Geographically Weighted Functional Multiple Regression Analysis: A Numerical Investigation

Yoshihiro YAMANISHI*
Graduate School of Natural Science & Technology, Okayama University 3-1-1 Tsushima-Naka, Okayama, Okayama 700-8530, Japan

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

Abstract
Functional regression analysis enables us to investigate the relationship among the variables over time. Sometimes, however, we meet the case where regression coefficients do not remain fixed over space, when we analyze spatial data. 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.

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

1. Introduction

Ramsay (1997) proposed a method of functional regression analysis and Shimokawa et al. (2000) extended it so that it can deal with more than one covariate. These methods enable us to investigate the relationship among the variables over time. Brunsdon et al. (1998) proposed geographically weighted regression (GWR). The aim of their method is to understand the relationship between the variables over space in applying regression analysis. Sometimes we meet the case where regression coefficients do not remain fixed over space when we analyze spatial data. The purpose in this study is to develop GWR in functional context in order to investigate the relationship among the variables not only over time but also over space. In section 2, we review functional regression analysis. In section 3, we formulate geographically weighted functional multiple regression model. In section 4, defining a statistic to assess the spatial variability of regression coefficient functions, we 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.

2. Functional regression analysis

2.1. Functional multiple regression model

Shimokawa et al. (2000) proposed a functional multiple regression model 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 g-th covariate function and

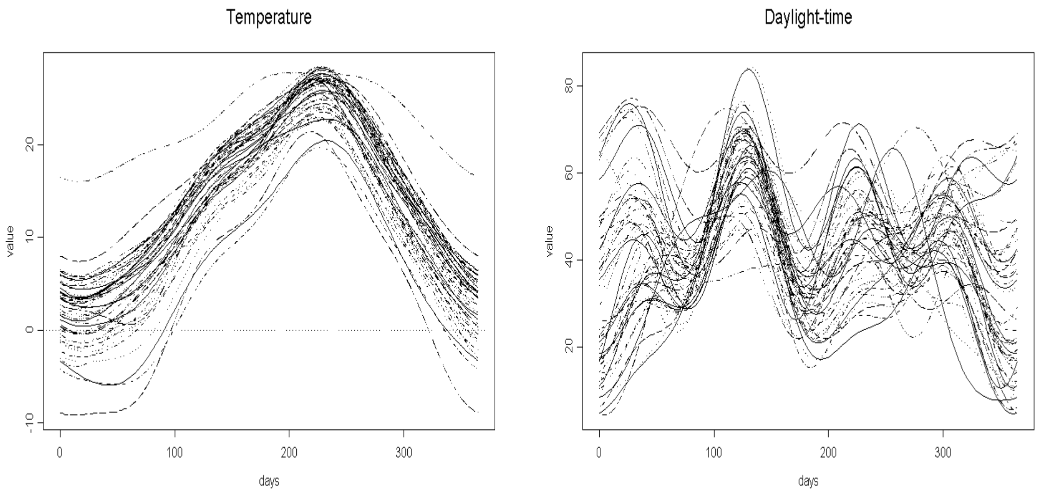

Fig. 1. Examples of functional data: temperature and daylight-time

$\epsilon(t)$ is a random error function. Then $\beta_g(s,t)$ is estimated by minimizing the sum of integrated squared residuals defined by

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

Figure 2 shows the process flow of applying the functional multiple regression analysis, where $G = 2$. The goodness of fit of a functional regression model can be assessed by the squared correlation function defined by

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

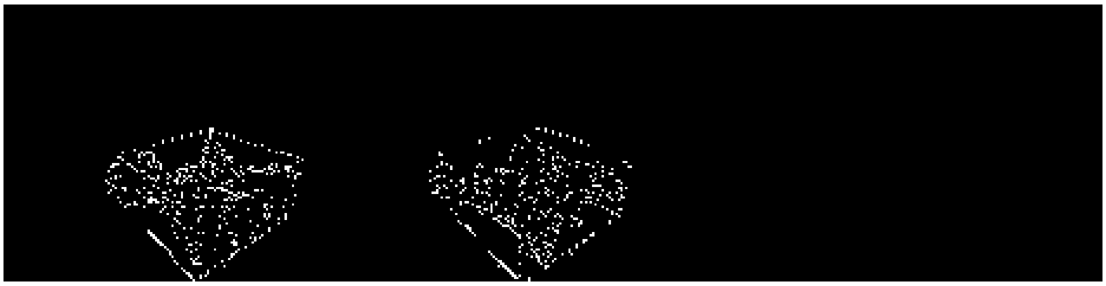

Fig. 2. Process flow of functional multiple regression

2.2. Algorithm of estimating regression coefficient functions

Suppose that functional data can be approximated with finite expansions of some 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 $x_g(s)$, $y(t)$ 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, $\mathbf{B}$ is a $K_\phi \times K_\psi$ coefficient matrix, and K_ϕ and K_ψ are the numbers of basis functions, respectively. 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 $IMISE$ leads to the following equation.

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

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 the variables, they introduce the following regression model with geographically variable 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 the regression coefficient for the j-th covariate at location p_i. This model makes it possible for us to understand the spatial variation of the regression coefficients 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 expressed by

$$\hat{\beta}(p_i) = (\mathbf{X}^T\mathbf{W}_i\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{W}_i y), \tag{8}$$

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 a candidate of the geographical weight, the following weight is used.

$$\alpha_{ik} = \exp(-d_{ik}/h), \tag{10}$$

where d_{ik} is the Euclidean distance between location i and location k. It 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.

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:

$$y_i(t) = \beta_0(t) + \sum_{g=1}^{G} \int x_{ig}(s)\beta_g(s,t,p_i)ds + \epsilon_i(t) \quad (i = 1, \cdots, N), \tag{11}$$

where p_i means the geographical location of the i-th observation, and $\beta_g(s,t,p_i)$ is a regression coefficient function. Here we define geographical weight α_{ik} in the similar manner as in ordinary GWR. Then, incorporating the weight matrix $\mathbf{W}_i$ into the procedure of estimating $\beta_g(s,t)$, we obtain the following equations.

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

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_i, y_i). 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

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

3.4. Numerical example

For the illustration of our method, the geographically weighted functional multiple regression analysis is applied to the daily meteorology data of 60 weather stations in Japan in 1999. The geographical weight is computed using the longitude and latitude of each location. The objective is to predict the temperature curve from the precipitation and daylight-time curves. 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. Figure 3 shows examples of regression coefficient functions at some 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. It is found that the relationship between the variables over time has some spatial variation. Figure 4 shows the comparison of the squared correlation function $R^2(t)$ between the cases when the geographical weight is introduced or not. Compared to the ordinary functional regression model, it is found that the goodness of fit improved apparently in the geographically weighted functional regression model.

4. Statistical inference

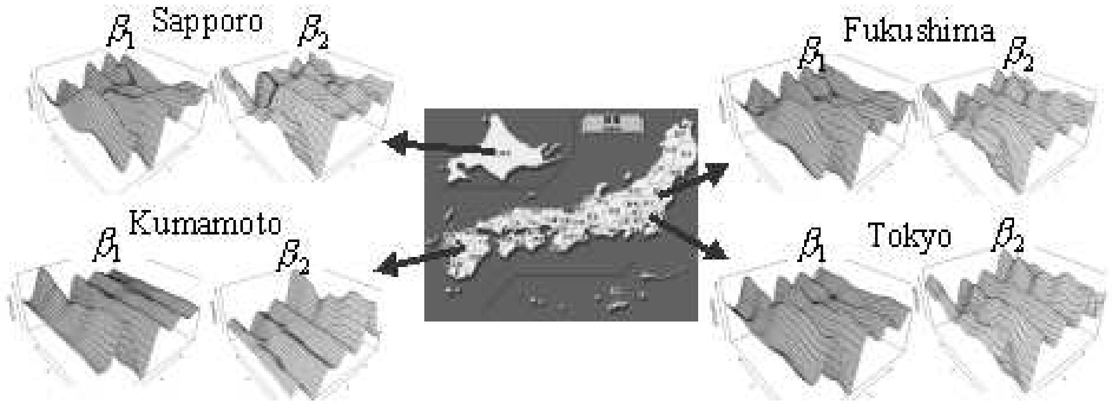

Fig. 3. Some examples of spatial variation of regression coefficient functions

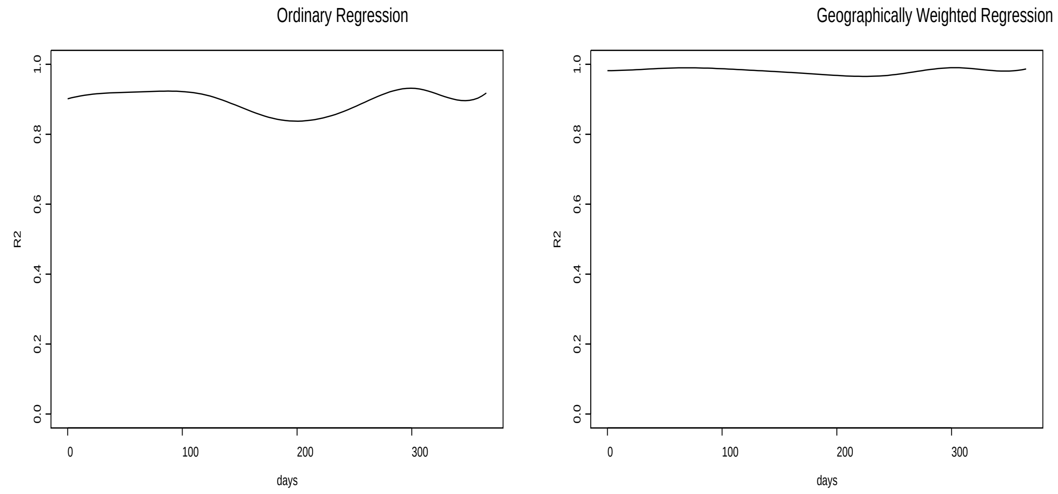

Fig. 4. Squared correlation function $R^2(t)$:
ordinary functional regression (left) and geographically weighted regression (right)

4.1. Assessing the spatial variability of 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. 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.

4.2. Test of the spatial variability

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

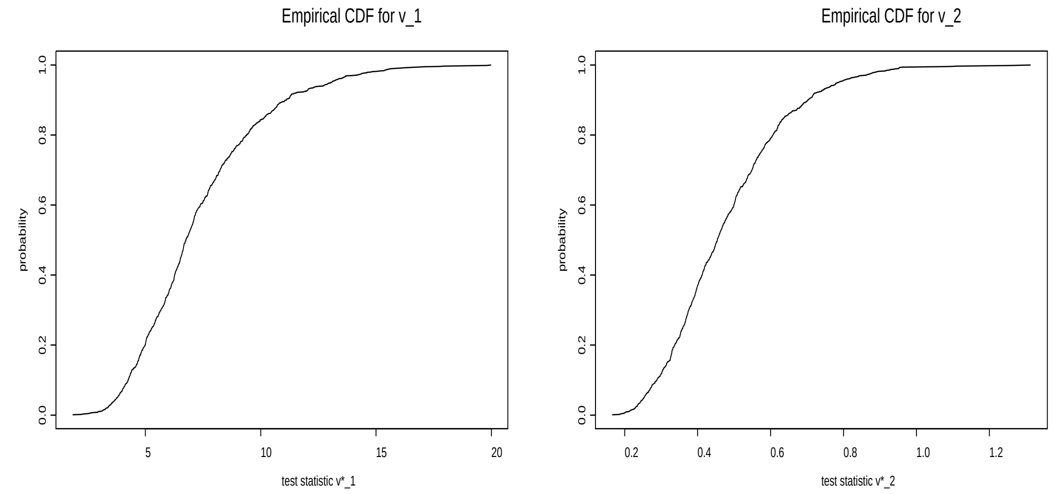

Fig. 5. Empirical cumulative distribution function for v_g: v_1 (left) and v_2 (right)

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

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

Step 3. Using the artificial 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 the R simulated sample $\{v_{g_1}^*, \cdots, v_{g_R}^*\}$ of the test statistic.

Step 5. Sort the above simulated sample and get the sample $\{v_{g(1)}^*, \cdots, v_{g(R)}^*\}$ in ascending order, and regard it as an approximate null distribution of v_g.

Step 6. Define 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.

Table 1. Monte Carlo test
90 %, 95 % and 99 % percentile 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

To investigate the accuracy of the prediction, we consider the confidence interval for the predicted functions in applying geographically weighted functional multiple regression analysis. 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 implies that ordinary functional regression analysis is applied to the data locally at the expense of the degree of freedom. In order to investigate the problem, we propose a bootstrap confidence interval for the predictors by the method of curve resampling. The procedure is as follows:

Step 1. Based on the initial sample $\{p_i, x_{ig}(s), y_i(t)\}_{i=1}^N$ estimate $\beta_g(s,t)\,(g = 1, \cdots, G)$.

Step 2. Draw a random sample $\{p_i^*, x_{ig}^*(s), y_i^*(t)\}_{i=1}^N$ with replacement from a set of original curves $\{p_i, x_{ig}(s), y_i(t)\}_{i=1}^N$.

Step 3. Based on the random sample $\{p_i^*, x_{ig}^*(s), y_i^*(t)\}_{i=1}^N$ estimate $\beta_g^*(s,t,p_i)\,(g = 1, \cdots, G)$.

Step 4. Compute $\hat{y}^*(t)$ using $\beta_g^*(s,t,p_i)$, and define $u^* = \hat{y}^*(t)$ as the bootstrap version of u.

Step 5. Repeat steps 2, 3 and 4 B times, and obtain the simulated sample $\{u_1^*, \cdots, u_B^*\}$.

Step 6. Sort the above simulated sample and get the sample $\{u_{(1)}^*, \cdots, u_{(B)}^*\}$ in ascending order.

Step 7. Compute the α and $1-\alpha$ quantiles of u^*, and finally get the following confidence limits $\left(u_{(B\alpha)}^*, \quad u_{(B(1-\alpha))}^*\right)$.

4.4. Numerical example (continued)

The Monte Carlo test is applied to the regression coefficient function $\beta_1(s,t,p_i)$ and $\beta_2(s,t,p_i)$ in the geographically weighted functional multiple regression model. Table 1 shows the $90\%, 95\%$ and 99% percentiles for v_1^* and v_2^* obtained by the simulation with size $R = 1000$ respectively. Figure 5 shows the empirical cumulative distribution functions (CDF) for v_1 and v_2. Next, the bootstrap confidence intervals are computed to investigate the accuracy of the prediction, where the number of resampling is 1000. Figure 6 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 6 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 that the width of confidence interval between upper and lower limits in geographically weighted functional regression is a bit broader than that in ordinary functional regression. It implies that the accuracy of the prediction in geographically weighted functional regression gets worse slightly compared to that in ordinary functional regression analysis.

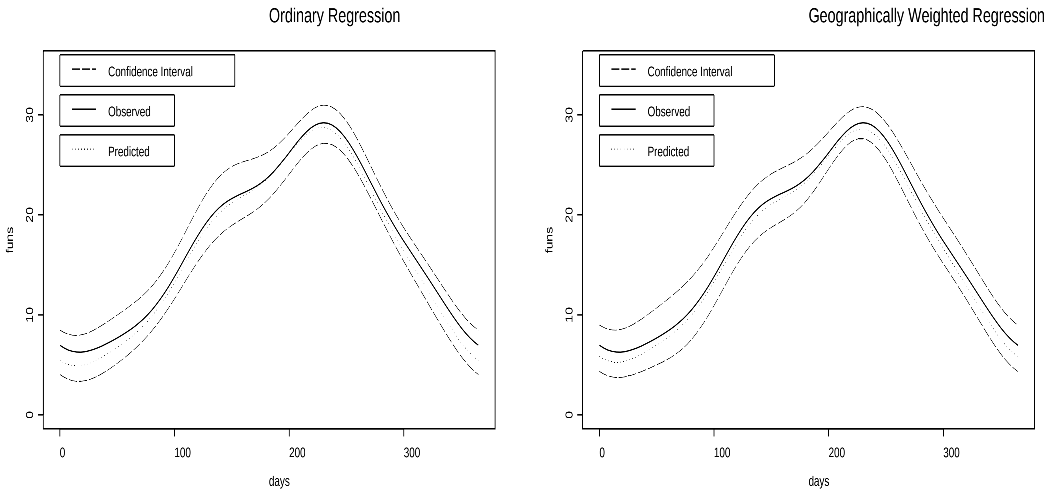

Fig. 6. Bootstrap confidence interval for the predictor (TOKYO): ordinary functional regression (left) and geographically weighted regression (right)

REFERENCES

Brunsdon, C., Fotheringham, S. and Charlton, M. (1998). Geographically weighted regression - modelling spatial non-stationarity, *J. Royal Statist. Soc.*, Ser.D, 47, 3, 431-443.

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. and Dalzell, C. (1991). Some tools for functional data analysis (with discussion), *J. Royal Statist. Soc.*, 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 (in Japanese). *J. Jpn. Applied Statist.*, 29-1, 27-39.

Wilhelm, A. and Steck, R. (1998). Exploring spatial data by using interactive graphics and local statistics, *J. Royal Statist. Soc.*, Ser.D, 47, 3, 423-430.