A Case Study: Exploring Industrial Agglomeration of Manufacturing Industries in Shanghai using Duranton and Overman's K-density Function

Siyu Tian [a], Jimin Wang [a], Zhipeng Gui [a,b,c] *, Huayi Wu [b,c], Yuan Wang[b,c]

[a] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China - (wilhelm_tian, jimin, zhipeng.gui)@whu.edu.cn
[b] The State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China - (zhipeng.gui, wuhuayi, yuan.wang)@whu.edu.cn
[c] Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan, China - (zhipeng.gui, wuhuayi, yuan. wang)@whu.edu.cn

Commission V, WG V/4

KEY WORDS: Industrial Agglomeration, Point Pattern Analysis, K-density function, Geoprocessing Services, Decision Support in Smart City

ABSTRACT: There has wide academic and policy attention on the issue of scale economy and industrial agglomeration, with most of the attention paid to industrial geography concentration. This paper adopted a scale-independent and distance-based measurement method, K-density function or known as Duranton and Overman (DO) index, to study the manufacturing industries localization in Shanghai, which is the most representative economic development zone in China and East Asia. The result indicates the industry has a growing tendency of localization, and various spatial distribution patterns in different distances. Furthermore, the class of industry also show significant influence on the concentration pattern. Besides, the method has been coded and published on *GeoCommerce*, a visualization and analysis portal for industrial big data, to provide geoprocessing and spatial decision support.

1. INTRODUCTION

Industrial agglomeration is now a very heated topic in both academia and politics. At least since 90s', the tendency of industrial agglomeration in certain areas seems to have been a conspicuous spatial pattern of most industries. Huge success of such industry clusters like the Silicon Valley has fascinated both economists and geographers. More recently, successful examples of these industry clusters have caught the attention of policy makers in mainland China. Following Shenzhen and some other cities' great success in *the Reform and Opening-up*, clusters are seen by many policy makers as the destined path for fast regional economic development, and more policies alike are put forward and carried out such as *Shanghai Free Trade Zone*, etc. Nevertheless, despite such success, the tendency for firms to co-locate and the real effects of these policies still raise a number of questions, including whether the agglomeration is beneficial to a certain industry's development, how distinct is the agglomeration and whether a certain industry has a tendency to co-locate in a certain local area, etc. These questions are crucial to answer for the benefit of measuring our policies effective or not and how to reasonably make new policies.

In the last ten years, most relevant research explanations of spatial clustering or industrial agglomeration rely on some form of external increasing returns which also figure prominently in theories of international trade, industrial organization and economic growth (Duranton and Overman, 2005). However, theoretical pre-assumptions and models seem to be a little far from practice, as we have learnt what concerns industrial agglomeration is far beyond the context of economy and geography (He et al, 2012; He et al, 2007; Yuan et al, 2010). Additionally, studies based on high accuracy micro-geography data is also lacked internationally (Scholl and Brenner, 2011). Thus, in this paper, to step back from theoretical concerns and perform a positivism study that concentrates on the spatial facts, a statistic methodology unbiased with respect to spatial scales, DO index, i.e. a statistic index introduced by Duranton and Overman in 2005, is adopted to measure the agglomeration. Using the data of Shanghai's manufacturing industrial registration, we tend to answer a series of questions: how strong is the tendency for agglomeration in manufacturing industries? Dose the truth suggests as the policy makers have planned? And put forward our own explanations with certain suggestions.

2. EXPERIMENT DATA

The dataset we used in the experiment is from a micro-geographical national database covers firm registration data from 1950 to 2015 which contains firms' names, industry categories (i.e. industry classification as per GB/T4754-94), legal person with contacts, and spatial location, etc. The raw data is collected by local *Administration for Industry and Commerce* and produced after processing by a workflow of data imputation based on naive-Bayes theory and geocoding tools that completes the missing values of industry level and spatial location (Li et al., 2017). After the data imputation, the complete rate of industry categories raised to 99.97% from 62.77%, and the accuracy of that is tested beyond 75%, which can be considered to meet the demands of this study.

For the need of our study, we use a subset of the original database that contains columns shown in Table 1 below. The data is of good quality in both quantity and geographic precision. This dataset covers manufacturing registration data in Shanghai from 1990 to 2010, with only year 2008 missing. Given the significant function and high status of Shanghai in China's economy, the point of choosing Shanghai is to make the study relatively representative. The data of Shanghai has more records in manufacturing industry compared to other huge metropolitans like Beijing or Guangzhou. Such data intensity can also reduce the impact of noise in the dataset as much. The specific column names with their explanations are shown in Table 1.

* Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-2-W7-149-2017 | © Authors 2017. CC BY 4.0 License.

Column	*Explanation*
ID	Identification of each row of record
InstitutionName	Name of the registered firm
RegisterDate	Registration date of the firm
Category in LV1	Industrial level 1 (per GB/T4754-94)
Category in LV2	Industrial level 2 (per GB/T4754-94)
Category in LV3	Industrial level 3 (per GB/T4754-94)
Latitude	Latitude of the registration position
Longitude	Longitude of the registration position

Table 1. Columns of used dataset

The spatial distribution of data used in this study is visualized by *GeoCommerce*, a Web GIS data visualization system we developed to support big geo-data spatial pattern discovery and analysis, in Figure 1.

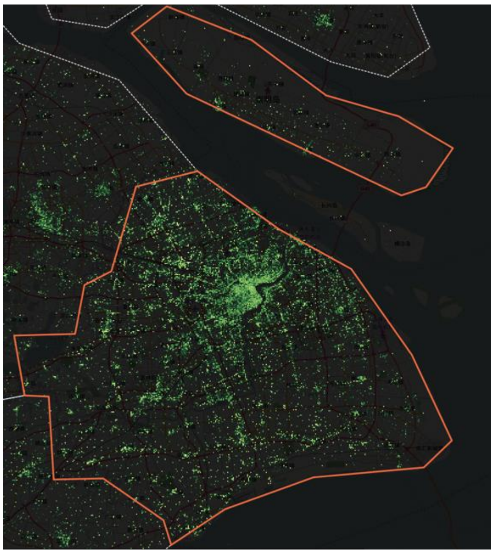


Figure 1. The spatial distribution of manufacturing registration data in Shanghai from year 1990 to 2010

In Figure 1, the original data is showcased as point vector with adjustable brightness in the visualization portal. In this visualized map, we can spot some simple spatial patterns like firm owners tend to register their business alongside the Yangtze River in the central functional zone of the city. While there exists certain firms located away from the city center like in Chongming Island or the countryside. In general, firms in Shanghai do seem to have a tendency to co-locate but in a rather limited space within the city while some firms appear to geographically spread out to the surroundings of the city or seemingly disperse from each other. What's interesting here is that the pattern seems to have drawn the outline and even road network of Shanghai city, which usually can only be detected and extracted from remote sensed images. This result reveals that with the development of social sensing technologies, big social and human activity data can also provide unique and useful data sources and approaches to support urban studies, such as urban function zone extraction and planning, urban road mapping, etc. When we try to explore deeper into the spatial distribution of the firm registration locations, patterns are more difficult to discern from a purely visual inspection of the data, thus requires us using statistic methodology to make more precise comparisons.

What might cause confusion is, our data is registration location data rather than real operating location data. That is, as a firm will consider more about policy and resource in choosing registration location, such data will more intuitively reflects the effects of policies and resources than real operating location data, which reaches the expectation of this study. More importantly, on a larger scale most firms have identical registration location with operating location. Considering the intensity and volume of our data, the real difference can be regarded as negligible noise in our study.

3. METHODOLOGY

The workflow of our study is shown in Figure 2. Firstly, we use the kernel estimation method in spatial analysis to smooth the distance of point pairs in a same industry. This is to introducing spatial autocorrelation to the study and reasonably consider the impact of a point with another and to the whole industry. Substituting the kernel estimated distances to K function formula, we can calculate the output of estimated K-densities. Then we construct our counterfactuals to testify the significance level of the output. By constructing a confidence band, of which the band width is calculated out by Monte Carlo simulation, we can decide whether an industry is localized or dispersed on a certain distance. Lastly, we compare and analyze the output of the process and draw our own conclusion and make explanations, and give our advices.

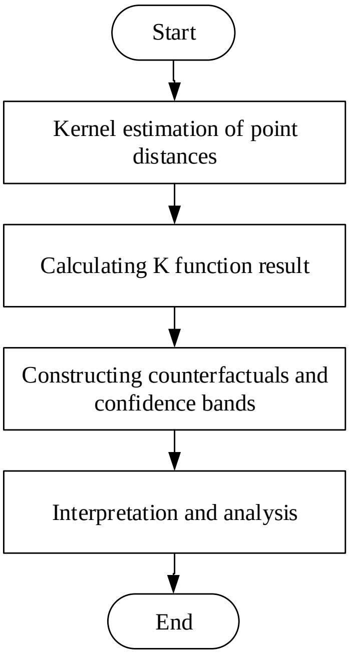

Start
Kernel estimation of point distances
Calculating K function result
Constructing counterfactuals and confidence bands
Interpretation and analysis
End

Figure 2. Workflow of data computation and analysis

3.1. Estimating K-densities

Consider all the entrants in the manufacturing industry. First, we calculate the Euclidian distance between every pair of entrants. For an industry with n entrants, there are $\frac{n(n-1)}{2}$ unique bilateral distances between entrants. The Euclidian distance is calculated through point latitude and longitude coordinates as formula (1).

$$C = \sin LatA * \sin LatB * \cos(LonA - LonB) + \cos LatA * \cos LatB$$
$$D = \frac{R * arccosC * \pi}{180°} \quad (1)$$

where A, B imply A(LatA, LonA) and B(LatB, LonB)
D is the value of Euclidian distance

Because Euclidian distance is only a proxy for true physical distance, we kernel-smooth to estimate the distribution of bilateral distances. More specifically, with n entrants, the estimator of the density of bilateral distances (henceforth K-density) at any distance d is calculated by formula (2)

$$\widehat{K}_d = \frac{1}{n(n-1)h}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} f\left(\frac{d-d_{ij}}{h}\right) \qquad (2)$$

where $\widehat{K}_d$ = K-density
n = total number of entrants
h = bandwidth set as per Section 3.4 (Silverman, 1986)
d = distance(scale) that the calculation is at
d_{ij} = Euclidian distance between entrants i, j
f(x) = $e^{\frac{-x^2}{2\sigma}}$, where σ is width of the function

With respect to the large amount of industries, in the experiment we perform the calculations on the Spark platform based on server cluster. Furthermore, for the ease of reuse and practical decision-making support, we package the K-density calculating API as web services published on *GeoCommerce*.

3.2. Constructing Confidence Bands

To testify the K-density results, we now need construct counterfactuals and confidence intervals.

3.2.1. Local Confidence Bands

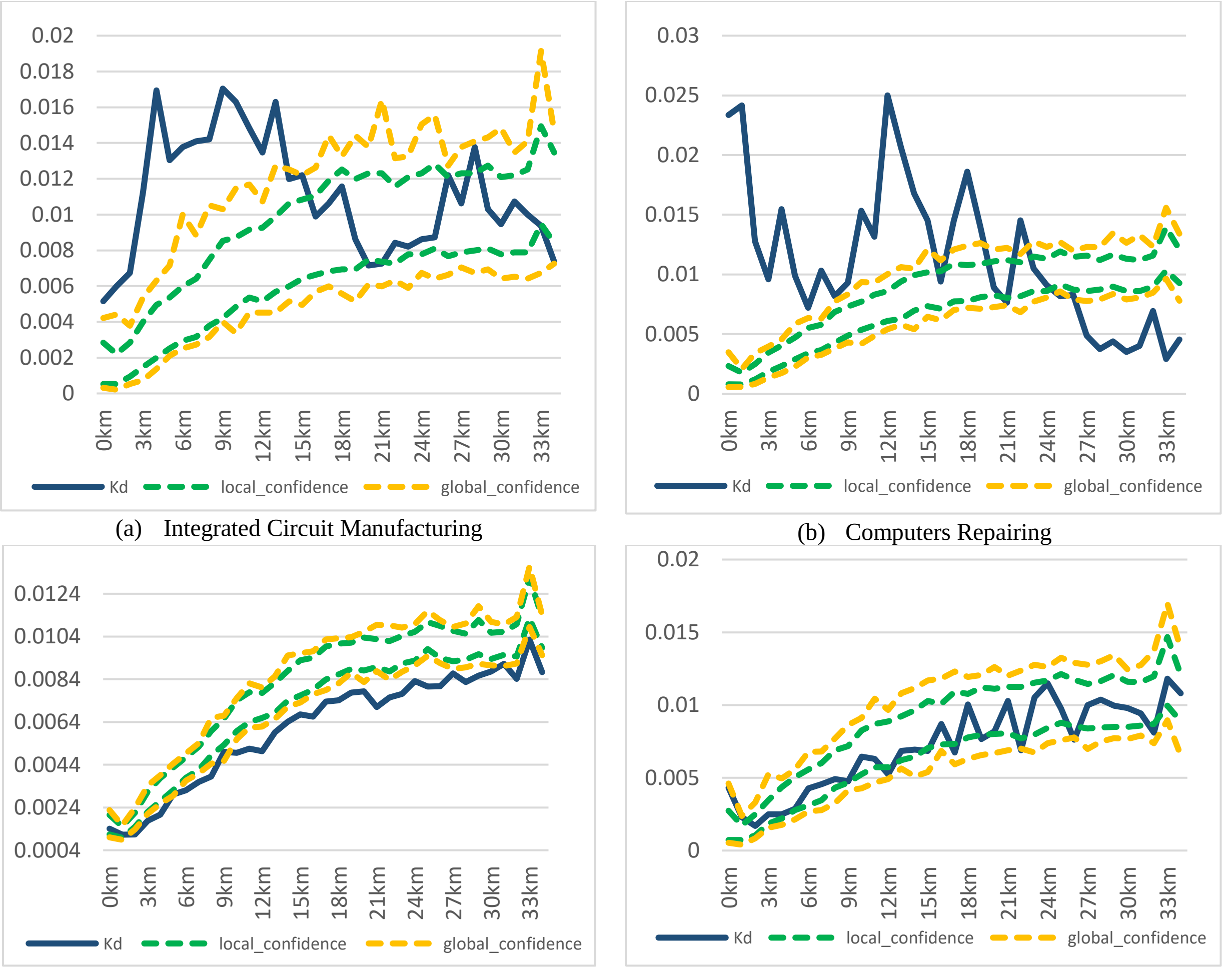
In our study, Monte Carlo Simulation is adopted to construct confidence bands. For an industry that has n entrants, randomly pick n locations from all the S locations, that is, redistribute n entrants randomly to all the S locations. For each industry do such simulation for 500 times and recalculate $\widehat{K}_d$ value. Rank all 500 $\widehat{K}_d$ values by its numeric value, set threshold value 0.95 and 0.05, then locate the values at the position of 95% and 5% of the rank as upper and lower envelopes, denoted as $\overline{K}_d$ and $\underline{K}_d$. Define localization and dispersion indicators α_d, β_d:

$$\alpha_d = \max(\widehat{K}_d - \overline{K}_d, 0)$$
$$\beta_d = \max(\underline{K}_d - \widehat{K}_d, 0) \qquad (3)$$

If $\alpha_d > 0$, then the industry is described as localized, when $\beta_d > 0$ indicates that the industry is dispersed. If both indicators are zero, that is to say, the industry has no prominent tendency to localize or disperse, rather it can be seen as close to random or natural distribution.

3.2.2. Global Confidence Bands

The α and β indicators can only reflect localization and dispersion on a local scale, with respect to the truth that those industries appear to be randomly distributed locally can be localized or dispersed in other certain distances. Hence, it's necessary to construct global confidence bands to describe the global tendency of industrial spatial distribution. The global confidence band is the joint estimation of local confidence band on different target distances. By interpolating extreme values on different target distances, global confidence bands are constructed. Denoting global confidence envelope as $\overline{\overline{K}}_d$ and $\underline{\underline{K}}_d$, define global localization and dispersion indicator Γ_d and Ψ_d :



(a) Integrated Circuit Manufacturing

(b) Computers Repairing

(c) Stainless Steel Products Manufacturing

(d) Ceramic Products Manufacturing

Figure 3. The spatial distribution patterns of different industries

$$\Gamma_d = \max\left(K_d - \overline{\overline{K}}_d, 0\right)$$

$$\Psi_d = \begin{cases} \max(\underline{\underline{K}}_d - K_d, 0), \; if \; \sum_{d=0}^{d_{max}} \Gamma_d = 0 \\ 0, \text{ otherwise} \end{cases} \qquad (4)$$

Likewise, Γ_d>0 suggests that the industry is globally localized while Ψ_d>0 suggests global dispersion.

3.3. Interpretation and Analysis

Comparison and analysis is needed to explore the spatial pattern revealed by the output that we calculated. First, we do curve fitting and draw line chart of series of K_d values with both local and global confidence envelopes to find certain industries' tendency of localization or dispersion on different target distances. Similarly we plot line chart of distances and numbers of localized industries to explore the pattern hiding behind it. Last but not least, time-series analysis method is adopted to study the change of spatial-temporal patterns on time series of industrial agglomeration.

4. RESULTS

4.1. Examples of Different Agglomeration Patterns

Sorting out calculating outputs, we spotted various agglomeration patterns, some of which are rather typical and shown in Figure 3 (a)-(d). Diagram (a) shows an industry with conspicuous tendency of small scale localization while diagram (b) shows an industry with localization tendency on a wider range. Noticing the tail of the curves in both (a) and (b) we can find industry (a) shows normalized distribution on long distances (over 30 km) in the city but (b) shows already dispersion on over 24 km. Diagram (c) indicates an industry dispersed in any distance, which means the industry distribution is strongly dispersed in the whole city. Diagram (d) shows an industry nearly normally distributed in the whole range of the main area of the city.

Such four space patterns are quite representative of all manufacturing industries we concerned in the study: industries have strong colocation tendency, industries show colocation in limited distances, industries normally distributed, and industries dispersed. What we want to highlight is, as the indicators' math characteristics, an industry which show strong colocation in short distances is bound to show strong dispersion on long distances, hence we limited the study range within the radius of major part of Shanghai to make sure the tail of the curves is of significance. For drawing the complete outline of the whole manufacturing industry, we counted up numbers of localized and dispersed industries in every distance.

4.2. General Situation of All Selected Industries

The numbers of manufacturing industry localized or dispersed in each distance are shown in Figure 3. The total number of the industries is 269 and the industrial categories is within the sub-categories of manufacturing industry, correspond to CategoryLV3 in Table 1.

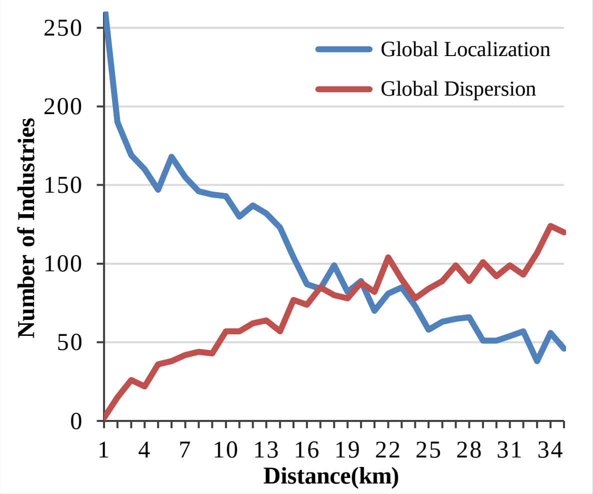


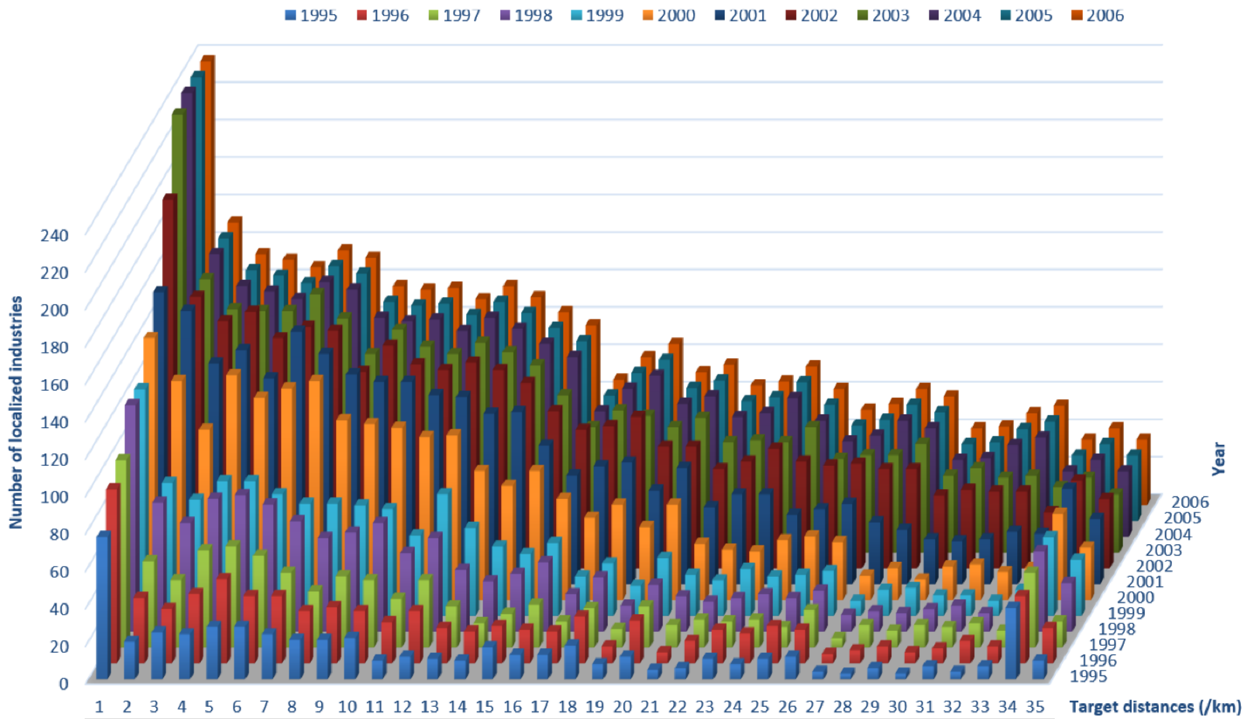
Figure 4. Number of industry showing tendency of global localization or dispersion in each target distance

In Figure 4, the obvious difference of localized industries and dispersed industries in short distances is noticeable. Such difference indicates that the manufacturing industrial agglomeration is significant in Shanghai within 15km. For distances longer than 15km, numbers of localized and dispersed industries show stable convergence. Till the distance over 32 or 33km the number of dispersed industries begin to come over that of localized industries. The result suggests that Shanghai does have an obvious industrial agglomeration phenomenon with respect to the fact that most industrial clusters emerge on a small scale within a quarter or one third of the city range, i.e. 8-12km.

Investigating the detailed scope of business of industries that exhibit different spatial patterns, we found that most short-range clusters appear in those industries that highly depend on labor power such as music instruments manufacturing, fermented food manufacturing and measuring vessels manufacturing, etc. However, when it comes to a longer range close to city radius, highly matured and large-scale industries still present localization. Such result is identical to economies of scale, and fits the view of spatial economy: point-axis distribution mostly develops towards network distribution (Ma et al., 2007).

4.3. Localization of All Selected Industries

In the case of Shanghai, we tend to focus more on localization as its significant meaning to industrial agglomeration. The number of localized industries in different years and distances, as shown in Figure 4, indicates that the localization tendency does exist and the distance range in which manufacturing industries tend to localize is growing stronger.



1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	
76	20	25	24	28	28	24	21	21	22	10	12	11	10	17	13	13	18	8	12	5	6	11	8	11	12	4	3	6	3	7	4	7	38	10	1995
93	35	29	37	45	36	36	28	30	28	22	28	19	17	20	18	17	25	9	23	6	12	18	16	20	18	5	7	9	6	8	12	9	36	19	1996
100	46	36	52	54	49	40	30	38	36	26	36	22	13	18	23	14	21	10	22	12	15	14	15	11	20	5	12	9	12	11	13	9	40	14	1997
121	69	58	71	73	68	59	50	53	58	42	50	33	27	31	37	20	29	14	25	19	16	18	20	18	22	9	11	10	12	14	10	10	43	26	1998
121	71	62	72	72	65	60	60	59	57	43	65	47	37	33	39	21	28	16	31	22	19	25	21	22	24	8	14	15	11	11	8	14	42	30	1999
140	117	91	120	108	113	117	96	94	92	87	88	69	61	69	54	44	51	39	51	30	27	26	32	34	31	13	17	11	18	19	15	17	46	28	2000
156	146	118	125	110	135	123	112	108	108	101	100	91	92	74	58	63	65	50	62	41	48	48	37	40	43	33	29	24	23	24	28	27	51	35	2001
197	145	132	137	123	129	127	105	119	109	106	110	106	99	84	74	76	81	65	65	53	57	64	57	55	56	53	53	39	42	41	41	30	47	37	2002
234	146	130	129	129	138	125	106	119	110	106	112	107	100	84	67	76	73	67	72	59	60	59	67	50	52	52	58	41	45	40	41	35	40	31	2003
237	151	134	131	127	136	132	117	115	116	110	117	111	103	96	67	79	86	71	75	64	66	74	62	51	54	62	58	41	42	49	53	35	41	35	2004
237	151	134	131	127	136	132	117	115	116	110	117	111	103	96	67	79	86	71	75	64	66	74	62	51	54	62	58	41	42	49	53	35	41	35	2005
237	151	134	131	127	136	132	117	115	116	110	117	111	103	96	67	79	86	71	75	64	66	74	62	51	54	62	58	41	42	49	53	35	41	35	2006

Figure 5. Number of industry showing tendency of localization in each distance and year

In the Figure 5 shown above, we can find that localization extent weakens as distance getting longer in each same year, but as year goes by, the extent of localization grows more and more obvious in any single target distance. This shows manufacturing industries do tend to form industrial clusters as economy develops within the range of Shanghai city. From another perspective, we notice the right diagonal line in the bar diagram (or left diagonal line in the table below) in the figure keeps a rather stable status, which shows that similar clusters form up on longer distances as economy develops, this is also a proof of that concentration continues forming and the tendency strengthening. Besides, some bloom pattern is also noticeable such as from year 1999 to year 2000 in distance from 2-10 km, 27-33 km. Such bloom shows industrial clusters were forming up at a high speed when we were welcoming the new century, which well-reflect the success of China's *Reform and Open* policy. The distinct division line between distance 26-27 km is very interesting that we can hardly explain its existence, but such division fades after year 2001 indicates that certain sub-industries of manufacturing have turned to focus on greater scale, which accounts for the localization grown-up in long distances close to city radius. For further information, we may need to date back to history and find more specific details.

5. CONCLUSION

This paper studies the industrial agglomeration in Shanghai by street-level accuracy industrial registration data. We adopt a widely used statistical method in geography, K-density function with kernel intensity estimation method from spatial analysis to measure the localization and dispersion. We also developed geoprocessing services and deployed it on our visualization portal to support customized calculating for analysis of industrial agglomeration. For most manufacturing industries in Shanghai we have these findings below:

a) Most industries have tendency of localization to form clusters within the range of the city. And most clustered industries show concentration between the distance from 1 km to 16km, which is roughly a quarter of the city range. Such distribution pattern confirms the economic growth tends to form a network layout from an initial point-axis layout.
b) The extent of localization or dispersion is skewed across various manufacturing industries that spatial distribution patterns various from industry to industry, but most industries present a short or medium distance localization.
c) Short distance localization appears mostly in immature small-scale industries that require craftsmanship or those do not have huge demands, while long distance localization takes place mostly in highly automated industries such as circuit and industrial equipment manufacturing.
d) The industrial agglomeration in Shanghai, especially manufacturing, continues to exist and develop that more industry clusters will keep forming up by the coming few

years and the extent of concentration will grow stronger in any distance.

Besides all these conclusions above, there are still many detailed issues remained to be investigated such as the co-localization between industries, measurement methods optimization considering firm scale and other factors, etc. For policy suggestions, as we can see from this study, the industry itself does have a tendency to localize, hence the essential point is to plan a reasonable layout for industries that may have different extent of localization or dispersion, to achieve efficient use of resources and city space, while considering other factors such as humanity and environment protection is also of vital importance in modern city growth.

The next step of this study may be to investigate further into concentrated industries, find their correlations and measure their influence on regional economic growth by big data integration. As a final point, note that such distance based and scale independent method widely used in geography can be also adopted beyond economic geography, as any data with micro-level geographical information needs to be measured on agglomeration or dispersion would suit such method fine. Such method provides a good perspective for geographers to test spatial agglomeration with non-biased and scale-free statistics.

6. ACKNOWLEDGEMENTS

This paper is supported by National Natural Science Foundation of China (No. 41501434 and No. 41371372). Thanks to those who offered generous help to this study, including Fa Li from LIESMARS, Wuhan Univerisity, along with Dehua Peng, Na Zhang and Shuang Liu, from School of Remote Sensing and Information Engineering, Wuhan University.

REFERENCES

Duranton, G. and Overman, H. G., 2005. Testing for localization using micro-geographic data. *Review of Economic Studies*, 72 (4). pp. 1077-1106.

He, Y., Liu, X., Li, R., 2012. On the Evolution of Spatial Agglomeration of China’s Manufacturing Industries and Its Driving Factors Based on Continuous Distance. *Journal of Finance and Economics*, 38(10). pp. 36-45.

He, C., Pan, F., Sun, L., 2007. Geographical Concentration of Manufacturing Industries in China. *ACTA Geographical Sinica*, 62(10). pp. 1254-1264.

Li, F., Gui, Z., Wu, H., Gong, J., Wang, Y., Tian, S., Zhang, J., 2017. Big enterprise registration data imputation: Supporting spatiotemporal analysis of industries in China. *Computers, Environment and Urban Systems. (Forthcoming)*

Ma, R., Gu, C., Pu Y., et al. Urban spatial sprawl pattern and metrics in South of Jiangsu Province along the Yangtze River, 2007. *ACTA Geographical Sinica*, 62(10). pp. 101.

Scholl, T., Brenner, T., 2011. Testing for Clustering of Industries – Evidence from micro geographic data. *Working Papers on Innovation and space*, 02(11). pp. 3-21.

Silverman, B. W., 1986. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York.

Yuan, F., Wei, Y., Chen, W., Jin, Z., 2010. Spatial Agglomeration and New Firm Formation in the Information and Communication Technology Industry in Suzhou. *ACTA Geographical Sinica*, 65(2). pp. 153-163.

https://doi.org/10.5194/isprs-archives-XLII-2-W7-149-2017 |