Testing Localization of Chinese Food Industries: Evidence from Microgeographic Data

Wenchao WU [a,1], Shaosheng JIN [b], and Suminori TOKUNAGA [c]

a. Graduate School of Life and Environmental Science, University of Tsukuba, Japan.
b. School of Management, Zhejiang University, China.
c. Faculty of Economics and Business Administration, Reitaku University, Japan.

1. Introduction

Ever since the publication of the *Principles of Economics* (Marshall, 1890), the agglomeration of economic activity has been fascinating generations of economists. As Krugman (1991) put it, "The most striking feature of the geography of economic activity… is surely concentration."

Generally speaking, the food industry is transportation intensive; therefore, the extent of localization is not high, in relative terms. Research in the UK (Duranton and Overman, 2005) and Japan (Nakajima, Saito and Uesugi, 2012) found very low percentages of food industries to be localized. Previous research also showed that the degree of industrial concentration in the food sector is low in China (Li, Shi and Jin, 2008). However, these conclusions are based on provincial aggregate data and discrete measurements.

Applying the continuous measurement of the kernel density function model, this study aims to assess the extent of agglomeration for the 50 four-digit Chinese food industries using microgeographic data. This introduction is followed by a. review of literature Section 3 presents the data and empirical strategy. Section 4 shows the main results and the last section concludes the study.

2. Literature review

To assess the geographic distribution of economic activity, economists have developed several indices. The representative of the first generation indices are Gini, Isard, and Herfindahl indices. These measurements are improved by the second generation, e.g., EG index (Ellison and Glaeser, 1997) and D index (Mori, Nishikimi and Smith, 2005).

However, these indices have drawbacks. Firstly, they rely on discrete, divided geographical units. In other words, the dots representing individual firms are transformed into units in boxes. Although computation is simpler by aggregating data, information is wasted. In addition, discrete space limits the analysis to only one spatial scale: usually county, region, or state. Consequently, results vary with spatial scales, rendering them incomparable. Motivated by satisfying the properties that an ideal agglomeration index should have (Combes, Mayer and Thisse, 2008), Duranton and Overman (2005) proposed a new index (DO index henceforth) which provides continuous space using microgeographic data. Some studies have already adopted this method (Duranton and Overman, 2008; Ellison, Glaeser and Kerr, 2010; Koh and Riedel, 2012; Nakajima, Saito and Uesugi, 2012; Barlet, Briant and Crusson, 2013).

Empirically, Akune and Tokunaga (2005) and Tokunaga and Akune (2005) measure agglomeration of the food industry in Japan using the EG index and analyze the dynamics of agglomeration from 1980 to 2000. Jin and Tokunaga (2009) evaluate agglomeration of the Chinese food industry using the D and EG indices. Li, Shi, and Jin (2008) measure agglomeration of the food industry in China using discrete measurements. The results indicate a low degree of concentration across the food sector in China. However, as pointed out earlier, measurements based on aggregate data and discrete space are proven to be biased.

3. Data and empirical strategy

3.1 Data

This research relies on data from the *Second National Economic Census of China*, conducted by the *National Bureau of Statistics of China* in 2008. The *National Economic Census Dataset* is, perhaps, the most comprehensive for Chinese manufacturing so far. It covers all existing firms regardless of size or ownership. The dataset contains 1,890,513 firms in the manufacturing sector. Each observation representing one firm provides information on the firm's location, post code, establishment year, number of employees, capital stock, revenue, ownership, business status, and industry classification code[2], including 2-digit, 3-digit, and 4-digit codes. Address and post codes help to match the geographic location of firms.

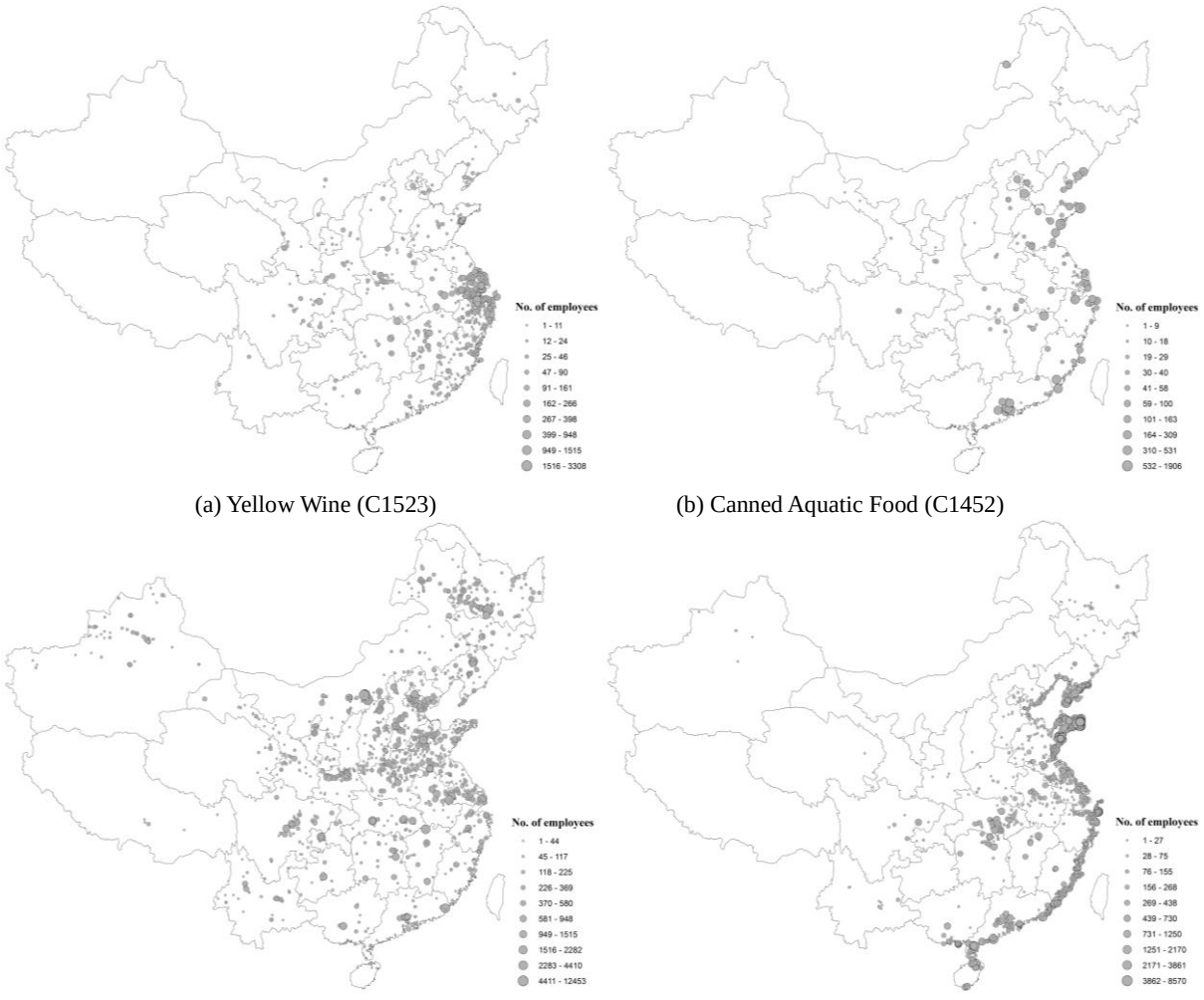
[1] Corresponding author. Email address: zju.wwc.2008@163.com

[2] Industry classifications follow the *Industrial Classification for National Activities* (ICNA) of China (GB/T 4754-2002).

In this study, we focus on three 2-digit industries, C13, C14, and C15, which represent food processing, food manufacturing, and beverage production, respectively. These 2-digit industries include 19 3-digit industries, which could be classified into 50 4-digit industries.

Since the calculation of DO's localization index requires exact geographic data, we use Google Maps API for geocoding, in which the address of each firm is converted to longitude and latitude. Approximately 0.65% of the observations are unable to be matched with a geographic location or have no employee number data. This left us with a population of 189,152 firms in the 3 two-digit food industries.

Figures 1(a)−(d) map this location information for our illustrative industries: Yellow Wine (C1523), Canned Aquatic Food (C1452), Liquid Milk and Milk Products (C1440), and Aquatic Products Freezing (C1361). Each dot on the map represents a firm, and the size of the dot reflects firm size. Here, we use the number of employees as a proxy for firm size. As we can observe from the map, the Yellow Wine (C1523) and Aquatic Products Freezing (C1361) industries seem to be localized, Liquid Milk and Milk Products (C1440) are dispersed, whereas Canned Aquatic Food (C1452) appears fairly random. We continue using these industries for illustrative purposes in the methodology section.



(a) Yellow Wine (C1523) (b) Canned Aquatic Food (C1452)

(c) Liquid Milk and Milk Products (C1440) (d) Aquatic Products Freezing (C1361)

Figure 1. Location of firms for illustrative industries

3.2 Empirical approach

This section briefly introduces the empirical approach. The general idea of DO's method is comparing the actual distribution of bilateral distances between any two firms with a randomly drawn set of bilateral distances.

Step 1: Estimating the kernel density

For industry *A* with *n* firms, the Euclidean distances between every pair of firms are calculated. This generates $n(n-1)/2$ unique bilateral distances. Also, the weight of each firm is considered. With e_i denoting the number of employee of firm *i*, the estimator of the density of bilateral distances at distance *d* is

$$K(d) = \frac{1}{h\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} e(i)e(j)} \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} e(i)e(j)\, f\left(\frac{d-d_{i,j}}{h}\right) \tag{1}$$

where $d_{i,j}$ is the Euclidean distance between firms *i* and *j*, *h* is the bandwidth, and *f* is the kernel function[3].

Step 2: Choosing counterfactuals

To report the significance of the result, we choose the relevant counterfactuals as the benchmark with which the kernel density should be compared. Since the exact distribution of the distance among the population is unknown, we have to rely on the Monte Carlo approach to construct the counterfactuals.

The analysis is only informative when controlling for the overall tendency of manufacturing to agglomerate. That is to say, the distribution of firms is inhomogeneous across space. Therefore, it is natural to consider the set of all existing "sites" *S* currently occupied by all the firms.

Following Duranton and Overman (2005) and Nakajima, Saito, and Uesugi (2012), we run 1,000 simulations for each investigation. In each simulation, we sample as many sites as there are firms in the group of interest. The sampling is done without replacement for each time. For an industry *A* with *n* firms, we generate the counterfactuals A_m for *m* = 1, 2, …, 1000 by sampling *n* elements without replacement from *S*, so that each trial could be regarded as a random relabeling of firm sites.

Step 3: Global confidence intervals

Next, we determine the confidence intervals. The distance under consideration is from zero to three hundred kilometers[4]. We neglect distances greater than 300 km, since any significantly high density of distance in this range could be interpreted as dispersion and thus be redundant for our purpose.

On the basis of the concept of a local confidence interval, we construct the global confidence interval. Since there are correlations between distances, the idea is to return to the simulated industries and look for the upper and lower local confidence intervals such that 5% of the randomly drawn densities lie above the upper band and 5% lie below the lower band[5].

Following this procedure, we define the upper global confidence band $\overline{\overline{K}}_A(d)$ and lower global confidence band $\underline{\underline{K}}_A(d)$. If, for an industry *A*, $K_A(d) > \overline{\overline{K}}_A(d)$ for at least one $d \in [0,300]$, then this industry is defined as globally localized at the 5% confidence level. On the contrary, when $K_A(d) < \underline{\underline{K}}_A(d)$ for at least one $d \in [0,300]$ and $K_A(d) \leq \overline{\overline{K}}_A(d)$ for every $d \in [0,300]$, industry A is said to be dispersed (at 5% confidence level). Similar to the local localization indices, we can define an index of global localization:

$$\Gamma_A(d) \equiv \max\left(K_A(d) - \overline{\overline{K}}_A(d), 0\right) \tag{2}$$

and an index of dispersion

$$\Psi_A(d) \equiv \begin{cases} \max\left(\underline{\underline{K}}_A(d) - K_A(d), 0\right) & \text{if} \quad \sum_{d=0}^{d=300} \Gamma_A(d) = 0 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

All computations, including the calculation of distance, estimation of kernel density, running simulations, and constructing the confidence bands, is manipulated using R. The whole procedure relies on the package "*dbmss*," developed and maintained by Marcon, Gabriel, Stephane, and Florence (2014).

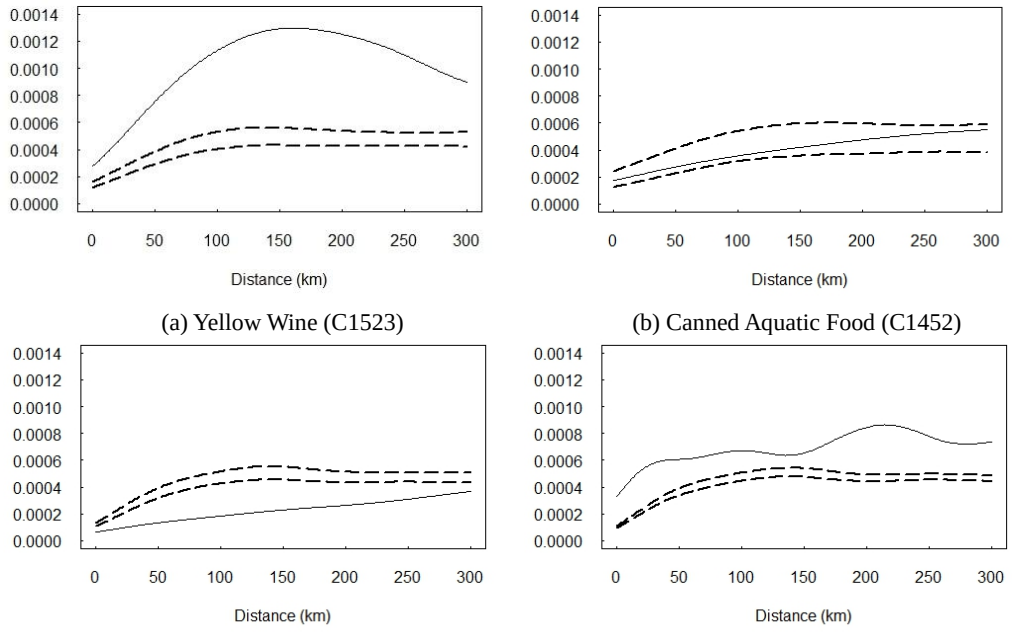
Figure 2 demonstrates the illustrative examples. The solid lines in the figures represent *k*-density estimation of observed industry, i.e., the real industry distribution. The two dashed lines are the upper and

[3] Following Silverman (1986), the Gaussian kernel with optimal bandwidth is adopted.

[4] Previous research in the UK and Japan investigate distances from 0−180 km. Considering the fact that China has a territory larger than the UK or Japan, we expand the range to 300 km.

[5] For more details on constructing the global confidence bands, see Duranton and Overman (2005, 2008) and the *KdEnvelope* functions in R package "dbmss" (Marcon, Gabriel, Stephane and Florence, 2013, 2014).

lower global confidence bands constructed from 1,000 simulations.



(a) Yellow Wine (C1523)

(b) Canned Aquatic Food (C1452)

(c) Liquid Milk and Milk Products (C1440)

(d) Aquatic Products Freezing (C1361)

Figure 2. Kernel density and global confidence bands for illustrative industries.

In Figures 2(a) and (d), we can see that the observed *k*-density lies over the upper global confidence band, which implies that the Sugar Refinery industry (C1340) exhibits a localization pattern. Intuitively, from Figure 1(a), we can see that most firms cluster in the Yangtze River delta. In Figure 2(b), the observed density is enveloped by the confidence bands. This suggests that the corresponding industry does not deviate significantly from randomness. In Figure 2(c), the solid line lies below the lower global confidence band. Therefore, the Liquid Milk and Milk Products (C1440) industry is defined as dispersed.

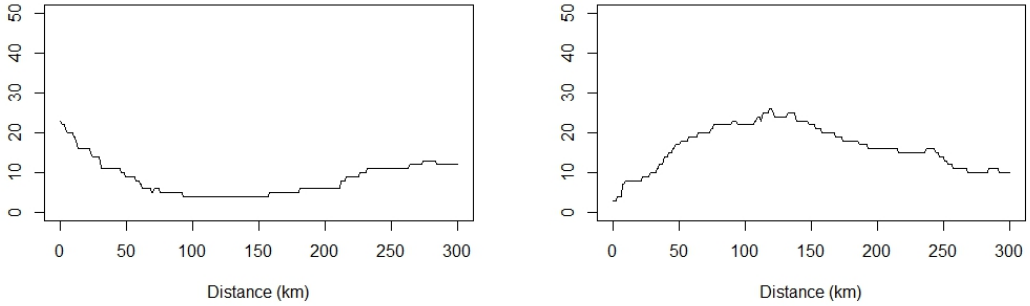
4. Results

We repeat the procedure above for each industry. Results show that 21 of the four-digit food industries are localized whereas 15 are dispersed. The remaining 14 do not deviate significantly from randomness. Compared with the results of the baseline analysis, the number of localized industries increases while the number of dispersed ones decreases. More industries are shown to be randomly distributed. Table 1 shows the number of localized and dispersed industries under each two-digit industry. Interestingly, localization of the food industry is significantly higher than that in some other countries. Duranton and Overman (2005) found that out of 30 four-digit food industries in UK, only one industry is localized. Similarly, among 40 four-digit food industries in Japan, only three exhibit localization (Nakajima, Saito and Uesugi, 2012).

Table 1. Number of localized and dispersed industries in each two-digit food industry

Two-digit industry		No. of four-digit industries	No. of localized industries	No. of dispersed industries
C13	Food Processing	17	12	3
C14	Food Manufacturing	20	5	6
C15	Beverage Production	13	4	6
Total		50	21	15

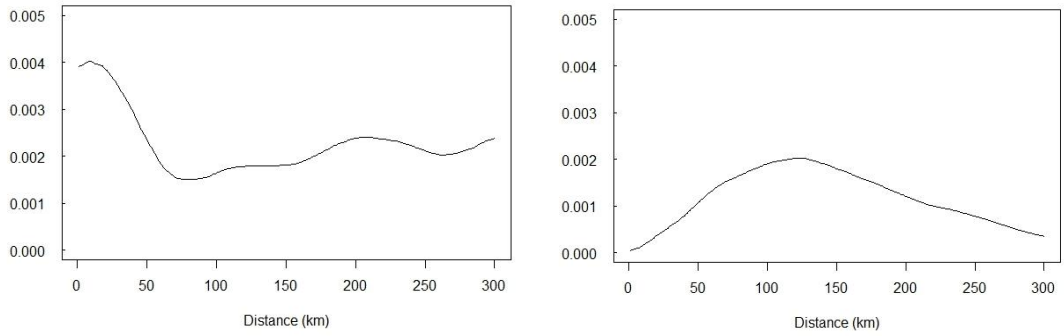
Figure 3 demonstrates the number of globally localized and dispersed industries at each distance. Although it seems more industries are globally dispersed at most distances, it does not contradict previous findings that more industries are localized (21 versus 15). The rationale is that one industry is defined as localized if it is localized for at least one unit distance.



(a) Global localization (b) Global dispersion

Figure 3. Number of global localized and dispersed industries

We define the measure of localization at each distance d as $\Gamma(d) \equiv \sum_A \Gamma_A(d)$. Similarly, the measure of the extent of cross-industry dispersion by distance is defined as $\Psi(d) \equiv \sum_A \Psi_A(d)$. Figure 4 demonstrates the index of global localization and dispersion by distance when accounting for the weight. Figure 4(a) shows that the extent of localization is much greater at small distances. The distances with significantly higher agglomeration are less than 70 km. Therefore, we can infer that the localization of food industries in China takes place within small areas.



(a) Global localization (b) Global dispersion

Figure 4. Index of global localization and dispersion by distance

To compare the degree of localization and dispersion among 4-digit industries, we employ the cross-distance measures of localization and dispersion. Following Duranton and Overman (2005), the cross-distance index of localization for industry A is defined as $\Gamma_A = \sum_{d=0}^{300} \Gamma_A(d)$ and cross-distance index of dispersion as $\Psi_A = \sum_{d=0}^{300} \Psi_A(d)$.

Table 2. Most localized and most dispersed four-digit food industries

Four-digit industry		Γ or Ψ
Most localized		
1523	Yellow Wine	0.220
1361	Aquatic Products Freezing	0.126
1340	Sugar Refinery	0.089
1494	Additive of Food and Fodder	0.070
1310	Grain Mill Products	0.027
Most dispersed		
1532	Drinking Water	0.037
1320	Feed Processing	0.031
1522	Beer	0.029
1440	Liquid Milk and Milk Products	0.024
1399	Processing of Other Food Not Listed	0.023

Table 2 lists the five most localized and five most dispersed industries. Yellow Wine industry (C1523) is the most localized industry. As we can see from Figure 1(a), most of the firms cluster in the Yangtze River delta. Aquatic Products Freezing (C1361) is the second most localized industry. This is consistent with the impression from Figure 1(d), where most of its production is located along the coastline of China.

5. Conclusion

On the basis of DO's model, this study tests the localization of Chinese food industries. Contrary to previous research based on aggregate data and discrete measurement, we found pronounced localization of Chinese food industries. The results show that, among the 50 four-digit food industries, 21 of them are localized when firm size is considered. Fifteen industries are dispersed and the remaining 14 are randomly located. Yellow Wine (C1523) and Aquatic Products Freezing (C1361) are the two most localized industries.

Reference

Akune, Y. and S. Tokunaga. 2005. "An Empirical Analysis of the Industrial Agglomeration of Food Industry from 1980 to 2000." *Studies in Regional Science*, 35(2), 625-35.

Barlet, M.; A. Briant and L. Crusson. 2013. "Location Patterns of Service Industries in France: A Distance-Based Approach." *Regional Science and Urban Economics*, 43(2), 338-51.

Combes, P.-P.; T. Mayer and J.-F. Thisse. 2008. *Economic Geography: The Integration of Regions and Nations*. Princeton University Press.

Duranton, G. and H. G. Overman. 2005. "Testing for Localization Using Micro-Geographic Data." *The Review of Economic Studies*, 72(4), 1077-106.

____. 2008. "Exploring the Detailed Location Patterns of UK Manufacturing Industries Using Microgeographic Data." *Journal of Regional Science*, 48(1), 213-43.

Ellison, G. and E. L. Glaeser. 1997. "Geographic Concentration in US Manufacturing Industries: A Dartboard Approach." *The Journal of Political Economy*, 105(5), 889-927.

Ellison, G.; E. L. Glaeser and W. R. Kerr. 2010. "What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns." *The American Economic Review*, 1195-213.

Jin, S. and S. Tokunaga. 2009. "Effects of Agglomeration on Production in the Chinese Food Industry: A Panel Data Analysis." *Studies in Regional Science*, 38(4), 1021-26.

Koh, H.-J. and N. Riedel. 2012. "Assessing the Localization Pattern of German Manufacturing and Service Industries: A Distance-Based Approach." *Regional Studies*, 1-21.

Krugman, P. R. 1991. *Geography and Trade*. Cambridge, MA: MIT press.

Li, N.; M. Shi and F. Jin. 2008. "An Empirical Study on the Spatial Agglomeration of China's Foods Industries." *Review of Management*, 20(1), 32-39.

Marcon, E.; L. Gabriel; T. Stephane and P. Florence. 2014. "Dbmss: Distance-Based Measures of Spatial Structures. R Package Version 2.0.5.,"

____. 2014. "Dbmss: Distance-Based Measures of Spatial Structures. R Package Version 2.1.0.,"

Marshall, A. 1890. *Principles of Economics*. London: Macmillan.

Mori, T.; K. Nishikimi and T. E. Smith. 2005. "A Divergence Statistic for Industrial Localization." *Review of Economics and Statistics*, 87(4), 635-51.

Nakajima, K.; Y. U. Saito and I. Uesugi. 2012. "Measuring Economic Localization: Evidence from Japanese Firm-Level Data." *Journal of the Japanese and International Economies*, 26(2), 201-20.

Silverman, B. W. 1986. *Density Estimation for Statistics and Data Analysis*. CRC press.

Tokunaga, S. and Y. Akune. 2005. "A Measure of the Agglomeration in Japanese Manufacturing Industries: Using an Index of Agglomeration by Ellison and Glaeser." *Studies in Regional Science*, 35(1), 155-75.