Geographically weighted regression approach for origin-destination flows

Kazuki Tamesue[1] **and Morito Tsutsumi**[2]

[1] Graduate School of Information and Engineering, University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8573, Japan
tamesue.kazuki@sk.tsukuba.ac.jp
[2] Faculty of Engineering, Information and Systems, University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8573, Japan
tsutsumi@sk.tsukuba.ac.jp

Abstract: Local regression methodology called geographically weighted regression (GWR) is used in a variety of fields to capture spatial variation or spatial non-stationarity of a regression analysis by allowing parameters to vary across space; however, application to spatial interaction model is much more complex. The complexity comes from the fact of measuring distances between flows, whereas each flows contains two regions. Because an origin-destination flow is a pairwise of origin and destination region, the weighting of both origins and destinations is preferable in order to consider geographical neighbors of flows in the GWR framework. In this paper, we propose GWR for an unconstrained gravity model, in which a weighting kernel is a mixture of origin-based and destination-based kernel. Weighting specification for the model is somewhat equivalent to that of geographically and temporally weighted regression, where a spatio-temporal kernel is constructed with a combination of spatial and temporal weight matrix; therefore the estimation procedure would be carried out in a standard manner of GWR. The estimation result using the interprefecture migration flow data of Japan is examined to see the effectiveness of our proposed method.

Keywords: spatial interaction; gravity model; origin-destination flows; geographically weighted regression

JEL Classification: C21, C51

1 Introduction

Since the development of the gravity model, spatial interaction models are often used for modeling flows between origins and destinations. Even though the gravity model is a classical one and has a long history, it has been widely used in a variety of fields, such as predicting flows in the transportation field and factor analyses in the economic field, and it is still recognized as a strong tool for modeling spatial interaction behavior. As noted by Sen and Smith (1995) in their monograph, the reasons for the popularity of the gravity model include the simplicity of its mathematical form and the intuitive nature of its underlying assumptions.

On the other hand, local regression methodology called geographically weighted regression (GWR) is used in a variety of fields for spatial data analysis to capture spatial variation or spatial nonstationarity of a regression analysis by allowing parameters to vary across space. However, the number of studies on GWR modeling of gravity model is only a few. Nakaya (2001) had applied the GWR approach to gravity model, however, gravity model used in the study was an origin-specific (or origin-constrained) gravity model, and it had only considered geographical weighting of a destination-base kernel.

The objective of this study is to develop the GWR approach for unconstrained gravity model with geographically weighting of origin-based and destination-based kernel. The complexity comes from the fact of measuring distances between flows, whereas each flows contains two regions. Because an origin-destination flow is a pairwise of origin and destination region, the weighting of not only both origins and destinations but also the combination of origin-destination is preferable in order to consider geographical neighbors of flows in the GWR framework. In this paper, we propose GWR for an unconstrained gravity model, in which a weighting kernel is a mixture of origin-based and destination-based kernel. Weighting specification for the model is somewhat equivalent to that of geographically and temporally weighted regression, where a spatio-temporal kernel is constructed with a combination of spatial and temporal weight matrix; therefore the estimation procedure would be carried out in a standard manner of GWR. We also show the validity of the methodology through an empirical application to the interprefectural migration flow data in Japan.

2 Geographically weighted regression

The GWR model can be expressed as follows:

$$y_i = \beta_{i0} + \sum_{k=1}^{p} \beta_{ik} x_{ik} + \varepsilon_i, \tag{1}$$

where y_i is the dependent variable at location i, x_{ik} is the kth explanatory variable at location i, β_{i0} is the intercept parameter at location i, β_{ik} is the coefficient pa-

rameter of kth explanatory variable at location i, and ε_i is the random disturbance at location i.

Therefore, unlike ordinal regression model, GWR allows coefficients to vary across locations. The coefficients are estimated via weighted least squares, where each observation is weighted according to the distance from location i, and matrix notation of the estimators is expressed as

$$\boldsymbol{\beta}(i) = [\mathbf{X}'\mathbf{W}(i)\mathbf{X}]^{-1}\mathbf{X}'\mathbf{W}(i)\mathbf{y} \tag{2}$$

where $\beta(i)$ is the $p+1$ dimensional vectors of local coefficients at location i, $\mathbf{X}$ is the $n \times (p+1)$ matrix of explanatory variables including 1s as the intercept, $\mathbf{y}$ is the n dimensional vector of dependent variables. $\mathbf{W}(i)$ is the $n \times n$ diagonal matrix at location i that put more weights on observations that are closer to the location point

$$\mathbf{W}(i) = \begin{bmatrix} w_{i1} & 0 & \cdots & 0 \\ 0 & w_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{in} \end{bmatrix} \tag{3}$$

which is calculated based on a given kernel function. There are various choices for the kernel functions, such as the bi-square function, the Gaussian function, and the exponential function. This study uses one of those famous and commonly used kernel functions; the Gaussian function

$$w_{ij} = \exp\left\{-\left(\frac{d_{ij}}{b}\right)^2\right\}, \tag{4}$$

where d_{ij} is the distance between location i and j, and b is the bandwidth of the kernel. Other than the coefficient parameter $\boldsymbol{\beta}$, the kernel bandwidth b is also an unknown parameter that has to be estimated in GWR model. The well-known method for finding the kernel bandwidth is by leave-one-out cross-validation. In this method, the bandwidth is estimated that minimizes cross-validation (CV) score

$$\text{CV} = \sum_{i=1}^{n}[y_i - \hat{y}_{\neq i}(b)]^2 \tag{5}$$

where $\hat{y}_{\neq i}(b)$ is a predictor at location i using data excluding location i. A drawback of CV method is that CV is an indicator of goodness-of-fit, and may lead to overfitting of the model. On the other hand, deriving the bandwidth by minimizing the Akaike Information Criterion (AIC) provides a trade-off between goodness-of-fit and degrees of freedom (Fotheringham et al., 2002). The AIC for GWR is:

$$\text{AIC} = 2n\log(\hat{\sigma}) + n\log(2\pi) + n\left\{\frac{n + \text{tr}(\mathbf{S})}{n - 2 - \text{tr}(\mathbf{S})}\right\} \tag{6}$$

where $\hat{\sigma}$ is the estimated standard deviation of the error term

$$\sigma = \sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i) \Big/ (n - 2\mathrm{tr}(\mathbf{S}) + \mathrm{tr}(\mathbf{S}'\mathbf{S}))} \tag{7}$$

and $\mathbf{S}$ is the hat matrix where each row is defined as

$$\mathbf{S}_i = \mathbf{X}_i(\mathbf{X}'\mathbf{W}(i)\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}(i). \tag{8}$$

Once the appropriate bandwidth is estimated, the estimated bandwidth is substituted to equation (4) to construct the weighting matrix at each location, and then local coefficients for each location are estimated with equation (2).

3 GWR for OD flows

The classical gravity model exhibits a flow from origin region i to destination region j with variables that indicate the characteristics of the origin and destination regions, respectively, and the distance between the two regions. The logarithmic transformation of the gravity model produces a least-square linear regression type of the model, as shown in the following equation:

$$\mathbf{y} = \alpha\mathbf{1}_n + \mathbf{X}_o\boldsymbol{\beta} + \mathbf{X}_d\boldsymbol{\gamma} + \theta\boldsymbol{d} + \boldsymbol{\varepsilon} \tag{9}$$

where $\mathbf{X}_o$ and $\mathbf{X}_d$ are $n^2 \times k$ matrices containing k explanatory variables for each origin and destination region, respectively, with associated $k \times 1$ parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Variable $\boldsymbol{d}$ is an $n^2 \times 1$ vector of distances and θ is an associated scalar that is known as a distance-decaying parameter, which reflects the effect of distance. The additional $n^2 \times 1$ vector of elements with the scalar parameter α is used to form a constant term. In this type of linear regression gravity model, the $n^2 \times 1$ disturbance vector $\boldsymbol{\varepsilon}$ is assumed to be *i.i.d.*

A point at issue regarding integration of GWR framework into the gravity model is the construction of the weight matrix $\mathbf{W}$. As explained in the previous section, GWR puts weight on each local model according to the distances between observation points, which can be carried out with ease in case of point data or areal data. In case of flow data, however, one flow observation comprises of two locations—origin and destination regions. As a consequence, definition of distance for flows is to consider both origin-based and destination-based distances, and combine them to form a flow distance. The idea of combining different sets of distance measures to a single distance measure can be often seen in spatio-temporal data analysis, and Huang *et al.* (2013) use this idea to extend GWR to geographically and temporally weighted regression. A point of difference is that whereas location and time are measured in different units in spatio-temporal data, origin-based and destination-based distances are measured in same units, and there is no need to encounter with different scale effects problem.

Following Huang *et al.* (2013), if the Gaussian function is used for construction of weight matrix for flows, we will have:

$$\begin{aligned} w_{ij}^{OD} &= \exp\left\{-\left(\frac{D_{ij}^{O}+D_{ij}^{D}}{b_{OD}}\right)^{2}\right\} \\ &= \exp\left\{-\left(\left(\frac{D_{ij}^{O}}{b_{O}}\right)^{2}+\left(\frac{D_{ij}^{D}}{b_{D}}\right)^{2}\right)\right\} \\ &= \exp\left\{-\left(\frac{D_{ij}^{O}}{b_{O}}\right)^{2}\right\}\times\exp\left\{-\left(\frac{D_{ij}^{D}}{b_{D}}\right)^{2}\right\}. \end{aligned} \tag{10}$$

Therefore, if the origin-based distance D_{ij}^{O} and the destination-based distance D_{ij}^{D} are given, the weight matrix for flows can be expressed as the product of origin-based and destination-based kernels. Now the only question remains to the specification of D_{ij}^{O} and D_{ij}^{D}.

For specifying distances for origin-based and destination-based, we employ the idea of LeSage and Pace (2008), whose interest was to construct spatial weight (contiguity) matrices of spatial econometric model for flow data. When a flow from origin region A to destination region B exists, we consider following flows as *neighbors* of the flow (Figure 1):

(a) Origin-based: flows from neighbor of origin A to destination B.

(b) Destination-based: flows from origin A to neighbor of destination B. and put weights according to distance between regions. Consequently, D_{ij}^{O} is the distance between A and origin of flow j if destinations of flow i and j coincide, and 0 otherwise. Similarly, D_{ij}^{D} is the distance between B and destination of flow j if origins of flow i and j coincide and 0 otherwise.

Slight modification is needed for equation(10) given the above D_{ij}^{O} and D_{ij}^{D}. When flow i and j do not have the same origin or destination, then $D_{ij}^{O}=D_{ij}^{D}=0$ and $w_{ij}^{OD}=1$, meaning that flow i would have a huge weight on flow j even though j are not considered to be a neighbor of flow i.

To avoid above problem, we modify equation (10) as shown below:

$$w_{ij}^{*OD}=\begin{cases} w_{ij}^{OD}, & \text{if } D_{ij}^{O}>0 \text{ or } D_{ij}^{D}>0 \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

Note that the kernel weighting Nakaya(2001) had applied is the special case when we only consider destination-based kernel ($b_O=0$), and will constrain gravity model equation (9) to the origin-specific model:

$$\mathbf{y}=\alpha\mathbf{1}_n+\mathbf{X}_d\boldsymbol{\gamma}+\theta\boldsymbol{d}+\boldsymbol{\varepsilon} \tag{12}$$

Similarly, when we consider origin-based kernel only ($b_D = 0$), the destination-specific model will be applied

$$\mathbf{y} = \alpha \mathbf{1}_n + \mathbf{X}_o \boldsymbol{\beta} + \theta \boldsymbol{d} + \boldsymbol{\varepsilon} \tag{13}$$

In the use of the unrestricted gravity model of euqation (9) in the GWR framework for flows, we implicitly assume $b_O > 0$ and $b_D > 0$.

4 Application

4.1 Data

In this section we present an application of our GWR for OD flows. The OD flow data this study uses is the interprefectural migration flow (logged) in 2006 from the basic resident register migration report of Statistics Bureau of Ministry of Internal Affairs and Communications in Japan. The study has excluded okinawa prefecture since it is an isolated island located far from other prefectures. Therefore, the number of prefectures in this data is 46, which makes the number of observed flows to be $46 \times 45 = 2070$, since we have excluded intraregional flows from the data. For explanatory variables indicating characteristics of origin and destination regions, the following six variables in 2005 are chosen: the log of population, the log of inhabitable land area (km^2), the proportion of the population under age 14 years, the degree of unemployment, the proportion of the tertiary industry employee, and the log of per capita income. We employ the Euclidean distances (logged) between the centers of population for each prefecture as the definition of the distance variable.

4.2 Estimates results

We have estimated four types of models: origin-specific model ($b_O = 0$), destination-specific model ($b_D = 0$), and two unrestricted model considering origin-based and destination-based kernel weighting. One of the unrestricted model assumes the bandwidths of origin-based and destination-based to be equal ($b_O = b_D$), and the other model does not have any restriction to the bandwidths. Note that the origin-specific model is the type of model Nakaya (2001) had used. Table 5 illustrates the estimated bandwidths of the four models in km and their AICs calculated by equation (7). It is clearly shown that AICs of unrestricted models are considerably lower than the origin-specific or the destiantion-specific model, indicating that taking origin and destination-based kernel weighting would substantially improve the model. To see whether the improvement of model is caused only by the increase of explajnatory power due to the increase of explanatory variables, Table 5 also illustrates AICs for each global regression model. From AICs of global models, it is true that the increase of explanatory variables contributes

to the model improvement; however, when we look at the ratios of AIC-global and AIC, the unrestricted models have higher ratios.

It is notable that the bandwidths of two unrestricted models do not vary much. Rather, the restricted bandwidth is approximately the mean of two bandwidths in the unrestricted bandwidths model. Also, the difference of AIC between those two models are negligible but unrestricted bandwidths model has slightly lower AIC. Therefore, when a researcher's interest is on the improvement of the model itself, then restricted bandwidth model may be used since deriving a bandwidth is easier and has less computational burden. On the other hand, unrestricted bandwidths model has the advantage of making an inference on nonstationarity of flows by comparing and examining derived bandwidths of origin-based and destination-based kernels. Table 5 shows parameter estimates of GWR with unrestricted bandwidths and standard global regression model for comparison. Estimates of GWR are shown in quantile summary for simplicity. Some of the variables have high variances, such as under age 14 population and unemployment population for both origin and destination. These might be occurred since a variable is not significant, and in fact those variables are not significant at 5% level on the global model.

5 Conclusion

This paper has considered the GWR framework for unconstrained gravity model. As with the geographically and temporally weighted regression framework of Huang *et al.* (2013), the study has constructed kernel weighting for flows as the combination of both origin and destination kernel weighting. The idea of making spatial weight matrices for spatial econometric model for OD flows (LeSage and Pace, 2008) is employed here for the specification of geographical neighbors of a flow. The application to interprefectural migration flow data in Japan clearly shows that taking both origin-based and destination-based kernel weighting into account would substantially improve the model.

More general specification of distance for flows is needed for future work. This paper only consider flows that have the same origin or destination are considered as neighbors, however, it would be desirable to consider flows from neighbor of origin A to neighbor of destination B. The specification of this type of contiguity is called origin-to-destination based in LeSage and Pace (2008), and it would be of our future work to examine whether this type of kernel weight could be applied in the GWR framework.

References

[1] Fotheringham, A.S., Brunsdon, C., Charlton, M. (2002) *Geographically Weighted Regression*, John Wiley & Sons Ltd, West Sussex.

[2] Huang, B., Wu, B., Barry, M. (2013) Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices, *International Journal of Geographical Information Science*, 24(3), pp.383-401.

[3] LeSage, J.P., Pace, R.K. (2008) Spatial Econometric Modeling of Origin-Destination flows, *Journal of Regional Science*, 48 (5), 941-967.

[4] LeSage, J.P., Fischer, M.M. (2010) Spatial econometric methods for modeling origin-destination flows, in *Handbook of Applied Spatial Analysis* (Eds. M.M. Fishcer, A. Getis), Springer-Verlag, Berlin, pp.409-433.

[5] Nakaya, T. (2001) Local spatial interaction modelling based on the geographically weighted regression approach, *The Geographical Journal*, 53, pp.347-358.

[6] Sen, A., Smith, T.E. (1995) *Gravity Models of Spatial Interaction Behavior*, Springer-Verlag, Berlin.

Table 1: Bandwidth and AIC

	Origin-specific	Destination-specific	Unrestricted	
			b_O=b_D	unrestricted
b_O		367.8529	433.5094	465.0094
b_D	369.3629		433.5094	392.4832
AIC	6005.944	6280.8	1944.076	1942.718
AIC-global	6049.563	6329.074	3612.682	3612.682

Table 2: Parameter estimates

	Local Model					Global Model	
Variable	Min	1st Q	Median	3rd Q	Max	Coefficient	p-value
Intercept	-64.795	-16.252	-5.636	2.964	30.478	-4.162	0.014
Origin_Pop	0.462	0.987	1.115	1.219	1.795	1.104	0.000
Origin_Area	-1.492	-0.085	0.117	0.283	0.996	0.087	0.077
Origin_U15Pop	-65.754	-13.996	-4.149	7.778	58.929	-2.311	0.227
Origin_UnempPop	-106.726	-13.476	6.425	24.206	106.115	7.268	0.041
Origin_Tertiary	-20.990	4.646	10.252	14.525	28.140	10.970	0.000
Origin_Income	-5.367	-1.762	-0.810	0.169	2.256	-1.316	0.000
Destination_Pop	0.443	1.014	1.131	1.243	2.082	1.124	0.000
Destination_Area	-2.031	-0.084	0.160	0.364	1.247	0.104	0.033
Destination_U15Pop	-53.127	-12.138	-2.033	13.166	75.974	2.332	0.223
Destination_UnempPop	-98.669	-19.812	3.265	27.994	111.128	6.978	0.050
Destination_Tertiary	-32.636	5.068	10.056	14.624	31.820	12.005	0.000
Destination_Income	-5.661	-1.375	-0.093	0.978	3.403	-0.610	0.000
Distance	-2.318	-1.489	-1.289	-1.090	-0.154	-1.141	0.000