Problem Chosen	2021	Team Control Number
B	ShuWei Cup Summary Sheet	202111097271

Summary

In order to cope with the fluctuating weather, we carry out an analysis related to precipitation and indicators affecting rainfall in selected cities.

For question 1, the data from the tables are first integrated by year and quarter. Considering the periodic nature of precipitation, a Fourier function was chosen to fit, and the result of the fit was $f(x) = 1.453e + 09 - 1.446e + 09cos(-4.708e - 05x) + 1.358e + 08sin(-4.708e - 05x)$ Precipitation variability has the characteristics of small periods, large ups and downs, and is prone to extreme precipitation. An M-K test was performed to analyse the points where sudden changes in precipitation occurred. The years with high precipitation were selected to analyse each indicator and the main indicators GUSP and WDSP were derived using the correlation and entropy weighting method, indicating that precipitation may be related to typhoons, cyclones.

For question 2, choosing Changsha and Taiyuan for quantitative analysis, yields that Changsha has high precipitation but natural cycles, while Taiyuan has consistently low precipitation but too small a cycle, and may be in a situation of both flood and drought resistance for a short period of time.

For question 3, build a prediction model about climate, reduce the data from 12 to 9 dimensions by principal component analysis, normalise to remove the magnitude. 1. neural network, use LSTM optimization, choose 96X3 hidden layer, better prediction effect, disadvantage is long training time, large computation. 2. use support vector machine. The presence or absence of heavy rainfall was transformed into a 0-1 variable to complete the clustering analysis 3. Using plain Bayesian theory, the probability of precipitation was estimated by expressing rainfall with as 0-1. Combined with the above analysis, the best prediction is the LSTM neural network, which is calculated using the data of prediction test effect to yield $R^2 = 0.89124$, but the disadvantage is that the computation is too large and difficult to be generalized on other models, followed by the plain Bayesian statistic $R^2 = 0.638751$. The shorter operation time facilitates the generalisation of the use.

For question 4, precipitation data around one month at the specified time were intercepted. Combining the climatic characteristics, it can be seen that the July rainstorm in Zhengzhou is mainly influenced by typhoons, the October rainstorm is mainly caused by the southward north wind, and the October rainstorm in Shanxi is mainly influenced by the subtropical high pressure, which is caused by climate warming.

For question 5, combined with the above models, each city should plan and build rationally to ensure that the drainage system in the city and the infrastructure in the countryside are adequate, while reducing emissions to ease the pressure from global warming.

Key word: Relevance, Manner-Kendall (M-K) - mutation test,Entropy method, SVM, nerual network,Plain Bayesian,Precipitation

Content

1. Introduction

1.1 Problem Background

Global warming has become an inescapable topic for us under the influence of both natural and man-made factors, and the background of global warming has led to a flurry of research into the problems we may face and the measures we can take to combat them. One of the direct effects of global warming is an increase in precipitation - the amount, intensity, frequency and type of precipitation that will affect our country in the future. Researchers estimate that by the end of the century, the amount of precipitation in the country will have increased by about 10 percent while the probability of precipitation will have increased by about 10 percent , more seriously, the probability of extreme precipitation will have increased significantly.

1.2 Problem Restatement

In recent years, this trend has already started to become apparent - many parts of China such as Henan, Shaanxi and Hubei have experienced rare heavy rainfall in the last two years, and some northern cities have been hit by historically rare snowstorms. These natural disasters can pose a serious threat to the lives, safety and property of local people. Take the heavy rainfall in Zhengzhou, for example, which accumulated an average of 449mm from 18:00 on 18 July to 00:00 on 21 July. From 16:00 to 17:00 on the 20th alone, the precipitation at Zhengzhou station exceeded the extreme value of hourly precipitation on land in China, reaching 201.9 mm. Specifically, the downpour began intermittently on July 17, and by the morning of July 20, the rain suddenly began to increase. In just half a day, many communities and roads in Zhengzhou were flooded with rain. And the amount of rainfall over the three days approached or even exceeded the previous annual average rainfall in Zhengzhou - 640.8 mm, which was described by meteorologists as a once-in-a-millennium event. As a city with a large population, Zhengzhou experienced huge losses in the safety of people's lives and property during this rainstorm - from the the result of the statistics ,as of 12:00 on July 23 a total of 395,989 people were urgently relocated and housed, and the serious impact on production and life - -Crop damage amounted to 44,209.73 hectares, as well as direct economic losses of RMB 65.5 billion. More noteworthy are the severe losses caused by secondary disasters, which have killed hundreds of people in the floods and torrential rains.

Against the backdrop of increased precipitation due to global warming and the increased likelihood of extreme weather, the country's large land area often presents different topographical features in different areas or different parts of the same region. Precipitation characteristics also vary from city to city. It is therefore imperative to develop predictive models and quantitative analysis of losses for cities with different potential extreme precipitation events .

Question 1: A correlation analysis of the annual variability of precipitation characteristics in the Zhengzhou region is carried out to filter out some years with high precipitation. Also, do a specific quantitative analysis of inundation events in the Zhengzhou area.
Question 2: Collect and collate precipitation data from more cities in China over a number of years and analyse precipitation trends?
Question 3: Forecast and analyse cities that are likely to experience extreme precipitation in the future based on different methods and compare and analyse the effectiveness of your forecasts?
Question 4: Are the characteristics of heavy rainfall in Zhengzhou in July 2021 the same as in other cities? Are the characteristics of heavy rainfall in Shanxi in July/October the same? What are the differences?
Question 5: Propose a long-term construction plan for the future city under extreme precipitation conditions?

2. Analysis of the Problem

This question requires a mathematical and physical analysis of urban precipitation using the data given in the annex. It involves the interception of valid data, the examination of data characteristics, correlation analysis and the evaluation of forecasts. Prior to analysis, the data is intercepted and pre-processed to detect any outliers, missing values and, if present, to reject and fill in the differences.

2.1 analysis of question one

Inter-annual variability: Inter-annual variability requires the integration of precipitation data for each year by year or by quarter. Different indicators are taken, focusing on the cyclical nature of precipitation and sudden changes in precipitation.
The years with high precipitation are filtered by annual precipitation and these are used to analyse the indicators affecting precipitation, either in the direction of correlation, weights, etc.

2.2 analysis of question two

For the analysis of precipitation data, simply expand the data to include precipitation data from other cities, in a similar way to Question 1.

2.3 analysis of question three

Question 3 requires consideration of the impact of indicators other than year on precipitation and prediction of cities with extreme precipitation in the future, taking into account the correlation between the indicators. The methods used include moving average, time series, exponential smoothing, ARIMA models and neural network models. 10% of the data will be used to test the predictions and 90% to learn and predict the model.

2.4 analysis of question four and five

Question 4: The question requires the analysis and evaluation of different sets of data and the drawing of relevant images to represent their relevant mathematical and physical meanings
Question 5: Integrate the above models and make recommendations for a typical city in China in terms of rainfall.

3. Assumptions and Justifications

The following are six common types of model assumptions.

1. The data given in the annexes are all true and reliable, with no outliers or missing values or small numbers that can be excluded.
2. Excluding the effects of factors other than those given in the Annex (e.g. artificial rainfall, unusual weather, anthropogenic climate change, changes in air quality, etc.)
3. The effects of anomalous phenomena are not considered, only the normal rainfall phenomena in the selected areas.
4. The data found are independent of the amount of precipitation that different cities receive from each other and the amount of precipitation that affects them, and do not take into account errors caused by the interaction of meteorological indicators between regions.

5. Excluding errors due to missing data caused by daily precipitation exceeding the upper statistical limit of 99 mm
6. Errors of less than two orders of magnitude due to special years not taken into account are not taken into account

4. Notations

For convenience, we use the following symbols in our models.

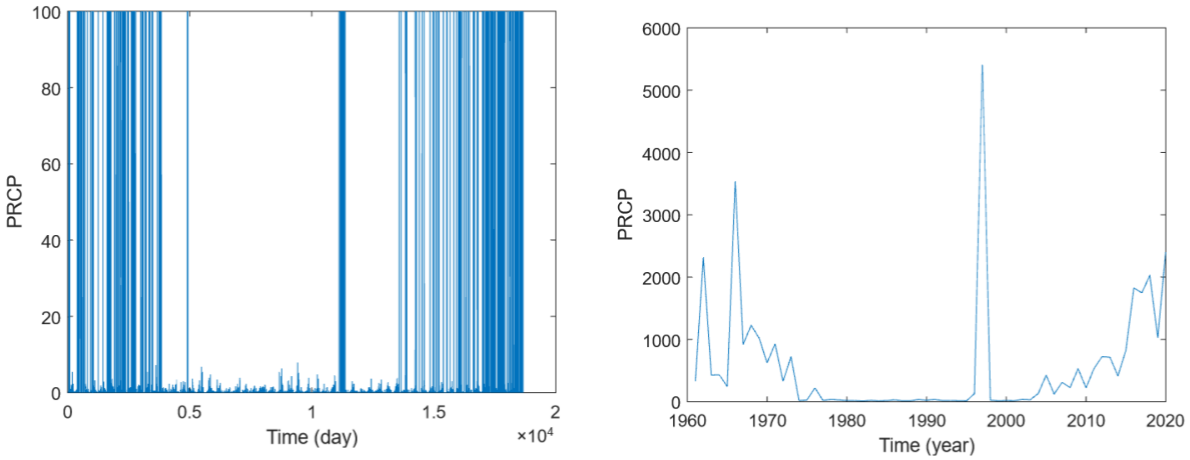
Symbol	Description	Unit
T_y	Time	years
T_d	Time	days
Y_n	Annual precipitation of n year	mm
D_n	Daily precipitation of day n	mm
η	eta factor	
r	Pearson coefficient	
r_s	Spearman factor	
$loss$	Loss function	
a, b	Fitting parameters or neural network parameters	
ω	Neural networks, hyperplane parameters	

Table 1 Notations

5. Modeling and solving of Problem 1

5.1 Correlation test

First of all, according to the data given in the question, the daily precipitation distribution in recent years shows a cyclical distribution, and the precipitation is higher in the early years, with the exception of 1991, when it was lower between 1970 and 2000.



(a) Daily precipitation

(b) Annual precipitation

5.1.1 Analysis of the amount of daily variation

Considering the more extreme distribution of daily precipitation, the distribution is mainly below 10mm or above 100mm. The D_n corresponds to the type of data for the fixed class of data, i.e. with and without precipitation. The other set of data yearsT_y is at a fixed distance, and for such data, the Eta coefficient η needs to be used to determine the correlation between the two.

$$\eta = \sqrt{\frac{\sum(D_n - \overline{D_n})^2 - \sum(D_n - \overline{D_n^k})^2}{\sum(D_n - \overline{D_n})^2}} \tag{1}$$

While $D_n$means the date,$\overline{D_n}$means the average. While the $D_n k$means average of dates with high rainfall.

Using the above formula, Using the Matlab code in Appendix 1,the calculated value is $\eta =$ 0.0422984,which can be seen that the correlation between rainfall and daily variation is very weak at the macro level, and it is not meaningful to fit the curve based on daily variation alone, so no further description will be given here.

5.1.2 Analysis of the annual volume of change

Based on the plotting of annual variation, it can be judged that annual precipitation is a constant distance variable, so the Pearson correlation coefficient needs to be used to test the two for analysis At the same time, considering the characteristics of the image, as precipitation decreases first and then increases relative to the year, there may be a secondary correlation, so the Pearson and Spearman correlation coefficient is used to test, the Spearman coefficient needs to first sort the data of year and rainfallT_y, D_n, so the correlation between the two is analysed according to the rank order of the sort.

$$r_s = [\frac{1}{n} r_i R_i - (\frac{n+1}{2})^2]/(\frac{n^2-1}{12}) \tag{2}$$

The Pearson coefficient determines whether the two indicators $T_y D_n$ are on a line and both reflect a linear relationship, so the square of year T_y and rainfall D_a is constructed as a statistic.

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{3}$$

Using matlab's corr toolkit, a value of approximately 0.8587 can be derived for the Pearson correlation coefficient, which is significantly correlated, so a quadratic fit can be used to analyse the annual variability characteristics of precipitation.

5.2 Fitting

With regard to fitting the curve, the first step is to use the scatter plot, while constructing the corresponding set of normal equations based on the class of fitting functions derived from the correlation analysis, so as to solve for the specific fitting function. The Cueve Fitting Toolkit for Matlab is used here, and the results are obtained by correlation analysis, after several comparisons, analysis and judgement of the R-squar values, and finally by selecting the Fourier fitting model that has a fitting effect and can reflect the trend of the function.

The function resulting from the use of Fourier fitting is as follows.

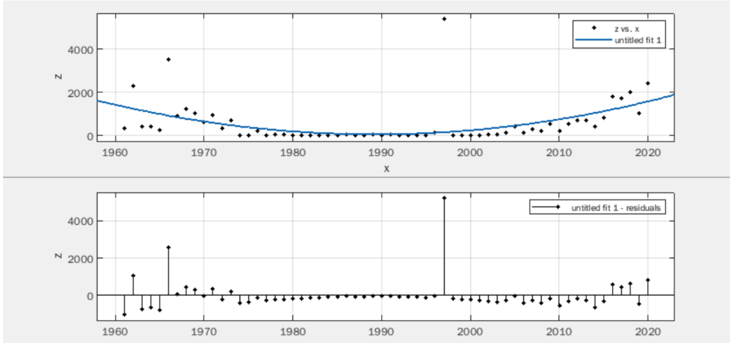


Figure 1 Fourier Fitting

$$f(x) = a_0 + a_1 cos(wx) + b_1 sin(wx) \tag{4}$$

$$a_0 = 1.453e + 09(-1.709e + 16, 1.709e + 16) \tag{5}$$

$$a_1 = -1.446e + 09(-1.709e + 16, 1.709e + 16) \tag{6}$$

$$b_1 = 1.358e + 08(-8.016e + 14, 8.016e + 14) \tag{7}$$

$$w = -4.708e - 05(-277, 277) \tag{8}$$

Goodness of fit:
SSE: 4.28e+07
R-square: 0.2115
Adjusted R-square: 0.1693
RMSE: 874.2

5.2.1 M-K test to determine mutations

In order to quantify the sudden changes in precipitation, the Manner-Kendal To quantify the abrupt changes in precipitation, the Manner-Kendal analysis was used to take the annual precipitation $Y_1, Y_2, Y_3 ... Y_n$ time series and successively find the positive and negative sequential outcome statistics.

$$S_k = \sum_{i=1}^{i=n} R_i \tag{9}$$

The R_i denotes the cumulative count of Y_i greater than Y_j (i<j<n).

Define the statistic UF_k in the case of random independence of time series

$$UF_k = \frac{s_k - E(s-k)}{\sqrt{VAR(s_k)}} \tag{10}$$

where VRK calculates the variance and mean, and UF1=0 when k=0, while defining $UF_k = -UB_k$.With confidence interval of 95

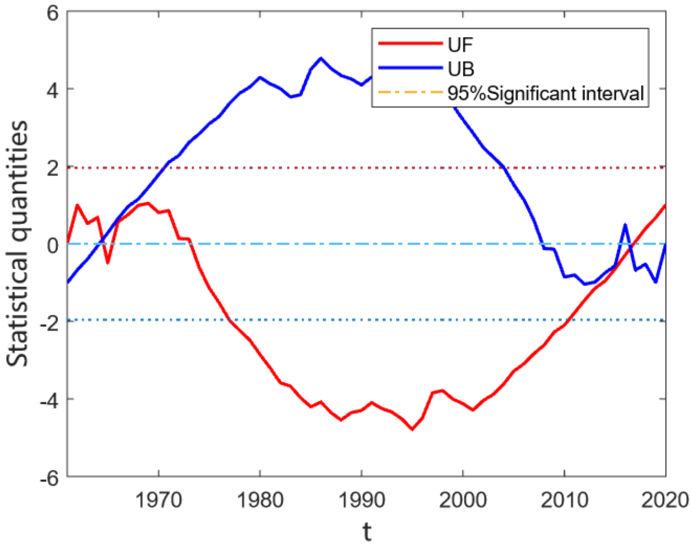


Figure 2 Manner-Kendall Features

From the nature of the MK statistic and the image, it is clear that precipitation around 1968 and around 2015 consists of a stronger sudden change.

5.3 Specific analysis of the Zhengzhou inundation

5.3.1 Selecting

For the years with high precipitation, the years with annual precipitation above 1000mm were selected from Figure 1, and the data for the years thus determined are shown in the table 2.

5.3.2 Overall assessment of the current round of rainfall

This problem requires a line intercept of the data given in the Annex, taking a total of one month around the outbreak date of the storm, 20 July. As the weights of each indicator need to be considered, an n-element linear fit was used to roughly analyse each indicator, using precipitation as the dependent variable.

1963	1967	1968	1969	1970
1971	1972	1974	1998	2010
2012	2013	2014	2016	2017
2018	2019	2020	2021	

Table 2 Years with high precipitation

Due to the inconsistent magnitudes of the indicators, normalisation was required to convert the magnituded data into dimensionless data.

$$x^* = \frac{x - \min}{\max - min} \tag{11}$$

where max and min represent the maximum and minimum values of this set of indicators for this scale, respectively. x^* is the result without the scale after processing, and x is the indicator with the scale before processing.

Using the normalized measure, regression analysis is performed using Matlab's regression function to construct the set of coefficients with the smallest residual values, using the coefficients to represent the impact weights of each indicator (code can be found in the appendix), by constructing a 1x12 matrix of coefficients ω, with all initial values filled to 1, and later using the gradient descent method to calculate the function of each indicator under this coefficient lift values to construct the loss function.

$$loss = \omega x - y \tag{12}$$

where x denotes the 33x12 matrix of independent variables and y is the 33x1 matrix of dependent variables, back propagation is performed and the learning rate $\alpha = 0.01$ is chosen, the resulting gradient value combined with the learning rate will ω be iteratively updated.

The resulting weights are shown in the table.

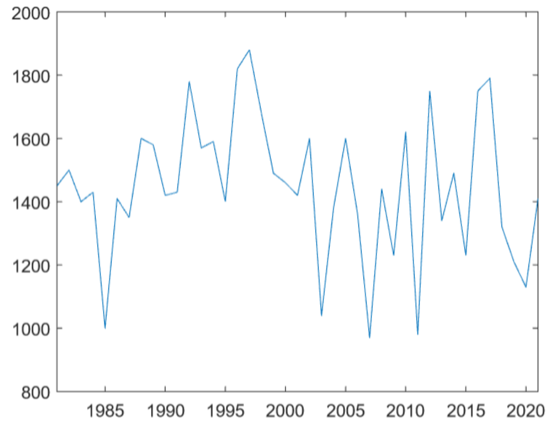
DEWP	FRSHIT	GUST	MAX	MIN	WDSP
0.66355	0.28373	0.83712	0.38715	0.428734	0.9189
MXSPD	SNDP	STP	TEMP	VISB	
0.83753	0.03463	0.684562	0.38421	0.375628	

Table 3 Weighting of indicators affecting this Zhengzhou rainstorm

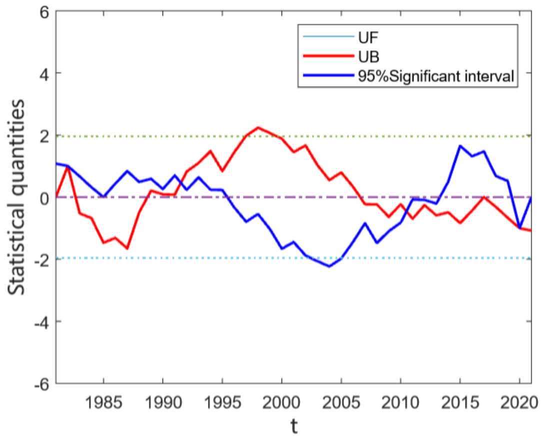
6. Modeling and solving of Problem 2

Question 2 was answered in the same way as question 1, and without going into too much detail here, we analysed three cities, Changsha, Taiyuan and Guangzhou (data source: http://data.cma.cn/), with the same algorithmic process and code as in the analysis of the previous question.

6.1 Changsha



(a) Annual precipitation



(b) MK-Features

Changsha has generally high precipitation, with inflection points in precipitation mainly occurring around 1983, 1993, 2009 and 2020, which is more in line with the Fourier fit and has a strong cyclical nature.

General model Fourier1:

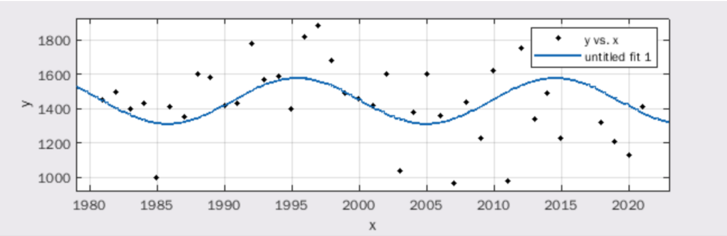


Figure 3 Fourier Fitting

f(x) = a0 + a1*cos(x*w) + b1*sin(x*w)

Coefficients (with 95% confidence bounds):

a0 = 1449 (1380, 1517)

a1 = 98.18 (-1.014e+04, 1.034e+04)

b1 = 89.28 (-1.118e+04, 1.136e+04)

w = 0.3278 (0.2705, 0.3852)

Goodness of fit:

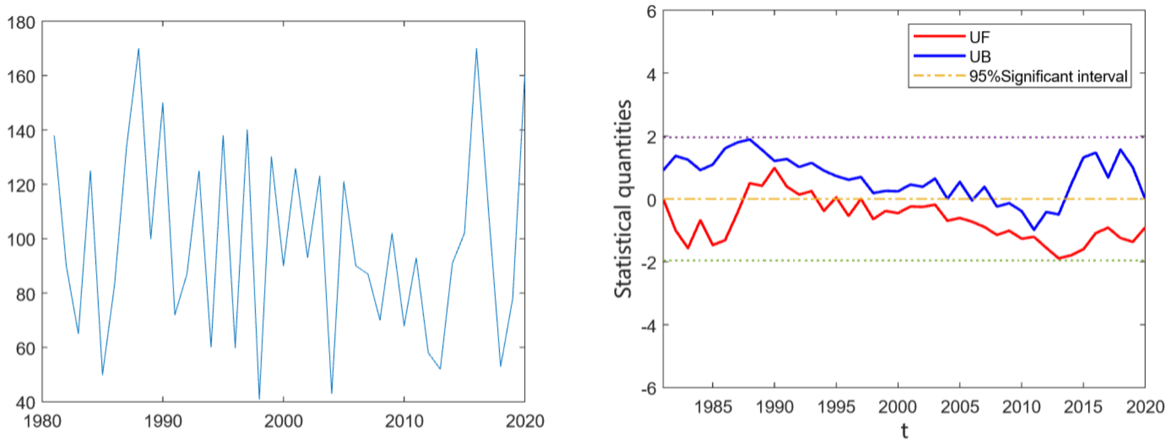
SSE: 1.731e+06

R-square: 0.1653

Adjusted R-square: 0.09758

RMSE: 216.3

6.2 Taiyuan



(a) Annual precipitation (b) MK-Features

Compared to Changsha, Taiyuan's precipitation was significantly lower, but the Fourier fit shows that the frequency of cycles was extremely high and the inflection point of rainfall was not reflected in the MK analysis, so unpredictable weather may have been the main cause of the flood casualties in Taiyuan.

General model Fourier1:

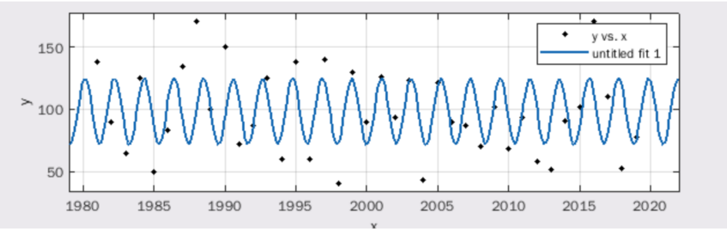


Figure 4 Fourier Fitting

f(x) = a0 + a1*cos(x*w) + b1*sin(x*w)

Coefficients (with 95% confidence bounds):
a0 = 98.47 (88.11, 108.8)
a1 = 26.32 (-33.47, 86.1)
b1 = -0.5985 (-2359, 2358)
w = 3.005 (2.96, 3.05)

Goodness of fit:
SSE: 3.756e+04
R-square: 0.2424
Adjusted R-square: 0.1793
RMSE: 32.3

7. Modeling and solving of Problem 3

Question 3 requires forecasting future precipitation using different methods and analysing the effects of the forecasts. The available data needs to be divided into a training group and a test group, and in order to ensure the timeliness of the forecasts, the data from the previous 10% of years is therefore chosen as the test material. The different methods are described separately here.

7.1 Data pre-processing

In order to standardise the scale, the data first needs to be normalised with dimensionality reduction using principal component analysis.

After normalisation and dimensionality reduction, from 12 dimensioned indicators to 9 dimensionless indicators,which was stored in Appendix named$'result_PCA.csv'$

7.2 Neural Networks

7.2.1 CNN

The neural network algorithm uses a neural network architecture with 9X1 implicit layers, a learning rate of 0.0001 and an epoch of 10000 to train the training group due to the large number of independent variables.

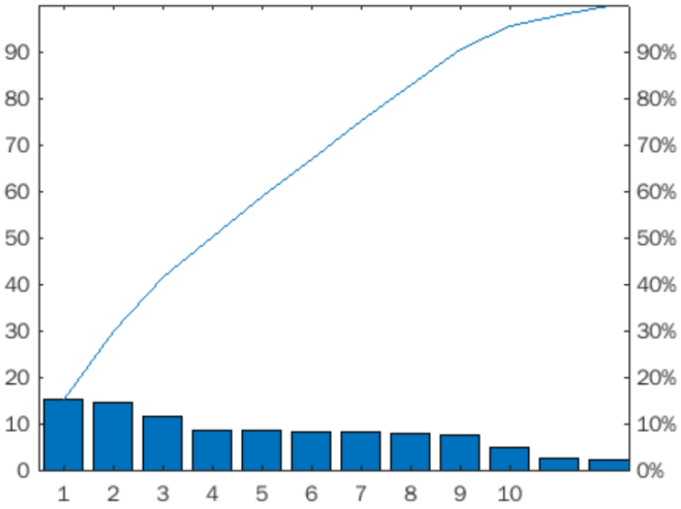


Figure 5 PCA rasult

Apparently, the training results were not satisfactory, mainly due to some deviations in the selection of hyperparameters and neural network models. The gradient value was kept around 0.04, which could not achieve the desired effect. After analysis of the test set, the Performance of the loss function was 157.5714.R=0.2936.

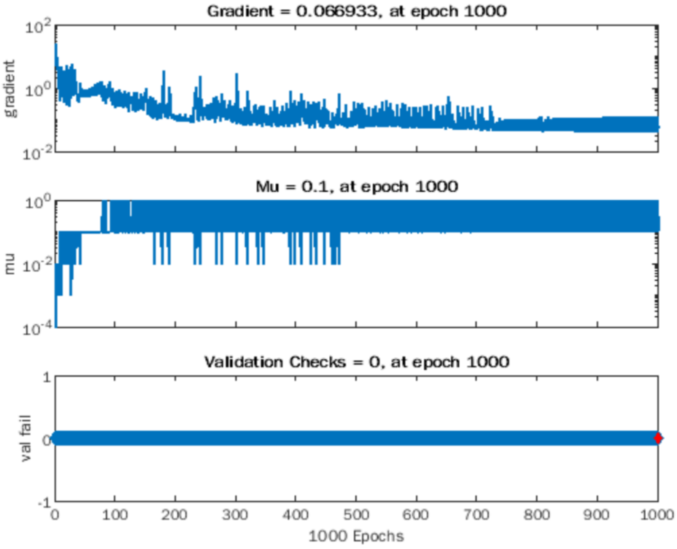


Figure 6 network result

7.3 Plain Bayesian classification

In order to predict the weather based on meteorological indicators, if we only need to determine whether there is very heavy rainfall weather, we can use 0-1 to indicate the presence or

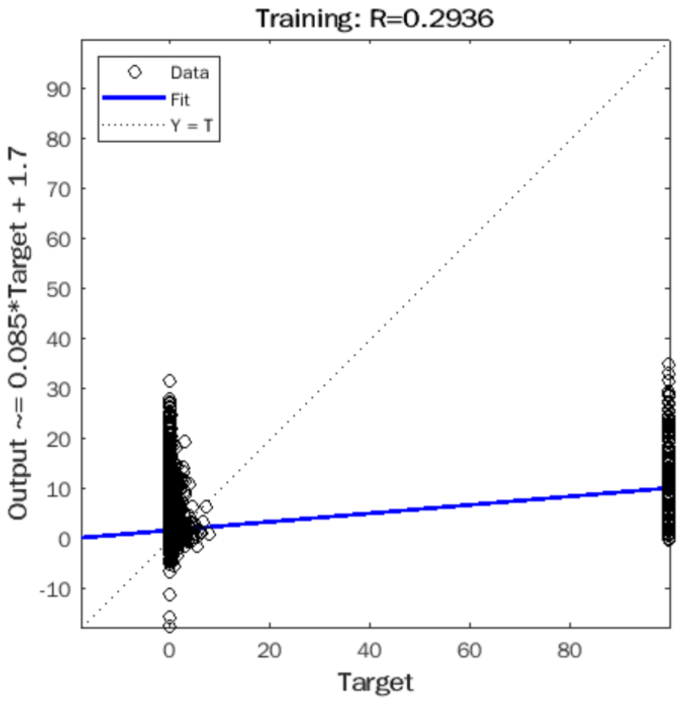


Figure 7 Correlation coefficient

absence of heavy rainfall, probability instead of value, as the indicator we want to predict, to determine the probability of rain, and plain Bayesian classification to assign unknown meteorological samples to two classes of whether it is raining or not. Using the full probability formula.

$$P(C/X) = \frac{P(X/C)P(C)}{P(X)} \tag{13}$$

It is also assumed that each categorical attribute follows a normal distribution, using

$$P(X/C)P(C) \geq P(X/C)P(C) \tag{14}$$

It follows that X is assigned to its largest class C.

The partial prediction results obtained are shown in the figure, the error is within the allowed range and the effect is due to the neural network.

Where blue dots are actual data and red dots are predicted data

7.3.1 LSTM nerual network

To optimise the neural network, an LSTM neural network architecture was used, utilising 96*3 hidden layers with a gradient threshold set to 1. An initial learning rate of 0.005 was specified, which was reduced after 125 training rounds by multiplying by a factor of 0.2. The predictions were denormalised using the previously calculated parameters. Comparing the real data with

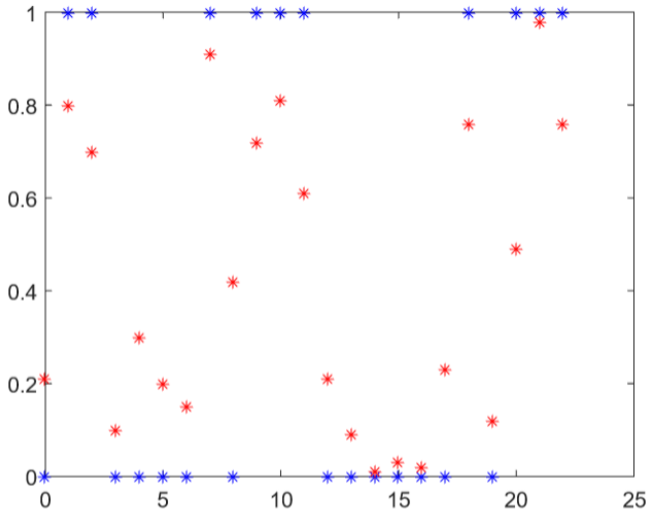
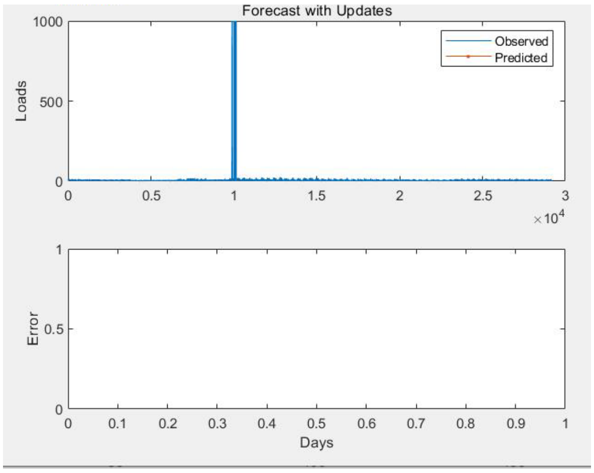
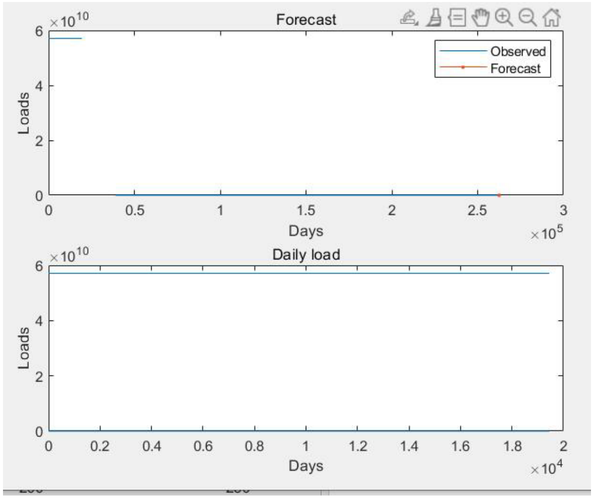


Figure 8 Bayesian prediction results

the predicted data, the predictions were significantly optimised. It can be seen that the neural network is significantly optimized under LSTM prediction.



(a) Annual precipitation



(b) MK-Features

7.4 Support vector machines

On the basis of plain Bayes, if the presence or absence of heavy precipitation is treated as a 0-1 variable, the problem can be transformed into a classification problem by using SVM classification ,Use the original indicator of i as x_i,and selecting $\omega$and$b$ on a 9-dimensional Euclidean space as the hyperplane, thus achieving a classification effect.Each sample will be divided into D_0 and D_1 categories based on whether or not rainfall has occurred

$$D_0 = \omega x_i + b > 0 \tag{15}$$

$$D_1 = \omega x_i + b < 0 \tag{16}$$

This leads to the construction of a hyperplane, and in order to make this hyperplane more robust, the best hyperplane needs to be found autonomously. The task at hand lies in the hyperplane that separates the two types of samples at the maximum interval, also known as the maximum interval hyperplane.

$$\begin{cases} \frac{\omega^T x+b}{||\omega||} \geq 1D_1 \\ \frac{\omega^T x+b}{||\omega||} \leq 1D_0 \end{cases})$$

The objective is to maximise the distance from the support vector $y(\omega^T x + b) = 1$, to solving max $\frac{2}{||\omega||}$ and then solving the inverse to convert it to a maximum value, combining the above analysis to solve the following linear programming problem by function linear regression in Matlab.

$$\max \frac{2}{||\omega||} s.t. y_i(\omega^T x_i + b) \geq 1 \tag{17}$$

The classification results are as follows Blue indicates the projection of the hyperplane in a given

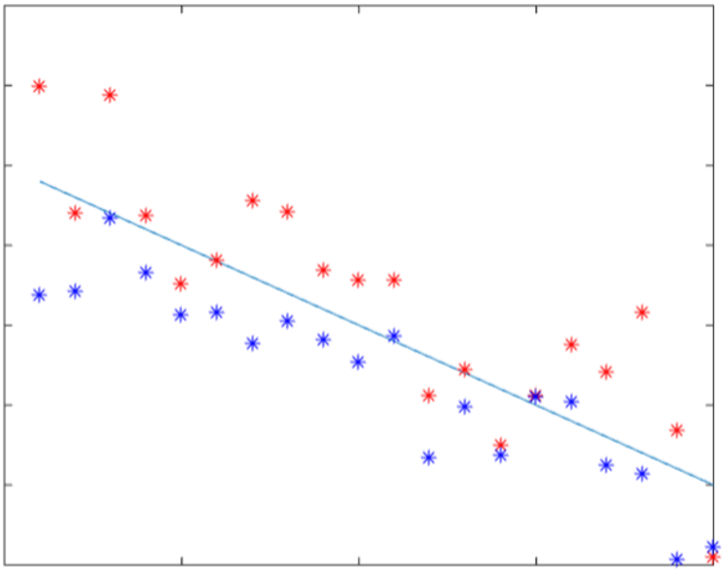


Figure 9 CVM classification

dimension, with red dots indicating points with more precipitation and blue dots indicating points with less precipitation.

8. Modeling and solving of Problem 4

For question four, correlation data and multivariate linear fit data were used to find the correlation and weight of each indicator with rainfall and to analyse the main factors that mainly affect this rainfall.

8.0.1 Taiyuan Rain on Oct.

Select data from the week before and after the rainfall to analyse correlation and indicator impact. The indicator with the highest correlation was P. The relevant weights for linear regression are as follows

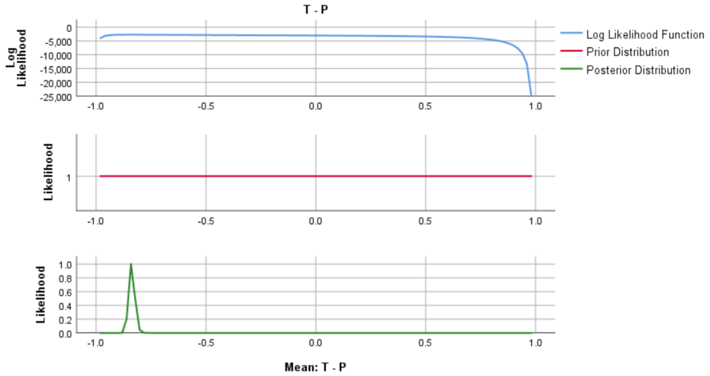


Figure 10 P, T correlation

P0	P	U	Ff	ff10	Td
0.45234	0.87318	0.82712	0.31812	0.0000	0.42132

Table 4 Weighting of indicators affecting this Zhengzhou rainstorm

The analysis of the China Meteorological Administration (CMA) indicates that there are three main factors leading to the current round of rainfall in Shanxi: firstly, the stable atmospheric circulation situation. The unusually strong western Pacific subtropical high pressure first stretched westward and northward and then steadily maintained in the Yellow and Huai regions, forming a stable east-high and west-low circulation situation in Shanxi with the low value system of the westerly wind belt, which is conducive to prolonged precipitation weather in Shanxi. This is consistent with our analysis that the indicator U has a large weight and correlation. Secondly, water vapour conditions are abundant. The southerly airflow and low-level southwestern rapids from the western side of the subtropical high pressure transport water vapour from the South

China Sea and the Bay of Bengal northwards through the southwest to the south-central region of Shanxi, providing an abundant source of water vapour for sustained precipitation in Shanxi, which is closely related to the influence of the weight p. Finally, low-level lifting conditions were maintained for a long time. The prolonged maintenance of the low-level shear convergence system and the repeated passage of precipitation echoes through central Shanxi under stable weather conditions, superimposed on the effect of the complex topography of the Luliang Mountains and Taihang Mountains in Shanxi on the precipitation increase of the easterly airflow, lead to the occurrence of extreme heavy precipitation in central Shanxi and the northern part of Linfen.

8.1 Zhengzhou Rain on Oct.

Compared to the precipitation in Taiyuan, the higher precipitation indicator for October in Zhengzhou was also Tx.

In addition to the factors considered in this question, according to the news media reports,

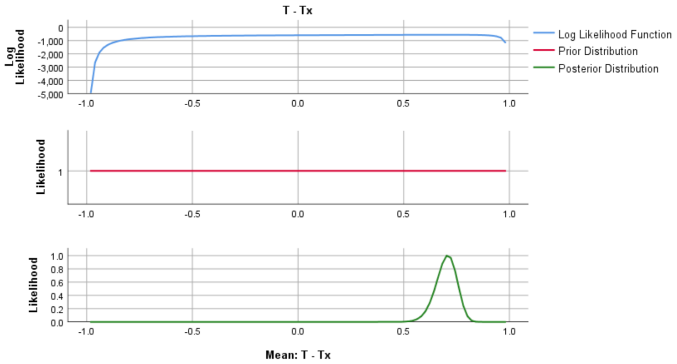


Figure 11 P, Tx correlation

"Shanxi has been fighting drought in previous years, but as a result, I didn't expect to encounter floods this year", this is what an affected woman said during the press interview, which may also be a reason for the delay in reporting the floods in Shanxi. The scope of the floods in Shanxi and Zhengzhou are also different. The floods in Shanxi occurred mostly in rural areas, while the storm in Zhengzhou occurred in urban areas, and these factors are the main reasons for the difference in impact and attention between the two places.

Regarding the rainfall in Zhengzhou in July and October, combining the analysis here with the analysis in the first question, the wind direction in July is different from that in October, and the intensity of convection is also different. It is thus inferred that the rainfall in July is probably due to typhoons, while the rainfall in October is frontal rain, the inevitable result of cold air moving south.

9. Problem 5

Cities need to judge the precipitation cycle based on annual precipitation, the annual variation in precipitation characteristics, and use the indicators affecting precipitation to make timely predictive analyses of extreme weather, especially in cities with short precipitation cycles, so that droughts in a given period do not neglect the possibility of flooding at any time. It is also necessary to strengthen urban and rural areas, both to improve drainage systems in cities and to address the poor infrastructure set up in the countryside to deal with floods.

10. Evaluation, improvement and promotion of models

There are many models used in this question and a brief explanation of the advantages and disadvantages of each model is given.

10.1 model strengths

1. Both annual and daily precipitation are analysed and evaluated together to help select the best evaluation and prediction model
2. Each model is optimised in a targeted way, taking into account the cyclical nature of precipitation and its sudden variability.
3. Data is utilised and segmented appropriately to ensure maximum data utilisation.

10.2 model drawbacks

1. Failure to account for human factors, extremes, etc. outside of the data
2. Neural networks, linear regression and other algorithmic tools are difficult to handle due to the large amount of data and the long computation time, and there is room for optimization.

10.3 improvements to the model

The usability of the model can be improved by practical activities such as distributing questionnaires to residents of various cities, and also by collecting some human factors other than meteorology and hydrology to improve the accuracy of the model. Improve algorithm efficiency and reduce time complexity in order to process more data.

11. References

[1]http://data.cma.cn/

[2]H. Chen, C. Wang, T. Chen and X. Zhao, "Feature selecting based on fourier series fitting," 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), 2017, pp. 241-244, doi: 10.1109/ICSESS.2017.8342905.

[3]H. Mosallaei and K. Sarabandi, "Periodic meta-material structures in electromagnetics: concept, analysis, and applications," IEEE Antennas and Propagation Society International Symposium (IEEE Cat. No.02CH37313), 2002, pp. 380-383 vol.2, doi: 10.1109/APS.2002.1016103.

[4]M. V. J. Reddy and B. Kavitha, "Neural Networks for Prediction of Loan Default Using Attribute Relevance Analysis," 2010 International Conference on Signal Acquisition and Processing, 2010, pp. 274-277, doi: 10.1109/ICSAP.2010.10

[5]A. Novikov, A. Levykin and E. Novikov, "(m, k)-Methods for Control Theory Problems," 2019 15th International Asian School-Seminar Optimization Problems of Complex Systems (OPCS), 2019, pp. 120-124, doi: 10.1109/OPCS.2019.8880174.

[6]A. Akandeh and F. M. Salem, $''SlimLSTMNETWORKS : LSTM_{6}andLSTM_{C}6,''$ 2019 IEEE 62nd International Midwest Symposium on Circuits and Systems (MWSCAS), 2019, pp. 630-633, doi: 10.1109/MWSCAS.2019.8884912.

[7]Y. Yang, J. Wang and Y. Yang, "Improving SVM classifier with prior knowledge in microcalcification detection1," 2012 19th IEEE International Conference on Image Processing, 2012, pp. 2837-2840, doi: 10.1109/ICIP.2012.6467490.

[8]I. Ntzoufras, V. Palaskas and S. Drikos, "Bayesian models for prediction of the set-difference in volleyball," in IMA Journal of Management Mathematics, vol. 32, no. 4, pp. 491-518, Aug. 2021, doi: 10.1093/imaman/dpab007.

12. Appendix

12.1 Appendix one

```
clc;
close all;
y = xlsread(’Attachment 1Data of three meteorological stations in
    zhengzhou annex.xlsx’,1,’I1:I18666’);
years = floor(18665/310);
i = 0;
z = zeros(1,years);
while (i<years)
    i = i+1;
    j = 310;
    while(j>0)
        m = (i-1)*310+j;
        z(1,i) = z(1,i)+y(m,1);
        j = j-1;
    end
end
x = 1961:1961+floor(18665/310)-1;
plot(x,z)
xlabel(’Time (year)’);
ylabel(’PRCP’);
r1=corr(x,y,’type’,’pearson’);
r2=corrcoef(x,y);
%------------------------------------------------------------------%
y = xlsread(’Attachment 1Data of three meteorological stations in
    zhengzhou annex.xlsx’,1,’I1:I18666’);
y = y’;
x = 1:18665;
plot(x,y)
xlabel(’Time (day)’);
ylabel(’PRCP’);

%%eta calculating
%%shift
i = 1;
n = 0;
```

```
m = 0;
while i<size(y)
    if y(i) > 90
        y(i) = 1;
        n = n+x(i);
        m = m+1;
    else
        y(i) = 0;
    end
    i = i+1;
end
%conbine
z = [x’ y’];
%calculate
a = 18065/2;
b = 0;
c = 0;
n = n/m;
i = 18665;
while i<size(y)
    b = b+(x(i)-a)^2;
    c = c+((y(i)*x(i))-n)^2;
end
eta = sqrt(1-c/b);
%%------------------------------------------------------------------%%
%%Fourier fitting
function [fitresult, gof] = createFit(x, z)
%CREATEFIT(X,Z)
%  Create a fit.
%
%  Data for ’untitled fit 1’ fit:
%      X Input : x
%      Y Output: z
%  Output:
%      fitresult : a fit object representing the fit.
%      gof : structure with goodness-of fit info.
%
%  See also FIT, CFIT, SFIT.
%  Auto-generated by MATLAB on 11-Nov-2021 11:54:22


%% Fit: 'untitled fit 1'.
[xData, yData] = prepareCurveData( x, z );

% Set up fittype and options.
ft = fittype( 'fourier1' );
opts = fitoptions( 'Method', 'NonlinearLeastSquares' );
opts.Display = 'Off';
opts.StartPoint = [0 0 0 0.106494666223383];

% Fit model to data.
[fitresult, gof] = fit( xData, yData, ft, opts );

% Create a figure for the plots.
figure( 'Name', 'untitled fit 1' );

% Plot fit with data.
subplot( 2, 1, 1 );
h = plot( fitresult, xData, yData );
legend( h, 'z vs. x', 'untitled fit 1', 'Location', 'NorthEast',
    'Interpreter', 'none' );
% Label axes
xlabel( 'x', 'Interpreter', 'none' );
ylabel( 'z', 'Interpreter', 'none' );
grid on

% Plot residuals.
subplot( 2, 1, 2 );
h = plot( fitresult, xData, yData, 'residuals' );
legend( h, 'untitled fit 1 - residuals', 'Zero Line', 'Location',
    'NorthEast', 'Interpreter', 'none' );
% Label axes
xlabel( 'x', 'Interpreter', 'none' );
ylabel( 'z', 'Interpreter', 'none' );
grid on
%%--------------------------------------------------------------------------------------
m = csvread('Annex1-Intercepted-data.csv', 1, 2);
```

```
i = 1;
j = 1;
while (i<=12)
    n = m(:,i);
    n = (n-min(n))/(max(n)-min(n));
    m(:,i) = n;
    i = i+1;
end
%%--------------------------------------------------------------------------------------
[b1,bint1,r1,rint1,stats1]=regress(y1,X);
%b1 is the coefficient of the multiple regression equation
b1,bint1,stats1
% plot the residuals
rcoplot(r1,rint1)

% predict and plot
z=b1(1)+b1(2)*x(:,1)+b1(3)*x(:,2)+b1(4)*x(:,3)+b1(5)*x(:,4)+b1(6)*x(:,5)+b1(7)*x(:,6)+b1
plot(X,y1, 'k+',X,z, 'g')
%%
% second time with SPREAD as the explanatory variable

[b2,bint2,r2,rint2,stats2]=regress(y2,X);
%b1 is the coefficient of the multiple regression equation
b2,bint2,stats2
% plot the residuals
rcoplot(r2,rint2)

z=b2(1)+b2(2)*x(:,1)+b2(3)*x(:,2)+b2(4)*x(:,3)+b2(5)*x(:,4)+b2(6)*x(:,5)+b2(7)*x(:,6)+b2
plot(X,y1, 'r+',X,z, 'o')
```

12.2 Appendix two

```
    % M-K test matlab code

    function [UFk,UBk] = mk(y)

    %y:original time series, could be temperature, precipitation, etc. Note
        the substitution!
```

```
%UFk:sequential results;UBk:inverse sequential results

n=length(y);

% Positive series calculation ---------------------------------

r=zeros(n,1);
%loop up
for i=2:n

    for j=1:i

        if y(i)>y(j)

            r(i)=r(i)+1;

        end

    end

end

sk=zeros(n,1);

for i=2:n

    for j=1:i

        sk(i)=sk(i)+r(j);

    end

end

UFk=zeros(n,1);
% define statistic UFk=(sk-E)/sqrt(var)

for i=2:n
```

```
    E=i*(i-1)/4;
    % Mean of Sk(i)

    Var=i*(i-1)*(2*i+5)/72;
    % variance of Sk(i)

    UFk(i)=(sk(i)-E)/sqrt(Var);

end

% End of positive sequence calculation ---------------------------------
% Inverse sequence calculation ---------------------------------

y2=flipud(y);

r2=zeros(n,1);
%double Loop
for i=2:n

    for j=1:i

        if y2(i)>y2(j)

            r2(i)=r2(i)+1;

        end

    end

end

sk2=zeros(n,1);
%Loop up
for i=2:n

    for j=1:i

        sk2(i)=sk2(i)+r2(j);
    end

end

UFk2=zeros(n,1);
%Loop up
for i=2:n

    E=i*(i-1)/4; % mean of Sk(i)

    Var=i*(i-1)*(2*i+5)/72; % variance of Sk(i)

    UFk2(i)=(sk2(i)-E)/sqrt(Var);

end
%define ubk
UBk=zeros(n,1);

UBk=flipud(-UFk2);

% End of inverse sequence calculation ---------------------------------



% Plot

plot(1961:1960+n,UFk,'r-','linewidth',1.5);

hold on;

plot(1961:1960+n,UBk,'b-','linewidth',1.5);

plot(1961:1960+n,0*ones(n,1),'-.' ,'linewidth',1); %plot horizontal 0-value line

plot(1961:1960+n,1.96*ones(n,1),':','linewidth',1); %plot 95% confidence interval bounds
plot(1961:1960+n,-1.96*ones(n,1),':','linewidth',1);%plot 95%
    confidence interval bounds

axis([1961,1960+n,-6,6]);%set the horizontal and vertical ranges

legend('UF','UB','95%Significant interval'); %set legend

xlabel('t','FontName','TimesNewRoman','FontSize',12);%set x-axis label

ylabel('Statistical
    quantities','FontName','TimesNewRoman','Fontsize',12);%set y-axis
    label

%plot(1:n,2.32*ones(n,1),':','linewidth',1); %plot 99% confidence
    interval bounds

%plot(1:n,-2.32*ones(n,1),':','linewidth',1);%plot 99% confidence
    interval bounds

% absolute value of MK test statistic 1.28/1.96/2.32 indicates
    90%/95%/99% confidence interval respectively



% Write the results of UF and UB calculations to an xlsx file:
    C:\test.xls

% Target form: Sheet1

% Target area: UFk from A1, UBk from B1

xlswrite('111.xlsx',UFk,'Sheet2','A1');

xlswrite('111.xlsx',UBk,'Sheet2','B1');
```

12.3 Appendix three

```
X1=xlsread('Attachment 1: Data of three meteorological stations in
    zhengzhou annex.xlsx',1,'c1:h18666');
X2=xlsread('Attachment 1: Data of three meteorological stations in
    zhengzhou annex.xlsx',1,'j1:o18666');
X = [X1,X2];
M = cov(z); % covariance
z=zscore(X); %data normalisation
[V,D]=eig(M); %find the eigenvectors, eigenroots of the covariance
    matrix
eig1=sort(d,'descend'); % sort the contributions by largest to smallest
    element
d=diag(D); %take out the eigenroot matrix column vectors (extract the
    contribution of each principal component)
v=fliplr(V); % rearrange the feature vectors according to D
i=0;
S=0;
while S/sum(eig1)<0.85
    i=i+1;
    S=S+eig1(i);
end % To find out the principal components with cumulative contribution
    greater than 85%
NEW=z*v(:,1:i);
%output the data under the new coordinates generated
W=100*eig1/sum(eig1);
figure(1);
pareto(W); %draw a histogram of the contribution
%%---------------------------------------------------------------------------------
%%------------------------------------------CNN------------------------------------
%%Function CNN

PCA;

P=NEW;

T=xlsread('Attachment 1Data of three meteorological stations in
    zhengzhou annex.xlsx',1,'I1:I18666');
P = P';
T = T';
```

```
net = newff(minmax(P),[9,1],{'tansig','purelin'},'trainlm');
net.trainParam.lr=0.00001;
net.trainParam.show=50;%
net.trainParam.lr=0.00001;
net.trainParam.goal=1e-5;
net.trainParam.epochs=1000;
[net,tr]=train(net,P,T);
net.b{1}%Hidden layer threshold
net.iw{1,1}%Hidden layer weights
net.b{2}%Output layer threshold
net.lw{2,1}%output layer weights
P2=[-1;2];
y3=sim(net,P2);
%%-----------------------------------------------------------------------------------
%%---------------------------------SVM----------------------------------------------
% Data set load
load flow_flow input output input_test output_test
train=input;
train_label=output;
test_label=output_test.
test=input_test;

[X_scale,temp] = scaleForSVM(train,test,1,2).
X_scale


tic
[bestmse,bestc,bestg] = SVMcgForRegress(train_label,train) %%
    regression search
toc
cmd = ['-c ',num2str(bestc),' -g ',num2str(bestg),' -s 3 -p 0.01'] .

% -c cost: Set the parameters (loss functions) for C-SVC, e -SVR and
    v-SVR (default 1).

% -g r(gama): Set gamma function in kernel function (for
    polynomial/rbf/sigmoid kernel functions) (default 1/k).

% -s Set the svm type.
```

```
    % 0 - C-SVC
    % 1 - v-SVC
    % 2 - single-class-SVM
    % 3 - -SVR
    % 4 - n - SVR

    % -p p: sets the value of the loss function p in e -SVR (default 0.1).

    %% SVM network training
    model = svmtrain(train_label, train, cmd);

    %% SVM network prediction
    tempzero = zeros(92,1) ;
    [predict_label] = svmpredict(tempzero, test, model)

    % Result Analysis
    Figure (1)
    plot(predict_label,'r*:')
    Keep
    plot(test_label,'bo--')
    xlabel('Point in time')
    ylabel('Precipitation')
    title('predict_value', 'Precipitation',12)
    legend({'predicted value', 'actual value'})
%%--------------------------------------------------------------------------------------
%%-----------------------------------------Bayes----------------------------------------
%%-------Function-attribute-------------%%
function y=attribute(X,n)
% function extracts a column of values corresponding to the nth attribute
    of the original dataset X
[M,N]=size(X);
for i=1:M
    temp{i}=X{i}{n}; %Save the specified column value as temp
end
y=temp';%transfer
%%-----Function-post_prob--------------%%
function [post_pro,post_name]=post_prob(E,D)
%E is the target attribute, D is the decision attribute, post_pro
    calculates the posterior probability of the target attribute
    corresponding to the decision attribute
%post_name is the name of the requested posterior probability variable
[M,N]=size(D);
decision=unique(D);%decision attribute type
attri=unique(E); %condition attribute type
[m1,n1]=size(decision);
[m2,n2]=size(attri);
temp=cat(2,E,D); %connect conditional and decision attributes
post_pro=zeros(m1,m2); % posterior probability initialization
for i=1:M
    for j=1:m2
        for k=1:m1
            post_name{k,j}=cat(2,{attri{j}},{decision{k}});
            if(isequal(temp(i,:),post_name{k,j}))
                post_pro(k,j)=post_pro(k,j)+1; % posterior probability of
                    conditional attributes (frequency)
            end
        end
    end
end
for i=1:m1
    post_pro(i,:)=post_pro(i,:)/sum(post_pro(i,:));% find the posterior
        probability of the conditional attribute
end
%%------Function-probality----------%%
function y=probality(E)
[M,N]=size(E);
class=unique(E);% find the class of the decision attribute
[m,n]=size(class);
p=zeros(m,1);%initialize the prior probability p
for i=1:M
    for j=1:m
        if(isequal(E{i},class{j}))
            p(j)=p(j)+1; %find the prior probability (frequency) of each
                sample
        end
    end
end
y=p/M;% get the probability of each sample
```

```
%%------Function-main-------------------%%
function out=my_bayes(X,Y)

x =csvread('result.csv',50,0);
y =xlsread('Attachment 1: Data of three meteorological stations in
    zhengzhou annex.xlsx',1,'I52:I18666');
X = [x,y];

Y = csvread('result.csv',0, 0, [0,0,50,9]);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% training section
[M,N]=size(X);
[m,n]=size(X{1});
decision=attribute(X,n); % extract decision attribute
Pro=probality(decision);% calculate the probability of individual
    components of the decision attribute
for i=1:n-1
    [post_pro{i},post_name{i}]=post_prob(attribute(X,i),decision); % find
        the posterior probability of each conditional attribute
end
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% prediction part
uniq_decis=unique(decision); % find the category of the decision attribute
P_X=ones(size(uniq_decis,1),1); %initialize the posterior probability of
    the decision attribute
[M,N]=size(Y);
k=1;
for i=1:M
    for j=1:n-1
        [temp,loc]=ismember(attribute({Y{i}},j),unique(attribute(X,j)));%
            decision attribute calculate posterior probability
        P_X=post_pro{j}(:,loc). *P_X;% product of posterior probabilities
            of each conditional attribute (Bayesian formula)
    end
    [MAX,I]=max(P_X);%find the maximum value
    out{k}=uniq_decis{I};%which class of decision attributes has the
        largest posterior probability, then the subsample belongs to that
        class
    k=k+1;
    P_X=ones(size(uniq_decis,1),1);%initialize the posterior probability of
        the decision attribute P_X again to prepare for the next sample
        calculation
end
out=out'; %output result (transposed form)
%%--------------------------------------------------------------------------------------
%%------------------------------------------LSTM---------------------------------------
data =xlsread('Attachment 1: Data of three meteorological stations in
    zhengzhou annex.xlsx');%Read rainfall data

%% The first 90% of the sequence is used for training and the last 10% for
    testing
numTimeStepsTrain = floor(0.9*numel(data));
dataTrain = data(1:numTimeStepsTrain+1);
dataTest = data(numTimeStepsTrain+1:end);



% Data preprocessing to normalize the training data to have zero mean and
    unit variance.
mu = mean(dataTrain);
sig = std(dataTrain);
dataTrainStandardized = (dataTrain - mu) / sig;



% input LSTM time series alternating one time step

XTrain = dataTrainStandardized(1:end-1);

YTrain = dataTrainStandardized(2:end);



%%
```

```
% Create the LSTM regression network, specifying the number of implicit cells in the LSTM layer 96*3

%Series prediction, so one dimensional input, one dimensional output

numFeatures = 1;

numResponses = 1;

numHiddenUnits = 96*3;

layers = [ ...

    sequenceInputLayer(numFeatures)

    lstmLayer(numHiddenUnits)

    fullyConnectedLayer(numResponses)

    regressionLayer];



% Specify training options, solver set to adam, 250 training rounds.
The %gradient threshold is set to 1. Specify an initial learning rate of 0.005, which is reduced after 125 training rounds by multiplying by a factor of 0.2.
options = trainingOptions('adam', ...
    'MaxEpochs',250, ...
    'GradientThreshold',1, ...
    'InitialLearnRate',0.005, ...
    'LearnRateSchedule','piecewise', ...
    'LearnRateDropPeriod',125, ...
    'LearnRateDropFactor',0.2, ...
    'Verbose',0, ...
    'Plots','training-progress');

% trainLSTM
net = trainNetwork(XTrain,YTrain,layers,options);
```

```
dataTestStandardized = (dataTest - mu) / sig;
XTest = dataTestStandardized(1:end-1);
YTest = dataTest(2:end);



net = resetState(net);
net = predictAndUpdateState(net,XTrain);



YPred = [];
numTimeStepsTest = numel(XTest);
for i = 1:numTimeStepsTest

   [net,YPred(:,i)] =
      predictAndUpdateState(net,XTest(:,i),'ExecutionEnvironment','cpu');

end

% Denormalize the prediction using the previously calculated parameters.

YPred = sig*YPred + mu;

% Calculate the root mean square error (RMSE).

rmse = sqrt(mean((YPred - YTest). ^2));
%Compare predicted values with test data.

figure
subplot(2,1,1)
plot(YTest)
hold on
plot(YPred,'. -')
hold off
```

```
legend(["Observed" "Predicted"])
ylabel("Loads")
title("Forecast with Updates")
subplot(2,1,2)
stem(YPred - YTest)
ylabel("Error")
xlabel("Days")
title("RMSE = " + rmse)

figure
subplot(2,1,1)
plot(dataTrain(1:end-1))
hold on
idx = numTimeStepsTrain:(numTimeStepsTrain+numTimeStepsTest);
plot(idx,[data(numTimeStepsTrain) YPred],’. -’)
hold off

ylabel("Loads")
xlabel("Days")

title("Forecast")
subplot(2,1,2)
legend(["Observed" "Forecast"])

plot(data)
xlabel("Days")
ylabel("Loads")
title("Daily load")
```