Bulgarian Journal of Agricultural Science, 19 (No 6) 2013, 1327-1336
Agricultural Academy

PREDICTION OF MULTIDIMENSIONAL TIME SERIES BASED ON GS-RSR-SVR AND ITS APPLICATION IN AGRICULTURAL ECONOMY

Y. G. XIE[1,2], H. Y. ZHANG[1,2,3*], H. Y. WANG[4], L. F. WANG[1,3] and ZH. M. YUAN[1,3*]
[1] *Hunan Provincial Key Laboratory of Crop Germplasm Innovation and Utilization, Changsha 410128, China*
[2] *Hunan Agricultural University, College of Information Science and Technology, Changsha 410128, China*
[3] *Hunan Provincial Key Laboratory for Biology and Control of Plant Diseases and Insect Pests, Changsha 410128, China*
[4] *Kansas State University, Department of Statistics, Manhattan, Kansas 66506, USA*

Abstract

XIE, Y. G., H. Y. ZHANG, H. Y. WANG, L. F. WANG and ZH. M. YUAN, 2013. Prediction of multidimensional time series based on GS-RSR-SVR and its application in agricultural economy. *Bulg. J. Agric. Sci.,* 19: 1327-1336

This paper proposes a method that creatively applies a Geo-statistics tool (GS) to complete fast and adequate order determination and introduces a novel algorithm, named Reasonable Sample Rejection (RSR) to realize rational sample selection. Then, combined with Support Vector Machine Regression (SVR), a high precision non-linear prediction method named GS-RSR-SVR is proposed for multidimensional time series. The main steps of the novel method includes: 1) determine the order for the dependent variable of the training samples based on one-dimensional GS aftereffect duration (range), 2) screen the independent variables according to Leave-One-Out Cross Validation (LOOCV) based on the minimum Mean Squared Error (MSE), 3) reject some oldest training samples based on the minimum correlation coefficient of fitting absolute relative error of training sets of different rejected sizes and sample number. Three real-world datasets was used to test the effectiveness of GS-RSR-SVR. The results show that GS-RSR-SVR has higher prediction precision and more stable prediction ability than MLR, ARIMA, CAR, BPNN, SVR and SVR-CAR.

Key words: multidimensional time series; geo-statistics tool; reasonable sample rejection; support vector machine regression; prediction

Abbreviations: GS: geo-statistics; RSR: reasonable sample rejection; SVR: support vector machine regression; LOOCV: leave-one-out cross validation; MSE: mean squared error; MLR: multi-level recursive; ARIMA: autoregressive integrated moving average; CARMA: controlled autoregressive integrating moving average; CAR: controlled autoregressive; ANN: artificial neural network; IGAOV: index of gross agricultural output value; MAPE: mean absolute percentage error; APE: absolute percentage error; LLD: log-linear de-trending

Introduction

There are a large number of multidimensional time series data with single dependent variable and multiple independent variables existing in agricultural science, economics and other fields. The accurate prediction of multidimensional time series is still facing great challenge, mainly because it requires multiple tasks including regression analysis of the effect of environmental factors, time series analysis of the massive implicit dynamic features of dependent variable, and nonlinear analysis to deal with the complicated nonlinearity (Wu and Hong, 1999, 2002). Accurate forecast is the basis of cognition and decision-making. It is important to develop highly accurate decision and prediction methods for multidimensional time series.

Early classical multidimensional time series analysis models, such as multi-level recursive prediction model (MLR) (Han, 2001), autoregressive integrated moving average model (ARIMA), controlled autoregressive integrating moving average (CARMA) and its simplified version controlled

* Corresponding authors: hongyan_zhang6@aliyun.com; zhmyuan@sina.com

autoregressive (CAR) models are linear methods that have limited application ability (Box and Jenkins, 1970; Hannan, 1980; Boker and Keviczky, 1982; Deng and Guo, 1985). Artificial neural network (ANN) has the merits of self-learning, self-adoption and excellent nonlinear approximation ability. However, ANN is often over-trained and easily falls into local minimum based on empirical risk minimization (Robert, 2008). Based on structural risk minimization, support vector machine regression (SVR) could effectively avoid the problems of local minimum and over-training (Vapnik, 1995; Corinna and Vladimir, 1995; Ping and Wei, 2005). However, SVR has not taken into account the temporal dynamic characteristics of the dependent variable for multidimensional time series, which often hinders accurate prediction.

Combining CAR with SVR, Yuan et al. (Yuan, 2008) and Zhang et al. (Zhang, 2007) developed a nonlinear prediction method SVR-CAR that integrates time series analysis and nonlinear modeling into regression analysis to determine the order and screens the variables nonlinearly with leave-one-out cross validation (LOOCV). In fact, the order is just like the embedding dimension of phase space reconstruction (Wang, 2006) and the order determination is the reconstruction process of independent variables. For example, for a dataset, if the order is estimated to be of order *a*, it indicates that the original samples are affected by their *a* previous samples. Therefore, the independent variables and dependent variables of previous samples should be embedded as the independent variables for the original samples. The prediction accuracy and generalization ability of SVR-CAR are apparently better than traditional SVR.

However, SVR-CAR still has the following drawbacks: 1) Repeated computation and gradual comparison of MSE based on LOOCV were used in SVR-CAR. This is too complicated to simultaneously determine the same order of both dependent variable and independent variables. It could easily lead to information redundancy and inadequate final order. Additionally, this kind of order determination method increases the modeling time and decreases prediction precision. Examination of one-dimensional GS semi-variogram aftereffect duration (corresponding to the spatial pattern range *a*) suggests that the current observed value is only relative to the *a* previous observed values, which provides a new idea to directly determine the final order for dependent variable. Meanwhile, in order to reduce the information redundancy, the independent variables are determined to be of order one by default. 2) The selection of training samples is beyond consideration in SVR-CAR. As is well known, SVR has excellent performance for small samples. For predicting time series by SVR-CAR with one-step prediction, the sequence stationarization, order determination, variable selection and training model construction need to be conducted independently in each step. Namely, the training set is always dynamically changing. The training set gets larger and larger as time goes on, which makes the training time unacceptable. More importantly, it is unreasonable to build a training model with all accumulated training samples for a certain test sample. Researches on prediction of parameters for chaotic time series show that simply increasing the number of training samples could decrease the generalization ability and prediction precision. The selection of training samples has a significant effect on time series forecasting (Wang, 2008). In this study, we propose a novel algorithm to sequentially remove some old samples of a certain step-length and then form several new training sets in different sample size. Through comparing the correlation coefficient of fitting absolute relative error and time order of these new training sets, some oldest training samples can be eliminated uninterruptedly, which provides a new idea for sample selection of time series.

The order determination and sample selection of training set are important but difficult in the field of time series prediction. In this work, we propose a multidimensional time series prediction method GS-RSR-SVR that integrates a new order determination method based on geo-statistics tools and a novel sample selection method RSR into SVR. We apply the proposed method to three real-world datasets to verify the feasibility and effectiveness of the method in comparison with competing methods.

Materials and Methods

Experimental data

Dataset A (Table 1): Index of gross agricultural output value (IGAOV) (y) and its correlative factors (agricultural labor force (x1, ten thousand persons), crop yield (x2, ten thousand tons) and agricultural taxes (x3, 100 million *yuan* RMB)) from 1952 to 1980 in China. Dataset B (Table 2): IGAOV (y) and its correlative factors (rural workers (x1, ten thousand persons), crop yield (x2, ten thousand tons) and agricultural taxes (x3, 100 million *yuan* RMB)) from 1978 to 2008 in China. Observed values of IGAOV in Dataset A and Dataset B are calculated regarding 1952(=100) as the base period. Due to the limitations in materials, 'rural workers' is used to replace 'agricultural labor force' in Dataset B. It is noteworthy that there are some nuances existing in IGAOV from 1978 to 1980 in dataset A and B as the result of constant revision of China statistical yearbooks.

Dataset C (Table 3): Grain yield (y, ten thousand tons) and its correlative factors (rural workers (x1, ten thousand persons), farm machinery production (x2, ten thousand kilowatt), effective irrigation area (x3, one thousand hectares), sown area of grain crops (x4, one thousand hectares), fertilizer use(x5, ten thousand tons), rural power consumption (x6, one hundred million kilowatt hours), afflicted area (x7, one thousand hectares)) from 1985 to 2011 in China.

Dataset A was quoted from DPS Data Processing System for Practical Statistics (Tang and Feng, 2002). Dataset B was adopted from the China Rural Statistical Yearbook 2009 and China statistical yearbooks over the years. Dataset C was adopted from the China Rural Statistical Yearbook 2012.

It is noteworthy that there are some nuances existing in IGVAO from 1978 to 1980 in dataset A and B as the result of constant revision of China statistical yearbooks.

Evaluation indicator

To avoid the contingency of a single test sample, this study considers prediction of 10 consecutive samples at the end of each dataset. In order to show the actual prediction ability of GS-RSR-SVR, one-step prediction is used for forecasting of each test sample. Specifically, to predict for a test sample *i*, this sample and the samples observed afterward cannot take part in the training process. However, the sample *i* should be entered into training set for predicting the sample *i+1*. Three indicators evaluated prediction results: mean squared error (MSE), mean absolute percentage error (MAPE), and absolute percentage error (APE).

$$MSE = \frac{\sum (y_i - y_i')^2}{n}, \qquad (1)$$

Table 1
Index of gross agricultural output value and its correlative factors from 1952 to 1980 in China

Year	y	x_1	x_2	x_3
1952	100	17317	16392	27
1953	103.1	17748	16683	27.1
1954	106.6	18152	16952	32.8
1955	114.7	18593	18394	30.5
1956	120.5	18545	19275	29.7
1957	124.8	19310	19505	29.7
1958	127.8	15492	20000	32.6
1959	110.4	16273	17000	33
1960	96.4	17019	14350	28
1961	94.1	19749	14750	21.7
1962	99.9	21278	16000	22.8
1963	111.6	21968	17000	24
1964	126.7	22803	18750	25.9
1965	137.1	23398	19453	25.8
1966	149	24299	21400	29.6
1967	151.2	25167	21782	29
1968	147.5	26065	20906	30
1969	149.2	27119	21097	29.6
1970	166.3	27814	23996	32
1971	171.4	28400	25014	30.9
1972	171.1	28286	24048	28.4
1973	185.5	28861	26494	30.5
1974	193.2	29222	27527	30.1
1975	202.1	29460	28452	29.5
1976	207.1	29448	28631	29.1
1977	210.6	29345	28273	29.3
1978	229.6	29426	30477	28.4
1979	249.4	29425	33212	29.5
1980	259.1	30211	32056	27.7

Table 2
Index of gross agricultural output value and its correlative factors from 1978 to 2008 in China

Year	*y*	x_1	x_2	x_3
1978	191.3	30638	30477	28.4
1979	204.2	31025	33212	29.5
1980	203.6	31836	32056	27.7
1981	217.4	32672	32502	28.4
1982	241.2	33867	35450	29.4
1983	261.7	34690	38728	32.8
1984	291.7	35968	40731	34.6
1985	291.2	37065	37911	42.1
1986	299.1	37990	39151	44.5
1987	318.3	39000	40298	49.4
1988	322.4	40067	39408	52.5
1989	330.4	40939	40755	84.9
1990	356.7	47708	44624	87.9
1991	360.1	48026	43529	90.7
1992	375.3	48291	44266	119.2
1993	394.9	48546	45649	125.7
1994	407.5	48802	44510	231.5
1995	439.7	49025	46662	278.1
1996	474	49028	50454	369.5
1997	495.2	49039	49417	397.5
1998	519.6	49021	51230	398.8
1999	542	48982	50839	423.5
2000	549.6	48934	46218	465.3
2001	569.4	49085	45264	481.7
2002	591.6	48960	45706	717.9
2003	591.6	48793	43070	871.8
2004	641.9	48724	46947	902.2
2005	668.2	48494	48402	936.4
2006	704.2	48090	49748	1084
2007	731.7	47640	50160	1439.1
2008	766.7	47270	52871	1689.4

$$MAPE = \frac{\sum |y_i - y_i'| / y_i}{n} \times 100\,, \tag{2}$$

$$APE(i) = \frac{|y_i - y_i'|}{y_i} \times 100\ , i = 1, 2, ..., n - 1, n, \tag{3}$$

where y_i is the observed value, y_i' is the predicted value, and n is the number of test samples. The MSE is only available for the same dataset while MAPE is also suitable for different datasets. The main evaluation indicator is MSE for the same dataset. This is because when MAPE and MSE give opposite rankings for two methods, it can be illustrated that the method with smaller MSE gives better prediction. APE is used to evaluate the prediction stability of a certain method.

In this study, LIBSVM2.9 (Chang and Lin, 2011) was used to conduct SVR, and Radial Basis Function was chosen to be the kernel function. GS-RSR-SVR was designed on MATLAB2009 by compiling program and calling LIBSVM. Other prediction methods to compare with GS-RSR-SVR include the MLR, CAR, ARIMA(p, d, q) (for dataset A and dataset B: p=d=q=2; for dataset C: p=d=q=4), BPNN (the hidden layer number is 1; the node number of input layer is equal to the number of independent variables *n*; the minimum training rate is 0.1; the permissible error is 0.0001; the maximum number of iterations is 1000; the number of the hidden layer nodes is 2**n*+1.) given by DPS6.55, SVR-CAR designed on MATLAB2009 by compiling program and calling LIBSVM.

Experimental Procedure

Sequence stationarization of the dependent variable

Sequence stationarization is an important premise for time series forecasting. It can greatly simplify the process of time series analysis and improve the estimation accuracy of

Table 3
Grain yield and its correlative factors from 1985 to 2011 in China

Year	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1985	37910.8	37065	20912.5	44035.9	108845	1775.8	508.9	44365
1986	39151.2	37990	22950.0	44225.8	110933	1930.6	586.7	47135
1987	40297.7	39000	24836.0	44403.0	111268	1999.3	658.8	42086
1988	39408.1	40067	26575.0	44375.9	110123	2141.5	712.0	50874
1989	40754.9	40939	28067.0	44917.2	112205	2357.1	790.5	46991
1990	44624.3	47708	28707.7	47403.1	113466	2590.3	844.5	38474
1991	43529.3	48026	29388.6	47822.1	112314	2805.1	963.2	55472
1992	44265.8	48291	30308.4	48590.1	110560	2930.2	1106.9	51332
1993	45648.8	48546	31816.6	48727.9	110509	3151.9	1244.9	48827
1994	44510.1	48802	33802.5	48759.1	109544	3317.9	1473.9	55046
1995	46661.8	49025	36118.1	49281.2	110060	3593.7	1655.7	45824
1996	50453.5	49028	38546.9	50381.4	112548	3827.9	1812.7	46991
1997	49417.1	49039	42015.6	51238.5	112912	3980.7	1980.1	53427
1998	51229.5	49021	45207.7	52295.6	113787	4083.7	2042.2	50145
1999	50838.6	48982	48996.1	53158.4	113161	4124.3	2173.4	49980
2000	46217.5	48934	52573.6	53820.3	108463	4146.4	2421.3	54688
2001	45263.7	48674	55172.1	54249.4	106080	4253.8	2610.8	52215
2002	45705.8	48121	57929.9	54354.9	103891	4339.4	2993.4	46946
2003	43069.5	47506	60386.5	54014.2	99410	4411.6	3432.9	54506
2004	46946.9	46971	64027.9	54478.4	101606	4636.6	3933.0	37106
2005	48402.2	46258	68397.8	55029.3	104278	4766.2	4375.7	38818
2006	49804.2	45348	72522.1	55750.5	104958	4927.7	4895.8	41091
2007	50160.3	44368	76589.6	56518.3	105638	5107.8	5509.9	48992
2008	52870.9	43461	82190.4	58471.7	106793	5239.0	5713.2	39990
2009	53082.1	42506	87496.1	59261.4	108986	5404.4	6104.4	47214
2010	54647.7	41418	92780.5	60347.7	109876	5561.7	6632.3	37426
2011	57120.8	40506	97734.7	61681.6	110573	5704.2	7139.6	32471

characteristic statistics (Tang and Xie, 2004). Proper methods should be chosen to stationarize the sequence. In this study, each of the three time series data has a significant rise, log-linear de-trending (LLD) can be used to stationarize the original data [14].

$$\ln y_t = a + b * t\,, \tag{4}$$

The stationarized time series is interpreted as equation:

$$y_t\,' = \ln y_t - (a + b * t)\,, \tag{5}$$

Order determination of the original independent variables

Time series has the characteristics of aftereffect. Specifically, the dependent variable y_t is not only related to the original independent variables $x_{t,\ k}$, $k=1,\ 2,\ \ldots,\ m$, but also relative to previous dependent variable y_{t-1}, y_{t-2}, ..., y_{t-e} and independent variables $x_{t-1,\ k}$, $x_{t-2,\ k}$, ..., $x_{t-e,\ k}$. Therefore, it is necessary to take into account previously observed dependent variable and independent variables as predictors for a certain prediction. There is another problem, however. The effect on y_t brought by independent variables $x_{t-i,\ k}$, $k=1,\ 2,\ \ldots,\ m$ is already partly reflected in y_{t-i}. Consequently, the proposed model prescribes that independent variables should be determined to be of order one.

Order determination of the original dependent variable by one-dimensional GS

Geo-statistics is a branch of applied statistics that concentrates on the description of spatial patterns and estimating values at unsampled locations. It is usually used to analyze natural phenomenon of both structural properties and randomness in spatial distributions and in the temporal domain based on the theory of regionalized variable and with semi-variogram function as the main tool. Underlying this approach is the expectation that, on average, samples close together have more similar values than those that are farther apart (Andrew et al., 1993). For a stationary dependent variable sequence $y(t)$, $t=1,\ 2,\ \ldots,\ n$, the semi-variogram $r(h)$ as a function of the separation distance h can be described as:

$$r(h) = \frac{1}{2N(h)} \sum_{t=1}^{N(h)} \left[y(t) - y(t+h) \right]^2 , \tag{6}$$

where $N(h)$ is the number of pairs of samples separated by h. The term $y(t)$ and $y(t+h)$ represent the observed values of the dependent variable at times t and $t+h$, respectively. There is a general rule for the semi-variogram construction in practical application, i.e., in order to ensure the $N(h)$ to be big enough, the max separation distance h should be less than half the width of the sampling space (Li et al., 1998; Dong, 2011).

For a time series with reducing autocorrelation as the separation distance increases, the semi-variogram function $r(h)$ has an increasing pattern. That is, $r(h)$ starts with small values for small h, and increases as the distance h increases, and then usually reaches a maximum value at some separation distance or becomes constant after some separation distance. The separation distance corresponding to the maximum value is called range a, which means that samples are unrelated or discontinuous when their separation distance is larger than a (Andrew et al., 1993). That is, y_t is only affected by the samples with separation distance no more than a. Therefore, the dependent variable should be determined to be of order a.

For instance, suppose $(y_t, x_{t,j})$ is the training set for a certain prediction, $t=1,2,\ \ldots\ldots,n$; $j=1,2,\ldots\ldots,m$. This training set contains n samples and m independent variables. First, independent variables are determined to be of order one, which results in n-1 samples and $2m$ independent variables in training, set. Further, the dependent variable is determined to be of order a according to one-dimensional GS. This turns the original training set into a new dataset with n-1-a samples and $2m+a$ independent variables. Comparing with original dataset, the new dataset takes the aftereffect of time series into account.

Data scaling

Data scaling aids the choosing of arguments during the training by SVR and the speed of solving SVM (Chang and Lin, 2011). Each independent variable was scaled into [-1, 1] according to equation:

$$x_i\,' = -1 + 2(x_i - x_{\min}) / (x_{\max} - x_{\min})\,, \tag{7}$$

where x_i' is the scaled data, x_i is the original data, x_{max} and x_{min} are the maximum and minimum in x_i respectively.

Nonlinear selection of independent variables

After order determination, the independent variables of the newly formed training set sharply increases leading to some redundant information. It is necessary to eliminate such independent variables, which may have negative impact on prediction results. To achieve this goal, we perform the following backward elimination procedure. We start by initializing Ω to be the collection of independent variables of the remaining training set after order determination. Ω will be updated as the algorithm below proceeds.

1) Denote k to be the number of variables in Ω.

2) With all k variables in Ω, obtain the MSE value with LOOCV using SVR. Denote the MSE value as MSE(k).

3) Leave out the ith variable and use the remaining k-1 variables in LOOCV with SVR to obtain the MSE_{-i}. Perform this for all $i=1,\ 2,\ \ldots,\ k$.

4) If $\min\{\text{MSE}_{-i},\ 1 \le i \le k\} > \text{MSE}(k)$, skip 5) and 6) and go to step 7).

5) Let *j* be the variable in the candidate list in *Ω* such that $\mathrm{MSE}_{-j} = \min\{\mathrm{MSE}_{-i}, 1 \le i \le k\}$. Remove the *j*th variable from the candidate list *Ω* and change the variable *k* to have value *k-1*.

6) Repeat steps 1) – 5).

7) Report all the variables that are in *Ω*. These are the final list of independent variables to be used for following analysis.

Sample selection by RSR

Samples that are far away in time from the sample to be predicted are usually not as helpful in prediction as those that are close. Sometimes, inclusion of such old samples can even deteriorate the prediction results. We perform the following procedure to identify which old samples to be discarded before further analysis. Initialize *n* to be the number of training samples.

1) Starting from the first sample of the training set, discard the contiguous *b* oldest samples successively to form a group of training set models of size *n, n-b,n-2b,……,n-i*b (i<=n/b-1)*, respectively. We call *b* the step-length, whose value can be assigned according to the size of training set. Generally, if the number of training samples is no more than 30, we suggest that *b* is assigned to be 1.

2) Fit the dependent variable of each training set model with the LOOCV by SVR.

3) Calculate the absolute relative error between the fitted and observed values of each training set model.

4) Obtain a set of correlation coefficients *R(j), j=1, 2, … , i, i+1* through correlation analysis of the absolute relative errors and sample number from all training set models. If *R(j)* is positive, it indicates that the model's fitting relative error increases over time. These models cannot be used. If *R(j) is* negative, it means that the model can fit better for new samples or future samples. These models are helpful. However, the correlation coefficients are not comparable due to different training sample sizes. So we recommend converting the correlation coefficient by equation:

$$r(j) = \frac{R(j)}{r_{0.01,n-2}} \qquad j = 1,2,...,i,i+1, \tag{8}$$

where *R(j)* is the correlation coefficient and $r_{0.01,n-2}$ is the correlation coefficient test critical value with *n-2* degrees of freedom at 0.01 significant level which can be obtained from Schedule 10, Methods of Experimental Statistics (Gai, 2007). The *r(j)* is the relative correlation coefficient independent of the training sample size. The smaller the r(j) is, the faster the absolute relative error decreases. The training set model corresponding to the min *r(j)* is chosen for further analysis.

Prediction by SVR

It has been known that some arguments(*c, g, p*) are important for SVR. These arguments will affect the accuracy and speed of prediction. In this paper, the 10-fold CV method is used for arguments optimization. Then the model fitted with the optimal arguments is used for prediction by SVR.

Results and Discussion

To predict the sample in 1971 of Dataset A, there are 19 samples and three independent variables in the original training set. The following analyses were conducted: (1) The Log-linear de-trending (LLD) method was used to stationarize the original dependent variable. Then, the semi-variogram function of the stationarized dependent variable from 1952 to 1970 is given in Figure 1.

The semi-variogram values *r(h)* are small for low values of *h*, and then increase with increasing distance *h* until the maximum is obtained when *h* equals to six. So the dependent variable was determined to be of order six. Meanwhile, the independent variables were determined to be of order one. Then, the training set is transformed into a new data set with 12 samples and 12 independent variables (see section ***Order determination of the original dependent variable by one-dimensional*** and ***Order determination of the original independent variables***). (2) Based on minimizing the MSE, the training set was screened and six independent variables were selected by SVR with LOOCV method. (3) According to RSR, the five oldest training samples were rejected reasonably with step-length 1. Therefore, the training set was reduced to contain only seven samples. The one-step prediction results of the last 10 years of Dataset A are given in Table 4.

To predict the sample in 1999 of dataset B, there are 21 samples and three independent variables in the original training set. (1) The original dependent variable was stationarized

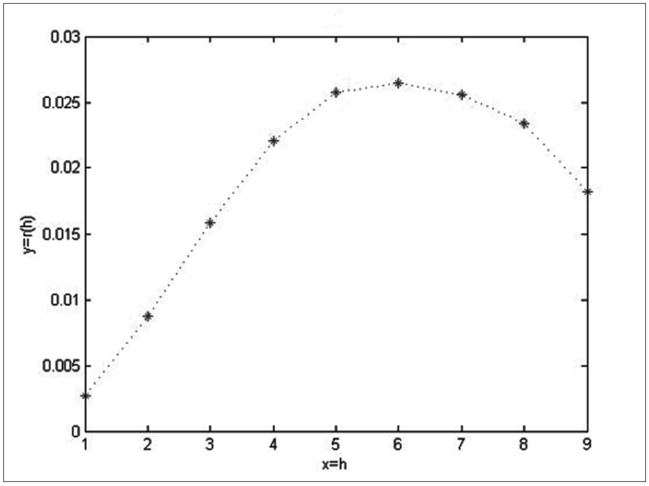

Fig. 1. The GS semi-variogram function curve of the dependent variable of IGVAO from 1952 to 1970 in China

with LLD too. Then, the semi-variogram function of the stationarized dependent variable from 1978 to 1998 is depicted in Figure 2.

The semi-variogram values *r(h)* reach the maximum as the *h* is equal to six. So the dependent variable was determined to be of order six and the independent variables were determined to be of order one. Then, the training set was turned into a new dataset with 14 samples and 12 independent variables. (2) The training set was screened and six independent variables were retained. (3) The four oldest samples were eliminated reasonably with step-length 2 according to RSR. The one-step prediction results from 1999 to 2008 of Dataset B are in Table 5.

To predict the sample in 2002 of dataset C, there are 17 samples and seven independent variables in the original training set. (1) The original dependent variable was stationarized with LLD too. Then, the semi-variogram function of the stationarized dependent variable from 1985 to 2001 is depicted in Figure 3.

The semi-variogram values *r(h)* reach the maximum as the *h* is equal to five. So the dependent variable was determined to be of order five and the independent variables were determined to be of order one. Then, the training set was turned into a new dataset with 11 samples and 19 independent variables. (2) The training set was screened and four independent variables were retained. (3) The three oldest

Table 4
Prediction results of the index of gross agricultural output value from 1971 to 1980

Obs	MLR		CAR		ARIMA		BPNN		SVR		SVR-CAR		GS-RSR-SVR	
	Pre	APE	Pre	APE	Pre	APE	Pre	APE	Pre	APE	Pre	APE	Pre	APE
171.4	176.1	2.74	189.1	10.33	175.1	2.16	161.3	5.89	171.3	0.6	176.1	2.74	173.2	1.05
171.1	167.3	2.22	169.9	0.70	173.4	1.34	164.5	3.86	163.9	4.2	166.9	2.45	168.8	1.34
185.5	185	0.27	186.1	0.32	168.1	9.38	168	9.43	173.6	6.42	180.9	2.48	186.6	0.59
193.2	191.8	0.72	167.8	13.15	192.3	0.47	178.6	7.56	179.3	7.19	193.2	0	187.3	3.05
202.1	198	2.03	201	0.54	195	3.51	186.5	7.72	198.4	1.83	199.9	1.09	198.7	1.68
207.1	200.3	3.28	206.2	0.43	203.3	1.83	192.8	6.9	196.2	5.26	204.5	1.26	206.6	0.24
210.6	199.7	5.18	207.8	1.33	206.4	1.99	197.1	6.41	197.3	6.32	207.1	1.66	209	0.76
229.6	217.5	5.27	222.7	3.01	209.1	8.93	205.1	10.67	205	10.71	219.3	4.49	227.6	0.87
249.1	242.2	2.77	249.3	0.08	235.5	5.51	216.5	13.19	234.1	6.13	240.4	3.61	250.7	0.64
259.1	236.9	8.57	248.5	4.09	256.3	1.08	232.3	10.34	226.8	12.47	248.9	3.94	248.5	4.09
MSE	91.2		112.7		102.6		369.6		257.9		37.4		17.7	
MAPE	3.31		3.39		3.63		8.2		6.05		2.4		1.42	

*Note: Obs is observed value, Pre is predicted value. Same obviations will be used in Table 5 and Table 6.

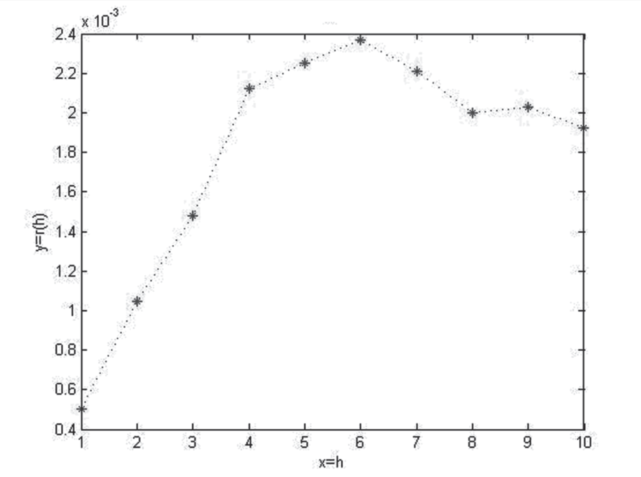

Fig. 2. The GS semi-variogram function curve of the dependent variable of IGVAO from 1978 to 1998 in China

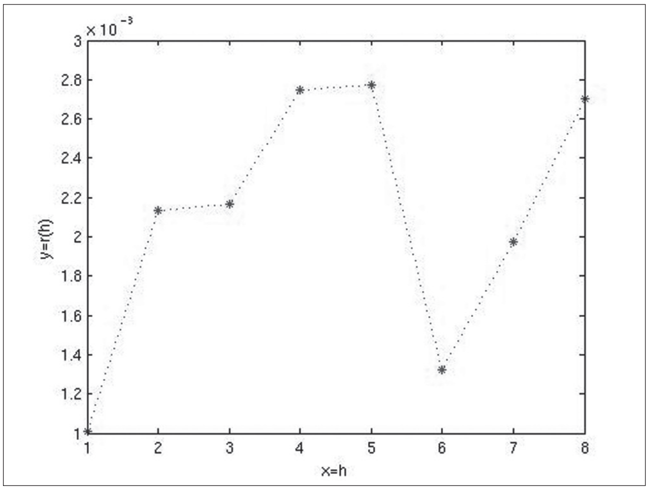

Fig. 3. The GS semi-variogram function curve of the dependent variable of grain yield from 1985 to 2001 in China

samples were eliminated reasonably according to RSR. The one-step prediction results from 2002 to 2011 of Dataset C are in Table 6.

According to Tables 4, 5 and 6, the MSE, MAPE and APE of GS-RSR-SVR are apparently better than those of all reference models. BPNN and SVR have poorer prediction accuracy than MLR, CAR and ARIMA for dataset A and dataset B. This indicates that the effect of independent variables on the dependent variable is more like a linear relationship in dataset A and dataset B. In addition, dataset C is more like a nonlinear data system. The results show that GS-RSR-SVR has good prediction accuracy for both linear and nonlinear data system, which has extensive application value.

CAR, SVR-CAR and GS-RSR-SVR simultaneously consider the effect of environmental factors and temporal dynamic characteristics of the dependent variable. Both CAR and SVR-CAR have to determine the order of the dependent variable and independent variables based on gradual comparison simultaneously. The processes are complicated and time-consuming. For prediction of the 30 test samples of the three datasets in this study, each training set was determined to be of order one by CAR and of orders 0~6 by SVR-CAR (only 2 training sets of Dataset B was determined to be of order six). On the other hand, GS-RSR-SVR stipulate that the original independent variables are determined to have order one and the order of the original dependent variable is deter-

Table 5
Prediction results of the index of gross agricultural output value from 1999 to 2008

Obs	MLR		CAR		ARIMA		BPNN		SVR		SVR-CAR		GS-RSR-SVR	
	Pre	APE	Pre	APE	Pre	APE	Pre	APE	Pre	APE	Pre	APE	Pre	APE
542	530.2	2.18	539.8	0.41	546.1	0.76	495.4	8.6	510.4	5.83	526.2	2.92	535.9	1.13
549.6	538.8	1.97	546.1	0.64	568.7	0.35	497.8	9.43	460.2	16.27	537.5	2.20	552	0.44
569.4	552.5	2.97	566.7	0.47	572.5	0.54	526.2	7.59	550.5	3.32	563.9	0.97	566.4	0.53
591.6	553	6.52	604	2.10	598	1.08	552.8	6.56	640.1	8.2	573.1	3.13	593.5	0.32
591.6	604.9	2.25	614.7	3.90	627.4	6.05	587.7	0.66	490.1	17.15	633	7	595.9	0.73
641.9	625.4	2.57	316.2	50.74	630.9	1.71	589.8	8.11	615	4.19	611.9	4.67	615.6	4.10
668.2	629.5	5.79	680.3	1.81	678	1.47	621.3	7.02	626.6	6.22	660.5	1.15	660.7	1.12
704.2	665.8	5.45	702.3	0.27	706.3	2.98	636.6	9.6	687	2.44	668.9	5.01	689.2	2.13
731.7	709.4	3.05	734.7	0.41	744.4	1.74	670.4	8.38	705.9	3.52	746.3	2	740.5	1.20
766.7	758.4	1.08	776.8	1.32	778.2	1.50	701.9	8.45	731.4	4.6	760.7	0.78	752.5	1.85
MSE	601.9		10705		222.9		2568.2		2666		493.7		132.6	
MAPE	3.38		6.21		1.86		7.44		7.18		2.98		1.35	

Table 6
Prediction results of the grain yield from 2002 to 2011 in China

Obs	MLR		ARIMA		BPNN		SVR		SVR-CAR		GS-RSR-SVR	
	Pre	APE	Pre	APE	Pre	APE	Pre	APE	Pre	APE	Pre	APE
45706	44803	1.98	45609	0.21	45611	0.21	44128	3.45	44673	2.26	44861	1.85
43070	43226	0.36	46012	6.83	43784	1.66	44177	2.57	41350	3.99	44389	3.06
46947	40230	14.31	43009	8.39	46602	0.73	45279	3.55	46693	0.54	46254	1.48
48402	45580	5.83	47148	2.59	47355	2.16	46925	3.05	47317	2.24	47659	1.54
49804	48120	3.38	49120	1.37	48627	2.36	48458	2.70	48403	2.81	48836	1.94
50160	48170	3.97	50357	0.39	49331	1.88	48746	2.82	50727	3.12	50283	0.25
52871	49230	6.89	50798	3.92	50460	4.56	52213	1.24	51048	1.75	52407	0.88
53082	51341	3.28	53452	0.7	52031	1.98	52775	0.58	53635	1.04	52432	1.22
54648	51793	5.22	53903	1.36	52967	3.41	53503	2.09	54581	0.12	53735	1.67
57121	55414	2.99	55203	3.36	54334	4.96	54649	4.33	54996	3.77	54458	4.66
MSE	8807031		3492000		2243390		2044500		1549177		1300100	
MAPE	4.82		2.91		2.39		2.64		2.17		1.85	

mined according to one-dimensional GS range. The process is simple and fast and the new independent variables have low information redundancy. In this study, the dependent variable of each training set in Dataset A was determined to be of order six in the 10 independent predictions. In dataset B, seven training sets were determined to be of order six, and three training sets were determined to be of order five. In dataset C, all training sets were determined to be of order five. In comparison to CAR and SVR-CAR, GS-RSR-SVR can provide a more stable and adequate order determination for the dependent variable. The actual independent prediction results further prove that order determination based on one-dimensional GS is feasible and reliable, and it has great significance to order determination of time series.

Theoretically, most machine learning methods give better prediction results when there are more samples in the training set. On the contrary, RSR advocates the reduction of redundant training samples. In this study, the sample size of almost half the training sets was reduced to less than 10 samples for the prediction of 20 test samples. This also provides an effective solution for time series forecasting with a limited number of samples.

Overall, GS-RSR-SVR generalizes better to future test samples with higher prediction precision and more stable prediction ability compared to present methods. This will lead to a promising future in the field of time series prediction.

Conclusion

GS-RSR-SVR is a combination of time series analysis, regression analysis and nonlinear modeling. It has the merit of fast and adequate order determination and reasonable sample selection. Compared to other time series prediction methods, GS-RSR-SVR has higher prediction precision, more stable prediction ability and stronger generalization ability. Three representative multidimensional time series datasets have been used to verify the feasibility of GS-RSR-SVR in this study. The prediction results show that the proposed method has great potential for multi-dimensional time series prediction in agricultural science, economics and so on. With the GS-RSR-SVR method, we found that reducing redundant samples by reasonable sample rejection could improve the prediction accuracy and stability. However, certain control over the number of sample rejections is necessary in practical applications since more sample rejection does not always produce better prediction result with the RSR. At least five or more training samples must be retained to ensure the statistical significance of the training set. Theoretically, independent variable selection affects training sample selection and vice versa. These two operations should be simultaneously conducted. Nevertheless, similar to other present methods, GS-RSR-SVR does not provide a reasonable solution to this problem yet. Possible association between the two processes is an important topic for future research.

Acknowledgements

This work was supported by grants from the Science Foundation for National Science and Technology Major Project (No. 2012BAD35B05), National Natural Science Funds of China (No. 61300130), PhD. Programs Foundation of Ministry of Education of China (No. 20124320110002) and Graduate Students'Scientific Research Innovation Project of Hunan Province (No. CX2012B307). The research of H. Y. Wang is supported by a grant from the Simons Foundation (#246077)..

Author contribution

Zhang and Yuan developed the algorithm; Xie designed the software and summarized the results; Xie and Wang L. F. drafted the manuscript; Zhang, Yuan and Wang H. Y. provided discussion and revised the manuscript. All authors have approved the final version of the manuscript.

References

Andrew, M. L., E. R. Richard and P. K. William, 1993. Geostatistics and geographic information systems in applied insect ecology. *Annu. Rev. Entomol.*, **38:** 303-327.

Box, Q. E. P. and G. M. Jenkins, 1970. Time Series Analysis: Forecasting and Control. *Holden-day Press, San Francisco*, 1pp. 1-175.

Boker, J. and L. Keviczky, 1982. Structural properties and structure estimation of vector difference equations. *International Journal of Control*, **36**: 461-476.

Corinna, C. and V. Vladimir, 1995. Support-vector networks. *Machine Learning*, **20:** 273-291.

Chang, C. C. and C. J. Lin, 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, **2**: 1-27.

Deng, Z. L. and Y. X. Guo, 1985. Dynamic System Analysis and Its Application, *Liaoning Science & Technology Press, Shenyang,* pp. 22-43 (Ch).

Dong, D. W., 2011. A study on temporal and spatial differentiation of land intensive use based on spatial sate mining and geo-statistical analyst--a case of Wuhan urban circle, *Master thesis*, *Central China Normal University*, *Wuhan,* China, pp. 27-39 (Ch).

Gai, J. Y., 2007. Methods of Experimental Statistics, *China Agriculture Press,* Beijing, 376 pp. (Ch).

Han, Z. G., 2001. The progress of theory and application of multi-level recursive method. *Control and Decision*, **16:** 129-132, 185 (Ch).

Hannan, E. J., 1980. The estimation of the order of an ARMA process. *Annals of Statistics*, **8**: 1071-1081.

Li, H. B., Z. Q. Wang and Q. C. Wang, 1998. Theory and methodology of spatial heterogeneity quantification. *Chinese Journal of Applied Ecology*, **9**: 651-657 (Ch).

Ping, F. P. and C. H. Wei, 2005. Support vector machines with simulated annealing algorithms electricity load forecasting. *Energy Conversion and Management*, **46:** 2669-2688 (Ch).

Robert, J. M., 2008. Application of partial mutual information variable selection to ANN forecasting of water quality in water distribution systems. *Environmental Modelling & Software*, **3**: 1-11.

Tang, Q. Y. and M. G. Feng, 2002. DPS data processing system for practical statistics, *Science Press*, *Beijing,* 989 pp. (Ch)

Tang, H. M. and Z. J. Xie, 2004. Stationarizing two classes of non-stationary processes by wavelet. *Acta Scientiarum Naturalium Universitatis Pekinensis*, **40:** 19-28 (Ch).

Vapnik, V. N., 1995. The Nature of Statistical Learning Theory. *Spring Verlag Press*, New York, pp. 3-47.

Wang, H. Y. and S. Lu, 2006. Nonlinear time series analysis and its applications. *Science Press*, *Beijing,* pp. 19-28 (Ch)

Wang, Y. S., 2008. Prediction of the chaotic time series from parameter varying systems using artificial neural networks. *Acta. Physica. Sinica.*, **57**: 6120-6131.

Wu, C. Z. and W. Hong, 1999. Multidimensional time series analysis on tree growth. *Chinese Journal of Applied Ecology*, **10:** 395-398 (Ch).

Wu, C. Z. and W. Hong, 2002. A proposed multidimensional time series model of individual age and diameter in tsuga longibracateatz. *Acta. Phytoecological Sinica.*, **26:** 403-407 (Ch).

Yuan, Z. M., Y. S. Zhang and J. Y. Xiong, 2008. Multidimensional time series analysis based on support vector machine regression and its application in agriculture. *Scientia. Agricultura. Sinica.*, **41:** 2485-2492 (Ch).

Zhang, Y. S., Z. M. Yuan, J. Y. Xiong and T. J. Zhou, 2007. Multidimensional time series analysis based on support vector regression and controlled autoregressive and its application in ecology. *Acta. Ecologica. Sinica.*, **27**: 2419-2424 (Ch).

Received January, 10, 2013; accepted for printing October, 2, 2013.