Title:Economic Development and Changes in Pollutant Emissions Based on Cluster Analysis

Faculty of Innovation Engineering

School of Computer Science and Engineering

Thesis for Degree of Bachelor of Science

MACAU UNIVERSITY OF SCIENCE AND TECHNOLOGY

Student Name: Pu Anyu

Student Number: 18098533-I011-0032

Supervisor Name: Yang Lei

Date: April,2022

Abstract

Clustering analysis has a long history of research, and its importance and cross-cutting nature with other research directions has been recognized for decades. Clustering is one of the most important aspects of data mining and pattern recognition research, and it plays an important role in identifying the intrinsic structure of data. Clustering is mainly applied to speech recognition and character recognition in pattern recognition, to image segmentation and machine vision in machine learning, and to data compression and information retrieval in image processing. In addition, cluster analysis is also widely used in statistical science. The definition of clustering given by Everitt<1>in 1974 is: entities within a class cluster are similar, entities in different class clusters are dissimilar; a class cluster is a convergence of points in a test space, the distance between any two points in the same class cluster is less than the distance between any two points in different class clusters; a class cluster can be described as A class cluster can be described as a connected region in a multidimensional space containing a relatively dense set of points that are separated from other regions (class clusters) by a region containing a relatively low-density set of points. So in fact, clustering is an unsupervised classification. Now I know that cluster analysis, an important unsupervised learning method, is widely recognised as a tool for data analysis in a variety of fields. Cluster analysis is an unsupervised learning process that finds sets of similar elements in a data set. Data sets from different application areas have different characteristics, and the purpose of clustering analysis varies, as does the method of clustering analysis depending on the data set and the purpose of use. In this paper, the purpose of cluster analysis can be commonly described as finding ”clusters” in a dataset, i.e. a collection of similar elements. Due to the large size of the dataset chosen for this study, it contains 31 provinces and cities across China. Therefore, the K-means algorithm, which has an outstanding advantage in terms of algorithmic complexity, is used in this study.

摘要

聚类分析研究有很长的历史, 几十年来, 其重要性及与其他研究方向的交叉特性得到人们的肯定. 聚类是数据挖掘、模式识别等研究方向的重要研究内容之一, 在识别数据的内在结构方面具有极其重要的作用. 聚类主要应用于模式识别中的语音识别、字符识别等, 机器学习中的聚类算法应用于图像分割和机器视觉, 图像处理中聚类用于数据压缩和信息检索。此外, 聚类分析还广泛应用于统计科学。但是，迄今为止, 聚类分析还没有一个学术界公认的定义。Everitt<1>在1974年关于聚类所下的定义是: 一个类簇内的实体是相似的, 不同类簇的实体是不相似的; 一个类簇是测试空间中点的会聚, 同一类簇的任意两个点间的距离小于不同类簇的任意两个点间的距离; 类簇可以描述为一个包含密度相对较高的点集的多维空间中的连通区域, 它们借助包含密度相对较低的点集的区域与其他区域(类簇)相分离。所以事实上, 聚类是一个无监督的分类。现在我知道聚类分析，是一种重要的无监督学习方法, 作为数据分析的工具, 其重要性在各个领域都得到了广泛的认可。聚类分析就是一个在数据集中寻找相似元素集合的无监督学习过程. 来自不同应用领域的数据集具有不同的特点, 人们对数据进行聚类分析的目的也不尽相同, 聚类分析的方法因数据集和使用目的而各不相同。本文聚类分析的目的可以通俗的说为寻找数据集中的“簇”，也就是相似元素的集合。由于本次选用的数据集较大，包含全国31省市。所以，本次研究在算法上选用在算法复杂度方面具有突出优势的K-means算法。

Contents

1 Introduction 3

1.1 Research Background 3

1.2 Basis of data selection 4

2 Analysis of economic development and total pollutant emissions 7

2.1 Relevant research literature for analysis with aggregate data 7

2.2 The Incremental versus the Aggregate Analysis 7

2.2.1 Analysis of total volume 7

2.2.2 Incremental analysis 8

2.2.3 The relationship between aggregate and incremental analysis 8

3 Cluster analysis methods 9

3.1 The Process of Cluster Analysis 9

3.2 Categories of clustering algorithms 9

3.3 Hierarchical clustering algorithm 10

3.3.1 Traditional aggregation rules 10

3.3.2 New hierarchical aggregation algorithm 10

3.4 Delineated clustering algorithm 11

3.4.1 K-means clustering 11

3.4.2 Strengths and weaknesses of the K-means algorithm 13

3.5 Graph theory-based clustering methods 13

3.6 Grid and density based clustering algorithms 13

3.7 Comparison of hierarchical aggregation algorithm and K-means algorithm . . . 14

4 Incremental data clustering analysis model building 15

4.1 Normalisation of incremental data 15

4.1.1 Definition of dimensionless 15

4.1.2 Dimensionless processing methods 15

4.1.3 Polarisation value method 15

4.1.4 Standardisation 15

4.1.5 Standardisation methods 16

4.2 Selection and processing of incremental data 17

4.3 Coefficient of variation method 19

4.4 Three incremental factors 21
4.4.1 Coefficient I: Incremental economic development factor 21
4.4.2 Coefficient II: Incremental environmental pollution factor 21
4.4.3 Coefficient III: Environmental Economic Increment Factor 21
4.5 Normalised results for incremental data 22

5 Incremental Clustering Joint Analysis 24
5.1 Cluster analysis of economic increments and pollutant increments 24
5.2 Joint analysis of economic increments and pollutant increments 24
5.2.1 Joint analysis of the incremental coefficient of economic development and the incremental coefficient of environmental pollution 24
5.2.2 Joint analysis of the incremental coefficient of economic development and the incremental coefficient of environmental economy 26
5.2.3 Joint analysis of the incremental coefficient of environmental pollution and the incremental coefficient of environmental economy 28

6 Appendix 29

7 Conclusion 34

8 Reference and Bibliography 35

9 Acknowledgement 37

1 Introduction

1.1 Research Background

According to the National Bureau of Statistics (NBS), published in the People's Daily on 29th August 2018, titled "NBS Releases Report on Economic and Social Development Achievements in the 40 Years of Reform and Opening Up", China's economy has undergone huge changes in the 40 years since the reform and opening up in 1978. In 1978<2>, China's GDP was only RMB 367.9 billion, but 40 years later, in 2017, it reached a historic RMB 827.122 billion. This means that China's economy in 2017 was 224.8 times larger than it was 40 years ago. But the price behind this magnificent economic figure is that China's environment is heavily polluted and consumes a lot of resources.According to a report on the official website of the Central People's Government of the People's Republic of China, "In the early 1980s<2>, the national investment in environmental pollution control was 2.5-3 billion yuan per year, and by the end of the 1980s the total annual investment exceeded 10 billion yuan. At the end of the Ninth Five-Year Plan (1996-2000), the total investment reached RMB 101 billion, accounting for the first time for more than 1% of the GDP in the same period. At the end of the 10th Five-Year Plan (2001-2005), the total investment reached 256.5 billion yuan, accounting for 1.37% of GDP in the same period; at the end of the 11th Five-Year Plan (2006-2010), the total investment reached 761.2 billion yuan, accounting for 1.84% of GDP in the same period. 2017<2>, China's total investment in environmental pollution control was 953.9 billion yuan, an increase of 7.2 times over 2001 an increase of 7.2 times, with an average annual growth rate of 14.0%. Among them, investment in the treatment of industrial pollution sources was RMB 68.2 billion, an increase of 2.9 times and an average annual growth rate of 8.9%; investment in environmental protection of RMB 277.2 billion was completed in the same year for environmental protection acceptance projects, an increase of 7.2 times and an average annual growth rate of 14.1%." According to the above data, the amount of money spent on environmental pollution control is rising at an alarming rate, which is even faster than the GDP growth rate, which means that China is facing a very serious environmental pollution problem. Therefore, environmental pollution and resource depletion are becoming the major problems that Chinese society will have to face in its next development. This model of economic growth at the expense of environmental pollution is not a healthy model of economic development, and it is difficult to sustain this model of economic development for a long time. It is therefore increasingly important to analyse the compatibility between the environment and the economy.

However, the analysis of the coherence between the environment and the economy is not a reasonable method of research in real life, because it does not eliminate the limitations imposed by the economic base, the size of the economy, on the results. In order to compensate for this limitation, we have chosen to base our research on incremental data, which can reflect the differences in development between subjects. By making vertical comparisons with our own past data and horizontal comparisons with other provinces, we can analyse the regularities between the development of the economy and the protection of the environment, so that we can better formulate the future direction of economic development.

1.2 Basis of data selection

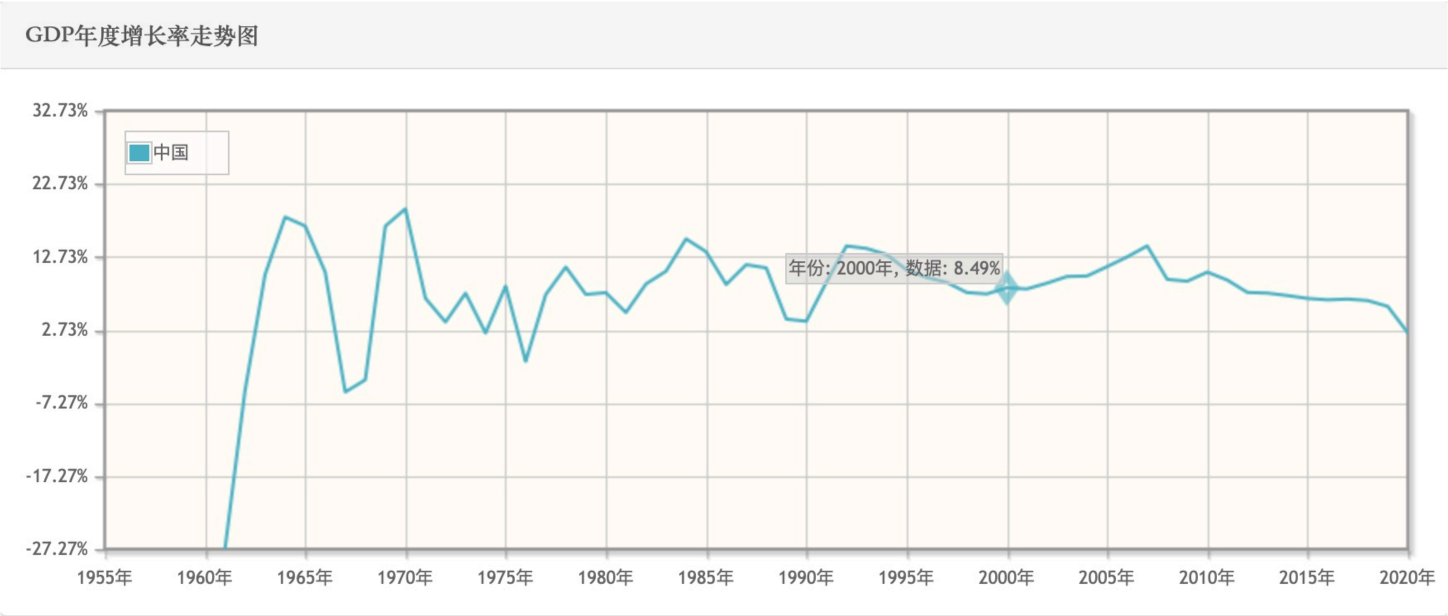
There are three main factors why the data for 2019 to 2020 was chosen for analysis in this thesis, rather than the latest data for 2020 to 2021.

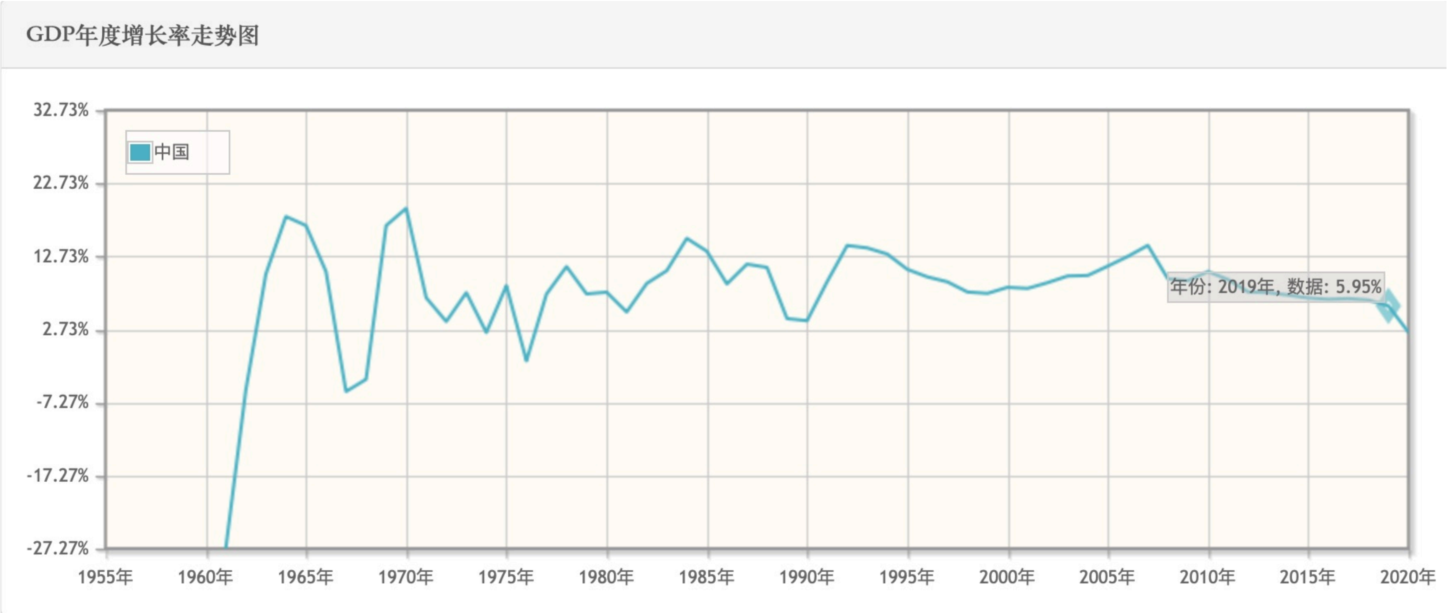
- Since entering the 21st century, the country has only had two years of GDP growth of less than 6% to date. These are: 2019 (5.95%) and 2020 (2.35%) respectively.

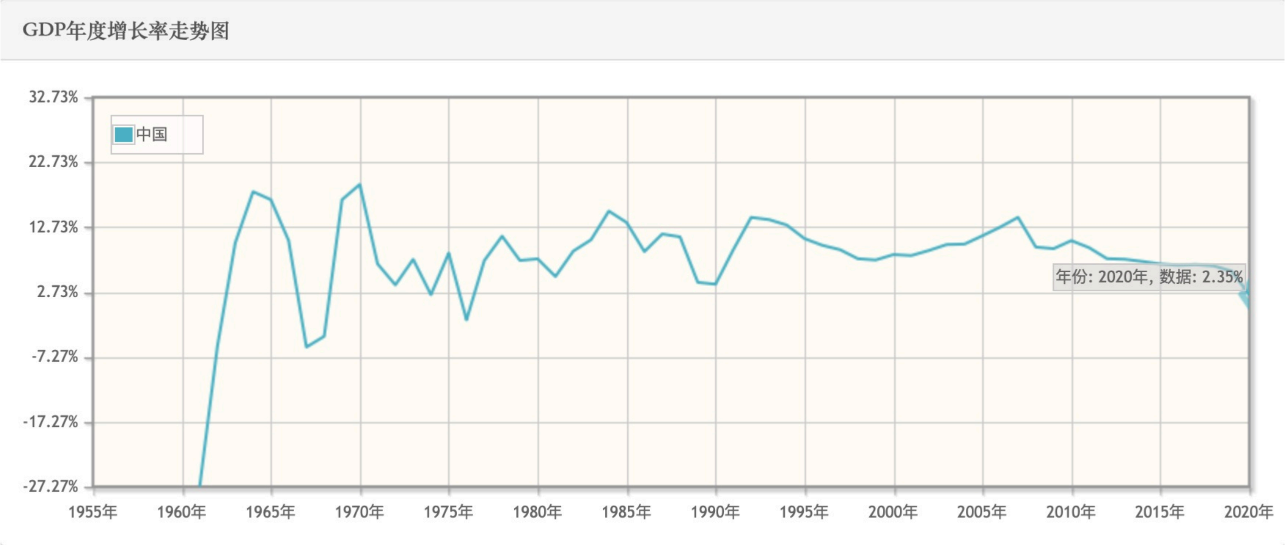
- China's GDP growth rate has seen a steady decline since the year 2019, so that 2019 is an important turning point

- As the Office of National Statistics generally publishes data with a time lag, the statistics for 2021 have not been published, although it is now 2022

Here are three following pictures: 2000 GDP growth rate, 2019 GDP growth rate and 2020 GDP growth rate for China.


GDP年度增长率走势图
中国
32.73%
22.73%
12.73%
2.73%
-7.27%
-17.27%
-27.27%
1955年
1960年
1965年
1970年
1975年
1980年
1985年
1990年
1995年
2000年
2005年
2010年
2015年
2020年
年份: 2000年, 数据: 8.49%


GDP年度增长率走势图
中国
32.73%
22.73%
12.73%
2.73%
-7.27%
-17.27%
-27.27%
1955年
1960年
1965年
1970年
1975年
1980年
1985年
1990年
1995年
2000年
2005年
2010年
2015年
2020年
年份: 2019年, 数据: 5.95%



In these three charts, it is clear that the GDP growth rate for 2019 is 5.95% and 2.35% for 2020. However, according to data released by the World Health Organization, the WHO<3> was first informed of the presence of ”viral pneumonia” in the People’s Republic of China on 31 December 2019 from the website of the Wuhan Municipal Health and Wellness Commission, and it was only on 3 January 2020 that Chinese officials officially provided the WHO with information on the occurrence of ”viral pneumonia” in Wuhan. It was only on 3 January 2020 that Chinese officials officially informed WHO of a cluster of cases of ”viral pneumonia of unknown origin” in Wuhan.<4> [So according to the timeline, the outbreak of the new coronavirus was only discovered at the end of 2019, but at a time when economic activity in China was not affected. issued Circular 1, which temporarily closed the airport and railway station from 10:00 a.m. for departures from Wuhan. This meant that the city of Wuhan was officially closed to the outside world. So the city closure that shocked the world has already passed almost two months in 2020. As I write this, my hometown is also under a city closure due to the New Crown epidemic. The entire city has come to a virtual standstill, with factories closing, shopping malls and restaurants shutting down, and public transport shutting down, which has had a huge impact on economic development. The city of Shanghai, China’s financial centre, has also seen a large number of confirmed cases. Although the city has not been closed, Shanghai is already experiencing a shortage of daily necessities in some areas. So many negative effects can be seen before the city is closed, let alone when the whole city stops functioning completely. However, Wuhan was already well into 2020 when the city was closed, and the city’s economic development in 2019 will not be affected by this factor. The city of Wuhan is also the source of the epidemic in China, so combined with the information from the above data, I believe that China’s economic development is not affected by the COVID-19 virus in 2019, so the 5.95% economic growth rate for the whole year in 2019 in China’s 31 provinces and cities is of research value.

2 Analysis of economic development and total pollutant emissions

2.1 Relevant research literature for analysis with aggregate data

Most of the studies on economic and environmental issues are now based on aggregate data for analysis. For example, Li Qinghua<5> chose the aggregate index of pollutant emissions to measure environmental pollution and concluded that the economic growth brought about through environmental pollution is short-lived and that the long-term effects of environmental pollution will constrain economic development; Feng Genfu<6> based on Theoretical and empirical analyses based on aggregate data, and the theoretical and empirical analyses by shifting the centre of gravity, conclude that the total control of environmental pollution should include both the total control of pollution variables and the total control of factors influencing the centre of gravity of environmental pollution; Zhao Jiahong's<7> analysis based on 2014 data on the total amount of relevant pollutants in 31 Chinese provinces and cities concluded that total control is as important as concentration control when conducting environmental and economic evaluations;Ding Jihong<8> examines these two relationships, focusing on the two-way relationship.In <An analysis of the relationship between economic growth and environmental pollution - an example from Jiangsu Province>, a two-way relationship between the two was found, focusing on the way economic growth is structured, i.e. the composition of inputs and outputs, to optimise the policy mix; Using aggregate data analysis, Wu Yuping<9> presents the importance of the environmental policies pursued by the government.

2.2 The Incremental versus the Aggregate Analysis

2.2.1 Analysis of total volume

Aggregate analysis refers to the analysis of the aggregate indicators of a research object and their patterns of change, with the main aim of studying the patterns of change in the aggregate indicators. This method examines the overall situation and structure of the object of study, so that the results of the study are important at a macro level for the whole. However, the limitation of aggregate analysis is that it tends to ignore the impact of individual differences on the overall results. At the same time, the 31 provinces and cities in China have different economic development bases and scales (the southeastern coastal region and the Yangtze River Delta region have earlier economic development and larger economies than the northwestern and southwestern regions), different geographical areas and different population sizes. So there are also significant differences in the economic and industrial structure. Plain regions are more likely to attract investment and have more developed transport networks than plateau regions, and regions with abundant water resources have larger populations and more developed industries than arid regions. Therefore, the analysis using aggregate data cannot eliminate the impact of such differences and the realistic limitations between actual development, and therefore the results obtained are not informative for all provinces and municipalities. This may lead to errors in the perception of the synergy between economic development and environmental pollution in the provinces and municipalities and may lead to a lag in action.

2.2.2 Incremental analysis

Incremental analysis is done by studying the data of a subject at different times and performing a difference operation. This method takes into account the differences in each subject and highlights the dynamic changes in each subject. The analysis based on incremental data reflects the development of each research subject and reflects objectively the coherence of its development and the gap with other subjects by comparing itself vertically and horizontally with other research subjects. Incremental analysis is an important aspect of the analysis of environmental economic systems. However, this method also has its shortcomings. It is limited in its ability to capture the overall situation of the sample at the macro level.

2.2.3 The relationship between aggregate and incremental analysis

Both aggregate and incremental analyses have their own strengths and weaknesses and are complementary to each other. In this study, for example, the purpose of both the aggregate and incremental analyses is to analyse data on pollutant indicators and economic development indicators in different provinces in order to compare the coordination of development in different places. The data sources for this experiment are all authoritative data published by the National Bureau of Statistics over a period of time. The aggregate analysis is the basis for the incremental analysis, while the incremental analysis complements and improves the aggregate analysis. It can be understood that incremental analysis is a dynamic representation of aggregate analysis. Both aggregate analysis and incremental analysis are complementary and are important foundations for environmental analysis, while incremental analysis has characteristics that aggregate analysis does not have. This study examines the relationship between pollutant emissions and economic development in 31 provinces and cities (excluding special administrative regions and Taiwan Province) based on incremental GDP data.

3 Cluster analysis methods

3.1 The Process of Cluster Analysis

A typical clustering process consists of data (or sample or pattern) preparation, feature selection and extraction, proximity calculation, clustering (or grouping), and evaluation of the validity of the clustering results.

Process of Cluster Analysis:

1) Data preparation: including feature normalisation and dimensionality reduction

2)Feature selection: select the most effective features from the initial set of features, and store them in a vector

3)Feature Extraction: New salient features are formed by transforming the selected features

4)Clustering (or grouping): first select a distance function (or construct a new distance function) of the appropriate feature type to measure proximity; then perform clustering or grouping

5)Evaluation of clustering results: is the evaluation of clustering results

3.2 Categories of clustering algorithms

There is no single clustering technique (clustering algorithm) that is universally applicable to reveal the wide variety of structures presented by various multidimensional data sets. There are various clustering algorithms based on the rules for the accumulation of data in clusters and the methods for applying these rules<8>:

1)Hierarchy-based clustering methods

2)Division-based clustering methods

3)Graph theory-based clustering methods

4)Density and grid-based clustering methods

3.3 Hierarchical clustering algorithm

Hierarchical clustering algorithms<10>, also known as tree clustering algorithms, use a hierarchical structure of data to repeatedly split or aggregate data in a hierarchical way to form a hierarchical sequence of clustering problem solutions.This algorithm is suitable for the classification of small data sets.

3.3.1 Traditional aggregation rules

The distance measure between two classes is an important part of traditional hierarchical aggregation algorithms, which consists of two important parameters: the similarity measure and the conjunction rule. Here, the Euclidean distance is used as the similarity measure, and the linkage rules include single linkage rule, complete linkage rule, inter-class average linkage rule, intra-class average linkage rule and Ward's method<11>.

3.3.2 New hierarchical aggregation algorithm

1)Binary-Positive method:

In 2007, Gelbard<12> et al. proposed a new hierarchical aggregation algorithm, called the binary-positive method. This method stores the data to be classified in positive binary form in a two-dimensional matrix, where the rows represent the records (objects) and the columns represent the possible values of their attributes. A record has a value of 1 or 0, indicating that it has a corresponding attribute value or that there is no corresponding attribute value, respectively. Therefore, the similarity distance is calculated only on the positive bits of the binary vector being compared, i.e. only between records (objects) with a value of 1. There are several Binary-Positve similarity measures represented by the Dice distance.

Gelbard<12> et al. experimented with 11 clustering algorithms using four datasets - Wine, Iris, Ecolic and Psychology balance - and showed that for any of the four datasets, the four methods such as Binary-Positive were the best in terms of overall accuracy of the clustering results. The results show that the four methods of Binary-Positive are the best in terms of the overall accuracy of the clustering results. They also concluded that converting the original data to positive binary would improve the correctness and robustness of the clustering results, especially for the hierarchical clustering algorithm.

2)Rough clustering of sequential data:

In 2007, Kumar<13> et al. proposed a new hierarchical clustering algorithm RCOSD based on indistinguishable coarse aggregation for continuous data. In this algorithm, the indistinguishable relations are extended to tolerant relations with non-strict transfer properties. The initial classes are formed using upper approximations of similarity, and subsequent classes are formed using the concept of upper approximations of constrained similarity, where a relative similarity condition is used as a merging criterion. The upper approximation technique is applied to obtain the upper approximation of coarse class clusters, where an element can belong to more than one class cluster. The algorithm introduces S3M as a similarity measure for Web data, which takes into account both the order of occurrence of items and the set content. The algorithm can merge two or more classes per iteration, thus speeding up hierarchical clustering. The algorithm can effectively mine continuous data and characterize the main features of class clusters, helping Web miners to characterize potential new Web user groups.

The experimental results of Pradeep Kumar<13> et al. on the essentially continuous MSNBC Web navigation dataset show that the RCOSD clustering algorithm is feasible compared to the traditional hierarchical clustering algorithm using sequence vector encoding. The description method given by the algorithm can help web miners to identify potentially meaningful groups of users.

3.4 Delineated clustering algorithm

A delimited clustering algorithm requires a pre-specified number of clusters or clustering centres, and iterative operations to gradually reduce the error value of the objective function, when the value of the objective function converges, the final clustering result is obtained.

3.4.1 K-means clustering

The K-means clustering algorithm (K-means) was first proposed by Mac Queen in 1967. The core idea of this algorithm is to find K cluster centres such that the sum of squares of each data point xi and its nearest cluster centre is minimised (this sum of squares is called the deviation D).

For example, given a training sample $x^{(1)}, ..., x^{(m)}$, for every $x^{(i)}$ has $x^{(i)} \epsilon R^n$, which means each sample element is an n-dimensional vector.

Description of the steps of the kmeans algorithm :

Step1: The k clustering prime points are randomly selected as μ_1, μ_2, ..., μ_k ϵ R^n

Step2: Repeat the following process until convergence:

For each sample i, calculate the class it should belong to

$$C^{(i)} := arg\,\underset{j}{min}\,||x^{(i)} - \mu_j||^2.$$

For each class j, recalculate the centre of mass of that class

$$\mu_j := \frac{\sum\limits_{i=1}^{m} 1[c^{(i)}=j]x^{(i)}}{\sum\limits_{i=1}^{m} 1[c^{(i)}=j]}$$

K is the number of clusters, $C^{(i)}$ represents the closest of sample i and the k classes, The value of $C^{(i)}$ is one of 1 to k. Quality Heart μ_j represents our guess on the centroids of samples belonging to the same class.To explain this in terms of the cluster model is to cluster all the stars into k clusters by first selecting k random points in the universe (or k stars) as the centre of mass of the k clusters.Then in the first step, for each star, the distance to each of the k centres of mass is calculated, and the cluster with the closest distance is selected as $C^{(i)}$. Thus, after the first step, each star has a cluster to which it belongs. In the second step, for each cluster, its centre of mass is recalculated μ_j. Then average the coordinates of all the stars in the cluster. Repeat the first and second steps until the centre of mass remains the same or changes very little.

For the explanation of the formula :

The arg in the first equation is the marker symbol, i.e. it indicates which sample parameter belongs to which class used, and the min minimization j that immediately follows is the function J that I will talk about next.

$$\mathrm{J} = \sum\limits_{i=1}^{n} \sum\limits_{j=1}^{k} r_{ij}||x^{(i)} - \mu_j||^2$$

This is the qualitative description in the kmeans algorithm, and the notation in the formula is still the same as described above. $x^{(i)} \epsilon R^{(n)}$, μ_j indicates the category to which the 1st sample belongs, r_{ij} indicates that the data point x(i) is 1 if it is classified to μ_j , otherwise it is 0.

3.4.2 Strengths and weaknesses of the K-means algorithm

Advantages of kmeans algorithm<14> :

Efficient classification of large datasets with a computational complexity of O (t Kmn) , where t is the number of iterations, K is the number of clusters, m is the number of feature attributes, n is the number of objects to be classified, usually, K, m, t«n. The K-means algorithm is much faster than the hierarchical clustering algorithm when clustering large datasets.

Disadvantages of kmeans algorithm<14> :

Usually terminates when a local optimum is obtained; only suitable for clustering numerical data; only suitable for data sets with convex clusters (i.e. class clusters are convex).

3.5 Graph theory-based clustering methods

The graph theory-based clustering method transforms the dataset to be clustered into a weighted undirected complete graph G = (V,E). where: the set of vertices V is the data points in the feature space, and the edge set and its weights are the linkage and similarity between any two data points. In this way, the clustering problem can be transformed into a graph partitioning problem, where the resulting subgraphs correspond to the clusters contained in the dataset. In recent years, representative research . The results include the GBR algorithm, the maximum distance subtree based clustering algorithm and the dominant set based point pair clustering algorithm.Most graph theory-based clustering methods use point pairs of data to represent interrelationships between data points, making them more suitable than other methods.These methods are more suitable for discovering irregularly shaped clusters in the data set than other methods.

3.6 Grid and density based clustering algorithms

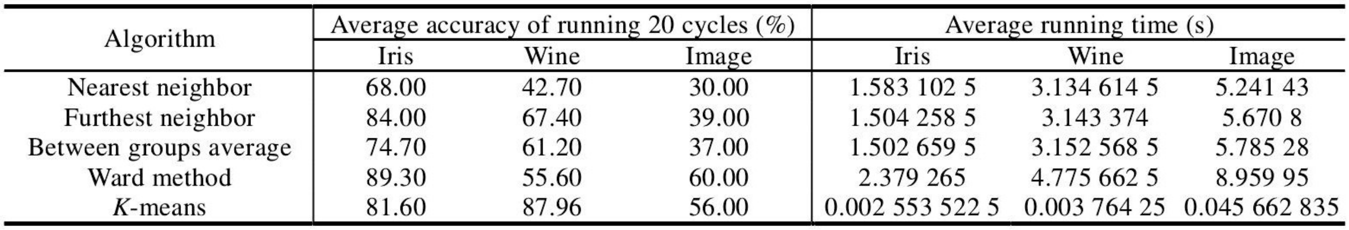
Grid<15> and density-based clustering methods are an important class of clustering methods that are widely used in many fields, including spatial information processing. In particular, with the recent development of scalable clustering methods for large data sets, they are becoming increasingly active in the subfield of spatial data mining research.

Differences from traditional clustering algorithms:

Density-based clustering algorithms use data density (number of instances per unit area) to discover arbitrarily shaped clusters; grid-based clustering algorithms use a grid structure to organize the value space around patterns divided by rectangular blocks, and achieve pattern clustering based on the distribution of the blocks. Grid-based clustering algorithms are often combined with other methods, in particular with density-based clustering methods.

3.7 Comparison of hierarchical aggregation algorithm and K-means algorithm

Iris, Wine, Image<16> conducted 20 randomized clustering experiments on the dataset in UCI using the single linkage method in the hierarchical aggregation algorithm, the complete linkage method, the average linkage between classes method, the Ward method and the K-means algorithm in the divisive clustering algorithm, respectively.The result are following:



Algorithm	Average accuracy of running 20 cycles (%)			Average running time (s)		
	Iris	Wine	Image	Iris	Wine	Image
Nearest neighbor	68.00	42.70	30.00	1.583 102 5	3.134 614 5	5.241 43
Furthest neighbor	84.00	67.40	39.00	1.504 258 5	3.143 374	5.670 8
Between groups average	74.70	61.20	37.00	1.502 659 5	3.152 568 5	5.785 28
Ward method	89.30	55.60	60.00	2.379 265	4.775 662 5	8.959 95
K-means	81.60	87.96	56.00	0.002 553 522 5	0.003 764 25	0.045 662 835

Figure 1:Result of experiments

The experimental results show that the traditional hierarchical aggregation algorithm does not give good classification results for the Wine dataset with good clustering structure, which is related to the poor reassignment ability of the traditional hierarchical aggregation algorithm (i.e. if some data are assigned to a class cluster in the initial stage, they cannot be reassigned to other class clusters). We also found that the clustering results are unpredictable and the correct clustering rate of the same algorithm may vary greatly for different data sets, and the clustering results and efficiency of different clustering algorithms for the same data set may also vary greatly. Therefore, in practice, the clustering algorithm should be chosen according to the type of data to be clustered and the clustering structure (if available) in order to achieve the best clustering results.

So,according to this experiment.K-means algorithm is a good choice.

4 Incremental data clustering analysis model building

4.1 Normalisation of incremental data

In order to try to eliminate the effects of differences in the panel data bases of the 31 provinces and cities, and also to ensure the scientific accuracy and rigour of the study results. I will treat the four parameters of GDP and NOx emissions, SO2 emissions and particulate matter emissions as dimensionless data.

4.1.1 Definition of dimensionless

Dimensionlessisation, also known as normalisation of data, refers to the fact that data is not comparable between different data indicators due to differences in scale (i.e. different units of measurement), so first the indicators need to be dimensionlessised to remove the effects of scale before proceeding to the next analysis.

4.1.2 Dimensionless processing methods

The main methods commonly used for dimensionlessization are:

- Polarisation value method

- Standardisation methods

- Averaging value method

4.1.3 Polarisation value method

Dimensionlessisation of variable data using the polarisation method involves converting the original data into data in a particular range by taking the maximum and minimum values of the variable, thereby eliminating the effects of magnitude and order of magnitude. As the polarisation method dimensionlessises a variable only in relation to the two extreme values of the variable, the maximum and the minimum, and not in relation to other values. This makes the polarisation method overly dependent on the two extreme values when changing the weights of each variable.

4.1.4 Standardisation

Before data can be analysed, it is often necessary to normalise the data and use the normalised data for data analysis. Data normalisation is also known as indexation of statistical data. Data normalisation consists of two main aspects, namely data homogenisation and dimensionless normalisation. The data homogenisation process mainly addresses the issue of different nature of data, the direct summation of different nature of indicators does not correctly reflect the comprehensive results of different forces, must first consider changing the nature of the inverse indicators data, so that all indicators on the assessment scheme of the force of the same convergence, and then summed to get the correct results.

4.1.5 Standardisation methods

Data dimensionless processing addresses the comparability of data. There are various methods<17> of standardising data, such as ’Min-max standardised data scaling’ and ’Z-score standardisation’. After the above standardisation process, the raw data is converted into dimensionless indicator measurements, i.e. the indicator values are all at the same quantitative level, and the data can then be analysed. Some of the common methods are:

- Min-max standardised data scaling

The min-max normalisation method is a process of linear transformation of the original data. Assuming minA and maxA are the minimum and maximum values of attribute A respectively, an original value x of A is mapped to a value x’ in the interval [0,1] by min-max normalisation with the formula: new data = (original data - minimum value)/(maximum value - minimum value)

- Z-score Standardisation<18>

The z-score normalisation method normalises data based on the mean and standard deviation of the original data. The original value of A, x, is normalised to x’ using z-score. z-score normalisation is used where the maximum and minimum values of attribute A are unknown, or where there are outliers outside the range of values. The data is subtracted from its mean by its attribute (done by column) and then divided by its variance. The final result obtained is that for each attribute/per column all data are clustered around 0 with a variance value of 1. The specific formula is shown below:

$$Z_{x_i} = \frac{x_i - mean[x_j];1\leq j\leq n}{std[x_j];1\leq j\leq n}$$

Z-score normalisation method formula

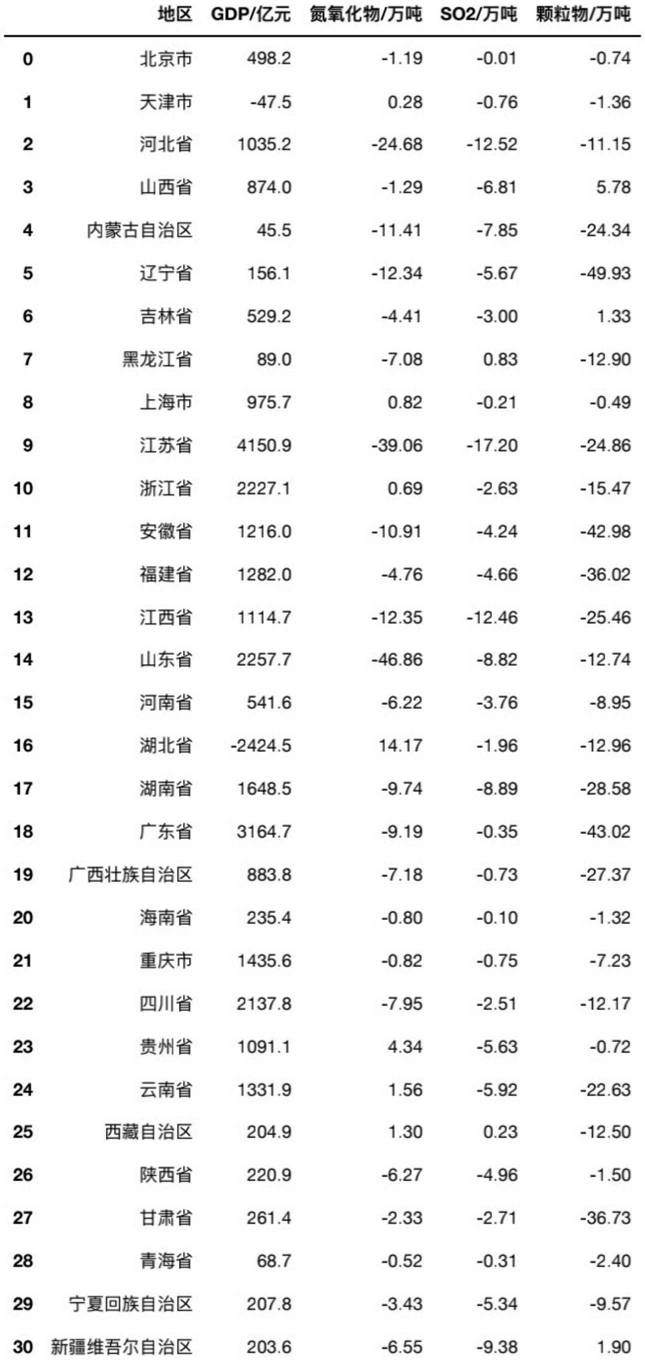
Combining the characteristics of each method, I decided to choose the z-score standardisation method to standardise the data for this paper. The formula for this study was derived based on the formula shown above and by bringing the data into the formula.The specific formula<18> is shown below:

$$x'_{ij} = \frac{x_{ij} - \overline{x}_j}{S_j}$$

Bringing the data into the Z-score normalisation method formula

4.2 Selection and processing of incremental data

This paper takes pollutant emissions and GDP as the object of study, and selects the data of SO2 emissions, particulate matter emissions, NOx emissions and GDP GDP of 31 provinces and municipalities (excluding special administrative regions and Taiwan Province) from 2019 to 2020, and obtains the corresponding incremental data of SO2 emissions, particulate matter emissions, NOx emissions and incremental data of economic growth after difference calculation, as shown in Table 1.



	地区	GDP/亿元	氮氧化物/万吨	SO2/万吨	颗粒物/万吨
0	北京市	498.2	-1.19	-0.01	-0.74
1	天津市	-47.5	0.28	-0.76	-1.36
2	河北省	1035.2	-24.68	-12.52	-11.15
3	山西省	874.0	-1.29	-6.81	5.78
4	内蒙古自治区	45.5	-11.41	-7.85	-24.34
5	辽宁省	156.1	-12.34	-5.67	-49.93
6	吉林省	529.2	-4.41	-3.00	1.33
7	黑龙江省	89.0	-7.08	0.83	-12.90
8	上海市	975.7	0.82	-0.21	-0.49
9	江苏省	4150.9	-39.06	-17.20	-24.86
10	浙江省	2227.1	0.69	-2.63	-15.47
11	安徽省	1216.0	-10.91	-4.24	-42.98
12	福建省	1282.0	-4.76	-4.66	-36.02
13	江西省	1114.7	-12.35	-12.46	-25.46
14	山东省	2257.7	-46.86	-8.82	-12.74
15	河南省	541.6	-6.22	-3.76	-8.95
16	湖北省	-2424.5	14.17	-1.96	-12.96
17	湖南省	1648.5	-9.74	-8.89	-28.58
18	广东省	3164.7	-9.19	-0.35	-43.02
19	广西壮族自治区	883.8	-7.18	-0.73	-27.37
20	海南省	235.4	-0.80	-0.10	-1.32
21	重庆市	1435.6	-0.82	-0.75	-7.23
22	四川省	2137.8	-7.95	-2.51	-12.17
23	贵州省	1091.1	4.34	-5.63	-0.72
24	云南省	1331.9	1.56	-5.92	-22.63
25	西藏自治区	204.9	1.30	0.23	-12.50
26	陕西省	220.9	-6.27	-4.96	-1.50
27	甘肃省	261.4	-2.33	-2.71	-36.73
28	青海省	68.7	-0.52	-0.31	-2.40
29	宁夏回族自治区	207.8	-3.43	-5.34	-9.57
30	新疆维吾尔自治区	203.6	-6.55	-9.38	1.90

Table1: 2019-2020 Total value of domestic production and incremental pollutant emissions, standardized results

4.3 Coefficient of variation method

The coefficient of variation, also known as the standard deviation rate, is another statistical measure of the degree of variation of observations in a data set. When comparing the degree of variation of two or more data, if the unit of measure is the same as the mean, the standard deviation can be chosen directly for comparison. If the subjects have different units, the standard deviation cannot be used to compare the degree of variation between subjects, but rather the ratio of the standard deviation to the mean. The specific formula<19> as follows:

$$(COV)_i = \frac{\delta_i}{\overline{x}_i}$$

Coefficient of variation formula

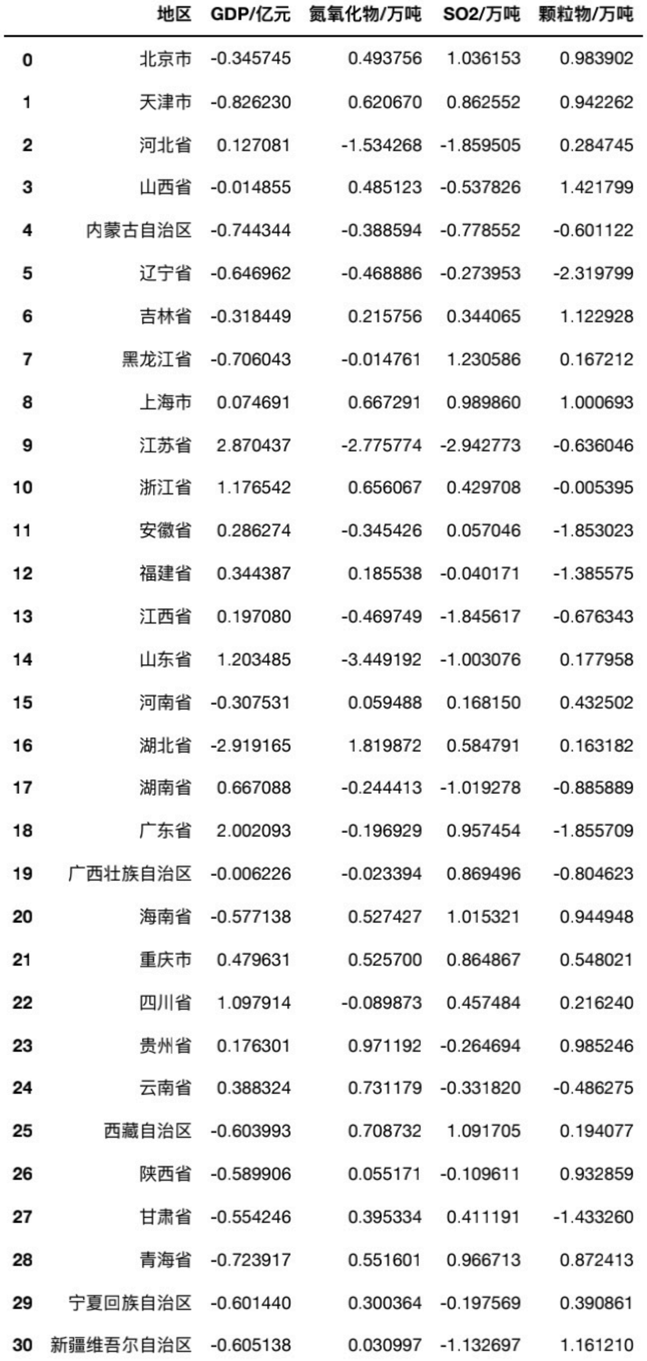
The next step formula<19> as shown:

$$\omega_i = \frac{(COV)_i}{\sum\limits_{i=1}^{n}(COV)_i}$$

Weighting percentage calculation formula

The numerator of the first formula is the standard deviation and the denominator is the mean. The results derived from formula one are then carried over into formula two to calculate the results and ultimately the weighting percentages.

In this formula: x_{ij} is the value of the jth variable in the i-th year. $\overline{x}_j$ is the arithmetic mean of the jth variable. S_j is the standard deviation of the jth variable. x'_{ij} is the normalised value. The incremental data in Table 1 were normalised to form Table 2:



	地区	GDP/亿元	氮氧化物/万吨	SO2/万吨	颗粒物/万吨
0	北京市	-0.345745	0.493756	1.036153	0.983902
1	天津市	-0.826230	0.620670	0.862552	0.942262
2	河北省	0.127081	-1.534268	-1.859505	0.284745
3	山西省	-0.014855	0.485123	-0.537826	1.421799
4	内蒙古自治区	-0.744344	-0.388594	-0.778552	-0.601122
5	辽宁省	-0.646962	-0.468886	-0.273953	-2.319799
6	吉林省	-0.318449	0.215756	0.344065	1.122928
7	黑龙江省	-0.706043	-0.014761	1.230586	0.167212
8	上海市	0.074691	0.667291	0.989860	1.000693
9	江苏省	2.870437	-2.775774	-2.942773	-0.636046
10	浙江省	1.176542	0.656067	0.429708	-0.005395
11	安徽省	0.286274	-0.345426	0.057046	-1.853023
12	福建省	0.344387	0.185538	-0.040171	-1.385575
13	江西省	0.197080	-0.469749	-1.845617	-0.676343
14	山东省	1.203485	-3.449192	-1.003076	0.177958
15	河南省	-0.307531	0.059488	0.168150	0.432502
16	湖北省	-2.919165	1.819872	0.584791	0.163182
17	湖南省	0.667088	-0.244413	-1.019278	-0.885889
18	广东省	2.002093	-0.196929	0.957454	-1.855709
19	广西壮族自治区	-0.006226	-0.023394	0.869496	-0.804623
20	海南省	-0.577138	0.527427	1.015321	0.944948
21	重庆市	0.479631	0.525700	0.864867	0.548021
22	四川省	1.097914	-0.089873	0.457484	0.216240
23	贵州省	0.176301	0.971192	-0.264694	0.985246
24	云南省	0.388324	0.731179	-0.331820	-0.486275
25	西藏自治区	-0.603993	0.708732	1.091705	0.194077
26	陕西省	-0.589906	0.055171	-0.109611	0.932859
27	甘肃省	-0.554246	0.395334	0.411191	-1.433260
28	青海省	-0.723917	0.551601	0.966713	0.872413
29	宁夏回族自治区	-0.601440	0.300364	-0.197569	0.390861
30	新疆维吾尔自治区	-0.605138	0.030997	-1.132697	1.161210

Table 2: Three incremental factors by province (municipality, autonomous region) for 2019-2020

4.4 Three incremental factors

4.4.1 Coefficient I: Incremental economic development factor

A dimensionless economic development coefficient is obtained by normalising the incremental GDP data for the 31 provinces and cities across the country. A larger value indicates a better economic development for the region in 2019 to 2020. Conversely, the smaller the economic development coefficient, the worse the region's economic development will be in 2019-2020.

4.4.2 Coefficient II: Incremental environmental pollution factor

By normalising the data increments of nitrogen oxide emissions, SO2 emissions and particulate matter emissions for each of the 31 provinces and municipalities in China, the indicators were assigned weights of 0.4 for nitrogen oxide, 0.4 for SO2 and 0.2 for particulate matter, and the data were normalised according to these weights to obtain the dimensionless environmental pollution increment coefficients for the 31 provinces and municipalities from 2019 to 2020. This coefficient can be used to indicate the environmental pollution situation in each of the 31 provinces and municipalities in 2019. When the value is greater than 0, a larger value indicates a more serious environmental deterioration in that region in 2019. Conversely, when the value is less than 0, a smaller value indicates a better environmental situation for that region in 2019.

4.4.3 Coefficient III: Environmental Economic Increment Factor

As can be determined from the name of coefficient III, this coefficient is the result of combining coefficient I with coefficient II. Firstly, in order to be able to combine coefficient one and coefficient two, it is necessary to determine separately the weights of Coefficient I and Coefficient II. Through research, I learned that the coefficient of variation method can be used to determine the specific weights to be assigned. Ultimately, the calculation determined that the weight of the incremental economic development coefficient was 0.6 and the weight of the incremental environmental pollution coefficient was 0.4. Based on the data from Coefficient I and Coefficient II and the weights assigned, a normalisation process was carried out to obtain the incremental environmental economic coefficient. Larger data indicates that the region has a better overall environmental economy in 2019. Conversely, smaller data indicates that the region's overall environmental economy is less developed in 2019.

In this table:

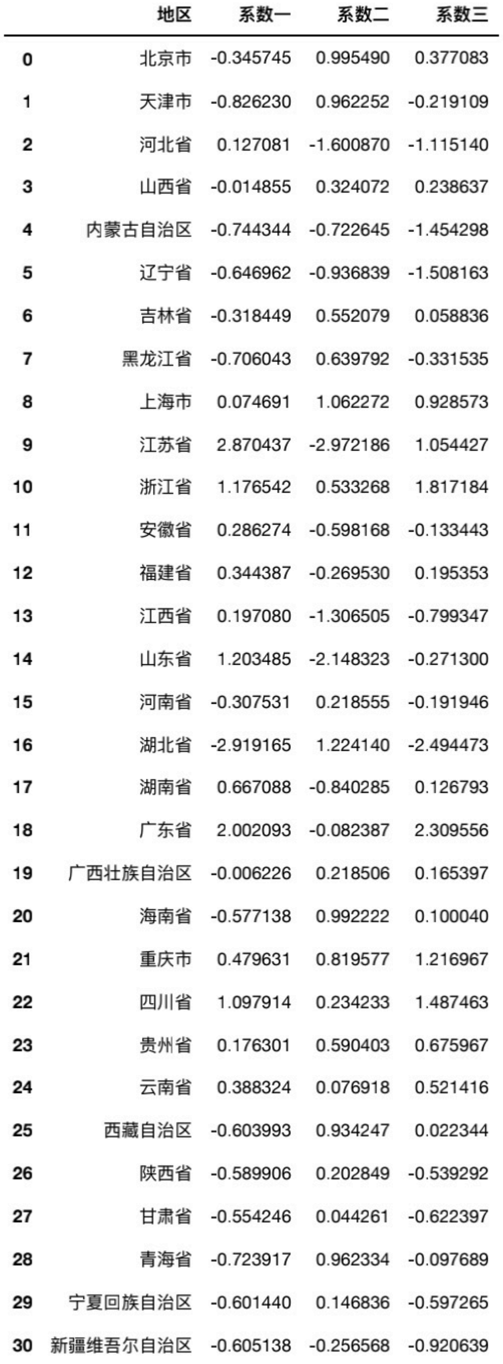
- Coefficient one is: incremental economic development coefficient

- Coefficient two is: incremental environmental pollution coefficient

- Coefficient three is: incremental environmental economic coefficient.

4.5 Normalised results for incremental data

In order to eliminate the influence of differences in the level of panel data across provinces (municipalities, autonomous regions) and to ensure the scientific accuracy and rigour of the results, the four parameters of domestic GDP, NOx emissions, SO2 emissions and particulate matter emissions need to be processed as dimensionless data, according to the formula in Figure 3, and the data in Table 2 have been standardised to form Table 3:



	地区	系数一	系数二	系数三
0	北京市	-0.345745	0.995490	0.377083
1	天津市	-0.826230	0.962252	-0.219109
2	河北省	0.127081	-1.600870	-1.115140
3	山西省	-0.014855	0.324072	0.238637
4	内蒙古自治区	-0.744344	-0.722645	-1.454298
5	辽宁省	-0.646962	-0.936839	-1.508163
6	吉林省	-0.318449	0.552079	0.058836
7	黑龙江省	-0.706043	0.639792	-0.331535
8	上海市	0.074691	1.062272	0.928573
9	江苏省	2.870437	-2.972186	1.054427
10	浙江省	1.176542	0.533268	1.817184
11	安徽省	0.286274	-0.598168	-0.133443
12	福建省	0.344387	-0.269530	0.195353
13	江西省	0.197080	-1.306505	-0.799347
14	山东省	1.203485	-2.148323	-0.271300
15	河南省	-0.307531	0.218555	-0.191946
16	湖北省	-2.919165	1.224140	-2.494473
17	湖南省	0.667088	-0.840285	0.126793
18	广东省	2.002093	-0.082387	2.309556
19	广西壮族自治区	-0.006226	0.218506	0.165397
20	海南省	-0.577138	0.992222	0.100040
21	重庆市	0.479631	0.819577	1.216967
22	四川省	1.097914	0.234233	1.487463
23	贵州省	0.176301	0.590403	0.675967
24	云南省	0.388324	0.076918	0.521416
25	西藏自治区	-0.603993	0.934247	0.022344
26	陕西省	-0.589906	0.202849	-0.539292
27	甘肃省	-0.554246	0.044261	-0.622397
28	青海省	-0.723917	0.962334	-0.097689
29	宁夏回族自治区	-0.601440	0.146836	-0.597265
30	新疆维吾尔自治区	-0.605138	-0.256568	-0.920639

Table 3: Three incremental factors by province (municipality, autonomous region) for 2019-2020

5 Incremental Clustering Joint Analysis

5.1 Cluster analysis of economic increments and pollutant increments

The economic development coefficient, environmental pollution coefficient and environmental economic value added coefficient of 31 provinces and cities were clustered by python, so that samples with similar characteristics were clustered together and those with greater differences were distinguished.

The results of the clustering of the incremental economic development coefficients are divided into three categories: Category 1 indicates that the incremental economic development coefficients are small and contain regions that will experience a small increase in economic growth in 2019 compared to other regions. Category 2 indicates that the economic growth rate in 2019 is at a medium level compared to other regions. Category 3 indicates that the economic growth rate in 2019 is at a significant advantage compared to other regions, with the fastest growth rate.

The clustering results set for the incremental environmental pollution coefficient are also divided into three categories: Category 1 indicates that the incremental environmental pollution coefficient is small and the level of environmental pollution resulting from regional economic development in 2019 is small. Category 2 indicates a moderate level of environmental pollution from regional economic development in 2019. Category 3 indicates the most serious environmental pollution resulting from regional economic development in 2019.

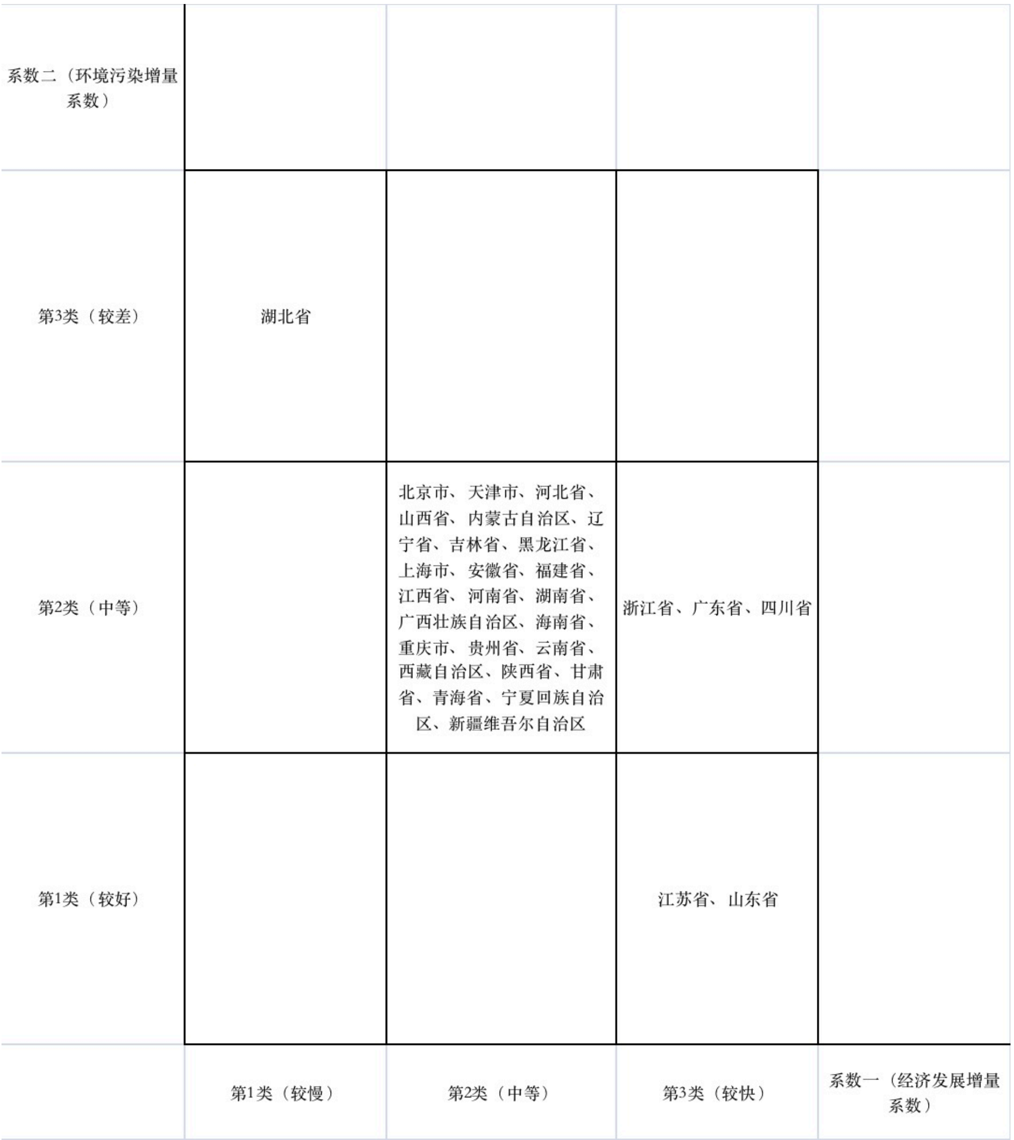
The results of the clustering of the incremental coefficients of the environmental economy are also divided into three categories: category 1 indicates the smallest incremental coefficient of the environmental economy, which means that the regional economy is the least coordinated with the environment. Category 2 indicates that the degree of coordination between the regional economy and the environment is at a medium level. Category 3 indicates the best level of coordination between the regional economy and the environment.

5.2 Joint analysis of economic increments and pollutant increments

After clustering and analysing the incremental economic development coefficient, the environmental pollution coefficient and the incremental environmental economic coefficient for each of the 31 provinces and cities in China, the three were then analysed jointly.

5.2.1 Joint analysis of the incremental coefficient of economic development and the incremental coefficient of environmental pollution

A joint analysis<20> of incremental economic development and incremental environmental pollution was first conducted and the joint analysis of the incremental economic development coefficient and incremental environmental pollution coefficient is shown in Figure 2:



系数二（环境污染增量系数）				
第3类（较差）	湖北省			
第2类（中等）		北京市、天津市、河北省、山西省、内蒙古自治区、辽宁省、吉林省、黑龙江省、上海市、安徽省、福建省、江西省、河南省、湖南省、广西壮族自治区、海南省、重庆市、贵州省、云南省、西藏自治区、陕西省、甘肃省、青海省、宁夏回族自治区、新疆维吾尔自治区	浙江省、广东省、四川省	
第1类（较好）			江苏省、山东省	
	第1类（较慢）	第2类（中等）	第3类（较快）	系数一（经济发展增量系数）

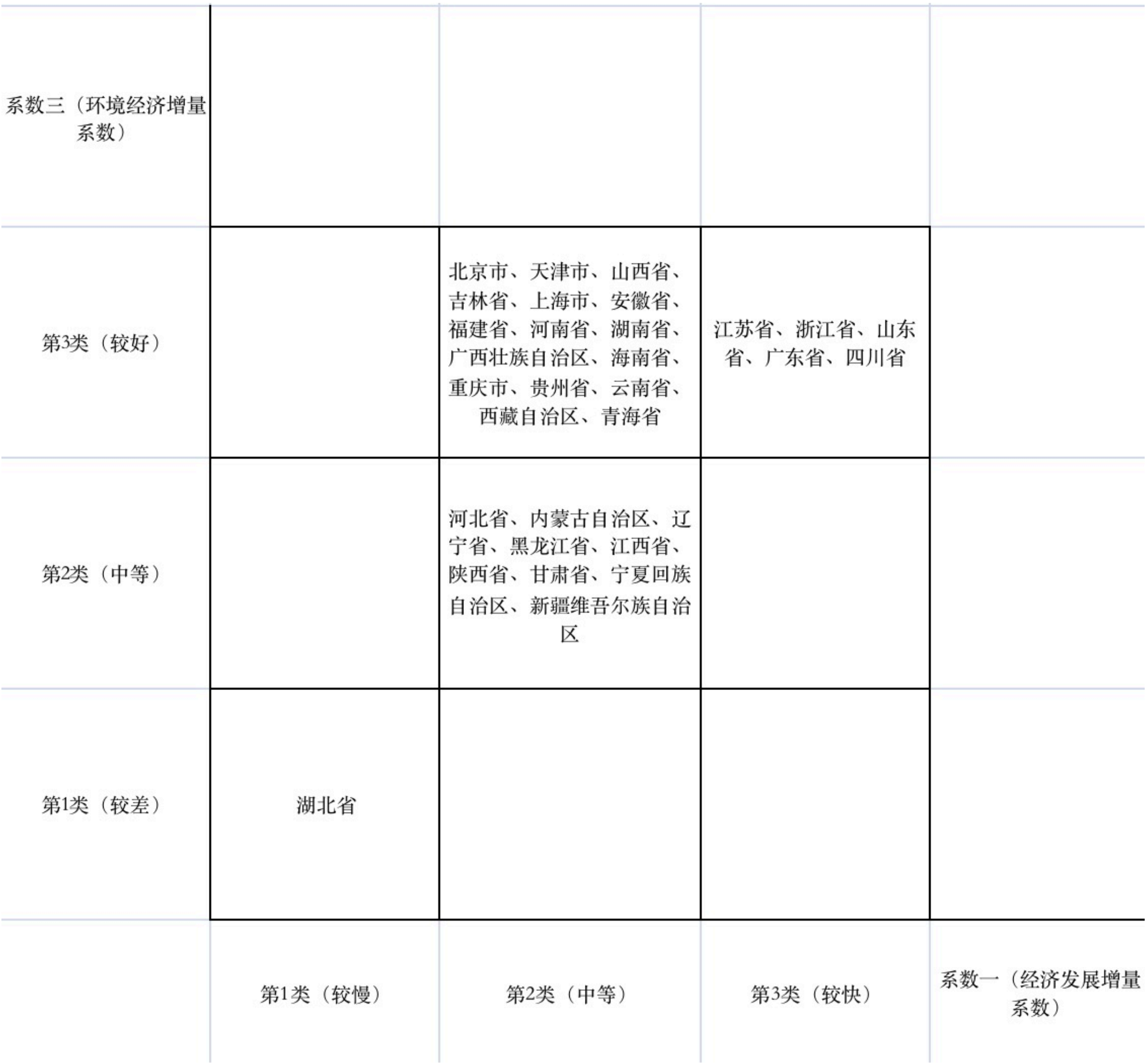
Figure 2: Joint analysis of coefficients I and II

As shown in the graph, the results of the classification of the two coefficients are divided into nine categories. The horizontal coordinates of the graph represent the incremental economic development coefficients, which are divided into category 1, category 2 and category 3 provinces. Among them, the category 1 provinces have a slower rate of economic development in 2019, which according to the graph can be seen in only one province, Hubei. Category 2 provinces have a 2019 economic development rate that is in the middle of the country, with 25 provinces, municipalities and autonomous regions. Category 3 provinces are those with the fastest economic development in 2019, with a total of five provinces. The vertical coordinates represent the incremental environmental pollution coefficients, and the provinces belonging to category 1, category 2 and category 3 from the origin to the far end of the axis are respectively the provinces belonging to this category. Among them, provinces belonging to category 1 have a small increment in incremental environmental pollutant emissions in 2019; provinces belonging to category 2 have a medium increment in incremental environmental pollutant emissions in 2019; and provinces belonging to category 3 have a large increment in incremental environmental pollutant emissions in 2019.

5.2.2 Joint analysis of the incremental coefficient of economic development and the incremental coefficient of environmental economy

The results of the classification of the two coefficients are divided into nine categories, as shown in the graph. The horizontal coordinates of the graph represent the incremental economic development coefficients, which are divided into category 1, category 2 and category 3 provinces. Of these, the 2019 economic development rate for category 1 provinces is at the slower end of the country. Category 2 provinces have a 2019 economic development rate that is in the middle of the country. Category 3 provinces' 2019 economic development rate is in the first tier nationally, faster and significantly faster than other provinces. The vertical coordinate represents the environmental economic increment coefficient, again divided into three categories, category 1, category 2 and category 3. Provinces in category 1 are in a poor position nationally in terms of environmental and economic coordination. The provinces in category 2 are in the middle of the country in terms of environmental and economic coordination. The provinces in category 3 have a good level of environmental and economic coordination nationally.

A joint analysis of the incremental economic development and incremental environmental economy coefficients is shown in Figure 3:

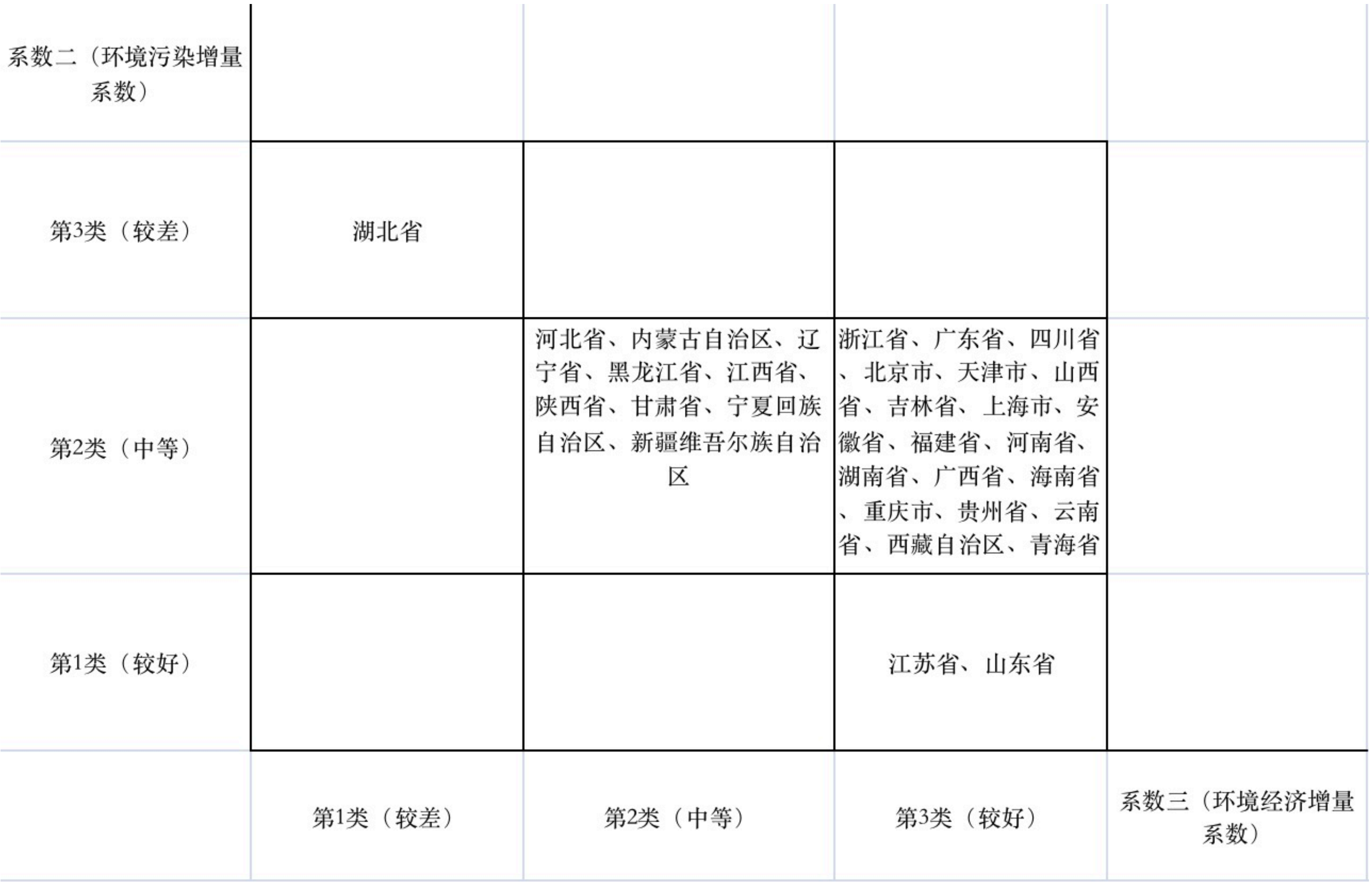


系数三（环境经济增量系数）				
第3类（较好）		北京市、天津市、山西省、吉林省、上海市、安徽省、福建省、河南省、湖南省、广西壮族自治区、海南省、重庆市、贵州省、云南省、西藏自治区、青海省	江苏省、浙江省、山东省、广东省、四川省	
第2类（中等）		河北省、内蒙古自治区、辽宁省、黑龙江省、江西省、陕西省、甘肃省、宁夏回族自治区、新疆维吾尔族自治区		
第1类（较差）	湖北省			
	第1类（较慢）	第2类（中等）	第3类（较快）	系数一（经济发展增量系数）

Figure 3: Joint analysis of coefficients I and III

5.2.3 Joint analysis of the incremental coefficient of environmental pollution and the incremental coefficient of environmental economy

A joint analysis of the incremental environmental pollution and incremental environmental economy coefficients was carried out and the joint analysis of the incremental environmental pollution and incremental environmental economy coefficients is shown in Figure 4:

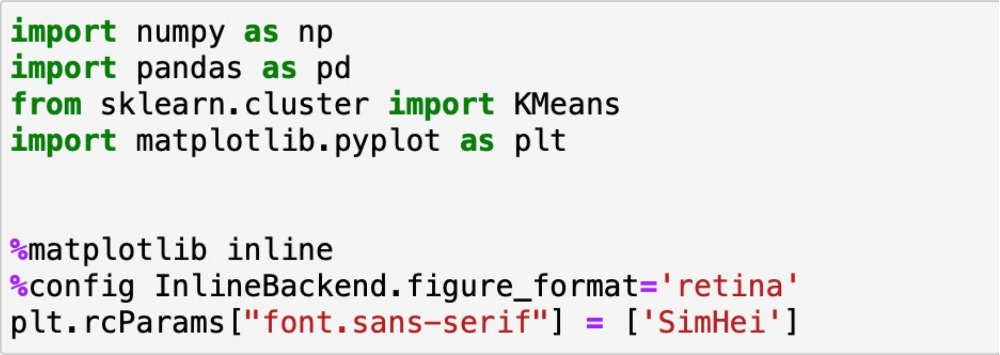


系数二（环境污染增量系数）				
第3类（较差）	湖北省			
第2类（中等）		河北省、内蒙古自治区、辽宁省、黑龙江省、江西省、陕西省、甘肃省、宁夏回族自治区、新疆维吾尔族自治区	浙江省、广东省、四川省、北京市、天津市、山西省、吉林省、上海市、安徽省、福建省、河南省、湖南省、广西省、海南省、重庆市、贵州省、云南省、西藏自治区、青海省	
第1类（较好）			江苏省、山东省	
	第1类（较差）	第2类（中等）	第3类（较好）	系数三（环境经济增量系数）

Figure 4: Joint analysis of coefficients II and III

6 Appendix

The numpy library for processing data, which supports a large number of arrays and matrices, is used to derive the differences between the two years of data through difference operations. At the same time, the results need to be labelled with the Chinese name of the corresponding province on the image. This is why the SimHei font is used here, to avoid a situation where the Chinese font cannot be displayed in the picture. The specific code is shown in Figure 5 below.

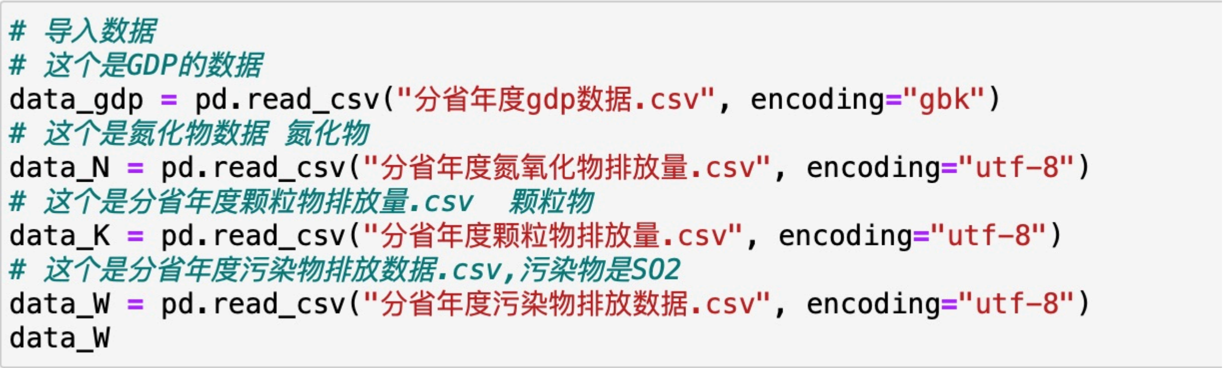


```python
import numpy as np
import pandas as pd
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt


%matplotlib inline
%config InlineBackend.figure_format='retina'
plt.rcParams["font.sans-serif"] = ['SimHei']
```

Figure 5: Using the numpy library to crunch the data

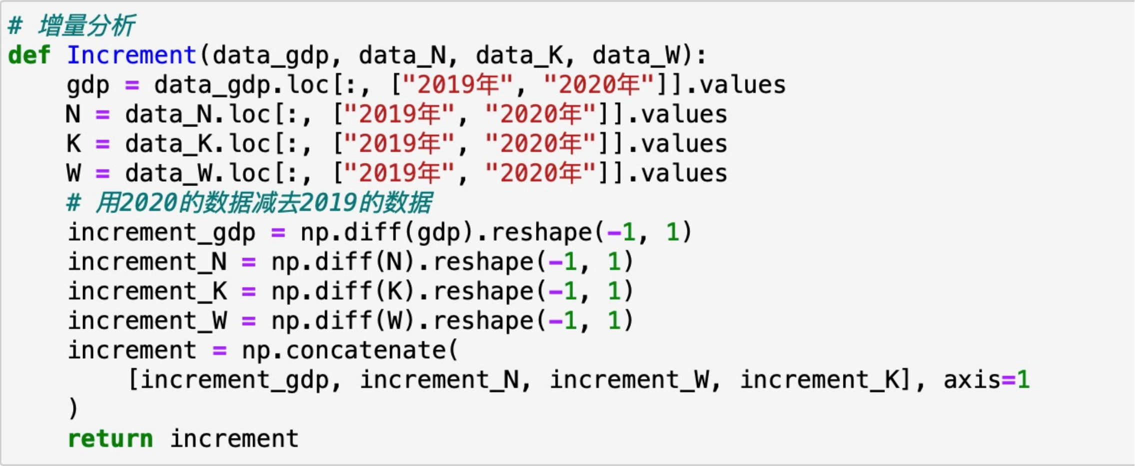
The next step was to import two years' worth of data separately, naming each of the three contaminant factors studied in this paper. In the encoding process, I had tried to use gbk or UTF-8 for all of them. However, neither worked properly, so I tried using a combination of the two and it eventually worked successfully. The code part of the data import process is shown in Figure 6.



```python
# 导入数据
# 这个是GDP的数据
data_gdp = pd.read_csv("分省年度gdp数据.csv", encoding="gbk")
# 这个是氮化物数据 氮化物
data_N = pd.read_csv("分省年度氮氧化物排放量.csv", encoding="utf-8")
# 这个是分省年度颗粒物排放量.csv  颗粒物
data_K = pd.read_csv("分省年度颗粒物排放量.csv", encoding="utf-8")
# 这个是分省年度污染物排放数据.csv,污染物是S02
data_W = pd.read_csv("分省年度污染物排放数据.csv", encoding="utf-8")
data_W
```

Figure 6: Importing data

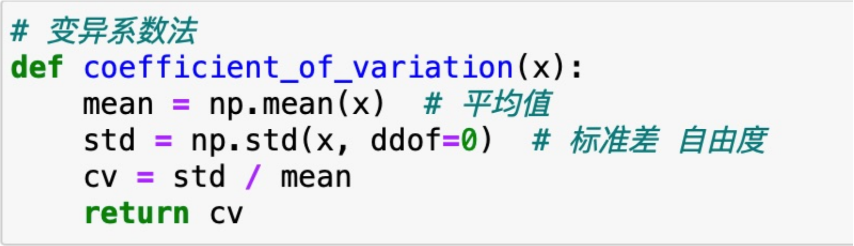
Once the data has been successfully imported, the next step is to discretize the two years of data. As the data sample is published by the National Statistics Office, the data sample contains many years of data. So here the reshape function is used to rearrange the two years of data, after which the numpy.diff function is used to calculate the difference between the two columns of data. The final step uses the concatenate function to join the differences of the three contaminants. The code for the data differencing process is shown in Figure 7.



```python
# 增量分析
def Increment(data_gdp, data_N, data_K, data_W):
    gdp = data_gdp.loc[:, ["2019年", "2020年"]].values
    N = data_N.loc[:, ["2019年", "2020年"]].values
    K = data_K.loc[:, ["2019年", "2020年"]].values
    W = data_W.loc[:, ["2019年", "2020年"]].values
    # 用2020的数据减去2019的数据
    increment_gdp = np.diff(gdp).reshape(-1, 1)
    increment_N = np.diff(N).reshape(-1, 1)
    increment_K = np.diff(K).reshape(-1, 1)
    increment_W = np.diff(W).reshape(-1, 1)
    increment = np.concatenate(
        [increment_gdp, increment_N, increment_W, increment_K], axis=1
    )
    return increment
```

Figure 7: Discrete processing

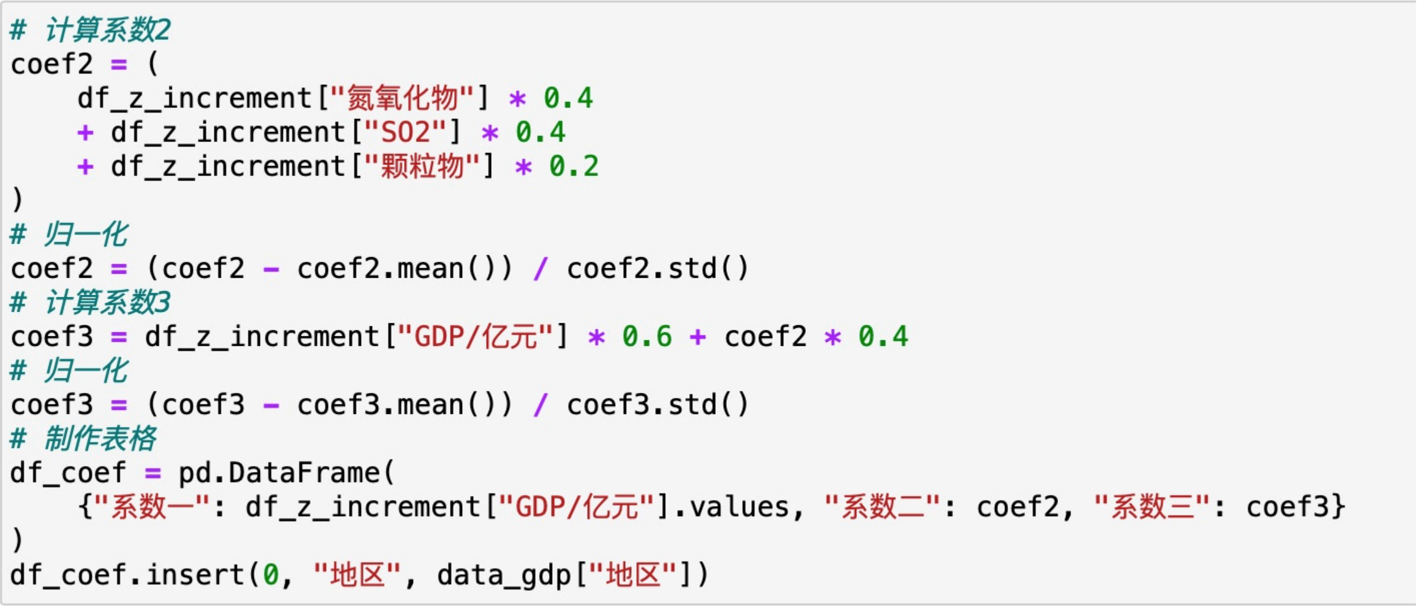
In order to determine the specific weights of the three coefficients, I learned that this could be achieved through the coefficient of variation method. The code implementation process for the coefficient of variation method is shown in Figure 8 below:



```python
# 变异系数法
def coefficient_of_variation(x):
    mean = np.mean(x)  # 平均值
    std = np.std(x, ddof=0)  # 标准差 自由度
    cv = std / mean
    return cv
```

Figure 8: Coefficient of variation method for determining weights

Combining the two formulas for the coefficient of variation, the weight of the incremental coefficient of economic development was finally determined to be 0.6 and the weight of the incremental coefficient of environmental pollution was 0.4. Based on the data of coefficient I and coefficient II and the weights assigned, data normalisation was carried out to obtain the incremental coefficient of environmental economy. The specific code implementation process is shown in Figure 9 below:



```
# 计算系数2
coef2 = (
    df_z_increment["氮氧化物"] * 0.4
    + df_z_increment["SO2"] * 0.4
    + df_z_increment["颗粒物"] * 0.2
)
# 归一化
coef2 = (coef2 - coef2.mean()) / coef2.std()
# 计算系数3
coef3 = df_z_increment["GDP/亿元"] * 0.6 + coef2 * 0.4
# 归一化
coef3 = (coef3 - coef3.mean()) / coef3.std()
# 制作表格
df_coef = pd.DataFrame(
    {"系数一": df_z_increment["GDP/亿元"].values, "系数二": coef2, "系数三": coef3}
)
df_coef.insert(0, "地区", data_gdp["地区"])
```

Figure 9: Three incremental coefficient code implementations

The trends in SO2 emissions, particulate matter emissions, incremental NOx emissions and incremental economic growth data for 31 provinces and municipalities (excluding Special Administrative Regions and Taiwan Province) from 2019 to 2020 are shown in Figure 10.

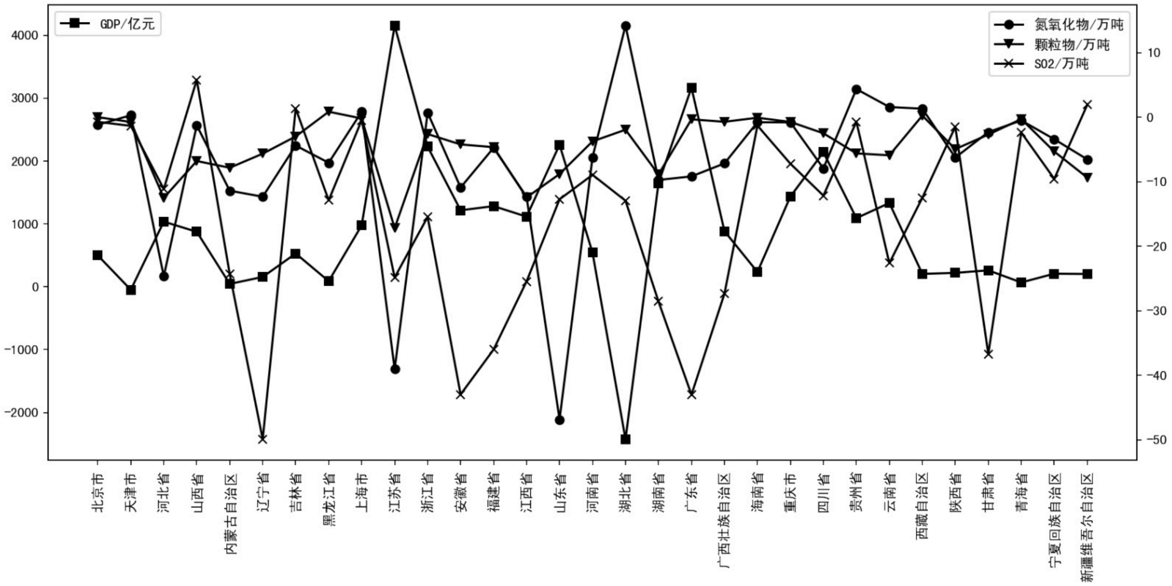


Figure 10: Incremental trends for the three pollutants

A Kmeans model was built to classify coefficient one, coefficient two and coefficient three into three categories respectively for cluster analysis.The corresponding Kmeans model is shown in Figure 11 below.

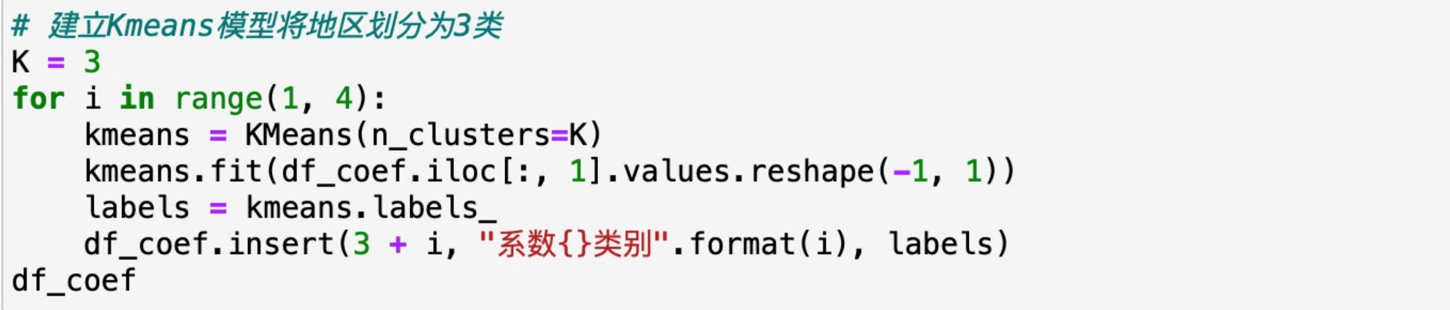
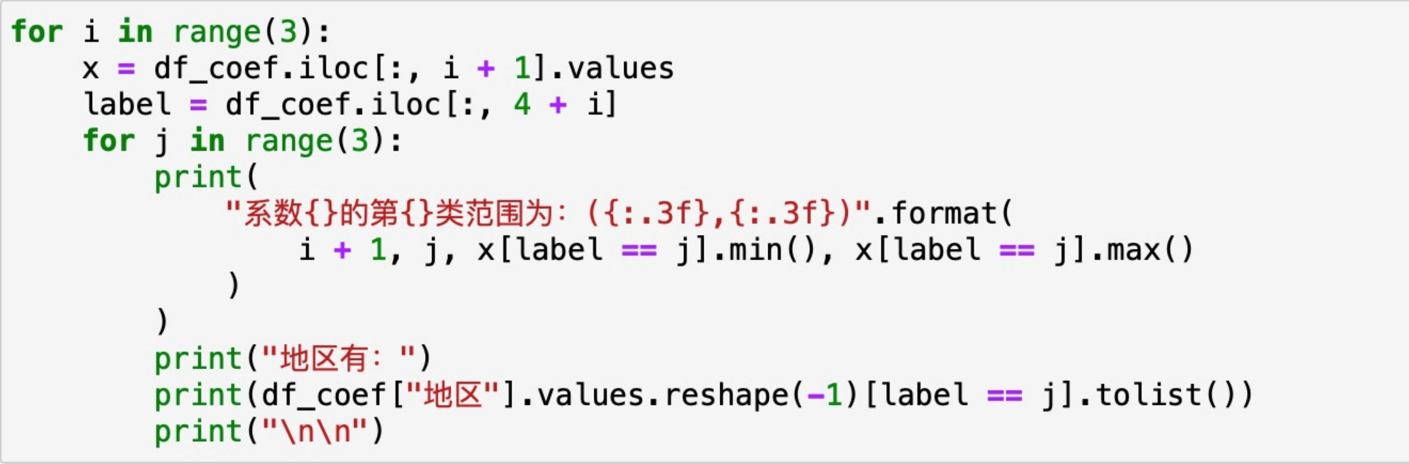

```
# 建立Kmeans模型将地区划分为3类
K = 3
for i in range(1, 4):
    kmeans = KMeans(n_clusters=K)
    kmeans.fit(df_coef.iloc[:, 1].values.reshape(-1, 1))
    labels = kmeans.labels_
    df_coef.insert(3 + i, "系数{}类别".format(i), labels)
df_coef
```

Figure 11: Kmeans model

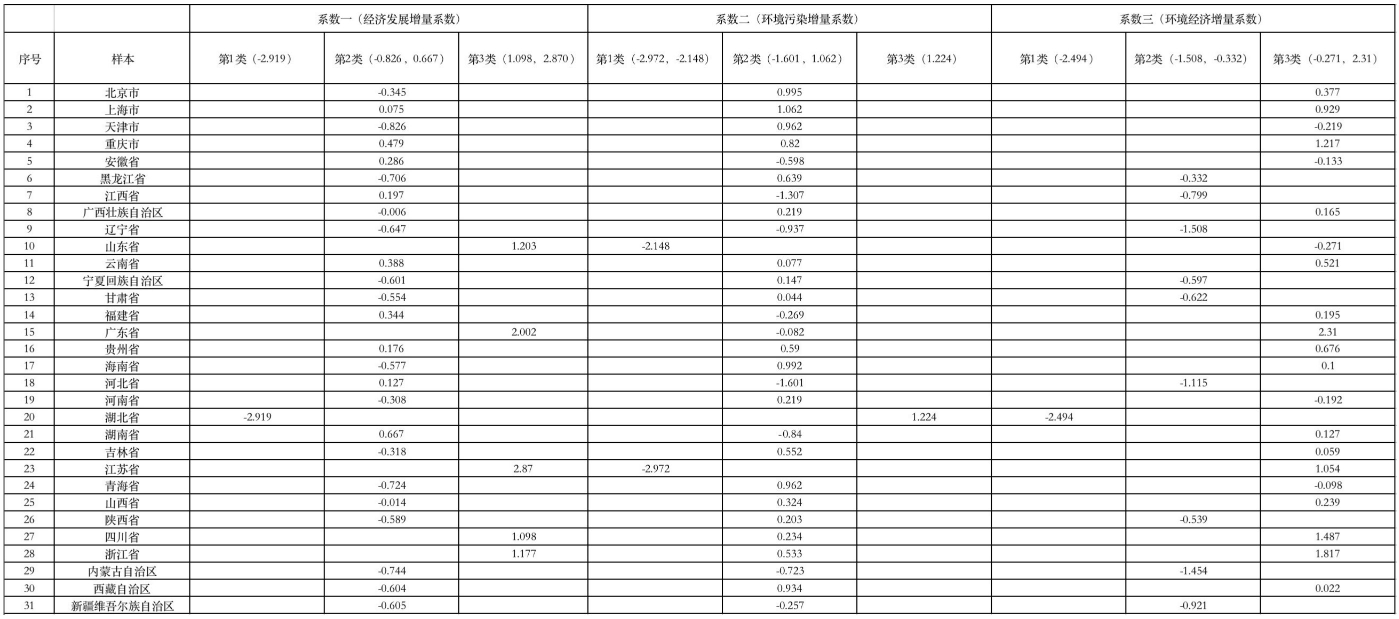
Each value is traversed by the range function to determine the range of the specific category, followed by formatted output using the format function.



```
for i in range(3):
    x = df_coef.iloc[:, i + 1].values
    label = df_coef.iloc[:, 4 + i]
    for j in range(3):
        print(
            "系数{}的第{}类范围为：({:.3f},{:.3f})".format(
                i + 1, j, x[label == j].min(), x[label == j].max()
            )
        )
        print("地区有：")
        print(df_coef["地区"].values.reshape(-1)[label == j].tolist())
        print("\n\n")
```

Figure 12: Range function

Figure 13 shows the results of the incremental clustering analysis:



		系数一（经济发展增量系数）			系数二（环境污染增量系数）			系数三（环境经济增量系数）		
序号	样本	第1类（-2.919）	第2类（-0.826，0.667）	第3类（1.098，2.870）	第1类（-2.972，-2.148）	第2类（-1.601，1.062）	第3类（1.224）	第1类（-2.494）	第2类（-1.508，-0.332）	第3类（-0.271，2.31）
1	北京市		-0.345			0.995				0.377
2	上海市		0.075			1.062				0.929
3	天津市		-0.826			0.962				-0.219
4	重庆市		0.479			0.82				1.217
5	安徽省		0.286			-0.598				-0.133
6	黑龙江省		-0.706			0.639			-0.332	
7	江西省		0.197			-1.307			-0.799	
8	广西壮族自治区		-0.006			0.219				0.165
9	辽宁省		-0.647			-0.937			-1.508	
10	山东省			1.203	-2.148					-0.271
11	云南省		0.388			0.077				0.521
12	宁夏回族自治区		-0.601			0.147			-0.597	
13	甘肃省		-0.554			0.044			-0.622	
14	福建省		0.344			-0.269				0.195
15	广东省			2.002		-0.082				2.31
16	贵州省		0.176			0.59				0.676
17	海南省		-0.577			0.992				0.1
18	河北省		0.127			-1.601			-1.115	
19	河南省		-0.308			0.219				-0.192
20	湖北省	-2.919					1.224	-2.494		
21	湖南省		0.667			-0.84				0.127
22	吉林省		-0.318			0.552				0.059
23	江苏省			2.87	-2.972					1.054
24	青海省		-0.724			0.962				-0.098
25	山西省		-0.014			0.324				0.239
26	陕西省		-0.589			0.203			-0.539	
27	四川省			1.098		0.234				1.487
28	浙江省			1.177		0.533				1.817
29	内蒙古自治区		-0.744			-0.723			-1.454	
30	西藏自治区		-0.604			0.934				0.022
31	新疆维吾尔族自治区		-0.605			-0.257			-0.921	

Figure 13: The results of the incremental clustering analysis

7 Conclusion

In this thesis, I have used the kmeans algorithm as the clustering algorithm for the study, as well as the normalisation method and the z-score method to appropriately process the data on emissions of the three pollutants as well as gdp for the 31 provinces and cities across China from 2019 to 2020. The coefficient of variation method was also applied to derive the three incremental coefficients after assigning appropriate weights. The three coefficients were analysed separately and jointly to produce the results of this study.

Finally I will classify the results of the analysis of these 31 provinces, cities and autonomous regions again, this time into four categories. ”fully in line with the scientific concept of development”, ”rapid economic growth with good environmental protection”, ”all aspects of economic growth and environmental protection need to be improved” and The four categories are ”major problems with the development model”.

Fully consistent with the concept of scientific development: Jiangsu Province, Shandong Province

Fast economic growth with good environmental protection: Zhejiang Province, Guangdong Province, Sichuan Province

Both economic growth and environmental protection need to be improved: Beijing, Tianjin, Hebei, Shanxi, Inner Mongolia, Liaoning, Jilin, Heilongjiang, Shanghai, Anhui, Fujian, Jiangxi, Henan, Hunan, Guangxi Zhuang Autonomous Region, Hainan, Chongqing, Guizhou, Yunnan, Tibet Autonomous Region, Shaanxi, Gansu, Qinghai, Ningxia Hui Autonomous Region, Xinjiang Uygur Autonomous Region Xinjiang Uyghur Autonomous Region

Significant problems with development patterns: Hubei Province

8 Reference and Bibliography

<1> Jain AK, Dubes RC.Algorithms for Clustering Data.Prentice-Hall Advanced Reference Series, 1988.1-334.

<2> National Bureau of Statistics Releases Report on Economic and Social Development Achievements in the 40 Years of Reform and Opening Up http://www.gov.cn/xinwen/2018-08/29/content5317294.html

<3> Timeline of WHO's response to the COVID-19 outbreak
https://www.who.int/zh/news/item/29-06-2020-covidtimeline

<4> Wuhan City Closure
https://baike.baidu.com/Wuhan City Closure /item/54212728

<5>Li Qinghua, Deng Pingping, Song Qin. An empirical analysis of the relationship between environmental pollution and economic growth in China Analysis, Resource Development and Markets, 2011 (2): 131 - 134.

<6>Feng Genfu, Jiang Wending, Huang Jianshan. Spatial Distribution of Economic Development and Environmental Pollution in China An empirical test of the relationship between economic development and environmental pollution in China, Contemporary Economic Science, 2011, 33 (6): 72 - 80.

<7>Zhao Jiahong, Dong Xiaolin, Wu Yang, Zhao Lijuan. Total Wastewater Discharge and Pollutant ConcentrationJoint Cluster Analysis of Total Wastewater Discharge and Pollutant Concentration — Example of 2014 Provincial Panel Data, Sichuan Environment, 2017, 36 (3): 66 - 73.

<8>Ding JH, Nian Y. An analysis of the relationship between economic growth and environmental pollution — Jiangsu Province as an example Example, Nankai University Economic Research, 2010 (2): 64 - 79.

<9>Wu YP, Dong Zuocheng, Song Jianfeng. Economic Growth and Environmental Pollution Levels in Beijing Quantitative Model Study, Geographical Studies, 2002 (2): 239 - 246.

<10> Clustering Algorithms Research
School of Computer Science and Technology, Jilin University;Key Laboratory of Symbolic Computing and Knowledge Engineering, Ministry of Education, Changchun 130012, Jilin, China

<11>Marques JP, Written;Wu YF, Trans.Pattern Recognition Concepts, Methods and Applications.2nd ed., Beijing:Tsinghua University Press, 2002.51-74 (in Chinese) .

<12> Gelbard R, Goldman O, Spiegler I.Investigating diversity of clustering methods:An empirical comparison.Data&Knowledge Engineering, 2007, 63 (1) :155-166.

<13> Kumar P, Krishna PR, Bapi RS, De SK.Rough clustering of sequential data.Data&Knowledge Engineering, 2007, 3 (2) :183-199.

<14> Huang Z.A fast clustering algorithm to cluster very large categorical data sets in data mining.In:Proc.of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery.Tucson,1997.146-151.http://www.informatik.uni-trier.de/ ley/db/conf/sigmod/sigmod97.html

<15>Zhao YC, Song J.GDILC:A grid-based density isoline clustering algorithm.In:Zhong YX, Cui S, Yang Y, eds.Proc.of the Internet Conf.on Info-Net.Beijing:IEEE Press, 2001.140-145.http://ieeexplore.ieee.org/iel5/7719/21161/00982709.pdf

<16> Karypis G, Han EH, Kumar V.CHANELEON:A hierarchical clustering algorithm using dynamic modeling.IEEE Computer, 1999, 2 (8) :68-75.

<17> Python data normalization common methods
https://cloud.tencent.com/developer/article/

<18>Sun Hongwei, Lv Chunyan, Qi Aiqin, et al. A study of the principles of data standardisation in integrated evaluation, China Health Statistics, 2015 (2) :342-344, 349.

<19>Coefficient of variation method for python
https://blog.csdn.net/qq_25990967/article/details/122801244

<20>ZHAO Li-juan,DONG Xiao-lin,WU Yang.Joint analysis of environmental pollution and economic development based on Z-score model. Key Laboratory of Ministry of Education of the Ecological Effect and Groundwater in Arid Areas, Chang'an University
Applied Chemicals 2017,46(09),1805-1809 DOI:10.16581/j.cnki.issn1671-3206.20170714.038

9 Acknowledgement

I have received a lot of enthusiastic help from my teachers and classmates from the selection of the topic and data collection to the completion of my thesis.

First of all, I would like to thank my supervisor, Ms. Yang Lei, for her valuable advice on the completion of my dissertation, which has given me a clear goal and direction in writing my dissertation. Ms Yang's rigorous attitude, knowledge, patience and charm have had a profound impact on me.

Secondly, I would like to thank all the teachers who have taught me and who have inspired me with their kindness and guidance. I would like to thank them all for helping me on my journey.

Finally I would like to thank my classmates and friends who have been by my side and accompanied me throughout the four years. I would like to thank them for their help in the process of writing my dissertation, which has enabled me to complete it successfully. I would like to thank them for the happy memories they have given me in my life. I would also like to thank my parents for nurturing and helping me over the past 21 years. I thank them for the support and help they have provided in my life, for the encouragement they have given me when I was in trouble, and for the careful teaching they have given me since I was a child.

My four years at MUST are coming to a close, but I still have regrets about the decisions I made in the past, about the people and things I missed, and about the hurt I caused others. But perhaps that is the essence of life, the best is when it is not yet complete. Although my undergraduate years were only four years, I believe that everything and everyone I experienced during those four years will be part of my life memories. With less than a month to go before I leave MUST, I will be embarking on a new journey. I would like to wish all the teachers I met at the university and my friends good health and all the best for the future.