Applicability of the Spatial Aggregation Model for Small Area Interpolation

Toshiaki TSUDA[1], Makoto TSUKAI[2]

Abstract

Recently, spatial data set for fine zone scale becomes more important in urban planning. However, social or economic activity data such as transportation or production output are often difficult to get in the fine zone scale, due to the difficulties in survey. Comparing to these data, demographic data obtained in national census is able to easily get, so then an area interpolation by a regression approach can be applied to estimate the value of objective variable in fine zone. On the other hand, Modifiable Areal Unit Problem (MAUP), often cause a serious problem such as biased estimation of the statistical model parameter due to inadequate treatment for spatial dependencies in the spatial dependency matrix. This study develops a novel spatial econometric model with areal interpolation by using spatial aggregation matrix and spatial decomposition matrix. The proposed model fulfills the pycnophylactic property (i.e. spatial aggregation condition) between fine and source zone.

We checked the performance of the proposed model by Monte Carlo simulation (MC) using hypothetical spatial data with the known value of parameters and the controlled error term. Moreover, we empirically examined the applicability of the proposed model for small area interpolation using the dataset in inter-regional passenger traffic. The estimated model proved that the proposed model can perform well for small area interpolation.

Keyword: spatial econometric models, spatial aggregation matrix, pycnophylactic property

JEL codes: C13(Estimation), C15(Statistical Simulation Methods), C51(Model Construction and Estimation), C52(Model Evaluation, Validation, and Selection)

[1] Graduate School of Engineering, Hiroshima University, 4-1, Kagamiyama 1 chome, Higashi-Hiroshima 739-8527, Japan. E-mail: m121388@hiroshima-u.ac.jp

[2] Graduate School of Engineering, Hiroshima University, 4-1, Kagamiyama 1 chome, Higashi-Hiroshima 739-8527, Japan. E-mail: mtukai@hiroshima-u.ac.jp

1 Introduction

Recently, geo-reference between different spatial data source becomes easier due development of GIS. In addition, we can get finer zone scale data, which would contribute to urban and regional planning. On the other hand, social or economic activity data such as transportation or production output are not available in the fine zone scale, due to the difficulties in survey. Thus, following case often occurs that analyst is only available for larger zone scale data than they really need. In such case, areal interpolation is applied for estimating fine zone scale data. Though many method of areal interpolation have been proposed so far, intelligent method referring to spatial property is accumulated in this field following to the expansion of geo-statistics (Sadahiro, 2000). For example, zonal data in demographic characteristics obtained in national census becomes easier to get in fine zone scale than that in economic activity or transportation, so then an area interpolation by a regression approach can be applied to estimate the objective variable in fine zone from the larger zone. Since the spatial data have a scale dependent property, interdependency between neighbor zones and heterogeneity of data differently appear depending on the spatial aggregation level (zone size). The former is called as "the first row of geography" (Tolber, 1970), and the latter is called as "Modifiable Areal Unit Problem (MAUP)" (Openshow, 1984). In the spatial econometric model, MAUP often causes a serious problem such as biased estimation in model parameter due to inadequate treatment in error term covariance or heteroscedasticity of variance.

This study develops a novel spatial econometric model with areal interpolation by using spatial aggregation matrix and spatial decomposition matrix. These matrices satisfy the pycnophylactic property about the spatial consistency of the data between fine (small) and source (large) zone.

2 Spatial aggregation model

2.1 Spatial aggregation matrix

In the following analysis, the spatial data for independent variables is only available in small (fine) scale zones (the number of zones is N), while the dependent variable is only available in large scale zone. Suppose the number of available data for dependent variable be M, and that for independent variable be N, so then $M < N$ is fulfilled. Hereafter, S and L denote small scale and large scale, respectively.

One of the most important consistencies required in spatial econometric models is the pycnophylactic property (Tolber, 1970). The property in a regression model is that the predicted value of dependent variable in large scale zones, calculated by aggregation of small zone scale, is identical to the observed value of large zone scale. Such property is expressed in Eq. (1).

$$y_l^L = \sum_j^N b_{ij} \hat{y}_{i \in l}^S \tag{1}$$

where, y_l^L is the observed value of large zone scale l, $\hat{y}_{i \in l}^S$ is the predicted value of small zones belonging to the large zone l, and b_{ij} is an element of the spatial aggregation matrix.

The spatial aggregation matrix $\boldsymbol{B}$ is $M \times N$ matrix, whose rows and columns is large zones and small zones, respectively b_{ij} is 1 if a small zone belongs to a large zone, otherwise b_{ij} is 0.
The spatial aggregation matrix $\boldsymbol{B}$ is described as Eq. (2).

$$\boldsymbol{B} = \begin{bmatrix} 1 & 1 & \cdots & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & 1 & 1 \end{bmatrix} \tag{2}$$

b_{ij} satisfies Eq. (3) and Eq. (4).

$$\sum_{i=1}^{M} b_{ij} = 1 \tag{3}$$

$$\sum_{j=1}^{N} b_{ij} = m_l \tag{4}$$

Where, m_l is the number of small zones belonging to the large zone l.

2.2 Model specification

Spatial econometric models can be classified into several types. For example, a spatial autoregressive model (SAR) considers the spatial correlation of dependent variables, and a spatial moving average model (SAM) considers error terms spatial correlation (Anselin, 1988).

Manski proposed a generalized model called in Manski model including the aboves as variants (Elhorst, 2010a). The Manski model is formulated in Eq. (5).

$$\begin{cases} \boldsymbol{Y} = \rho \boldsymbol{W}\boldsymbol{Y} + \boldsymbol{\gamma} + \boldsymbol{X}\boldsymbol{\beta} + \varphi \boldsymbol{W}\boldsymbol{X} + \boldsymbol{u} \\ \boldsymbol{u} = \lambda \boldsymbol{W}\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon} \end{cases} \tag{5}$$

Where, ρ, φ and λ are spatial correlation parameter, respectively, $\boldsymbol{W}$ is the $N \times N$ spatial proximity matrix, $\boldsymbol{\beta}$ is the $k \times 1$ structural parameter vector, $\boldsymbol{\gamma}$ is the $N \times 1$ constant terms vector, and $\boldsymbol{\varepsilon}$ is

the $N \times 1$ error term vector.

In this study, for simplicity, we define φ as Eq. (6) using μ (spatial parameter about independent variables).

$$\varphi = \mu\boldsymbol{\beta} \tag{6}$$

In order to discuss the regression based areal interpolation, we integrate the zone size difference with Manski model. Firstly, Small Zone model (SZ model) is also defined as spatial econometric model using data of small zone. It is formulated in Eq. (7).

$$\boldsymbol{Y}_S = (\boldsymbol{I}_N - \rho\boldsymbol{W}_S)^{-1}\boldsymbol{\gamma}_S + (\boldsymbol{I}_N - \rho\boldsymbol{W}_S)^{-1}(\boldsymbol{I}_N + \mu\boldsymbol{W}_S)\boldsymbol{X}_S\boldsymbol{\beta} + (\boldsymbol{I}_N - \rho\boldsymbol{W}_S)^{-1}(\boldsymbol{I}_N + \lambda\boldsymbol{W}_S)\boldsymbol{\varepsilon}_S \tag{7}$$

The spatial aggregation model is a regression model with an aggregation from *N* zones to *M* zones, multiplying both sides in Eq. (7) by $\boldsymbol{B}$. It is formulated in Eq. (8).

$$\begin{aligned} \boldsymbol{Y}_L &= \boldsymbol{B}\boldsymbol{Y}_S \\ &= \boldsymbol{B}(\boldsymbol{I}_N - \rho\boldsymbol{W}_S)^{-1}\boldsymbol{\gamma}_S + \boldsymbol{B}(\boldsymbol{I}_N - \rho\boldsymbol{W}_S)^{-1}(\boldsymbol{I}_N + \mu\boldsymbol{W}_S)\boldsymbol{X}_S\boldsymbol{\beta} + \boldsymbol{B}(\boldsymbol{I}_N - \rho\boldsymbol{W}_S)^{-1}(\boldsymbol{I}_N + \lambda\boldsymbol{W}_S)\boldsymbol{\varepsilon}_S \end{aligned} \tag{8}$$

Where, $\boldsymbol{I}_N$ is the $N \times N$ identity matrix. For simplicity, we define Eq. (9).

$$\begin{cases} \boldsymbol{A} = \boldsymbol{I} - \rho\boldsymbol{W} \\ \boldsymbol{F} = \boldsymbol{I} + \lambda\boldsymbol{W} \\ \boldsymbol{G} = \boldsymbol{I} + \mu\boldsymbol{W} \end{cases} \tag{9}$$

Eq. (8) can be simplified by using Eq. (9).

$$\begin{aligned} \boldsymbol{Y}_L &= \boldsymbol{B}\boldsymbol{Y}_S \\ &= \boldsymbol{B}\boldsymbol{A}_S^{-1}\boldsymbol{\gamma}_S + \boldsymbol{B}\boldsymbol{A}_S^{-1}\boldsymbol{G}_S\boldsymbol{X}_S\boldsymbol{\beta} + \boldsymbol{B}\boldsymbol{A}_S^{-1}\boldsymbol{F}_S\boldsymbol{\varepsilon}_S \end{aligned} \tag{10}$$

2.3 Model estimation procedure

Eq. (10) cannot be estimated by OLS due to the existence of heterogeneous variance and correlation of errors, so then the maximum likelihood estimation (MLE) is required for estimating Eq. (10). The log-likelihood function is obtained in Eq. (11).

$$\ln L = -\frac{M}{2}\ln(2\pi) - \frac{M}{2}\ln(\sigma^2) - \frac{1}{2}\ln\left|\left(\boldsymbol{B}\boldsymbol{A}_S^{-1}\boldsymbol{F}_S\right)\left(\boldsymbol{B}\boldsymbol{A}_S^{-1}\boldsymbol{F}_S\right)'\right| - \frac{1}{2\sigma^2}\left(\boldsymbol{F}_S^{-1}\boldsymbol{A}_S\boldsymbol{B}^{-1}\boldsymbol{B}\boldsymbol{A}_S^{-1}\boldsymbol{F}_S\boldsymbol{\varepsilon}_S\right)'\left(\boldsymbol{F}_S^{-1}\boldsymbol{A}_S\boldsymbol{B}^{-1}\boldsymbol{B}\boldsymbol{A}_S^{-1}\boldsymbol{F}_S\boldsymbol{\varepsilon}_S\right) \tag{11}$$

The spatial decomposition matrix $\boldsymbol{B}^{-1}$ appearing in Eq. (11) decomposes $\boldsymbol{Y}_L$ into $\boldsymbol{Y}_S$. The element of spatial decomposition matrix b^{ij} is required to satisfy Eq. (12).

$$\sum_{i=1}^{N} b^{ij} = 1 \quad \forall j \tag{12}$$

b^{ij} is specified in Eq. (13), considering Eq. (12).

$$b^{ij} = \frac{\hat{y}_{i\in l}^{S}(\rho,\lambda,\mu,\boldsymbol{\beta})}{\sum_{i=1}^{m_l}\hat{y}_{i\in l}^{S}(\rho,\lambda,\mu,\boldsymbol{\beta})} \tag{13}$$

The predicted dependent value in small zones $\hat{\boldsymbol{Y}}_S$ can be calculated in following two ways. One is using $\boldsymbol{B}^{-1}$, which is formulated as the function of $\hat{\rho}$, $\hat{\lambda}$, $\hat{\mu}$ and $\hat{\boldsymbol{\beta}}$.

$$\hat{\boldsymbol{Y}}_S = \boldsymbol{B}^{-1}\left(\hat{\rho},\hat{\lambda},\hat{\mu},\hat{\boldsymbol{\beta}}\right)\boldsymbol{Y}_L \tag{14a}$$

The other is to directly calculate the small zone value from SZ model as Eq. (14b).

$$\hat{\boldsymbol{Y}}_S = (\boldsymbol{I}_N - \hat{\rho}\boldsymbol{W}_S)^{-1}\boldsymbol{\gamma}_S + (\boldsymbol{I}_N - \hat{\rho}\boldsymbol{W}_S)^{-1}(\boldsymbol{I}_N + \hat{\mu}\boldsymbol{W}_S)\boldsymbol{X}_S\hat{\boldsymbol{\beta}} \tag{14b}$$

3 Monte Carlo simulation using hypothetical spatial data

In this section, we check a performance of the proposed model by Monte Carlo simulation (MC) using hypothetical spatial data under the known value of parameters and the controlled error term.

3.1 Generating of hypothetical spatial data

We set a hypothetical spatial mesh area, and distance of adjacency mesh is 1 for all. In this simulation, let the small zone be 900(N = 30×30 = 900), which is original small zone, and let the large zone be 60 (M = 60), which is obtained by aggregating small zone. In addition, we set the generation model for

hypothetical data in small zone as following by SAR type in Eq. (15).

$$\boldsymbol{Y}_S = \rho \boldsymbol{W}_S \boldsymbol{Y}_S + \boldsymbol{X}_S \boldsymbol{\beta} + \boldsymbol{\gamma}_S + \boldsymbol{\varepsilon}_S \tag{15}$$

Generation of hypothetical data process is shown below.

i. First, we generate independent variables, with spatial correlation. Independent variables are calculated by Eq. (16).

$$\boldsymbol{X}_{S0} = (\boldsymbol{I}_N - \rho_0 \boldsymbol{W}_S)^{-1} \boldsymbol{Z} \tag{16}$$

Where, ρ_0 is spatial correlation parameter, **Z** is the random variable matrix with $N \times k$. Using $\boldsymbol{X}_{S0}$, we generate the expected value of dependent variable by Eq. (17).

$$\boldsymbol{Y}_{S0} = (\boldsymbol{I}_N - \rho_0 \boldsymbol{W}_S)^{-1} \boldsymbol{X}_{S0} \boldsymbol{\beta}_0 + (\boldsymbol{I}_N - \rho_0 \boldsymbol{W}_S)^{-1} \boldsymbol{\gamma}_0 \tag{17}$$

Where, $\boldsymbol{\beta}_0$ is the $k \times 1$ structural parameter vector, $\boldsymbol{\gamma}_0$ is the $N \times 1$ constant term vector.

ii. Secondly, we add error term, with spatial correlation. Error term is calculated by Eq. (18).

$$\boldsymbol{\varepsilon}_{S0} = (\boldsymbol{I}_N - \rho_0 \boldsymbol{W}_S)^{-1} \boldsymbol{v} \tag{18}$$

Where, $\boldsymbol{v}$ is the $N \times 1$ random vector.

iii. Finally, we obtain hypothetical spatial data with errors as the true value of dependent variable.

$$\boldsymbol{Y}_S = \boldsymbol{Y}_{S0} + \boldsymbol{\varepsilon}_{S0} \tag{19}$$

Using above data, we check the performance of proposed spatial model.

3.2 Model diagnostic indices

This section summaries some diagnostic indices for the estimated model.

3.2.1 Test for spatial dependency

For testing the spatial dependency based on the estimated residuals, Moran's I statistic is applied. Moran's I statistics is formulated in Eq. (20).

$$\tilde{M} = \frac{\hat{\boldsymbol{\varepsilon}}\left\{\frac{1}{2}\left(\boldsymbol{W}' + \boldsymbol{W}\right)\right\}\hat{\boldsymbol{\varepsilon}}'}{\hat{\boldsymbol{\varepsilon}}'\,\hat{\boldsymbol{\varepsilon}}} \tag{20}$$

Where, $\hat{\boldsymbol{\varepsilon}}$ is error terms, $\boldsymbol{W}$ is spatial proximity matrix. Eq. (20) standardizes Eq. (21) for hypothesis testing.

$$M = \frac{\tilde{M} - E\left[\tilde{M}\right]}{\sqrt{Var\left[\tilde{M}\right]}} \tag{21}$$

In this study, we use M in Eq. (21) for testing spatial dependency. In this simulation, we test four types of spatial dependency: true error term in small zone which is calculated by Eq. (18), true error term in large zone which is aggregating Eq. (18) to large zone by using spatial aggregation matrix $\boldsymbol{B}$, small zone residual, residual in large zone which is aggregating small zone residual to large zone by using spatial aggregation matrix $\boldsymbol{B}$. Moran's I statistics for true error term $\boldsymbol{\varepsilon}_{S0}$ is obtained by Eq. (22).

$$M_0 = \frac{\tilde{M}_0 - E\left[\tilde{M}_0\right]}{\sqrt{Var\left[\tilde{M}_0\right]}} \quad \text{where, } \tilde{M}_0 = \frac{\boldsymbol{\varepsilon}_{S0}\left\{\frac{1}{2}\left(\boldsymbol{W}_S' + \boldsymbol{W}_S\right)\right\}\boldsymbol{\varepsilon}_{S0}'}{\boldsymbol{\varepsilon}_{S0}'\,\boldsymbol{\varepsilon}_{S0}} \tag{22}$$

Moran's I statistics for small zone residuals $\hat{\boldsymbol{\varepsilon}}_S\left(=\boldsymbol{Y}_S - \hat{\boldsymbol{Y}}_S\right)$ is obtained from Eq. (23).

$$M_1 = \frac{\tilde{M}_1 - E\left[\tilde{M}_1\right]}{\sqrt{Var\left[\tilde{M}_1\right]}} \quad \text{where, } \tilde{M}_1 = \frac{\hat{\boldsymbol{\varepsilon}}_S\left\{\frac{1}{2}\left(\boldsymbol{W}_S' + \boldsymbol{W}_S\right)\right\}\hat{\boldsymbol{\varepsilon}}_S'}{\hat{\boldsymbol{\varepsilon}}_S'\,\hat{\boldsymbol{\varepsilon}}_S} \tag{23}$$

Moran's I statistics for true error term in large zone $\boldsymbol{\varepsilon}_{L0}\left(=\boldsymbol{B}\boldsymbol{\varepsilon}_{S0}\right)$ is obtained from Eq. (24).

$$M_{0a} = \frac{\tilde{M}_{0a} - E\left[\tilde{M}_{0a}\right]}{\sqrt{Var\left[\tilde{M}_{0a}\right]}} \quad \text{where, } \tilde{M}_{0a} = \frac{\boldsymbol{\varepsilon}_{L0}\left\{\frac{1}{2}\left(\boldsymbol{W}_L' + \boldsymbol{W}_L\right)\right\}\boldsymbol{\varepsilon}_{L0}'}{\boldsymbol{\varepsilon}_{L0}'\,\boldsymbol{\varepsilon}_{L0}} \tag{24}$$

Moran's I statistics for residual in large zone $\hat{\boldsymbol{\varepsilon}}_L\left(=\boldsymbol{B}\hat{\boldsymbol{\varepsilon}}_S\right)$ is obtained from Eq. (25).

$$M_{1a} = \frac{\tilde{M}_{1a} - E\left[\tilde{M}_{1a}\right]}{\sqrt{Var\left[\tilde{M}_{1a}\right]}} \quad \text{where, } \tilde{M}_{1a} = \frac{\hat{\boldsymbol{\varepsilon}}_L \left\{\frac{1}{2}\left(\boldsymbol{W}_L' + \boldsymbol{W}_L\right)\right\} \hat{\boldsymbol{\varepsilon}}_L'}{\hat{\boldsymbol{\varepsilon}}_L' \hat{\boldsymbol{\varepsilon}}_L} \tag{25}$$

3.2.2 Model fitness

The adjusted determination coefficient for reproducibility of regression model is formulated Eq. (26).

$$R^2 = 1 - \frac{\sum_i \left(y_i - \hat{y}_i\right)^2 / (N-k-1)}{\sum_i \left(y_i - \bar{y}\right)^2 / (N-1)} \tag{26}$$

Where, y_i is observed value, $\hat{y}_i$ is predicted value, and $\bar{y}$ is average of observed value. Adjusted determination coefficient is calculated for 4 ways. The first one is between $\boldsymbol{Y}_{S0}$ and $\boldsymbol{Y}_S$ which is obtained in Eq. (27).

$$R_0^2 = 1 - \frac{\sum_i \left(y_i^{S0} - y_i^S\right)^2 / (N-k-1)}{\sum_i \left(y_i^{S0} - \bar{y}^{S0}\right)^2 / (N-1)} \tag{27}$$

The second one is between $\boldsymbol{Y}_S$ and $\hat{\boldsymbol{Y}}_S$ which is obtained in Eq. (28).

$$R_1^2 = 1 - \frac{\sum_i \left(y_i^S - \hat{y}_i^S\right)^2 / (N-k-1)}{\sum_i \left(y_i^S - \bar{y}^S\right)^2 / (N-1)} \tag{28}$$

The third one is between $\boldsymbol{Y}_{L0}$ and $\boldsymbol{Y}_L$ which is obtained in Eq. (29).

$$R_{0a}^2 = 1 - \frac{\sum_i \left(y_i^{L0} - y_i^L\right)^2 / (M-k-1)}{\sum_i \left(y_i^{L0} - \bar{y}^{L0}\right)^2 / (M-1)} \tag{29}$$

The final one is between $\boldsymbol{Y}_L$ and $\hat{\boldsymbol{Y}}_L$ which is obtained in Eq. (30).

$$R_{1a}^2 = 1 - \frac{\sum_i \left(y_i^S - \hat{y}_i^S\right)^2 / (M-k-1)}{\sum_i \left(y_i^S - \bar{y}^S\right)^2 / (M-1)} \tag{30}$$

In addition, we use the log-likelihood for fitness index of model estimation. The log-likelihood of spatial aggregation model and that of general spatial econometric model is obtained in Eq. (11), Eq. (31), respectively.

$$\ln L = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) + \ln|\boldsymbol{A}| - \ln|\boldsymbol{F}| - \frac{1}{2\sigma^2}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \tag{31}$$

3.3 Results of Monte Carlo simulation

We check the stability of parameter by Monte Carlo simulation. In this simulation, we set three independent variables (k = 3) and give true value of each parameter as $\rho_0 = 0.5$, $\boldsymbol{\beta}_0' = (0.1, 0.2, 0.3)$, $\gamma_0 = 10$, respectively. Moreover, we estimate the following models as each model type: SARMA (μ = 0), SAR (λ = 0, μ = 0), OLS (ρ = 0, λ = 0, μ = 0) by MLE, with the dataset made in Small Zone and Large Zone, respectively. In addition, we adjust R_0^2 to be 0.7 by controlling the variance of error term. Table 1 to table 3 show the results of simulation. First of all, the value of R_0^2 about all models mostly accord with the designed true value. Therefore, this simulation seems successful. We consider the result of simulation about each type of model.

At first, consider about SAR type. In the spatial aggregation model, although spatial correlation parameter appears under estimation, structural parameters close to the true value. Moreover, R_{0a}^2 and R_{1a}^2 show high value and the log-likelihood is the highest of the three models. On the other hand, M_1 shows that spatial dependency in residual is still left. In SZ model, parameters are almost same with the true value. Moreover, M_1 shows that this model is able to remove spatial dependency in residual. In LZ model, parameters are under-estimated and the constant shows different sign with SZ model. Therefore, we can conclude that it is difficult to correctly reproduce the true value when we predict small zone data from the estimated parameters in LZ model.

Next, consider about SARMA type. In the SARMA type, we annoyed the non-convergence of spatial correlation parameter λ. Specifically, non-convergence in parameter estimation occurs 25 times in the spatial aggregation model, 14 times in the SZ model, and 2 times in the LZ model, respectively. In the following results in table 2, these are excluded from simulation results. Although the sign of spatial correlation parameter λ in spatial aggregation model is different from that in SZ model, other parameter estimation and result of model diagnostic indices are similar to SAR type.

Finally, consider about OLS (non-spatial) type. In the spatial aggregation model, although structural parameters close to the true value, the gap is bigger than that of SAR and SARMA. In the SZ model, structural parameters mostly accord with true value. On the other hand, judging from M_1, spatial dependency is still left in residuals. Tendency in under or over estimation in parameters or diagnostic indices in LZ model is similar to SAR type.

To sum up, the proposed model shows higher fitness in each type: SAR, SARMA, and OLS. Besides, the model performance with spatial dependency such as SAR and SARMA is superior to non-spatial one.

Table 1: Monte Carlo simulation for SAR

parameter	spatial aggregation (SA) model		small zone (SZ) model		large zone (LZ) model	
	average	variance	average	variance	average	variance
ρ [0.5]	0.326	3.36E-02	0.476	8.58E-03	0.062	1.56E-04
λ	-	-	-	-	-	-
β_1 [0.1]	0.116	8.34E-04	0.099	5.47E-05	0.395	9.41E-04
β_2 [0.2]	0.221	1.99E-03	0.201	8.17E-05	0.967	1.08E-03
β_3 [0.3]	0.314	1.37E-03	0.300	9.31E-05	1.044	1.23E-03
γ [10]	16.050	4.17E+01	10.949	1.30E+01	-26.763	5.57E+01
σ	1.573	3.43E-02	1.517	1.27E-03	26.366	1.30E+00
M_0	4.845	3.50E+00	4.573	3.79E+00	4.674	3.77E+00
M_1	42.645	1.77E+01	0.010	6.20E-01	-	-
M_{0a}	1.041	1.24E+00	-	-	1.078	1.13E+00
M_{1a}	0.130	4.96E-01	-	-	0.589	9.59E-02
R_0^2 [0.7]	0.695	1.87E-04	0.693	2.33E-04	0.695	2.68E-04
R_1^2	0.712	2.57E-04	0.694	2.26E-04	-	-
R_{0a}^2	0.999	1.91E-08	-	-	0.999	1.49E-08
R_{1a}^2	0.999	1.78E-08	-	-	0.992	4.40E-07
$L.L$	-0.216	4.71E-05	-1.838	5.46E-04	-4.690	1.85E-03
Samples	60		900		60	

Note:

[ ] : true value

ρ : spatial correlation parameter about dependent variables

λ : spatial correlation parameter about error terms

β_k : structural parameter

γ : constant parameter

σ : standard deviation parameter about error of small scale zones

M_0 : Moran's I statistic for true error (Eq. (22))

M_1 : Moran's I statistic for predicted error (Eq. (23))

M_{0a} : Moran's I statistic for true error (Eq. (24))

M_{1a} : Moran's I statistic for predicted error (Eq. (25))

R_0^2 : adjusted determination coefficient (Eq. (27))

R_1^2 : adjusted determination coefficient (Eq. (28))

R_{0a}^2 : adjusted determination coefficient (Eq. (29))

R_{1a}^2 : adjusted determination coefficient (Eq. (30))

$L.L$: the log-likelihood

Table 2: Monte Carlo simulation for SARMA

parameter	spatial aggregation (SA) model		small zone (SZ) model		large zone (LZ) model	
	average	variance	average	variance	average	variance
ρ [0.5]	0.364	4.30E-02	0.477	9.92E-03	0.055	1.26E-04
λ	-0.113	9.53E-02	0.010	4.69E-02	0.273	3.57E-02
β_1 [0.1]	0.112	9.72E-04	0.098	3.16E-05	0.391	8.14E-04
β_2 [0.2]	0.219	1.66E-03	0.198	5.59E-05	0.992	1.23E-03
β_3 [0.3]	0.313	1.48E-03	0.301	7.80E-05	1.021	1.36E-03
γ [10]	14.647	5.29E+01	10.946	1.53E+01	-22.374	4.75E+01
σ	1.617	4.06E-02	1.506	9.55E-04	26.463	1.27E+00
M_0	5.142	3.29E+00	5.021	3.22E+00	4.539	3.48E+00
M_1	44.412	7.31E-01	0.177	3.14E-03	-	-
M_{0a}	1.119	1.07E+00	-	-	0.996	1.07E+00
M_{1a}	0.155	5.22E-01	-	-	-0.049	2.72E-02
R_0^2 [0.7]	0.694	2.34E-04	0.695	1.88E-04	0.692	1.87E-04
R_1^2	0.711	3.89E-04	0.696	1.90E-04	-	-
R_{0a}^2	0.999	1.11E-08	-	-	0.999	3.44E-07
R_{1a}^2	0.999	1.43E-08	-	-	0.992	4.43E-07
L.L	-0.215	4.51E-05	-1.831	4.44E-04	-4.685	1.72E-03
Samples	60		900		60	

Table 3: Monte Carlo simulation for SAR OLS (non-spatial)

parameter	spatial aggregation (SA) model		small zone (SZ) model		large zone (LZ) model	
	average	variance	average	variance	average	variance
ρ [0.5]	-	-	-	-	-	-
λ	-	-	-	-	-	-
β_1 [0.1]	0.139	9.79E-04	0.106	6.73E-05	0.401	7.46E-04
β_2 [0.2]	0.268	1.75E-03	0.205	6.80E-05	0.985	1.33E-03
β_3 [0.3]	0.337	9.36E-04	0.307	7.92E-05	1.014	1.72E-03
γ [10]	27.426	9.72E-01	29.455	7.05E-02	8.324	2.36E+00
σ	1.771	3.81E-02	1.554	1.34E-03	26.703	9.91E-01
M_0	4.660	3.18E+00	4.403	2.13E+00	4.411	2.76E+00
M_1	51.048	4.55E+01	5.178	3.54E+00	-	-
M_{0a}	1.112	1.27E+00	-	-	0.965	1.05E+00
M_{1a}	0.518	7.60E-01	-	-	0.880	1.25E-01
R_0^2 [0.7]	0.696	2.18E-04	0.694	2.06E-04	0.694	1.73E-04
R_1^2	0.703	3.04E-04	0.687	1.98E-04	-	-
R_{0a}^2	0.999	1.56E-08	-	-	0.999	2.73E-07
R_{1a}^2	0.999	1.86E-08	-	-	0.992	3.26E-07
L.L	-0.132	5.40E-05	-1.860	5.55E-04	-4.703	1.39E-03
Samples	60		900		60	

4 Empirical analysis

4.1 Summaries in the dataset

We conduct an empirical analysis for checking applicability of the proposed model. In this analysis, we use the dataset available for both small and large scales, in order to check the prediction performance with the true value in small zone. In this analysis, we use a trip attraction dataset obtained in the net passenger traffic survey in Japan (2005). Here, the number of small zone is 194 sub-regions, which covers all Japan except Okinawa or staggered small islands. The number of large zone is the 46 prefectures. The dependent variables Y is the trip attraction of each zone, and the independent variables X are 1) the employees in the private sector, 2) the accommodation and food industries, and 3) the zonal average of generalized cost (including time and fare) to the other areas. In addition, we adopt spatial proximity matrix $\boldsymbol{W}$ which elements are standardized for each row.

4.2 Results of the estimation

In this chapter, we compare SAR type with SARMA type formulated as the spatial aggregation model. In addition, we estimate non-spatial OLS model for additional comparison. The spatial aggregation model of SAR, SARMA and OLS types are formulated in Eq. (32a) ~ (32c), respectively.

$$\boldsymbol{Y}_L = \boldsymbol{B}\boldsymbol{Y}_S = \boldsymbol{B}\boldsymbol{A}_S^{-1}\boldsymbol{\gamma}_S + \boldsymbol{B}\boldsymbol{A}_S^{-1}\boldsymbol{X}_S\boldsymbol{\beta} + \boldsymbol{B}\boldsymbol{A}_S^{-1}\boldsymbol{\varepsilon}_S \tag{32a}$$

$$\boldsymbol{Y}_L = \boldsymbol{B}\boldsymbol{Y}_S = \boldsymbol{B}\boldsymbol{A}_S^{-1}\boldsymbol{\gamma}_S + \boldsymbol{B}\boldsymbol{A}_S^{-1}\boldsymbol{X}_S\boldsymbol{\beta} + \boldsymbol{B}\boldsymbol{A}_S^{-1}\boldsymbol{F}_S\boldsymbol{\varepsilon}_S \tag{32b}$$

$$\boldsymbol{Y}_L = \boldsymbol{B}\boldsymbol{Y}_S = \boldsymbol{B}\boldsymbol{\gamma}_S + \boldsymbol{B}\boldsymbol{X}_S\boldsymbol{\beta} + \boldsymbol{B}\boldsymbol{\varepsilon}_S \tag{32c}$$

The model performance is also compared to the SZ model. The SZ model of SAR type, SARMA type and OLS type are formulated in Eq. (33a) ~ (33c), respectively.

$$\boldsymbol{Y}_S = \boldsymbol{A}_S^{-1}\boldsymbol{\gamma}_S + \boldsymbol{A}_S^{-1}\boldsymbol{X}_S\boldsymbol{\beta} + \boldsymbol{A}_S^{-1}\boldsymbol{\varepsilon}_S \tag{33a}$$

$$\boldsymbol{Y}_S = \boldsymbol{A}_S^{-1}\boldsymbol{\gamma}_S + \boldsymbol{A}_S^{-1}\boldsymbol{X}_S\boldsymbol{\beta} + \boldsymbol{A}_S^{-1}\boldsymbol{F}_S\boldsymbol{\varepsilon}_S \tag{33b}$$

$$\boldsymbol{Y}_S = \boldsymbol{\gamma}_S + \boldsymbol{X}_S\boldsymbol{\beta} + \boldsymbol{\varepsilon}_S \tag{33c}$$

In this analysis, model diagnostic indices are Moran's I statistics, the determination coefficient, the adjusted determination coefficient, the root mean squared error and the log-likelihood. Table 4 shows the results of the parameter estimation in each model.

Table 4: The result of model estimation

	spatial aggregation (SA) model			small zone (SZ) model		
	SAR	SARMA	OLS (non-spatial)	SAR	SARMA	OLS (non-spatial)
parameter	Coefficient	Coefficient	Coefficient	Coefficient	Coefficient	Coefficient
ρ	-0.115	-0.375 +	———	0.005	-0.198 *	———
λ	———	0.490 +	———	———	0.762 **	———
β_1	0.026 **	0.022 **	0.028 **	0.028 **	0.025 **	0.028 **
β_2	-0.241 *	-0.155	-0.287 **	-0.258 **	-0.202 **	-0.257 **
β_3	-2.257 **	-2.842 **	-1.979 **	-1.306	-2.242 +	-1.327
γ	3521.037 **	4859.001 **	2986.119 **	1961.977 +	3427.056 *	1998.107 *
σ	4150.795 **	3967.380 **	4012.427 **	3024.939 **	2900.427 **	3024.970 **
[small zone]						
M.I	11.642	8.198	10.683	5.856	0.955	5.926
R^2	0.940	0.941	0.935	0.887	0.892	0.887
R'^2	0.930	0.930	0.927	0.884	0.888	0.884
RMSE	2267.319	2246.801	2339.107	3027.626	2955.818	3024.970
N.P.	3 / 194	5 / 194	3 / 194	1 / 194	4 / 194	1 / 194
[large zone]						
M.I	3.251	2.580	3.200	3.035	3.078	3.107
R^2	0.859	0.856	0.856	0.859	0.866	0.859
R'^2	0.837	0.830	0.838	0.837	0.842	0.842
RMSE	6826.321	6880.997	6899.538	7048.890	6838.762	7041.331
N.P.	0 / 46	0 / 46	0 / 46	0 / 46	0 / 46	0 / 46
L.L	-2.433	-2.425	-2.434	-9.434	-9.367	-9.434
Samples	46	46	46	194	194	194

+ significant at 10% level, * significant at 5% level, ** significant at 1% level.

Note: *M.I* : Moran's I statistic, R^2 : determination coefficient, R'^2 : adjusted determination coefficient, *RMSE* : root mean squared error, *N.P.* : the number of negatively predicted dependent variables, *L.L* : log-likelihood.

In the SAR type of proposed model, spatial correlation parameter ρ is insignificant even at the 10% level. In addition, spatial dependency still remains in the residuals in both small and large zones. On the other hand, other parameters are significant and the diagnostic indices are better than that of the SZ model. In the SARMA type of proposed model, β_2 is insignificant even at the 10% level. It would be caused by multicollinearity between X_1 and X_2. In addition, even though spatial correlation parameters ρ and λ are significant, spatial dependency still remains in residuals in both small and large zones, while it is not found in residuals in SZ model. Moreover, the diagnostic indices in small zone are superior to other two models. In the OLS type of proposed model, all parameters are significant, while spatial dependency still remains in residuals in both small and large zones. It is as same as other two models.

To sum up, except for M.I, the diagnostic indices show that the proposed model is superior to SZ

model. In the three type of proposed model, SARMA type performs well rather than other two models based on diagnostic indices.

4.3 Reproducibility of proposed model

In this section, we check the reproducibility of the proposed model using SARMA type. The predicted values of each small zone $\hat{\boldsymbol{Y}}_S$ are calculated from Eq. (14a). Figure 1 shows the scatter plot between the observed value $\boldsymbol{Y}_S$ and the predicted values of the small zones $\hat{\boldsymbol{Y}}_S$.

Reproducibility of the proposed model is fairly good. The slope of regression is 0.908, and the correlation coefficient is 0.970. On the other hand, 5 zones are negatively predicted. 4 out of 5 zones are less than 400 trips, and the other is about 3,000 trips in $\boldsymbol{Y}_S$. Therefore, the reproducibility for the small observation is not good. In order to check the model performance, we classify all the zones into two groups, referring to the median of the observed trip. The upper group with more than 1,595 trips is referred as A1, and the other is as A2. Figure 2 shows the scatter plot of A1 in predicted and observed, and Figure 3 shows the plot of A2.

In Figure 2, the reproducibility for A1 is good, and the slope of regression line is 0.906, and the correlation coefficient is 0.968. On the other hand, in Figure 3, reproducibility for A2 is not good, and the slope of regression line is 1.589 so that over estimation appears. The correlation coefficient for A2 is 0.534.

To sum up, the model performance is significantly different whether $\boldsymbol{Y}_S$ belongs to large or small observation group. An adequate treatment for small observations to prevent negative predictions (a constraint) is required.

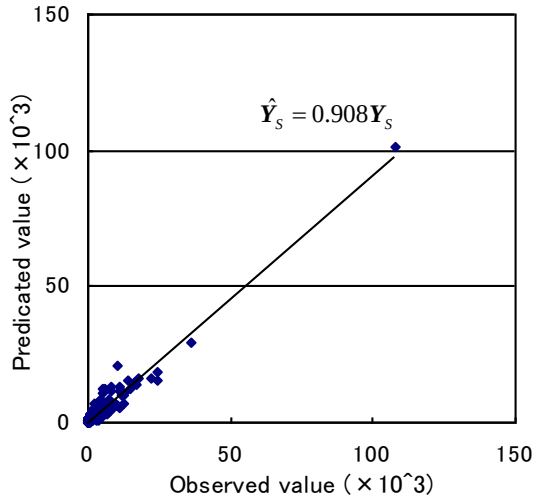

Figure 1: Scatter plot of the predicted and the observed values in the proposed model.

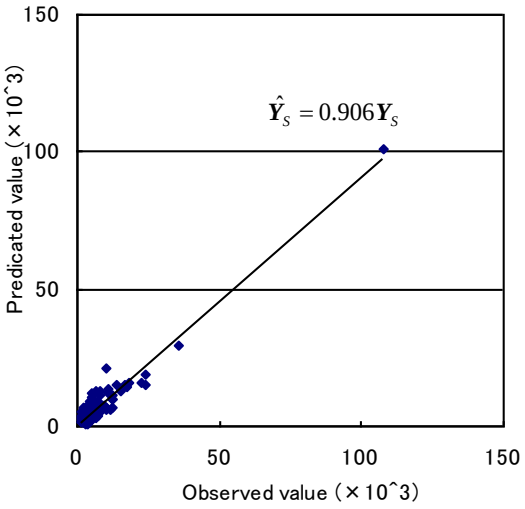

Figure 2: Scatter plot of the predicted and the observed values for group A1

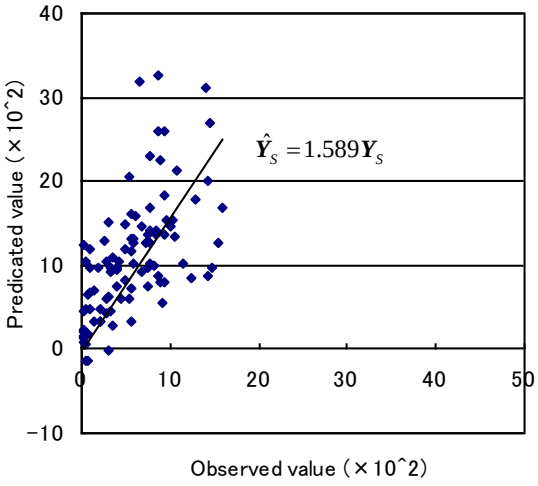

Figure 3: Scatter plot of the predicted and the observed values for group A2

5 Conclusion

In this study, we developed a spatial aggregation model able to make area interpolation. This model is formulated using spatial aggregation matrix and spatial decomposition matrix, that satisfy the pycnophylactic property to fulfill spatial aggregation condition between small and large zone. Further, we checked the performance of the proposed model in two ways; by Monte Carlo simulation (MC) using hypothetical spatial data, and by empirically analysis using trip attraction data of the net passenger traffic survey in Japan. MC showed that the parameters in the proposed model are stable in any variant of Manski-spatial econometric models. Moreover, the model with spatial dependency performs well rather than non-spatial model so then the proposed model can approximate well the data generation process. In empirically analysis, the proposed model enables small area interpolation with high accuracy. The reproducibility of the proposed model is higher than that of small zone model using observed data of dependent variables in small zone. However, the interpolation accuracy strongly depends on the value of dependent variable. The reproducibility is fairly good in a large observation group, while it is not good in a small observation group. Therefore, the issues to be overcome in future study are almost same with the conventional regression analysis, such as multicollinearity, the selection of independent variables and the choice of describing spatial (set of spatial proximity matrix). Tackling these issues will also improve the areal interpolation performance of the proposed approach.

References

Anselin, L. (1988) *Spatial Econometrics: Methods and Models*, Kluwer Academic Publishers, Dordrecht, Netherlands.

Elhorst, J.P. (2010a) “Applied spatial econometrics: Raising the bar,” *Spatial Economic Analysis*, Vol.5, No.1, pp.9-28.

Openshaw, S. (1984) “The modifiable areal unit problem,” Geo-books, CATMOG 38.

Sadahiro, Yukio (2000) “Evaluation of data accuracy estimated by areal interpolation,” *City Planning Review*, No.225, pp. 75-81.

Tobler, W. R. (1970) “A computer movie simulating urban growth in the Detroit region,” *Economic geography*, vol.46, pp.234-240.