Journal of Korean
Data & Information Science Society
2002, Vol. 13, No.2 pp. 261 270

Spatial-Temporal Modelling of Road Traffic Data in Seoul City

Sangyeol Lee[1], Soohan Ahn[2], Changyi Park[3] and Jongwoo Jeon[4]

Abstract

Recently, the demand of the Intelligent Transportation System (ITS) has been increased to a large extent, and a real-time traffic information service based on the internet system became very important. When ITS companies carry out real-time traffic services, they find some traffic data missing, and use the conventional method of reconstructing missing values by calculating average time trend. However, the method is found unsatisfactory, so that we develop a new method based the spatial and spatial-temporal models. A cross-validation technique shows that the spatial-temporal model outperforms the others.

Key words: Intelligent Transportation System (ITS), spatial model, spatial-temporal model, spatial neighborhood structure, time series model, cross-validation method.

1. Introduction

In recent years, the demand of the ITS has been rapidly increased, and the task of collecting real-time traffic information has become crucial in running the ITS business. For obtaining accurate information, ITS companies have made efforts for analyzing the traffic data in lines with the internet business. However, such an effort often ends up with undesirable results. The ROTIS, an ITS company located in the metropolitan areas of Seoul city, has been collecting the real-time traffic data. The data consists of the information obtained through the beacon-based infrastructure at every 15 minutes. It provides customers with useful

1. Associate Professor, Department of Statistics, Seoul National University, Seoul, 151-742. E-mail: sylee@stats.snu.ac.kr
2. Post doctor, Department of Statistics, Seoul National University, Seoul, 151-742.
3. Ph.d student, Department of Statistics, Seoul National University, Seoul, 151-742.
4. Professor, Department of Statistics, Seoul National University, Seoul, 151-742.

traffic information via calculating the traveling velocity between two preassigned locations, measured from the road-side beacons and probe cars equipped with in-vehicle module. However, some portions of data are found missing since the traffic informations are collected only from probe cars.

A simple way to solve the missing data problem is to use a time trend. For illustration, let us assume that during 7:01-7:15 p.m., the traffic data on **A** road is missing. In this case, a conventional method to estimate the missing data is to use a time trend, calculated as the average of observed past data corresponding to 7:01-7:15 p.m. on **A** road. Another way is to use the traffic data on **A** road obtained during, say, 6:46-7:00 p.m. and prediction algorithm of fitted time series model for data on **A** road. Actually, this is the way that the ROTIS handles the problem. However, in utilizing the method one usually finds a large gap between the real and predicted values. The negative result is due to the fact that the traffic flow does not always follow one pattern in time. In that case, an extra information should be taken for compensation from another source. Therefore, we consider utilizing the data on other roads to set up more sophisticated statistical model, especially the data on the neighboring roads. From this reasoning, a spatial model is taken into consideration. So is the spatial-temporal model since traffic data has a stochastic trend both in time and space.

The spatial models are widely used in many application fields. Cliff and Ord [9] suggested the spatial autoregressive, moving average and regressive model and analyzed the data Casetti and Semple [5]. Ali [2] analyzed the STAR (spatial temporal auto-regression) process with unit spatial order. Pfeifer and Deutsch [12], [13] and Deutsch and Pfeifer [14] extended the STAR model to the STARMA (spatial temporal auto-regressive moving average) model. See also Nui and Tiao [11], Cressie and Majure [7] and Huang and Cressie [10].

This paper is organized as follows. In Section 2, we introduce a spatial model with certain spatial neighborhood structure suitable for traffic data. Based on this, we also introduce a spatial-temporal model, and develop a prediction procedure. In Section 3, we compare six prediction procedures through real data analysis. They are the procedures using the average time trend, the time series model, the spatial model, and the spatial-temporal model. The data used here is the traffic velocity data at Gangnam-Gu, a downtown area of Seoul city with 64 links. It is shown that the method based on spatial-temporal model outperforms the other methods.

2. Prediction of missing data

Before we proceed, we introduce some notations and terminologies originated from the transportation engineering (cf. Banks [3]). A crossroad is called a *node* and is given a unique id number. A road between two contiguous nodes is called a *link*. It is important to notice that there are two links between contiguous nodes. For example, for nodes A and B, there is the link from A to B (A-B link) and

the one from B to A (B-A link). In the meantime, we define N_i at link i as the family of neighboring links either through which cars can reach the link i or which one can reach through the link i. Everytime they pass a link, probe cars send the control center a velocity data. Now let $X_i(t)$, $i=1,\cdots,L$, $t=1,\cdots,J$, denote the observed velocity (km/h) data at link i and time t. In our analysis, we consider the mean-adjusted r.v.'s

$$Y_i(t) = X_i(t) - \mu_i(t),\ i=1,\cdots L,,\ t=1,\cdots,J, \tag{1}$$

where $\mu_i(t)$ is the time trend at link i time t.

2.1 Spatial model

2.1.1 First order Neighborhood structure

For $t=1,\cdots,96$, we consider the spatial model:

$$Y_i(t) = \sum_{j\in N_i} c^1_{i,j}(\theta^1(t))\, Y_j(t) + \delta_i(t), \quad i = 1,\cdots,L, \tag{2}$$

cf. Cressie [6]) where $\{\delta_i(t),\ i = 1,\ldots,L\}$ is a family of independent r.v.'s with mean 0 and variance σ^2_δ. Note that $c^1_{i,j}(\theta^1(t))$, which are the spatial dependence parameter, can be defined as the (i,j)-th component of the symmetric $L\times L$ matrix $C^1(\theta^1(t))$:

$$C^1(\theta^1(t)) = \theta^1_1(t)\, C^1_1 + \theta^1_2(t)\, C^1_2 + \theta^1_3(t)\, C^1_3, \tag{3}$$

where $\theta^1(t) = (\theta^1_1(t), \theta^1_2(t), \theta^1_3(t))'$ and the (i,j)-th element of C^1_1, C^1_2, C^1_3 is as follows:

$$(C^1_1)_{i,j} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are linked by going straight} \\ 0 & o.w. \end{cases}$$

$$(C^1_2)_{i,j} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are linked by turning to the right} \\ 0 & o.w. \end{cases} \tag{4}$$

$$(C_3^1)_{i,j} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are linked by turning to the left} \\ 0 & o.w. \end{cases}$$

Then, we can write

$$[I - C^1(\theta^1(t))]\mathbf{Y}(t) = \delta(t) \tag{5}$$

where $\mathbf{Y}(t) = (Y_1(t),\cdots,Y_L(t))'$ and $\delta(t) = (\delta_1(t),\cdots,\delta_L(t))'$.

2.1.2 Second order Neighborhood structure

Here we consider the spatial model for each time t :

$$Y_i(t) = \sum_j c^2_{i,j}(\theta^2(t))\, Y_j(t) + \eta_i(t), \quad i = 1,\cdots,L, \tag{6}$$

where $\{\eta_i(t);\ i = 1,\cdots,L\}$ is a family of independent r.v.'s with mean 0 and variance σ^2_η, and $c^2_{i,j}(\theta^2(t))$ is the (i,j)-th component of the following symmetric $L \times L$ matrix :

$$C^2(\theta^2(t)) \qquad C^2(\theta^2(t)) = \sum_{k=1}^{9} \theta^2_k(t)\, C^2_k \tag{7}$$

where $\theta^2(t) = (\theta^2_1(t),\cdots,\theta^2_9(t))'$ and the (i,j)-th element of the symmetric matrices $C^2_1,\ldots,C^2_9$ are defined as follows:

$$(C_1^2)_{i,j} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are linked by going straight} \\ 0 & o.w. \end{cases}$$

$$(C_2^2)_{i,j} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are linked by turning to the right} \\ 0 & o.w. \end{cases} \tag{8}$$

$$(C_3^2)_{i,j} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are linked by turning to the left} \\ 0 & o.w. \end{cases}$$

$$(C_4^2)_{i,j} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are linked by going straight twice} \\ 0 & o.w. \end{cases}$$

$$(C_5^2)_{i,j} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are linked by turning to the right and one-time going strait} \\ 0 & o.w. \end{cases}$$

$$(C_6^2)_{i,j} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are linked by turning to the left and then going straight} \\ 0 & o.w. \end{cases}$$

$$(C_7^2)_{i,j} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are linked by turning to the right and then turning to the left} \\ 0 & o.w. \end{cases}$$

$$(C_8^2)_{i,j} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are linked by turning to the right twice} \\ 0 & o.w. \end{cases}$$

$$(C_9^2)_{i,j} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are linked by turning to the left twice} \\ 0 & o.w. \end{cases}$$

Then, we obtain

$$[I - C^2(\theta^2(t))]\mathbf{Y}(t) = \eta(t) \tag{9}$$

where $\mathbf{Y}(t) = (Y_1(t), \cdots, Y_L(t))'$ and $\eta(t) = (\eta_1(t), \cdots, \eta_L(t))'$

2.1.3 Parameter estimation and prediction

The estimators $\hat{\theta}^v(t)$ of $\theta^v(t)$ v=1, 2, are obtained as the ones that minimize the error sum of squares, that is,

$$\hat{\theta}^v(t) = \text{argmin } \mathbf{Y}'(t)[I - C(\theta^v(t))]'[I - C(\theta^v(t))]\mathbf{Y}(t) \tag{10}$$

If we set $N_1 = 3$ and $N_2 = 9$, it follows that

$$\hat{\theta}^v(t) = (A^v)^{-1} b^v, \quad v=1,2, \tag{11}$$

where A^v is an $N_v \times N_v$ matrix and its (l,m)-th component is $Y'(t)(C_l^v)'(C_m^v)Y(t)$ and

$$b^v = (b_1^v, \cdots, b_{N_v}^v)', \quad b_k^v = Y'(t)(C_k^v)'Y(t), \quad k = 1, \cdots, N_v, \quad v = 1,2.$$

Now let us consider the case that data missing occurs. Let M denote the set of links where the data is missing, and let $Y^{(M)}(t)$ denote the vector of observed data only. Then, taken into consideration $Y^{(M)}(t)$, redefine the matrix C^v, $v = 1,2$, in Equations (3) and (7), and the matrices C_k^v, $k = 1,\cdots,N_v$, $v = 1,2$, in Equations (4) and (8). That is, if $i \in M$, then we remove the i-th column and row from the matrices $C^{v'}s$ and $C_k^{v'}s$. Denote them by $C^{v,M}$ and $C_k^{v,M}$, $k = 1,\cdots,N_v$, $v = 1,2$, and find the estimators $\hat{\theta}^{v(M)}$ by minimizing the error sum of squares:

$$\hat{\theta}^{v(M)} = \text{argmin}\ (Y^{(M)}(t))'(I - C^{v,M})'(I - C^{v,M})\ Y^{(M)}(t) \tag{12}$$

Finally, we predict the missing data $X_i(t)$ of the link $i \in M$ as follows:

$$\hat{X}_i^{v,S}(t) = \mu_i(t) + \sum_{j \notin M, j \in N_i} c_{i,j}^v(\hat{\theta}^{v(M)}(t))\, Y_j(t), \quad v = 1,2. \tag{13}$$

2.2 Spatial-temporal model

In this subsection, we consider the spatial-temporal model:

$$Z_i^v(t) = Y_i(t) - \sum_{j \in N_i} c_{i,j}^v(\theta^v(t))\, Y_i(t),$$

$$Z_i^v(t) - \sum_{l=1}^{p} \gamma_{i,l}^v Z_i^v(t-l) = \varepsilon_i(t) + \sum_{m=1}^{q} \phi_{i,m}^v \varepsilon_i^v(t-m), \quad v = 1,2 \tag{14}$$

where $\{Z_i^v(t)\}$ are independent of $\{Z_j^v(t)\}$ for $j \neq i$, and $\{\varepsilon_i^v(t), t = 1,2,\cdots,J\}$, a family of i.i.d. r.v.'s with the mean 0 and variance $(\sigma_i^v)^2$, are independent of $\{\varepsilon_j^v(t), t = 1,2,\cdots,J\}$ for $j \neq i$. We call $Z_i(t)$ the spatial residual of link i at time t.

Note that the first equation in (14) is the spatial model considered in Section 2.1 and the second equation forms a stationary ARMA (p, q) time series model. The model indicates that the spatial residuals obtained after removing the effects of neighboring links still have a correlation, but only in time not in space. For estimating the parameters, we use a sort of the plug-in method. That is, we adjust the model in (14) as follows:

$$Z_i^{v*}(t) = Y_i(t) - \sum_{j \in N_i} c^v_{i,j}(\widehat{\theta^v}(t))\, Y_i(t),$$

$$Z_i^{v*}(t) - \sum_{l=1}^{p} \gamma^v_{i,l} Z_i^{v*}(t-l) = \varepsilon_i(t) + \sum_{m=1}^{q} \phi^v_{i,m} \varepsilon^v_i(t-m), \quad v = 1,2 \tag{15}$$

where $\widehat{\theta}^v(t)$ is the one defined in Equation (11). In handling missing values, we replace $\widehat{\theta}^v(t)$ by $\widehat{\theta^v}^{(M)}(t)$ defined in Equation (12).

For each link, we determine the orders p and q using spatial residuals $\{Z_i^{v*}(t),\ t = 1,2,\cdots,J\}$ defined in Equation (15) based on Akaike's Information Criterion (AIC). That is, we consider ARMA (p,q) models, $p,q = 1,\cdots,6$, at each link and choose the ARMA (p_i^v, q_i^v) model which has the smallest AIC value among those models. Once we choose an optimal model, we can estimate the parameters $\gamma's, \phi's$ and $\sigma's$ based on the least squares method.

Now, in prediction of missing values we use the Kalman filter algorithm. For this, we need a form of a state space model for the second equation of Model (15) (cf. Harvey [1]). Then, similarly to (13), we can predict missing data. The details are omitted for brevity.

3. Data Analysis

From the year 1998, the ROTIS has been collecting the real-time traffic informations in Seoul metropolitan areas using about 15,000 links through the beacon-based infrastructure. At present, about 5 million number of data are gathered everyday. In this

section, we compare our proposed prediction procedure with those based on the time trend method and time series analysis. For the comparison, we consider only data observed in the 64 links of the Gangnam-Gu area in Seoul from February 7 to February 28 in 2001 year except Saturdays and Sundays, 16 days in total. We first obtain the time trend $\mu_i(t)$ in Equation (1) and the time series parameters in Section 2.2 for each link using data.

$\gamma's$ and $\phi's$ and $\sigma^{2}'s$

A. **Average time trend model**

This model predicts $X_i(t)$ only using $\mu_i(t)$ of data obtained at link i and time t.

T. **Time series model**

For each link i, we choose an optimal ARMA (p, q) model out of the ARMA models with $p, q = 0, \cdots, 6$, for $\{Y_i(t)\}$, $i = 1, \cdots, L$, where $\{Y_i(t)\}$ is defined in Equation (1). If $Y_i(t|t-1)$ is the one-step predictor, the predicted value $\widehat{X}_i^{\,T}(t)$ of $X_i(t)$ is given as follows (cf. Harvey [1]):

$$\widehat{X}_i^{\,T}(t) = \mu_i(t) + Y_i(t|t-1). \tag{16}$$

S1. **Spatial model with the first order neighborhood structure**

We predict the missing data using spatial model with the first order neighborhood structure in Section 2.1. The predictor is $\widehat{X}_i^{\,1,S}(t)$ as in Section 2.1.

S2. **Spatial model with the second order neighborhood structure**

We predict the missing data using spatial model with the second order neighborhood structure in Section 2.1. The predictor is $\widehat{X}_i^{\,2,S}(t)$ as in Section 2.1.

ST1. **Spatial-temporal model with the first order neighborhood structure**

We predict the missing data using spatial-time model with the first order neighborhood structure in Section 2.2. The predictor is $\widehat{X}_i^{\,1,ST}(t)$ as in Section 2.2.

ST2. **Spatial-temporal model with the second order neighborhood structure**

We predict the missing data using spatial-time model with the second order neighborhood structure in Section 2.2. We denote the predictor by $\widehat{X}_i^{\,2,ST}(t)$.

	A	T	S1	ST1	S2	ST2
CRV	7.24	7.13	6.84	6.78	6.24	6.19
ratio of CRV(%)	100	98.6	94.5	93.7	86.3	85.5

Table1.: Comparison of cross-Validation statistics of 6 models

For the comparison, we use the cross-validation technique for the above 6 models. If we let $X_{i,d}(t)$, $i = 1, \cdots, 64$, $t = 1, \cdots, 96$, $d = 1, \cdots, 16$, be the observation at t time and i link on d-th day, the cross-validation is implemented by removing the data $X_{i,d}(t)$ and predicting it from the remaining data. Here we use the cross-validation statistics (CRV) in Huang and Cressie [10]

as follows:

$$CRV_{L,i} = \left[\frac{1}{\sum_{d=1}^{16} |M_{i,d}|} \sum_{d=1}^{16} \sum_{t=1, t \in M_{i,d}}^{96} [X_{i,d}(t) - \hat{X}_d^{(-i,-t)}]^2 \right]^{\frac{1}{2}}, \quad (17)$$

$$CRV = \left[\frac{1}{\sum_{d=1}^{16} \sum_{l=1}^{64} |M_{i,d}|} \sum_{d=1}^{16} \sum_{l=1}^{64} \sum_{t=1, t \in M_{i,d}}^{96} [X_{i,d}(t) - \hat{X}_d^{(-i,-t)}]^2 \right]^{\frac{1}{2}} \quad (18)$$

where $\hat{X}_d^{(-i,-t)}$ denotes the predictor of $X_{i,d}(t)$ by removing $X_{i,d}(t)$ and $M_{i,d}, |M_{i,d}|$ are the set of times when observations are not missing at link i on d-th day and the number of elements of $M_{i,d}$, respectively.

In Table 1, we summarized the CRV values of all those methods and the ratios of the CRV of each method to that of Method A. Table 1 shows that the spatial-temporal model using the second order neighborhood structure, is better than the others. In average, there was about 14% improvement in CRV when using the spatial-temporal approach with second order neighborhood structure. This strongly recommends to use a spatial-temporal model for handling missing values.

4. Concluding Remarks

In this paper, we proposed an algorithm using the spatial-temporal model to reconstruct the missing values. A significant improvement in CRV was seen in using the spatial model compared with a time series model. As we can see in Section 3, the spatial-temporal model outperforms the pure spatial model, but the difference of improvement in CRV between two models is not remarkable. This indicates that spatial effects dominate temporal effects in our traffic data analysis, and the concept of spatial modeling should be taken into consideration. To our knowledge, our work is the first attempt to use spatial-temporal model in ITS related fields in Korea. We believe that the result obtained here gives a functional tool to solve the missing value problem.

Acknowledgements. We are grateful to the ROTIS Inc. for providing data and relevant traffic informations. We also wish to thank the editor and the two referees. This work was supported by Korea Research Foundation Grant (KRF-99-042-D0021 D1200).

References

1. A.C. Harvey (1993). *Time Series models*. The MIT press, Cambridge, Messachusetts.
2. Ali, M. M. (1979). Analysis of stationary spatial-temporal processes : Estimation and prediction. *Biometrika*, **66**, 513-518.
3. Banks, J. H. (1997) *Introduction to Transportation Engineering*. New-York, McGraw-Hill.
4. Besag, J. (1974). Spatial interactions and the statistical analysis of lattice data. *Journal of the Royal Statistical Society* B, **36**, 192-225.
5. Casettie, E. and Semple, R. K. (1969). Concerning the testing of spatial diffusion hypotheses. *Geographical Analysis*, **1**, 254-259.
6. Cressie, N. (1993). *Statistics for Spatial Data*. Revised edition, New York, Wiley and Sons.
7. Cressie, N. and Majure, J. J. (1997). Spatio-Temporal Statistical Modeling of Livestock Waste in Streams. *Journal of Agricaltural, Biological, and Environmental Statistics*, **2**, 24-47.
8. Cressie, N., Kaiser, M. S., Daniels, M. J.,Aldworth, J., Lee, J., Lahiri, S. N., and Cox, L. H. (1998). Spatial analysis of particulant matter in an urban environment. Unpublished manuscript.
9. Cliff, A. D. and Ord, J. K. (1975). Space-time modeling with an application to regional forecasting. *Transactions and Papers, Institution of British Geographers*, **66**, 119-128.
10. Huang, H. C. and Cressie, N. (1996). Spatio-temporal prediction of snow water equivalent using the Kalman filter. *Computational Statistics and Data Analysis*, **22**, 159-175.
11. Niu, X. and Tiao, G. C. (1995). Modeling satellite ozone data. *Journal of the American Statistical Association*, **90**, 969-983.
12. Pfeifer, P. E. and Deutsch, S. J. (1980). A three-stage iterative procedure for space-time modeling. *Technometrics*, **22**, 35-47.
13. Pfeifer, P. E. and Deutsch, S. J. (1980). Identification and interpretation of first order space-time ARMA models. *Technometrics*, **22**, 397-408.
14. Deutsch, S. J. and Pfeifer, P. E. (1981). Space-time ARMA modeling with contemporaneously correlated innovations. *Technometrics*, **23**, 401-409.

[2002년 10월 접수, 2002년 10월 채택]