Journal of the Korean Data & Information Science Society
2017, **28**(1), 227–235

http://dx.doi.org/10.7465/jkdi.2017.28.1.227
한국데이터정보과학회지

Geographically weighted least squares-support vector machine[†]

Changha Hwang[1] · Jooyong Shim[2]

[1]Department of Applied Statistics, Dankook University
[2]Department of Statistics, Inje University

Received 26 December 2016, revised 7 January 2017, accepted 10 January 2017

Abstract

When the spatial information of each location is given specifically as coordinates it is popular to use the geographically weighted regression to incorporate the spatial information by assuming that the regression parameters vary spatially across locations. In this paper, we relax the linearity assumption of geographically weighted regression and propose a geographically weighted least squares-support vector machine for estimating geographically weighted mean by using the basic concept of kernel machines. Generalized cross validation function is induced for the model selection. Numerical studies with real datasets have been conducted to compare the performance of proposed method with other methods for predicting geographically weighted mean.

Keywords: Generalized cross validation function, geographically weighted regression, kernel machine, least squares-support vector machine, model selection, spatial information.

1. Introduction

Geographical weighted regression (GWR) model was proposed first by Fotheringham *et al.* (1996) for the study with respect to the spatial heterogeneity, which is an alternative approach to the spatial autoregressive regression (Anselin, 1992). GWR extends the classical regression model to allow local rather than global regression parameters to be estimated by incorporating the spatial information to satisfy the assumption that the regression parameters vary spatially across locations of interest. Brunsdon and Fotheringham (1999) have studied the relationship between the house price and the area, and mentioned several questions GWR faced: the selection of the input variables, bandwidth parameter, and the spatial autocorrelation of errors. Zhang (2004) utilized GWR model to study the height of crown

[†] This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2015R1D1A1A01056582, NRF-2014R1A1A2054917). This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2015S1A3A2046715).

[1] Professor, Department of Applied Statistics, Dankook University, Yongin 16890, Korea.
[2] Corresponding author: Adjunct professor, Institute of Statistical Information, Department of Statistics, Inje University, Kimhae 50834, Korea. E-mail: ds1631@hanmail.net

and the result showed that GWR could bring better significance and residual error than the ordinary least squares regression estimation.

Support vector machine (SVM), first developed by Vapnik (1995) and his group at AT&T Bell Laboratories, is known to solve the weak point of the artificial neural network (Rosenblatt, 1958) such as the local minima existence in the area of structural risk minimization and statistical learning theory. SVM has been successfully applied to lots of real world problems related to regression and classification problems. One of prominent advantages of SVM is the use of kernels to utilize the nonlinear transforms without knowing the specific transforms. According to this idea, other authors proposed a class of kernel-based algorithms, such as kernel Fisher discriminant analysis (Mika *et al.*, 1999), least square-SVM (LS-SVM, Suykens and Vandewalle, 1999), kernel ridge regression (Saunders *et al.*, 1998) and the kernel minimum squared error model (Xu *et al.*, 2001). LS-SVM is the modified version of SVM in least squares senses. The kernel minimum squared error model is a generalization of the conventional minimum squared error model to yield a new type of nonlinear model, which was devised by using the theory of reproducing kernels and adding different penalty terms. Xu *et al.* (2001) argued that LS-SVM can be viewed as a special case of the kernel minimum squared error model. Smola *et al.* (1998) developed a semiparametric SVM which is useful in the case where the domain knowledge exists about functions to be estimated or emphasis is put onto comprehensibility of the given model. Shim *et al.* (2011) proposed a semiparametric LS-SVM for accelerated failure time model. LS-SVM has been extended to recurrent models and use in optimal control problems. See for further details, Suykens and Vandewalle (1999), Suykens *et al.* (2001), Hwang and Shim (2016) and Hwang *et al.* (2016).

In this paper we present the geographically weighted LS-SVM (GWLS-SVM) that combines ideas from the architectures of GWR with those of LS-SVM. GWLS-SVM can be applied efficiently when the functional form of the relationship between the response and input variables is left unspecified.

The rest of this paper is organized as follows. In Section 2 we review GWR briefly. In Section 3 we propose the geographically weighted LS-SVM (GWLS-SVM) for geographically weighted mean estimation and induce the generalized cross validation function. In Section 4 and 5 we present the case studies and conclusion, respectively.

2. Geographically weighted regression

Given a training dataset $\{\boldsymbol{x}_i, \boldsymbol{u}_i, y_i\}_{i=1}^n$ with each input vector $\boldsymbol{x}_i \in R^d$, $\boldsymbol{u}_i \in R^2$ is the spatial coordinate vector (longitude, latitude) and corresponding response $y_i \in R$, GWR can be represented as follows:

$$\boldsymbol{y}_i = \beta_0(\boldsymbol{u}_i) + \sum_{k=1}^{d} \beta_k(\boldsymbol{u}_i) x_{ik} + e_i,$$

where $\beta_0(\boldsymbol{u}_i)$ and $\beta_k(\boldsymbol{u}_i)$'s are the bias (intercept) and the slope parameter which depend on the ith spatial coordinate vector $\boldsymbol{u}_i$.

In GWR, the weight of the ith observation is affected by the proximity to the spatial coordinates $\boldsymbol{u}_i$, which leads the weight of the ith observation varies along with the change of $\boldsymbol{u}_i$, and $\boldsymbol{y} = (y_1, \cdots, y_n)'$ given $\boldsymbol{x}$ at location $\boldsymbol{u}_j$ is assumed to follow a normal distribution

$N(\mathbf{0}, W_j^{-2})$ so that the negative log-likelihood can be expressed as follows:

$$L_j = \sum_{i=1}^{n} w_{ji}(y_i - \beta_0(\boldsymbol{u}_i) - \sum_{k=1}^{d} \beta_k(\boldsymbol{u}_i)x_{ik})^2, \tag{2.1}$$

where w_{ji} is the weight function which is the decreasing in the distance from the jth location $\boldsymbol{u}_j$ to the ith location $\boldsymbol{u}_i$, and W_j is the diagonal matrix composed of w_{ji}'s.

Generally the exponential function of the distance from $\boldsymbol{u}_j$ to other location $\boldsymbol{u}_i$ is used for the weighted function, which is $w_{ji} = \exp(-d_{ji}/h)$, where $h > 0$ is the bandwidth parameter. From the negative log-likelihood (2.1), the estimators of $\boldsymbol{\beta}(\boldsymbol{u}_j) = (\beta_0(\boldsymbol{u}_j), \cdots, \beta_d(\boldsymbol{u}_j))'$ as follows:

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{u}_j) = (\boldsymbol{X}'W_j\boldsymbol{X})^{-1}\boldsymbol{X}'W_j\boldsymbol{y},$$

where $\boldsymbol{X} = \{1, \boldsymbol{x}_i\}_{i=1}^n$ is the $n \times (d+1)$ input matrix and $\boldsymbol{y} = (y_1, \cdots, y_n)'$.

The estimated regression function given $(\boldsymbol{x}_j, \boldsymbol{u}_j)$ can be obtained as

$$\widehat{f}(\boldsymbol{x}_j, \boldsymbol{u}_j) = \boldsymbol{X}_j\widehat{\boldsymbol{\beta}}(\boldsymbol{u}_j) = H_j\boldsymbol{y},$$

where $H_j = \boldsymbol{X}_j(\boldsymbol{X}'W_j\boldsymbol{X})^{-1}\boldsymbol{X}'W_j$.

The predicted regression function given $\boldsymbol{x}_t$ at new location $\boldsymbol{u}_t$ is given as follows:

$$\widehat{f}(\boldsymbol{x}_t, \boldsymbol{u}_t) = H_t\boldsymbol{y},$$

where $H_t = \boldsymbol{X}_t(\boldsymbol{X}'W_t\boldsymbol{X})^{-1}\boldsymbol{X}'W_t$, W_t is the diagonal matrix of $(w_{t1}, \cdots, w_{tn})$, $w_{ti} = \exp(-d_{ti}/h)$, d_{ti} is the distance between location $\boldsymbol{u}_t$ and $\boldsymbol{u}_i$ for $i = 1, \cdots, n$.

The performance of GWR is affected by the bandwidth parameter in the weight function. To select the optimal values of the bandwidth parameter, we consider the leave-one-out cross validation (LOO-CV) function as follows:

$$CV(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{f}_i^{(-i)}(\boldsymbol{\theta}))^2,$$

where $\boldsymbol{\theta}$ is a candidate set of bandwidth parameters and $\widehat{f}_i^{(-i)}(\boldsymbol{\theta})$ is the predicted value of $f(\boldsymbol{x}_i, \boldsymbol{u}_i)$ obtained from data without the ith observation, which can be obtained as follows:

$$\widehat{f}(\boldsymbol{x}_i, \boldsymbol{u}_i)^{(-i)} = \boldsymbol{X}_i(\boldsymbol{X}^{(-i)'}W_i^{(-i)}\boldsymbol{X}^{(-i)})^{-1}\boldsymbol{X}^{(-i)'}W_i^{(-i)}\boldsymbol{y}^{(-i)},$$

where $\boldsymbol{X}^{(-i)} = (\boldsymbol{X}_1', \cdots, \boldsymbol{X}_{i-1}', \boldsymbol{X}_{i+1}', \cdots, \boldsymbol{X}_n')'$, $W_i^{(-i)}$ is the diagonal matrix of $(w_{i1}, \cdots, w_{i,i-1}, w_{i,i+1}, \cdots, w_{1n})$ and $\boldsymbol{y}^{(-i)} = (y_1, \cdots, y_{i-1}, y_{i+1}, \cdots, y_n)'$.

Since for each candidate set of bandwidth parameters, $\widehat{f}_i^{(-i)}(\boldsymbol{\theta})$ for $i = 1, \cdots, n$, should be evaluated, selecting parameters using LOO-CV function is computationally burdensome. By using leaving-out-one lemma (Wahba, 1990) and the first order Taylor expansion, the ordinary cross validation function is obtained as follows:

$$OCV(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1 - \widehat{f}_i(\boldsymbol{\theta})}{1 - \frac{\partial\widehat{f}_i}{\partial y_i}}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1 - \widehat{f}_i(\boldsymbol{\theta})}{1 - h_{ii}(\boldsymbol{\theta})}\right)^2, \tag{2.2}$$

where $h_{ii}(\boldsymbol{\theta})$ for $i = 1, \cdots, n$, is the ith diagonal element of the hat matrix H such that $\widehat{\boldsymbol{f}} = H\boldsymbol{y}$, which is composed of H_j's in (2.4) such that $H = (H_1', \cdots, H_n')'$. By averaging the residuals in (2.2) by $(1 - trace(H)/n)$, the generalized cross validation (GCV) function is obtained as follows:

$$GCV(\boldsymbol{\theta}) = \frac{n \sum_{i=1}^{n} (1 - \widehat{f}_i(\boldsymbol{\theta}))^2}{(n - trace(H))^2}.$$

3. Geographically weighted LS-SVM

3.1. LS-SVM for regression

LS-SVM has been successfully applied to statistical problems like regression and classification. The LS-SVM model for regression estimation has the following representation in feature space,

$$f(\boldsymbol{x}) = \boldsymbol{\omega}'\phi(\boldsymbol{x}) + b,$$

where $\boldsymbol{\omega} \in R^{d_f}$ is a weight vector corresponding to $\phi(\boldsymbol{x})$.

Given a training dataset $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ with each input $\boldsymbol{x}_i \in R^d$ and the corresponding response $y_i \in R$, the optimization problem in the primal weight space is considered as follows:

$$L(\boldsymbol{\omega}, b, \boldsymbol{e}) = \frac{1}{2}\boldsymbol{\omega}'\boldsymbol{\omega} + \frac{C}{2}\sum_{i=1}^{n} e_i^2$$

subject to equality constraints

$$y_i = \boldsymbol{\omega}'\phi(\boldsymbol{x}_i) + b + e_i, \ i = 1, \cdots, n.$$

The cost function with squared error and penalty corresponds to a form of ridge regression. To find minimizers of the objective function, we can construct the Lagrangian function as follows:

$$L(\boldsymbol{\omega}, b, \boldsymbol{e}; \boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\omega}'\boldsymbol{\omega} + \frac{C}{2}\sum_{i=1}^{n} e_i^2 - \sum_{i=1}^{n} \alpha_i(\boldsymbol{\omega}'\phi(\boldsymbol{x}_i) + b + e_i - y_i)$$

where α_i 's are the Lagrange multipliers. Then, the conditions for optimality are given by

$$\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{\omega}} &= \boldsymbol{0} \rightarrow \boldsymbol{\omega} = \sum_{i=1}^{n} \alpha_i \phi(\boldsymbol{x}_i) \\
\frac{\partial L}{\partial b} &= 0 \rightarrow \sum_{i=1}^{n} \alpha_i = 0 \\
\frac{\partial L}{\partial e_i} &= 0 \rightarrow e_i = \frac{1}{C}\alpha_i, \ i = 1, \cdots, n \\
\frac{\partial L}{\partial \alpha_i} &= 0 \rightarrow y_i - b - \boldsymbol{w}'\phi(\boldsymbol{x}_i) - e_i, \ i = 1, \cdots, n
\end{aligned}$$

After eliminating e_i and $\boldsymbol{w}$, we could have the solution by the following linear equations

$$\begin{bmatrix} \boldsymbol{K}+\frac{1}{C}\boldsymbol{I} & \mathbf{1} \\ \mathbf{1}' & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ b \end{bmatrix} = \begin{bmatrix} \boldsymbol{y} \\ 0 \end{bmatrix}, \tag{3.1}$$

where is $\mathbf{1}$ is a $n \times 1$ vector of 1's, $\boldsymbol{I}$ is a $n \times n$ identity matrix and $\boldsymbol{K} = K(\boldsymbol{x},\boldsymbol{x})$ is the kernel matrix constructed with $\boldsymbol{x}$.

From the linear equation (3.1) the bias estimate and optimal values of Lagrangian multipliers, $\widehat{b}$ and $\widehat{\alpha}_i$'s can be obtained. The predicted regression function given $\boldsymbol{x}_t \in R^d$ is obtained as

$$\widehat{f}(\boldsymbol{x}_t) = K(\boldsymbol{x}_t, \boldsymbol{x})\widehat{\boldsymbol{\alpha}} + \widehat{b} = H_t \boldsymbol{y},$$

where $H_t = (K(\boldsymbol{x}_t, \boldsymbol{x}), \mathbf{1})H_0$, $\boldsymbol{x} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)' \in R^{n \times d}$, $\boldsymbol{y} = (y_1, \cdots, y_n)' \in R^n$, $K = K(\boldsymbol{x}, \boldsymbol{x})$ and $H_0 = \begin{pmatrix} (K+I/C)^{-1} - (K+I/C)^{-1}\mathbf{1}(\mathbf{1}'(K+I/C)^{-1}\mathbf{1})\mathbf{1}'(K+I/C)^{-1} \\ (\mathbf{1}'(K+I/C)^{-1}\mathbf{1})^{-1}\mathbf{1}'(K+I/C)^{-1} \end{pmatrix}$.

The performance of LS-SVM is affected by hyperparameters, which are the penalty parameter and the kernel parameter. To select the optimal values of hyperparameters of LS-SVM, we consider LOO-CV function as follows:

$$CV(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{f}_i^{(-i)}(\boldsymbol{\theta}))^2, \tag{3.2}$$

where $\boldsymbol{\theta}$ is a candidate set of hyper-parameters and $\widehat{f}_i^{(-i)}(\boldsymbol{\theta})$ is the predicted value of $f(\boldsymbol{x}_i)$ obtained from data without the ith observation. From LOO-CV function in (3.2) GCV function is obtained as follows:

$$GCV(\boldsymbol{\theta}) = \frac{n\sum_{i=1}^{n}(1 - \widehat{f}_i(\boldsymbol{\theta}))^2}{(n - trace(H))^2},$$

where $\widehat{f}_i(\boldsymbol{\theta})$ is the estimated regression function given $\boldsymbol{x}_i$ and $H = H(\boldsymbol{x}, \boldsymbol{x})$ is the hat matrix such that $\widehat{\boldsymbol{f}} = H\boldsymbol{y}$.

3.2. Geographically weighted LS-SVM

Given a training dataset $\{\boldsymbol{x}_i, \boldsymbol{u}_i, y_i\}_{i=1}^n$ with each input vector $\boldsymbol{x}_i \in R^d$, $\boldsymbol{u}_i \in R^2$ (longitude, latitude) and corresponding response $y_i \in R$. We want predict the regression function given $\boldsymbol{x}_t$ at location $\boldsymbol{u}_j$ as follows:

$$f(\boldsymbol{x}_t, \boldsymbol{u}_j) = \boldsymbol{\omega}_j'\phi(\boldsymbol{x}_t) + b_j.$$

We assume that $\boldsymbol{y}$ given $\boldsymbol{x}$ at location $\boldsymbol{u}_j$ follows a normal distribution $N(\mathbf{0}, W_j^{-2})$ so that the penalized negative log-likelihood can be expressed as follows:

$$L_j = \sum_{i=1}^{n} w_{ji}(y_i - \boldsymbol{\omega}_j'\phi(\boldsymbol{x}_i) - b_j)^2 + \frac{\lambda}{2}||\boldsymbol{\omega}_j||^2, \tag{3.3}$$

where $\lambda > 0$ is a penalty parameter, W_j is the diagonal matrix of weight, w_{ji} is the weight function which is an exponential function of the distance between $\boldsymbol{u}_j$ and $\boldsymbol{u}_i$.

According to the representation theorem by Kimeldorf and Wahba (1971), the optimal values of $\boldsymbol{\omega}_j$ can be written as expansions over training observations, such that

$$\boldsymbol{\omega}_j' \phi(\boldsymbol{x}_i) = K(\boldsymbol{x}_i, \boldsymbol{x})\boldsymbol{\alpha}_{\cdot j}$$

for some weight vector $\boldsymbol{\alpha}_{\cdot j}$. The penalized negative log-likelihood (3.3) can be reexpressed as follows:

$$\boldsymbol{L} = \sum_{i=1}^{n} w_{ji}(y_i - K_i\boldsymbol{\alpha}_{\cdot j} - b_j)^2 + \frac{\lambda}{2}\boldsymbol{\alpha}_{\cdot j}' K \boldsymbol{\alpha}_{\cdot j} \tag{3.4}$$

where $K = K(\boldsymbol{x}.\boldsymbol{x})$ and K_i is the ith row of K .

Taking derivative of (3.4) with respect to $(\boldsymbol{\alpha}_{\cdot j}, b_j)$, estimate of $(\boldsymbol{\alpha}_{\cdot j}, b_j)$ can be obtained from the linear equations as follows:

$$\begin{pmatrix} \widehat{\boldsymbol{\alpha}}_{\cdot j} \\ \widehat{b}_j \end{pmatrix} = \begin{pmatrix} W_j K + \lambda \mathrm{I} & W_j \mathbf{1} \\ \mathbf{1}' W_j K & \mathbf{1}' W_j \mathbf{1} \end{pmatrix}^{-1} \begin{pmatrix} W_j \\ \mathbf{1}' W_j \end{pmatrix} \boldsymbol{y} = S_j \boldsymbol{y}.$$

Then the estimated regression function given $\boldsymbol{x}_j$ at location $\boldsymbol{u}_j$ is given as follows:

$$\widehat{f}(\boldsymbol{x}_j, \boldsymbol{u}_j) = K(\boldsymbol{x}_j, \boldsymbol{x})\boldsymbol{\alpha}_{\cdot j} + b_j = (K(\boldsymbol{x}_j, \boldsymbol{x}), 1) S_j \boldsymbol{y}, \tag{3.5}$$

From (3.5) the estimated regression function vector of training data can be expressed as follows:

$$\widehat{\boldsymbol{f}} = \begin{pmatrix} \widehat{f}(\boldsymbol{x}_1, \boldsymbol{u}_1) \\ \widehat{f}(\boldsymbol{x}_2, \boldsymbol{u}_2) \\ \vdots \\ \widehat{f}(\boldsymbol{x}_n, \boldsymbol{u}_n) \end{pmatrix} = \begin{pmatrix} H(\boldsymbol{x}_1, \boldsymbol{u}_1) \\ H(\boldsymbol{x}_2, \boldsymbol{u}_2) \\ \vdots \\ H(\boldsymbol{x}_n, \boldsymbol{u}_n) \end{pmatrix} \boldsymbol{y} = \widetilde{H} \boldsymbol{y}, \tag{3.6}$$

where $H(\boldsymbol{x}_j, \boldsymbol{u}_j) = K(\boldsymbol{x}_j, \boldsymbol{x}) S_j$ for $j = 1, \cdots, n$.

Using (3.6) the predicted regression function given $\boldsymbol{x}_t$ at new location $\boldsymbol{u}_t$ is given as follows:

$$\widehat{f}(\boldsymbol{x}_t, \boldsymbol{u}_t) = (K(\boldsymbol{x}_t, \boldsymbol{x}), 1) S_t \boldsymbol{y},$$

where $S_t = \begin{pmatrix} W_t K + \lambda \mathrm{I} & W_t \mathbf{1} \\ \mathbf{1}' W_t K & \mathbf{1}' W_t \mathbf{1} \end{pmatrix}^{-1} \begin{pmatrix} W_t \\ \mathbf{1}' W_t \end{pmatrix}$, W_t is the diagonal matrix of weight function $w_{ti} = \exp(-d_{ti}/h)$, d_{ti} is the distance between location $\boldsymbol{u}_t$ and $\boldsymbol{u}_i$ for $i = 1, \cdots, n$. The functional structures of the GWLS-SVM is characterized by the hyperparameters, which are the bandwidth parameter of the weight function, penalty parameter and kernel parameter. To choose optimal values of hyperparameters we use LOO-CV function as follows:

$$CV(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{f}_i^{(-i)}(\boldsymbol{\theta}))^2,$$

where $\boldsymbol{\theta}$ is a set of hyper-parameters, that is, kernel and penalty parameters, and $\widehat{f}_i^{(-i)}(\boldsymbol{\theta})$ is the predicted value of $f(\boldsymbol{x}_i)$ obtained from data without the ith observation. For example $\widehat{f}_1^{(-1)}$ is obtained as follows:

$$\widehat{f}_1^{(-1)} = (K(\boldsymbol{x}_1, \boldsymbol{x}^{(-1)}), 1) S_1^{(-1)} \boldsymbol{y}^{(-1)},$$

where $\boldsymbol{x}^{(-1)} = \{\, \boldsymbol{x}_i \}_{i=2}^{n}$, $S_1^{(-1)} = \begin{pmatrix} W_1^{(-1)}\boldsymbol{K}^{(-1)} + \lambda \mathrm{I} & W_1^{(-1)}\mathbf{1} \\ \mathbf{1}'W_1^{(-1)}K^{(-1)} & \mathbf{1}'W_1^{(-1)}\mathbf{1} \end{pmatrix}^{-1} \begin{pmatrix} W_1^{(-1)} \\ \mathbf{1}'W_1^{(-1)} \end{pmatrix}$, $\boldsymbol{y}^{(-1)} = (y_2, \cdots, y_n)'$, $K^{(-1)} = K(\boldsymbol{x}^{(-1)}, \boldsymbol{x}^{(-1)})$, $W_1^{(-1)}$ is the diagonal matrix of weight function $w_{1i} = \exp(-d_{1i}/h)$, d_{1i} is the distance between location $\boldsymbol{u}_1$ and $\boldsymbol{u}_i$ for $i = 2 \cdots, n$.

Thus GCV function is obtained as follows:

$$GCV(\boldsymbol{\theta}) = \frac{n \sum_{i=1}^{n} (y_i - \widehat{f}_i(\boldsymbol{\theta}))^2}{(n - tr(\widetilde{H}))^2},$$

where $\widehat{f}_i(\boldsymbol{\theta})$ is the estimated regression function given $(\boldsymbol{x}_i, \boldsymbol{u}_i)$ and $\widetilde{H}$ is the hat matrix such that $\widehat{\boldsymbol{f}} = \widetilde{H}\boldsymbol{y}$ in (3.6).

4. Case studies

In this section, we illustrate the prediction performance of GWLS-SVM for geographically weighted mean estimation. It is meaningful to compare the existing methods such as LS-SVM and GWR which shows good performance on geographically weighted mean estimation. The experiments were conducted in MATLAB environment. We use the acid neutralizing capacity (ANC) dataset (Salvati *et al.*, 2011) and the growth dataset (Fernadez *et al.*, 2001). ANC dataset was obtained by a sample of 338 lakes drawn from a total population of 21,026 lakes in the northeastern states of the USA. The survey was carried out between 1991 and 1995. The input variables of ANC dataset are the elevation and the geographical coordinates of the centroid of each lake (longitude and latitude), and the response is ANC of each lake. The input variables of the growth dataset are the initial GDP (gross domestic product) level and the geographical coordinates (longitude and latitude), and the response is GDP growth rate 1960-1980 for 72 countries.

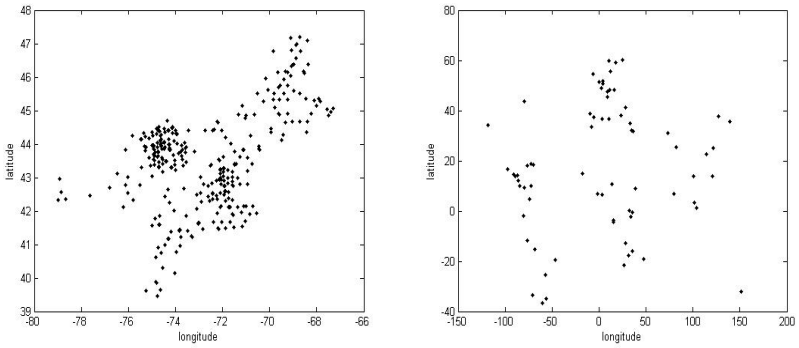

Figure 4.1 Locations of 338 lakes sampled in the northeastern states of the USA (Left) and locations of 72 countries (Right)

To compare the prediction performance we split datasets into 67% training dataset (225 lakes, 48 countries) and 30% test dataset (113 lakes, 24 countries). We replicate the above process 100 times. For the standard LS-SVM, the elevation, longitude and latitude were used as the input variables. The response, input variables and the geographical coordinates were standardized respectively as follows:

$$y_i = \frac{y_i - \text{mean}(\boldsymbol{y})}{\text{std}(\boldsymbol{y})},\ x_{ik} = \frac{x_{ik} - \min(\boldsymbol{x}_{.k})}{\max(\boldsymbol{x}_{.k}) - \min(\boldsymbol{x}_{.k})} \text{ and } u_{ik} = \frac{u_{ik} - \min(\boldsymbol{u}_{.k})}{\max(\boldsymbol{u}_{.k}) - \min(\boldsymbol{u}_{.k})}$$

where $\boldsymbol{x}_{.k} = (x_{1k}, \cdots, x_{nk})'$ and $\boldsymbol{u}_{.k} = (u_{1k}, \cdots, u_{nk})'$.

As the prediction performance metric, the predicted mean squared error is utilized,

$$PMSE = \frac{1}{100}\sum_{i=1}^{100}(y_i - \widehat{f}_i)^2.$$

The Gaussian kernel function and exponetial function are utilized in the case studies.

$$K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\left(-\frac{1}{\sigma_1^2}||\boldsymbol{x}_1 - \boldsymbol{x}_2||^2\right) \text{ and } w_{12} = \exp\left(-\frac{1}{\sigma_2^2}||\boldsymbol{u}_1 - \boldsymbol{u}_2||\right).$$

In LS-SVM (GWLS-SVM) the optimal values of the penalty parameter, kernel parameter (and the bandwidth parameter) are obtained from training dataset by GCV function. In GWR the optimal value of the bandwidth parameter is obtained from training dataset by GCV function. The results are shown in Table 4.1 and Figure 4.2. We can see that GWLS-SVM overally shows better prediction performance than the others. However we need to check whether or not it is statistically significant. To do this we obtain one tailed p-value of paired t-test for the nulll hypothesis that PMSE of LS-SVM or GWR is smaller than or equal to that of GWLS-SVM. From Table 4.2 we can see that PMSE of GWLS-SVM is smaller than PMSE of other methods.

Table 4.1 Averages of 100 PMSEs and their standard errors

ANC dataset			
	LS-SVM	GWR	GWLS-SVM
average	0.9953	0.6837	0.6579
std error	0.0250	0.0177	0.0180
Growth dataset			
	LS-SVM	GWR	GWLS-SVM
average	0.9959	0.9655	0.4662
std error	0.0251	0.0871	0.0156

Table 4.2 Test statistics of paired t-test for PMSE (one-tailed p-value in parenthesis)

ANC dataset		Growth dataset	
LS-SVM	GWR	LS-SVM	GWR
13.9777 (<0.001)	7.2926 (<0.001)	26.2778 (<0.001)	5.8119 (<0.001)

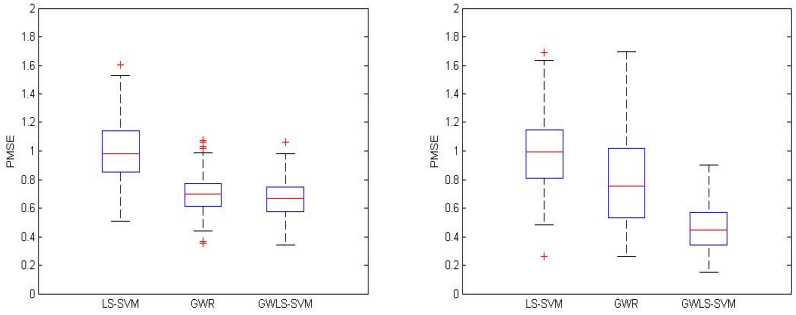

Figure 4.2 Boxplots of 100 PMSEs obtained by 3 methods for ANC dataset (Left) and Growth dataset (Right)

5. Conclusion

In this paper, we have studied how LS-SVM based method work for geographically weighted mean estimation. The proposed method take over all advantages of LS-SVM that capture nonlinearities in the data, that have good prediction ability, and that are useful tools when the functional form of the relationship between the response and input variables is left unspecified (linear or nonlinear) and the data are characterized by complex patterns of spatial dependence. In particular, the proposed method can be easily used without heavy computations under high-dimensional covariate settings.

Through the case studies, we conclude that LS-SVM based method derives the satisfying solutions to estimate the geographically weighted mean.

References

Anselin, L. (1992). *Spatial econometrics: Method and models*, Kluwer Academic Publishers, Boston.

Brunsdon, C. and Fotheringham, A. S. (1999). Some notes on Parametric significance tests for geographically weighted regression. *Journal of Regional Science*, **39**, 497-524.

Fotheringham, A. S., Brunsdon, C. and Charlton, M. (2002). *Geographically weighted regressio*, John Wiley and Sons, Chichester, UK.

Fotheringham, A. S., Charlton, M. E. and Brunsdon, C. (1996). The geography of parameter space: An investigation of spatial non-stationarity. *International Journal of Geographical Information Science*, **10**, 605-627

Hwang, C. and Shim, J. (2016). Deep LS-SVM for regression. *Journal of the Korean Data & Information Science Society*, **27** 827-833.

Hwang, C., Bae, J. and Shim, J. (2016). Robust varying coefficient model using L1 penalized locally weighted regression. *Journal of the Korean Data & Information Science Society*, **27**, 1059-1066.

Kimeldorf, G. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, **33**, 82-95.

Mika, S., Ratsch, G., Weston, J., Schddoto lkopf, B. and Muller, K. R. (1999). *Fisher discriminant analysis with kernels. IEEE International Workshop on Neural Networks for Signal Processing IX*, Madison, WI, August, 41-48.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, **65**, 386-408.

Salvati, N., Ranalli, M. G. and Pratesi, M. (2011). Small area estimation of the mean using non-parametric M-quantile regression: A comparison when a linear mixed model does not hold. *Journal of Statistical Computation and Simulation*, **81**, 945-964.

Saunders, C., Gammerman, A. and Vork, V. (1998). Ridge regression learning algorithm in dual variable. *Proceedings of the 15th International Conference on Machine Learning*, San Fransisco, CA, Morgan Kaufmann, 515-521.

Shim, J. Kim, C. and Hwang, C. (2011). Semiparametric least squares support vector machine for accelerated failure time model. *Journal of Korean Statistical Society*, **40**, 75-83.

Smola, A. J., Friess, T. T. and Schlkopf, B. (1998). Semiparametric support vector and linear programming machines. *Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems*, Cambridge, MA, MIT Press, 585-591.

Suykens, J. A. K. and Vandewalle, J. (1999). Least squares support vector machine classiers. *Neural Processing Letters*, **9**, 293-300.

Suykens, J. A. K., Vandewalle, J. and DeMoor, B. (2001). Optimal control by least squares support vector machines. *Neural Networks*, **14**, 23-35.

Vapnik, V. (1995). The nature of statistical learning theory, Springer, Berlin.Wahba, G. (1990). *Spline models for observational data, CMMS-NSF Regional Conference Series in Applied Mathematics*, **59**, SIAM, Philadelphia.

Xu, J., Zhang, X. and Li, Y. (2001). Kernel MSE algorithm: A unified framework for KFD, LS-SVM. *Proceedings of the International Joint Conference on Neural Networks, IJCNN 2001*, Washington, DC, IEEE, 1486-1491.

Zhang, L. J. (2004). Modeling spatial variation in tree diameter-height relationships. *Forest Ecology and Management*, **189**, 317-329.