Analysis of global temperature influencing factors based on spearman correlation coefficient method and grey correlation theory

Yufei Zheng[1], Yinghao Meng[2], Erzhen Lou[3], Yanjian Li[4], Likai Dong[1, *]

[1]School of Information Science and Engineering, University of Jinan, Jinan, China

[2]College of Mechanical and Electrical Engineering, Wenzhou University, Wenzhou, China

[3]College of Mathematics and Physics, Wenzhou University, Wenzhou, China

[4]School of Mathematical Sciences, University of Jinan, Jinan, China

*Corresponding author: ise_donglk@ujn.edu.cn

Abstract. In recent years, mankind has experienced the hottest period in its history as greenhouse gas concentrations have increased. As a hot topic for many years, the issue of global warming is gaining more and more attention.By exploring and analyzing the main factors affecting global temperature, we can take targeted measures to improve the global warming situation.In this paper, firstly,by building the ARMA model, we found that the global average temperature is increasing year by year.Secondly, we analyze that latitude and longitude are negatively correlated with global average temperature and time is positively correlated with global temperature by Spearman's correlation coefficient method, and year, CO2, population, forest area, land area, wetland area are all correlated with global temperature by gray correlation theory method, and the gray correlation of year, CO2 and land area is over 0.9.Finally, we give some effective suggestions to improve global warming,such as expanding forest area and reducing carbon emissions.

Keywords: Global Warming, Spearman, Grey Correlation Theory.

1. Introduction

In recent years, all regions of the world have been warming at different rates. The report Climate Change 2022: Mitigating Climate Change, published by the United Nations Intergovernmental Panel on Climate Change (IPCC), shows that global anthropogenic greenhouse gas emissions continue to grow from 2010-2019, with the highest annual average emissions in human history for this decade, proving that the world is on the "fast track" to disaster [1]. Undoubtedly, global warming has become a "now or never" problem from which no one can escape. As a social issue, global warming is expected to affect everything from natural resources to crop production to economic development around the world [2-4]. There are many factors that contribute to global warming, including greenhouse gas emissions, natural disasters, and population growth rates. We need to know the factors that have a significant impact on global warming so that we can take effective measures as soon as possible.

2. ARMA model and Spearman Correlation Coefficient Method and Grey Correlation Theory

2.1. The basic funamental of ARMA model

The autoregressive sliding average model is an important method for studying time series and consists of a "mixture" of an autoregressive model (AR model) and a moving average model (MA model). It is commonly used to study data with seasonal variations.

Considering the characteristics of the data and the excellent prediction effect of ARMA, we chose to use ARMA as the first prediction model.

Autoregressive moving average model ARMA(p,q), if x satisfies:

$$x_t - \phi_1 x_{t-1} \ldots - \phi_p x_{t-p} = \varepsilon_t - \theta_1 \varepsilon_{t-1} \ldots - \theta_q \varepsilon_{t\text{-}q} \tag{1}$$

Then the time series x is said to obey the (p, q) order autoregressive moving average model ARMA (p, q).

ARMA model requires the series to satisfy the smoothness, by viewing the ADF test results, based on the analysis of t-values, we judged that the hypothesis that the series is not smooth can be significantly rejected ($p<0.05$).

By looking at the comparison graph of the data before and after differencing, we judged that the data is smooth (not much up and down fluctuation), while the time series is biased (autocorrela- tion analysis), and its p and q values are estimated according to the truncated tails.

ARMA model requires the model to have pure randomness, that is, the model residuals are white noise, view the model test table, and test the model white noise according to the p-value of Q statistic ($p>0.05$), and derive the model formula combined with the time series analysis graph for comprehensive analysis to get the backward prediction order results.

2.2. Spearman's correlation coefficient

2.2.1 The basic funamental of Spearman's correlation coefficient

Spearman's rank correlation coefficient, or simply Spearman's correlation coefficient, is a nonparametric measure of rank correlation. The Spearman correlation can be considered as a nonparametric version of the Pearson correlation. While Pearson correlation is a statistical measure of the strength of the linear relationship between two random variables, Spearman correlation examines the strength of the monotonic relationship between the two, which in layman's terms is the extent to which they maintain the same pace in terms of their tendency to become larger or smaller, even if they do not maintain a proportional relationship. Pearson's correlation coefficient is calculated using the data sample values themselves, while Spearman's correlation coefficient is calculated using the data sample rank order values (sometimes the data itself is the rank order value, and sometimes the data itself is not the rank order value, then the rank order value is calculated before calculating Spearman's correlation coefficient)[5-6].

$$r_s = \frac{\sum (R_X - \overline{R_X})(R_Y - \overline{R_Y})}{\sqrt{\sum (R_X - \overline{R_X})^2 \sum (R_Y - \overline{R_Y})^2}} = \frac{\sum R_X R_Y - \frac{(\sum R_X)(\sum R_Y)}{n}}{\sqrt{(\sum R_X{}^2 - \frac{(\sum R_X)^2}{n})(\sum R_Y{}^2 - \frac{(\sum R_Y)^2}{n})}} \tag{2}$$

Where:

R(x) and R(y) are the digits of x and y, respectively

$\overline{R(x)}$ and $\overline{R(y)}$ denote the mean rank, respectively

If there are no rank knots (or fewer rank knots), a simpler and easier formula is shown below.

$$r_s = 1 - \frac{6\sum_{i=1}^{n} [R(X_i) - R(Y_i)]^2}{n(n^2-1)} = 1 - \frac{6\sum_{i=1}^{n} d_i^{\,2}}{n(n^2-1)} \tag{3}$$

Where:

d_i denotes the difference of the place value of the ith data pair

n is total number of observed samples

2.2.2 Scope of application

The Spearman's rank correlation coefficient can be used only if the observations of the two variables are paired rank-rated data or rank data obtained by transforming the observations of continuous variables, regardless of the overall distribution pattern of the two variables and the size of the sample size [7].

2.2.3 Establishment of spearman model

The usual method for calculating the spearman correlation coefficient is to sort the variables X and Y by rank from smallest to largest respectively, denoted by rank RX and RY. The phenomenon of equal data resulting in the same rank is called phase holding (tie), when the average rank is taken as the rank of each data. rs, the Spearman correlation coefficient, is calculated as follows.

Compared with the formula of Pearson correlation coefficient, the formula of Spearman correlation coefficient only replaces X and Y in the formula of Pearson correlation coefficient with RX and RY.

We did the correlation analysis according to the following steps.

1.Test whether there is a statistically significant relationship ($p<0.05$) between X, Y.

2.Analyze the positive and negative direction of the correlation coefficient as well as the degree of correlation.

3.Summarize the results of the analysis.

2.3. Grey Correlation Theory

2.3.1 The basic funamental of Grey Correlation Theory

Grey system theory is a kind of system science theory (Grey Theory) pioneered by the famous scholar Professor Deng Julong, in which the grey correlation analysis is Grey Relation Analysis (GRA), which is a kind of multi-factor statistical analysis method. It can quantitatively describe and compare the development and change dynamics of a system. The basic idea is to determine the degree of similarity of the geometric shapes of the reference data columns and a number of comparison data columns to determine whether the connection is strong or not, and it reflects the degree of correlation between curves [8-10].

2.3.2 Development of Grayscale Correlation Theory Model

Step 1: Determine the analysis sequence.

Determine the reference series reflecting the characteristics of the system behavior and the comparison series affecting the system behavior. The data series reflecting the characteristics of the system behavior is called the reference series. The data series consisting of factors that affect the system behavior is called the comparison series.

(1) Parent series (also called reference series, parent indicator)

The data series that reflects the behavior of the system, similar to the dependent variable Y, here x0

(2) Sub-series (also known as comparison series, sub-indicators)

A data series consisting of factors affecting the behavior of the system, similar to the dependent variable X, which is denoted as (x0, x1, x2,...,xm)

Step 2: Dimensionlessization of variables

Since the data of each factor column in the system may be of different sizes, it is not easy to compare them, or it is difficult to get the correct conclusion when comparing them. Therefore, when performing gray correlation analysis, we need to normalize them in a dimensionless way. This operation is generally known as normalization in the field of data processing, i.e., reducing the differences in the absolute values of the data, unifying them into an approximate range, and then focusing on the changes and trends.

The basic idea is that mapping a sequence of definitions from

$$\mathrm{x} = (x(1), x(2),...x(n)) \tag{4}$$

Into

$$f : x \to y \tag{5}$$

$$f(x(k)) = y(k), k = 1,2,..., n \tag{6}$$

There are two main methods of normalization.

(1) Initialization:

As the name implies, the data of this one sequence is uniformly divided by the very first value. Since the order of magnitude of the sequence of the same factor does not differ much, these values can be sorted out to around the order of 1 by dividing by the initial value.

$$f(x(k)) = \frac{x(k)}{x(1)} = y(k), x(1) \neq 0 \tag{7}$$

(2) Averaging:

As the name implies, it means dividing the data of this sequence by the mean value. Since the mean value of a sequence with a large order of magnitude is larger, it can be normalized to near the order of 1 after dividing.

$$f(x(k)) = \frac{x(k)}{\bar{x}} = y(k), \bar{x} = \frac{1}{n}\sum_{k=1}^{n} x(k) \tag{8}$$

Step 3: Calculate the correlation coefficient.

$\rho \in (0, \infty)$, is called the resolution factor. In general, the smaller ρ is, the larger the resolution is. Usually, $\rho = 0.5$ is taken.

$$\zeta_i(k) = \frac{\min_i \min_k \left|x_0(k) - x_i(k)\right| + \rho \cdot \max_i \max_k \left|x_0(k) - x_i(k)\right|}{\left|x_0(k) - x_i(k)\right| + \rho \cdot \max_i \max_k \left|x_0(k) - x_i(k)\right|} \tag{9}$$

Step 4: Calculate the correlation degree

After calculating the correlation coefficients of the X i(k) sequence and the Y(k) sequence, the average value of each type of correlation coefficients is calculated and the average value r iscalled the correlation degree of Y(k) with Xi(k).

Since the correlation coefficient is the value of the degree of correlation between the comparison series and the reference series at each moment (i.e., each point in the curve), it has more than one number, and the information is too scattered to facilitate a holistic comparison.

Therefore, it is necessary to concentrate the correlation coefficients of each moment (i.e., each point in the curve) into one value, i.e., to find its average value, as a quantitative representation of the degree ofcorrelation between the comparison series and the reference series, and the correlation degree formula is as follows:

$$r_{0i} = \frac{1}{m}\sum_{k=1}^{m} \zeta_i(k) \tag{10}$$

If each indicator plays a different role in the comprehensive evaluation, we can find the weighted average of the correlation coefficients.

$$r_{0i} = \frac{1}{m}\sum_{k=1}^{m} W_k \cdot \zeta_i(k), (k = 1, ..., m) \tag{11}$$

W_k is the weight of each indicator

Step 5: Relevancy ranking

The correlations are ordered by size, and if r 1 < r 2 , then the reference series y is more similar to the comparison series x2.

3. Results

3.1. Global average temperature forecast base on ARMA model

We did ARMA time series projections using global average temperature data from 1973-2021 as shown in Table.1 :

Table.1. ADF Checklist

ADF Checklist							
Variables	Difference order	t	P	AIC	Threshold Value		
					1%	5%	10%
tempera	0	-0.117	0.948	-28.756	-3.597	-2.933	-2.605
	1	-4.836	0.000***	-30.752	-3.597	-2.933	-2.605
	2	-5.901	0.000***	-20.276	-3.616	-2.941	-2.609

Note: ***, **, * represent 1%, 5%, 10% significance levels, respectively

The results of this series test show that based on the variable tempera: At the difference into 0th order, the significance P-value is 0.948, do not present significance at the level, the original hypothesis cannot be rejected and the series is an unsteady time series.

At the difference of order 1, the significance P-value is 0.000***, the level of significance is presented, the original hypothesis is rejected, and the series is a smooth time series.

At the difference of order 2, the significance P-value is 0.000***, which presents significance at the level, and the original hypothesis is rejected, and the series is a smooth time series.

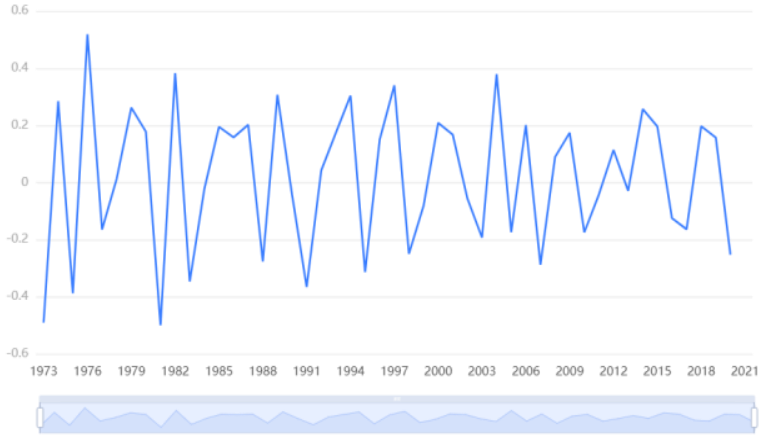


Figure 1. Timing diagram of the original data after 1st order differencing

Since the series is a smooth time series when the difference is of order 1. So Figure 1 shows the time series plot of the original data after the 1st order difference.

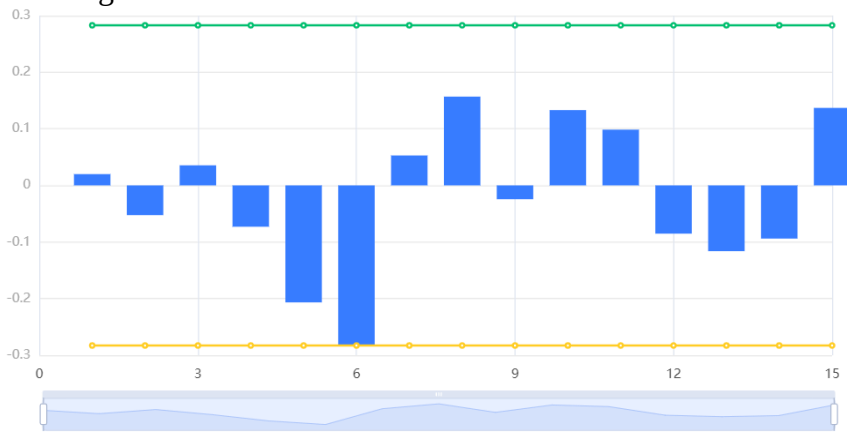


Figure 2. Model residual autocorrelation plot (ACF)

Figure 2 shows the autocorrelation plot (ACF) of the model residuals, including the coefficients, upper and lower confidence limits. The correlation coefficients are all within the dashed lines, the autoregressive model (AR) residuals are white noise series, and the time series requires the model residuals to be white noise series.



Figure 3. Model residual bias autocorrelation plot (PACF)

Figure 3 illustrates the residual partial autocorrelation plot (PACF) of the model, including the coefficients, upper confidence limits and lower confidence limits. Most of the correlation coefficients are within the dashed line, the sliding average model (MA) residuals are white noise series, and the time series requires the model residuals to be white noise series.

Table.2. ARIMA Model (1,1,2) Test Table

ARIMA Model (1,1,2) Test Table		
Item	Sign	Value
Number of samples	Df Residuals	44
	N	49
Q statistic	Q6(P value)	0.02(0.887)
	Q12(P value)	7.501(0.277)
	Q18(P value)	11.408(0.494)
	Q24(P value)	20.816(0.289)
	Q30(P value)	25.264(0.392)
Information Criteria	AIC	-19.487
	BIC	-10.131
Goodness of fit	R^2	0.834

Note: ***, **, * represent 1%, 5%, 10% significance levels, respectively

The system automatically finds the optimal parameters based on the AIC information criterion, and the model results are Table.2.ARIMA model (1,1,2) test table based on the variable: tempera, and from the analysis of the Q statistic results, we can get: Q6 does not present significance at the level, and the hypothesis that the residuals of the model are white noise series cannot be rejected, while the goodness-of-fit R^2 of the model is 0.834, and the model performs excellent, and the model basically meets the requirements.

Table.3. Model Parameters Table

Model Parameters Table						
	Coefficient	Standard Deviation	t	P>|t|	0.025	0.975
Constants	0.028	0.002	16.602	0	0.025	0.031
ar.L1.D.tempera	-0.98	0.067	-14.618	0	-1.112	-0.849
ma.L1.D.tempera	-0.055	0.137	-0.397	0.691	-0.324	0.215
ma.L2.D.tempera	-0.945	0.136	-6.94	0	-1.212	-0.678

Note: ***, **, * represent 1%, 5%, 10% significance levels, respectively

We find the optimal parameters and the model results in Table.3 ARIMA model (1,1,2) test table and based on 1-difference data with the following model equation.

$$y(t) = 0.028 - 0.98 * y(t-1) - 0.055 * \varepsilon(t-1) - 0.945 * \varepsilon(t-2) \quad (12)$$

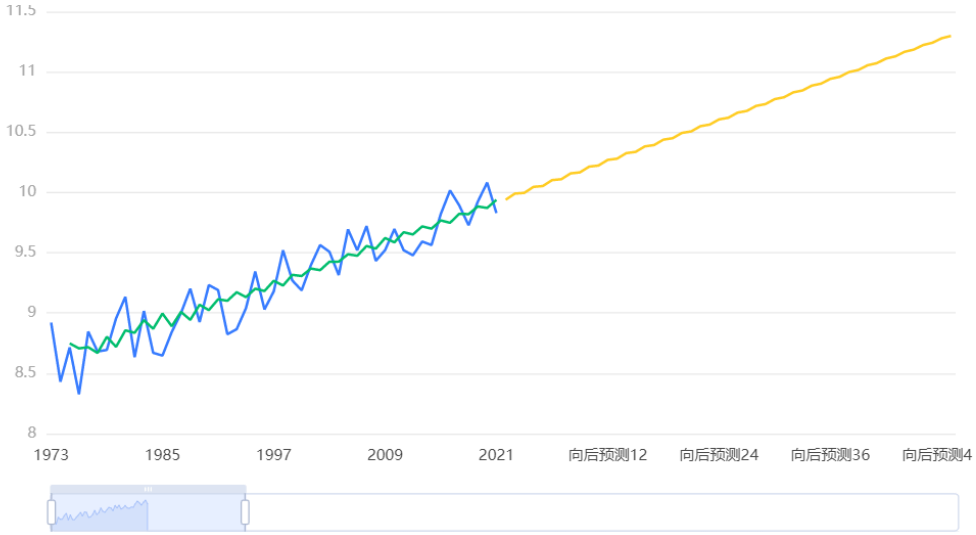


Figure 4. Time Series Chart

The three values in Figure 4 represent the true, fitted and predicted values once.

By building the ARMA model, we found that the global average temperature is increasing year by year, so we must pay attention to the problem of global warming and find the factors that affect the temperature change, and take timely measures to curb global warming.

3.2. Analysis of influence of geographic location and time on global temperature change base on Spearman's correlation coefficient

We made a correlation coefficient heat map as shown as Figure 5, which is shown to demonstrate the value of the correlation coefficient. This type of graph uses color shades to indicate the magnitude of the values, the darker the color, the larger the value. By comparing the four we found that there is a correlation between them.

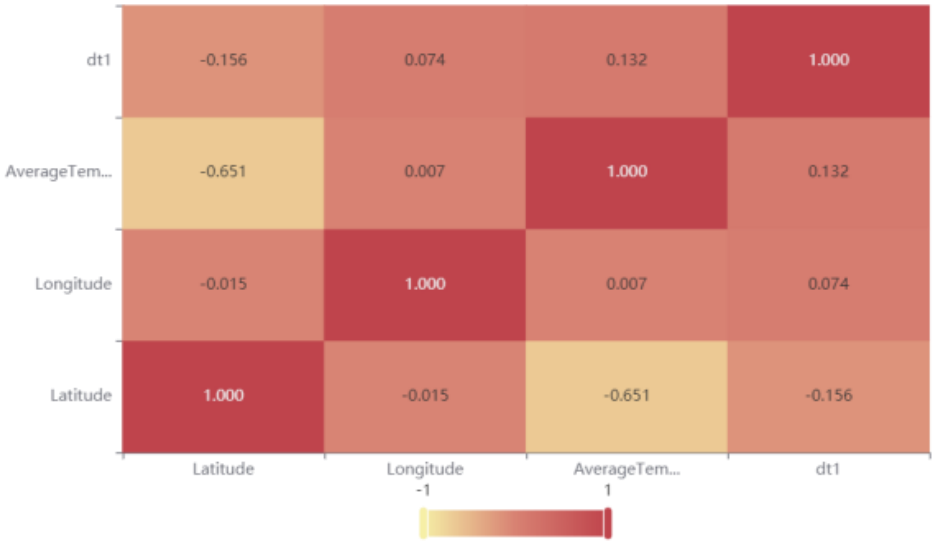


Figure 5. Heat Map of Correlation Coefficient

Where, the variable dt1 denotes the time.

We also made a spearman correlation coefficient table as shown in Table.4:

It can be found in the table that the p-values are all less than 0.05, i.e., they show significance, and there is a correlation between both variables. latitude is negatively correlated with global av- erage temperature, longitude is correlated with global average temperature, but the correlation is small, and time is positively correlated with global temperature.

Table.4. Results of spearman-based analysis of global temperature influencing factors

	Latitude	Longitude	AverageTemperature_1	dt1
Latitude	1(0.000***)	-0.015(0.000***)	-0.651(0.000***)	-0.156(0.000***)
Longitude	-0.015(0.000***)	1(0.000***)	0.007(0.001***)	0.074(0.000***)
AverageTemperature_1	-0.651(0.000***)	0.007(0.001***)	1(0.000***)	0.132(0.000***)
dt1	-0.156(0.000***)	0.074(0.000***)	0.132(0.000***)	1(0.000***)

Note: ***, **, *represent 1%, 5%, 10% significance levels, respectively

According to the Spearman correlation coefficient results, we can find that global temperature is related to time (including latitude and longitude) and location. Among them, the passage of time is positively correlated with global temperature, while longitude and latitude are inversely correlated with global temperature.

3.3. Analysis of influence of CO2、population、forest area etl. on global temperature change base on Grey Correlation Theory

We collected the average value of CO2, population, forest area, land area, wetland area and global temperature data for the period 1990-2020, and analyzed the correlation coefficients corresponding to the characteristic series and the rest as the parent series through gray correlation analysis to determine which are the main causes of global temperature change.
With the first four steps we get in Table.5:

Table.5. Correlation coefficient results

Correlation coefficient results						
	year	land	co2	wetland	population	forest
1	0.954	0.935	0.894	0.334	0.692	0.854
2	0.942	0.924	0.911	0.342	0.716	0.848
3	0.857	0.843	0.989	0.369	0.796	0.783
4	0.865	0.854	0.997	0.384	0.811	0.795
5	0.903	0.892	0.969	0.403	0.801	0.832
6	0.978	0.967	0.905	0.399	0.769	0.902
7	0.898	0.890	0.998	0.443	0.850	0.839
8	0.933	0.924	0.964	0.440	0.844	0.875
9	0.979	0.986	0.898	0.430	0.802	0.962
10	0.953	0.947	0.971	0.452	0.875	0.905
11	0.931	0.927	0.999	0.487	0.920	0.870
12	0.985	0.981	0.956	0.526	0.898	0.945
13	0.972	0.974	0.927	0.635	0.887	0.994
14	0.988	0.989	0.957	0.671	0.927	0.983
15	0.959	0.9565	0.980	0.869	0.993	0.936

As can be seen from the above table, a gray correlation analysis was conducted for 6 evaluation items (year, land, CO2, wetland, population, forest) and 31 data items, and the correlation (correlation degree) between 6 evaluation items (year, land, CO2, wetland, population, forest) and tem was studied using tem as the "reference value" (parent series).

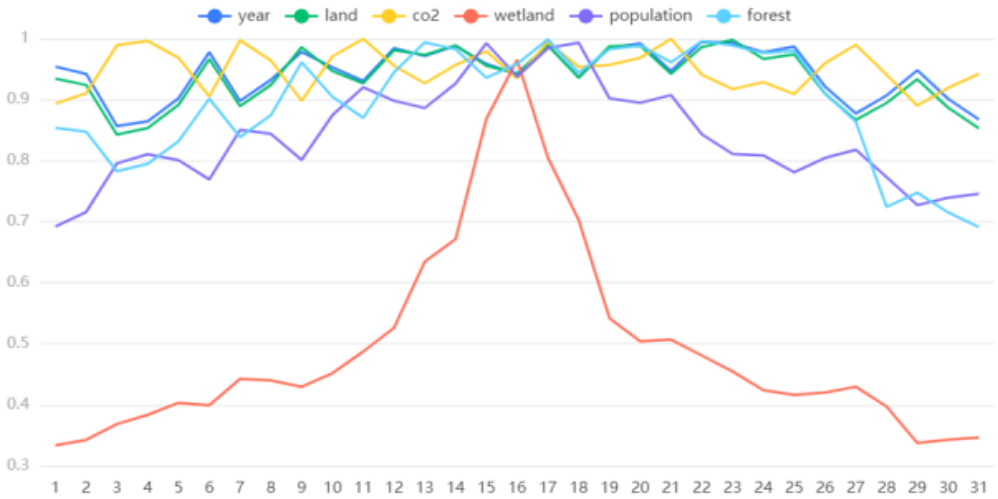


Figure 6. Correlation Coefficient Chart

In Figure 6,the correlation coefficient represents the value of the degree of correlation between this subseries year, land, co2, wetland, population, and forest pairs and the corresponding dimension of the parent series (the larger the number, the stronger the correlation).

Table.6. Results of Weighted ranking of correlation coefficients

Rating Items	Correlation	Rankings
CO2	0.951	1
year	0.944	2
land	0.937	3
forest	0.895	4
population	0.841	5
wetland	0.492	6

The correlation coefficients are weighted by the above correlation coefficients, and the correlation coefficients are used to rank the 6 evaluation items; the correlation coefficients are between 0 and 1, and the larger the coefficients are, the stronger the correlation between the coefficients and the "reference values" (parent series), which means that the coefficients are rated higher. From the Table 6, we can see that for the six evaluation items, CO2 has the highest evaluation (correlation: 0.951), followed by year (correlation: 0.944).

Therefore, we can judge that the main causes of global temperature change are CO2 emissions, the passage of time, and land and forest area and so on.

4. Conclusions

In this paper, we found that the global temperature is increasing year by year by ARMA model, and it is very important to find the factors that affect its change.

And then we investigated the correlation between geographical location and time on global temperature by the spearman correlation coefficient method. We found that latitude was negatively correlated with global average temperature, longitude was positively correlated with global average temperature, but the correlation was small, and time was positively correlated with global temperature.

In addition, we also studied the effects of carbon dioxide, population, forest area, land area, and wetland area on global temperature in the past 30 years using gray correlation theory, and we found that carbon dioxide emissions had the highest correlation with average temperature, and the correlation between year, land, forest, and population and average temperature were all over 0.5.

By studying the correlation of global temperature influencing factors, we advocate that the increase of global temperature can be slowed down by reducing CO2 emissions and increasing forest area.

References

[1] IPCC, 2022: Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [P.R. Shukla, J. Skea, R. Slade, A. Al Khourdajie, R. van Diemen, D. McCollum, M. Pathak, S. Some, P. Vyas, R. Fradera, M. Belkacemi, A. Hasija, G. Lisboa, S. Luz, J. Malley, (eds.)] [R]. Cambridge University Press, Cambridge, UK and New York, NY, USA. doi: 10.1017/9781009157926.

[2] Wu Xiaojian, Shi Yan, Zhu Jianing, Sun Linjun, Ma Xiangqing. The effects of global warming on forest litter [J/OL]. World forestry research: 1-7 [2023-01-10]. DOI: 10.13348 / j.carol carroll nki sjlyyj. 2022.0079. Y.

[3] [Lu J, Guo X F. Impacts of global warming on agriculture in eastern coastal areas of China [J]. Shanxi agricultural economy, 2022 (13): 1-4. DOI: 10.16675 / j.carol carroll nki cn14-1065 / f 2022.13.001.

[4] Li Wenting, Duan Hongbo. Global warming impact on China's economic research: regional and industrial perspective [J]. Journal of China university of geosciences (social science edition), 2022, 22 (4): 128-141. The DOI: 10.16493 / j.carol carroll nki. 42-1627 / c. 2022.04.006.

[5] Yu Qun, Huo Shoudong, He Jian, Li Lin, Zhang Jianxin, Feng Yuyao. Trend prediction of power grid outages in China based on Spearman correlation coefficient and system inertia [J/OL]. Chinese Journal of Electrical Engineering: 1-12 [2023-01-11].

[6] Chen Xiaoshang. Analysis of water quality trends in the Ulansu Sea by Spearman's rank correlation coefficient method [J]. Tannery and Environmental Protection Science and Technology,2022,3(07):128-130. DOI: 10.20025/j.cnki.CN10-1679.2022-07-43.

[7] Xie W. H. Variable screening method for Spearman correlation coefficient [D]. Beijing University of Technology,2015.

[8] Guo Kilian, Dong Yanfei, Zhang Hengxi. misconceptions of grey correlation analysis in variable screening applications [J]. Systems Engineering Theory and Practice, 2002(11): 126-128.

[9] Du Xianhan, Li Yan. Forecasting railroad passenger traffic based on gray correlation theory and multiple linear regression model [J]. Transportation Technology and Economy, 2013, 15(04): 90-93. DOI: 10.19348/j.cnki.issn1008-5696.2013.04.026.

[10] Wang Lili, Wang Keren. Fatigue tagging analysis of controllers based on gray correlation theory [J]. Journal of Safety and Environment, 2022, 22(02): 810-818. DOI: 10.13637/j.issn.1009-6094.2020.1560.