

Big Data Industrial Agglomeration Promoting Regional Innovation: Comparison between Guangzhou and Zhaoqing in China

Yuliang Zhou^{1,a}, Jinfeng Li^{2,b*}

¹School of Finance, Guangdong University of Finance & Economics, Guangzhou, China

²School of Liberal Arts, Zhaoqing University, Zhaoqing, China

Abstract—This paper selects the data of big data industry in China's "Guangzhou Development Zone Big Data Industrial Park" and "Zhaoqing Big Data Cloud Service Industrial Park" from 2014 to 2018, uses the improved knowledge production function to establish an OLS model, and compares the impact of MAR and Jacobs external aggregation on the R&D input and patent output in Guangzhou and Zhaoqing. It is found that: (1) MAR externality is not conducive to the technological innovation of the two cities, and has a stronger negative effect on innovation in Zhaoqing; Jacobs externality can actively promote the innovation of the two cities, and has a stronger positive effect on innovation in Guangzhou. (2) In the impact of Jacobs externality on innovation output of the two cities, R&D plays a part of intermediary effect, and the effect on Guangzhou is stronger; in the impact of MAR externality on innovation output of the two cities, R&D only plays a part of negative intermediary effect in Zhaoqing. The conclusions show that the MAR and Jacobs agglomeration in big data industry all play more effective roles in promoting technological innovation in economically developed cities.

1 Introduction

The basis of big data industry is cloud computing, internet of things and other key technologies, which is the development direction of scientific and technological innovation in the new era. Big data industry and technological innovation complement each other. On the one hand, a large number of scientific and technological innovation practices have accumulated a large number of data, providing materials for the construction of big data platform and big data research; on the other hand, the construction of big data platform reduces the exploration cost of technological innovation, and provides convenience for technological innovation.

Practice shows that the development of big data industry help to guide the input and output of regional technological innovation, but the construction of big data industry is still in the primary stage, there are not many practical experiences and research results for reference. Therefore, this paper discusses the impact of big data industry on the input and output of regional innovation by referring to the relevant research of traditional industrial agglomeration and regional innovation. In the study of the influence of traditional industrial agglomeration on regional innovation, there are two common problems. (1) Generally, it is based on the

"innovation of two-stage value chain" (Guan and Chen, 2010), but the relationship between innovation input and innovation output is rarely discussed (Feldman and Audretsch, 1999; Paci and Usai, 1999), that is to say, the endogenous problems in the two stages need to be solved better. (2) There are regional differences in the development level of big data industry. So, are there differences in the impact of relevant variables based on big data industry on regional innovation? It is necessary to clarify this issue.

2 Data, variables and models

2.1 Data Sources

In this paper, the sample range is limited to Guangdong province, China, which is a strong province of big data industry development in China. At present, there are 16 provincial big data industrial parks in Guangdong province. In order to find out whether there are differences in the impact of big data industry on regional innovation, this paper makes a comparative analysis of two cities with large economic strength gap in the province. We choose "Guangzhou Development Zone Big Data Industrial Park" and "Zhaoqing Big Data Cloud Service Industrial Park" as the research objects. The reason why we choose these two cities is that they are

^ahefly_2007@126.com

*Corresponding author: ^blijinfeng_1989@163.com

the first tier cities and the third tier cities respectively, and there are great differences in their technological innovation capabilities. Then, exploring the impact of big data on the technological innovation of the two cities may have different economic results, which is an interesting comparative perspective.

The data of big data industry comes from the website of China's "State Intellectual Property Office" ^①. Advanced retrieval is carried out in the "patent retrieval" module to obtain the relevant data of the two industrial parks in 2014-2018. The key words are: ("big data" or "map reduce" or "hadoop" or "data mining" or "cloud computing" or "structured" or "K-means" or "spark" or "mahout" or "slope" or "association rule" or "multi fork tree"). Other data are from Guangzhou statistical yearbook, Guangzhou science and technology statistical yearbook, Zhaoqing statistical yearbook and Zhaoqing science and technology statistical yearbook.

2.2 Variables Setting

2.2.1 Explanatory variables

Glaeser et al. (1992) puts forward "MAR externality" and "Jacobs externality": "MAR externality" emphasizes specialized knowledge spillover within the same industry, that is, knowledge spillover within the same industry is an important source to promote regional innovation and economic growth; "Jacobs externality" considers that knowledge spillover between different industries is the main driving force to promote regional innovation and economic growth. Based on this, this paper focuses on the difference between the specialization and diversity externalities of big data industry in the two cities on the role of local technological innovation. The calculation formula of specialized externality variables is shown in formula (1):

$$MAR_{i,t} = \sum_k |S_{i,k,t} - S_{j,k,t}| \quad (1)$$

Where, t , i and k respectively represents year, city and industry. $S_{i,k,t} = E_{i,k,t} / \sum_k E_{i,k,t}$, $S_{j,k,t} = \sum_{j \neq i} E_{j,k,t} / \sum_k \sum_{j \neq i} E_{j,k,t}$. $E_{i,k,t}$ represents the total output value of industrial production. $MAR_{i,t}$ measures the difference between the specialization structure of urban industrial knowledge and the average level of the other city. The larger the variable value is, the stronger the effect of industrial knowledge specialization is. If the MAR is significantly positive, it shows that the specialization of local industry can help to improve the level of local innovation.

Refer to the measurement method of industrial diversification level (Duranton and Puga, 2000), the calculation formula of Jacobs index is as shown in formula (2):

$$Jacobs_{i,t} = 1 / \sum_k |S_{i,k,t} - S_{k,t}| \quad (2)$$

Among them, $S_{i,k,t} = E_{i,k,t} / \sum_k E_{i,k,t}$, $S_{k,t} = \sum_i E_{i,k,t} / \sum_k \sum_i E_{i,k,t}$. The meaning of $E_{i,k,t}$ is the same as (1). $Jacobs_{i,t}$ is the level of industrial diversification. The greater the value, the stronger the diversification and the greater the degree of decentralization. If the coefficient of Jacobs is significantly positive, it indicates that the level of local industrial diversification help to improve the level of innovation in the region.

2.2.2 Control variables

(1) The patent application count in big data industry (represented by L). This variable can examine the stimulating effect of big data patent applications on promoting local R&D investment. (2) Fixed asset investment (represented by F).

2.2.3 Explained variable

We use local R&D input (R&D) and patent authorization (Lngpt) as explained variables, they embody the idea of two-stage value chain of regional innovation.

3 Empirical analysis

3.1 Improved Knowledge Production Model

This paper draws lessons from Kluge and Lehmann's (2013) department knowledge production function with constant returns to scale, and improves it as shown in formula (3) according to the research needs.

$$INV_{i,t} = A_{i,t}(X)L_{i,t}^\alpha F_{i,t}^\eta \quad (3)$$

Among them, $INV_{i,t}$ is innovation level, including R&D input and patent output. $L_{i,t}$, $F_{i,t}$ respectively represents the patent application counts for big data industry, the amount of fixed asset investment, and $A_{i,t}$ is the technical level, namely regional total factor productivity (TFP), which X is a variable parameter, α and η is the corresponding output elasticity.

The regional technological progress value $A_{i,t}(X)$ of formula (3) can not be directly observed. In this paper, with reference to Martin et al. (2011), it is assumed that the technological progress is in the form of formula (4):

$$A_{i,t}(X) = U_{i,t} e^{g(MAR_{i,t}, Jacobs_{i,t})} \quad (4)$$

Formula (4) shows that technological progress depends on the specialization $MAR_{i,t}$ and diversification $Jacobs_{i,t}$ of knowledge externality agglomeration, $U_{i,t}$ represents the regional characteristics. Among them, it represents the possible functional relationship between the degree of industrial specialization and diversification. In this paper, it is assumed that its logarithmic form is formula (5):

^① <http://pss-system.cnipa.gov.cn/sipublicsearch/portal/uiIndex.shtml>

$$g(MAR_{i,t}, Jacobs_{i,t}) = \beta \ln(MAR_{i,t}) + \gamma \ln(Jacobs_{i,t}) \quad (5)$$

Then, after the logarithm of equation (3), the general form can be obtained, as shown in equation (6):

$$\ln(Inv_{i,t}) = C_i + \beta \ln MAR_{i,t} + \gamma \ln Jacobs_{i,t} + \alpha \ln L_{i,t} + \eta F_{i,t} + \varepsilon_{i,t} \quad (6)$$

Equation (6) is similar to the Innovation Factor Mode (IFM) of Zhou and Li (2020), and its derivation process is consistent.

3.2 Effects of Big Data Industry on Regional Innovation

OLS regression is carried out for equation (6) by software Eviews 7.0, and the results are shown in Table 1:

Table1 Impact of big data industry on regional innovation

	Guangzhou		Zhaoqing	
	R&D input	Patent output	R&D input	Patent output
MAR	-0.578 (1.003)	0.126** (0.784)	0.317* (0.862)	0.391*** (0.557)
Jacobs	0.616** (1.337)	0.201** (1.098)	0.248** (0.832)	0.145* (0.815)
L	0.225*** (0.301)	0.403** (0.394)	0.213** (0.304)	0.138** (0.456)
F	0.144* (0.278)	0.093* (0.361)	0.086* (0.222)	0.050* (0.295)
Adj. R ²	0.279	0.255	0.369	0.387
F Value	31.654***	33.771***	24.156***	20.080***

Note: *, ** and *** are significant at the levels of 10%, 5% and 1% respectively.

Table 1 shows that:

(1) MAR externality has a negative impact on the R&D input and patent output of the two cities, indicating that the more specialized the big data industry is, the worse the local innovation level is. Due to the high similarity between industries and the lack of motivation to learn from each other, the level of R&D investment is also weak. In contrast, the negative impact is more obvious in Zhaoqing city, which shows that specialized agglomeration makes the innovation power of the small city weaker. Jacobs externality has a significant positive impact on the innovation input and output of the two cities, indicating that the more diversified the big data

industry is, the stronger the local innovation level is. Diversified industries promote the power of mutual learning, so R&D investment level is higher, and knowledge spillover further promotes patent output.

(2) The greater the patent application counts of big data industry, the stronger the R&D input and patent output capacity of the two cities, indicating that the innovation capacity of big data industry has led to the improvement of regional innovation level. In addition, the fixed asset input of big data industry has a significant positive impact on innovation input and output of the two cities, indicating that the fixed asset input of big data industrial park has a positive guiding effect on regional innovation, or the fixed asset input of big data industry innovation and regional innovation has a linkage effect.

3.3 Intermediary Effect of R&D

As mentioned above, R&D input in regional innovation will affect patent output, so there may be endogenous in the regression model. Therefore, this paper chooses R&D investment as the intermediary variable to judge whether the two kinds of big data industry cluster externalities have intermediary effects on regional innovation output. The methods of setting up the function of intermediary variable and judging the intermediary effect are as follows:

First, the basic explanatory variable has a significant impact on the intermediary variable; Second, the basic explanatory variable has a significant impact on the explained variable; Third, the intermediary variable is included in the explanatory variable, and the intermediary and explanatory variable have both significant impacts on the explained variable.

According to the coefficients and significance levels of the latter two explanatory variables, the intermediate effect results are confirmed. If the coefficient of the basic explanatory variable in the third step is smaller than that in the second step, there is partial intermediary effect; if the basic explanatory variable in the third step is not significant, there is complete intermediary effect (Baron and Kenny, 1986). As shown in equation (7), (8), (9):

$$\ln(R \& D_{i,t}) = C_i + \beta \ln MAR_{i,t} + \gamma \ln Jacobs_{i,t} + \alpha \ln L_{i,t} + \eta F_{i,t} + \varepsilon_{i,t} \quad (7)$$

$$\ln(Lngpt_{i,t+1}) = C_i + \beta \ln MAR_{i,t} + \gamma \ln Jacobs_{i,t} + \alpha \ln L_{i,t} + \eta F_{i,t} + \varepsilon_{i,t} \quad (8)$$

$$\ln(Lngpt_{i,t+1}) = C_i + \beta \ln MAR_{i,t} + \gamma \ln Jacobs_{i,t} + \ln(R \& D_{i,t}) + \alpha \ln L_{i,t} + \eta F_{i,t} + \varepsilon_{i,t} \quad (9)$$

The regression results are sort out as shown in Table 2:

Table 2 intermediary effects of R&D

	Guangzhou			Zhaoqing		
	<i>R&D</i>	<i>Lngpt_{t+1}</i> ^① (8)	<i>Lngpt_{t+1}</i> (9)	<i>R&D</i>	<i>Lngpt_{t+1}</i> (8)	<i>Lngpt_{t+1}</i> (9)
<i>MAR</i>	-0.396 (1.234)	-0.208** (1.052)	-0.271* (0.843)	-0.213* (0.886)	-0.376** (0.873)	-0.300** (1.118)
Jacobs	0.273** (0.777)	0.329** (0.712)	0.235** (0.860)	0.147** (0.538)	0.225* (0.403)	0.118* (0.462)
<i>R&D</i>	/	/	0.209*** (0.557)	/	/	0.123*** (0.364)
Adj. R ²	0.295	0.327	0.277	0.218	0.185	0.273
F Value	30.442***	31.671***	19.174***	25.910***	21.347***	19.685***

Note: for brevity, the regression results of control variables are omitted.

① Taking R&D investment as the current variable (t period, 2017), then, R&D investment has an impact on patent output (t+1 period, 2018), and there is a time sequence relationship. This setting can avoid endogenous problem.

Table 2 shows that: (1) for the regression results of Guangzhou, the influence of MAR externality on R&D is not significant, indicating that there is no intermediary effect. In contrast, for the regression results of Zhaoqing, MAR externality has significant influence on the explained variables, and the significance level of model (8) and model (9) is the same, what's more, the absolute value of the latter coefficient is smaller than the former, indicating that R&D plays a part of intermediary effect in the influence of MAR externality on patent output of Zhaoqing area. It needs to be noted that this intermediary effect is a negative effect, indicating that the higher the level of professional agglomeration of big data industry, the stronger the inhibition of R&D investment in relatively underdeveloped regions, which is not conducive to their patent output. Of course, the professional agglomeration of big data industry not only plays a role in innovation output through R&D, but also directly inhibits the exchange and learning between industries in relatively backward regions, which is not conducive to patent output.

(2) The Jacobs externality of big data industry has a significant positive impact on R&D input and patent output at the level of 0.10 or 0.05 respectively. Further judgment shows that: for two cities, the regression coefficient of Jacobs externality to model (9) is smaller than that of model (8), and the significance level is the same, indicating that R&D variable plays a part of intermediary effect about the impact of Jacobs externality on patent output. That is to say, the diversified big data industry agglomeration further promotes the patent output in part by stimulating the R&D input of local relevant industries, but the patent output is not all created by the R&D input, the diversified agglomeration itself can also stimulate the learning and exchange between local relevant industries, thus promoting the patent output.

4 Conclusions and policy implications

This paper selects two big data industrial parks in Guangzhou and Zhaoqing, Guangdong province, China as the research objects, and uses the improved knowledge production function to analyze the impact of the externality of specialized and diversified industrial agglomeration on the technological innovation of the two regions, and makes a comparison. It is found that: (1) MAR externalities are not conducive to the R&D input and patent output of the two cities, in which the negative impact on Zhaoqing city is stronger; on the contrary, Jacobs externalities have significant positive impact on the innovation input and output of the two cities, and there is no significant difference between the two cities. (2) The more patent applications in big data industry, the more stimulating innovation input and output of the two cities; the more fixed assets input, the more innovation input and output. (3) In the process of the impact of the externalities of two kinds of big data industry clusters on the innovation output of the two cities, the R&D investment has both intermediary effects. Obviously, the conclusions of this study support the theory of Jacobs' externality, which is consistent with the conclusions of Glaeser et al. (1992), Feldman and Audretsch (1999), Batisse (2002), Van der Panne (2004), Antoniotti and

Cainelli (2011), Tavassoli and Carbonara (2014), Murata et al. (2015), Zhou and Li (2020).

According to the above conclusions, the relevant policy recommendations are as follows: First, the specialized agglomeration of big data industry is not conducive to regional scientific and technological innovation, and excessive specialization should be avoided in the process of developing big data industry, especially in the relatively backward regions. Second, the diversity agglomeration of big data industry is conducive to regional innovation, especially in the relatively developed regions, and should promote the development of industrial diversity. Third, in the process of regional innovation development, the government should reasonably guide and support the R&D investment, so as to make the process of big data industry agglomeration promoting regional innovation output more smoothly.

In this paper, only two cities are selected as the research objects. The following research should examine the impact of big data industry agglomeration on regional innovation and its spatial effect.

Acknowledgment

Thank Professor Zhang Caijiang of the School of Economics and Commerce, South China University of Technology for guidance on modeling in this paper.

Fund

Characteristic innovation projects of general universities in Guangdong province in 2019: "Research on the influence of social capital on the bank loan of SMEs based on three stages of innovation" (2019WTSCX036).

References

1. J. C. Guan, K. H. Chen, "Measuring the Innovation Production Process: A Cross-Region Empirical Study of China's High-tech Innovations", *Technovation*, 2010, vol. 30, pp. 348-358.
2. M. P. Feldman, D. B. Audretsch, "Innovation in Cities: Science-based diversity, Specialization and localized Competition", *European Economic Review*, 1999, vol. 43, pp. 409-429.
3. R. Paci, S. Usai, "Externalities, Knowledge Spillovers and the Spatial Distribution of Innovation", *GeoJournal*, 1999, vol. 49, pp. 381-390.
4. E. L. Glaeser, H. Kallal, J. A. Scheinkman, et al. "Growth in Cities", *Journal of Political*, 1992, vol. 100, pp. 1126-1152.
5. G. Duranton, D. Puga, "Diversity and Specialization in Cities: Why, Where and When Does it Matter?", *Urban Studies*, 2000, vol. 37, pp. 533-555.
6. J. Kluge, R. Lehmann, "Marshall or Jacobs? New insights from an interaction model", *Review of Regional Research*, 2013, vol. 33, pp. 107-133.

7. P. Martin, T. Mayer, F. Mayneris, “Spatial concentration and plant-level productivity in France”, *Journal of Urban Economy*, 2011, vol. 69, pp. 182-195.
8. Y. L. Zhou, J. F. Li, “Impact of Industrial Agglomeration on Regional Innovation: Innovation Factor Model and Empirical Test”, *International Conference on Economic Management and Model Engineering*, 2020, vol. 57, pp. 255-259.
9. R. M. Baron, D. A. Kenny, “The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical consideration”, *Journal of Personality and Social Psychology*, 1986, vol. 51, pp. 1173-1182.
10. C. Batisse, “Dynamic externalities and local growth: A panel data analysis applied to Chinese provinces”, *China Economic Review*, 2001, vol. 13, pp. 231-251.
11. G. Van der Panne, “Agglomeration externalities: Marshall versus Jacobs”, *Journal of Evolutionary Economics*, 2004, vol. 14, pp. 593-604.
12. R. Antonietti, G. Cainelli, “The role of spatial agglomeration in a structural model of innovation, productivity and export: a firm-level analysis”, *Annals of Regional Science*, 2011, vol. 46, pp. 577-600.
13. S. Tavassoli, N. Carbonara, “The Role of Knowledge Variety And Intensity for Regional Innovation”, *Small Business Economics*, 2014, vol. 2, pp. 493-509.
14. Y. Murata, R. Nakajima, R. Okamoto, R. Tamura, “Localized knowledge spillovers and patent citations: a distance-based approach”, *Review of Economics and Statistics*, 2015, vol. 96, pp. 967-985.