Global Warming Trends and Driving Factors Based on Grey Prediction and BP Neural Network Analysis

Team Number :	apmcm2209885
Problem Chosen :	C

2022 APMCM summary sheet

With the increasing number of extreme heat events around the world, the survival of human beings has been seriously threatened. This has prompted us to think about whether global warming is the cause of the extreme heat phenomenon, so we need to find the internal connection with the global temperature, natural disasters and other historical data sets, and we need to make corresponding measures to improve the natural ecological environment and build a natural environment suitable for human survival.

For question 1, we first test the global temperature data in March 2022 by using the **Mann-Kendall mutation** test method, and make a mutation test chart of global average temperature, tropical, southern temperate, and northern temperate mean temperature, which shows that none of the mutation points are near March 2022, which denies the idea that the global temperature increase in March 2022 led to a greater increase than the past decade. We then use the **time series model** to describe the past, and we can see that the future global temperature level is generally on an upward trend, and then use the model to forecast, but the test results reject the model, so we refer to the **gray prediction model** for forecasting, resulting in the time response equation, and the forecast results are shown in the Appendix, and then carry out quasi-exponential law test, which proves the reasonableness of the model. In order to make the prediction results more accurate, we then use the **BP neural network model** for prediction, in which the data use the global average temperature observation point data, that is, the global land average temperature, the year as the input quantity, the average temperature as the input quantity, in which the Bayesian Regularization algorithm is selected to train the model, and then the required prediction of the year with the sim function The model is trained with Bayesian Regularization algorithm, and then the year to be predicted is simulated with sim function to get the predicted average temperature of the corresponding year. Finally, we think that the BP neural network model is more accurate, and it predicts that the annual average temperature will reach 20.66 degrees Celsius in 2050 and 21.17 degrees Celsius in 2100.

For problem 2, we split the time data into months and years, and based on latitude and longitude, we divided the data into four groups, namely NE, NW, SE and SW, and then performed **Spearman correlation coefficient** calculation based on their regional average temperatures from 1899 to 2012, thus analyzing the relationship between global temperature, time and location. In order to make the analysis more accurate, we then used **multiple linear regression** to explore the relationship between the relevant factors, we used year, month, latitude and longitude as the independent variables, and the average temperature as the independent variable, and used the least squares method to establish a multiple linear regression model for the influencing factors, and the results obtained are shown in Tables 5 to 8. Finally, we concluded that global temperature is most influenced by latitude, and regions with different latitude Different heat conditions lead to a greater impact on temperature. Secondly, the month, different months, the angle of sunlight is different, resulting in different area and time of sunlight, the difference of sunlight, resulting in the fluctuation of temperature changes. For longitude, the effect on temperature is smaller, probably because different longitudes are located in different geographical areas, there are differences between land and ocean ground, and different warming conditions. Several steps of the year do not affect the global average temperature level. We then used visual analysis to select three typical cities, namely COVID-19 in the U.S., forest fires in Australia and volcanic eruptions in Indonesia, screened their data and accessed the time of natural disasters in the three countries online, and first used SPSS to statistically describe the data, and found that the annual average temperature in Indonesia was relatively stable, while the U.S. and Australia The average annual temperature peaks and valleys in the United States and Australia differed significantly. Then we plotted their line graphs with tableau and then **visualized the descriptions**, we found that when there were natural disasters, all three countries had large sudden change-type transitions in temperature, indicating that natural disasters have an impact on global temperature levels, with a sudden drop in temperature when COVID-19 erupted in the United States, a sudden rise in temperature when forest fires occurred in Australia, and a sudden drop in temperature when volcanic eruptions occurred in Indonesia. Then, we then use **principal component analysis** to study the factors affecting global temperature. By finding online information, we use global CO2, O3, API, PM2.5, SO2, NO2, CO content and global average temperature 2015-2021 data to conduct principal component analysis, and we conclude that CO2 concentration and O3 concentration content have the most influence on global temperature. Combining the two models we can see that the main factors affecting temperature are human activities, latitudinal position and abrupt changes in the environment (including the effects of natural disasters), and global warming mitigation initiatives are described below.

Finally, we have prepared a non-technical paper explaining our model, as well as the model's findings and recommendations for global temperature change.

Keywords: Mann-Kendall mutation; time series model; gray prediction model; BP neural network model; Spearman correlation coefficient; multiple linear regression; principal component analysis

Contents

1 Problem Restatement

1.1 Background of the problem

With the rise of the industrial revolution, more and more factories have been built, and with them, more and more greenhouse gases such as carbon dioxide have been emitted. The earth's atmospheric system is unable to absorb all these gases, and the concentration of greenhouse gases in the atmosphere continues to accumulate and increase, resulting in the consequence that the greenhouse effect continues to intensify, and new high temperature crises have emerged around the world one after another. This situation has seriously threatened the living environment of human beings, so we should use the accumulated global temperature data over the years to build models to analyze the truth of global warming behind these phenomena, and take urgent action to build a better natural ecological environment.

1.2 Problem Requirement

Based on the above background, answer the following questions with the help of the data given in the title and the data set collected by yourself.

(1) Based on the attached data and online search data, evaluate the impact of global temperature increase in March 2022 and build a mathematical model to describe the past data and predict the global average temperature in 2050 and 2100, solve for when the global average temperature will reach 20°C, and finally compare and evaluate the model built.

(2) Using the above results and the obtained data set, construct a mathematical model to analyze the correlation between global temperature, time and place, and then collect historical data related to natural disasters and evaluate whether natural disasters will have an impact on global temperature.

(3) Write a non-technical paper for the organizing committee, covering the data processing, the findings from the model building, and the future recommendations for the global warming situation.

2 Problem Analysis

2.1 Analysis of Problem 1

We use the Mann-Kendall mutation test method to test the global temperature data in March 2022, and make mutation test plots of global mean temperature, tropical, southern temperate, and northern temperate mean temperature, and then draw conclusions. We then use SPSS to make a time series plot to find the data pattern, and then use SPSS to build a time series model and then draw the prediction results, but due to the poor test, we do not use this method for predicting the data and only use it to describe the past global temperature level. Then we invoke the gray prediction model to make predictions, derive the time response equation, and make a global average temperature fit graph, from which we analyze the global average temperature level, and we then perform a quasi-exponential law test on it, and make a data smoothness analysis image, from which we justify the model. In order to make the prediction results more accurate, we then use the BP neural network model for prediction, I which data using the global average temperature observation point data, that is, the global land average temperature, the year as the input, the average temperature as the input, where the Bayesian Regularization algorithm is selected to train the model, and then the required prediction of the year with the sim function for The simulation is performed to obtain the predicted average temperature of the corresponding year.

2.2 Analysis of Problem 2

We split the time data into months and years, and based on the latitude and longitude, we divided the data into four groups, namely NE, NW, SE and SW, and then performed the Spearman correlation coefficient calculation based on their regional average temperatures from 1899 to 2012, thus analyzing the relationship between global temperature, time and location. To make the analysis more accurate, we then used multiple linear regression to explore the relationship between the relevant factors. We used the year, month, latitude and longitude as independent variables and the average temperature as the independent variable, and used least squares to build a multiple linear regression model for the influencing factors, and drew the final conclusion based on the constructed coefficients combined with the Spearman's correlation coefficient. We then used visual analysis to select three typical cities, namely COVID-19 in the United States, forest fires in Australia, and volcanic eruptions in Indonesia, screened their data, and accessed the time of natural disasters in the three countries on the Internet, and first used SPSS to describe the data statistically, and then used tableau to plot their line graphs, thus making the data more intuitive, combined with We analyzed the causes of natural disasters and global temperature changes by combining the times of natural disasters in the three countries reviewed. Then, we then use principal component analysis to study the factors that affect global temperature. Through online information finding, we use the global data of CO2, O3, API, PM2.5, SO2, NO2, CO content and global average temperature from 2015-2021 to conduct principal component analysis and come up with the factors that have a greater impact on global temperature. Finally, we combine the two models of the question to conclude the factors that have the greatest impact on global temperature and propose initiatives to mitigate global warming.

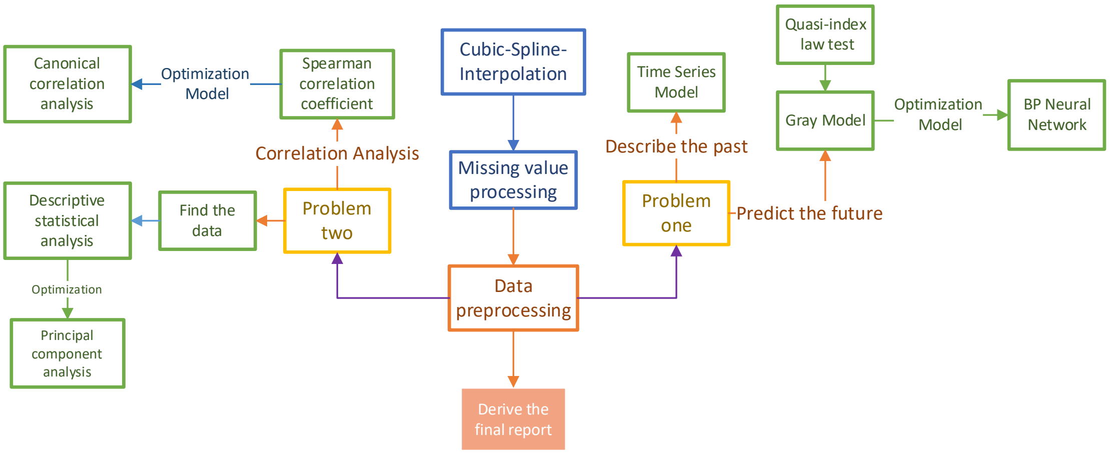


Figure 1 Flowchart

3 Model assumptions

1. It is assumed that unknown future factors have a small and negligible impact on global temperature change.

2. It is assumed that the factors we selected play a decisive role in the global temperature change, and other factors have negligible impact on the model.

3. Assume that the average temperature data of global observation points are real and valid.

4 Symbol Description

Symbol	Description
e_i	**Standardized variables**
α_x	**Standard deviation of data**
α	**Significance level**
$\hat{a}$	**Development of gray number**
$\hat{u}$	**Endogenous control of the number of ash**
$x^{(0)}$	**Statistical global average annual temperature data**
$x^{(1)}$	**The grey prediction model is then based on the cumulative global average annual temperature**

5 Pre-processing of data

5.1 Missing value processing

In this data analysis process, we found that there are missing data in the mean temperature and mean temperature uncertainty columns, and the missing values only exist as null values, in order to make the data more accurate, we use the method of three spline interpolation to process the missing data, we calculate each city separately, and then three spline interpolation according to the different months of each city each year, from which we can derive the final missing values, we establish the following construction function.

$$F(x)=\begin{cases} F_0(x)\ ,\ x\ \epsilon\ [x_0,x_1], \\ F_1(x)\ ,\ x\ \epsilon\ [x_1,x_2], \\ \vdots \\ F_{n-1}(x)\ ,\ x\ \epsilon\ [x_{n-1},x_n]; \end{cases} \quad F_i(x)\epsilon C^3([x_i,x_{i+1}])$$

The following conditions are also satisfied.

$$\begin{cases} F(x_i)=f_i \\ F_{i-1}=F_i(x_i) \\ F'_{i-1}(x_i)=F'_i(x_i) \\ F''_{i-1}(x_i)=F''_i(x_i) \end{cases}$$

With this model, we can make the data more accurate and complete, and the final derived data is shown in Annex 1. We then select one of the three spline interpolation graphs for visual display, see the following figure.

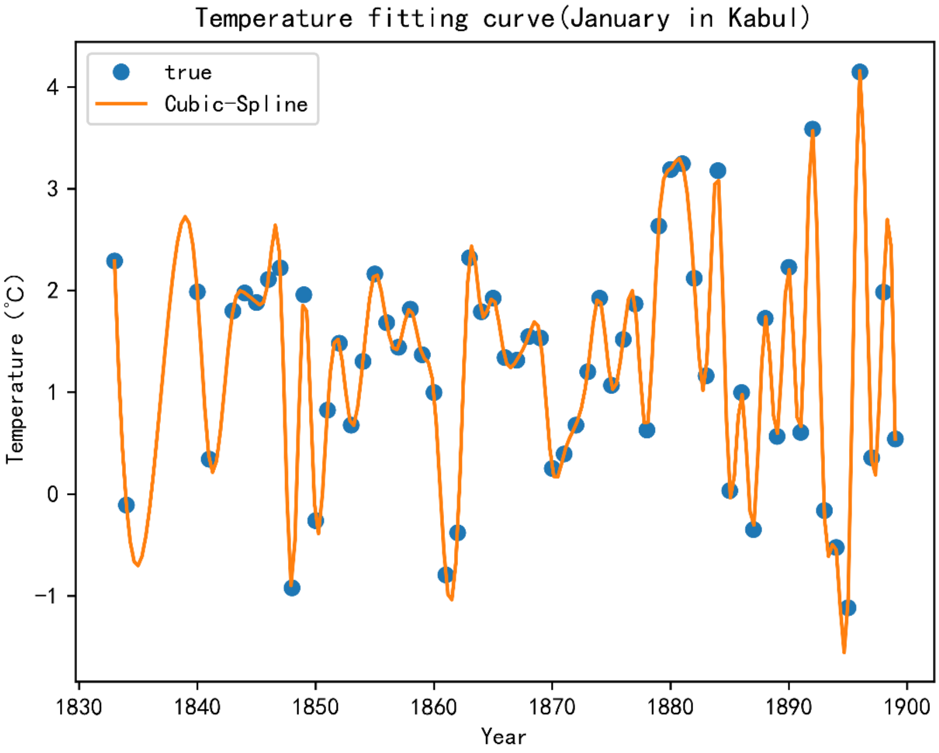


Figure 2 Three-time spline interpolation graph

5.2 Data normalization

In order to eliminate the influence of the scale on the data for the calculation, we standardize the data on the basis of the previous deletion of invalid data, so that the data can be more accurate in the calculation. Each sample data in the table is processed according to the formula: , where denotes the standardized variable, denotes the mean of the group, and denotes the standard deviation of the data. The qualitative variables such as city and country are converted into quantitative variables by dividing the frequency of the total data by the data in that row. The resulting standardized partial data table is shown in the figure below, and the complete standardized data table is shown in Annex 2.

5.3Calculation of average temperature by region

The table is supplemented by searching for temperature data from2012 to2020 available online. At the same time, each region is classified into the northern hemisphere and the southern hemisphere; Tropical, southern temperate, northern temperate and four special typical countries: Australia, Indonesia, the United States and China. and calculated the average monthly temperature of 1899 to 2020 for each region. The data sheet is provided in Annex 3

6 Problem 1 Modeling and Solving

6.1 Mann-Kendall mutation analysis of temperature increase

Disagree that the global temperature increase in March 2022 leads to a greater increase than in the past decade.

First, the analysis concludes that if the global temperature increase in March 2022 leads to a greater increase than during the past decade, that is, the global temperature produces a larger increase in March 2022, our group uses the Mann-Kendall mutation test method to test the global temperature data in March 2022 to determine whether March 2022 is the global temperature mutation point in the past decade.

In this subproblem we use the Mann-Kendall mutation test, which is a nonparametric statistical test that has band you in that it is not only convenient to calculate, but also allows to specify the moment point at which the mutation starts and indicates the mutation time period region[1].

For a time series X with n each sample size, construct an order column.

$$S_k = \Sigma_{i_1=1}^{k} r_i$$

We construct the order column with global average temperature data from January 2012 to October 2022.

$$r_i = \begin{cases} +1, \text{ if } x_i > x_j \\ 0, \text{ else} \end{cases} (j = 1, 2, \ldots, i)$$

It can be seen that the order series S_k is the cumulative value of the ith moment value greater than the number of moment values.

Under the assumption of stochastic independence of the time series, the statistic is defined as

$$UF_k = \frac{[s_k - E(s_k)]}{\sqrt{\operatorname{Var}(s_k)}} (k = 1,2, \ldots, n)$$

where $UF_1 = 0$, $E(s_k)$and $\operatorname{Var}(s_k)$are the mean and variance of the cumulative s_k. When $x_1, x_2, \ldots, x_n$ are independent of each other and have the same continuous distribution column, they can be calculated by the following equation.

$$E(s_k) = \frac{n(n-1)}{4} \operatorname{Var}(s_k) = \frac{n(n-1)(2n+5)}{72}$$

UF_i is the standard normal-terminus distribution, which is a sequence of statistics calculated in the order of time series $x_1, x_2, \ldots, x_n$. Given the significance level α, checking the normal distribution table, if$|UF_i| > U_\alpha$, it indicates that there is a significant trend change in the series.

Generate the time series x into its corresponding inverse series $x_n, x_{n-1}, \ldots, x_1$, and repeat the above calculation process, while making

$$UB_k = -UF_k, k = n, n-1, \ldots, 1, UB_1 = 0.$$

first we compute the order column of the sequential time series, calculating UF_k according to the above formula.

then calculate the order column of the inverse-order time series, calculating UB_k according to the above formula.

And given the significance level, α=0.01, for the critical value of $U_{0.05} = \pm 1.96$, the two statistical series curves of UF_k and UB_k are plotted on a plane right angle coordinate system with two straight lines of $U_{0.05} = \pm 1.96$.

The plotted UF_k and UB_k curves were analyzed to find their intersection points within the critical line, and their corresponding moments were the burst start times. The test plots of the global mean temperature, tropical, southern temperate, and northern temperate mean temperature bursts are based on the following.

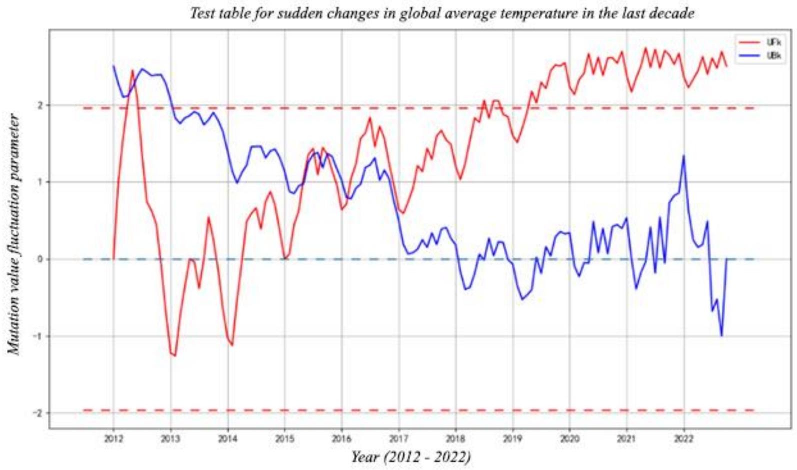


Figure4 Test table for sudden changes in global average temperature in the last decade

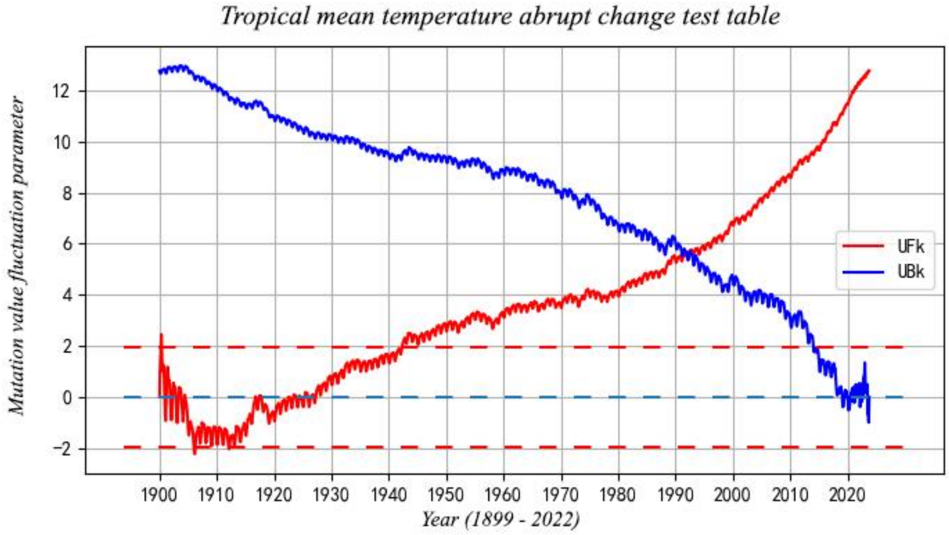


Figure5 Tropical mean temperature abrupt change test table

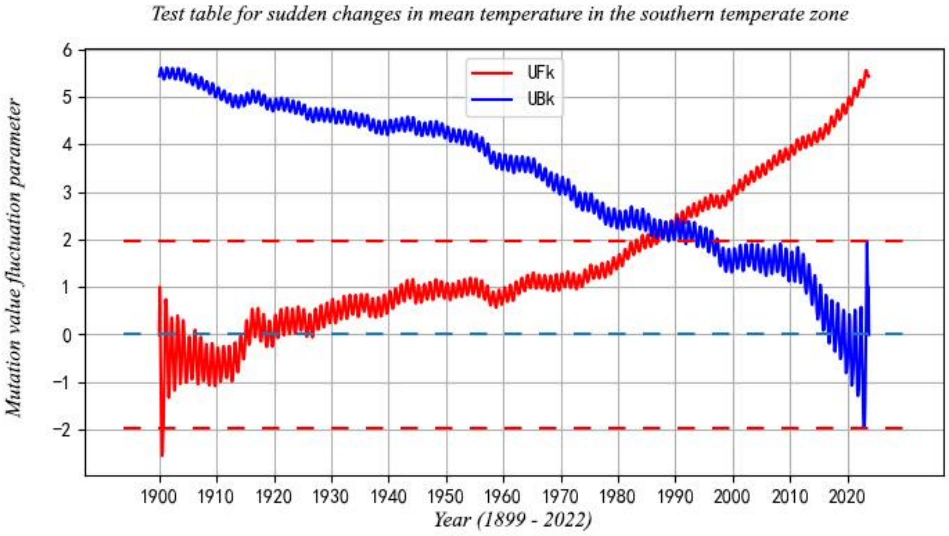


Figure6 Test table for sudden changes in mean temperature in the southern temperate zone

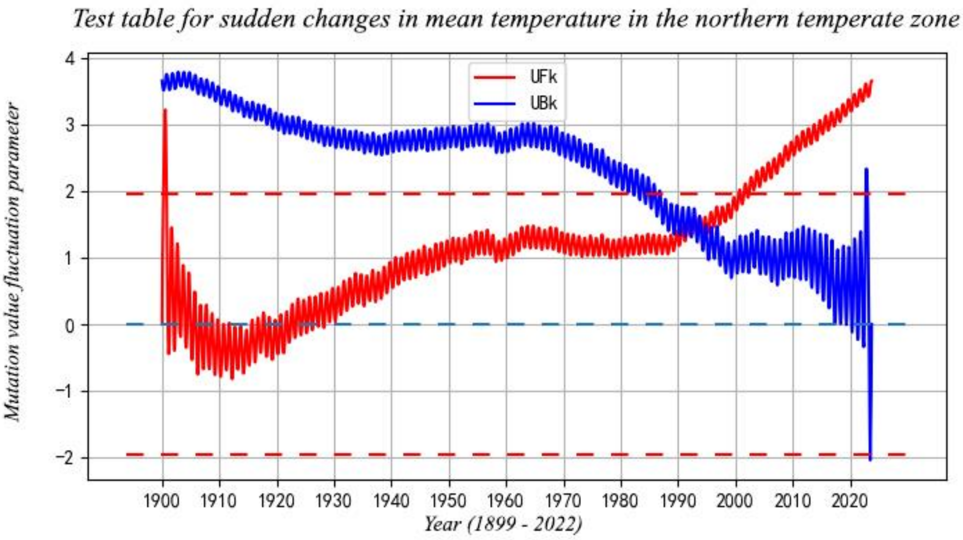


Figure7 Test table for sudden changes in mean temperature in the northern temperate zone

Ultimately, it was concluded that none of the mutation points were near March 2022, so this group disagreed that the global temperature increase in March 2022 led to a greater increase than in the past decade.

6.2 Predictive Models Analyze Future Global Temperature Levels

6.2.1 Time series model to describe the past and predict the future

Due to the requirements of the title, the past global temperature level is described, so we make statistics according to the data after data preprocessing and processing the data incomplete value, and then combine the global temperature level and the global observation point temperature level in recent years to obtain the global average temperature data and the average temperature data of the observation point from 1882 to 2015 The specific data are shown in Annexes 4 and 5, which are analyzed by a time series model. We first perform a time series graph analysis, and the established graph is shown in the figure below：

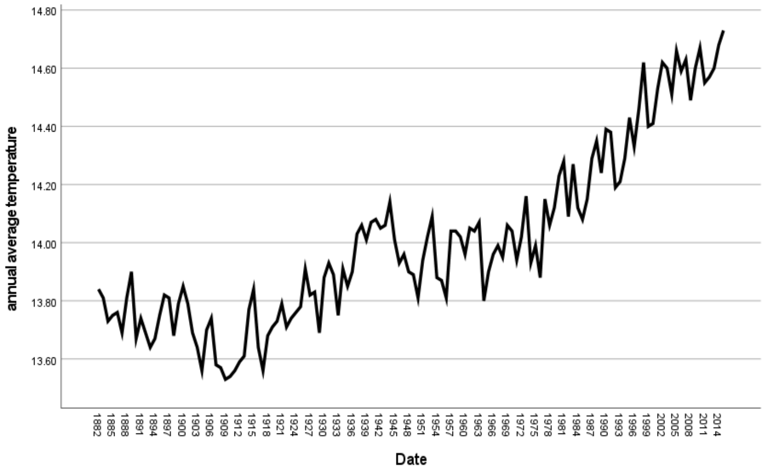
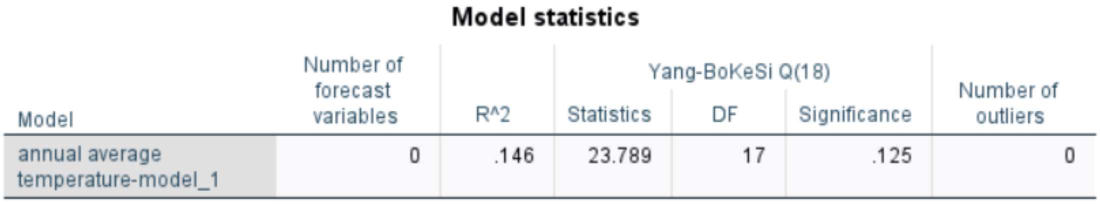


Figure 8 Time series chart

We can find that the data are roughly smooth series according to the figure, and the data fluctuate up and down around the mean value without obvious trend and seasonality. We can find that the global temperature level in the past is generally on an upward trend with fluctuations and ups and downs in the middle, indicating a general increase in global temperature and a significant global warming phenomenon. Among them, the global temperature change from 1882 to 1909 was not obvious and showed a slight downward trend, indicating that the level of industrialization before the 20th century was not high, carbon emissions were low and climate change was not significant. And the rising trend was the largest from 1975 to 2015, which showed that the high development of various countries and the increase of carbon emission led to the warming of the climate and the significant climate warming phenomenon. We then used SPSS to build an expert modeler to derive the best time series model, and we performed a time series model analysis based on the global average temperature in "years", and came up with the final prediction results. Then we can get its statistical table, as shown in Table 1.

Table 1 Model statistics table



Model statistics

Model	Number of forecast variables	R^2	Yang-BoKeSi Q(18)			Number of outliers
			Statistics	DF	Significance	
annual average temperature-model_1	0	.146	23.789	17	.125	0

We found a significance of 0.125, indicating that this prediction conclusion accepts the original hypothesis and there is a large error for the prediction effect, and then our resulting sample ACF and PACF plots are shown in the following figure.

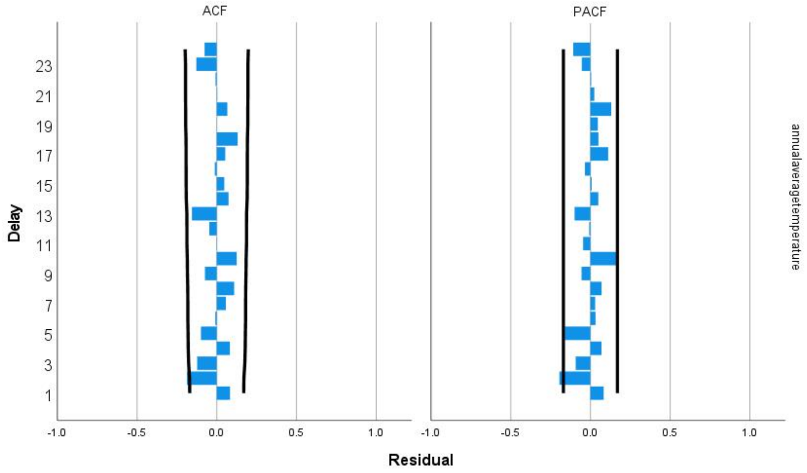


Figure 9 ACF and PACF diagram

Where ACF is a complete autocorrelation function that provides us with the autocorrelation value of any series with lagged values. It describes the degree of correlation between the current value of that series and its past values. ACF takes all components into account when finding correlations and it describes the autocorrelation between one observation and another, including both direct and indirect correlation information. pacf is a partial autocorrelation function or partial autocorrelation function. Basically, instead of finding lagged and current correlations like the ACF, it finds the residuals. Thus, if there is any hidden information in the residuals that can be modeled by the next lag, we may get good correlation and we will use the next lag as a feature when modeling[2]. From the resulting values we know that it has a good correlation and only some areas are outside the accepted range, indicating that the results are more accurate, but there is still some error. The results of the fit as well as the predictions we have obtained are shown in the following figure.

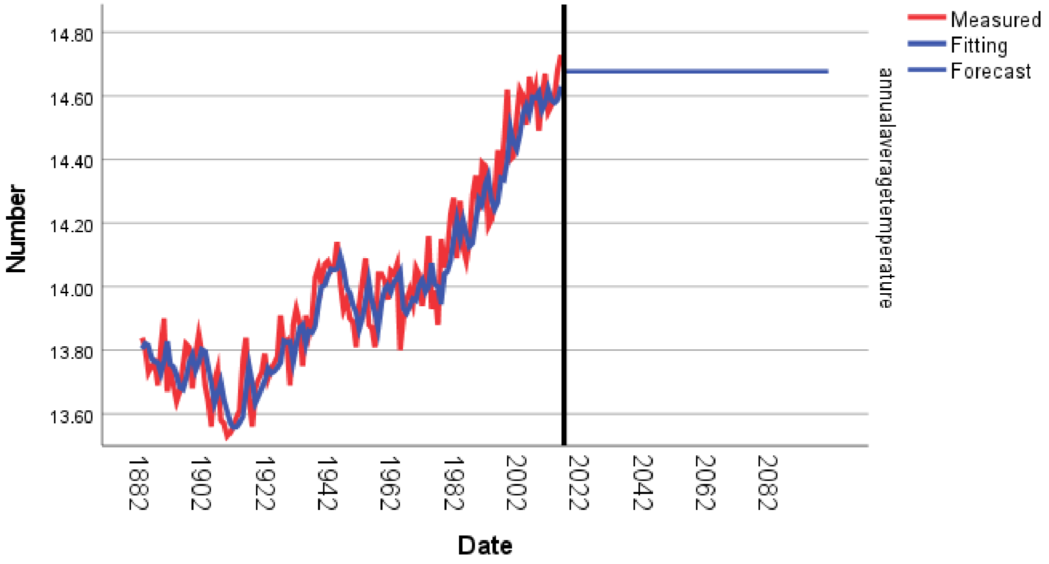


Figure 10 Fitted graph of time series analysis

From the obtained images we can see that the fitting results are more approximate, indicating that the fit is obviously better, but the prediction results are not satisfactory, because the time series model is not suitable for long-term data prediction, and the operation of the prediction results are more complex, so we only use the time series analysis model to describe the past global temperature level, after which the global temperature level prediction method is described below.

6.2.2 Optimization of prediction results by gray prediction model and testing

Since there is a large error in the prediction results given by the time series model, we then use the gray prediction model to optimize the prediction results. Firstly, we build a data table according to the national temperature table given in the topic and the global average temperature data collected by ourselves in recent years, and the specific data are shown in Table 2.

Table 2 Global average temperature table

Year	**1882**	**1883**	**1884**	**…**	**2025**
Serial number	1	2	3	…	134
$x^{(0)}$	13.84	13.81	13.73	…	14.73
$x^{(1)}$	13.84	27.65	41.38	…	1876.59

$$x^{(1)}(i)=\sum_{j=1}^{i}x^{(0)}(j)\ ,\ i=1,2,\cdots,n$$

With this data we can build the matrix as follows.

$$Y=\begin{bmatrix} x^{(0)}(2)\\ x^{(0)}(3)\\ \cdots\\ x^{(0)}(139)\end{bmatrix},\ B=\begin{bmatrix} -\frac{1}{2}[x^{(1)}(2)+x^{(1)}(1)] & 1\\ -\frac{1}{2}[x^{(1)}(3)+x^{(1)}(2)] & 1\\ \cdots & \\ -\frac{1}{2}[x^{(1)}(134)+x^{(1)}(133)] & 1\end{bmatrix}$$

We can use the least squares method to obtain estimates of the parameters a , b as

$$\hat{U}=\begin{pmatrix}\hat{a}\\ \hat{u}\end{pmatrix}=(B^TB)^{-1}B^TY$$

Where, $\hat{a}$ denotes $\hat{u}$ the developmental gray number and denotes the endogenous control gray number. From the Matlab code we can solve for $\hat{a}$=0.00049885 and $\hat{u}$=13.5394.

From this, we can substitute the values and introduce the time response equation as

$$x^{(1)}(k+1)=\left[x^{(1)}(1)-\frac{\hat{u}}{\hat{a}}\right]e^{-\hat{a}k}+\frac{\hat{u}}{\hat{a}}\ ,\ k=1,2,\cdots,n-1$$

The fitted image of the future global average temperature is solved as shown below.

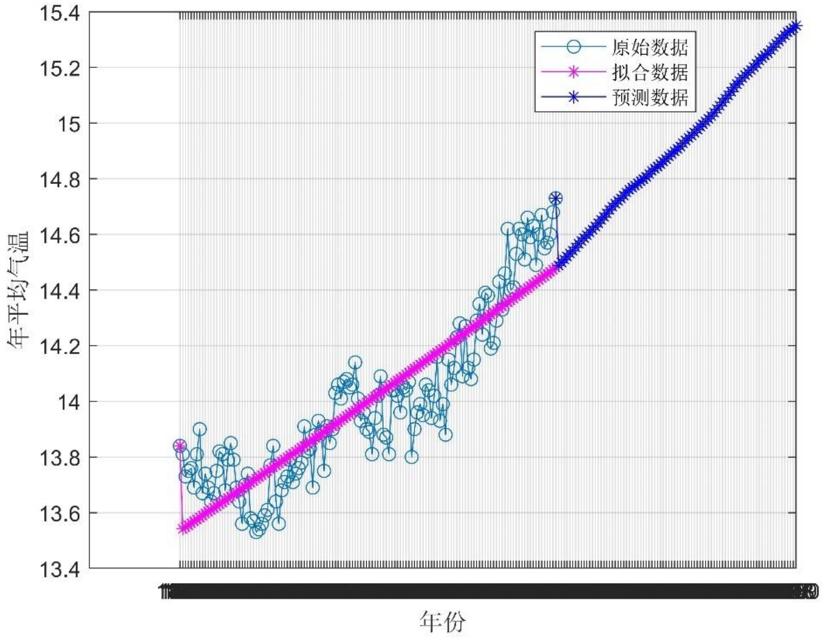


Figure 11 Grey prediction fitting diagram

The obtained predicted values of future global observation point average temperature are shown in Annex 7. In order to check the accuracy of the model predictions, we performed a quasi-exponential law test on the model, and the test procedure was as follows.

(1) Data with quasi-exponential laws is the theoretical basis for modeling using gray systems.

(2) The sequence of cumulants r times is: $x^{(r)}=(x^{(r)}(1),x^{(r)}(2),\cdots,x^{(r)}(n))$,Define the rank ratio

$\sigma(k)=\dfrac{x^{(r)}(k)}{x^{(r)}(k-1)}$, $k=2,3,\cdots,n$.

(3) If $\forall k,\sigma(k)\in[a,b]$,and the interval length$\delta=b-a<0.5$,then the sequence after accumulating r times is said to have a quasi-exponential law.

(4) Specifically in the GM(1,1) model, we only need to determine whether the sequence after accumulating once$x^{(1)}=(x^{(1)}(1),x^{(1)}(2),\cdots,x^{(1)}(n))$ has a quasi-exponential law.

(5) According to the above formula: the level ratio of the sequence is:

$\sigma(k)=\dfrac{x^{(1)}(k)}{x^{(1)}(k-1)}=\dfrac{x^{(0)}(k)+x^{(1)}(k-1)}{x^{(1)}(k-1)}=\dfrac{x^{(0)}(k)}{x^{(1)}(k-1)}+1$, defined

$\rho(k)=\dfrac{x^{(0)}(k)}{x^{(1)}(k-1)}$ as the smooth ratio of the original sequence $x^{(0)}$, noting that

$\rho(k)=\dfrac{x^{(0)}(k)}{x^{(0)}(1)+x^{(0)}(2)+\cdots+x^{(0)}(k-1)}$, assuming that the non-negative sequence (almost all common time series in life meet the non-negativity), then with the increase, the final $\rho(k)$ will gradually approach 0, so to make $x^{(1)}$ with quasi-exponential law, that is $\forall k$, the length of the interval $\delta<0.5$, only need to ensure that can be $\rho(k)\in(0,0.5)$, at this time the level ratio of $x^{(1)}$ the sequence $\sigma(k)\in(1,1.5)$.

We can obtain two indicators of the quasi-exponential law test by Matlab operation of the above table data as follows.

Indicator 1: the proportion of data with smooth ratio less than 0.5 is 50%.

Indicator 2: excluding the first two periods, the proportion of data with smooth ratio less than 0.5 is 100%.

The smoothness of the original data is shown in Fig. 12. After comparison, the above data passed the quasi-exponential law test, so the problem can be established as a gray prediction model.

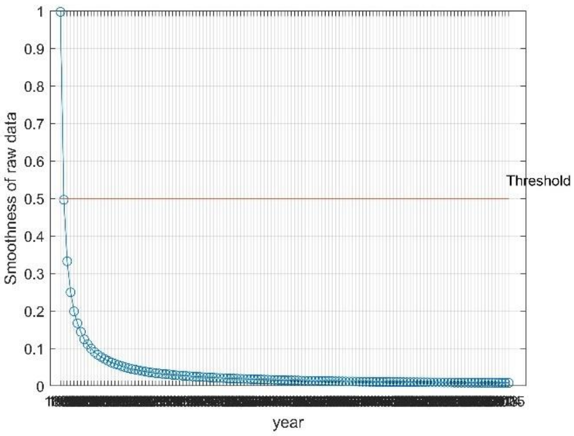


Figure 12 Indicator smoothness test chart

6.2.3 Optimization of the model: BP neural network prediction

Since the accuracy of the results obtained by the above method is not high, we use BP neural network to optimize the model. We use the BP neural network model to achieve more accurate prediction results by training the data set. We use the original data to train the neural network model and run it on the dataset we need to predict to get the desired prediction data[3].

We take the year column of data3（观测点数据） data as the input and the global average annual temperature as the output (shown in Figure 13 below) and take the number of neural hidden layers as 10 to train the neural network.

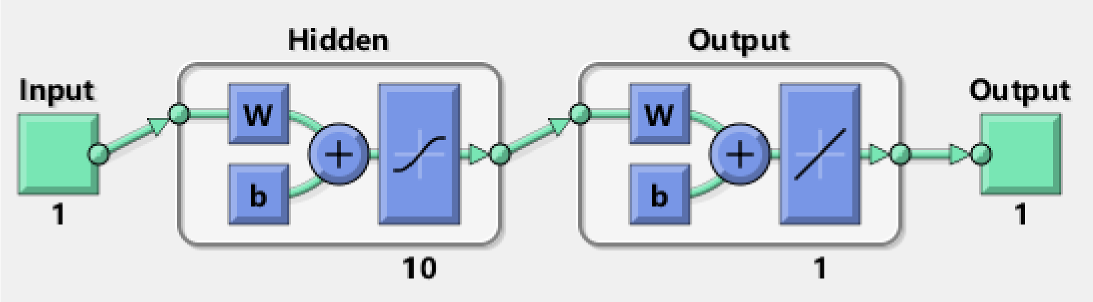


Figure 13 Neuronal network diagram

We let X be the data of 136 years from 1880 to 2015, Y be the data of global annual average temperature from 1880 to 2015, and new_x be the data of global annual average temperature from 2016 to 2100. We used Matlab's neural network fitting toolbox to train the data first, Matlab randomly selected samples according to the ratio we gave 70% of the training set, 15% of the validation set and 15% of the test set, we chose Bayesian Regularization method for training because this algorithm has the ability to modify the parameters when executing to reach the combination of Bayesian Regularization algorithm and its advantage of avoiding overfitting.

We derived the neural network training model after training, as shown in Figure 14 below. We can see from the figure that the mean square error MSE of the neural network model is minimized after 32 training sessions, so the best model corresponds to the smallest MSE, i.e., the thirty-second training model.

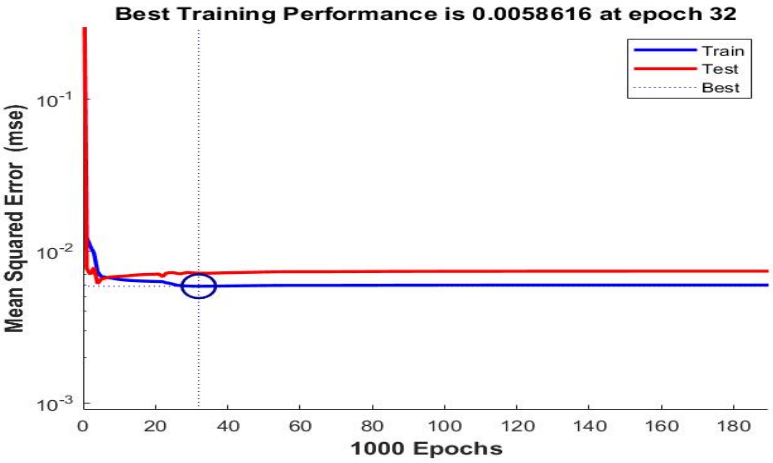


Figure 14 Graph of MSE variation for different training stages

We regressed the fitted values obtained from this model on the initial data after training on the true values and obtained the results shown in Figure 15 below, and according to the principle that the higher the goodness of fit is, the better for the fitting effect, we can see that the fitting effect is relatively good.

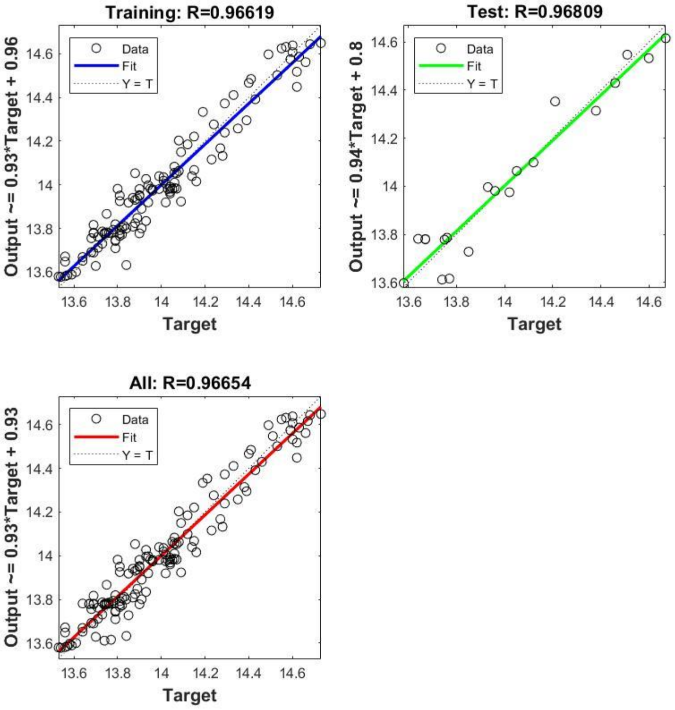


Figure 15 Fitted regression curve

We applied the sim function to the data corresponding to the year dataset to be predicted and substituted it into this training model for the simulation solution, and the resulting prediction results are shown in Annex 8.

Table 3 Prediction results of the BP neural network algorithm

6.3 Discourse of predictive models and evaluation of models

6.3.1 Description of the prediction model

Our results based on the gray prediction model are that the global average temperature will reach 14.84°C in 2050 and 15.35°C in 2100, and we expect that the global average temperature will reach a high temperature of 20°C by about 2400. Our results using BP neural network model are that the global average temperature will reach 20.66°C in 2050 and 21.17°C in 2100. The results are shown in Table 4 below.

Table 4 Overview chart of the prediction model

Models	Gray prediction model	Neural Network Model
Year 2050	14.84℃	20.66℃
Year 2100	15.35℃	21.17℃
Year of reaching 20°C	Year 2400	

6.3.2 Evaluation of the model

The time series model describes the past function relatively well, but for data prediction, it is not suitable for prediction of long-term continuous data and there is a large bias when the external world encounters huge changes. For the gray prediction model, in dealing with less eigenvalue data, it does not require a large enough sample space of data, it can solve the problems of less historical data, the integrity of the sequence and low reliability, and it can generate the irregular original data to get a strong regular generating sequence, but it is only applicable to the short and medium term forecasting, and only suitable for the prediction of the approximate exponential growth, so there is still error in the question. We believe that the BP neural network prediction model is more accurate, it has a strong nonlinear mapping ability, and it is suitable for solving problems with complex internal mechanisms due to more factors influenced by the future, and it can extract the "reasonable rules" between the output and output data through learning during training, and adaptively memorize the learning content in the network It has a high degree of self-learning and self-adaptive ability. Therefore, we believe that the BP neural network model is more accurate.

7 Problem 2 Modeling and Solving

7.1 Analyzing the factors affecting temperature variation

7.1.1 Spearman's correlation coefficient to analyze the factors of temperature change

In order to better explore the relationship between global temperature change and each factor, we use Spearman's correlation coefficient to explore the relationship between each variable separately, where in order to better distinguish the influence of each factor on global temperature change, we divide them into northern and southern hemispheres according to latitude, and according to longitude division rules (i.e., 20 degrees west longitude to 160 degrees east longitude) we divide them into Eastern Hemisphere and Western Hemisphere, for the temperature change we take the average temperature of each region as the determination criterion, we divide the data into four groups and calculate their Spearman correlation coefficients, i.e., Northern Hemisphere Eastern Hemisphere, Northern Hemisphere Western Hemisphere, Southern Hemisphere Eastern Hemisphere and Southern Hemisphere Western Hemisphere regions.

Since Spearman correlation coefficient is used to describe the correlation between two fixed-order data sets, we first transformed the variables into a fixed-order data series, i.e., ranking the size of indicator ratings, and used this to calculate Spearman correlation coefficient, which is calculated as follows: $r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$, where denotes the difference in ratings between two indicators. Since the sample size is >30, we use the statistic $r_s\sqrt{n-1}$~$N(0,1)$. Let H_0: $r_s = 0$,H_1: $r_s \neq 0$, from which the test value $z^* = r_s\sqrt{n-1}$, is calculated to determine the data significance, according to the final results obtained, it is found that the data significance is good, indicating that the model is more accurate, and the specific Spearman correlation coefficient heat map is shown in the following figure.

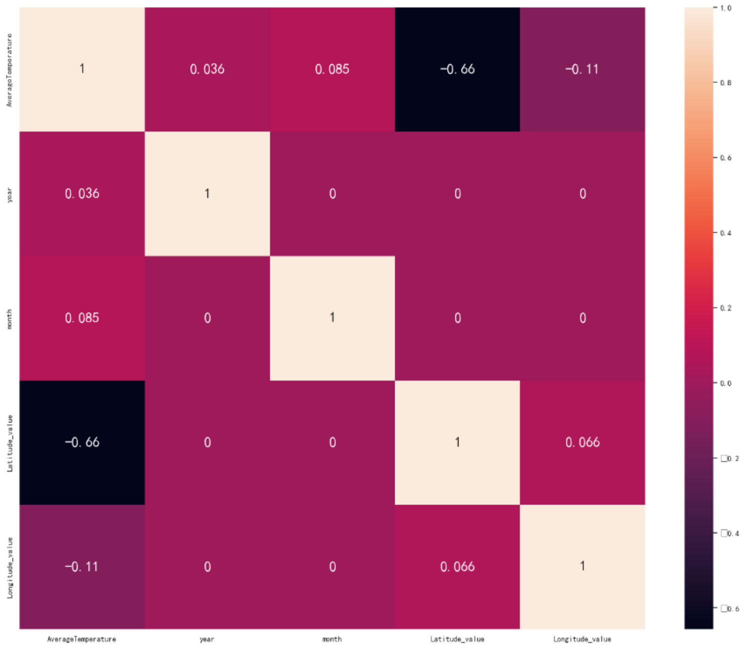


Figure 16 Northern Hemisphere Eastern Hemisphere correlation coefficient heat map

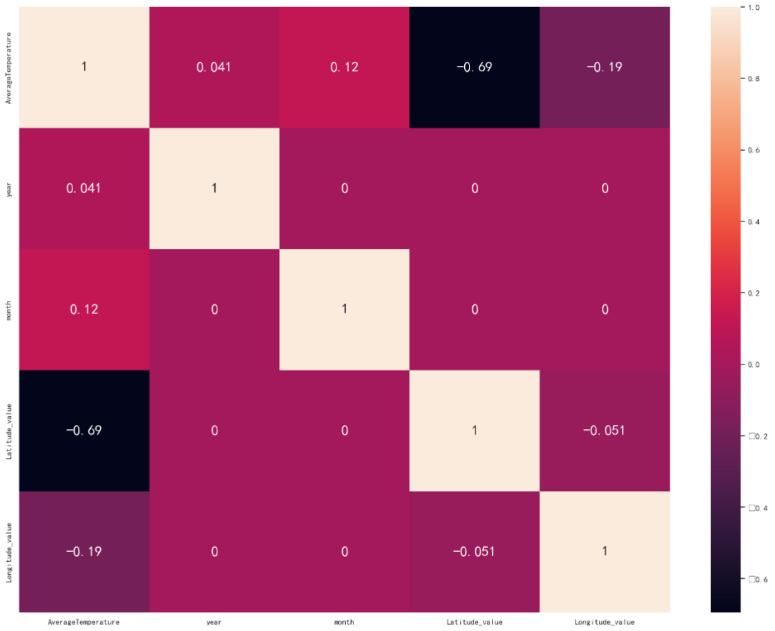


Figure17 Heat map of correlation coefficients in the Northern Hemisphere Western Hemisphere

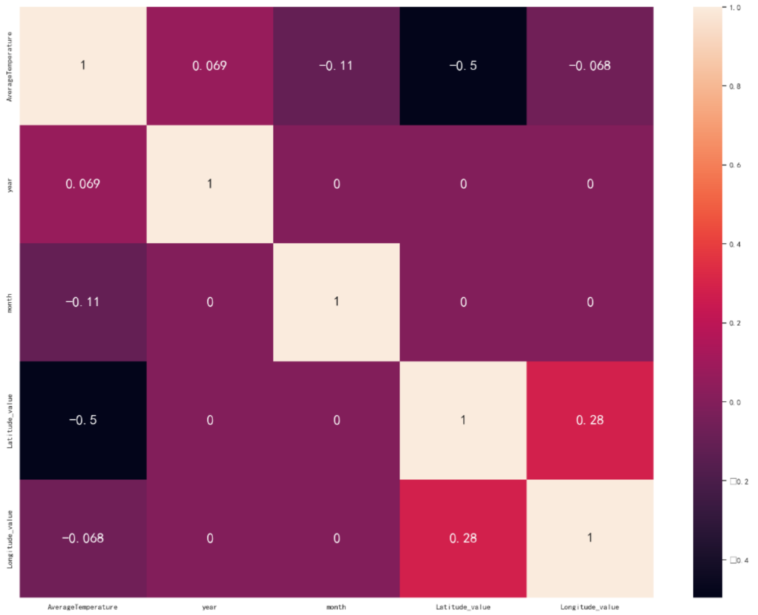


Figure 18 Southern Hemisphere Eastern Hemisphere correlation coefficient heat map

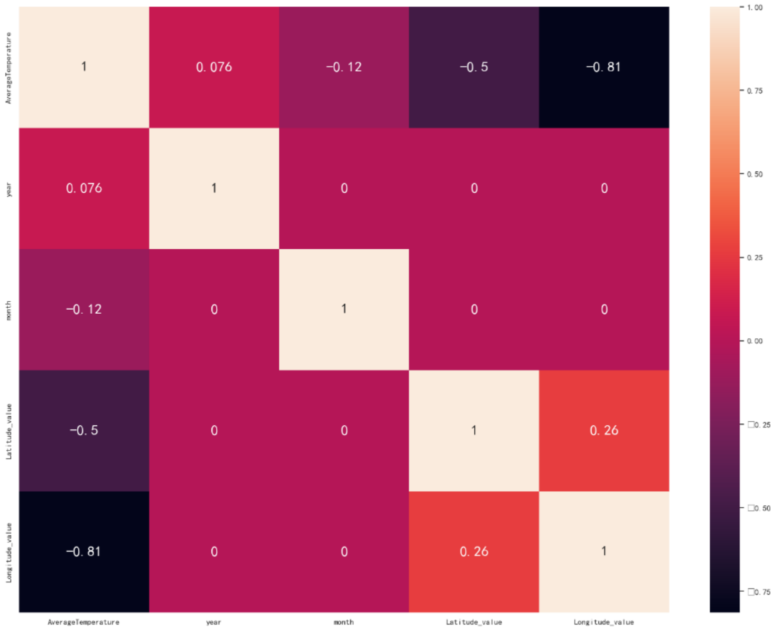


Figure 19 Heat map of correlation coefficients for the Southern Hemisphere Western Hemisphere

Based on the results of the four plots above we can see that, overall, the latitude at which a region is located has the greatest effect on temperature, followed by longitude and month, while year has nearly no relationship with global temperature. For northern hemisphere regions, both in the eastern and western hemispheres, latitude has a greater effect on temperature, followed by longitude, while month and year have a relatively small effect on temperature. For the southern hemisphere region, the degree of influence is slightly different between the eastern and western hemispheres. For the eastern hemisphere region of the southern hemisphere, latitude has the greatest influence, followed by month, while year and longitude have less influence on temperature; while for the western hemisphere region of the southern hemisphere, the greatest difference, unlike any hemisphere, is that the most influential factor is the longitude in which it is located, followed by latitude again, and then month and year both have less influence on temperature.

7.1.2 Multiple linear regression to optimize the correlation factor analysis

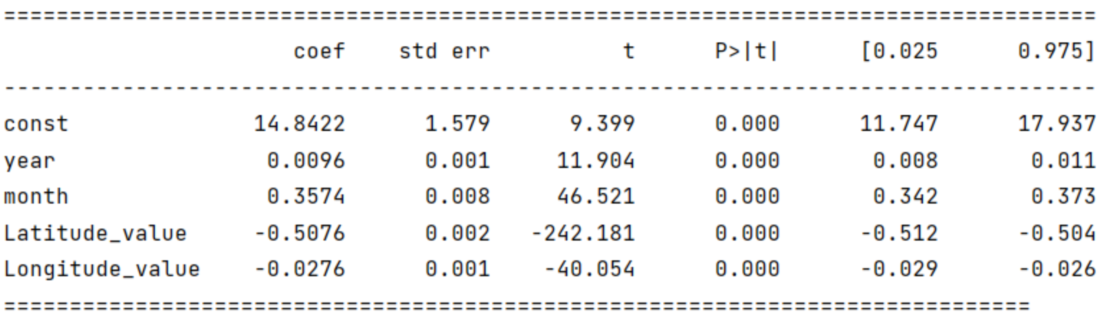
In order to make the relationship results more accurate, we use multiple linear regression for a more precise analysis. The model equation of multiple linear regression.

$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k + e$, b0 is a constant term，b1, b2 , ... bk are the regression coefficients .

We select the regional average temperature data from January 1899 to December 2012 and other relevant data according to the data given in the question. In order to make the data more accurate, we adopt the same grouping method as the Spearman coefficient and divide the data into four groups, which are subjected to multiple linear regressions. Where x_1 is assumed to be the year, x_2 is the month, x_3 is the latitude, x_4 is the longitude, x_1、x_2、x_3、x_4 are all independent variables; y is the regional average temperature as the dependent variable. Python is then used to perform multivariate linear forecasting.

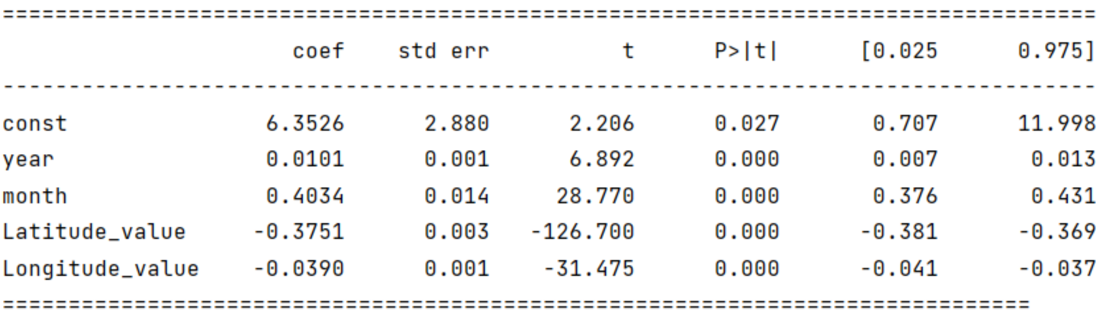
We used least squares to build multiple linear regression models for the influencing factors[4], and used Python to build least squares models for the four factors to obtain least squares linear regression models for each of the four data sets, and the results of the linear regression analysis are tabulated in the following four tables.

Table 5 Northern Hemisphere Eastern Hemisphere Linear Regression Analysis Table



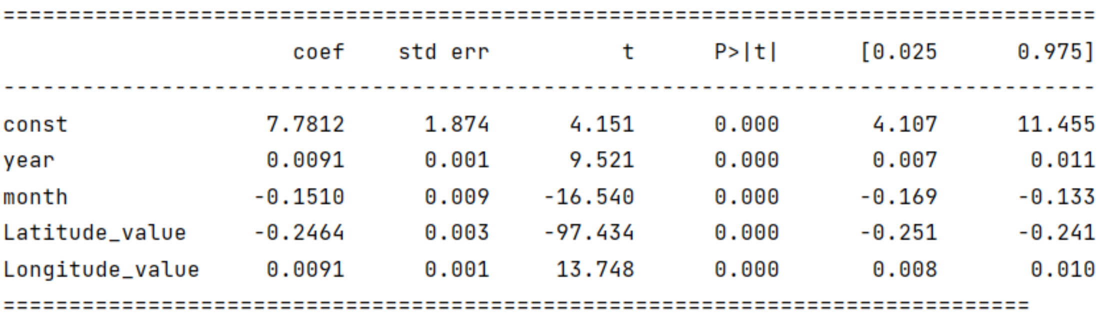
```
==============================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const              14.8422      1.579      9.399      0.000      11.747      17.937
year                0.0096      0.001     11.904      0.000       0.008       0.011
month               0.3574      0.008     46.521      0.000       0.342       0.373
Latitude_value     -0.5076      0.002   -242.181      0.000      -0.512      -0.504
Longitude_value    -0.0276      0.001    -40.054      0.000      -0.029      -0.026
==============================================================================
```

Table 6 Northern Hemisphere Western Hemisphere Linear Regression Analysis Table



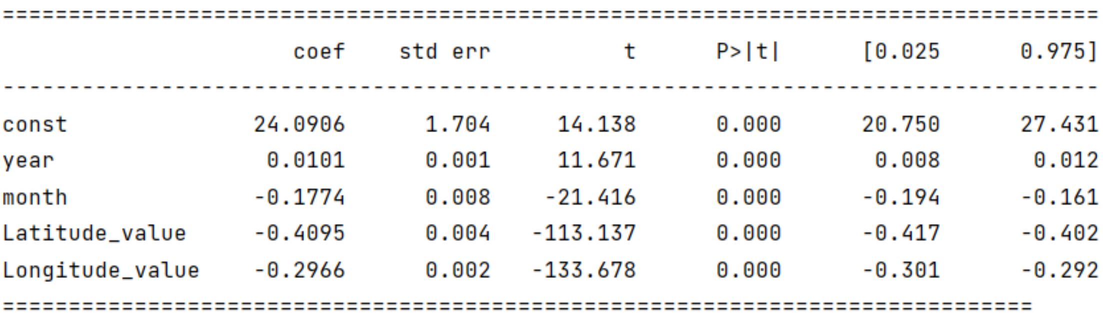
```
==============================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const               6.3526      2.880      2.206      0.027       0.707      11.998
year                0.0101      0.001      6.892      0.000       0.007       0.013
month               0.4034      0.014     28.770      0.000       0.376       0.431
Latitude_value     -0.3751      0.003   -126.700      0.000      -0.381      -0.369
Longitude_value    -0.0390      0.001    -31.475      0.000      -0.041      -0.037
==============================================================================
```

Table 7 Southern Hemisphere Eastern Hemisphere Linear Regression Analysis Table



```
==============================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const               7.7812      1.874      4.151      0.000       4.107      11.455
year                0.0091      0.001      9.521      0.000       0.007       0.011
month              -0.1510      0.009    -16.540      0.000      -0.169      -0.133
Latitude_value     -0.2464      0.003    -97.434      0.000      -0.251      -0.241
Longitude_value     0.0091      0.001     13.748      0.000       0.008       0.010
==============================================================================
```

Table 8 Table of linear regression analysis for the Southern Hemisphere Western Hemisphere



```
==============================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const              24.0906      1.704     14.138      0.000      20.750      27.431
year                0.0101      0.001     11.671      0.000       0.008       0.012
month              -0.1774      0.008    -21.416      0.000      -0.194      -0.161
Latitude_value     -0.4095      0.004   -113.137      0.000      -0.417      -0.402
Longitude_value    -0.2966      0.002   -133.678      0.000      -0.301      -0.292
==============================================================================
```

We can find that the analysis of the results from the F-test can be obtained, the significance P-value is roughly 0, the level presents significance, rejecting the original hypothesis that the regression coefficient is 0, and for its performance of variable covariance, VIF are less than 10, so the model does not have the problem of multiple covariance, so we believe that the model is more reasonably constructed, the specific linear expressions are as follows.

Northern Hemisphere Eastern Hemisphere.:

$$y = 0.010x_1 - 0.357x_2 - 0.508x_3 - 0.028x_4 + 14.842$$

Northern Hemisphere Western Hemisphere.:

$$y = 0.010x_1 - 0.403x_2 - 0.375x_3 - 0.039x_4 + 6.353$$

Southern Hemisphere Eastern Hemisphere:

$$y = 0.009x_1 - 0.151x_2 - 0.41x_3 - 0.246x_4 + 7.781$$

Southern Hemisphere Western Hemisphere：

$$y = 0.010x_1 - 0.177x_2 - 0.41x_3 - 0.297x_4 + 24.091$$

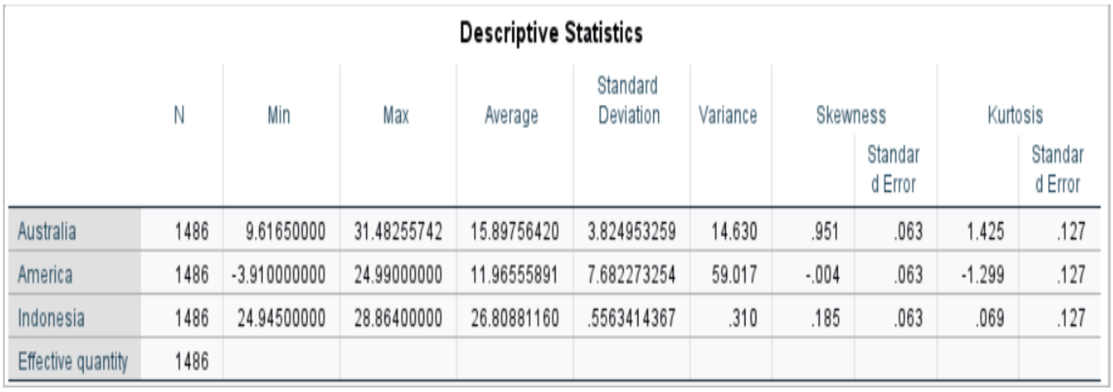
According to the above formula we can find that, in general, the global temperature is most influenced by latitude, and the different heat conditions in different latitudes lead to a greater impact on the temperature. Next is the month, different months, the angle of sunlight is different, resulting in different sunlight area and time, the difference in sunlight, resulting in temperature changes fluctuate. For longitude, the effect on temperature is smaller, probably because different longitudes are located in different geographical areas, there are differences between land and ocean ground, and different warming conditions. Finally, the year has a negligible effect on the temperature, and the possible influence may be the sudden natural disasters that occur every year leading to a slight variation in global temperature.

7.2 Impact of natural disasters on global temperature

7.2.1 Visual analysis of natural disaster impacts

Since we want to study the impact of natural disasters on global temperature, we choose three typical cities and analyze them to draw conclusions. The three cities include, COVID-19 in the U.S., forest fires in Australia and volcanic eruption in Indonesia, we consult the data, the time period when these disasters broke out in each of the three cities, and statistics in word, the specific statistical results are shown in Annex 9. Then we filter the data of the three countries, the U.S., Australia and Indonesia, respectively, according to the data given in the topic, and analyze them from 1899 for visual analysis of the average temperature. In order to make the results more intuitive, we first performed statistical descriptions of the data for the three countries using SPSS, and the results are shown in Table 9.

Table 9 Statistical description table



Descriptive Statistics

	N	Min	Max	Average	Standard Deviation	Variance	Skewness		Kurtosis	
								Standard Error		Standard Error
Australia	1486	9.61650000	31.48255742	15.89756420	3.824953259	14.630	.951	.063	1.425	.127
America	1486	-3.910000000	24.99000000	11.96555891	7.682273254	59.017	-.004	.063	-1.299	.127
Indonesia	1486	24.94500000	28.86400000	26.80881160	.5563414367	.310	.185	.063	.069	.127
Effective quantity	1486									

From this data, we can find that the annual average temperature in Indonesia is relatively stable, while the difference between the peak and trough of the annual average temperature in the United States and Australia is large, indicating that there are other factors that have a large impact on the annual average temperature in these two cities. In order to make the data more visual and clear, we present a line graph of the annual average temperature for each of the three cities based on the information we have accessed, and the results are shown in Figures 20, 21 and 22.

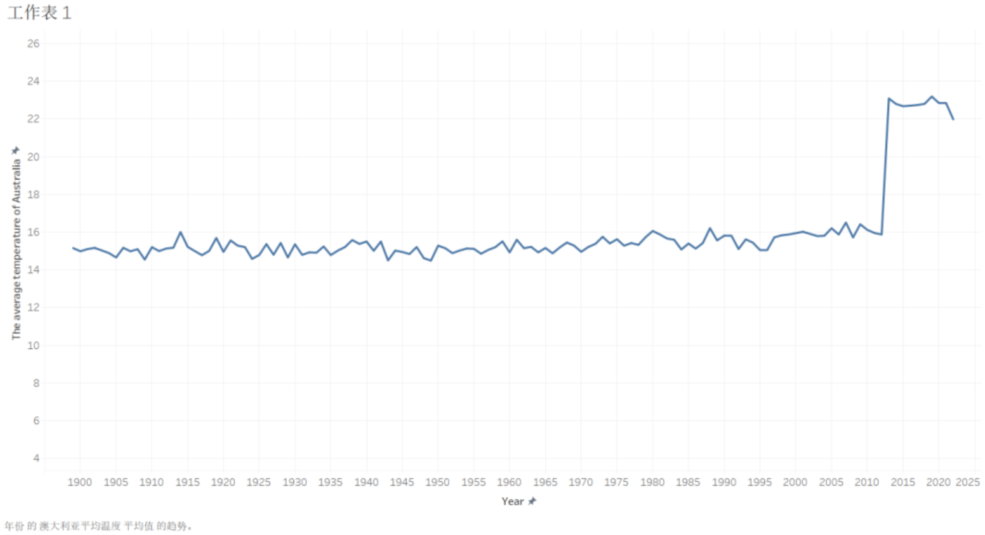


Figure 20 Folding line of annual average temperature in Australia

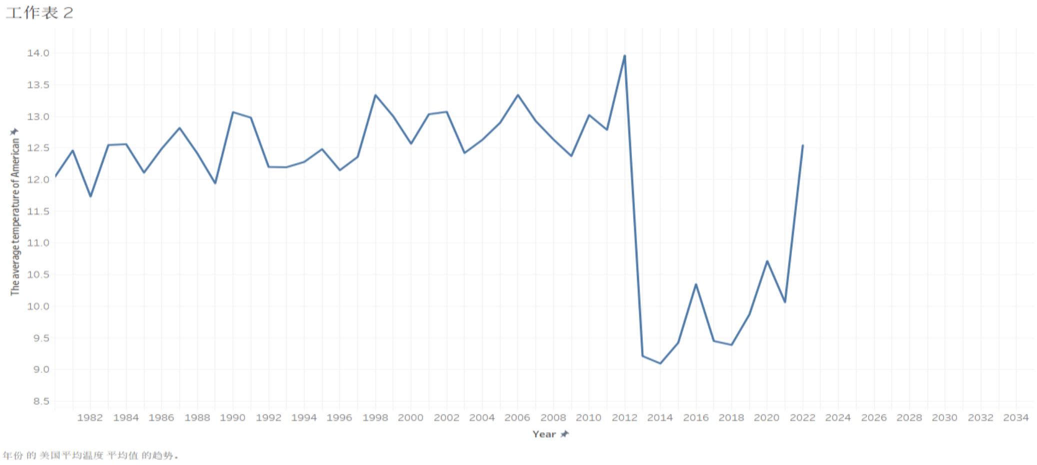


Figure 21 Folding line graph of annual average temperature in the United States

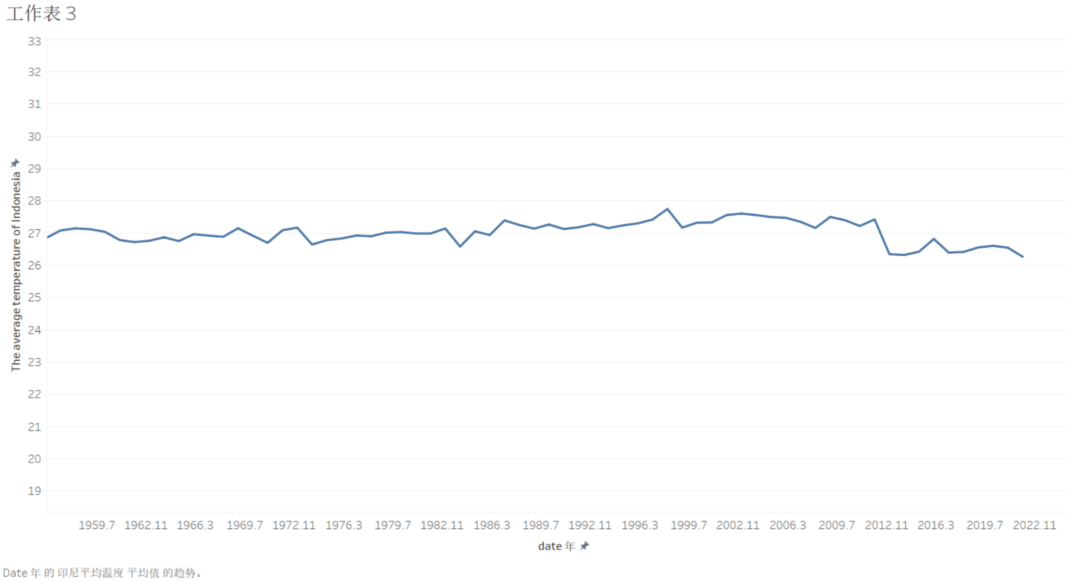


Figure 22 Folding line of annual average temperature in Indonesia

According to the data collected in front of the results of the three visualization tables we found that: for the United States, the emergence of the new crown epidemic in 2020, resulting in a sudden drop in temperature by 2021, the possible reason is that the epidemic led to a stagnant level of industrialization in the United States, reducing industrial carbon emissions, as well as the restrictions on people's lives, resulting in a sudden drop in temperature, and the turnaround in temperature from 2021 to 2022, precisely after the epidemic improved and people Life returned to regularity before the annual average temperature in the U.S. gradually returned to normal, indicating the existence of an impact of the epidemic on global temperature change. For Australia, there were major forest fires from 2019 to 2020, and we found that in those two years the temperature in Australia rose sharply, and the occurrence of forest fires was a signal of rising global temperatures, because the rise in temperature and the weather becoming hot and dry led to forests being more prone to fires and caused local temperatures to rise during the fires, which also contributed to the rise in temperature. For Indonesia, we find that it had volcanic eruptions in 1963 to 1964, 1982, 1983, 2012 and 2014, and these times we can find a sharp drop in temperature in Indonesia because volcanic eruptions release large amounts of hot ash and gases that enter the atmosphere that can block solar radiation and lead to lower temperatures. And the large size and weight of volcanic ash particles will be stagnant in the atmosphere, and it will also affect the atmospheric circulation system, causing the global temperature to drop. In summary, we can find that natural disasters affect global temperature change, and the impact is greater when natural disasters occur.

7.2.2 Principal component analysis to study the factors affecting temperature change

The Earth's surface temperature is determined by the amount of solar radiation hitting the Earth's surface and the amount of infrared radiation reflected back from the Earth's surface. Gases such as carbon dioxide, water vapor, ozone, carbon monoxide, and sulfur dioxide in the atmosphere can easily pass through solar short-wave radiation, but can absorb long-wave radiation reflected from the surface. Therefore, more than30 gases such as carbon dioxide are also called greenhouse gases, which will cause the greenhouse effect. Let's examine the effects of gases such as carbon dioxide on temperature.

We use PCA principal component analysis to study the effect size of seven data of CO2, O3, API, PM2.5, SO2, NO2, and CO on the average temperature of continental observation sites. The data see Annex6.

We take n samples, p indicators, and form a sample matrix x of size n*p.

$$x = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = (x_1, x_2, \cdots, x_p)$$

First we standardize the data.

The mean value $\overline{x_j} = \frac{1}{n}\sum_{i=1}^{n} \alpha\, x_{ij}$ and standard deviation $S_j = \sqrt{\frac{\sum_{i=1}^{n} x(x_{ij}-\overline{x_j})^2}{n-1}}$ are calculated by column to obtain the standardized data $X_{ij} = \frac{x_{ij}-\overline{x_j}}{S_j}$, and the original sample matrix is normalized into:

$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} = (X_1, X_2, \cdots, X_p)$$

Calculate the covariance matrix of the standardized sample.

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix}$$

Among them: $r_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(X_{ki} - \overline{X_i})(X_{kj} - \overline{X_j}) = \frac{1}{n-1}\sum_{k=1}^{n} \alpha\, X_{ki}X_{kj}$

$$R = \frac{\sum_{k=1}^{n}(x_{ki} - \overline{x_i})(x_{kj} - \overline{x_j})}{\sqrt{\sum_{k=1}^{n} \alpha\,(x_{ki} - \overline{x_i})^2 \sum_{k=1}^{n} \alpha\,(x_{kj} - \overline{x_j})^2}}$$

The covariance matrix is obtained as follows.

```
[[ 1.01219512  0.93451855  0.96274348  0.94706008  0.4947746   0.8289641
  -0.73993363 -0.47000926]
 [ 0.93451855  1.01219512  0.93135026  0.94108115  0.70271783  0.9121182
  -0.92194057 -0.55781483]
 [ 0.96274348  0.93135026  1.01219512  0.93856079  0.60843717  0.80345429
  -0.75669791 -0.46301209]
 [ 0.94706008  0.94108115  0.93856079  1.01219512  0.65272263  0.76174401
  -0.82273325 -0.52197664]
 [ 0.4947746   0.70271783  0.60843717  0.65272263  1.01219512  0.55416208
  -0.77722805 -0.45895254]
 [ 0.8289641   0.9121182   0.80345429  0.76174401  0.55416208  1.01219512
  -0.86764895 -0.50426159]
 [-0.73993363 -0.92194057 -0.75669791 -0.82273325 -0.77722805 -0.86764895
   1.01219512  0.54783506]
 [-0.47000926 -0.55781483 -0.46301209 -0.52197664 -0.45895254 -0.50426159
   0.54783506  1.01219512]]
```

The eigenvalues and eigenvectors of R are calculated, and the eigenvalues are sorted in descending order, and the results are as follows.

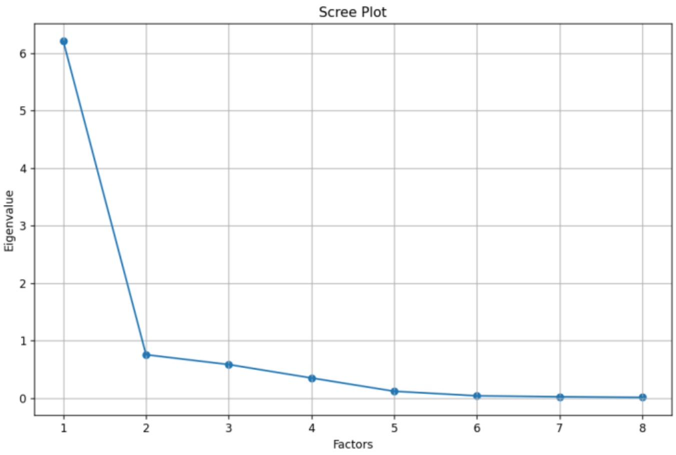
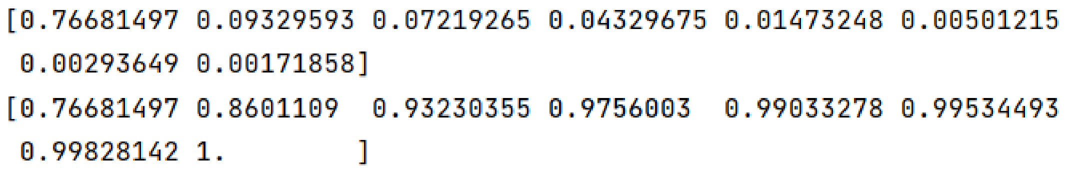


Figure23 Eigenvalue Scree Plot

The principal component contribution and cumulative contribution were calculated and the results were as follows.



```
[0.76681497 0.09329593 0.07219265 0.04329675 0.01473248 0.00501215
 0.00293649 0.00171858]
[0.76681497 0.8601109  0.93230355 0.9756003  0.99033278 0.99534493
 0.99828142 1.        ]
```

The heat diagram of the correlation coefficient is plotted as follows.

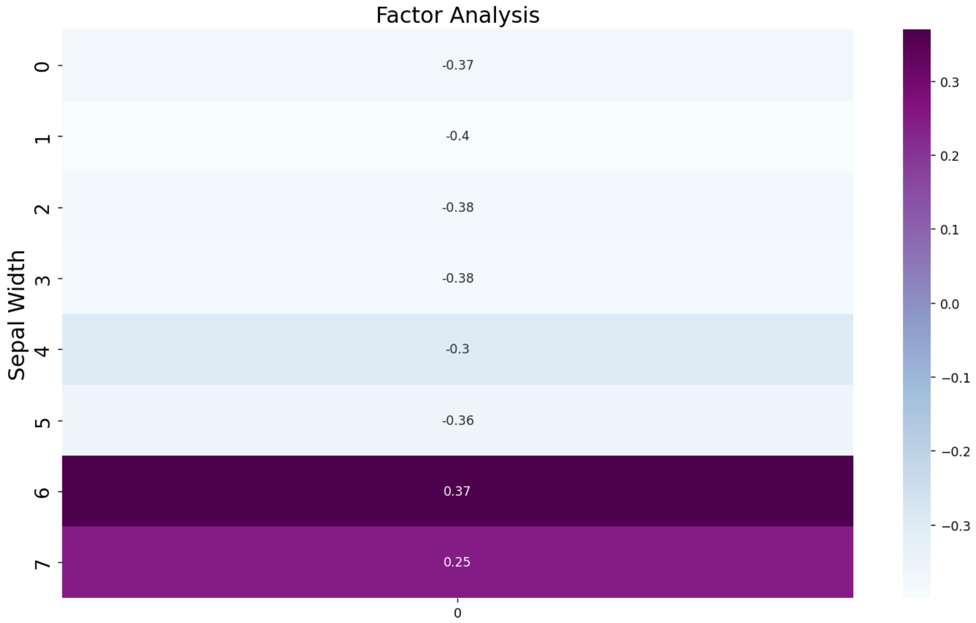
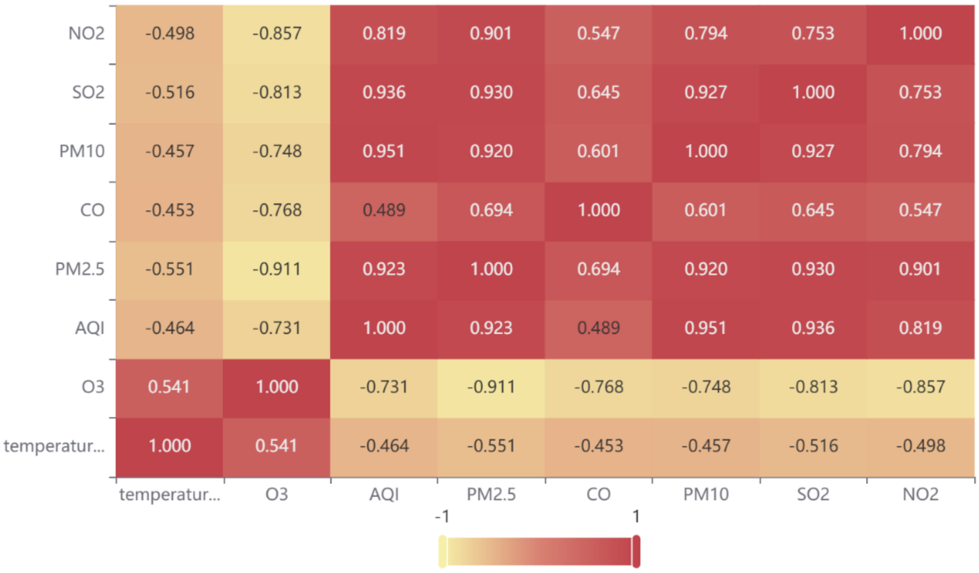


Figure 24 Correlation coefficient percentage graph



	temperatur...	O3	AQI	PM2.5	CO	PM10	SO2	NO2
NO2	-0.498	-0.857	0.819	0.901	0.547	0.794	0.753	1.000
SO2	-0.516	-0.813	0.936	0.930	0.645	0.927	1.000	0.753
PM10	-0.457	-0.748	0.951	0.920	0.601	1.000	0.927	0.794
CO	-0.453	-0.768	0.489	0.694	1.000	0.601	0.645	0.547
PM2.5	-0.551	-0.911	0.923	1.000	0.694	0.920	0.930	0.901
AQI	-0.464	-0.731	1.000	0.923	0.489	0.951	0.936	0.819
O3	0.541	1.000	-0.731	-0.911	-0.768	-0.748	-0.813	-0.857
temperatur...	1.000	0.541	-0.464	-0.551	-0.453	-0.457	-0.516	-0.498

Figure 25 Heat map of correlation coefficient

From the above analysis, it is concluded that CO2 concentration and O3 concentration have the greatest influence on the global average temperature.

7.2.3 Suggested measures to slow down global warming

Comprehensive analysis above, we believe that there are three main factors affecting the temperature, the first one is the anthropogenic cause, including the surge of population and a series of industrialization activities of the population, etc. Because of human activities, it leads to the surge of gas emissions such as CO2 and O3, which leads to the global temperature warming. The second is the latitude factor, the fundamental source of heat in the atmosphere is solar radiation, in low latitudes, there is more solar radiation and high temperatures, in high latitudes, there is less solar radiation and low temperatures, so latitude affects global temperature change. The third is the sudden change of environment, such as natural disasters volcanic eruptions, forest fires and epidemics, all of which lead to sudden changes in the environment, the land is destroyed, meteorological conditions receive changes, and human activities are affected, all of which lead to global temperature changes.

Therefore, we propose the following recommendations.

1. Strictly control human activities, such as car travel and factory emissions that produce large amounts of CO2 or oxy-nitride gases.
2. Reduce the use of fossil fuels.
3. Plant trees to prevent wind and consolidate soil.
4. Protect the natural environment and reduce the occurrence of extreme natural disasters.

8 Evaluation and improvement of the model

9.1 Advantages of the model

1. BP neural network is used to predict the data, which makes the results more accurate.
2. For the correlation analysis of relevant factors, we not only use Spearman correlation coefficient, but also cite multiple linear regression to better reflect the relationship between factors through regression equation.
3. The time series model is used to describe the past and the visual analysis is used to describe the impact of natural disasters, which makes the thesis statement more intuitive.

9.2 Disadvantages of the model

1. When building the model, the impact of some future factors on natural variability is ignored.
2. When studying the impact of natural disasters on global temperature levels, most of the cases found are typical cases, and there is a lack of universal cases.

9.3 Improvement of the model

1. More comprehensive access to information can be made, especially for natural disasters, so that the correlation between natural disasters and global average temperature can be better found.
2. When building the prediction model, we can consider adding the future part of the factors that affect the future global average temperature to increase the accuracy of the prediction.

9 Article

The problems caused by global warming are already appearing more and more around the world in different ways, and we should pay great attention to them. We have analyzed the extent of temperature increase through 10 years of temperature data and found that the global warming trend has become more and more obvious and the temperature increase can no longer be ignored. With the help of the average temperature data of the observation points from the late 19th century to the present, we found that the temperature of the land observation points will reach 20.66℃ in 2050 and 21.17℃ in 2100, and according to such a trend for each of us, it will only feel hotter every year. Then we have to find out the cause of global warming and take appropriate measures for it.

So, we first analyzed the specific factors of temperature change in the collected data, and we found that the latitude and month are different, but these factors play a bigger role in the temperature change. In the hot summer and dry autumn, natural disasters such as volcanic eruptions and forest fires are likely to occur, and the damage caused by natural disasters will in turn increase the temperature, creating a vicious circle. Therefore, we urgently need to take measures to improve the situation.

In this regard, our team decided to make the following recommendations.

1. advocate low-carbon life, raise awareness of energy saving and reduce carbon footprint

At present, people do not know much about low-carbon economy, and the concept of high-carbon consumption is still deeply rooted in people's hearts. The first step is to establish a low-carbon concept and awareness of energy conservation, and to create a strong social atmosphere for environmental protection through TV, newspapers, the Internet and other media, so as to raise people's awareness of environmental protection.

2. Carry out afforestation and increase forest carbon sinks

First of all, we should protect the existing forest resources, expand the forest protection area and reduce the destruction of forests. Secondly, we mobilize the general public to actively plant trees to expand the area of forest vegetation. Increase the greening area by improving grassland, treating degraded, sandy and alkaline grassland, and treating desertified land.

3. Strengthen scientific and technological innovation, save traditional energy, develop new energy, and optimize energy structure

Vigorously promote the development and research of energy-saving technology, pay attention to international cooperation and exchange, and improve the efficiency of traditional energy use. Develop China's green energy sources: solar energy, wind energy, water energy, nuclear energy, tidal energy, geothermal energy, etc. Reduce the proportion of coal in the energy consumption structure, further optimize the energy structure, and promote the diversification of the energy structure.

4. vigorously promote key projects to reduce emissions

Implement emission reduction tasks and responsibilities, implement mandatory measures and penalty mechanisms, and set up special funds and technical support in promoting key emission reduction projects to ensure the successful completion of the tasks.

10 Reference

[1] Zhu R. J. Mann-Kendall-based analysis of winter temperature trends and abrupt changes in Xianyang City [J]. Journal of Xianyang Normal College, 2020, 35(6):8.

[2] Alex Tech Bolg, [Time series] How to understand ACF and PACF, https://blog.csdn.net/qq_41103204/article/details/105810742, November 26, 2022

[3] He Shengjia. Research on deep foundation pit slope top displacement prediction based on GA-BP neural network model [J]. Engineering Technology Research, 2022, 7(15):4.

[4] Li Bingxiao, Zhang Shiwei, Zheng Shuyu, et al. Research on hydroelectric power prediction based on combined multiple linear regression and ARIMA model[J]. Science and Technology Innovation, 2022(33):4.

Appendix

Appendix 1:

Description: A list of documents supporting materials

1. Annex 1: Data table after processing missing values by cubic spline interpolation
2. Annex 2: Data table after data standardization
3. Annex 3: Average temperature data table by region
4. Annex 4: Global average temperature data table
5. Annex 5: Observation Point Average Temperature Data Table
6. Annex 6: Data on air quality indicators in China
7. Annex 7: Grey projections of future global average temperature at observations
8. Annex 8: BP neural network predictions of future global average temperature at observation points
9. Annex 9: Timing of natural disasters in three cities

Appendix 2:

Description: List of specific code files

1. Attachment 1: Python implements cubic spline interpolation code
2. Annex 2: Python implements standardized data code
3. Annex 3: MATLAB calculates the average temperature codes for each region
4. Annex 4: Python implements the Mann-Kendall mutation analysis code
5. Annex 5: Time Series Model and Grey Forecast Code
6. Annex 6: BP Neural Network Prediction Code
7. Annex 7: Spearman correlation analysis code
8. Annex 8: Multiple Linear Regression Optimization Model Code
9. Annex 9: Principal Component Analysis code

Due to the huge amount of code, only some of the code is shown below, please see the attachment for the detailed code.

Standardize data codes：

```
from sklearn import preprocessing
df = preprocessing.scale(df)
print(df)
```

Cubic spline interpolation code：

```
def calculateEquationParameters(x):
# parameter 为二维数组，用来存放参数，sizeOfInterval 是用来存放区间的个数
parameter = []
sizeOfInterval = len(x) - 1;
i = 1
# 首先输入方程两边相邻节点处函数值相等的方程为 2n-2 个方程
while i < len(x) - 1:
    data = init(sizeOfInterval * 4)
    data[(i - 1) * 4] = x[i] * x[i] * x[i]
    data[(i - 1) * 4 + 1] = x[i] * x[i]
    data[(i - 1) * 4 + 2] = x[i]
    data[(i - 1) * 4 + 3] = 1
    data1 = init(sizeOfInterval * 4)
    data1[i * 4] = x[i] * x[i] * x[i]
    data1[i * 4 + 1] = x[i] * x[i]
    data1[i * 4 + 2] = x[i]
    data1[i * 4 + 3] = 1
    temp = data[2:]
    parameter.append(temp)
    temp = data1[2:]
    parameter.append(temp)
    i += 1
# 输入端点处的函数值。为两个方程，加上前面的 2n - 2 个方程，一共 2n 个方程
data = init(sizeOfInterval * 4 - 2)
data[0] = x[0]
data[1] = 1
parameter.append(data)
data = init(sizeOfInterval * 4)
data[(sizeOfInterval - 1) * 4] = x[-1] * x[-1] * x[-1]
data[(sizeOfInterval - 1) * 4 + 1] = x[-1] * x[-1]
data[(sizeOfInterval - 1) * 4 + 2] = x[-1]
data[(sizeOfInterval - 1) * 4 + 3] = 1
temp = data[2:]
parameter.append(temp)
```

Python implements the Mann-Kendall mutation analysis code:

```
    def mktest(inputdata):
import numpy as np
inputdata = np.array(inputdata)
n = inputdata.shape[0]
Sk = np.zeros(n)
UFk = np.zeros(n)
r = 0
for i in range(1, n):
    for j in range(i):
        if inputdata[i] > inputdata[j]:
            r = r + 1
    Sk[i] = r
    E = (i + 1) * i / 4
    Var = (i + 1) * i * (2 * (i + 1) + 5) / 72
    UFk[i] = (Sk[i] - E) / np.sqrt(Var)

Sk2 = np.zeros(n)
UBk = np.zeros(n)
inputdataT = inputdata[::-1]
```

```
    r = 0
    for i in range(1, n):
        for j in range(i):
            if inputdataT[i] > inputdataT[j]:
                r = r + 1
        Sk2[i] = r
        E = (i + 1) * (i / 4)
        Var = (i + 1) * i * (2 * (i + 1) + 5) / 72
        UBk[i] = -(Sk2[i] - E) / np.sqrt(Var)
    UBk2 = UBk[::-1]
    return UFk, UBk2
```

Time Series Model and Grey Forecast Code：

```
# 1、数据均值化处理
x_mean = x.mean(axis=1)
for i in range(x.index.size):
    x.iloc[i, :] = x.iloc[i, :] / x_mean[i]

# 1、数据差值化处理
x = (x - x.min()) / (x.max() - x.min())
x = x.T

# 1、数据初值化处理
x_mean = x.mean(axis=1)
for i in range(x.index.size):
    x.iloc[i, :] = x.iloc[i, :] / x.iloc[i, 0]

# 2、提取参考队列和比较队列
ck = x.iloc[0, :]
print("参考队列：", ck)
cp = x.iloc[1:, :]
print("参考队列：", cp)

# 比较队列与参考队列相减
t = pd.DataFrame()
for j in range(cp.index.size):
    temp = pd.Series(cp.iloc[j, :] - ck)
    t = t.append(temp, ignore_index=True)

# 求最大差和最小差
mmax = t.abs().max().max()
mmin = t.abs().min().min()
rho = 0.4
```

```
# 3、求关联系数
ksi = ((mmin + rho * mmax) / (abs(t) + rho * mmax))

# 4、求关联度
r = ksi.sum(axis=1) / ksi.columns.size

# 5、关联度排序，得到结果
result = r.sort_values(ascending=False)
```

Spearman correlation analysis code:

```
myfont = FontProperties(fname=r'C:\Windows\Fonts\simhei.ttf', size=40)
sns.set(font=myfont.get_name(), color_codes=True)
# corr = df.corr(method='pearson')  # 使用皮尔逊系数计算列与列的相关性
# corr = df.corr(method='kendall')  # 肯德尔秩相关系数
data_corr = data.corr(method='spearman')  # 斯皮尔曼秩相关系数
# data_corr = data.corr(method='pearson')  # 使用皮尔逊系数计算列与列的相关性
plt.figure(figsize=(20, 15))  # figsize可以规定热力图大小
fig = sns.heatmap(data_corr, annot=True, fmt='.2g', annot_kws={'fontsize': 20})  # annot为热力图上显示数据；fmt='.2g'为数据保留两位有效数字
print(fig)
fig.get_figure().savefig(location + '_S.png')  # 保留图片
```

Principal Component Analysis code:

```
# 求解系数相关矩阵
covX = np.around(np.corrcoef(df.T), decimals=3)
# print(covX)

# 求解特征值和特征向量
featValue, featVec = np.linalg.eig(covX.T)  # 求解系数相关矩阵的特征值和特征向量
# print(featValue, featVec)
# 去中心化
data_adjust = df - avgs
# print(data_adjust)

# 协方差阵
covX = np.cov(data_adjust.T)  # 计算协方差矩阵
# print('协方差矩阵')
# print(covX)

# 计算协方差阵的特征值和特征向量
```

```
featValue, featVec = np.linalg.eig(covX)  # 求解协方差矩阵的特征值和特征向量
# 绘制散点图和折线图
# 同样的数据绘制散点图和折线图
plt.scatter(range(1, df.shape[1] + 1), featValue)
plt.plot(range(1, df.shape[1] + 1), featValue)

# 显示图的标题和 xy 轴的名字
# 最好使用英文，中文可能乱码
plt.title("Scree Plot")
plt.xlabel("Factors")
plt.ylabel("Eigenvalue")

plt.grid()  # 显示网格
plt.show()  # 显示图形

# 求特征值的贡献度
gx = featValue / np.sum(featValue)

# print(gx)

# 求特征值的累计贡献度
lg = np.cumsum(gx)
# print(lg)

# 选出主成分
k = [i for i in range(len(lg)) if lg[i] < 0.85]
k = list(k)
# print(k)

# 选出主成分对应的特征向量矩阵
selectVec = np.matrix(featVec.T[k]).T
selectVe = selectVec * (-1)
# print(selectVec)

# 主成分得分
finalData = np.dot(data_adjust, selectVec)
print(finalData)
```