*American Journal of Computational Mathematics*, 2012, 2, 56-60
http://dx.doi.org/10.4236/ajcm.2012.21007 Published Online March 2012 (http://www.SciRP.org/journal/ajcm)

# An Improved Kriging Interpolation Technique Based on SVM and Its Recovery Experiment in Oceanic Missing Data*

**Zhisong Huang, Huizan Wang, Ren Zhang**
Institute of Meteorology, PLA University of Science and Technology, Nanjing, China
Email: hzsong123@126.com

Received December 24, 2011; revised January 25, 2012; accepted February 3, 2012

## ABSTRACT

In Kriging interpolation, the types of variogram model are very finite, which make the variogram very difficult to describe the spatial distributional characteristics of true data. In order to overcome its shortage, an improved interpolation called Support Vector Machine-Kriging interpolation (SVM-Kriging) was proposed in this paper. The SVM-Kriging uses Least Square Support Vector Machine (LS-SVM) to fit the variogram, which needn't select the basic variogram model and can directly get the optimal variogram of real interpolated field by using SVM to fit the variogram curve automatically. Based on GODAS data, by using the proposed SVM-Kriging and the general Kriging based on other traditional variogram models, the interpolation test was carried out and the interpolated results were analyzed contrastively. The test show that the variogram of SVM-Kriging can avoid the subjectivity of selecting the type of variogram models and the SVM-Kriging is better than the general Kriging based on other variogram model as a whole. Therefore, the SVM-Kriging is a good and adaptive interpolation method.

**Keywords:** Least Square Support Vector Machine; Kriging Interpolation; Variogram; SVM-Kriging

## 1. Introduction

Kriging is a method of interpolation which predicts unknown values from data observed at known locations, and it minimizes the error of predicted values which are estimated by spatial distribution of the predicted values. Kriging uses variogram to express the spatial variation. The key problem of Kriging is selection of variogram model, which determines the spatial interpolation accuracy. Variogram model includes linear model, exponential model, Gaussian model, spherical model and so on. In a general way, the reasonable variogram model is selected based on the cloud pictures of variogram distribution. However, this general method of variogram selection is subjective, and may not select the optimal variogram model.

In order to overcome the shortcoming of variogram model selection, an improved interpolation method called Support Vector Machine-Kriging interpolation (SVM-Kriging) was proposed. SVM-Kriging uses least square support vector machine (LS-SVM) to fit the variogram, which needn't select the basic variogram model and can directly get the optimal variogram of real interpolated field by using SVM to fit the variogram curve automatically. The variogram of SVM-Kriging come from the real data, so it can avoid the subjectivity and arbitrariness of selecting the type of variogram models and improve the interpolated results. Based on GODAS data, the proposed SVM-Kriging was compared with other general vario- gram models in this paper.

*Project supported by the National Natural Science Foundation of China (No. 41276036).

## 2. General Kriging

### 2.1. Basic Idea

Let $Z(x)$ be the value of the variable $Z$ at a point $x$. Given the $n$ measurements $Z(x_1),\cdots,Z(x_n)$ at known locations $x_1,\cdots,x_n$, you want to obtain an estimate of $Z^*$ at an unsampled location $x_0$.

The Kriging estimator is given by weighed linear combinations of the available samples [1]:

$$Z^*(x_0)=\sum_{i=1}^{n}\lambda_i Z(x_i) \tag{1}$$

Considering the unbiasedness condition yields:

$$\sum_{i=1}^{n}\lambda_i=1 \tag{2}$$

Under this condition, the variance of estimate error of expression can be simplified as follows:

 

$$\begin{aligned} S &= Var\left[Z^*(x_0) - Z(x_0)\right] \\ &= 2\sum_{i=1}^{n}\lambda_i\gamma(x_i - x_0) - \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\gamma(x_i - x_j) \end{aligned} \tag{3}$$

where $\gamma(h)$ is variogram. Under the restricted condition (2), in order to make the estimate variance minimum, by introducing Lagrange multiplier, the Kriging linear equations, by which the weight can be calculated, is derived as follows:

$$\begin{cases} \sum_{i=1}^{n}\lambda_i\gamma(x_i - x_j) + \mu = \gamma(x_j - x_0)\ (j = 1, \cdots, n) \\ \sum_{i=1}^{n}\lambda_i = 1 \end{cases} \tag{4}$$

where $\gamma(x_i - x_j)$ is the value of variogram between location $x_i$ and location $x_j$. All weights $\lambda_i$ and Lagrange multiplier $\mu$ can be calculated, and then $Z^*$ can be obtained by (1).

## 2.2. Variogram

The key problem of Kriging is to determine the law of variable changed with space and then to estimate the unknown value based on the known samples. This law is variogram. Variogram is used to describe the spatial structure of variable.

The variogram of samples, which is also called experimental variogram, can be calculated by the following formula:

$$\gamma^*(h) = \frac{1}{2N_h}\sum_{i=1}^{N_h}\left[Z(x_i + h) - Z(x_i)\right]^2 \tag{5}$$

where $N_h$ is the number of pairs separated by vector $h$, vector $h$ is lag distance, $x_i$ is the starting location and $x_i + h$ is the ending location. If $\gamma$ is only dependent on the length of lag distance but not its direction, $\gamma$ is isotropic, also the variable $Z$ is isotropic. For the sake of simplicity, we only consider isotropy of Kriging.

Generally speaking, after the experimental variogram is computed by (5), we usually observe the distribution of variogram and then identify a reasonable variogram model. After that we use least square method to fit variogram in accordance with the principle of minimum variance estimate, which yields fitting curve called empirical variogram. Variogram model is usually a basic model or a linear combination of several basic models. The common theoretical model of variogram mainly includes linear model, spherical model, exponential model, Gaussian model and so on. Their mathematic expressions are as follows:

a) Linear model:

$$\gamma(h) = C_0 + C_1 h$$

b) Spherical model:

$$\gamma(h) = \begin{cases} C_0 + C_1[1.5(h/a) - 0.5(h/a)^3], & 0 \le h \le a \\ C_0 + C_1, & h > a \end{cases}$$

c) Exponential model:

$$\gamma(h) = C_0 + C_1\left(1 - e^{-h/a}\right)$$

d) Gaussian model:

$$\gamma(h) = C_0 + C_1\left[1 - e^{-(h/a)^2}\right]$$

where $C_0, C_1, a$ are unknown parameters that should be identified by least squares.

## 2.3. Existing Problem

At present, there are not very good methods to select the variogram models in general interpolation. In a general way, the reasonable variogram model is often identified based on the comparison of different variogram models. However, this method is time-consuming (because it need compute Kriging interpolation several times) and the types of variogram models are very finite, which make the variogram very difficult to describe the spatial distributional characteristics of true data .The general methods contain some subjectivity and arbitrariness. In order to overcome the existing problem, the least-square Support Vector Machine (LS-SVM) was introduced to fit the experimental variogram, and then the shortcoming of variogram model selection can be avoided. Based on least-square support vector machine, it does not need to identify the type of basic variogram models but to fit the experimental variogram according to its own distribution picture directly.

# 3. Least Squares Support Vector Machine

Support Vector Machines, as a novel learning machine developed by Vapnik and his coworkers in 1995 [2], have been introduced for pattern recognition and regression. Least squares support vector machine (LS-SVM), originally proposed by Suykens in 2001 [3], is one kind of SVM. LS-SVM transforms inequality constraints of standard SVM to equality constraints.

Given a training data set of samples, $(x_1, y_1), \cdots, (x_l, y_l) \in R^n \times R$, where $x_i \in R^n$ is the *i*-th input data. The LS-SVM approach aims at identify the parameters of the model:

$$f(x) = w^T\Phi(x) + b \tag{6}$$

where $w$ is weight vector, $\Phi(x)$ is a function which maps the input data into a higher dimensional feature space.

LS-SVM is to solve the following optimization problem:

 

$$\begin{cases} \min \quad \frac{1}{2}\|w\|^2 + C\frac{1}{2}\left(\sum_{i=1}^{p}\xi_i^2\right) \\ y_i = w^T\Phi(x_i) + b + \xi_i \quad \xi_i \geq 0, i = 1,\cdots,l \end{cases} \tag{7}$$

where $\xi_i \in R$ denotes regression error for sample $x_i$, $b$ is a bias scalar, and $C$ is a given positive constant. After introducing Lagrangian multipliers $a_i$, based on Karush-Kuhn-Tuchker conditions, we obtain the nonlinear function based on LS-SVM:

$$f(x) = w^T\Phi(x) + b = \sum_{i=1}^{l} a_i K(X_i, X) + b \tag{8}$$

where $K(X_i, X_j)$ is kernel function.

A Large number of imitation tests have shown that Radial Basis Function (RBF) kernel function is more effective than others as a whole, so we select RBF kernel as the kernel of LS-SVM.

$$K(x_i, x_j) = \exp\left(-\sigma\|x_i - x_j\|^2\right), \sigma > 0 .$$

Note that $\sigma$ and $C$ are two parameters. They can be optimized by Genetic Algorithms [4].

## 4. Support Vector Machine-Kriging Method

There are mainly three steps in SVM-Kriging method as follows:

1) Use (5) to compute experimental variogram $\gamma^*(h)$;

2) Use LS-SVM with parameters optimized by Genetic Algorithm to fit the experimental variogram $\gamma^*(h)$, and then get $\gamma(h)$;

3) Use (4) to get the weights $\lambda_1,\cdots,\lambda_n$ for every point $x_0$ and then obtain the estimated value $Z^*(x_0)$ at $x_0$ by using (1).

## 5. Application in Oceanic Missing Data Recovery

In order to test the effect of improved Kriging based on LS-SVM, this paper takes data derived Global Ocean Data Assimilation System (GODAS) as the experimental data. GODAS is developed at National Centers for Environmental Prediction (NCEP) Centers, and GODAS data are time series of monthly average derived from GODAS operational datasets. The area coverage is [120.5˚E-71.5˚W, 60˚S-58˚N]. We selected four representative months January, April, July, and October of 2006 as test time, sea surface salinity (5 m deep in this paper) and the sea surface height relative to Geoid (sshg) as variables.

### 5.1. Interpolation Process

As the process of different months and different variables are similar, we take sshg in January 2006 as an example to introduce interpolation in detail and draw a compareson of different variogram model.The area coverage of sshg is [120.5˚E-71.5˚W, 60˚S-58˚N], and the spatial resolution is 2˚ × 2˚.The number of total grid points are 5100 (85 × 60), including a total of 4215 points in ocean availably. We selected 75% of them (3161) randomly as cross-validation data, and the remaining data (1054) are taken as known observed data. **Figure 1** shows the remaining data after take out of 75% of available data.

Firstly, compute the experimental variogram. The experimental variogram was computed by (5) based on 1054 known data (**Figure 2**). Secondly, obtain empirical variogram. The empirical variogram was obtained by fitting based on different variogram models (**Figure 3**). At last, obtain the estimated values. Kriging interpolations

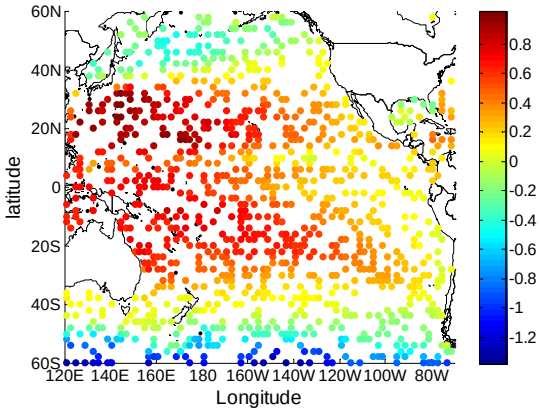


**Figure 1. The remaining available sshg data.**

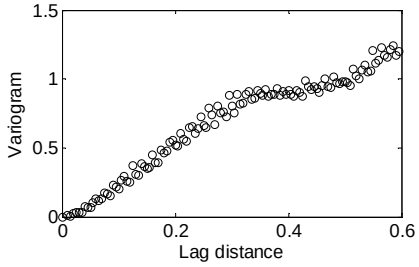


**Figure 2. Experimental variogram.**

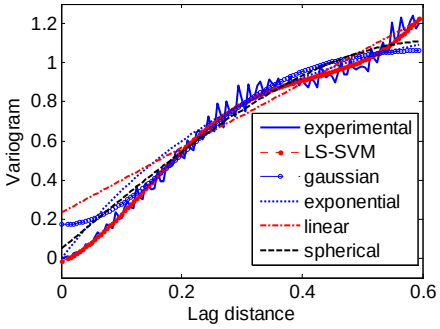


**Figure 3. Experimental and different model variograms.**

 

were carried out by using the obtained empirical variograms of different variogram models.

As the other interpolation processes of different months and variables are similar, the descriptions about them are omitted.

## 5.2. Cross-Validation Results

In order to analyze the interpolation quality, an evaluation by cross validation has been carried out. The cross validation starts by eliminating some available sample points randomly, the Kriging methods are then applied to estimate the missing value on basis of the remaining known sample points. The errors between estimated values and the observed values at missing points are calculated. The kinds of quantitative Error calculated mainly contain mean error (ME), mean absolute error (MAE), root mean square prediction error (RMSPE).The definitions of ME, MAE and RMSPE are as follows:

$$\mathrm{ME} = \frac{1}{n}\sum_{k=1}^{n}\left[Z^*(x_k) - Z(x_k)\right] \tag{9}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{k=1}^{n}\left|Z^*(x_k) - Z(x_k)\right| \tag{10}$$

$$\mathrm{RMSPE} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left[Z^*(x_k) - Z(x_k)\right]^2} \tag{11}$$

RMSPE is a quantity used to compare the quality of different interpolation methods. RMSPE is smaller, the interpolation method is better. Based on differential variogram models, the Kriging interpolations with selected data are carried out. Following tables are error results of different months and different variables by cross-validation (See **Table 1** and **Table 2**). The values with underline denote that they are minimal in their columns.

The error results have shown that the RMSPE of SVM-Kriging is smaller than others as a whole and the SVM-Kriging method is of adaptive advantage for real data of different variables at different time.

## 6. Conclusion

Kriging interpolated results are dependent on the selection of variogram model largely, and different variogram model will lead to different results. In this paper, the sea surface salinity and the sea surface height relative to Geoid are applied to test the interpolation effect. Tests have shown that the empirical variogram based on LS-SVM model improve the Kriging results by contrast with other variogram models, and it also avoids the subjectivity of selecting the type of basic variogram models. Therefore, the improved SVM-Kriging is a good and adaptive interpolation method for the real data, especially for the

**Table 1. The comparison of interpolation quality based on different variogram models to the sea surface height relative to Geoid (meter).**

(a)

| January | ME($\times10^{-3}$) | MAE($\times10^{-2}$) | RMSPE($\times10^{-2}$) |
|---|---|---|---|
| Gaussian model | –0.9866 | 5.6032 | 8.0570 |
| exponential model | –0.5896 | 3.0390 | 5.5112 |
| linear model | –0.6265 | 3.0344 | 5.4946 |
| spherical model | –0.5960 | 3.0323 | 5.4929 |
| LS-SVM | –0.5070 | 3.0263 | 5.4650 |

(b)

| April | ME($\times10^{-3}$) | MAE($\times10^{-2}$) | RMSPE($\times10^{-2}$) |
|---|---|---|---|
| Gaussian model | 0.5304 | 4.8351 | 7.1637 |
| exponential model | 3.3320 | 2.5465 | 4.7073 |
| linear model | 3.2360 | 2.5436 | 4.6932 |
| spherical model | 3.2675 | 2.5436 | 4.6933 |
| LS-SVM | 3.2381 | 2.5364 | 4.6640 |

(c)

| July | ME($\times10^{-3}$) | MAE($\times10^{-2}$) | RMSPE($\times10^{-2}$) |
|---|---|---|---|
| Gaussian model | 1.0249 | 5.6143 | 8.0398 |
| exponential model | –1.0010 | 2.6297 | 4.7588 |
| linear model | –1.0566 | 2.6295 | 4.7624 |
| spherical model | –1.0286 | 2.6268 | 4.7582 |
| LS-SVM | –1.0159 | 2.6279 | 4.7582 |

(d)

| October | ME($\times10^{-3}$) | MAE($\times10^{-2}$) | RMSPE($\times10^{-2}$) |
|---|---|---|---|
| Gaussian model | –1.0656 | 5.7940 | 8.4461 |
| exponential model | 0.5359 | 2.8382 | 5.0358 |
| linear model | 0.4651 | 2.8382 | 5.0338 |
| spherical model | 0.4993 | 2.8375 | 5.0323 |
| LS-SVM | 0.5095 | 2.8376 | 5.0332 |

(e)

| Average in total | ME($\times10^{-3}$) | MAE($\times10^{-2}$) | RMSPE($\times10^{-2}$) |
|---|---|---|---|
| Gaussian model | 0.9019 | 5.4617 | 7.9266 |
| exponential model | 1.3646 | 2.7633 | 5.0033 |
| linear model | 1.3460 | 2.7614 | 4.9960 |
| spherical model | 1.3479 | 2.7601 | 4.9942 |
| LS-SVM | 1.3176 | 2.7570 | 4.9801 |

 

**Table 2. The comparison of interpolation quality based on different variogram models to sea surface salinity (Kg/kg).**

(a)

| January | ME($\times10^{-6}$) | MAE($\times10^{-5}$) | RMSPE($\times10^{-4}$) |
|---|---|---|---|
| Gaussian model | –3.1830 | 11.392 | 2.2437 |
| exponential model | –1.4831 | 6.3555 | 1.3900 |
| linear model | –1.5149 | 6.3536 | 1.3891 |
| spherical model | –1.4191 | 6.3435 | 1.3878 |
| LS-SVM | –1.2199 | 6.3259 | 1.3856 |

(b)

| April | ME($\times10^{-6}$) | MAE($\times10^{-5}$) | RMSPE($\times10^{-4}$) |
|---|---|---|---|
| Gaussian model | 6.2451 | 10.928 | 1.8567 |
| exponential model | 8.1825 | 6.7105 | 1.4267 |
| linear model | 7.7905 | 6.7071 | 1.4299 |
| spherical model | 7.7854 | 6.6926 | 1.4265 |
| LS-SVM | 7.1287 | 6.6496 | 1.4206 |

(c)

| July | ME($\times10^{-6}$) | MAE($\times10^{-5}$) | RMSPE($\times10^{-4}$) |
|---|---|---|---|
| Gaussian model | 3.0329 | 10.568 | 2.0281 |
| exponential model | 0.0949 | 6.4795 | 1.4361 |
| linear model | –0.2326 | 6.4566 | 1.4316 |
| spherical model | –0.1731 | 6.4548 | 1.4306 |
| LS-SVM | 0.0353 | 6.3882 | 1.4111 |

(d)

| October | ME($\times10^{-6}$) | MAE($\times10^{-5}$) | RMSPE($\times10^{-4}$) |
|---|---|---|---|
| Gaussian model | 10.092 | 11.410 | 2.2117 |
| exponential model | 8.6852 | 6.6331 | 1.5610 |
| linear model | 8.7131 | 6.6254 | 1.5585 |
| spherical model | 8.7829 | 6.6176 | 1.5573 |
| LS-SVM | 9.2950 | 6.6164 | 1.5540 |

(e)

| Average in total | ME($\times10^{-6}$) | MAE($\times10^{-5}$) | RMSPE($\times10^{-4}$) |
|---|---|---|---|
| Gaussian model | 5.6382 | 11.075 | 2.0851 |
| exponential model | 4.6114 | 6.5446 | 1.4535 |
| linear model | 4.5628 | 6.5356 | 1.4523 |
| spherical model | 4.5401 | 6.5271 | 1.4505 |
| LS-SVM | 4.4197 | 6.4950 | 1.4428 |

data containing complicated spatial structure

## REFERENCES

[1] R. D. Zhang, "Spatial Variability Theory and Its Application," Science Press, Beijing, 2005

[2] V. N. Vapnik, "The Nature of Statistical Learning Theory," Springer-Verlag, New York, 1995.

[3] J. A. K. Suykens, J. Vannderwalle and B. D. Moor, "Optimal Control by Least Squares Support Vector Machine," *Neural Network*, Vol. 14, No. 1, 2001, pp. 23-35. doi:10.1016/S0893-6080(00)00077-0

[4] K. F. Liu and R. Zhang, "Minimal Risk Based on the Structure of the Support Vector Machine Methods and Its Application in Numerical Forecast Optimization of Subtropical High," *Journal of Basic Science and Engineering*, Vol. 14, No. 3, 2006, pp. 384-389.

 