J. Jpn. Soc. Comp. Statist., **xx**(200x), 1–9

GEOGRAPHICALLY WEIGHTED FUNCTIONAL MULTIPLE REGRESSION ANALYSIS: A NUMERICAL INVESTIGATION

Yoshihiro Yamanishi* **and Yutaka Tanaka†**

ABSTRACT

Functional regression analysis enables us to investigate the relationship among variables over time. Sometimes, however, we meet the case where regression coefficients do not remain fixed over space, when we analyze spatial data. The present paper proposes a method of geographically weighted functional regression analysis to analyze the relationship among variables which varies over space as well as over time, borrowing the idea of Brunsdon *et al.* (1998) in which geographical weight is considered in ordinary regression. Monte Carlo and bootstrap methods are used to perform the statistical test for spatial variability and to evaluate the reliability of the prediction. The proposed methods are illustrated using a real data set.

1. Introduction

Ramsay *et al.* (1991) proposed a method of functional regression analysis and Shimokawa *et al.* (2000) extended it so that it can deal with more than one explanatory variable. These methods enable us to investigate the relationship among the variables over time when they are applied to some sets of time series data. Sometimes we meet the case where regression coefficients do not remain fixed over space when we analyze spatial data. Brunsdon *et al.* (1998) proposed geographically weighted regression (GWR). The geographical weight based on the distance between the observations is introduced, and spatially varying regression coefficients are estimated and used for understanding the spatial variation of the effect of explanatory variables on the response variable. They showed the usefulness of their method as a tool of spatial data analysis by applying it to a census data set. Related to the GWR is Wilhelm *et al.* (1998), who discussed the concept of local statistics, and claimed that local statistics help us to investigate local patterns and to detect unusual observations. It is expected that we can develop a method that enables us to understand the relationship among variables over time and space if the GWR is combined with functional regression. The main purpose in this study is to develop GWR in functional data analysis (FDA) in order to investigate the relationship among the variables not only over time but also over space. In section 2, we review an ordinary functional regression analysis. In section 3, we formulate a geographically weighted functional multiple regression model. In section 4, we define a statistic to assess the spatial variability of regression coefficient functions, and propose a Monte Carlo test to confirm the existence of the spatial variability. Beside that, we propose a bootstrap confidence interval (or curve) to investigate the accuracy of the prediction.

*Graduate School of Natural Science and Technology, Okayama University, 3-1-1, Tsushima-naka, Okayama 700-8530, Japan Address for correspondence: Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan E-mail: yoshi@kuicr.kyoto-u.ac.jp

†Department of Environmental and Mathematical Sciences, Okayama University, 3-1-1 Tsushima-Naka, Okayama 700-8530, Japan

Key words: Functional data; Regression analysis; Spatial data analysis; Spatial non-stationarity

2. Functional regression analysis

2.1. Functional multiple regression model

A functional multiple regression model is formulated as follows:

$$y_i(t) = \beta_0(t) + \sum_{g=1}^{G} \int x_{ig}(s)\beta_g(s,t)ds + \epsilon_i(t) \quad (i = 1, \cdots, N), \tag{1}$$

where $\beta_0(t)$ is a mean function, G is the number of functional covariates, N is the number of observations, $\beta_g(s,t)$ is a regression coefficient function for the g-th covariate function and $\epsilon(t)$ is a random error function. The objective of this model is to predict a functional response $y(t)$ from some functional covariates $x_g(s)$. The model corresponds to that of Simokawa *et al.* (2000). Then $\beta_g(s,t)$ is estimated by minimizing the sum of integrated squared residuals defined by

$$LMISE = \sum_{i=1}^{N} \int \left[y_i(t) - \beta_0(t) - \sum_{g=1}^{G} \int x_{ig}(s)\beta_g(s,t)ds \right]^2 dt. \tag{2}$$

The goodness of fit of the model can be assessed by the squared correlation function defined by

$$R^2(t) = 1 - \sum_{i=1}^{N} \{\hat{y}_i(t) - y_i(t)\}^2 / \{y_i(t) - \bar{y}(t)\}^2. \tag{3}$$

2.2. Algorithm of estimating regression coefficient functions

Suppose that functional data $\{x_{ig}(s)\}_{i=1}^{N}$ and $\{y_i(t)\}_{i=1}^{N}$ can be approximated with finite expansions of sets of basis functions $\phi_g(s) = (\phi_{g1}(s), \cdots, \phi_{gK_\phi}(s))^T$ and $\psi(t) = (\psi_1(t), \cdots, \psi_{K_\psi}(t))^T$. For simplicity, it is assumed that $x_g(s)$, $y(t)$ are centered. Then $x_g(s)$ and $y(t)$ and $\beta_g(s,t)$ can be expanded as

$$x_g(s) = \mathbf{C}_g\phi_g(s), \quad y(t) = \mathbf{D}\psi(t), \quad \beta_g(s,t) = \phi_g(s)^T\mathbf{B}_g\psi(t), \tag{4}$$

where $\mathbf{C}_g$ is an $N \times K_\phi$ coefficient matrix, $\mathbf{D}$ is an $N \times K_\psi$ coefficient matrix, $\mathbf{B}_g$ is a $K_\phi \times K_\psi$ coefficient matrix, and K_ϕ and K_ψ are the numbers of basis functions, respectively. Here for simplicity the number of the terms is common for $\{x_g(s)\}, g = 1, \cdots, G$. Then $LMISE$ defined in eq.(2) is expressed by

$$LMISE = trace\left\{ \left(\mathbf{D} - \sum_{g=1}^{G} \mathbf{C}_g\mathbf{J}_{\phi_g}\mathbf{B}_g\right) \mathbf{J}_\psi \left(\mathbf{D} - \sum_{g=1}^{G} \mathbf{C}_g\mathbf{J}_{\phi_g}\mathbf{B}_g\right)^T \right\}, \tag{5}$$

where $\mathbf{J}_{\phi_g} = \int \phi_g(s)\phi_g(s)^T ds$ and $\mathbf{J}_\psi = \int \psi(t)\psi(t)^T dt$. Choosing $\mathbf{B}_g$ which minimizes $LMISE$ leads to the following equation.

$$(\mathbf{C}_g\mathbf{J}_{\phi_g})^T \left(\sum_{g=1}^{G} \mathbf{C}_g\mathbf{J}_{\phi_g}\mathbf{B}_g\right) \mathbf{J}_\psi = (\mathbf{C}_g\mathbf{J}_{\phi_g})^T \mathbf{D}\mathbf{J}_\psi. \tag{6}$$

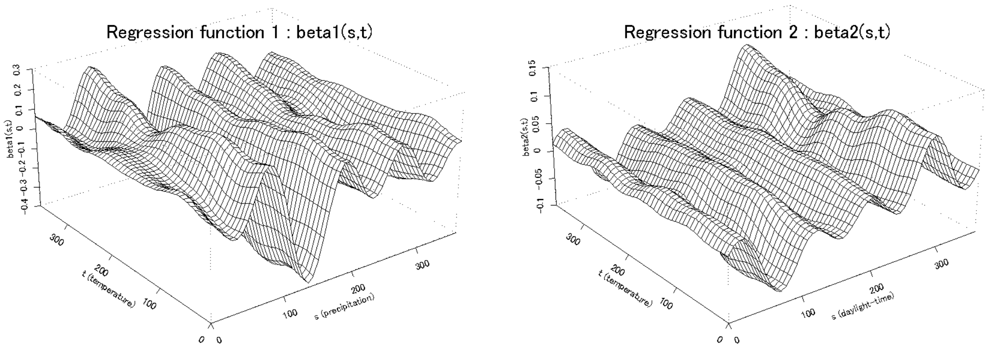

Fig. 1: Regression coefficient functions:
$\beta_1(s,t)$ *and* $\beta_2(s,t)$

2.3. Numerical example

For the illustration, the functional multiple regression analysis is applied to the daily meteorology data of 60 weather stations in Japan in 1999. The objective is to predict the temperature curve from the precipitation and daylight-time curves. Figure 1 shows the regression coefficient functions $\beta_1(s,t)$ and $\beta_2(s,t)$. In this case, $\beta_1(s,t)$ represents the relationship between precipitation and temperature over time, while $\beta_2(s,t)$ represents the relationship between daylight-time and temperature over time. From these figures, we can discern several aspects of the effect of precipitation on temperature, and that of daylight-time on temperature as follows: Precipitation around March has a strong positive association with spring and winter temperature. Precipitation around May is negatively associated with temperature throughout the year. Daylight-time around October has a strong positive association with spring and winter temperature and has a moderate positive association with summer and fall temperature.

3. Geographically weighted functional regression analysis

3.1. Ordinary geographically weighted regression model

Brunsdon *et al.* (1998) proposed geographically weighted regression (GWR) as a tool of spatial data analysis. Taking account of the spatial variation of the relationship among variables, they introduce the following regression model with spatially varying regression coefficients

$$y_i = \beta_0 + \sum_{j=1}^{G} \mathbf{X}_{ij}\beta_j(p_i) + \epsilon_i \quad (i = 1, \cdots, N), \tag{7}$$

where p_i means the geographical location of the i-th observation, $\beta_j(p_i)$ is a regression coefficient for the j-th explanatory variable at location p_i. This model makes it possible for us to understand the spatial variation of the β_j and to gain some understanding of the spatial patterns of the dependency of the response on explanatory variables. Let α_{ik} be the weight for the k-th obervation in predicting the i-th observation, and suppose weight α_{ik} $(k = 1, \cdots, N)$, which is the k-th diagonal element of a diagonal matrix $\mathbf{W}_i$, is defined on the basis of the distance between locations i and k. Then the estimator of $\beta(p_i)$ is computed by the following equation

$$(\mathbf{X}^T\mathbf{W}_i\mathbf{X})\hat{\beta}(p_i) = (\mathbf{X}^T\mathbf{W}_i\mathbf{y}), \tag{8}$$

where

$$\mathbf{W}_i = \begin{pmatrix} \alpha_{i1} & 0 & \dots & 0 \\ 0 & \alpha_{i2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \alpha_{iN} \end{pmatrix}. \tag{9}$$

As candidates for the geographical weight, Brunsdon *et al.* (1998) proposed several kernel functions such as $\alpha_{ik} = \exp(-d_{ik}/h)$, where d_{ik} is the Euclidean distance between location i and location k.

3.2. Geographically weighted functional multiple regression model

To deal with the spatial non-stationarity of the regression coefficient functions, we propose a geographically weighted functional multiple regression model as follows:

$$y_i(t) = \beta_0(t) + \sum_{g=1}^{G} \int x_{ig}(s)\beta_g(s,t,p_i)ds + \epsilon_i(t) \quad (i = 1, \cdots, N), \tag{10}$$

where p_i means the geographical location of the i-th observation, and $\beta_g(s,t,p_i)$ is a regression coefficient function for the g-th covariate at location p_i. In the similar manner as in ordinary GWR, we define geographical weight as follows:

$$\alpha_{ik} = \exp(-d_{ik}/h). \tag{11}$$

This implies that the more the distance between location i and location k, the less the value of α_{ik}, and that we can control the intensity of the geographical variability by varying the value of h. The idea of geographical weight comes from a theoretical exponential variogram, which is a measure of spatial correlation in spatial statistics (see Cressie, N., 1991). Then, incorporating the weight matrix $\mathbf{W}_i$ into the procedure of estimating $\beta_g(s,t)$, we obtain the following equation.

$$\left(\mathbf{C}_g\mathbf{J}_{\phi_g}\right)^T \mathbf{W}_i \left(\sum_{g=1}^{G} \mathbf{C}_g\mathbf{J}_{\phi_g}\mathbf{B}_{ig}\right) \mathbf{J}_\psi = \left(\mathbf{C}_g\mathbf{J}_{\phi_g}\right)^T \mathbf{W}_i\mathbf{D}\mathbf{J}_\psi \quad (i = 1, \cdots, N). \tag{12}$$

3.3. Choice of parameter h by cross-validation

Let $\alpha^{[-i]}(t), \beta_g^{[-i]}(s,t,p_i)$ be the estimates for the constant term and regression coefficient functions at p_i based on the data set except for $(x_{ig}(s), y_i(t))$. Then we can define the predictor of $y_i(t)$ as

$$\hat{y}_i^{[-i]}(t) = \alpha^{[-i]}(t) + \sum_{g=1}^{G} \int x_{ig}(s)\beta_g^{[-i]}(s,t,p_i)ds, \tag{13}$$

and choose h which minimizes the cross-validation score $CV(h)$ defined as

$$CV(h) = \sum_{i=1}^{N} \int \left\{ y_i(t) - \hat{y}_i^{[-i]}(t) \right\}^2 dt. \tag{14}$$

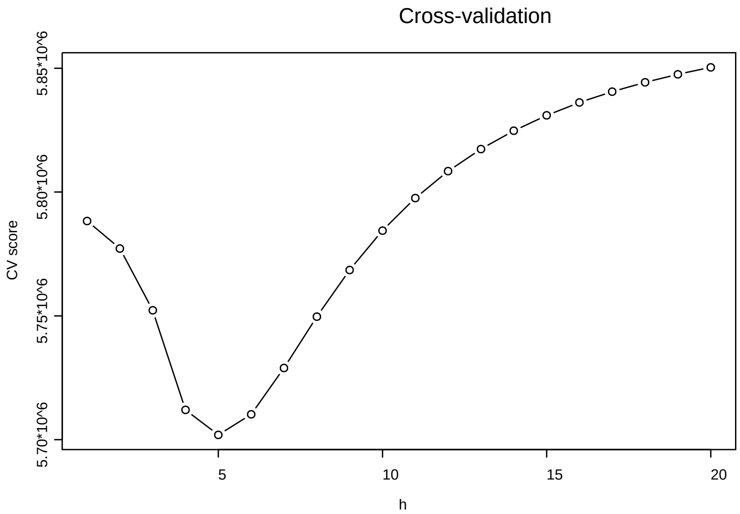

Fig. 2: Cross-validation for determining parameter h
Cross-validation scores

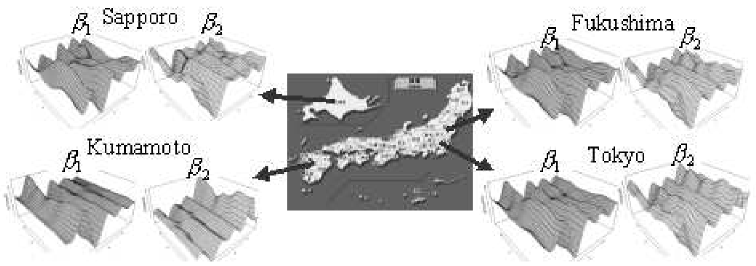

Fig. 3: Some examples of spatial variation of regression coefficient functions

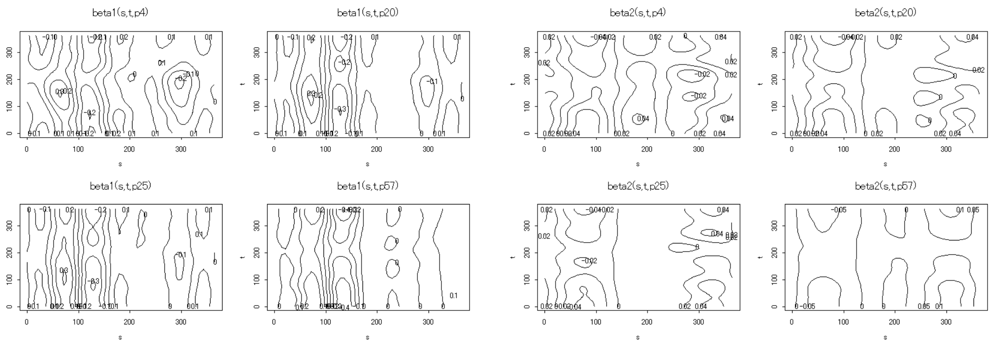

Fig. 4: Contour plots of spatially varying regression coefficient functions: $\beta_1(s,t,p_i)$ *and* $\beta_2(s,t,p_i)$ *at Sapporo, Fukushima, Tokyo, and Kumamoto*

3.4. Numerical example

For the illustration of our method, the geographically weighted functional multiple regression analysis is applied to the meteorology data. The geographical weight is computed using the Euclidean distance based on the longitude and latitude of each location. Figure 2 shows the cross-validation scores computed by varying the value of h with step 1. As a result, it gives $h = 5$ as the optimum value. Figure 3 shows examples of spatially varying regression coefficient functions at different geographical locations, where we can see $\{\beta_1(s,t,p_i)\}$ and $\{\beta_2(s,t,p_i)\}$ at locations No.4 Sapporo, No.20 Fukushima, No.25 Tokyo, and No. 57 Kumamoto. Figure 4 shows the corresponding contour plots of the regression coefficient functions. It is found that the relationship between the variables over time has some spatial variation. For example, looking at the $\{\beta_2(s,t,p_i)\}$, there are strong negative effect around March of daylight-time at locations No.4, No.20, and No.25 , while there are not so much effect around March of daylight-time at location No.53. Figure 5 shows the comparison of the squared correlation function $R^2(t)$ between the cases when the geographical weight is introduced or not. The goodness of fit improves in the geographically weighted functional regression model, because appropriate regression functions are constructed at each location.

4. Statistical inference

4.1. Assessing the spatial variability of the regression coefficient functions

Here we propose a statistic to assess the variability of $\beta_g(s,t,p_i)$ as i varies for a fixed g. A similar technique is used by Brunsdon (1998) in ordinary GWR. Our statistic is the following integrated variance of $\beta_g(s,t,p_i)$ across i:

$$v_g = \frac{1}{N}\sum_{i=1}^{N}\int\int\{\beta_g(s,t,p_i) - \beta_g(s,t,\cdot)\}^2 dtds \quad (g = 1,\cdots,G) \tag{15}$$

where a dot denotes averaging over subscript i. This implies that the higher the value of v_g, the stronger the evidence that the regression coefficient $\beta_g(s,t)$ has a large spatial variation.

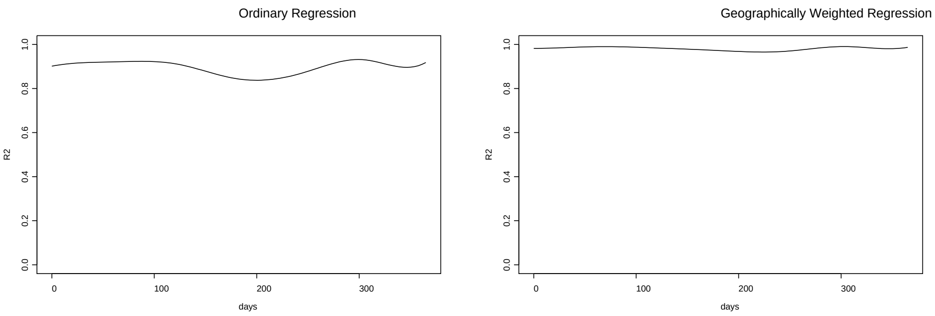

Fig. 5: Squared correlation function $R^2(t)$:
ordinary functional regression (left) and geographically weighted regression (right)

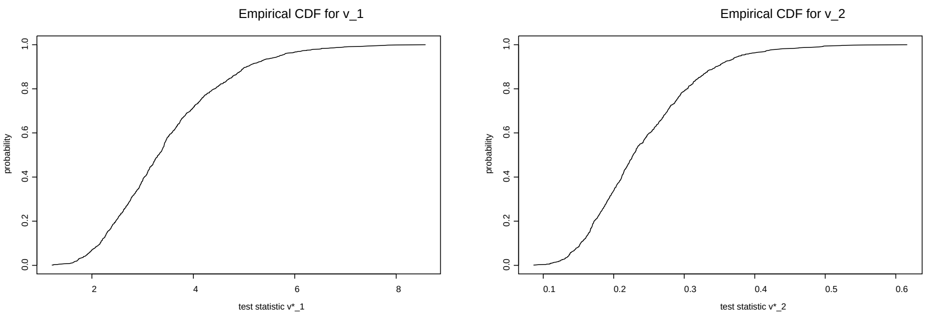

Fig. 6: Empirical cumulative distribution function for v_g:
v_1 *(left) and* v_2 *(right)*

4.2. Test of the spatial variability

To confirm the existence of the spatial variability of the regression coefficient functions, we propose to use a Monte Carlo test based on the v_g defined by eq.(15) for testing

$$H_0 : \beta_g(s, t, p_i) = \beta_g(s, t) \quad against \quad H_1 : \beta_g(s, t, p_i) \neq \beta_g(s, t) \quad for\,some\,i. \tag{16}$$

Under hypothesis H_0 we assume the regression coefficient function $\beta_g(s, t, p_i)$ do not vary with location i. This implies that, if the location of each individual were not taken into account, the regression coefficient function $\beta_g(s, t, p_i)$ would not extremely vary across location i. A test can be conducted if a null distribution of v_g is approximately created. We consider the following test procedure as an extension of the Monte Carlo test of Brunsdon *et al.* (1998). Related to the test procedure is Davison *et al.* (1997), who proposed a bootstrap test and permutation test based on Monte Carlo simulation.

Step 1. Using the original sample $\{p_i, x_{ig}(s), y_i(t)\}_{i=1}^N$, calculate v_g $(g = 1, \cdots, G)$ as the test statistic.

Step 2. Draw a random sample $\{p_i^*\}_{i=1}^N$ without replacement from $\{p_i\}_{i=1}^N$, and make a data set $\{p_i^*, x_{ig}(s), y_i(t)\}_{i=1}^N$.

Step 3. Using the data set $\{p_i^*, x_{ig}(s), y_i(t)\}_{i=1}^N$, calculate v_g^* $(g = 1, \cdots, G)$.

Step 4. Repeat steps 2 and 3 R times, and obtain R simulated values $\{{v_g}_1^*, \cdots, {v_g}_R^*\}$ of the test statistic.

Step 5. Sort the above simulated values in ascending order and get their order statistics $\{{v_g}_{(1)}^*, \cdots, {v_g}_{(R)}^*\}$. Then, regard them as a sample from the null distribution of v_g.

Step 6. Evaluate the p-value by $\frac{1}{R}\sum_{r=1}^R I\{v_g < {v_g}_{(r)}^*\}$, where $I\{\cdot\}$ is an indicator function.

Step 7. The null hypothesis H_0 is rejected when the value of the test statistic v_g is larger than the $1 - \alpha$ quantile of the above simulated distribution.

Steps 2 and 3 in the procedure imply that the spatial dependency of the observations is dispelled by randomly allocating the geographical information to them.

Table 1: Result of Monte Carlo test
test statistic v_g and 90, 95 and 99 % percentiles of v_g^*

	v_g	90%	95%	99%
β_1	3.51	5.05	5.71	6.97
β_2	0.46	0.34	0.38	0.49

4.3. Confidence interval for the predictor

So far, the reliability of the prediction based on the functional regression model has not been revealed. Here we approach the problem from the viewpoint of confidence intervals for the predicted functions. In the numerical example in the previous section, we confirmed that the goodness of fit improved by introducing the geographical variation of regression coefficients. From a different perspective, the application of geographically weighted functional regression analysis is equivalent to fitting locally ordinary functional regression models to every data point at the expense of the degrees of freedom. In order to investigate the problem, we propose a bootstrap confidence curve for the predicted functions using a method of curve resampling. The reason to use a bootstrap method is that the theory of probability of functional context has not been developed yet. Our idea of bootstrap confidence curves stems from the percentile confidence interval of Davison *et al.* (1997). The procedure is as follows:

Step 1. Draw a random sample $\{p_i^*, x_{ig}^*(s), y_i^*(t)\}_{i=1}^N$ with replacement from the original sample $\{p_i, x_{ig}(s), y_i(t)\}_{i=1}^N$.

Step 2. Based on the random sample $\{p_i^*, x_{ig}^*(s), y_i^*(t)\}_{i=1}^N$, estimate $\beta_g^*(s,t,p_i)$. Then, compute $\hat{y}^*(t)$ using $\beta_g^*(s,t,p_i)$.

Step 3. Repeat steps 1 and 2 B times, and obtain the simulated predicted functions $\{\hat{y}_1^*(t), \cdots, \hat{y}_B^*(t)\}$.

Step 4. For every knots $t_j (j = 1, \cdots, T)$, sort the above simulated predicted functions in ascending order. Then, get the order statistics of the predicted functions $\{\hat{y}_{(1)}^*(t), \cdots, \hat{y}_{(B)}^*(t)\}$.

Step 5. Compute the α and $1-\alpha$ quantiles of $\hat{y}^*(t)$, and finally get the following confidence limits $\left(\hat{y}_{(B\alpha)}^*(t), \qquad \hat{y}_{(B(1-\alpha))}^*(t)\right)$.

4.4. Numerical example (continued)

The Monte Carlo test is applied to the regression coefficient functions $\beta_1(s,t,p_i)$ and $\beta_2(s,t,p_i)$ in the geographically weighted functional multiple regression model. Figure 6 shows the empirical cumulative distribution functions (CDF) for v_1 and v_2. Table 1 shows the test statistic v_1 and v_2, and the 90, 95 and 99% percentiles for v_1^* and v_2^* obtained by the simulation with size $R = 1000$ respectively. Take 95% percentiles, for example. Since $v_1 = 3.51 < v_{1,0.95}^* = 5.71$, the null hypothesis $H_0 : \beta_1(s,t,p_i) = \beta_1(s,t)$ is accepted. This implies that the spatial variability of the relationship between the precipitation and temperature is not significant in terms of statistical inference. On the other hand, since $v_2 = 0.46 > v_{2,0.95}^* = 0.38$, the null hypothesis $H_0 : \beta_2(s,t,p_i) = \beta_2(s,t)$ is rejected. This implies that the spatial variability of the relationship between the daylight-time and temperature is significant in terms of statistical inference. Next, the bootstrap confidence intervals are computed to investigate the accuracy of the prediction, where the number of resampling is 1000. Figure 7 shows the bootstrap confidence intervals (or curves) for the predicted temperature curve of No.25 Tokyo, where the solid curve indicates the observed temperature curve, the dotted curve indicates the predicted temperature curve, the dashed curves indicate the upper and under confidence limits. The left figure in Figure 7 shows the case of the ordinary functional regression analysis, while the right figure shows the case of the geographically weighted functional regression analysis. From these figures, it seems

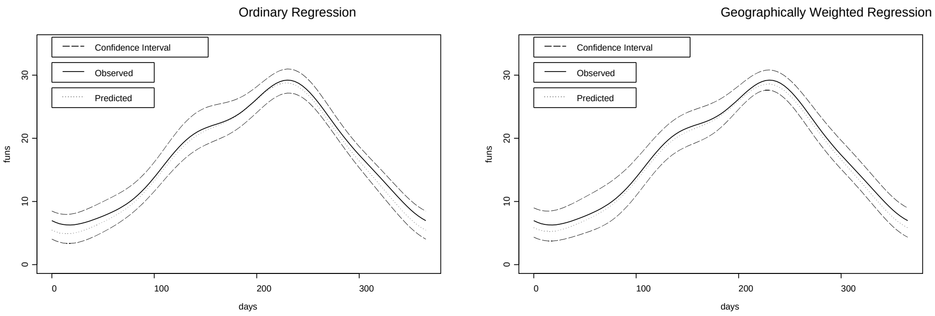

Fig. 7: Bootstrap confidence interval for the predictor (TOKYO):
ordinary functional regression (left) and geographically weighted regression (right)

that the width of the confidence interval between upper and lower limits is almost the same in both cases. This is a warning that the accuracy of the prediction in geographically weighted functional regression does not improve, although the goodness of fit improves by the introduction of geographical weight. We should take care about that point.

5. Concluding remarks

We have proposed a method of geographically weighted functional regression analysis. It is assumed that the relationship between the variables over time does not remain fixed over space and the regression functions vary over space. A procedure for evaluating the spatial variability of the regression functions is established using Monte Carlo simulation. The advantage of the method is that it enables us to understand the relationship among the variables over space as well as over time. In this study, a kernel function was used to define the geographical weight in terms of spatial correlation. However, as well known in the field of spatial statistics, there are several candidates of mathematical expressions for modeling the spatial correlation. The relationship among the choice of kernel functions, the power of Monte Carlo test, and the accuracy of the prediction is our important future work.

6. Acknowledgements

This research was partly supported by Japan Society of the Promotion of Science (Grant-in-Aid for Scientific Research (C) 13680374).

REFERENCES

Brunsdon, C., Fotheringham, S. and Charlton, M. (1998). Geographically weighted regression - modelling spatial non-stationarity, *Journal of The Royal Statistical Society*, Ser.D, **47**, 431–443.

Cressie, N. (1991). Statistics for Spatial Data, *New York:Wiley.*

Davison, A.C. and Hinkley, D.V. (1997). *Bootstrap Methods and Their Application* , Cambridge University Press.

Japan Meteorological Agency (1999). *Annual report of Automated Meteorological Data Acquisition System*, Japan Meteorological Business Support Center (JMBSC).

Ramsay, J. O. and Dalzell, C. (1991). Some tools for functional data analysis (with discussion), *Journal of The Royal Statistical Society*, Ser.B, **53**, 539–572.

Ramsay, J. O. and Silverman, B.W. (1997). *Functional Data Analysis* , Springer.

Shimokawa, M., Mizuta, M. and Sato, Y. (2000). An Expansion of Functional Regression Analysis. *Japanese Journal of Applied Statistics*, **29**, 27–39.

Wilhelm, A. and Steck, R. (1998). Exploring spatial data by using interactive graphics and local statistics, *Journal of The Royal Statistical Society*, Ser.D, **47**, 423–430.