International Journal of Finance and Investment
ISSN: 3078-6002 (Print), ISSN: 3078-2295 (Online) | Vol. 2, No. 3, 2025

Theoretical Measurement and Analysis of Chinese Data Factor Development Index

Zi Huang

School of Business, Hunan University of Science and Technology, Xiangtan 411100, China

Abstract: Data factors are the basic resources, strategic resources and important productivity in the digital economy era. This paper combines two methods of index system and text analysis, and constructs data factor measurement indicators from the provincial macro level and the enterprise micro level. This study found that from 2011 to 2020, the level of development of data factors generally showed a continuous upward trend, of which the upward trend was particularly obvious from 2016 to 2018. Judging from the development situation of various regions, there are still obvious differences in the development levels of data factors in various regions of China. The development level of data factors in the eastern region is the highest and far higher than the national average. The development trend of data factors in the central and western regions is almost the same, and the development level of data factors in the northeast region is the lowest. Judging from the differences in the manufacturing industry, advanced manufacturing industries with high technology intensiveness such as computer, communications and other electronic equipment manufacturing industries, electrical machinery and equipment manufacturing industries, and automobile manufacturing industries have the highest comprehensive index of data factors. The research in this article provides a reference for better evaluating the effectiveness and characteristics of data factors and for the government to formulate relevant policies.

Keywords: Data factor; Text Analysis; Regional Analysis.

1. Introduction

Data factors are the basic resources, strategic resources and important productivity in the digital economy era. In 2019, the Fourth Plenary Session of the 19th Central Committee of the Communist Party of China listed data as a production factor for the first time. In 2020, the "Opinions on Building a More Perfect System and Mechanism for Market-based Allocation of Factors" officially proposed to take data as one of the five major production factors, the data factors were elevated to a national-level strategy, and the top-level design of systematized data factors was officially launched. Driven by this series of policy documents, the scale of data factors has increased significantly. Data shows that China's data production in 2022 reached 8.1ZB, a year-on-year increase of 22.7%, accounting for 10.5% of the world, ranking second in the world. In addition, the contribution rate of data factors to GDP growth that year increased from 12.25% in 2016 to 14.7% in 2021, and data factors play an increasingly powerful role in promoting economic growth. However, at present, there are no authoritative indicators widely recognized in the academic community for measuring data factors. What kind of cutting-edge scientific data and methods to more accurately and comprehensively measure data factors is an important issue worthy of study. By scientifically measuring the development level of data factors at a more micro level, we can better evaluate the development performance and characteristics of data factors, and provide a reference for the government to formulate relevant policies.

2. Literature Review

At present, the academic community's measurement indicators for data factors mainly include two methods: building a relevant indicator system and text analysis. Specifically: On the one hand, the measurement indicators for building an indicator system include the release of comprehensive data factors by some domestic and foreign research institutions, such as the national-level data factor indicators released by the IMD World Competitiveness Center in Switzerland, the big data development index released by the key laboratory of the big data strategy, and the China data factor market-oriented development index released by the National Industrial Information Security Development Research Center. This type of indicator is authoritative, can fully reflect the development status of data factors, and has obvious advantages. However, most of these indicators only measure data from recent years, and most of them are macro-level data, making it difficult to examine the dynamic evolution process of more micro-level data factors from a long-term perspective.

In fact, some scholars have tried to build a data factor index system based on their understanding of the concepts and connotations of data factors. For example, Shi Dan and Sun Guanglin (2022) construct a big data development index from the two dimensions of institutional conditions and market conditions; Wang Renxiang and Xie Wenjun (2024) construct a data factor index system from four dimensions: data generation, data management, data dissemination and data application; Pan Hongliang et al. (2024) construct a data factor development evaluation index system from three dimensions: data foundation support, data capability transformation and data industry application; Xu Xiang et al. (2024) measure the scale and level of data factors in various regions of my country based on the economic value of data factors, namely the labor cost of production data and the cost of data carriers. However, most of these indicators are also regional data and cannot reflect the degree of utilization of data factors at the micro level of the enterprise.

On the other hand, there are also researches that use text analysis methods to measure data factors from the micro level of enterprises, and use the total number of word frequency related to data factors in the annual report to reflect the degree of application of enterprise data factors (Saunders and Tambe,

2013; Zhang et al., 2021; Tang et al., 2022; Wang and Chao, 2024). The measurement indicators of this type of method can reflect information at the micro level of the enterprise and have obvious advantages; however, the word frequency data used in this type of method may have some noise.

3. Specific Measurement Methods of Data Factors

Given that the index system and text analysis method have their own advantages and disadvantages; in order to more accurately measure data factors, this article draws on the idea of Tang et al. (2022) to build an enterprise digital economy indicator system, combines the two methods to construct data factor measurement indicators from the provincial macro level and the enterprise micro level. The specific steps are:

The first step is to use the Big Data Development Index in the "Big Data Blue Book: China Big Data Development Report" as the overall index for measuring the macro-level data factors (denoted by the symbol *Sbigdata*). Since this series of reports only evaluates the big data development index at the 31 provinces, cities and district levels from 2016 to 2020, here, the data from Zhao et al. (2019) are used to calculate the average annual growth rate at the provincial level, and then calculates the big data development index data from 2011 to 2015.

The second step is to use text analysis to construct the enterprise-level data factor development and utilization index (denoted by the symbol *Inddata*), that is, the number of keywords related to the annual report data factors of manufacturing listed companies is used to measure the development and utilization of enterprise-level data factors. Specifically, first, based on relevant policy documents such as the "14th Five-Year Plan for Digital Economy Development" and the series of research reports of the "Data Factor White Paper", through Python word segmentation processing and manual identification, the relevant keywords of data factors are selected from four dimensions: data factor stock, data development capabilities, data-driven commercial applications, and data value monetization. Secondly, use Python software to search and match the keywords about data factors in the annual report of listed companies, and at the same time eliminate keywords with negative meaning expressions, summarize the total word frequency of valid keywords, and add 1 to it and take the natural logarithm.

The third step is to calculate the comprehensive index of enterprise-level data factors (*Data*) based on the two-level indicators calculated above. The specific approach is to first match the provincial-level data factor overall index (*Sbigdata*) to the enterprise level, and then arithmetic average of the enterprise-level data factor overall index (*Sbigdata*) and the enterprise-level data factor development and utilization index (*Inddata*) to calculate the enterprise-level data factor comprehensive index (*Data*). It should be noted that before arithmetic averaging, the two indicators were standardized to eliminate the differences between dimensionality and orders of magnitude.

Table 1. Three Scheme comparing

Dimension	Keywords
Data factor inventory	Big data, data integration, data fusion, data information, data management, data assets, digitalization
Data development capabilities	Automation, 5G, intelligence, robots, machine learning, 3D printing, 3D technology, 3D tools, AI, Internet of Things, edge computing, cloud computing, cloud services, cloud, digital technology, digital technology, computer technology, information age, information technology, information integration, information communication
Data-driven commercial applications	O2O, B2B, C2C, P2P, C2B, B2C, Electronic Technology, Electronic Technology, Online, Network, Online and Offline, Internet, E-commerce, Cross-border E-commerce, E-commerce Platform, Smart Era, Smart Construction, Smart Business, Digital Operation, Digital Terminal, Digital Economy, Digital System, Digital Supply Chain, Digital Marketing
Data value monetization	Digital currency, blockchain, digital trade

4. Analysis of Measurement Results of Data Factors

This section analyzes the two subdivided indicators of the provincial-level overall index of data factors, the enterprise-level data factor development and utilization index, and the measurement results of the data factor comprehensive index obtained by weighting.

4.1. Analysis of Measurement Results of Big Data Development Index

Figure 1 shows the mean change trend of China's big data development index. From 2011 to 2020, the average value of China's big data development index showed a continuous upward trend overall, and the average value of the national big data development index rose from 25.631 in 2011 to 35.998 in 2020, of which the upward trend between 2016 and 2019 was particularly obvious. This shows that the development speed of China's big data is accelerating, especially after the implementation of the "National Big Data Strategy", the big data industry has ushered in unprecedented development opportunities and has a faster development speed. The average value of the big data development index in 2020 has declined because: on the one hand, the big data development index index system of the "Big Data Blue Book: China Big Data Development Report (2021 Edition)" has been adjusted, resulting in the overall index values being lower than in previous years; on the other hand, affected by the current economic situation of insufficient domestic demand and high fiscal pressure in China, especially in relatively backward areas, insufficient financial support has led to slow progress in the construction and upgrading of data infrastructure and insufficient ability to realize the value of data factors, resulting in a decline in the level of big data development.

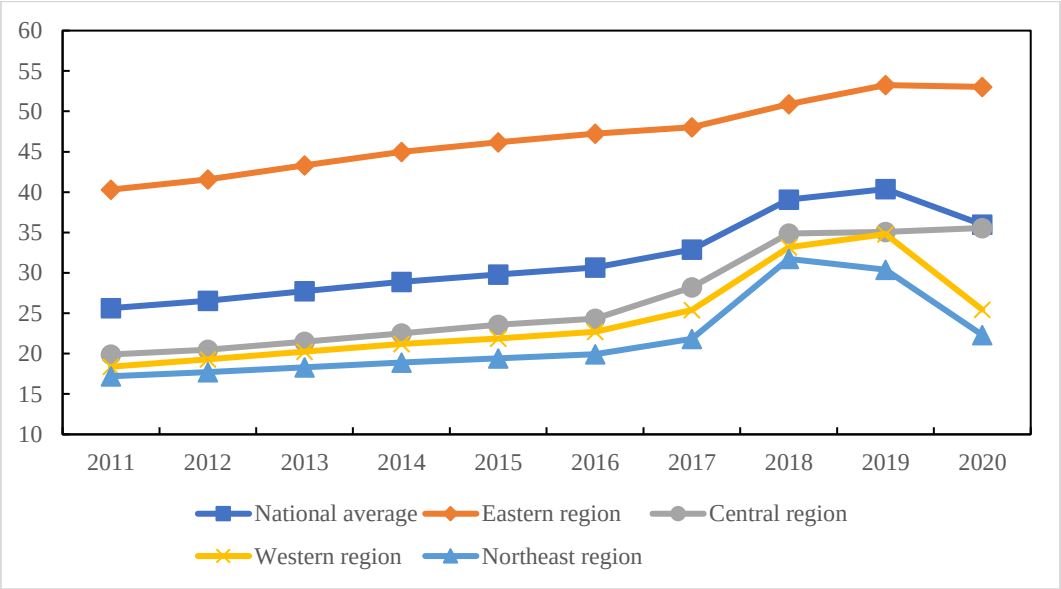

Figure 1. Trends in regional average value of big data development in China, 2011-2020

Judging from the development situation of each region, there are obvious differences in the development level of big data in various regions of China. The average value of big data development in various regions shown in Figure 1 shows that the level of big data development in the eastern region is higher than the national average, the level of development in the central region is comparable to that in the western region, while the level of big data development in the northeast region is the lowest. In addition, the development of big data in the eastern and central regions is relatively stable, showing an upward trend year by year. The western region and the northeast region developed rapidly from 2016 to 2019, which may be due to the key support of national and regional policies; but during the 2019 to 2020, the level of big data development in the region declined, which may be due to the relatively backward development of data infrastructure and insufficient financial support.

4.2. Analysis of Measurement Results of Data Factor Development and Utilization Index

Figure 2 shows the trend of the mean and standard deviation changes of the data factor development and utilization index. From the measurement results, from 2011 to 2020, the overall average of the data factor development and utilization index showed a continuous upward trend, and the average of the data factor development and utilization index rose from 2.283 in 2011 to 3.626 in 2020, of which the upward trend between 2013 and 2020 was particularly obvious. This shows that China's data scale is constantly expanding and its data value is constantly being released, especially after 2013, the development and utilization of data factors has increased day by day. On the one hand, this is due to the optimization of the policy environment. The state and local governments have introduced a series of policies to promote the reform of the market-oriented allocation of data factors and the construction of the data factor ecological system, providing institutional guarantees for the development and utilization of data factors. On the other hand, the construction of digital infrastructure and the wide application of digital intelligent technology provide technical support for the development and utilization of data factors. The new data infrastructure with "high storage capacity, high computing power, high transportation capacity, high safety and high energy efficiency" has effectively improved the quality of data supply, and has continuously innovated and developed big data solutions, analysis tools and related technologies, promoting the improvement of data storage, processing and analysis capabilities.

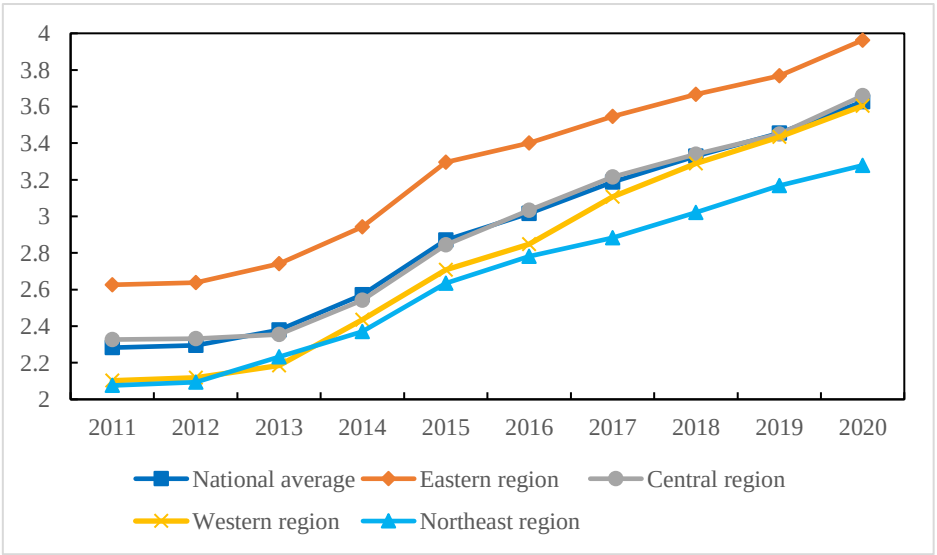

Figure 2. Regional mean change trend of data development and utilization index in China, 2011-2020

Judging from the development situation of various regions, there are obvious differences in the data development and utilization levels of enterprises in various regions of China. The average of enterprise data development and utilization indexes in various regions shown in Figure 2 shows that the data development and utilization level of enterprises in the eastern region is the highest and has maintained a high-speed growth trend; the data development and utilization level of enterprises in the central region is basically consistent with the national average level; the data development and utilization level of enterprises in the western region is low, but the development speed is relatively fast, almost reaching the national average from 2018 to 2020; while the data development and utilization level of enterprises in the northeast region is the lowest. Although certain development results have been achieved, there is still a big gap compared with other regions.

4.3. Analysis of Measurement Results of Comprehensive Index of Data Factors

Figure 3 shows the mean change trend of the comprehensive index of data factors. From the measurement results, from 2011 to 2020, the comprehensive data factor index generally showed a continuous upward trend, and the average of the comprehensive data factor index rose from 0.167 in 2011 to 0.261 in 2020. Among them, the upward trend was particularly obvious from 2016 to 2018, while from 2019 to 2020, the increase rate tended to be flat. The reason may be that in 2015, my country proposed to implement the "National Big Data Strategy" and promulgated the "Action Outline for Promoting the Development of Big Data" to conduct top-level design and coordinated deployment for the development of big data in my country, thereby starting the construction of a data power in my country and accelerating the development of data factors. During the period from 2019 to 2020, the development of data factors has slowed down, which may be because the development of the data factor market is still in its initial stage, and the data circulation rules and data supply and demand docking mechanism have not been effectively established, resulting in inactive on-site trading market; in addition, there are also legal and regulatory issues that need to be improved during the data transaction process, such as unclear data ownership and difficult data security, which has led to most companies with data resources not daring or unwilling to participate in data transactions, which to a certain extent hindered the development of the data factor industry.

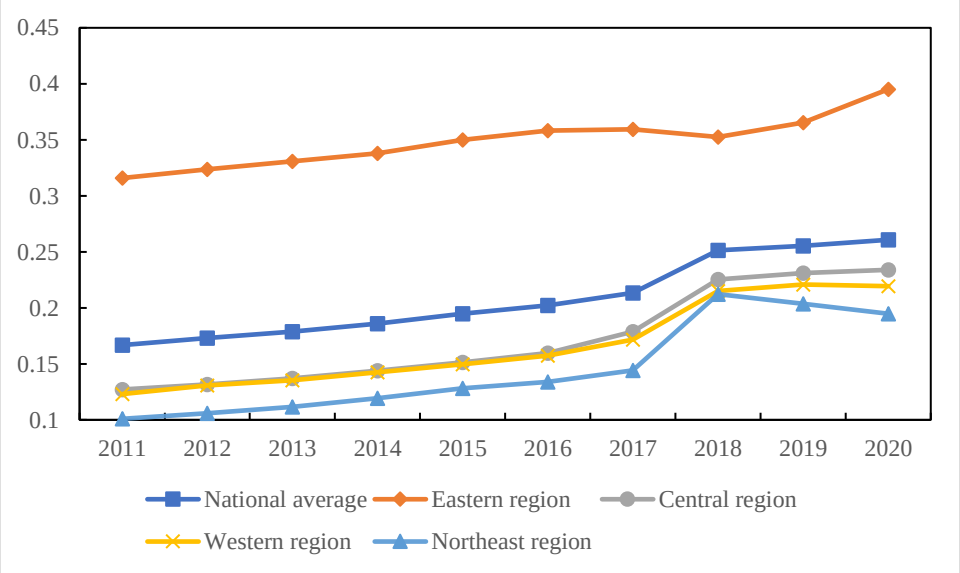

Figure 3. Trend of regional mean value of comprehensive index of data factor in China, 2011-2020

Judging from the development situation of various regions, there are still obvious differences in the development levels of data factors in various regions of China. Figure 3 shows the mean change trend of the comprehensive index of data factors in each region. The results show that from 2011 to 2020, the comprehensive data factor indexes in various regions of my country showed a steady growth trend, but the development differences between regions were relatively obvious. The comprehensive index of data factors in the eastern region is the highest and is far higher than the national average. The development trend of the average data factor comprehensive index in the central and western regions is almost consistent, but between 2017 and 2020, the average data factor comprehensive index in the central region was slightly higher than that in the western region, indicating that the development of data factors in the central region was slightly better than that in the western region during this period. The average level of the comprehensive index of data factors in the Northeast region is the lowest, and it is worth noting that after a brief and rapid growth between 2016 and 2018, it showed a slight downward trend between 2019 and 2020. This shows that the development potential of data factors in the Northeast region is insufficient, and it is necessary to focus on strengthening policy support and guidance, strengthening digital infrastructure construction, etc., to effectively improve the development potential and competitiveness of data factors in the Northeast region.

4.4. Industry Differences Analysis of Comprehensive Data Factor Index

Figure 4 shows the industry mean difference in the comprehensive data factor index. [10] There are obvious differences in the development level of data factors among each two-digit code manufacturing industry. Computer, communications and other electronic equipment manufacturing industries (C39) ranked first with an average of 0.429. This is because products in this industry, such as computers, communication equipment, smart devices, etc., all

can collect, process and transmit data, and have inherent advantages in data collection, processing and application. Following closely behind are electrical machinery and equipment manufacturing (C38) and automobile manufacturing (C36). Industries with high level of development of data factors are advanced manufacturing industries with high technology intensiveness. The core competitiveness of this type of enterprise lies in its technological innovation capabilities, and data has become a key production factor for enterprise technological innovation and industrial upgrading. Therefore, the technology-intensive manufacturing industry has a more urgent need for data and can make full use of the value of data factors.

The bottom three are petroleum processing, coking and nuclear fuel processing industries (C25), ferrous metal smelting and rolling processing industries (C31), and nonferrous metal smelting and rolling processing industries (C32). These types of manufacturing industries are industries with low technology intensiveness, and digital transformation started late, and their emphasis on the application of emerging technologies and data factors is not as high as that of emerging industries, resulting in a lower level of development of data factors.

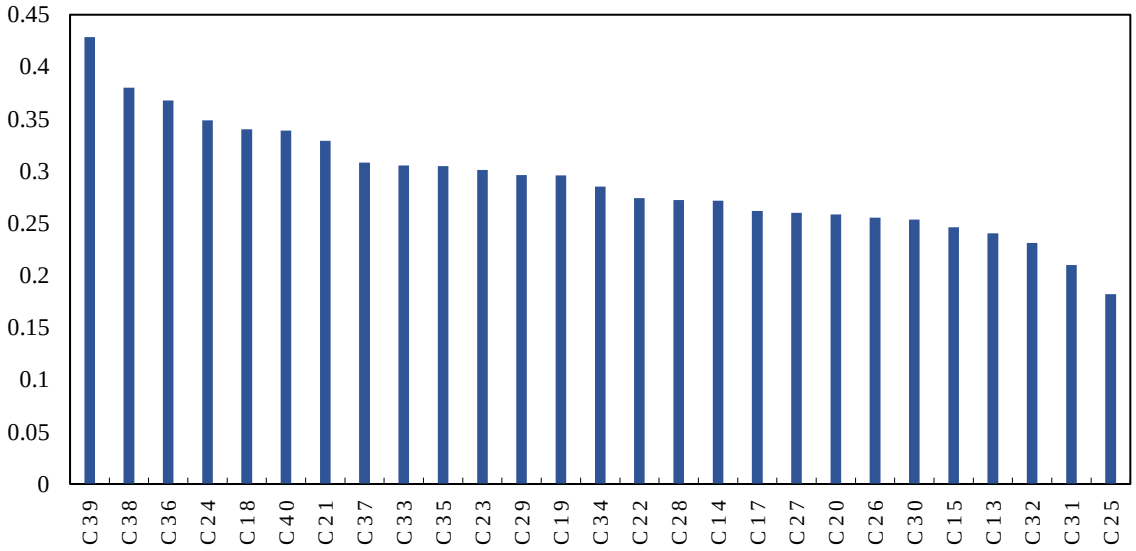

Figure 4. The mean value of the binary manufacturing industry of the comprehensive index of data factor

5. Conclusion

Based on the basic characteristics of data factors, this paper innovatively calculates the China Data Factor Development Index. The analysis found that whether it is the China Big Data Development Index at the macro level, the data factor development and utilization index measurement at the micro level, or the overall data factor comprehensive index measurement results show that the overall development level of data factor showed a continuous upward trend between 2011 and 2020, among which the upward trend was particularly obvious from 2016 to 2018. Judging from the development situation of various regions, there are still obvious differences in the development levels of data factors in various regions of China. The development level of data factors in the eastern region is the highest and far higher than the national average. The development trend of data factors in the central and western regions is almost the same, and the development level of data factors in the northeast region is the lowest. Judging from the differences in manufacturing industries, the development level of data factors among each two-digit code manufacturing industry has obvious differences. Advanced manufacturing industries with high technology intensiveness such as computer, communications and other electronic equipment manufacturing industries (C39), electrical machinery and equipment manufacturing industries (C38), automobile manufacturing industries (C36), etc. have the highest comprehensive index of data factors.

References

[1] D. Shi, G.L. Sun: Influence Mechanism of Big Data Development on the Total Factor Productivity of Manufacturing Enterprises, Finance & Trade Economics, Vol. 43(2022), 85-99.

[2] R. X. Wang, W.J. Xie: How do Data Elements Affect the Structure of Finance and Technology Coupling? R&D Management, Vol. 5(2024), 26-38.

[3] H.L. Pan, L.X. Zhao, L. Ye: Measurement and spatiotemporal evolution research on China' s data elements development, Studies in Science of Science, Vol. 1(2025), 205-216.

[4] X. Xu, X. X. Tian, A. B. Like, et al. A Scale Estimation and Structural Analysis of China's Data Factor: From the Perspective of Information Value Chain, Contemporary finance and economics, Vol. 4(2024), 3-16.

[5] A. Saunders, P. Tambe: A measure of firms' information practices based on textual analysis of 10-K filings. Working. Paper, 2013.

[6] Y.Q. Zhang, Y. Lu, T.Y. Li: Effects of Big Data on Firm Value in China: Evidence from Textual Analysis of Chinese Listed Firms' Annual Reports, Economic Research Journal, Vol. 12(2021), 42-59.

[7] Y.J. Tang, Y. Wang, C. H. Tang: Digital Economy, Market Structure and Innovation Performance, China industrial economics, Vol. 10(2022), 62-80.

[8] C.W. Wang, X.J. Chao: How the Application of Data Elements Affects Distributed Innovation in Enterprises: Empirical Evidence from the Micro Perspective of Listed Companies, Modern Finance and Economics-Journal of Tianjin University of Finance and Economics, 2024, 44(11):3-21.

[9] Y.H. Zhao, Z. Zhang, T.W. Feng, et al. Big data development, institutional environment and government governance efficiency, Management World, Vol. 11(2019), 119-132.

[10] Binary manufacturing industry code and corresponding industry: Agricultural and sideline food processing industry (C13); Food manufacturing industry (C14); Alcohol, beverage, and refined tea manufacturing industry (C15); Tobacco products industry (C16); Textile industry (C17); Textile and apparel industry (C18); Leather, fur, feathers and their products industry (C19); Wood processing and wood, bamboo, rattan, palm, and grass products industry (C20); Furniture manufacturing industry (C21); Paper and Paper Products Industry (C22); Reproduction of printing and recording media (C23); Manufacturing of cultural, educational, artistic, sports, and entertainment products (C24); Petroleum processing, coking, and nuclear fuel processing industry (C25); Chemical raw material and chemical product manufacturing industry (C26); Pharmaceutical manufacturing industry (C27); Chemical fiber manufacturing industry (C28); Rubber and plastic products industry (C29); Non metallic mineral products industry (C30); Black metal smelting and rolling processing industry (C31); Nonferrous metal smelting and rolling processing industry (C32); Metal products industry (C33); General equipment manufacturing industry (C34); Specialized equipment manufacturing industry (C35); Automobile manufacturing industry (C36); Railway, ship, aerospace and other transportation equipment manufacturing industry (C37); Electrical machinery and equipment manufacturing industry (C38); Computer, communication, and other electronic equipment manufacturing industry (C39); Instrumentation manufacturing industry (C40).