the cross-semivariogram is calculated for the combination of TomTom and PeMS data.

4 Results and discussion

4.1 Semivariogram and cross-semivariogram

Several examples of the results of coordinate transformation using isometric embedding are presented in Table 1. After projecting the observation points into a non-Euclidean space, the network distance between any pair of points can be obtained in the variogram modeling.

The plots of semivariogram for each variable and cross-semivariogram for the combination of both variables are displayed in Figure 2. Compared to the semivariogram that only considers a single variable (TomTom or PeMS), the empirical values of the cross-semivariogram fit better with the true values.

4.2 OCK interpolation

The traffic speed interpolation results using OCK are presented in Figure 3. The interpolated maps are fragmented maybe because of the use of network distances, and more works should be conducted to fix this problem. From the maps we produce, it is found that the traffic speed varies significantly across the Bay Area, where its southeast and north have observed a relatively lower average traffic speed, and the traffic speed at the center of this region is higher. The variance of interpolation is higher in the central regions, denoting a higher uncertainty of the interpolated results.

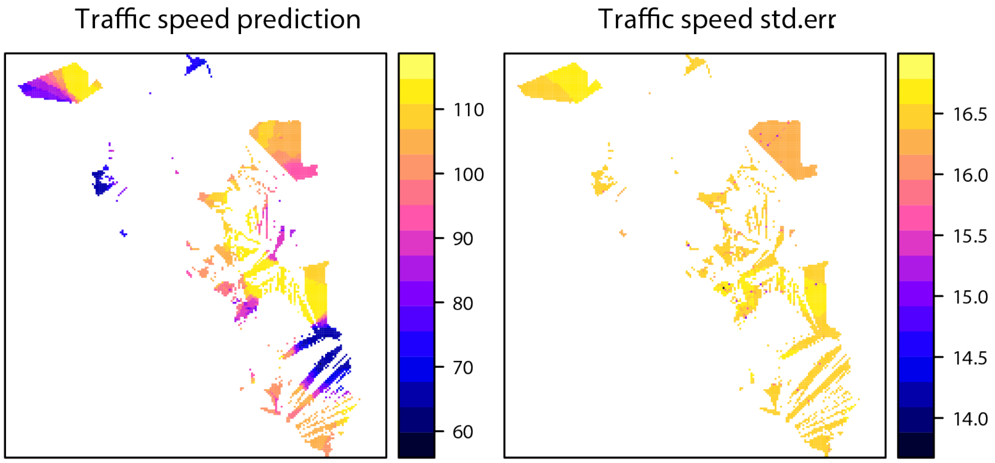


Figure 3. OCK interpolation results. The outputs include the predicted values for traffic speed and the corresponding variance.

5 Conclusions

In this project, we perform traffic speed interpolation using an improved OCK model. We extend the OCK model by introducing a network distance metric developed based on the isometric embedding theory, which is more applicable for the interpolation of networks. To implement the interpolation, traffic flow information are collected from two data sources, where

TomTom serves as the primary data source and PeMS serves as the secondary data source. The results show that the southeastern and northern areas of the Bay Area have a relatively low traffic speed.

There are still some limitations in this project. First, the PeMS data (secondary variable) mainly lies on several arterial roadways, which may not be helpful for the interpolation of the collector streets. Second, the introduction of network distance metric may cause distortions and result in a fragmented interpolated map. Therefore, in the future, more works should be conducted to address these problems.

References

Balakrishnan, S., Zhang, Z., Machemehl, R., Murphy, M.R., 2020. Mapping resilience of Houston freeway network during Hurricane Harvey using extreme travel time metrics. Int. J. Disaster Risk Reduct. 47, 101565. https://doi.org/10.1016/j.ijdrr.2020.101565

Lowry, M., 2014. Spatial interpolation of traffic counts based on origin-destination centrality. J. Transp. Geogr. 36, 98–105. https://doi.org/10.1016/j.jtrangeo.2014.03.007

Marcotte, D., 1991. Cokriging with matlab. Comput. Geosci. 17, 1265–1280. https://doi.org/10.1016/0098-3004(91)90028-C

Meng, C., Yi, X., Su, L., Gao, J., Zheng, Y., 2017. City-wide traffic volume inference with loop detector data and taxi trajectories, in: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems. ACM, Los Angeles Area, CA, USA. https://doi.org/10.1145/3139958.3139984

Miller, H.J., Goodchild, M.F., 2015. Data-driven geography. GeoJournal 80, 449–461. https://doi.org/10.1007/s10708-014-9602-6

Pinto, J.A., Kumar, P., Alonso, M.F., Andreão, W.L., Pedruzzi, R., Espinosa, S.I., de Almeida Albuquerque, T.T., 2020. Kriging method application and traffic behavior profiles from local radar network database: A proposal to support traffic solutions and air pollution control strategies. Sustain. Cities Soc. 56, 102062. https://doi.org/10.1016/j.scs.2020.102062

Qiang, Y., Xu, J., 2019. Empirical assessment of road network resilience in natural hazards using crowdsourced traffic data. Int. J. Geogr. Inf. Sci. https://doi.org/10.1080/13658816.2019.1694681

Santi, P., Resta, G., Szell, M., Sobolevsky, S., Strogatz, S.H., Ratti, C., 2014. Quantifying the benefits of vehicle pooling with shareability networks. Proc. Natl. Acad. Sci. U. S. A. 111, 13290–13294. https://doi.org/10.1073/pnas.1403657111

Yang, H., Yang, J., Han, L.D., Liu, X., Pu, L., Chin, S.M., Hwang, H.L., 2018. A Kriging based spatiotemporal approach for traffic volume data imputation. PLoS One 13, 1–11. https://doi.org/10.1371/journal.pone.0195957

Zhu, L., Yu, F.R., Wang, Y., Ning, B., Tang, T., 2019. Big data analytics in intelligent transportation systems: A survey. IEEE Trans. Intell. Transp. Syst. 20, 383–398. https://doi.org/10.1109/TITS.2018.2815678

Zou, H., Yue, Y., Li, Q., Yeh, A.G.O., 2012. An improved distance metric for the interpolation of link-based traffic data using kriging: A case study of a large-scale urban road network. Int. J. Geogr. Inf. Sci. 26, 667–689. https://doi.org/10.1080/13658816.2011.609488