Quantitative analysis on the sustainable development of four municipalities in China

Q R Shen[1], H Tian[1], X Y Han[1], H Zhang[2,*] and W Sun[1,*]

1 Key Lab of Membrane Science and Technology, College of Chemical Engineering, Beijing University of Chemical Technology, Beijing 100029, China
2 School of Chemistry and Chemical Engineering, Southwest University, Chongqing, 400715, China
E-mail: sunwei@mail.buct.edu.cn

Abstract. Due to the geographical location, development history and many other social-economic factors, the economic development of each city displays significant difference from each other. It is hard to characterize the economic development of a city by only looking at one or two measurements, which is only one aspect of a city. A snap shot of a city with all of its data information could provide its holographic image. With its evolution along time, the image can be even more impressive, but is still hard to compare with its peer cities given its multivariate nature. From system point of view, no matter how complicate the appearance of a system is, it will be always associated to its intrinsic characteristics. In this work, data from four municipalities in China are analysed quantitatively to discuss different development patterns by Pearson Correlation Coefficient and Hierarchical Clustering Analysis methods, correlation of factors and urban development patterns are clearly obtained and explained.

1. Introduction

Sustainability usually refers to the balance among economic development, social development and environment protection within a geographic region, especially the interaction between economic development and environment protection attract more attention along with the rapid urbanization in China. At the beginning, economic progress dominated while the environment cost was neglected. As the negative effect of environment became significant, such as poor air quality, people start to pay attention on the development planning with the consideration of environment constraints. During the urbanization, industry development is almost equivalent to the economic development, which gets even truer when industry is generalized to include agriculture as the first category of industry. Gross Domestic Product, GDP, is usually the measure of economic development. As the outcome of a city development, GDP is usually related to the energy cost, the composition of industry, the workforce, sometimes, the total population, while it is also associated to environment cost, most importantly, air pollution. Due to different development history and geographic condition, it is hard to evaluate the development of a city by a unique index, but the quantitative analysis still can reveal certain characteristics for a city when more data are available from various sources.

Data analysis methods are well-developed in commercial and engineering applications, among which Pearson correlation (PC) is a straightforward method to quantifying the linear relation of pair variables [1]. Zhou et al proposed an enhanced particle filter based on PCC to compare the difference of observation path and true path of particles and further provided validation that the degeneracy and sample disadvantage can be solved [2]. Xia et al evaluated the performance of thermal power unit by successfully applying PCC to assess analog signal [3]. Liu et al used PCC to identify the main factors

that influence willingness to buy low-carbon products [4]. However, the correlation can only be investigated in each pair variables by PCC, while all variables are inevitably correlated together respect to a city development. In order to extract the correlation among multiple variables, clustering analysis is another straightforward choice. According to availability of object attributes, there are two clustering methods in practice, supervised clustering and unsupervised clustering, typical represented by k-means and hierarchical clustering analysis, HCA. Once the attribute of data is known, the data can be properly labeled, so called labeled data, and vice versa. Given the data from different cities, no priori information can be used to supervise the clustering, HCA is more suitable to analyse the data without labeling. Beaver et al analysed weather and atmospheric data by HCA, and successfully obtained the correlation between weather pattern and local air quality [5]. Ma et al classified nine districts of Beijing into three groups by HCA based on economy index such as industrial and agricultural income [6].

In this work, PCC and HCA analysis are applied to multiple data sets regarding economic development and environment conditions from four municipalities in China. The paper is organized as follows: the data is described in Section 2; the methods are shown in Section 3; the results are presented and analysed in the Section 4; the conclusion is given in Section 5.

2. Data

2.1. Data source

The data are obtained from the web of National Bureau of Statistics of China (http://data.stats.gov.cn/).

2.2. Details of data

The data considered in this work in 2012-2016 are listed in Table 1, including GDP, Urban population density, pollutant emissions, etc., from Beijing, Shanghai, Tianjin and Chongqing.

Table 1. Data information.

Variables	Description	Unit
A	GDP	USD
B	Fossil energy consumption	Ton
C	Electric power consumption	Kw ḥ
D	Urban population density	per square kilometer
E	Sulfur dioxide emissions	Ton
F	Nitrogen oxide emissions	Ton
G	Soot emissions	Ton
H	Number of vehicles	/

The definition of eight variables in the Table 1 was described as follow:

A- Gross Domestic Product (GDP) refers to the sum of all the final goods and services produced in a period of time. GDP estimates are commonly used to determine the economic performance of a country or region.

B- Fossil energy consumption refers to the amount of energy consumed, including coal, Coke, crude oil, gasoline, kerosene, fuel oil, natural gas.

C- Electric power consumption refers to the total annual electricity consumption in the region.

D- Urban population density refers to the density of population in urban areas which is calculated as the sum of urban population and urban temporary population divided by urban area.

E- Sulfur dioxide emissions refer to the sum of industrial SO_2 emissions and domestic SO_2 emissions over a period of time.

F- Nitrogen oxide emissions refer to the total NOx mass discharged into the atmosphere by the enterprise during fuel combustion and production processes over a period of time.

G- Soot emissions refer to the total mass of smoke, dust and industrial dust discharged into the atmosphere by the enterprise during fuel combustion and production process over a period of time.

The soot or industrial dust emission can be obtained by multiplying the exhaust air volume of the dust removal system and the dust concentration at the exit of the dust removal equipment.
H- Number of vehicles refers to the vehicles registered with the public security traffic administration department over a period of time and licensed as vehicles in certain area.

3. Methodology

Pearson correlation coefficient (PCC) and Hierarchical clustering analysis (HCA) are well applied in practice, and employed in this work. They are briefly introduced in this section.

3.1. Pearson correlation coefficient (PCC)

In statistics, the Pearson correlation coefficient is a measurement of the linear correlation between two variables *X* and *Y*, which was proposed by Karl Pearson [7].
Given two sample sets, *X* and *Y* with the same set sizes, Pearson correlation coefficient is defined as:

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{1}$$

where ρ refers to the Pearson correlation coefficient, cov is the covariance of *X* and *Y*, σ_X is the standard deviation of *X*, and σ_Y is the standard deviation of Y.
For one dataset $X=\{x_1, \ldots, x_n\}$ containing n values and another dataset $Y=\{y_1, \ldots, y_n\}$containing n values, r is calculated as:

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

where *n* is the sample size, x_i and y_i are the individual sample points indexed with *i*, and $\bar{x}, \bar{y}$ are the sample mean. The value of *r* is between −1 and +1, where −1 represents total negative linear correlation, 0 means no linear correlation, and 1 is corresponding to total positive linear correlation.
By PCC, the correlation degree between pair variables can be calculated and the corresponding economic-environment interaction can be investigated.

3.2. Hierarchical clustering analysis (HCA)

In the method of hierarchical clustering analysis, clusters are represented through a dendrogram with node layers, where each node represents a cluster. There are two main approaches of hierarchical clustering, agglomerative and divisive, which are shown in Figure 1. For the agglomerative hierarchical clustering, each sample is initially considered as one cluster, and subsequently pairs of cluster are merged, while in divisive hierarchical clustering it starts with one cluster that includes all samples, and then recursive splits are performed. In this work, the agglomerative cluster approach is used. The most common type of hierarchical clustering algorithm proceeds as follows:
(1) Treat every single node as one cluster and calculate each pair of nodes by using Euclidean distance, which is shown as follows:

$$\mathrm{dist}(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{3}$$

where x_i is the variable in cluster *X*, y_i is the same variable in cluster *Y*, and *n* is the total number of variables.
(2) Find the smallest value (closest Euclidean distance) from above and agglomerate them as a new cluster.
(3) Calculate Euclidean distance of the pairs of this new cluster from (2) and other clusters.
(4) Repeat (2) and (3) until all clusters are agglomerated as one.

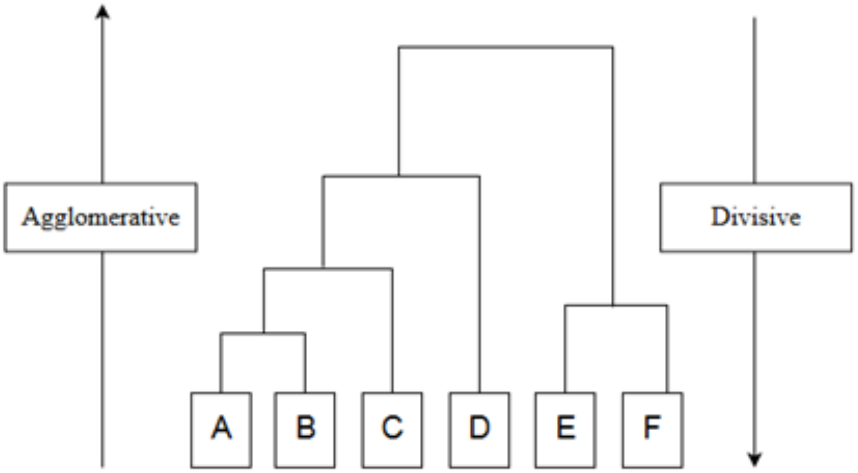

Figure 1. Hierarchical cluster analysis.

For hierarchical clustering analysis in Figure 1, there are six samples from A to F. A and B are merged to form a cluster while E and F to form another one. Then the cluster of A and B are combined with C to form a cluster and D is added to form a new cluster. Finally, this cluster united with the cluster of E and F to be one cluster.

The HCA algorithm evaluates the similarity of two clusters by calculating the Euclidean distance between them. The smaller the distance, the higher the similarity. A hierarchical clustering tree is formed by combining the two most similar clusters and repeat the process. In a clustering tree, the original data points of different categories are the lowest layer of the tree and similar clusters are combined to form upper layers. Based on HCA result, the similarity between two clusters is identified, which can be applied to the analysis of city development data to compare the pattern of each city.

4. Results

4.1. Pearson correlation coefficient

Table 2. The correlation coefficients among different variables of Beijing.

r	a	b	c	d	e	f	g	H
a	1	-0.9866	0.9894	-0.6490	-0.9485	-0.9807	-0.9820	0.9681
b		1	-0.9701	0.6191	0.9236	0.9567	0.9789	-0.9524
c			1	-0.7263	-0.9724	-0.9890	-0.9877	0.9601
d				1	0.8558	0.7806	0.7631	-0.5143
e					1	0.9912	0.9818	-0.8688
f						1	0.9914	-0.9163
g							1	-0.9208
h								1

As shown in Table 2, the Pearson correlation test of Beijing data shows negative correlations in 15 groups of variables, among which, the correlation coefficients of 12 groups of variables, such as GDP and fossil energy consumption, GDP and soot emissions, electric power consumption and sulfur dioxide emissions are close to 1. On the other side, there are 13 groups of variables with positive correlation, among which, 10 groups of variables such as GDP and electric power consumption, fossil energy consumption and sulfur dioxide emissions, urban population density and number of vehicles are strongly correlated.

It is easy to understand that GDP and number of vehicles, Electric power consumption and Urban population density are positively correlated, but some result of Pearson correlation test in Table2

needs more investigation for an acceptable explanation. In Beijing, negative correlation exists between GDP and fossil energy consumption, Urban population density, pollutant emission. The possible reason could be that the increase of GDP is not contributed by the industry with fossil fuel as raw material. According to the statistics data, the economic identities of Beijing are mainly government service, finance, education, IT and other tertiary industries, in which electricity is the major form of energy consumption. In recent years, Beijing has strictly controlled population and the high-tech technology is developed faster. Driven by the high-tech technology and finance, GDP growth is increasing, leading Urban population density and GDP negatively correlated.

Table 3. The correlation coefficients among different variables of Shanghai.

r	a	b	c	d	e	f	g	h
a	1	-0.6904	0.8519	0.6480	-0.9631	-0.9771	0.0358	0.9985
b		1	-0.2193	-0.5119	0.6038	0.6356	-0.6525	-0.6834
c			1	0.4825	-0.8914	-0.8817	-0.4540	0.8485
d				1	-0.4948	-0.5229	0.4253	0.6456
e					1	0.9981	0.1570	-0.9515
f						1	0.1091	-0.9678
g							1	0.0420
h								1

As shown in Table 3, the Pearson correlation test of Shanghai data shows negative correlations in 14 groups of variables, among which, the correlation coefficients of 10 groups of variables, such as GDP and fossil energy consumption, urban population density and sulfur dioxide emissions, electric power consumption and nitrogen oxide emissions are close to 1. On the other side, there are 14 groups of variables with positive correlation, among which, 10 groups of variables such as GDP and electric power consumption, fossil energy consumption and sulfur dioxide emissions, nitrogen oxide emissions and soot emissions are strongly correlated.

Similarly, it's straightforward that GDP and electric power consumption, urban population density and the number of vehicles are positively correlated, but it is noticeable that there is a negative correlation between GDP and pollutants such as sulfur dioxide emissions and nitrogen oxide emissions. It could be interpreted that an inverted U-shape relation exists between economic development and environment pollutants. When economy develops, the environment firstly worsens and then becomes good, which is also called Environmental Kuznet Curve [8]. As Shanghai is a highly developed region, regulations and techniques regarding pollution control are also upgraded along with its economic development, therefore, it is not too difficult to explain why sulfur dioxide and nitrogen oxide emissions have an opposite trend with GDP changes.

Table 4. The correlation coefficients among different variables of Tianjin.

r	a	b	c	d	e	f	g	h
a	1	-0.8908	0.9338	0.9614	-0.8258	-0.9327	0.0651	0.8446
b		1	-0.6727	-0.9119	0.9190	0.9621	0.1981	-0.5124
c			1	0.8676	-0.6021	-0.7546	0.2815	0.9800
d				1	-0.7585	-0.8851	0.1833	0.7646
e					1	0.9718	0.3814	-0.4495
f						1	0.2185	-0.6161
g							1	0.3887
h								1

Pearson correlation test of Tianjin data is shown in Table 4, in which one can find negative correlations in 12 groups of variables, among which, the correlation coefficients of 5 groups of variables, such as GDP and fossil energy consumption, GDP and Nitrogen oxide emissions, electric power consumption and number of Vehicles are close to 1. On the other side, there were 16 groups of variables with positive correlation, among which, 8 groups of variables such as GDP and urban

population density, fossil energy consumption and polluted gas emissions, urban population density and number of vehicles are strongly correlated.

Table 5. The correlation coefficients among different variables of Chongqing.

r	a	b	c	d	e	f	g	h
a	1	-0.9419	0.8782	0.9902	-0.8862	-0.9282	-0.5490	0.9991
b		1	-0.6727	-0.9737	0.9168	0.9389	0.6674	-0.9465
c			1	0.8046	-0.6179	-0.6853	-0.1843	0.8749
d				1	-0.9324	-0.9640	-0.6458	0.9888
e					1	0.9918	0.8651	-0.8739
f						1	0.8204	-0.9182
g							1	-0.5286
h								1

As shown in Table 5, the Pearson correlation test of Chongqing data shows a negative correlation in 16 groups of variables, among which, the correlation coefficients of 9 groups of variables, such as GDP and fossil energy consumption, urban population density and nitrogen oxide emissions, sulfur dioxide emissions and number of vehicles are close to 1. On the other side, there were 12 groups of variables with positive correlation, among which, 11 groups of variables such as GDP and urban population density, fossil energy consumption and nitrogen oxide emissions, urban population density and number of vehicles are strongly correlated.
Table 4 and 5 show that it's clear Tianjin and Chongqing are quite similar in correlation results, especially in the groups of electric power consumption and sulfur dioxide emissions, nitrogen oxide emissions and soot emissions. Correlation coefficients of the three pairs of variables obtained in Tianjin and Chongqing are larger than those in Beijing. This can be perfectly explained with the different industrial structures. In an industrial structure with high percent of tertiary industry like Beijing, electric power accounts the most of the energy consumption so that the need for heavy industry is comparatively low, as a result, electric power is highly negatively correlated with pollutants emissions, while in a relatively low percent of secondary and tertiary industry like Tianjin and Chongqing, the need for heavy industry is much more which results in a smaller absolute value in correlation with electric power consumption and pollutants emissions, i.e., less correlated.

4.2. Hierarchical clustering analysis

The hierarchical clustering result about the development of the four municipalities in 2012-2016 is shown in Figure 2. The information of each sample of the clustering result is shown in Table 6. It can be seen in Figure 2 that the development of Chongqing in 2012-2016 is different from other cities, the development of Shanghai in the recent years has been kept at a steady state, while that of Beijing and Tianjin in changes since 2016. Combined with the industrial ratio variation diagram (Figure 3-Figure 6), the secondary industry of Chongqing accounts for an important part, while the tertiary industry takes a leading role in Beijing and Shanghai. In 2016, tertiary industry of Beijing contributed more than 80 percent for the first time, which may be the reason that it is different from others, while the tertiary industry of Tianjin in 2016 grows greatly compared to that in 2012-2015.The development of Shanghai was steady from 2012 to 2016, with the secondary industry gradually decreasing and the tertiary industry gradually rising. The result above shows that the development of the secondary industry is still the pillar industry, but experience a slow decrease, especially in Beijing and Shanghai. The increase of tertiary industry is more independent of the fossil energy, on the other side, requires significant amount of electricity, which is free of local emission.

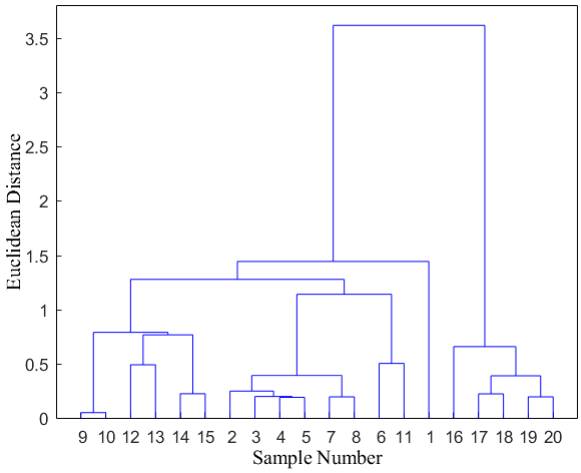

Figure 2. Hierarchical clustering analysis results.

Table 6. Sample description.

Sample	Description	Sample	Description
1	Beijing 2016	11	Tianjin 2016
2	Beijing 2015	12	Tianjin 2015
3	Beijing 2014	13	Tianjin 2014
4	Beijing 2013	14	Tianjin 2013
5	Beijing 2012	15	Tianjin 2012
6	Shanghai 2016	16	Chongqing 2016
7	Shanghai 2015	17	Chongqing 2015
8	Shanghai 2014	18	Chongqing 2014
9	Shanghai 2013	19	Chongqing 2013
10	Shanghai 2012	20	Chongqing 2012

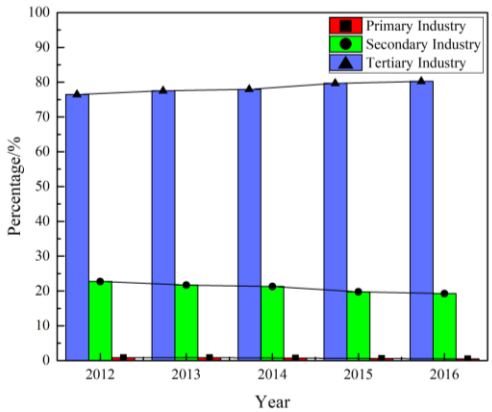

Figure 3. Industrial structure of Beijing.

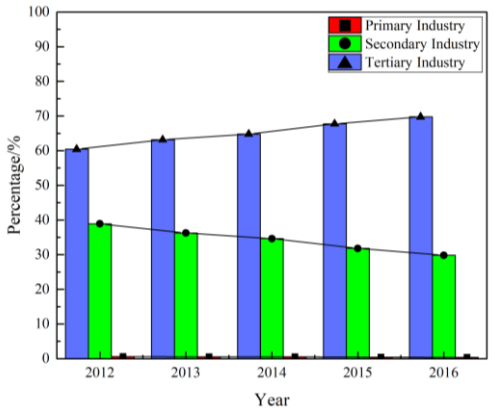

Figure 4. Industrial structure of Shanghai.

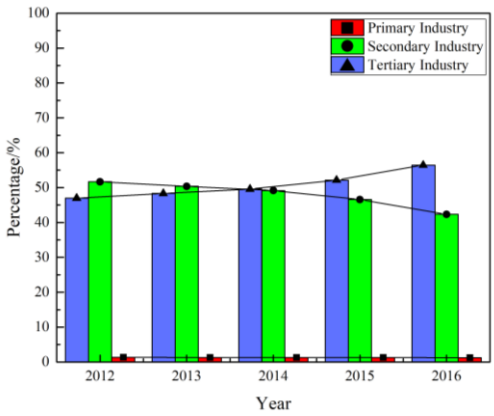

Figure 5. Industrial structure of Tianjin.

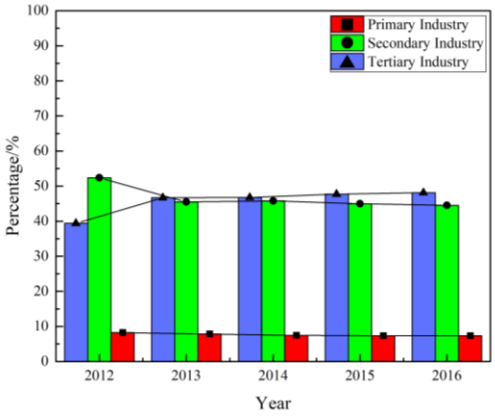

Figure 6. Industrial structure of Chongqing.

Primary and secondary industries are major part in industry structure but is confined by environment capacity because they rely heavily on resource supply and inevitably discharge to environment, while tertiary industry are mainly services, education, IT, etc. in which research and development are strongly motivated. Without any doubt, the tertiary industry provides technical support to the sustainable development of urban economy with no expense in environment. The tertiary industry is beneficial to coordinate industry development, improve the efficiency of resource allocation, and guarantee stability and economic development, in terms of society economy and environment. Due to the high pressure of metropolis such as overpopulation and overconsumption and limited natural resources, the best way for future development is to increase the proportion of tertiary industry, which is highly agree with the idea of sustainability.

5. Conclusion

In this work, the PCC and HCA are used to analyse the data of four municipalities from 2012 to 2016, and correlation within a city and urban development pattern were obtained respectively. These four cities (Beijing, Shanghai, Tianjin and Chongqing) show quite different development patterns in recent years, high-tech development of Beijing is more comprehensive. Shanghai has been kept a steady development with an increasing tertiary industry, while the secondary industry still play an important role. In Tianjin and Chongqing, the secondary industry still dominates, but their tertiary industry started to increase slightly. With the limitation of fossil fuel supply and the restriction of air emission, tertiary industry is becoming a better choice for the sustainable development in these municipal cities.

6. References

[1] Joseph L R and W A Nicewander 1988 Thirteen Ways to Look at the Correlation Coefficient *The American Statistician* **42(1)** 59-66
[2] Zhou H M, Deng Z H, Xia Y Q and Fu M Y 2016 A new sampling method in particle filter based on Pearson correlation coefficient *Neurocomputing* **216** 208-215
[3] Zhi X, Song Y X, Ma J, Zhou L J and Dong Z J 2017 Research on the Pearson correlation coefficient evaluation method of analog signal in the process of unit peak load regulation *In Electronic Measurement & Instruments (ICEMI) 13th IEEE International Conference on* pp 522-527
[4] Li Q W, Long R Y and Chen H 2017 Empirical study of the willingness of consumers to purchase low-carbon products by considering carbon labels: A case study *Journal of Cleaner Production* 161
[5] Beaver S and Palazoğlu A 2006 A cluster aggregation scheme for ozone episode selection in the San Francisco, CA Bay Area *Atmospheric Environment* **40(4)** 713-725
[6] Ma G, Li H X and Luo K 2005 Application of clustering in regional economy *International Conference on Electronic Commerce* pp 48-51

[7] Pearson K 2006 Note on Regression and Inheritance in the Case of Two Parents *Proceedings of the Royal Society of London* **58** 240-242
[8] Tajul Ariffin Masron and Yogeeswari Subramaniam 2018 The environmental Kuznets curve in the presence of corruption in developing countries *Environmental Science and Pollution Research* **25(13)** 12491-12506