Temperature time series prediction based on autoregressive integrated moving average model

Huanhuan Zheng*, Yuxiu Bai, Yaqiong Zhang

School of Information Engineering, Yulin University, Yulin 719000, China

497759682@qq.com

ABSTRACT. *This paper establishes a prediction model for land and ocean temperature time series based on the improved autoregressive integrated moving average (ARIMA) model. First, the temperature time series was normalized and differenced before passing the stationarity test by augmented Dickey-Fuller (ADF) method, while the model parameters were determined by the autocorrelation coefficient and the partial autocorrelation coefficient. After that, the model was trained by the historical temperature data series, and applied to predict the temperatures in future. To validate the model, several experiments were conducted using the average land and ocean temperature data of Lawrence Berkeley National Laboratory. The results of the ARIMA-based model were contrasted against those of the support vector regression (SVR) and the random forest (RF). The comparison shows that the ARIMA-based model was 10%~30% smaller than the SVR and the RF in the values of RMSE and MAE, and 1%~10% higher in the value of* R^2*. This means our model outperformed the two benchmark algorithms.*

RÉSUMÉ. *Cet article établit un modèle de prévision pour les séries temporelles de température des terres et des océans sur la base du modèle amélioré à moyenne mobile intégré et autorégressif (ARIMA). Premièrement, les séries temporelles de température ont été normalisées et différenciées avant de réussir le test de stationnarité par la méthode améliorée Dickey-Fuller (ADF), tandis que les paramètres du modèle ont été déterminés par le coefficient d'autocorrélation et le coefficient d'auto-corrélation partielle. Après cela, le modèle a été formé par la série de données historiques de température et appliqué pour prévoir les températures dans le futur. Pour valider le modèle, plusieurs expériences ont été menées à l'aide des données de température moyenne des terres et des océans du Laboratoire national Lawrence Berkeley. Les résultats du modèle basé sur ARIMA ont été comparés à ceux de la régression vectorielle (SVR) et de la forêt aléatoire (RF). La comparaison montre que le modèle basé sur ARIMA était inférieur de 10% à 30% aux valeurs de SVR et de RF dans les valeurs de RMSE et MAE, et supérieur de 1% à 10% dans la valeur de* R^2*. Cela signifie que notre modèle a dépassé les deux algorithmes de référence.*

KEYWORDS: *autoregressive integrated moving average (ARIMA) model, temperature prediction, time series analysis, difference, stationarity test.*

MOTS-CLÉS: *modèle à moyenne mobile intégré et autorégressif (ARIMA), prévision de la température, analyse des séries temporelles, différence, test de stationnarité.*

DOI:10.3166/I2M.17.443-453

1. Introduction

The accurate measurement of land and ocean temperatures enable meteorologists to forge a better understanding of the meteorological changes and their influencing factors. For instance, the causes of steady-state temperature anomaly and the sea surface temperature departure in the western Pacific warm pool area can be derived inversely from the temperature time series (Mohammadi *et al.*, 2015; Mesbah & Soroush, 2016). The key to these studies lies in the analysis and prediction of the temperature time series. The common methods for this research point include extreme learning machine (Xue & Forman, 2015), vector machine regression (Erdemir & Ayata, 2016), data assimilation (Xu *et al.*, 2017), neural network (Korteby *et al.*, 2016; Shirvani *et al.*, 2015), and autoregressive integrated moving average (ARIMA) model (Das M., Ghosh, 2017; Wang *et al.*, 2016). For time sequence analysis in the general sense, the existing strategies range from Bayesian network (Rheinwalt *et al.*, 2016), echo state network (Tu and Yi, 2017), nonlinear time series (Grigorievskiy *et al.*, 2013), vector autoregression, multi-core extreme learning machines. Inspired by the previous research, this paper establishes an ARIMA-based prediction model for land and ocean temperatures according to the unique features of land and ocean temperature series. The model was proved effective and reliable through sufficient experiments.

2. ARIMA-based prediction model

For simplicity, the temperature time series is denoted as $\{x_i\}$ $(1 \le i \le n)$ with x_i being the temperature at time t_i.

2.1. Data preprocessing

(1) Normalization

All data were normalized at the beginning. The normalization limits the data to a certain scale, such that the algorithm can achieve desirable result. In addition, the normalization facilitates the performance evaluation of different algorithms, for different algorithms have the same data scale on different datasets. The normalization function can be expressed as:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

where x is the original value; x_{min} and x_{max} are the minimum and maximum values of the dataset, respectively; x' is the normalized value. After normalization, all temperature data fall in [0~1].

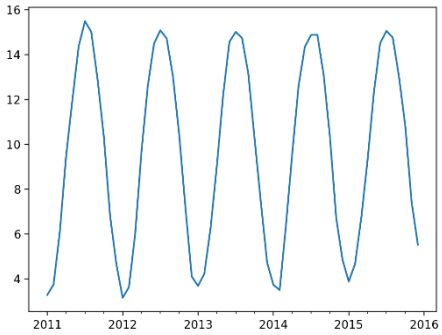

Figure 1. Average land temperature from 2011 to 2015

(2) *k*-step difference

Figure 1 shows an obvious periodicity in the data of the temperature series. The period is 12 months, exactly one year.

To fully extract the periodic information and trend effect from the temperature series, the data in the original series were subjected to *k*-step difference ($k = 12$ for the period is 12 months):

$$\nabla x_i = x_{i+k} - x_i \tag{2}$$

where $1 \le i \le n\text{-}12$ with n being the size of the dataset; $\{x_i\}$ is the original dataset; $\{\nabla x_i\}$ is the differenced dataset. It can be seen from Figure 2 that data is not stable.

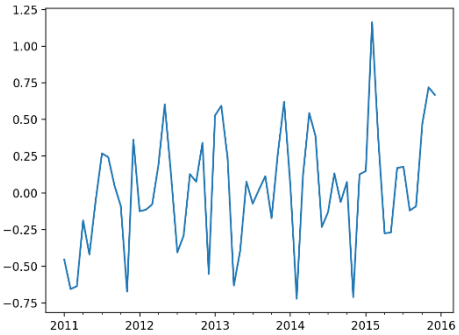

Figure 2. Average land temperature after 12-step difference

2.2. Stationarity test

The stationarity test is a must because only stationary series is applicable to the prediction model. Here, the augmented Dickey-Fuller (ADF) method is adopted for stationarity test:

For any time series process:

$$x_i = \phi_1 x_{i-1} + ... + \phi_p x_{i-p} + \varepsilon_i \tag{3}$$

The corresponding feature equation is:

$$\lambda^p - \phi_1 \lambda^{p-1} - ... - \phi_p = 0 \tag{4}$$

If all the eigenvalues of the equation are in a circle, i.e. $|\lambda_i|<1, i=1,2,\ldots p$, then the series is stationary; otherwise, the series is not stationary and satisfies $\sum_{i=1}^{p} \phi_p = 1$.

Assuming that $\rho=\phi_1+\phi_2+\ldots+\phi_p-1$, the hypotheses for the test can be established as:

H_0: $\rho=0$ (The series $\{x_i\}$ is not stationary.)

H_1: $\rho<0$ (The series $\{x_i\}$ is stationary.)

The ADF test statistic can be defined as:

$$\tau = \frac{\rho}{S(\rho)} \tag{5}$$

where $S(\hat{\rho})$ is the sample standard deviation of parameter ρ. The stationarity test results of Python's ADFuller are listed in Table 1 below.

Table 1. Stationarity test results

index		value
Test statistic τ		-15.788
P-test probability		1.109×10^{-28}
Lag K		24
Sample size		3119
Test critical values:	1%level	-3.4324
	5%level	-2.8625
	10%level	-2.5673

As shown in the table, the test probability was way below 1%, and the test statistic was smaller than the three critical values corresponding to the significance levels of 10%, 5%, and 1%. Thus, the series after the 12-step difference is stationary.

2.3. ARIMA-based temperature time series prediction model

The AR (p) model can be expressed as:

$$x_i = \varphi_0 + \phi_1 x_{i-1} + ... \phi_p x_{i-p} + \varepsilon_i \tag{6}$$

The MA (q) model can be expressed as:

$$x_i = \mu + \varepsilon - \theta_1 \varepsilon_{i-1} - \theta_2 \varepsilon_{i-2} - ... \theta_q \varepsilon_{i-q} \tag{7}$$

The centralized ARIMA (p, q, d) model can be expressed as:

$$x_i = \phi_1 x_{i-1} + ... \phi_p x_{i-p} + \varepsilon_i - \theta_1 \varepsilon_{i-1} - \theta_2 \varepsilon_{i-2} - ... \theta_q \varepsilon_{i-q} \tag{8}$$

where x_i is the temperature at time t_i; ε_i is the white noise; AR is the autoregression; p is the autoregressive coefficient; MA is the moving average; q is the number of moving average terms; d is the difference order. Among them, p and q can be determined according to the following table.

Table 2. Reference table for model parameter determination

model	Autocorrelation coefficient	Partial autocorrelation coefficient
AR(p)	trailing	P order truncated
MA(q)	Q order truncated	trailing
ARMA(p,q)	trailing	trailing

Each parameter in the above formulas should be trained by the training dataset to derive the optimal values of ϕ_i and θ_i. Then, these optimal values, coupled with the data of the historical moments, can be used to predict the temperatures at the other moments by the formulas.

The p and q values in the model can be determined by the autocorrelation coefficient and the partial autocorrelation coefficient. The autocorrelation coefficient can be expressed as:

$$r_k = \frac{\sum_{i=1}^{n-k}(x_i - \bar{x})(x_{i+k} - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{9}$$

where

$$k = 1,2,3\ldots n-1;\quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

The partial autocorrelation coefficient can be expressed as:

$$\psi_k = \frac{D_k}{D},\quad k = 1,2,3\ldots n-1 \tag{10}$$

where

$$D_k = \begin{vmatrix} 1 & r_1 & \cdots & r_{k-1} \\ r_1 & 1 & \cdots & r_{k-2} \\ \vdots & \vdots & & \vdots \\ r_{k-1} & r_{k-2} & \cdots & 1 \end{vmatrix};\quad D = \begin{vmatrix} 1 & r_1 & \cdots & r_1 \\ r_1 & 1 & \cdots & r_2 \\ \vdots & \vdots & & \vdots \\ r_{k-1} & r_{k-2} & \cdots & r_k \end{vmatrix}.$$

3. Experimental verification

3.1. Dataset

To ensure the effectiveness of our experiments, the data compiled and updated by the Berkeley Earth team under Lawrence Berkeley National Laboratory are adopted in this research. The data includes 3,192 entries on each month of the years 1750~2015. Each entry covers the average land temperature and the average ocean temperature of that month.

Two datasets were developed to fully validate our model, namely, the average land temperature forecast dataset and the average land and ocean temperature dataset. For the former dataset, the average monthly land temperatures of the years 1750~2010 were taken as the training dataset, while those of the years 2011~2015 were taken as the test dataset. For the latter dataset, the average monthly land and ocean temperatures of the years 1750~2010 were taken as the training dataset, while those of the years 2011~2015 were taken as the test dataset.

3.2. Benchmark algorithms

(1) Support vector regression (SVR)

The SVR hypothesis can tolerate any deviation less than ε between the model output f(x) and the true value y, creating an interval with a width of 2ε. If the training sample falls into this interval, the prediction is considered as correct. Hence, the SVR problem can be formalized as:

$$f(x) = \min_{w,b,\xi_i,\hat{\xi}_i} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{m} (\xi_i + \hat{\xi}_i) \tag{11}$$

S.t.

$$\begin{aligned} & f(x_i) - y_i \le \varepsilon + \xi_i, \\ & y_i - f(x_i) \le \varepsilon + \hat{\xi}_i, \\ & \xi_i \ge 0, \hat{\xi}_i \ge 0, i = 1,2,...,m \end{aligned} \tag{12}$$

where C is the regularization constant; ξ_i and $\hat{\xi}_i$ are relaxation variables.

(2) Random forest (RF)

The RF is an integrated learning algorithm involving a series of regression trees. The prediction values of all trees are averaged as the output of the algorithm. Through self-service resampling, this algorithm overcomes the over-fitting of regression trees and greatly enhances the model performance. In addition, the RF can process high-dimensional data and is thus suitable for numerical and categorical variables.

3.3. Evaluation indices

To compare the effects of different algorithms, the prediction results were evaluated by such three indices as the root mean square error (RMSE), the mean absolute error (MAE) and the coefficient of determination (R^2). For simplicity, the prediction value and the true value are respectively denoted as f_i and y_i in the calculation formulas of these indices.

(1) RMSE

The RMSE refers to the expected value of the square of the difference between the estimated value and the true value of the parameter. It can reflect the degree of change of the data. The RMSE value is negatively correlated with the accuracy of the prediction model based on the test data. The calculation formula of the RMSE can be expressed as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - y_i)^2} \tag{13}$$

(2) MAE

As the average of absolute errors, the MAE can accurately demonstrate the actual error in the prediction results. The calculation formula of the MAE can be expressed as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left| f_i - y_i \right| \tag{14}$$

(3) R^2

R^2 is the ratio of the regression sum of the squares to the total sum of squares. As a fitness statistic of the regression equation, this index reflects how much the deviation of the dependent variable y is explained by the estimated regression equation. The closer the value of R^2 is to one, the greater the ratio of the regression sum of the squares to the total sum of squares, the closer the regression line is to each observation point, the better the explanation of the true value deviation by the predicted value variation, and the higher the fitness of the regression. The calculation formula of R^2 can be expressed as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - f_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{15}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the average of the true values.

3.4. Experimental results and analysis

Figures 3 and 4 respectively compare the values predicted by the ARIMA-based model using the said two datasets against the true values. Tables 3 and 4 respectively list the values of the three evaluation indices obtained by the ARIMA-based model, the SVR and the RF using the said to datasets.

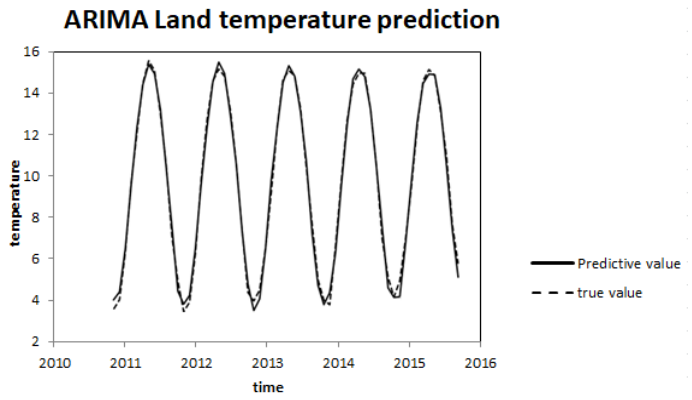

Figure 3. Land temperatures predicted by the ARIMA-based model

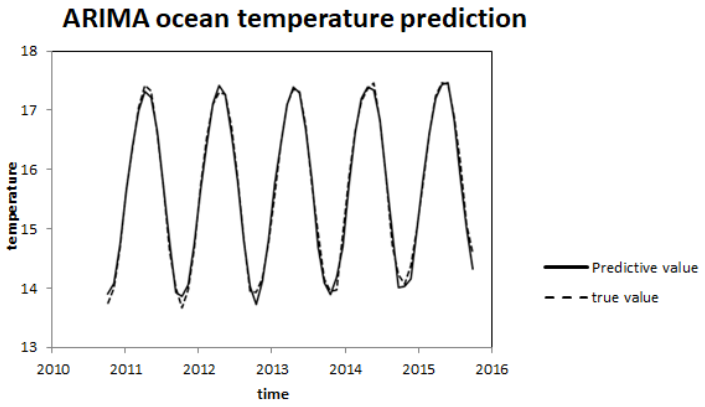

Figure 4. Ocean temperatures predicted by the ARIMA-based model

Table 3. Land temperatures predicted by obtained by the ARIMA-based model, the SVR and the RF

	ARIMA	SVR	RF
RMSE	0.3144	0.4662	0.3630
MAE	0.2580	0.4000	0.2800
R^2	0.9944	0.9876	0.9925

Table 4. Ocean temperatures predicted by obtained by the ARIMA-based model, the SVR and the RF

	ARIMA	SVR	RF
RMSE	0.1129	0.2562	0.3223
MAE	0.0860	0.2030	0.2200
R2	0.9919	0.9346	0.8925

From Figures 3 and 4, it can be seen that the land and ocean temperatures predicted by the ARIMA-based model were close to the true values, indicating that our model fulfills the prediction purpose. Tables 3 and 4 show that the ARIMA-based model was 10%~30% smaller than the SVR and the RF in the values of RMSE and MAE, and 1%~10% higher in the value of R^2. This means our model outperformed the two benchmark algorithms.

4. Conclusions

In view of the features of land and ocean temperature time series, this paper normalizes and differences the research data, and realizes the accurate prediction of land and ocean temperatures by the improved ARIMA algorithm. Compared with popular machine learning algorithms like the SVR and the RF, the proposed ARIMA-based prediction model achieves improvement in all evaluation indices, namely, the RMSE, the MAE and R^2. In future research, the influencing factors of land and ocean temperatures will be evaluated and included to enhance the prediction accuracy of our model.

Acknowledgement

This paper is made possible thanks to the generous support from High-Level Talent Research Startup Fund of Yulin University (Grant No.: 14GK46); Yulin Science and Technology Plan Project (Grant No.: 2016CXY-12-6); Special Scientific Research Plan of Shaanxi Provincial Department of Education (Grant No.: 18JK0902)

Reference

Das M., Ghosh S. K. (2017). Sembnet: A semantic Bayesian network for multivariate prediction of meteorological time series data. *Pattern Recognition Letters*, pp. 93. https://doi.org/10.1016/j.patrec.2017.01.002

Erdemir D., Ayata T. (2016). Prediction of temperature decreasing on a green roof by using artificial neural network. *Applied Thermal Engineering*, pp. 112. https://doi.org/10.1016/j.applthermaleng.2016.10.145

Grigorievskiy A., Miche Y., Ventelä A. M., Séverin E., Lendasse A. (2013). Long-term time series prediction using OP-ELM. *Neural Networks*, Vol. 51, pp. 50-56. https://doi.org/10.1016/j.neunet.2013.12.002

Korteby Y., Mahdi Y., Azizou A., Daoud K., Regdon G. (2016). Implementation of an artificial neural network as a PAT tool for the prediction of temperature distribution within a pharmaceutical fluidized bed granulator. *European Journal of Pharmaceutical Sciences*, pp. 88. https://doi.org/10.1016/j.ejps.2016.03.010

Mesbah M., Soroush E. (2016). Development of a least square support vector machine model for prediction of natural gas hydrate formation temperature. *Chinese Journal of Chemical Engineering*, Vol. 25, No. 9, pp. 1238-1248. https://doi.org/10.1016/j.cjche.2016.09.007

Mohammadi K., Shamshirband S., Motamedi S., Petković D., Hashim R., Gocic M. (2015). Extreme learning machine based prediction of daily dew point temperature. *Computers and Electronics in Agriculture*, Vol. 117, pp. 214-225. https://doi.org/10.1016/j.compag.2015.08.008

Rheinwalt A., Boers N., Marwan N., Kurths J., Hoffmann P., Gerstengarbe F. W., Werner P. (2016). Non-linear time series analysis of precipitation events using regional climate networks for Germany. *Climate Dynamics*, Vol. 46, No. 3-4, pp. 065-1074. https://doi.org/10.1007/s00382-015-2632-z

Shirvani A., Nazemosadat S. M. J., Kahya E. (2015). Analyses of the Persian gulf sea surface temperature: Prediction and detection of climate change signals. *Arabian Journal of Geosciences*, Vol. 8, No. 4, pp. 2121-2130. https://doi.org/ 10.1007/s12517-014-1278-1

Tu Y. D., Yi Y. P. (2017). Forecasting cointegrated nonstationary time series with time-varying variance. *Journal of Econometrics*, Vol. 196, No. 1. https://doi.org/10.1016/j.jeconom.2016.09.012

Wang L., Wang Z. G., Liu S. (2016). An effective multivariate time series classification approach using echo state network and adaptive differential evolution algorithm. *Expert Systems with Applications*, Vol. 43, pp. 237-249. https://doi.org/10.1016/j.eswa.2015.08.055

Xu B., Dan H. C., Li L. (2017). Temperature prediction model of asphalt pavement in cold regions based on an improved BP neural network. *Applied Thermal Engineering*, pp. 120. https://doi.org/10.1016/j.applthermaleng.2017.04.024

Xue Y., Forman B. A. (2015). Comparison of passive microwave brightness temperature prediction sensitivities over snow-covered land in North America using machine learning algorithms and the Advanced Microwave Scanning Radiometer. *Remote Sensing of Environment*, Vol. 170, pp. 153-165. https://doi.org/10.1016/j.rse.2015.09.009