WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 4, Number 1, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v4n1.35
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Predictive Modeling of Global Temperature Change Based on Artificial Intelligence Algorithms

Jincheng Zhang *, Shichang Sun, Runqing Wang

College of Computer Science and Engineering, Dalian Minzu University, Dalian, China
*Corresponding Author: Jincheng Zhang

ABSTRACT

As the greenhouse effect intensifies, global temperature changes are increasingly affecting all aspects of economic life. The purpose of this paper is to analyze the influencing factors of temperature to establish an analytical model of global temperature change. The Berkeley Earth Climate Database provides access to the data needed for the study. Pearson correlation analysis was used to quantify the strength of the correlations between temperature, location and other factors. After that, global temperatures were modelled and predicted by artificial intelligence algorithms such as ARIMA and linear regression prediction models, respectively. The results test that the model has good stability and can achieve accurate prediction function. The model can better help predict and respond to global temperature changes. The application of this AI algorithm can be used as a reference in other fields.

KEYWORDS

Global Warming; ARIMA Model; Linear Regression Models; Artificial Intelligence (AI)

1. INTRODUCTION

At present, the global warming problem is becoming more and more serious, which has become a major problem in the world, and the leading factors have also become the focus of the world [1-3]. In this paper, artificial intelligence algorithms are employed to investigate the factors affecting global temperature and to construct a predictive model for their analysis. The quantitative relationship between global temperature and variables such as temporal and spatial factors is revealed through Pearson correlation analysis. Two artificial intelligence algorithms, Autoregressive Integrated Moving Average (ARIMA) and Univariate Linear Regression Models, are utilized for modeling and forecasting changes in global temperature. By leveraging these methodologies, the model is able to achieve temperature predictions more consistently and accurately. This research not only establishes a scientific foundation for future predictions of global temperature changes but also offers insights into the application of artificial intelligence algorithms within related domains.

2. ANALYSIS OF FACTORS INFLUENCING GLOBAL TEMPERATURES

2.1. Data Collection and Processing

First, the factors influencing global temperatures can be analyzed, which asked us to find data and build a mathematical model to analyze the relationship between global temperature, time, and location, and we chose to use correlation analysis to calculate the relationship between global temperature and time and location. First, we found the city-related temperature data stored in txt. text on the official

website of Berkeley Earth Climate Database and used the code to implement the text into excel [4]. By obtaining the data for data cleaning, we divided the location into six temperature zone, which are the average temperature of the southern hemisphere, the average temperature of the northern hemisphere, the average temperature of the tropics, the average temperature of the southern temperate zone and the average temperature of the northern temperate zone, and the data were merged into one data set.

2.2. Model Selection and Implementation

Here we choose Pearson correlation analysis as the analysis model, Pearson correlation analysis can be used to study the relationship between quantification and how strong the relationship is, and the size of the correlation coefficient can measure how strong the correlation is. If P-value < 0.05, it means there is significance, and if P-value > 0.05, it means there is no significance between quantifications. If the correlation coefficient is positive, the variables are said to be positively correlated, if the correlation coefficient is negative, it is called negative correlation and the two variables have opposite changes, if the Pearson correlation coefficient is 0, the variables are said to be nonlinearly correlated with each other.

For the derived Pearson correlation coefficient, if it is between -1 and 1. The larger the absolute value of the coefficient γ Then the higher the correlation, if the absolute value of r is closer to 0, the weaker the correlation between the two. Define the covariance of the two variables x, y divided by the product of their standard deviations.

$$\rho(X,Y)=\frac{\operatorname{cov}(X,Y)}{\sigma X \sigma Y}=\frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma X \sigma Y} \tag{1}$$

The Pearson correlation coefficient equation is as follows:

$$\gamma=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}} \tag{2}$$

Then the Pearson correlation analysis is constructed as shown in Table.1, that the correlation is global temperature significant correlated with time and location.

Table 1. Comparison of power load forecasting of 403 line

	year	southavertem	northavertem	tropicalavertem	northtemzoneavertem	southtemzoneavertem
year	1(0.000**)	0.183 (0.000*")	0.056 (0.038")	0.255 (0.000"")	0.043 (0.109)	0.096 (0.000**)
southavertem	0.183(0.000**)	1 (0.000*")	-0.9 (0.000*)	-0.313 (0.000")	-0.919 (0.000*)	0.978 (0.000***)
northavertem	0.056(0.038")	-0.9 (0.000*")	1 (0.000"")	0.650 (0.000")	0.998 (0.000")	-0.962 (0.000*")
tropicalavertem	0.255(0.000"*")	-0.313 (0.000*)	0.650 (0.000"*")	1 (0.000"")	0.607 (0.000**")	-0,452 (0.000***)
northtemzoneavertem	0.043(0.109)	-0.919 (0.000***)	0.998 (0.000**)	0.607 (0.000***)	1 (0.000***)	-0.964 (0.000*")
southtemzoneavertem	0.096(0.000"""")	0.978 (0.000"")	-0.962 (0.000"""")	-0.452 (0.000")	-0.964 (0.000**)	1 (0.000**")

Where ***, **, * represent 1%, 5%, 10% significant levels.

For the relationship between global temperature and time and location before, we illustrate it by plotting a heat map, the darker the color the greater the correlation. As shown in Figure 1.

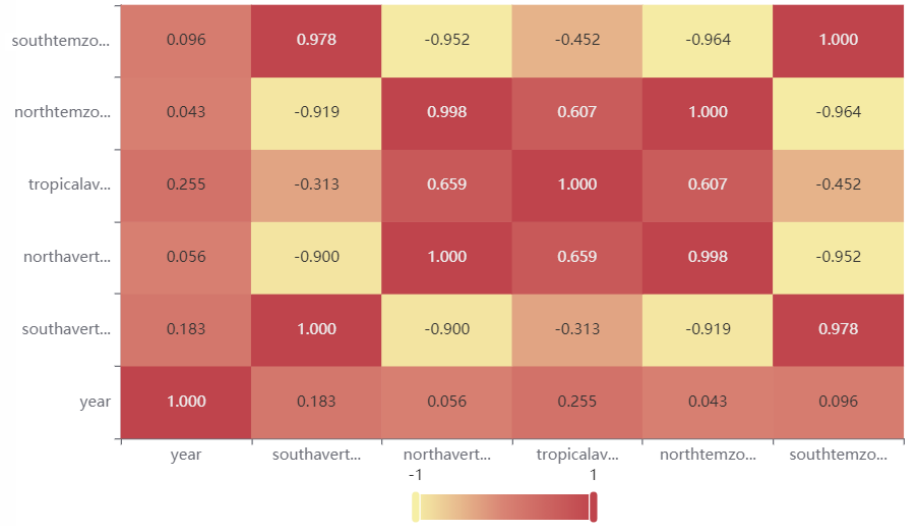

Figure 1. The heat map of the relationship between global temperature and time and location

The images clearly show that the annual average temperature at different locations is positively correlated with time, and the significant value $P<0.05$, so we conclude that global temperature is positively correlated with time and location, and the temperature in different latitudes increases with time.

3. MODEL ESTABLISHMENT

3.1. Selection of Models

We choose the autoregressive integrated moving average (ARIMA) forecasting model and Linear Regression forecasting model for the following reasons:

(1) Time series model is one of the commonly used methods to predict the future data. Time series refers to the data composed in the order of time occurrence under the same index and forecasts the future according to the existing data. Considering the stability of data, we choose the autoregressive integrated moving average model (ARIMA) to predict the future temperature trend.

(2) The application of univariate linear regression model is called the regression prediction of the relationship between a single independent variable and a dependent variable by the least square function. When we use a set of sample toilet paper and the least square method to get the regression equation of the sample, if the regression equation of the sample passes various tests, we can use the regression equation of the sample to predict.

3.2. Autoregressive Integrated Moving Average Model (ARIMA)

(1) Analysis of the model

Due to the low stability of the series, we cannot use ARMA directly, here we need to go through differential transformation, ARIMA can be split into three parts: AR (Auto Regression), I (Integration), MA (Moving Average), respectively: autoregressive, differential and moving average, where the meanings they indicate are: 1) Auto Regression: autoregressive 2) Integration: difference method, said single integer order, in order to successfully build the model, we want to meet the time series smooth this requirement, if the time series is not smooth, then the need to make the series

smooth by difference transformation, the need for n times difference transformation, we will say that this is n order single integer; 3) Moving Average: said average moving model [5, 6].

Autoregressive model (AR), moving average model (MA) and difference method (I) combined, we get the difference autoregressive moving average model ARIMA (p,d,q), where d is the number of orders that need to be differenced on the data, ARIMA is the ARMA model after differencing.

Time series generally consist of stationary trends, seasonal variations, and stochastic factors. It is difficult to model this part of the time series if the stochastic factors of the time series are completely independent without any linkage at each time point. Fortunately, for general time series, after excluding the fixed trend and seasonal effects, the time series is correlated at different time points, and this autocorrelation feature is the basis for our modeling of the time series.

In statistics, we use the correlation coefficient to express the correlation between two variables, and an important issue in time series is to study how the series affect each other. Similar to the statistical correlation coefficient, a similar approach is used in time series analysis to represent the autocorrelation characteristics of a time series. The autocorrelation function plot gives a clear view of the autocorrelation characteristics of the time series, and such characteristics are the basis for modeling.

Autocorrelation Function (ACF): The correlation coefficient measures the linear correlation of two vectors, and in a smooth time series $\{\gamma_t\}$, we sometimes want to know the linear correlation of γ_t, with its past value γ_{t-i}. This is when we extend the concept of correlation coefficient to autocorrelation coefficient.

The correlation coefficient between γ_t and γ_{t-i} is called the autocorrelation coefficient for γ_t, with interval l, usually denoted as ρ_l specifically.

$$\rho_l = \frac{\mathrm{cov}(\gamma_t, \gamma_{t-i})}{\sqrt{Var(\gamma_t)Var(\gamma_{t-i})}} = \frac{\mathrm{cov}(\gamma_t, \gamma_{t-i})}{Var(\gamma_t)} \tag{3}$$

The property of weakly smooth series is used:

$$Var(\gamma_t) = Var(\gamma_{t-1}) \tag{4}$$

Then the function: $\hat{\rho}_1, \hat{\rho}_2, \hat{\rho}_3 \ldots$(ACF), when all the values in the autocorrelation function are 0, we consider the sequence to be completely uncorrelated; therefore, we often need to test whether multiple autocorrelation coefficients are 0.

Describing the relationship between current and historical values, using the variable's own historical time data to make predictions about itself, the autoregressive model must satisfy the requirement of smoothness, and in general, the p-order autoregressive process is defined by the formula:

$$\mathrm{y}_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-1} + \varepsilon_t \tag{5}$$

y_t is the current value, μ is the constant term, p is the order, γ_i is the autocorrelation coefficient, and ε_t is the error white noise. From the formula, the current value is predicted by the historical value,

and p is an order in the autoregressive model, indicating that the historical value of several periods is used to predict the current value.

(2) Data pre-processing and smoothness test

First for the problem we performed a date ascending process, selecting data from 1900-2012 for training, finding the global annual average and calculating the global maximum temperature, with the global minimum temperature added to the data to visualize the five data in the tail. Construct timing and autocorrelation diagrams to visualize the original data timing as in Figure 2.

The unit root test (ADF) was used to test the smoothness of the time series and the results are (0.4693107079130492, 0.9839098592415724, 6, 106, {'1%': -3.4936021509366793, '5%': -2.8892174239808703, '10%': -2.58153320754717}, -47.536342444450696) as the p-value is greater than the significance level (0.05), the original hypothesis is accepted as set. The original series is tested as non-smooth.

(3) First-order difference series test

Since the series test result is a non-stationary series, we have to carry out a difference test to make the series stationary, we first carry out a difference and draw the time series after the difference to find the series into a stationary state, the result is (-6.696195437849027, 3.993986436993228e-09, 5, 106, {'1%': -3.4936021509366793, '5%': -2.8892174239808703, 10%': -2.5815332075471 7}, -52.18240273846345)

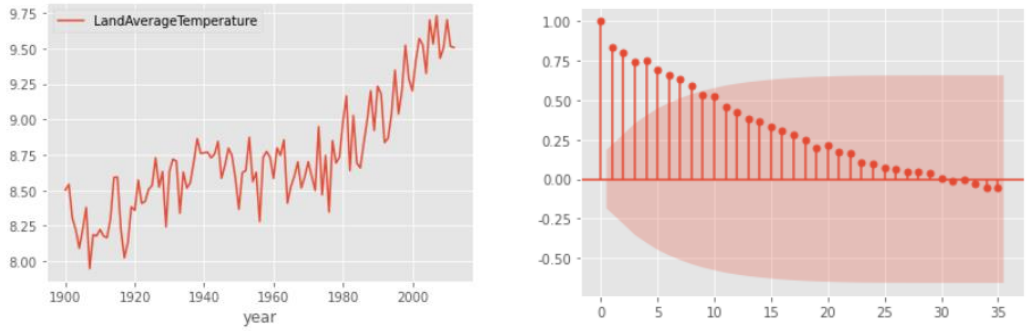

Figure 2. The time series after the difference

The result of the white noise test of the resulting first-order difference series is (array ([23.71717065]), array ([1.11582556e-06])), which returns the statistic P-value. If the P-value is less than 0.05, the original hypothesis that the series is a random series is rejected, indicating that the first-order difference series is non-white noise. That is, the first-order difference-transformed sequence is a smooth sequence.

(4) Fixing the order and tuning the parameters

We choose the model, according to the autocorrelation plot shown in Figure 3. after the first order is falling in the confidence interval, is ending with the first order while the partial autocorrelation plot can be seen to be oscillatory and off dimensional, so we choose ARIMA (p,d,q) = ARIMA (0,1,1) model to fit.

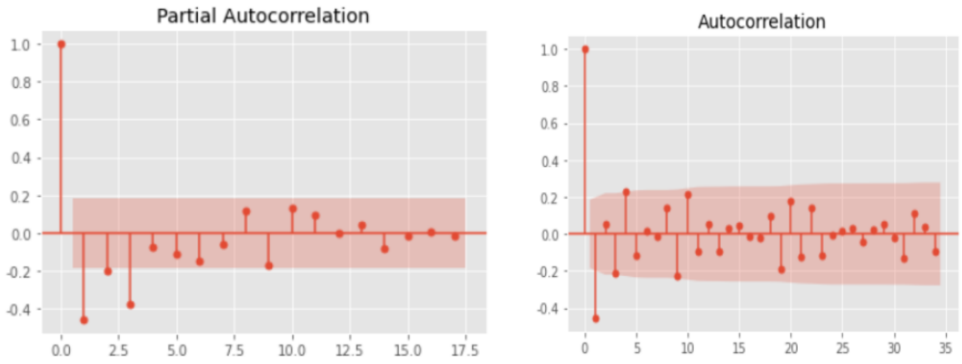

Figure 3. The autocorrelation plot and the partial autocorrelation plot

The following parameter tuning, here using BIC for parameter tuning; we use different combinations of p,q to derive different BIC stored in matrix form, using the stacking approach to get the optimal p value = 0 and q value = 1, so the result ARIMA (p,d,q) = ARIMA (0,1,1)

(5) Modeling and prediction

Next, the model is created, the report of the model is viewed, and then the residual test is performed, as shown in Figure 4. we can find that the correlation is very small.

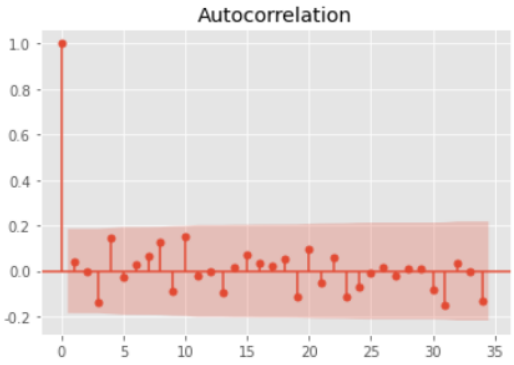

Figure 4. The residual test

A QQ plot is then made, and the data can be found to be normally distributed, as in Figure 5.

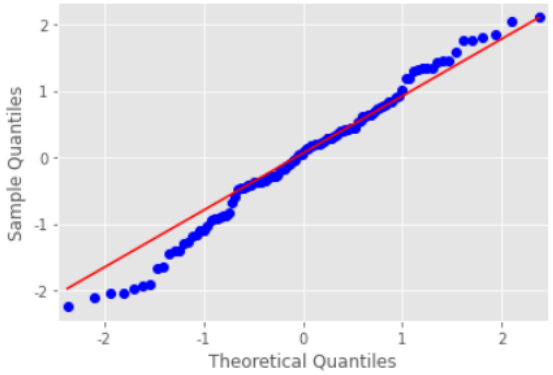

Figure 5. The QQ plot

The D-W test and the Ljung-Box test were used to test whether the residuals were white noise, whether the data were random.

The D-W test result is: 1.9185200526212165.

The result of the Ljung-Box test for white noise of the residual series is: (array ([0.17590671]), array ([0.67491481]))

The residuals pass the white noise test, so the ARIMA model holds, and let's predict the temperature for the next twenty years (13-32): where 13-22 are to test the accuracy of the model predictions, and 23-33 are to make future temperature predictions. The results are shown in Table 2.

Table 2. Comparison of power load forecasting of 403 line

Out [30]	
2013-12-31	9.582117
2014-12-31	9.593252
2015-12-31	9.604388
2016-12-31	9.615524
2017-12-31	9.626659
2018-12-31	9.637795
2019-12-31	9.648931
2020-12-31	9.660066
2021-12-31	9.671202
2022-12-31	9.682338
2023-12-31	9.693473
2024-12-31	9.704609
2025-12-31	9.715745
2026-12-31	9.726880
2027-12-31	9.738016
2028-12-31	9.749152
2029-12-31	9.760287
2030-12-31	9.771423
2031-12-31	9.782558
2032-12-31	9.793694
Freq: A-DEC, dtype: float64	

3.3. Unary Linear Regression Model

(1) Model analysis

The relationship between Y and X in the overall regression function can be either linear or nonlinear. There are two interpretations of "linearity" in a linear regression model [7, 8].

Explanation 1: linear in terms of variables, the conditional mean of Y is a linear function of X.

Explanation 2: linear in terms of the parameters, the conditional mean of Y is a linear function of the parameters.

Linear regression models are mainly "linear" with respect to the parameters, because as long as they are linear with respect to the parameters, their parameters can be estimated in a similar way.

Parameter estimation - least squares method

For a univariate linear regression model, suppose that n sets of observations $(x_1, y_1),(x_2, y_2),...(x_n, y_n)$ are obtained from the overall population. For these n points in the plane, an infinite number of curves can be used to fit. The sample regression function is required to fit this set of values as well as possible. Taken together, it makes the most sense for this line to be at the center of the sample data. The criteria for selecting the best-fitting curve can be determined as such that the total fitting error (i.e., the total residual) is minimized. We chose to use the principal criterion of least squares. In addition to being easier to calculate with the least squares method, the estimates obtained have excellent properties. This method is very sensitive to outliers.

The most commonly used arithmetic here is Ordinary Least Squares (OLS): the regression model chosen should minimize the sum of squares of residuals for all observations. (Q is the sum of squared residuals)

The sample regression model is:

$$Y_i = \beta_0 + \beta_1 X_i + e_i \rightarrow e_i = Y_i - \beta_0 + \beta_1 X_i \tag{6}$$

With the confidence interval of 1- , the prediction interval of the overall mean $E(Y \mid X_0)$.

$$\hat{Y}_0 - t_{\frac{\alpha}{2}} \times S_{\hat{Y}_0} < E(Y \mid X_0) < \hat{Y}_0 + t_{\frac{\alpha}{2}} \times S_{\hat{Y}_0} \tag{7}$$

(2) Calculate the parameters of the regression model according to the formula

The value of the regression parameter a: 0.012

The value of the regression parameter b: -14.515

(3) Modeling and prediction

The formula of the model is as follows: $y = b + a * year$. The prediction results of the univariate linear regression model are shown in Figure 6.

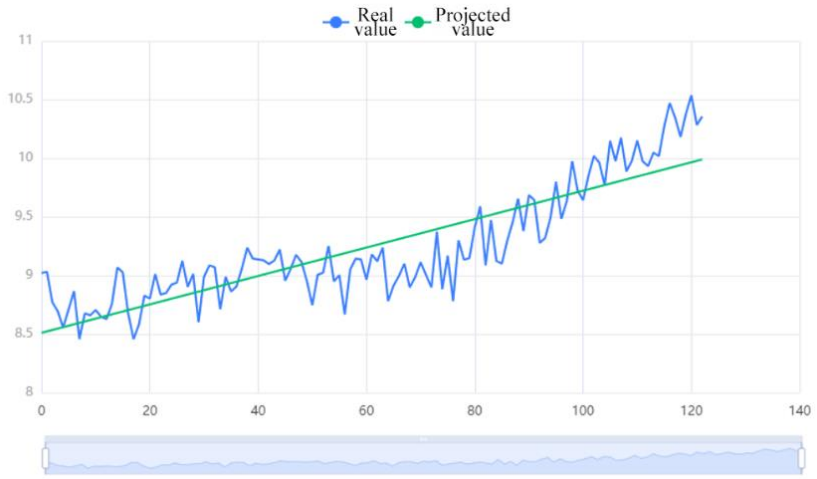

Figure 6. The prediction results of the univariate linear regression model

4. CONCLUSIONS

Throughout the study, it became evident that the average annual global temperature has been increasing year by year, rather than in a steady phase, with temperatures varying across different geographic locations over time. Numerous factors influence global temperature change, including geographic latitude and longitude as well as human activities. Pearson correlation analysis quantifies the impact of these factors on global temperature. Predictive models constructed using artificial intelligence algorithms, such as the ARIMA model and the Univariate Linear Regression model, forecast future global temperature changes with greater accuracy and stability. Achieving harmony between humanity and nature remains a long-term goal, and artificial intelligence algorithms offer new perspectives and tools to address the global greenhouse effect.

REFERENCES

[1] Loiy Al Ghussain. "Global warming: review on driving forces and mitigation." Environmental Progress & Sustainable Energy 38.1(2019). doi:10.1002/ep.13041.

[2] Qin Li, Zhou Xin. Possible impact of human activities on temperature change in Beijing area since the industrial revolution [J] 2010, 55(06):522-525.

[3] Xiao Yuanjun, Li Baoshan, Song Wendan, Cheng Yongxiang, Huang Jingfeng. Effects of PM (2.5) concentration on temperature in China from 1998 to 2016 [J]. Journal of Meteorology and Environment, 2022, 38(03):85-92.

[4] Wu Chunde, Sun Yankun. Heilongjiang Meteorology,2014,31(01):39-41. DOI:10.14021/j.cnki.hljqx.2014.01.019.

[5] Li-Ping Tian, et al. "Nonlinear-Model-Based Analysis Methods for Time-Course Gene Expression Data." The Scientific World Journal 2014. (2014). doi:10.1155/2014/313747.

[6] Liu Fang, GE Ruiting. Prediction and analysis of winter minimum temperature based on ARIMA [J]. Electronic Technology and Software Engineering, 2022(12):184-188.

[7] Zhou Suiyuan, XU Xin, XU Xiaoqian. Model analysis of carbon dioxide demand prediction based on univariate linear regression method [J]. Guangdong Chemical Industry, 2020, 47(23):64-66+75.

[8] Cheng Liu and Yixiao Sun. "A simple and trustworthy asymptotic t test in difference-in-differences regressions." Journal of Econometrics 210.2(2019). doi:10.1016/j.jeconom.2019.02.003.