sustainability

MDPI

Article

Spatial Analysis of Big Data Industrial Agglomeration and Development in China

Yanru Lu and Kai Cao *

Department of Geography, National University of Singapore, Singapore 117570, Singapore
* Correspondence: geock@nus.edu.sg

Received: 29 December 2018; Accepted: 14 March 2019; Published: 25 March 2019

Abstract: Nowadays, our daily life constantly creates and needs to utilize tremendous amounts of datasets. Fortunately, the technologies of the internet, both in software and hardware, have the capability to transmit, store, and operate big data. With China being the most populous country in the world, developing the big data industry is, therefore, seen as an urgent task. As generating industrial agglomeration is important for forming a mature industry, this study aims to characterize the phenomenon of big data industrial agglomeration in China, and to identify the factors for developing the big data industry using spatial analysis approaches and GIS technology from a geographer's perspective. The problems and strengths of these representative cities are discussed, from which the solutions and the possible directions for the future are also provided. The findings argued that China is still at the primary stage of the development in the big data industry. Only several cities had the presence of a strong agglomeration, but the intercity space spillover was weak. However, comparing the changes in industry distribution, the trend of agglomeration have appeared, and the benefits of industrial agglomeration have also worked. The principal factors of the big data industry and its agglomeration include the support of government and the outstanding higher education agglomeration. In addition, it was also noted that each city has its own characteristics and potentials to attract more big data enterprises, talent, and investment.

Keywords: industrial agglomeration; big data industry; GIS; spatial statistics; China

1. Introduction

With the rapid development of information technology, the cost of data storage, data operation, and data transmission are decreasing continuously, which provides the possibility for generating and processing big data. Big data refers to neo-technologies or neo-architectures that are able to extract value in high veracity from high volumes of data in high-variety by using high-velocity methods of capturing, discovering, and analyzing [1], which are also called the '5V's of big data: value, veracity, volume, variety, and velocity. With technologies and resources of big data, the efficiency of business or research activities of different domains, including finance, telecommunication, public service, manufacturing, agriculture, healthcare, education, retail, and real estate, etc., have witnessed distinct improvement.

As a nation with a population of nearly 1.4 billion and world's second-largest economy combined with rapid economic and technological development during the past few decades (1978 to 2018), China's digital economy has reached 27.2 trillion CNY (≈4.34 trillion USD). Its nominal growth is more than 20.3%, which is significantly higher than the nominal GDP growth (11.1%) and its added value accounts for 32.9% of the gross domestic product (GDP) (82.7 trillion CNY ≈ 13.18 trillion USD). As of September 2017, China has 0.75 billion netizens and 1.39 billion smartphone users [2], which reflects the tremendous intra market for big data industry in China.

Big data is essential for the digital economy, which supports the process of 'Digital Transformation 2.0' in the 'ABC Era' ('Artificial Intelligence + Blockchain + Cloud'). In 2014, Big data was included for the first time in the Report on the Work of the Government [3]. In 2015, the state council released the Action Outline for Promoting the Development of Big Data that emphasizes the critical position of the big data industry [3]. The foundation of new local big data enterprises significantly increased from 2007 to 2014, and reached 223 in 2014 (Figure 1). In addition, the number of new companies in 2015 had a slight decrease, probably caused by incomplete statistics. The incomplete statistics of the new companies in 2016 also did not affect the trend of industrial agglomeration because agglomeration depends more on the accumulation of numbers.

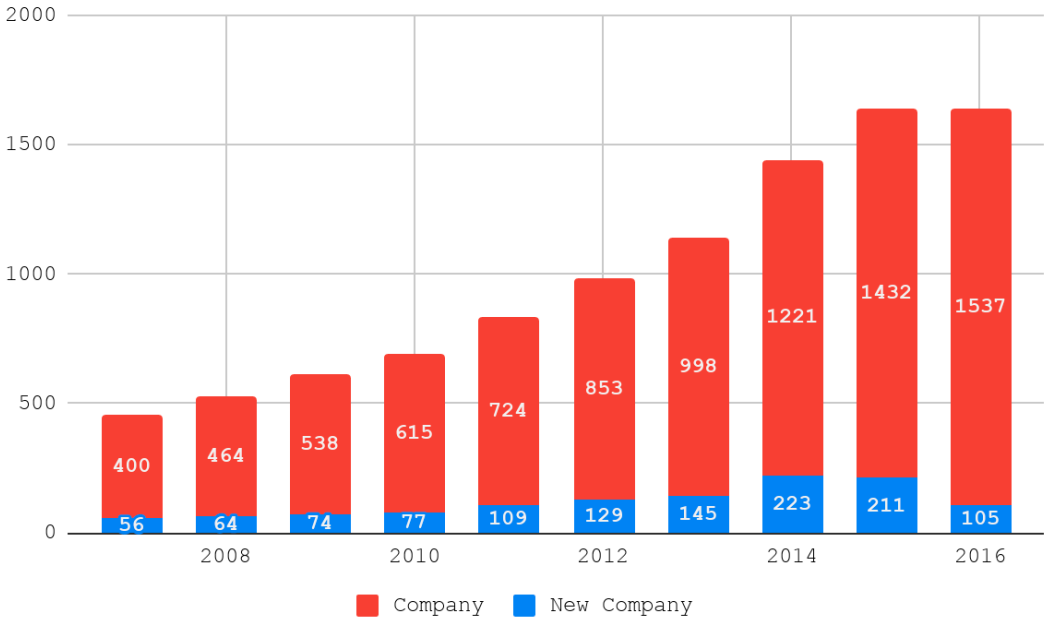

Figure 1. The number of local companies in the Big Data industry [4].

The digitalization of industrial processes provides individual companies with sustainability and innovative strength and also forces the world to rise to the challenge of this transformation from a sustainability point of view [5]. In this study, we take the big data industry as an example. It shows significant unbalanced development in different provinces across mainland China, the differences of which are even more noticeable than economic development. Even in one relatively developed province, cities have considerable gaps in big data industry development, although some of these cities have a similar level of population and development in terms of the economy. Because of these gaps, those cities are diverse largely in particular leading fields: IoT, AI, Blockchain, mobile payment, automotive, etc., and the development lagging in these fields would sharply reduce the progress of local modernization construction.

Guizhou Province, the Beijing-Tianjin-Hebei Region, the Pearl River Delta Region, the Yangtze River Delta Region, Henan Province, Chongqing, Shenyang Province, and Nei Mongol were approved by the National Development and Reform Commission, the Ministry of Industry and Information Technology, and the Central Cyberspace Affairs Commission for constructing the Big Data Trail Zone in 2016 (Figure 2).

The development of the big data industry is vital to the nation's economy, manufacturing, and people's standard of living. Industrial agglomeration enables the congregation of relevant talents and capitals. There is no strict or precise definition of the big data industry. In this study, referring to the definition of Annual Government Work Report (2014), there are two requirements for defining the big data industry: (1) using basic techniques of big data, e.g., Hadoop, Spark, Hive, Pandas, Storm, etc.; (2) taking big data and the value behind big data as the key resources to gain profit and value.

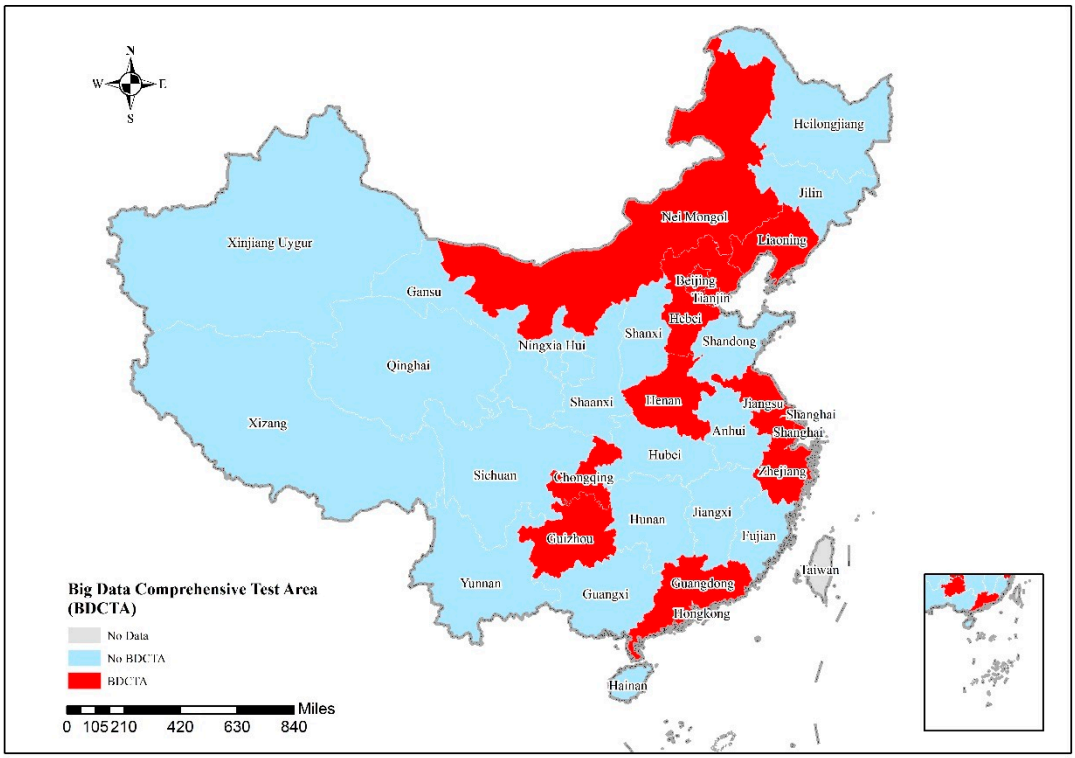

Figure 2. The map of the Big Data Comprehensive Trial Zone (Source: Assessment of Big Data Industry Development, 2018).

Industrial agglomeration is an economic phenomenon of the spatial congregation of firms that require similar technical support or provide relevant services. This kind of agglomeration could decrease the cost of transportation, enable the exchange of technique and knowledge, enrich and extend the industry chain, which is conducive to the development of the industrial economy. A good industry agglomeration not only drives the economy but also accelerates the innovation of those relative enterprises [6]. Fast and sustainable development of this industry depends highly on the understanding of the complicated mechanism behind the industrial agglomeration of big data. In addition, there has not been any study on this aspect. Hence, it is very meaningful to discover the factors and characteristics of big data industry agglomeration in China.

Two significant research questions have been addressed in this research, including: What are the characteristics of big data industrial agglomeration in China and what factors cause big data industrial agglomeration? The remaining manuscript includes a relevant literature review, the introduction to the research area and data collected, methodology, the analyzed results, as well as discussion and conclusion of this research.

2. Literature Review

In this literature review, to better address the research questions mentioned above, we first reviewed the relevant studies on industrial agglomeration, and then followed it by a review of some specific studies on the influencing factors of big data industry. Due to the limitation of existing studies on big data industry agglomeration, some studies on the characteristics of agglomeration in the relevant industry have also been reviewed to support this research.

2.1. Industrial Agglomeration

Industrial agglomeration has been studied significantly in the past decades in various industries [7–11]. In this section, the classic theories of industrial agglomeration will be reviewed, in

which the ways and the perspectives of scholars studying the problem of industrial agglomeration will be introduced so that we can understand industrial agglomeration in-depth and systematically.

The most obvious change in recent studies is that the emphasis of macro-level analysis on the industrial agglomeration has been shifting to medium or micro-level analyses. Regional economists began to discuss agglomeration economy from the perspective of investment locations [12–14]. The locational factors fall into regional factors and agglomeration factors, and the causes of agglomeration generation are classified into two parts: the concentration of enterprises leads to the economy of scale, and the cooperation and infrastructure that also cause the spatial concentration [15]. Hoover [16] first divided agglomeration economy problems into the internal economy of scale, localization economy, and urbanization economy. Marshall [17] proposed the concept of industrial districts and the theory of industrial districts, and Marshall argued that machine and professionals would be more effective when they are in industrial agglomeration.

During the 1970s and 1980s, it was quite popular to use capitalist macroeconomics theory in the discussion of the spatial changes of production activity [18]. Lundvall [19] combined innovation, technological change, economic growth, and commerce to study the new industrial agglomeration innovation system from a Neo-Schumpeterian perspective. Markusen [20] studied five categories of industrial agglomeration, including the Marshallian industrial district, Italianate industrial districts, hub-and-spoke industrial districts, state-anchored districts, and satellite platforms, and analyzed the role of government and large enterprises, embeddedness, regional potential, regional economic structure and their interactions to discuss the characteristics of each type of agglomeration.

Although most of the industrial agglomerations are self-forming, when the agglomeration growth is at a nascent stage or meets with a market failure, the government function plays an essential role in the process of agglomeration growth, especially in China, including institutional supply, public goods supply, financial support, market supervision, and macroeconomic guidance [21]. The institution provides a basic economic incentive structure, the development of which leads to the change of economy [22]. However, Li [21] mentioned that excessive institutional supply or arbitrary market intervention may lead the industry to miss out on the economic benefits. Regarding public goods supply, it is one of the most remarkable benefits that the industrial agglomeration will bring to enterprises [17]. The high geographical concentration of the industrial clusters is able to attract venture capital, professional labor, producers, and users. However, those enterprises usually lack abilities and motivation to produce public goods, such as traffic, telecommunication, culture, entertainment, medicine, finance, advertisement, and law. Consequently, on the one hand, the government's role includes making up the market failure, improving the supply of public goods, promoting the development of infrastructure, and constructing the industrial development supporting system. On the other hand, the government is also supposed to strengthen the development of a soft environment of industry and social service system within the industry cluster.

2.2. Influencing Factors on Big Data Industry

The big data industry is one of the emerging industries, the development of which might be affected by multiple factors and the factors would influence the industry by forming a complex system. The major influencing factors in big data industry are technology, human capital, business modes, capital, and policy [23]. However, the factors of infrastructure and market have less effect on developing the big data industry. As the internet helps to reduce the cost of space and time, accompanied by the fact that the industrial chain is dispersedly distributed all over the country, the influence of infrastructure is weakened. Similarly, due to the boundary of the market being enormously expanded, the local market is also not one of the vital factors in big data industry development. The characteristics of the big data industry are different from the traditional information industry. Although the industry does not depend much on geographical location to foster or develop big data industry, it is still vital in the generation of industrial agglomeration [24].

2.3. Characteristics of Agglomeration in Relevant Industry

Because of the similarities of human capital structure, daily operation, and production mode in the reserch and development (R&D) industry and big data industry and the lack of researches in big data industry and its agglomeration, the methodologies of exploring the characteristics of R&D industry and its agglomeration is also valuable for this study although the objectives of those researches are not directly associated with the big data industry. In today's society, the knowledge economy and innovation are not only necessary but also the principal drivers of the prosperity of a society, which is why governments around the world encourage the development of innovation and high technology industries. The mechanism of industrial agglomeration is the core of the industrial economy [25]. Knowledge-intensive industries invariably have characteristics of spatial agglomeration, because adjacent locations make generating and distributing knowledge more effective [26]. As it is different from traditional industries, such as basic chemical manufacturing, which relies on land, raw material, and traffic, the likes of the internet industry and the big data industry rely more on the high-quality labor, research institutions, and potential customers.

The agglomeration of science and technology-oriented industry and the development of the economy have similar spatial distributions, which means, to some extent, technological progress is a by-product of economic growth [27]. Wang divided the study area of China into three parts and employed Moran's I, local indicators of spatial association, and other spatial statistical tools to find the obvious differences among Eastern, Central, and Western parts for regional technology development and R&D industrial agglomeration, which aligns with the distribution of economic development. Zheng [28] analyzed the changes of locational factors and location potential in decision-making. The results showed that information, knowledge, and innovation are new factors of industrial agglomeration in the era of information, and technology is the important factor that builds the spatial structure of society and the economy. Varga [29] examined the relative influence of static and dynamic agglomeration effects empirically as well as networking on regional R&D productivity in the Europe Union. They argued that agglomeration is an important determinant of R&D productivity in the case of science-driven (Pasteur-type) research. Moreover, those two determinants are never jointly significant, which indicates that agglomeration and scientific networking are not substitutes or complements and they only operate in the knowledge production process. Han [30] set the cloud computing industry as a research object. He argued that the supply factors of emerging industries comprise production factors, including technology, capital, and human talent; demand factors that mainly refer to the domestic market; policy environment factors, including policy, laws and regulations, and industry standards; and also, infrastructure factors, such as the internet, industrial clustering, and resources. Sun [31] also studied R&D Industrial Agglomeration where the study area is located in the Yangtze River Delta region in China mainly by spatial analysis, such as the kernel density analysis and Ripley's K. For R&D industrial agglomeration, the author argued that the main factors of such generation are market size, knowledge spillovers, size of government and transportation.

3. Study Area and Data

3.1. Study Area

In this research, the study mainly focuses on the entire mainland of China. The level of province, city and county data will be analyzed. The population from the eastern part of China is more than the western part of China, and so is the industry density. Moreover, the economy of the eastern part is much more developed than the western part.

3.2. Data

Demographic Data about the companies in the big data industry in China is one of the most important data in this study, which consists of the company's name, launch time, address, city, and service type. This data were collected from the website of Data Technology Innovation Industry

Institute that was updated in October, 2017. Other important data, including the university in 211 program, permanent population, GDP, and fiscal expenditure of each city were obtained from the China Statistics Bureau. After collecting these data, geocoding and table join were employed to process the data in a GIS environment. Some important missing data, for example, the location and launch year of the companies, were fixed either by manual searching or automatic searching in the internet using Python programming, and each data were double-checked one by one.

The big data industry in the data source consists of five types, which have complementarities and similarities: (1) industry use, including advertising industry, transportation industry, financial industry, real estate, tourism, medical and health industry, human resource, agriculture and many other kinds of companies that use big data technology or data resources in their daily operations (e.g., production, marketing); (2) big data tool services, which contains social network analysis, artificial intelligence, business intelligence, e-commerce analysis, big data visualization services, and other kinds of third-party analysis services; (3) Infrastructure suppliers, such as information transmission, data storage, data security, cloud service; (4) big data source suppliers, which includes government data, geographic data, data mart, physiological data, social network data, and data collection service; (5) peripheral services: industry alliances, big data research institutions, education institutions, professional media, the professional community, and big data competition platforms.

4. Methodology

In this research, to better look into the characteristics of big data industry agglomeration at various levels, a few spatial analysis approaches were employed, specifically including kernel density analysis, global Moran's I, local indicators of spatial association (LISA) clusters, spatial lag/error models, which are commonly used approaches in spatial analysis related studies. In addition, the workflow of this research can also be seen below (Figure 3), which can provide a more systematic view of the methodological framework of this research.

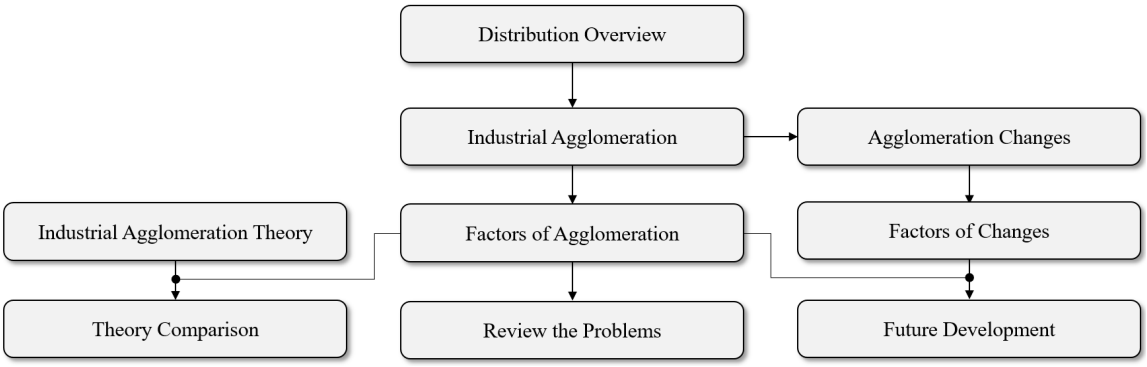

Figure 3. Workflow of the research.

4.1. Methods of Exploring Industrial Agglomeration Phenomenon and Changes

4.1.1. Kernel Density

Kernel density calculates the density of points in a neighborhood around each output raster cell, which represents the characteristics of points distributions. It is based on the quartic kernel [32] with the help of ESRI ArcGIS 10.4 to visualize and represent the distribution and agglomeration level of big data companies. Kernel density is a commonly used tool in approaching the status of spatial density. We have compared the kernel density maps for 2007 and 2016, to find the changes in industrial agglomeration. The function of calculating kernel density follows.

$$f(j) = \frac{1}{h^2} \sum_{i=1}^{n} \left[\frac{3}{\pi} \left(1 - \frac{{d_{ij}}^2}{h^2} \right)^2 \right] \quad (1)$$

where $f(j)$ is the density at the center of the cell; h is search radius; n is the number of points within the search radius. and d_{ij} is the distance between the point of interest and center of the cell.

4.1.2. Spatial Autocorrelation Analysis

Spatial autocorrelation analysis is a series of methods that use aggregated data to analyze, to some extent, whether the issue has an effect and to show that the spatial aggregation would improve the effects for the surrounding area [33]. Generally, spatial autocorrelation analysis can be divided into two types, namely global spatial autocorrelation analysis and local spatial autocorrelation.

Global spatial autocorrelation analysis only indicates the significance of spatial autocorrelation, but do not point out the location of clusters, which is a good way to describe the general tendency of spatial autocorrelation. It is calculated based on covariance. In statistics, covariance measures the joint variability of two variables, for two jointly distributed real-valued random variables X and Y, which have finite second moments and the expectations of them are $E(X)$ and $E(Y)$, respectively. The covariance of them is defined as [34]:

$$cov(X, Y) = E[(X - E(X))(Y - E(Y))] \tag{2}$$

Global Moran's I is calculated as [35]:

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j}(x_i - \overline{x})(x_j - \overline{x})}{W \sum_{i=1}^{n} (x - \overline{x})^2} \tag{3}$$

in which n is the total number of data units; x_i represents the value of each unit; W_{ij} is the spatial weight of two neighbor units, and W is the sum of weights.

The result of global Moran's I is between -1 and 1. The positive value represents positive spatial correlation, and the negative value means negative spatial correlation. The larger global Moran's I it has, the more significant of spatial autocorrelation it has, in other words, it contains more significant phenomenon of spatial agglomeration. On the contrary, a small value of global Moran's I reflects low spatial autocorrelation, and when the value tends to be 0, the spatial distribution tends to be random.

Local indicators of spatial association (LISA) is able to estimate the extent of spatial association [36]. Given a set of weighted features, the univariate local Moran's I identifies clusters of features with values similar in magnitude. The tool also identifies spatial outliers by calculating a Local Moran's I value, a Z score, a p-value, and a code representing the outlier type (COType) for each feature. The Z score and p-value represent the statistical significance of the computed index value.

The local Moran's I [36] is calculated as:

$$I_i = \frac{x_i - \overline{X}}{{S_i}^2} \sum_{j=1, j \neq i}^{n} w_{ij}(x_i - \overline{X}) \tag{4}$$

where x_i is feature i's attribute; $\overline{X}$ is the mean of the corresponding attribute; W_{ij} is the spatial weight between feature i and j, for S_i^2:

$${S_i}^2 = \frac{\sum_{j=1, j \neq i}^{n} w_{ij}}{n - 1} - \overline{X}^2 \tag{5}$$

with n that represents the total number of features.

The Z_{I_i} score is computed as:

$$z_{I_i} = \frac{I_i - E[I_i]}{\sqrt{V[I_i]}} \tag{6}$$

where

$$\mathrm{E}[I_i] = \frac{\sum_{i=1.\ j\neq i}^{n} W_{ij}}{n-1} \tag{7}$$

and

$$\mathrm{v}[I_i] = \mathrm{E}\left[{I_i}^2\right] - \mathrm{E}[I_i]^2 \tag{8}$$

The *p*-values are approximate values of the area under the curve for a known distribution and are limited by the test statistic. When the feature is surrounded by the features of similar values, local Moran's I shows a positive value. Similarly, when the feature is surrounded by features with dissimilar values, the index shows a negative value, and the feature surrounded is called an outlier. The local Moran's I can be meaningless without the context of the Z score or *p*-value because the local Moran's I is a relative measure. The COType distinguishes between a statistical significance of 95 percent cluster of high values (HH), cluster of low values (LL), outliers with high value which are mainly surrounded by low values (HL), and outliers with the low value which are mainly surrounded by high values (LH).

In this case, by calculating Anselin local Moran's I, to find the outliers and clusters for the number of big data enterprises on the city-level, it reflects the unbalanced development and distributions of big data industry among the cities. Those cities with HH indicates that industrial agglomeration does exist among the cities.

4.2. Methods of Exploring the Factors of Agglomeration

There are a lot of factors generating the agglomeration of big data industry. However, due to the special property of this industry where it does not require special natural resources, a transport network support or local market demands, the most important factors that need to be considered include the following aspects: human resources, support of government, the local market, and the local economy (Figure 4). In this study, the number of top universities, government fiscal expenditure in science, permanent population, and GDP of each city were utilized to represent these four factors, respectively (Table 1). The number of local big data enterprises was set to be the dependent variable to represent the development level of industrial agglomeration in each city.

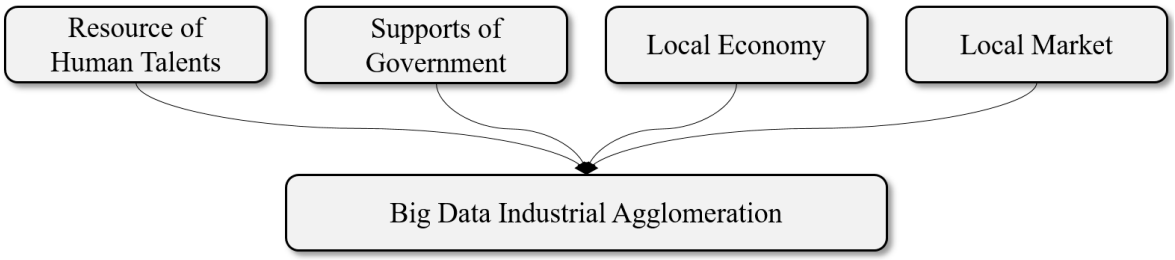

Figure 4. Model of Main Factors.

Table 1. Independent Variables of Regression.

Variable	Definition
GDP	Normalized GDP of each city in 2015
FE_Sc	Normalized fiscal expenditure in science of each city in 2015
Pop	Normalized permanent resident of each city in 2015
Top_Univ	Normalized number of universities in the Double First Plan of each city

There are three regression analysis models relative to this question. For spatial lag model (SLM) and spatial error model (SEM), we used the Lagrange multiplier (LM) and the robust Lagrange multiplier (R-LM) to test which model was better for the case. If SLM has a higher LM value, SLM would be the better choice and vice versa. If both of them show non-significant LM values, ordinary least squares (OLS) becomes the best choice of these three models. If both of them have significant

values of LM, then using the model which has a higher significance of R-LM value will be a better choice [37].

4.2.1. Ordinary Least Squares (OLS)

Ordinary least squares is also called linear least squares, which is a statistical method to estimate unknown parameters. It is commonly used in linear regression. The smaller the difference it shows, the better the model fits the data. A simple formula would represent the resulting estimator:

$$\mathrm{y} = \mathrm{a} + B_1 X_1 + B_2 X_2 + \cdots + e \tag{9}$$

where a is an intercept, B is regression coefficient of X which is the independent variable and e is the residue.

4.2.2. Spatial Lag Model (SLM)

SLM is also called spatial autoregressive model (SAR), which describes spatially substantive dependence and it reflects the spatial interactive effects. When the spatial independence among the explained variables is so crucial in causing the spatial correlation, the relationship could be described as:

$$\mathrm{y} = \lambda W y + Z\beta + u \quad |\lambda| < 1 \tag{10}$$

where Z refers to the explained variable; Wy refers to the lag term; λ refers to the coefficient of spatial autoregression, and u refers to error term vector [37,38].

4.2.3. Spatial Error Model (SEM)

SEM, which is also called spatial autocorrelation model (SAM) or spatial residual autoregressive model (SRAM), describes spatial nuisance dependence and spatially global dependence. The equations are as follows:

$$\begin{gathered} \mathrm{y} = Z\beta + u \\ u = \rho W u + \varepsilon \quad |\rho| < 1 \end{gathered} \tag{11}$$

where Z refers to explained variables; the non-stochastic weights Wu refers to the stochastic error vector; ρ refers to the coefficient of spatial autoregression, and ϵ refers to the normally distributed stochastic error vector [37,38].

4.3. Review the Status Quo of Cities in Factors Terms

After finding the major factors in generating the big data industry and its agglomeration, as well as understanding the development mechanism of this industry, this study also reviewed the status quo of the cities from the perspective of factors to make the theory and data results a reality. Some cities may have capabilities to attract more big data enterprises and have a more developed industry in the future, and the weaknesses they have are worth being concerned about. The tool of 3D scatter plot in Geoda was used to assist in finding and highlighting the points of interest.

In 3D scatter plot, there are three axes: X-axis, Y-axis, and Z-axis, which represent government expenditure in science, GDP, and the top universities in each city, respectively. The cube inside helps us to select and brush the data in the map as well as in the table so that it is possible to review the status quo of the city, which includes house price, policy, labor, infrastructure, local market, and the exact data values of the major factors.

5. Results

5.1. Spatial Characteristics of Big Data Industry Distribution

The number of local big data enterprises represents the development of each province (Figure 5). From the figure, although the provinces of central and southeastern China show higher development than the western and northeastern part, Jiangxi province that is surrounded by comparatively higher development provinces is still not such flourishing in the industry of big data. However, this figure can only reflect the data on province-level, which may not be as accurate and precise for each city inside the province.

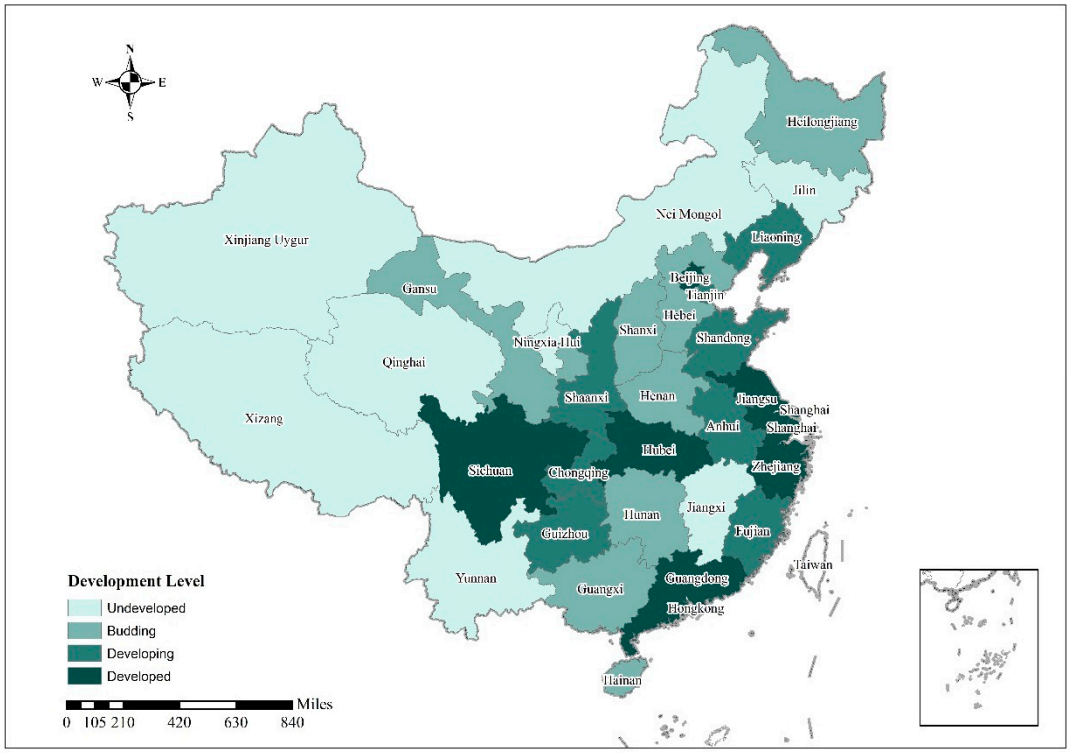

Figure 5. Development Level of Each Province in 2016.

Between 2008 and 2016, the global Moran's I of big data companies shows a low positive value and a general growing trend. From 2009 to 2013, the index presents an increasing trend, from 0.0356741 and up to 0.0480922 (Figure 6). That means, although the county-level big data industrial agglomeration is not that common, a congregation trend has been noted to emerge.

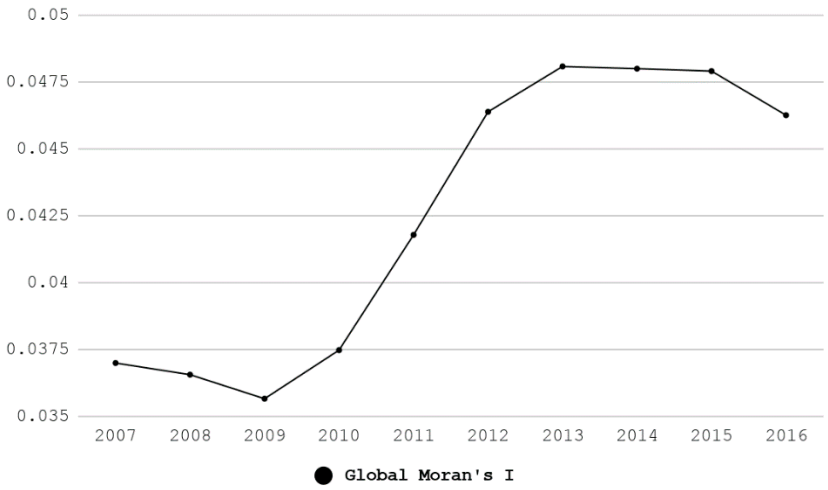

Figure 6. Global Moran's I from 2007 to 2016.

The kernel density map of 2007 indicates that Beijing, Shanghai, and Shenzhen were the three super cores, with two peripheral cores lying in Chengdu and Wuhan (Figure 7). Among them, Beijing is the most powerful core, and the core in Shanghai is the largest in scale extending to south areas of the Jiangsu province, the strong core of Hangzhou in Zhejiang province, and some regions of Anhui province. The rest of the kernel density map reveals comparatively weak kernel density, and for most of provinces, especially in western China, there are extremely low agglomeration densities.

The kernel density map of 2016 presents a phenomenon of a great increase in overall national KD, more than three times as compared to 2007 (Figure 8). In 2016, except for the existing areas of a relatively high density, a certain scale of cores arose in some parts of Guiyang, Chongqing, and Shenyang. In this year, there existed three super cores, and more than half of the provinces manifested the comparatively evident agglomeration phenomenon.

The map of local Moran's I in 2016 shows more details about agglomeration distribution at the county-level (Figure 9). In this study, HH, LL, LH, and HL are interpreted as the agglomeration region, depressed region, hollowing region, and island region of big data industry, respectively. For this section, HH, LH, and HL are the regions that are worth paying attention to. As can be seen from the map, HH clusters appear to have similar distributions with the kernel density map in 2016. There are significant differences in the cluster's distribution across the research area. Most of the counties in Beijing, Shanghai, the Pearl River Delta Region, and partial counties in Chengdu, Hangzhou, Nanjing, and Wuhan have shown high-level agglomeration, and the majority of them are far from each other. Comparing the clusters of agglomeration quantitatively, the cluster that are located in the Yangtze River Delta region tops the list. Moreover, around those agglomeration regions, there are always some hollowing regions. Apart from those places, most provinces only exist as isolated islands.

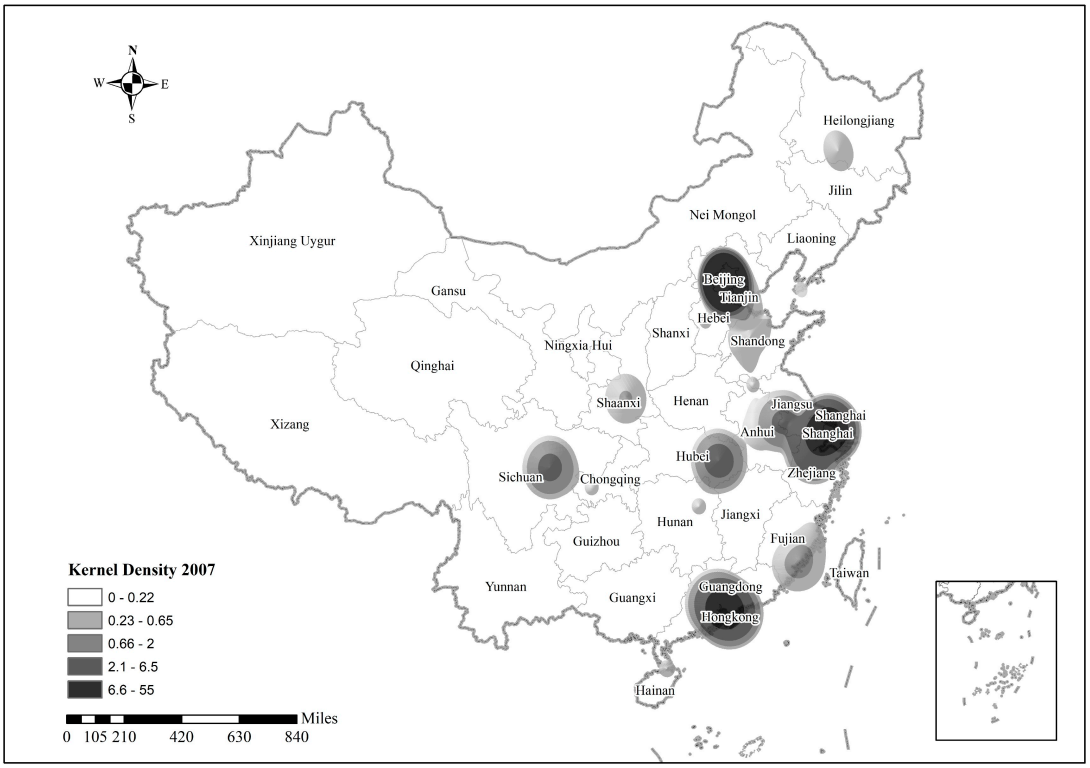

Figure 7. Kernel density map of 2007.

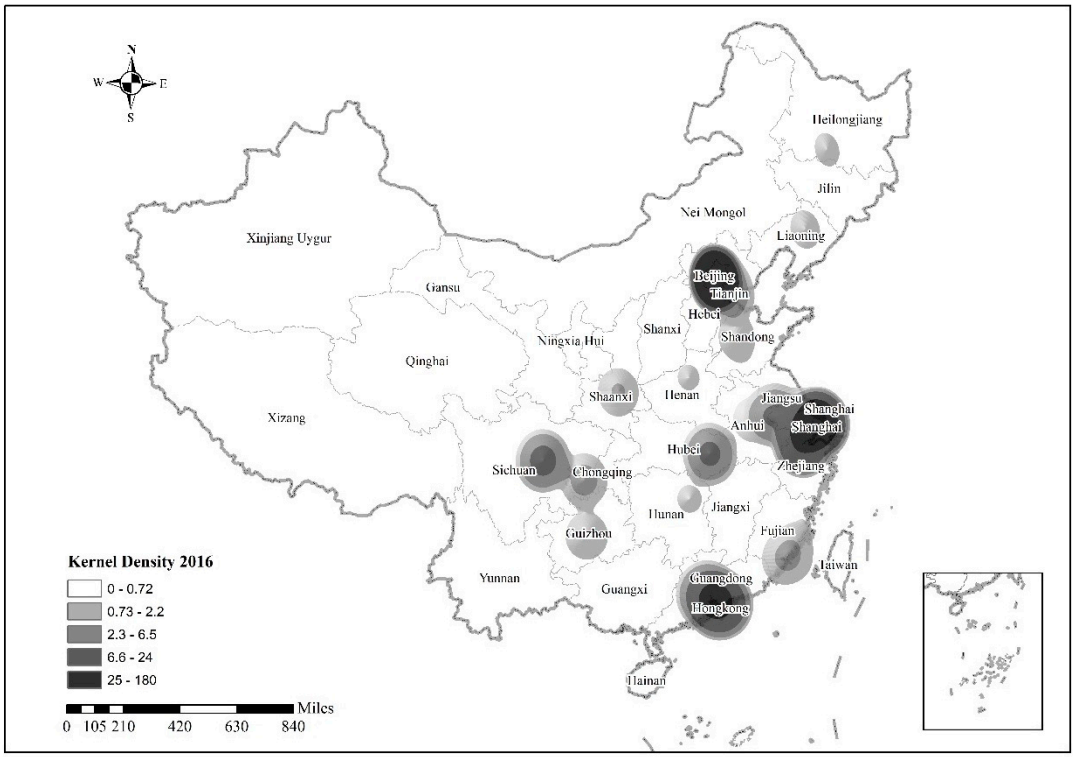

Figure 8. Kernel density map of 2016.

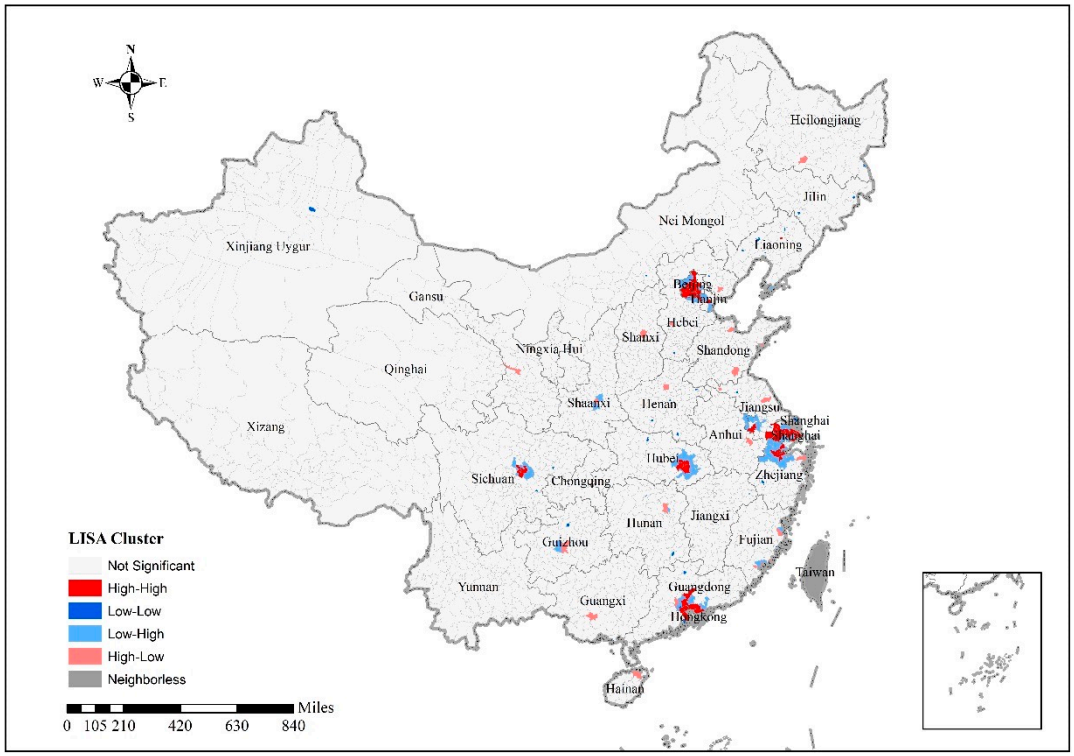

Figure 9. Local indicators of spatial association (LISA) cluster map in 2016.

5.2. Factors of Big Data Industry Agglomeration

The results of the OLS model for 214 observations reveal the acceptable fitness of 0.88 and the high significance of the four factors (Table 2). Among these three factors, fiscal expenditure of science, population, and top university show positive correlations with big data industry agglomeration at the city scale and GDP shows a negative correlation. And for spatial dependence diagnostics, both the Lagrange multiplier (error) value and Lagrange multiplier (lag) value are not significant. As such, in this study, the ordinary square model was used (Table 3).

Table 2. Ordinary square model results.

Variable	Coefficient	Std. Error	t-Statistic	Probability
GDP	−0.304531	0.03677	−8.28203	0.00000
FE_Sc	0.433443	0.0379156	11.4318	0.00000
Pop	0.066763	0.0265546	2.51418	0.01268
Top_Univ	0.638184	0.0336114	18.9871	0.00000
R-squared	0.886880	Adjusted R-squared	0.884716	

Table 3. Diagnostics for spatial dependence.

TEST	MI/DF	Value	Probability
Moran's I (error)	−0.0148	−0.1678	0.86674
Lagrange Multiplier (lag)	1	1.4647	0.22619
Robust LM (lag)	1	1.4516	0.22826
Lagrange Multiplier (error)	1	0.0941	0.75908
Robust LM (error)	1	0.0810	0.77591
Lagrange Multiplier (SARMA)	2	1.5457	0.46169

According to the results of the ordinary square model and the diagnostics of spatial dependence, there are five findings listed below:

(a) Spatial spillovers among the cities were not significant. The cities that have a comparatively developed economy in big data industry did not show a significant impact on surrounding areas. It reflects that big data industry was still at the primary stage of development.

(b) The negative value of the regression coefficient of GDP corresponds to the new generated industrial agglomeration that appeared in central inland areas but not on the most developed coastal areas (Figures 7 and 8). It means that compared with the areas that have a developed economy and the other fields of industries with solid foundations, those cities that were relatively less developed attempted to change the status quo and develop industries that do not require a high level of an industry base or natural resources, such as big data industry.

(c) The government fiscal expenditure for supporting science was important and helpful for a city to attract big data companies to launch in. That is why some undeveloped cities with a lower industry base could be more attractive to entrepreneurs.

(d) The more permanent residents in the city, the larger local market it will have. However, just as mentioned above, the local market was not the most important factors in big data industry or its agglomeration. The coefficient of 'Pop' was only 0.066763 as the lowest among these factors.

(e) The spillover of knowledge that was led by top universities will bring high-technology industrial agglomeration for that city. The coefficient of 'Top_Univ' was the largest among the four factors, which means this factor had the strongest impact on big data industry development and agglomeration.

5.3. Problems in Development of Big Data Industry

By reviewing the main factors of the big data industry, there were eight representative cities suggested. They represented different types and industrial development of cities. In this section, the cities were classified into four types and the relevant factors and problems were also been analyzed (Table 4).

Table 4. Conclusion of Representative Cities.

Type	I	II	III	IV
Description	They have developed big data industrial agglomeration as well as the best industrial bases.	They have the trend of generating competitive big data industrial agglomeration.	They had a relatively good base but are facing bottle-neck in big data industry development.	They had a poor industry base but are facing with the opportunities of big data industry.
Representative City	Beijing, Shanghai	Hangzhou, Shenzhen	Wuhan, Xi'an, Nanjing	Guiyang
Industrial Agglomeration [1]	5	4	3	2
Main Positive Factors [2] (FE_Sc, GDP, University)	(5,5,5)	(4,4,1)	(3,(3~4),4)	(2,2,1)
Other Factors	Infrastructure, Industrial base, Top companies, Big market	Infrastructure, Location, Resident Policy, Industrial Base, Big market	Provincial capital city, Infrastructure	Provincial capital city, Resident policy, Low electricity price
Problem	High cost, High competition	High cost, Fewer top universities	Industrial structure, Industry base	Economy, Fewer top universities

Note: [1] Industrial agglomeration level (1–5): 5 represents the highest agglomeration, 1 is the lowest; [2] Main factors (1–5): 5 represents the highest value, 1 is the lowest.

5.3.1. Type I: Beijing and Shanghai

Beijing and Shanghai were the two cities that had the highest values in all the three axes of 214 cities in China (Figures 9 and 10). Although they looked like the outliers in the 3D scaVer plot (Figure 10), from another perspective, this was why they became the central points of the big data industry as well

as in this study. Both the cities enjoy the best resources in the country, in which Beijing is the country's political, economic, and educational center, and Shanghai is a global financial center, transport hub as well as an education center in China. It seemed that the high cost of living and production did not cause the loss of development of big data industry, because of the sufficient resource of higher education, developed economy, good infrastructure, and robust industrial base. Beijing possessed more than half of the domestic top 100 Big Data enterprises in the country (Table 5). Zhongguancun in Beijing is also called 'China's Silicon Valley', in which there were many high-technology and innovation enterprises or startups. The Zhangjiang Hi-Tech Park is a technology park located in the Pudong district of Shanghai that specializes in life sciences, software, semiconductors, and information technology. The good milieu of innovation and the favorable policy in Beijing and Shanghai were also at the top in China. However, these two metropolises had high population density, skyrocketing prices and are relatively weak in policy on human capital, compared with cities such as Hangzhou. If these trends continue without the intervention of policy regulation, it might cause the industry be lost to other first-tier cities and top second-tier cities.

Table 5. The number of top 100 Big Data enterprises in the city (source: Xinhuanet).

City	Number
Beijing	55
Shanghai	10
Hangzhou	7
Shenzhen	8
Nanjing	1
Wuhan	2
Xi'an	2
Guiyang	1
	(12)

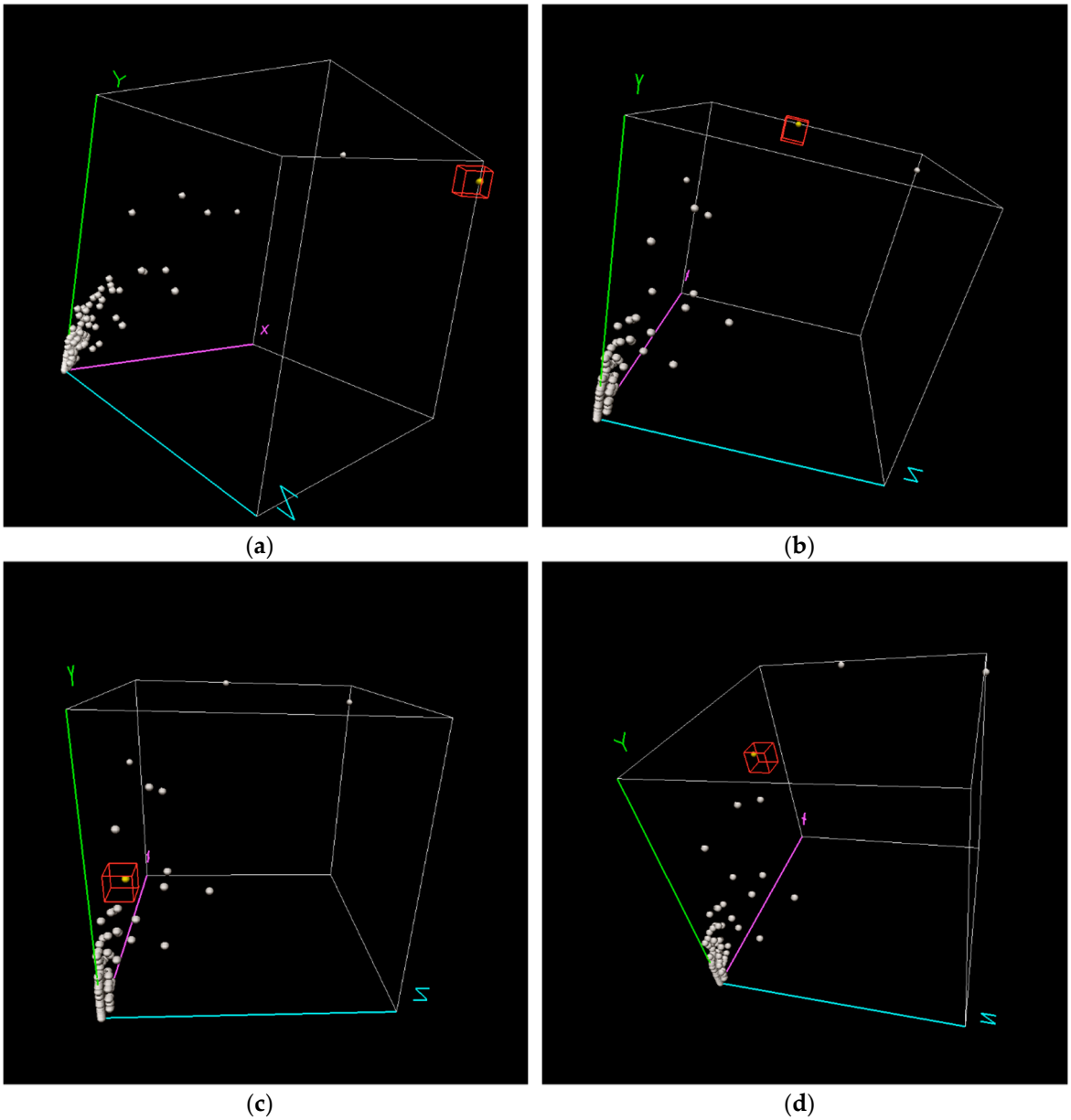

Figure 10. *Cont.*

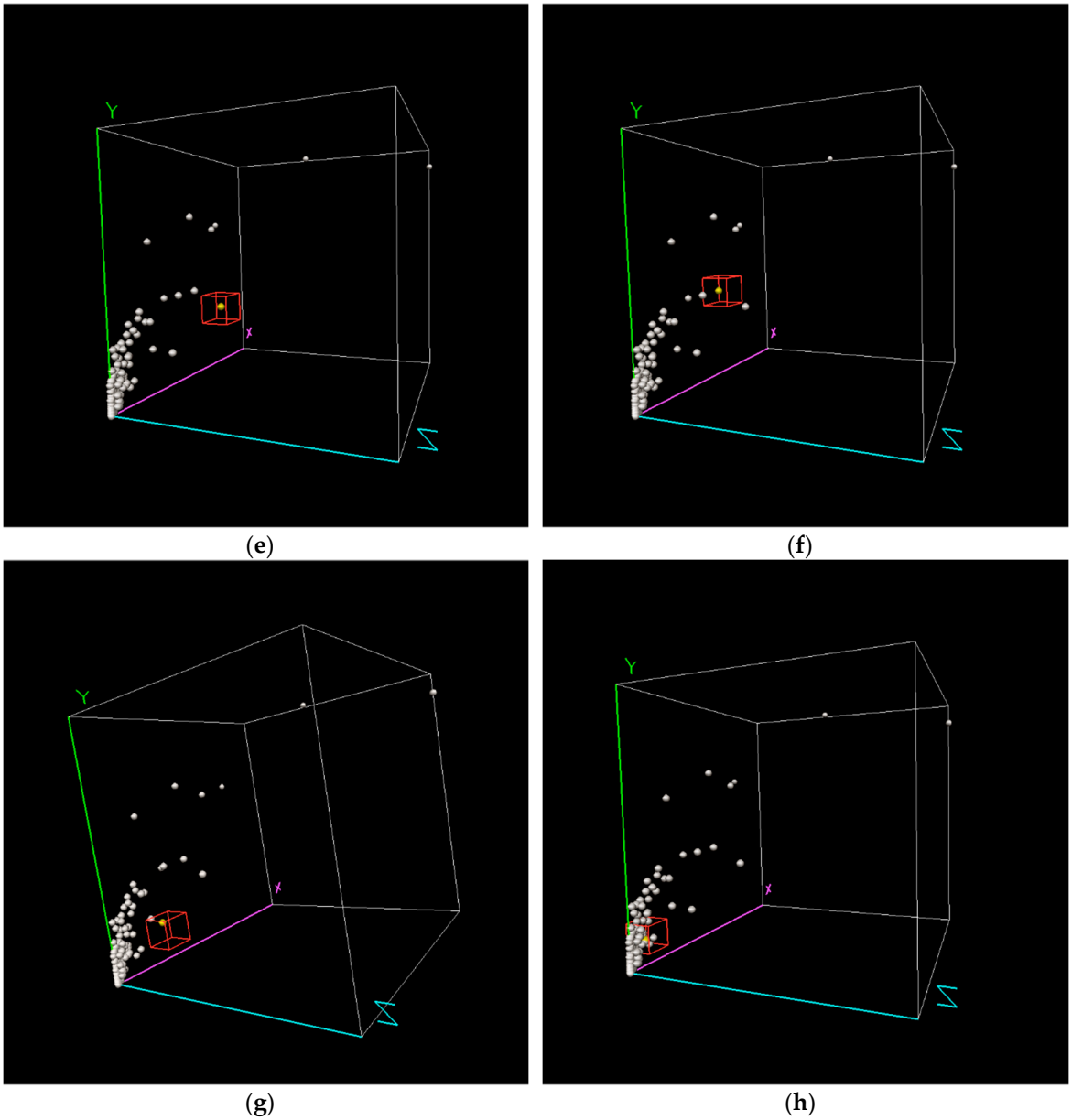

Figure 10. 3D scatter plot for eight representative cities. For each axis: X: Fiscal expenditure in science; Y: Gross domestic product (GDP); Z: Top university; (**a**) Beijing; (**b**) Shanghai; (**c**) Hangzhou; (**d**) Shenzhen; (**e**) Nanjing; (**f**) Wuhan; (**g**) Xi'an; (**h**) Guiyang.

5.3.2. Type II: Hangzhou and Shenzhen

Hangzhou and Shenzhen were the cities that were competing with Beijing and Shanghai. Both of them have a developed economy, good infrastructure and industrial base. The Shenzhen Hi-Tech Industrial Park was founded in 1996. Shenzhen was the first Chinese "special economic zone" and is also the incubation ground for many startups. There are a significant number of giant and successful hi-tech companies in Shenzhen, which bring experience and the milieu of innovation to the city. Hangzhou is the capital city in Zhejiang province, and it became an emerging technology hub in recent years. Because few top universities are in these two cities, Hangzhou and Shenzhen released policies to attract professionals, scholars, and graduates. Fortunately, the cities are in the Yangtze River Delta Region and the Pearl River Delta Region, where the human capital is more abundant than other regions.

5.3.3. Type III: Nanjing, Wuhan, and Xi'an

Nanjing, Wuhan, and Xi'an are the capital cities of Jiangsu, Hubei, and Shaanxi provinces, respectively, which have the best resources in the region. They possess top universities and have well-managed infrastructure. However, representative industry in these cities did not really match big data industry, so without a solid industrial base, they could only develop big data industry by following favorable industrial policies that may allow for the construction a new industry agglomeration in the cities (Table 6).

Table 6. Representative industry—Nanjing, Wuhan, and Xi'an.

	Nanjing	Wuhan	Xi'an
Representative industry	electronic, automobile, petrochemical, iron and steel	iron and steel, automobile, photoelectron, biochemical, food	automobile, aerospace, chemical, food, new energy

5.3.4. Type IV: Guiyang

Although Guizhou is one of the poorest provinces in China, it is undeniable that Guiyang had a rapid development in recent years. In 2016, it was named as China's Best-Performing City [39]. From 2014, big data industry began to bloom in Guiyang. China's first Big Data Industry Development Agglomeration District, the first Big Data Transaction Institute, and the first Big Data Strategy Key Lab are all located in Guiyang. Furthermore, the local government also introduced supporting policies for big data enterprises. The policies include the supporting of founding, operating, innovation, talent, and so on. However, for Guiyang, there is still a long way to go. To strengthen the weak links, such as infrastructure and talent, is the urgent task to develop the city as a 'black horse' in the big data industry.

6. Discussion

In this research, it seemed that half of the provinces were at the developing phase and the agglomerations only exist in minority of these cities. From the analysis of the phenomenon of big data industrial agglomeration in China, it could be found that there was imbalanced development between east and west, and comparatively large scale of agglomerations only existed in Beijing-Tianjin, the Yangtze River Delta Region, and the Pearl River Delta Region. It seemed that the agglomeration was matched with the development of economy on the surface, but from the view of city or county, the economy only appeared in a few districts, with some cities that had similar economic development showing a totally different development level for the big data industry. Moreover, the industrial fault lines around the agglomeration region and many industrial islands scattered across the nation reflected that the big data industry was only at the embryonic stage in China. The agglomeration of the industry will bring the city better resources and development of the industry, and it will also draw on the resources from surrounding cities that may cause industrial fault lines. Furthermore, there were only several cities in the Yangtze River Delta Region and the Pearl River Delta Region that showed spatial spillover, which meant there was still room for China to promote the development of big data industry. It has been noted that cities that have big data industry agglomerations are dispersed with the phenomenon of absorbing the human talent and capital from the surrounding regions, which leads to unbalanced regional development.

According to the analysis of major factors in agglomeration, fiscal support and the gathering of top universities are the main positive factors in appealing to big data enterprise and the generation of industrial agglomeration, while local markets had a lower effect on the development. Government's support provides a region which has a relatively undeveloped economy and talent market with the conditions to achieve a breakthrough in the big data industry. Munificent policy attracts entrepreneurs, high-quality labor, and capital enters into the market, which is a feasible approach to the formation

of an industrial agglomeration with a virtuous cycle. As big data industry relies on high-technology people and knowledge distribution, having top universities resolves the critical problem of human talent. Human resource to big data industry is what soil and environment are to agriculture. Top universities gather a great number of scholars, researchers, and graduates, a strong attraction for a lot of innovative and mutual big data companies to locate their business locally. This observation is close to the vital function of innovation in Porter's diamond model where innovation is one of the ways to sustainable productivity growth [40].

By reviewing the development of big data industry in China, there were roughly four kinds of cities that would have relatively higher development potential, of which some of the cities had shown high-level agglomeration. However, other factors might still limit the stable development of big data industry, such as high densities of people and high cost. Some cities were performing well among the 214 cities, although they each had their own weakness. Finding the key features of developing a big data industry will help the city to improve the industry and generate more industrial agglomeration. Another kind of cities was also identified and these cities did not have a good industrial base, but they depended on the guidance of policy and their specific advantages to attract investors and entrepreneurs. If these cities could seize the opportunity, they will also be able to develop big data industry and generate industrial agglomeration, which in turn can become the driving force to their surrounding cities. If a city does not have inherent advantages, such as high-quality labor, a developed economy or good base of the industry, but has relatively low house and electricity prices with preferential policies and a suitable natural environment, it can also have attractions for the field of data storage and big data transaction in the big data industry. As for cities with less development, they tended to be more active in building up a big data industry than the cities that were relatively more development and more matured industries. Moreover, giant enterprises have a great need of human capital, which is also the bane of the big data industry. Multiple factors systematically form a mechanism to develop a big data industry, which indicates that after having a breakthrough in the big data industry, the more critical task is to remedy the situation in the weakness and build a healthy environment for the industry's growth.

This research is the first study that employed the spatial analysis approaches to analyze big data industry agglomeration in China, and will encourage more studies in this direction. In addition, this research can also contribute to the discussion of big data industry [41] research from a different perspective.

On the other hand, there are also a few limitations to this research. First, because the state government and local government introduced a lot of favorable policy to promote the development of the big data industry, the big data industry might also have some recent changes. There is a lack of statistical data for the specific industry contributions to the economy, which cannot be precisely reflected by this study due to the data limitation. Defining the mechanism from agglomeration to the growth of the big data industry will be our future work based on data that can truly reflect the reality. Second, due to the large scale of the study area, the research could not go through the internal structure of all the cities to have a more detailed and micro-level understanding. Both the limitations could be addressed in our future research once the datasets are available. In addition, a more in-depth and specific qualitative perspective will also be the direction of our future research on this subject.

7. Conclusions

Nowadays, big data plays a more and more critical role in our life and production. Scholars, government, and citizens are concerned about the development and industrial agglomeration of the big data industry. This study analyzed the phenomenon of industrial agglomeration of the big data industry in China from a geographer's perspective. We employed a few spatial analysis approaches to look into the characteristics of agglomeration from various levels of data based on the approaches of kernel density analysis, global Moran's I, LISA clusters, spatial lag/error models, and other GIS tools. In detail, based on the existing literature and datasets, the study examined and determined four

important factors in city-level industrial agglomeration of big data. In addition, from the perspectives of various factors, this research also reviewed the problems of industrial development in each city and summarized four types of representative cities according to the current situation and recent development trend. Finally, the study clarified and analyzed the strengths, weaknesses, and industrial structure of each type of city, respectively, and provided advice and solutions to address problems that the cities have encountered. There are several valuable findings for China to develop the big data industry sustainably:

(1) The ongoing development of big data industry was still at a primary stage, with scarce agglomeration over the nation and with weak space spillover among the neighboring cities, although the trend and the benefits of industrial agglomeration have appeared;
(2) Fiscal support and the high level human capital are the two main factors to build the big data industry and its industrial agglomeration;
(3) For the cities that have a comparatively developed big data industry, their driving force in surrounding cities, such as professional talent, knowledge, advanced technology, and equipment, will make them play an important role in developing big data industry regionally and nationally;
(4) We should not ignore the potential of those relatively undeveloped regions, because they also possess particular value for the big data industry.

Author Contributions: Conceptualization, K.C.; Data collection, Y.L.; Formal analysis, Y.L. and K.C.; Investigation, K.C.; Methodology, Y.L. and K.C.; Resources, K.C.; Supervision, K.C.; Visualization, Y.L.; Writing—original draft, Y.L. and K.C.; Writing—review & editing, Y.L. and K.C.

Funding: This research was supported by Xiamen University of Technology Digital Fujian Big Data Research Institute of Natural Disaster Monitoring Open Fund (NDMBD2018006) and Singapore Ministry of Education (MOE) Academic Research Fund Tier 1 Grant (R-109-000-229-115).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Vesset, D. Worldwide big data technology and services 2012–2015 forecast. *IDC Rep.* **2012**, *233485*.
2. Ministry of Industry and Information Technology (MIIT). *Report of Users of Telephone and Broadband by Province in September 2017*; Ministry of Industry and Information Technology of People Republic of China: Beijing, China, 2017.
3. The Shanghai Cooperation Organization. *Action Outline for Promoting the Development of Big Data*; The Shanghai Cooperation Organisation: Shanghai, China, 2015.
4. Chen, X. *Success in Big Data: The Bluebook of Big Data Development in China*; Publishing House of Electronic Industry: Beijing, China, 2017.
5. Beier, G.; Niehoff, S.; Ziems, T.; Xue, B. Sustainability Aspects of a Digitalized Industry—A Comparative Study from China and Germany. *Int. J. Precis. Eng. Manuf. Green Technol.* **2017**, *4*, 227–234. [CrossRef]
6. Jin, H. Review of the Theory for Industrial Agglomeration. *Times Financ.* **2015**, *11*, 192–193.
7. Dos Santos Silvestre, B.; Dalcol, P.R.T. Geographical proximity and innovation: Evidences from the Campos Basin oil & gas industrial agglomeration—Brazil. *Technovation* **2009**, *29*, 546–561.
8. Fan, C.C.; Scott, A.J. Industrial agglomeration and development: A survey of spatial economic issues in East Asia and a statistical analysis of Chinese regions. *Econ. Geogr.* **2003**, *79*, 295–319. [CrossRef]
9. Malmberg, A.; Maskell, P. Towards an explanation of regional specialization and industry agglomeration. *Eur. Plan. Stud.* **1997**, *5*, 25–41. [CrossRef]
10. O'Donoghue, D.; Gleave, B. A note on methods for measuring industrial agglomeration. *Reg. Stud.* **2004**, *38*, 419–427. [CrossRef]
11. Storper, M.; Christopherson, S. Flexible specialization and regional industrial agglomerations: The case of the US motion picture industry. *Ann. Assoc. Am. Geogr.* **1987**, *77*, 104–117. [CrossRef]
12. Boudier-Bensebaa, F. Agglomeration economies and location choice: Foreign direct investment in Hungary 1. *Econ. Trans.* **2005**, *13*, 605–628. [CrossRef]

13. Du, J.; Lu, Y.; Tao, Z. Economic institutions and FDI location choice: Evidence from US multinationals in China. *J. Comp. Econ.* **2008**, *36*, 412–429. [CrossRef]
14. Guimaraes, P.; Figueiredo, O.; Woodward, D. Agglomeration and the location of foreign direct investment in Portugal. *J. Urb. Econ.* **2000**, *47*, 115–135. [CrossRef]
15. Weber, A. *Theory of the Location of Industries*; University of Chicago Press: Chicago, IL, USA, 1929.
16. Hoover, E.M. *Location Theory and the Shoe and Leather Industries*; Harvard University Press: Cambridge, MA, USA, 1937.
17. Marshall, A. *Principles of Economics: An Introductory Volume*; Macmillan London: London, UK, 1937.
18. Scott, A.J. Flexible production systems and regional development: The rise of new industrial spaces in North America and Western Europe. *Int. J. Urb. Reg. Res.* **1988**, *12*, 171–186. [CrossRef]
19. Lundvall, B.-Å. Explaining inter-firm cooperation and innovation: Limits of the transaction cost approach. In *Explaining Inter-Firm Cooperation and Innovation*; Routledge: Abingdon-on-Thames, UK, 1992.
20. Markusen, A. Sticky places in slippery space: A typology of industrial districts. In *The New Industrial Geography*; Routledge: Abingdon-on-Thames, UK, 2002; pp. 120–146.
21. Li, C. The Analysis of the Government Function in Industrial Agglomeration Development. *J. Part. Sch. Cent. Comm. CPC* **2009**, *3*, 42–47.
22. Li, X.; Li, Z.; Rong, J. *The Review of Contemporary Foreign Economists in Market Economics*; CPC Central Party School Press: Beijing, China, 1994.
23. Lei, T. *Research on the Influence Factors of the Development of Big Data Industry in China*; Beijing Jiaotong University: Beijing, China, 2017.
24. Wang, Q.; Cui, W.; Qi, W. Design and Application of Regional Competitiveness Model of Big Data Industry. *Electron. Sci. Technol.* **2017**, *5*, 109–113.
25. Zhang, M. *Studies on Industrial Agglomeration and Reginal Developing*; Economic Press China: Beijing, China, 2008.
26. Audretsch, B. Agglomeration and the location of innovative activity. *Oxf. Rev. Econ. Policy* **1998**, *14*, 18–29. [CrossRef]
27. Wang, X. *Study on the Spatial Structure and Collaboration Network of Regional S & T Development Based on GIS*; Dalian University of Technology: Dalian, China, 2009.
28. Zheng, F. *Regional Structure in Information Age*; The Commercial Press: Beijing, China, 2004.
29. Varga, A.; Pontikakis, D.; Chorafakis, G. Metropolitan Edison and cosmopolitan Pasteur? Agglomeration and interregional research network effects on European R&D productivity. *J. Econ. Geogr.* **2012**, *14*, 229–263.
30. Han, Q. A Research on Cloud Computing Industrialization Factor in China. *Rev. Ind. Econ.* **2014**, *7*, 11–18.
31. Sun, X. *Study on Research and Development Industrial Agglomeration in the Yangtze River Delta Region*; Shanghai Normal University: Shanghai, China, 2016.
32. Silverman, B.W. *Density Estimation for Statistics and Data Analysis*; Routledge: Abingdon-on-Thames, UK, 1998.
33. Odland, J. *Spatial Autocorrelation*; Sage Publications: Newbury Park, CA, USA, 1988; Volume 9.
34. Upton, G.; Cook, I. *Oxford Dictionary of Statistics*; Oxford University Press: Oxford, UK, 2006.
35. Moran, P.A. The interpretation of statistical maps. *J. Royal Stat. Soc. Ser. Methodol.* **1948**, *10*, 243–251. [CrossRef]
36. Anselin, L. Local indicators of spatial association—LISA. *Geogr. Anal.* **1995**, *27*, 93–115. [CrossRef]
37. Anselin, L. Lagrange multiplier test diagnostics for spatial dependence and spatial heterogeneity. *Geogr. Anal.* **1988**, *20*, 1–17. [CrossRef]
38. Kelejian, H.H.; Prucha, I.R. A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *J. Real Estate Finance Econ.* **1998**, *17*, 99–121. [CrossRef]
39. Wong, P.; Lin, M.C.Y.; Jackson, J. *Best Performing Cities, China 2016*; Milken Institute: Santa Monica, CA, USA, 2016.

40. Porter, M.E. The Role of Location in Competition. *Int. J. Econ. Bus.* **1994**, *1*, 35–40. [CrossRef]
41. Lytras, M.D.; Raghavan, V.; Damiani, E. Big data and data analytics research: From metaphors to value space for collective wisdom in human decision making and smart machines. *Int. J. Semant. Web Inf. Syst.* **2017**, *13*, 1–10. [CrossRef]