RIETI Discussion Paper Series 13-E-014

# A Spatial Approach to Identifying Agglomeration Determinants

**MORI Tomoya**
RIETI

**Tony E. SMITH**
University of Pennsylvania

RIETI
Research Institute of Economy, Trade & Industry, IAA
The Research Institute of Economy, Trade and Industry
http://www.rieti.go.jp/en/

RIETI Discussion Paper Series 13-E-014

**March 2013**

# A Spatial Approach to Identifying Agglomeration Determinants

MORI Tomoya[1],[2] and Tony E. SMITH[3]

## Abstract

Typical analyses of industrial agglomerations start with some aggregate measure of the “agglomeration degree” for each industry, and attempt to explain differences in these values across industries by regressing them on sets of industrial attributes. But this aggregation makes it difficult to capture the *spatial* aspects of individual agglomerations. In the present paper, we develop a more explicit spatial approach to identifying agglomeration determinants by means of a two-stage analysis. First, we detect individual spatial clusters of each industry on a map. We then attempt to explain differences in these cluster patterns between industries by employing an appropriate regression framework. Here, cluster employment sizes are regressed on selected regional attributes for each industry-cluster pair, and significant differences between industries are captured in terms of industry-level interactions with these attributes. This modeling approach is then applied to the three-digit manufacturing industries in Japan.

*Keywords :* Industrial agglomeration, Cluster analysis
*JEL Classifications* : R12, C21, L60

**Acknowledgement:** The authors benefited from discussions with Gilles Duranton, Laurent Gobillon, Wen-Tai Hsu, Yoko Konishi, Kentaro Nakajima, Yasuhiro Sato, and Ayumu Tanaka. This study is conducted as part of the Project, “The formation of economic agglomerations and the emergence of order in their spatial patterns: Theory, evidence, and policy implications,” undertaken at the Research Institute of Economy, Trade and Industry (RIETI), and has been partially supported by The Kajima Foundation, The Grant in Aid for Research (Nos. 21243021, 21330055, 22330076, and the Global COE Program “Raising Market Quality”) of Ministry of Education, Culture, Sports, Science and Technology of Japan.

[1] Institute of Economic Research, Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501 Japan. Email: mori@kier.kyoto-u.ac.jp.

[2] Research Institute of Economy, Trade and Industry (RIETI), 11th floor, Annex, Ministry of Economy, Trade and Industry (METI) 1-3-1, Kasumigaseki Chiyoda-ku, Tokyo, 100-8901 Japan.

[3] Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA. Email: tesmith@seas.upenn.edu.

# 1 Introduction

Typical analyses of industrial agglomeration start with some scalar measure of the “degree of agglomeration”for each industry, and attempt to explain differences in these values across industries by regressing them on appropriate sets of industrial attributes (refer to Rosenthal and Strange [22] for a survey). But it is difficult to capture the *spatial* aspects of industrial agglomeration in terms of any such aggregate measure. In the present paper, we develop a more explicit spatial approach to identifying industrial agglomeration by means of a two-stage analysis. In the initial *cluster-detection stage*, we attempt to identify individual spatial clusters (or agglomerations) of each industry on a map. In the subsequent *cluster-analysis stage*, we attempt to explain differences in these cluster patterns between industries by identifying those local regional factors that may induce each industry to agglomerate.

Here, we again use regression methods. But the key difference is that the relevant sample units are now *clusters* for each industry, where the dependent variable is the size of each cluster (measured by industrial employment), and the candidate explanatory variables consist of local regional attributes defined with respect to each cluster. By treating individual industries as fixed effects, we can then include interaction effects to identify differences between the sets of regional attributes most relevant for each industry.

Within this regression framework, there are a number of key issues that must be addressed. First there is a question of defining meaningful regional attributes at the cluster level. In the cluster-detection stage (summarized in Section 2 below), individual clusters are characterized in terms of “spatially coherent”sets of contiguous municipalities. Hence local regional attributes can in principle be defined in terms of appropriate averages of municipality attributes within each cluster. In our application to the Japanese manufacturing industries in Section 3, for instance, we have municipality data on populations, incomes, education levels and the natural conditions (coastal access and climate conditions). So by using this data together with inter-industry transactions data for the whole of Japan, we are able to construct a candidate set of local regional attributes. However, it should be emphasized that given the limited nature of available data, the present analysis is necessarily partial in nature. Hence our main objective is to illustrate how this methodology can be used to extract richer information about the spatial determinants of agglomeration within each industry.

A second key issue relates to the use of clusters as the appropriate level of spatial aggregation.[1] Why not take the smallest regional units in which the relevant data are available (municipalities in our case) as the basic spatial units, and use this data directly? Alternatively, why not aggregate data to a larger prefecture (or state) level where a much

[1] It should be noted that some researchers (most notably Combes and Overman [3] and Duranton and Overman [5]) questioned spatial aggregation at any degree, and proposed to use the point pattern of firm locations directly. We return to this issue in Section 2.3 below.

richer variety of data is readily available? Because this issue is of central importance for the present analysis, we undertake a systematic comparison of regressions at these three levels of aggregation in Section 4 below. In brief, our results at the *municipality level* indicate (not surprisingly) that distinctions between adjacent municipalities are so slight that all potentially relevant relations are dominated by spatial autocorrelation effects within industries. This autocorrelation results from the fact that a municipality is typically much smaller than an individual agglomeration. While it might be argued that spatial regression approaches should be of some help here, it seems clear that with respect to industrial agglomeration, these municipalities are simply too small for meaningful relations to be captured.[2]

In contrast, results at the *cluster level* exhibit relatively little spatial autocorrelation within industries, reflecting the fact that the spatial extent of an individual agglomeration is well approximated by that of the corresponding cluster. More importantly, the regression results at this level of aggregation appear to be very meaningful, both in terms of economic and structural differences between the industries studied.

Spatial autocorrelation effects are even weaker at the *prefecture level*. But as discussed in Section 4 below, this seems to have more to do with the statistical properties of large aggregates than with any interesting economic relations at the prefecture level. In particular, while regressions at the prefecture level do identify sets of significant explanatory variables which are comparable in number to those at the cluster level, the compositions of these sets are drastically different. Hence, the results obtained at the cluster level cannot be approximated by those obtained at the prefecture level. In short, this limited comparison suggests that clusters identified in our first stage of analysis provide natural spatial units for comparative analyses of industrial agglomeration.

Finally, it should be stressed that in addition to identifying spatial determinants of cluster size across industries, an important goal of this research is to determine whether such clusters are significantly related to *industrial productivity*. The relation between industrial productivity and agglomeration has long been studied by economists (see Rosenthal and Strange [22] and Melo, Graham and Noland [15] for surveys. For a more recent study of industries in France, see Combes et al. [4]). The standard approach in this literature is to start with a given regional subdivision (such as counties in the US, or employment areas and urban areas in France) and to compare agglomeration and productivity levels among such regions. In particular, agglomeration levels are typically identified in terms of general employment density (for example, above-median versus below-median density levels), and are taken to be common to all industries. In contrast,

[2] With respect to spatial regression approaches, it is also worth noting that the between-industry effects of primary interest here are (at a minimum) extremely awkward to model in terms of standard spatial error or spatial lag formulations.

our present strategy is to identify specific areas of agglomeration for each industry, as in the cluster-detection stage above. In this context, the most direct approach to productivity comparisons is simply to test whether productivity in a given industry is higher for clustered establishments than for non-clusered ones. In Section 5 below, we compute total factor productivity of individual establishments, and carry out Mann-Whitney tests of mean-productivity differences. It is shown that establishments within industrial clusters are generally more productive than those not in clusters.

All methodologies developed here are illustrated by applications to the three-digit manufacturing industries for Japan in 2001 (of which 163 industrial types are present in the regional system chosen for analysis).[3] The organization of rest of the paper is as follows. First, to develop our basic two-stage procedure, we begin in Section 2 below with the cluster-detection stage. Since this procedure has been detailed in Mori and Smith [19], it suffices to give a brief summary of the main ideas in terms of our Japanese data. The main body of the paper focuses on the cluster-analysis stage in Section 3, which constitutes the primal contribution of this paper. Here we begin by identifying a number of regional attribute variables as possible determinants of agglomeration, which, combined with identified clusters, are in turn utilized in the regression. We then analyze the desirability of clusters as regional units, relative to municipalities and prefectures, from the view point of spatial autocorrelations among regression residuals in Section 4. Finally, in Section 5, we analyze relations between clustering and productivity of firms. The paper concludes with brief discussions of related research in Section 6.

# 2 Identification of Industrial Clusters

We begin with a brief description of our cluster-detection methodology in Section 2.1 below. In Section 2.2, we then compare the results of this method (for the case of Japan) with the scalar "degree of agglomeration" approach alluded to in the Introduction. Finally, in Section 2.3, we also give a brief comparison of this method with the "distance density" approach of Duranton and Overman [5].

## 2.1 Cluster-Detection Methodology

To motivate our approach to cluster detection, we begin by observing that recent theoretical results on equilibrium location patterns in continuous space (e.g., Tabuchi and Thisse [23], Ikeda et al. [10], Hsu [9]) suggest that there is remarkable commonality among possible

[3] Industrial types are based on the Japanese Standard Industry Classification (JSIC) in 2001. The establishment counts across these industries are taken from the *Establishment and Enterprise Census of Japan* in 2001. The mean and median establishment counts per industry are respectively 3958 and 1825. In addition, 147 (90%) of these industries have more than 100 establishments, and 125 (77%) have more than 500 establishments.

equilibrium patterns of agglomeration within each industry. In particular, the number, size and spacing of agglomerations are shown to be well preserved under a variety of stable equilibria. From this perspective, our objective is to identify these common features. To do so, we treat such equilibria as *stationary states*, and develop a probabilistic model of location behaviour within such stationary states. In particular, while individual location decisions may be based on the prevailing steady-state distribution, they can nonetheless be treated as statisitically *independent* events, i.e, as random samples from this distribution. This simplification of course precludes any questions about the process of cluster formation, or even the economic rationale for clustering. Rather, our goal here is to provide a simple statistical framework within which the most salient features of these equilibrium cluster patterns can be identified.

To develop this framework, we assume that the relevant location space for establishments is a set, $R$, of *basic regions* (municipalities) which are sufficiently small to ensure that all meaningful clusters for industries can be modeled as contiguous sets of basic regions. Figure 1 shows the map of municipalities in Japan which are used as basic regions in the present study.[4] In order that a given set, $C$, of municipalities qualify as a "cluster", it is required to be "approximately convex" in an appropriate sense.[5]

[Figure 1]

With this setting, each spatial pattern of clusters for an industry can be characterized formally as a *cluster scheme*, $\mathbf{C} = (R_0, C_1, .., C_{k_{\mathbf{C}},})$, that partitions $R$ into one or more disjoint clusters, $C_1, .., C_{k_{\mathbf{C}},}$ together with the residual set, $R_0$, of all non-cluster regions.

If we now let $I$ denote the given set of relevant industries, then the problem of cluster detection for each industry, $i \in I$, amounts to determining the cluster scheme, $\mathbf{C}_i^*$, that "best fits"the observed (stationary) distribution of industry establishments in $R$. To do so, we start with the basic idea that meaningful clusters should in some sense correspond to local "peaks" in this regional distribution of establishments. To formalize this notion, we first associate with each possible cluster scheme, $\mathbf{C}$, a family of possible location probability models, $p_{\mathbf{C}} = [p_{\mathbf{C}}(j) : j = 1, .., k_{\mathbf{C}}]$, called *cluster probability models*, where $p_{\mathbf{C}}(j)$ denotes the probability that a randomly sampled establishment will be located

[4] We use the municipalities in Japan as of October 1, 2001. Out of a total of 3363 municipalities, we only consider 3207 (as shown in Figure 1) which are *geographically connected* to the major islands of Japan (Honshu, Hokkaido, Kyushu and Shikoku) via a road network. This avoids the need for ad-hoc assumptions regarding the effective distance between non-connected regions. The only exception here is Hokkaido, which is one of the four major islands, but is disconnected from the road network covering the other three.

[5] More precisely, we first approximate the original planar regional space by a discrete network of municipalities in which adjacent pairs of municipalities are connected by shortest road-network paths between them (assuming municipal offices as their representative locations). An approximate convex set in this context is then represented by a *convex solid set* with respect to the metric on $R$ induced by this municipality network. See Mori and Smith [19] for details.

in cluster $C_j$. If $\mathbf{C}$ does correspond to the peaks in this regional distribution, then one would expect the best fitting cluster probability model for $\mathbf{C}$ to have relatively large cluster probabilities, $p_{\mathbf{C}}(j)$, compared to the "non-cluster" probability, $p_C(R_0) = 1 - \Sigma_j p_{\mathbf{C}}(j)$. In summary, our basic approach to cluster identification is to hypothesize that the prevailing steady-state distribution of establishments in each industry $i$ is well approximated by some cluster probability model as defined above. This, together with our independence assumption above, implies that the observed pattern of $i$-establishment frequencies must be multinomially distributed with respect to these cluster probabities.

Given this family of candidate multinomial models, the task is then to find the cluster scheme, $\mathbf{C}_i^*$, with a corresponding cluster probability model that best fits the observed distribution of industrial establishments in the sense of maximum likelihood. However, it turns out that such a procedure always favors cluster schemes with larger numbers of clusters. So to achieve more robustness, some "penalty"is needed to avoid excessively large numbers of clusters. While there are a number of methods for doing so, our investigations have led us to conclude that the *Bayesian Information Criterion* ($BIC$) is the most promising method for purposes of industrial cluster identification. Hence our operational procedure amounts to an algorithm (detailed in Mori and Smith [19]) for finding a cluster scheme, $\mathbf{C}_i^*$, for each industry $i$ that maximizes this criterion.[6]

However, it is important to emphasize that even if the distribution of establishments for industry $i$ were random, this procedure would necessarily produce some "best fitting" cluster scheme, $\mathbf{C}_i^*$. So to determine whether or not there is any *meaningful* clustering for industry $i$, it is crucial to test these results against the null hypothesis, $H_0$, of completely random locations. Here, random locations are characterized by the uniform probability distribution over economic area $P_0 \equiv [P_0(r) : r \in R]$ with $P_0(r) = a_r / \sum_{j \in R} a_j$ where $a_r$ is the size of economic area in region $r$.[7] Hence the final step in this procedure is to simulate the distribution of $BIC$ for a substantial number of random location patterns (here 1000 are used), and determine whether or not $H_0$ can be rejected (at the 0.05 level of significance). If so, then $\mathbf{C}_i^*$ can be taken to represent a "significant" clustering pattern for industry $i$. Otherwise, $\mathbf{C}_i^*$ is taken to represent "spurious clustering" (i.e., not distiguishable from random clustering).

[6] As with most algorithms, this procedure is only guaranteed to find a local maximum of $BIC$. Hence the actual procedure used is to start with the "minimal" cluster scheme containing a single cluster consisting of a single basic region with the highest $BIC$ value, and proceed by incremental steps to find a local maximum, $\mathbf{C}_i^*$, with respect to this starting point. However, our investigations so far indicate that this always produces reasonable results (Mori and Smith [19, S 5.4.1]).

[7] To represent the areal extent of each municipality we adopt the notion of "economic area", obtained by subtracting forests, lakes, marshes and undeveloped area from the total area of the region. The data are available from the *Toukei de Miru Shi-Ku-Cho-Son no Sugata* in 2002 and 2003 (in Japanese) by Statistical Information Institute for Consulting and Analysis of Japan. The economic area of Japan as a whole (120,205km$^2$) amounts to only 31.8% of total area in Japan. Among individual municipalities this percentage ranges from 2.1% to 100%, with a mean of 48.5%.

For purposes of the second stage of our approach, we are not interested in explaining such spurious cluster patterns. These industries are thus excluded from the regression analysis below. In the present case, only 9 of the 163 three-digit manufacturing industries in Japan were spurious. So the final set of manufacturing industries, $I$, for the present analysis consists of the 154 industries exhibiting significant clustering. Moreover, as discussed in detail in Mori and Smith [19], these 9 industries all involve special locational conditions or constraints that would qualify them as outliers in any reasonable type of cluster analysis.

## 2.2 Comparison with the Scalar-Measure Approach

The single most dominant approach to agglomeration comparisons between industries has been in terms of scalar measures of the overall *degree* of industrial agglomeration (see, e.g., Rosenthal and Strange [22] for a survey). These indices are computed by measuring the discrepancy between the spatial distribution of establishments within an industry and a given reference distribution representing "complete dispersion" of establishments.[8] But, not surprisingly, such scalar measures often yield similar values for industries with very different spatial patterns of agglomeration (or dispersion).

This can be illustrated in terms of the *D index* developed in Mori et al. [16], which for a given industry $i$ is defined as the Kullback-Leibler [12] divergence of its establishment location probability distribution, $P_i \equiv [P_i(r) : r \in R]$, from purely random establishment location patterns, $P_0 \equiv [P_0(r) : r \in R]$, as defined in Section 2.1 above. By using the sample estimate of $P_i$, namely, $\widehat{P}_i = [\widehat{P}_i(r) : r \in R]$ with $\widehat{P}_i(r) \equiv n_r/n$, a corresponding estimate of this $D$ index is given by

$$D(\widehat{P}_i|P_0) = \sum_{r\in R} \widehat{P}_i(r) \ln\left(\frac{\widehat{P}_i(r)}{P_0(r)}\right). \tag{1}$$

The intuition behind this particular index is that it provides a natural measure of distance between probability distributions. So by taking uniformity to represent the complete absence of clustering, it is reasonable to assume that those distributions "more distant" from the uniform distribution should involve more agglomeration. Note also that since both $D$ and $BIC$ are based on similar log-likelihood measures of "distance from uniformity", our cluster identification procedure is closer in spirit to this scalar measure than other possible choices. Hence $D$ provides a natural candidate for comparing the advantages of this approach over scalar measures in general.

Given these general observations, the histogram of $D$ values for the 154 three-digit

[8] In Ellison and Glaeser [7] and Duranton and Overman [5], this reference distribution is defined in terms of the aggregate distribution of all industries, while in Mori et al. [16], it is taken to be the distribution of economic area.

industries in Japan is shown in Figure 2 below, and is seen to range from $D = 0.47$ to 5.98.

[Figure 2]

To illustrate different patterns with similar values of $D$, it is first necessary to ask how similar these values must be in order to qualify as "significantly similar"in a statistical sense. One straightforward approach is to consider the sampling distribution of $D$ values for all industries, and simply ask how close two independent sample values, $D_1$ and $D_2$, must be in order that there be less than a 5% chance of drawing a pair this similar.[9] This question can in fact be answered exactly for finite sampling distributions of size $n$ by simply computing $\Delta = |D_1 - D_2|$ for all possible $n^2$ realizations of $(D_1, D_2)$ and identifying the threshold value, $\Delta_{.05}$, that bounds the smallest 5% of all $\Delta$ values. In the present case of Figure 2, with $n = 154$, this value is given by $\Delta_{.05} = 0.0786$. By employing this similarity criterion, we now consider a range of example comparisons, starting from very low values of $D$ and proceeding to higher values. These selected comparisons will also serve to illustrate the regression results obtained in Section 3.3 below.

Starting at the low end of the $D$ scale, we first note that there is little qualitative difference between the cluster patterns of industries exhibiting low divergence values. This is well illustrated by "bakeries and confectionery products" (JSIC127) and "sliding doors and screens" (JSIC173). These industries have respective $D$ values, 1.031 and 0.760, which are at the low end in Figure 2, but which are nonetheless considerably more different than the threshold level, $\Delta_{.05}$, above. Their corresponding cluster patterns are shown in Panels (a) ad (b) in Figure 3, respectively, where each enclosed gray area represents an individual cluster, and darker color indicates larger concentration of employment. As is clear from the figure, these industries are typically *ubiquitous*, so that their clusters are found densely over the country.

[Figure 3]

But as cluster patterns become more diversified for larger values of $D$, larger qualitative differences begin to appear. Panels (a) through (d) in Figure 4 show the cases of "textile outer garments and shirts" (JSIC151), "seafood products" (JSIC122), "communication equipments" (JSIC304) and "paper containers" (JSIC185), respectively. Their $D$ values are 1.611, 1.646, 1.756 and 1.822, respectively, where the first two, and the last two values are statistically indistinguishable from one another (refer to Figure 2). But, unlike the ubiquitous industries discussed above, these cluster patterns are concentrated in certain

[9] This procedure is of course only meaningful with respect to a *given* sampling distribution, and in particular, implicitly assumes that this distribution is sufficiently dense to ensure the presence of some "very similar"distinct pairs. An examination of Figure 2 suggests that this assumption is not unreasonable in the present case.

parts of the country that are specific to each industry. In Section 3.3, we will see that such variations in cluster localizations can be well distinguished in terms of local regional attributes of clusters.

[Figure 4]

For industries with even higher values of $D$, the ambiguity of scalar indices becomes severer. Panels (a) and (b) in Figure 6 show the cluster patterns of "musical instruments" (JSIC342) and of "steel with rolling facilities" (JSIC263), respectively. Their $D$ values are 2.953 and 3.000, respectively, and again are statistically indistinguishable. But on the one hand, the former is seen to be *globally more concentrated* in the central region of the country [encompassing Hamamatsu, Nagano and Tokyo, as in Panel (a)], and at the same time is *locally more dispersed* (relatively ubiquitous) within this region. On the other hand, clusters of steel industries are spiker, i.e., *locally more concentrated*, and spread over a wider area, i.e., *globally more dispersed*.

[Figure 5]

A further increase in $D$ can be realized either by more global concentration or by more local concentration. Panels (a) and (b) in Figure 6 show the cluster patterns of "printing industries" (JSIC192) and "iron smelting without blast furnaces" (JSIC262), respectively. Although they have statistically indistinguishable values, 4.272 and 4.237, of $D$, again there is difference in spatial scale of agglomerations. Namely, the clusters of the former are globally more concentrated toward the major metro areas along the Pacific industrial belt between Tokyo and Fukuoka, while locally more dispersed over suburban areas around the metro areas (e.g., Tokyo, Nagoya, Osaka, and Fukuoka as indicated in the figure) along the belt, whereas those of the latter are locally more concentrated but scattered all over the country.

[Figure 6]

Such differences in spatial scales of agglomeration and dispersion between industrial patterns with very similar $D$ values are often more readily explained by differences in the local regional attributes most relevant for each industry. We return to this issue in Section 3.3 below.

## 2.3 Comparison with the Distance-Based Approach

An alternative *distance-based approach*, proposed by Duranton and Overman [5], focuses on the frequency distribution of bilateral distances among establishments.[10] The primary

[10] Macon and Puech [14] propose an alternative cumulative density approach based on the bilateral distance among establishments.

motivation for this approach was to avoid the Modifiable Areal Unit Problem (MAUP) by focusing on point pattern data (see, e.g., Combes and Overman [3]),[11] Here agglomeration within an industry is essentially characterized by the presence of significantly high frequencies of short distances between its firms. But while this approach is useful for identifying the spatial scale of such agglomerations, it provides no information about their locations. Hence the most important advantage of our cluster-based approach over this distance-based approach is its ability to pin down both the location and size of individual agglomerations. In particular, this allows one to study relations between agglomeration size and local regional attributes that may serve as possible determinants of agglomeration.

However, it should also be noted that the distance-based approach offers advantages in studying other aspects of agglomeration. In particular, Duranton and Overman [6] have used this approach to study co-localization behavior of establishments belonging to different categories (such as firms in different industries, or existing firms and new entries within the same industry). While it is in principle possible to contruct tests for co-localization within our prensent framework, their approach appears to be far simpler for this purpose.

# 3 Regression Analyses of Industrial Clusters

Given the cluster patterns identified above, the second stage of our analysis seeks to explain differences in these cluster patterns between industries in terms of local regional factors that may be more relevant for some industries than others. In particular, we employ regression analysis to explain differences in cluster employment sizes between industries in terms of selected local regional attributes. In Section 3.1 below we employ both municipality-level data (population, income, education levels, and climate) together with national-level data on inter-industry transaction linkages to construct a range of candidate explanatory variables to be used in the regressions. This is followed in Section 3.2 with a formulation and specification of the final regression model, and with results in Section 3.3.

## 3.1 Possible Local Determinants of Agglomeration

Below we consider five possible local determinants of agglomeration. Before doing so, however, it is important to stress that while this terminology is convenient for our purposes, and has become quite standard in the literature (as, e.g., in Rosenthal and Strange [21]), it is not meant to imply any notion of "causality". Our objective here is much more limited

[11] Note that our cluster-detection procudure can also be used at very fine regional levels, such as ZIP codes, where the MAUP is less of an issure. Hence this problem is not an intrinsic limitiation of our approach.

in nature, and focuses rather on the identification of local regional attributes that are (*i*) statistically significant correlates of cluster employment size in certain industries, and (*ii*) help to account for differences in spatial employment patterns between industries. Such attributes are here designated as *determinants of agglomeration*.

Before developing candidate determinants, it is useful to establish certain notational conventions. If the cluster scheme obtained for industry, $i \in I$, in the first stage above is denoted by $\mathbf{C}_i = (R_{i0}, C_{i1}, .., C_{ik_{\mathbf{C}_i,}})$, then the relevant *sample units* for the regressions to follow are taken to be industry-cluster pairs, $(ik : i \in I, k = 1, .., k_{\mathbf{C}_i})$. Where no ambiguity arises, we refer to such sample units simply as *clusters*. Hence the task at hand is to construct explanatory variables for each cluster, *ik*, that constitute potentially relevant determinants of agglomeration for indusries. We begin with the key market-oriented variables that appear to be relevant for virtually all industries, and then consider a number of labor-oriented and the first-nature variables.

### 3.1.1 Market Access

Access to consumers is perhaps the single most influential determinant of establishment locations. To model this, we adopt the *population share*, $P_r$, of each municipality, $r$, as a surrogate for the local consumer market in $r$ (where $\sum_{r \in R} P_r = 1$).[12] But the relevant consumer market for manufacturing establishments in $r$ may of course be much larger. Following standard methods, we model accessibility of establishments in $r$ to *all* population (in Japan) by introducing an exponential *distance decay function*,[13]

$$\alpha_{rs} \equiv \exp[-\tau d(r,s)] \, , \tag{2}$$

and taking the effective market access at $r$ to be representable by *population access*:[14]

$$A_r^P = \sum\nolimits_{s \in R} \alpha_{rs} P_s \, , \;\; r \in R \, . \tag{3}$$

While many choices for $\alpha$ are of course possible, there seems to be general agreement that effective bandwidth is more important than particular functional forms. In the present case, since $\alpha$ has maximum value 1 at zero distance, we take the *effective bandwidth*, $d_\tau$, for each value of the *decay parameter*, $\tau$, to be the distance (in kilometers) at which $\alpha$ falls to .01.

[12] Since only the *relative* attractiveness of each municipality is relevant for our purposes, such "share" variables will be used throughout.

[13] It should be noted that access to foreign markets versus domestic market is less of a locational issue in Japan than in countries with large interiors, such as the US. Virtually all major cities in Japan are on the coast, so that access to ports, for example, is virtually the same as access to population centers.

[14] It could be argued *total income* in each municipality is a more appropriate local market measure. But, if *income access* is then defined in the same manner as the population access, their correlation turns out to be 0.990. Hence from a practical viewpoint, these two accessibilities are statistically indistinguishable.

Here we consider three values of this parameter, $\tau \in \{0.20, 0.10, 0.05\}$ [as shown in Figure 7 below] with associated effective bandwidths $d_{.20} = 23.03, d_{.10} = 46.05$, and $d_{.05} = 92.10$, corresponding roughly to $25, 50$ and 100 kilometers, respectively.

[Figure 7]

In particular, the middle value, $\tau = 0.10$, corresponds to the standard definition of a "metropolitan area" given by Kanemoto and Tokuoka [11], and (as will be shown below) yields the single best representative value for our present accessibility measures. A map of the population access levels in (3) for all municipalities is shown in Figure 8 below for the case of $\tau = 0.10$.

[Figure 8]

These market access measures for each municipality are next employed to construct an appropriate market access measure for each cluster, $ik$. Here it seems natural to consider the *expected* market access for a randomly located establishment in cluster $ik$. If we denote the set of municipalities comprising cluster $ik$ by $R_{ik} \subseteq R$, and assume that all locations in the economic area, $a_r$, of each municipality, $r \in R_{ik}$, are equally likely, then the chance that a randomly located establishment is found in municipality, $r$, is given by the share, $p_r^{ik} = a_r / \Sigma_{s \in R_{ik}} a_s$, of economic area in $r$. Hence the desired measure for $ik$ is given by *expected population access*,

$$A_{ik}^P = \sum\nolimits_{r \in R_{ik}} p_r^{ik} A_r^P \; . \tag{4}$$

### 3.1.2 Transactions Access

In addition to consumer markets, industrial factor markets constitute a major determinant of establishment locations, both in terms of forward and backward linkages. The relative importance of each link for industry establishments can be captured by the transaction frequency data.[15] In particular, if the total reported transaction linkages between establishments in industries, $i, j \in I$, is denoted by $\ell_{ij}$, then for each industry $i \in I$, the *transactions share*,

$$T_{ij} = \frac{\ell_{ij}}{\sum_{h \in I - \{i\}} \ell_{ih}} \; , \quad j \in I - \{i\} \; , \tag{5}$$

denotes the fraction of all transactions of industry $i$ with industry $j$ (both forward and backward). Hence in terms of inter-industry linkages, it is reasonable to suppose that

[15] We use inter-industry transactions data, *Kigyo Sokan File*, reported in 2006 by Tokyo Shoko Research, Ltd. This questionnaire-based data set for 2005 contains a total of 4,066,704 transaction partners (suppliers, customers, major shareholders) as claimed by 146,135 individual firms in the manufacturing industries of the present study. In particular, the supplier, customer and major shareholder categories account, respectively, for 49.3%, 46.7% and 4.0% of all transactions. Finally, while each firm also lists the order of importance of these transaction partners, we give equal weight to all partners.

the relative importance of each industry $j$ for $i$ is represented by the transactions share, $T_{ij}$. Using these shares, we may construct a measure of transactions access appropriate for establishments in industry $i$ as follows. First, if $n_{rj}$ denotes the share of industry-$j$ establishments in municipality, $r$, (where $\sum_{r\in R} n_{rj} = 1$ for all $j \in I$), then by employing the distance decay function in (2), we can measure the *j-industry access* of municipality $r$ in a manner paralleling (3) as

$$A_{rj} = \sum\nolimits_{s\in R} \alpha_{rs} n_{sj}\ ,\quad r \in R\ ,\ j \in I\ . \tag{6}$$

By combining (5) and (6), we can then summarize the transactions access of muncipality $r$ relevant for establishments in industry $i$ by the *transactions access* measure,

$$A^T_{ri} = \sum\nolimits_{j\in I-\{i\}} A_{rj} T_{ij}\ ,\quad r \in R\ ,\ i \in I\ . \tag{7}$$

Finally, we can extend these to transactions access measures for clusters, *ik*, paralleling (4). If we again focus on the expected transactions access for a randomly located establishment in cluster *ik*, then by using the same notational conventions in (4), it follows that for each cluster *ik* the appropriate *expected transactions access* is given by

$$A^T_{ik} = \sum\nolimits_{r\in R_{ik}} p^{ik}_r A^T_{ri}\ . \tag{8}$$

But while these variables are perfectly meaningful for each industrial cluster, it turns out that for the vast majority of industries, expected transactions access across clusters is strongly correlated with expected access to all industries across clusters.[16] Thus, even if industry-specific access to transactions partners is a relevant predictor of cluster size, it is difficult to distinguish this effect statistically from access to industries as a whole. However, by considering transactions access *relative* to general industry access, one can obtain much sharper distinctions between industries. These relativized variables (used in our final regressions) can be defined as follows. First, if the total manufacturing establishment share in municipality $r$ is denoted by $n^I_r$ (where $\sum_{r\in R} n^I_r = 1$), then the overall *industrial access* of municipality $r$ is given by

$$A^I_r = \sum\nolimits_{s\in R} \alpha_{rs} n^I_s\ ,\quad r \in R \tag{9}$$

and the corresponding *expected industrial access* for cluster *ik* is given by

$$A^I_{ik} = \sum\nolimits_{r\in R_{ik}} p^{ik}_r A^I_r\ . \tag{10}$$

[16] The correlation between $\ln A^T_{ik}$ and $\ln A^I_{ik}$ defined by (10) below is 0.943 for the baseline case of $\tau = 0.1$.

In these terms, the *relative transactions access* for cluster *ik* is given by

$$F_{ik}^{T} = A_{ik}^{T} / A_{ik}^{I} \,. \tag{11}$$

This measure for each cluster essentially reflects the *density* of transaction partners with respect to all industries at nearby locations. For cluster sizes in particular, higher densities might serve to increase the visibility of nearby transaction partners, thus rendering certain clusters more attractive.

The distinction among $A_{ik}^{I}$, $A_{ik}^{T}$ and $F_{ik}^{T}$ can be illustrated by the industry, $i =$ "musical instruments" (JSIC342), as shown in Figure 9. Panel (c) shows the distribution of log *relative transactions access*, $\ln F_{ir}^{T} \equiv \ln(A_{ir}^{T}/A_{r}^{I})$ [as defined by (11) above],[17] for $i$ across municipalities, $r$. This is compared with both the numerator [log *transactions access*, $\ln(A_{ir}^{T})$] in Panel (b) and the denominator [log *industry access*, $\ln(A_{r}^{I})$] in Panel (a). Here it is clear that the numerator is much more correlated with the denominator than with the ratio.[18] Moreover, the ratio in this case turns out to be much more informative about the special features of transaction linkages in the musical instruments industry than is the numerator, as discussed further in Section 3.3.

[Figure 9]

### 3.1.3 Labor Access

In addition to consumer markets and inter-industry factor markets, it should be clear that *labor market* access is also a major determinant of establishment locations. Moreover, in a manner similar to transaction linkages, the relevant labor markets for each industry can be quite different. Perhaps the most basic distinction is between "high tech" and "low tech" industries, which tend to exhibit sharp differences in required skills. While the specifics of these requirements can be quite complex, such distinctions can be broadly captured in terms of education levels. For Japan, data is available on shares of population in each municipality with maximum eduction level corresponding to $J =$ "Junior highschool", $H =$ "Highschool"and $U =$ "University", where the last represents all educations levels beyond highschool.[19] These can in principle serve to differentiate the labor-skill requirments for each industry.

In a manner similar to general population access, we can formalize access to populations with various education levels as follows. If we denote a generic level of education by

[17] We use logs to reflect the fact that all variables are logged in the regressions to follow.

[18] Not surprisingly, both the numerator and denominator are also highly correlated with the log of general population access in (3).

[19] The total population of each municipality, together with sub-population totals of four education levels (junior-highschool, highschool, junior or technical college, university and above) are available from the *Population Census of Japan* in 2000.

$e \in Ed = \{J, H, U\}$, then as a refinement of the total population share, $P_r$, for each municipality, $r$, [as in (3) above], we denote the population share with education level $e$ in municipality $r$ as the *e-population share*, $P_r^e$ , in $r$.[20] As a parallel to (3) we may then define the corresponding *e-population access* of municipality $r$ by

$$A_r^e = \sum\nolimits_{s \in R} \alpha_{rs} P_s^e \;, \quad e \in Ed \;, \; r \in R \;. \tag{12}$$

Similarly, for each cluster $ik$, we can employ the same concepts and notation in (4) to define the corresponding *expected e-population access* of cluster $ik$ as follows:

$$A_{ik}^e = \sum\nolimits_{r \in R_{ik}} p_r^{ik} A_r^e \;. \tag{13}$$

In a manner similar to inter-industry transactions above, it is not surprising that these expected $e$-population access levels are all highly correlated with expected population access in (4). So by again taking ratios, and defining associated *relative e-population access*,

$$F_{ik}^e = A_{ik}^e / A_{ik}^P \;, \tag{14}$$

these relative values again yield sharper distinctions between $e$-populations which are useful for our present analysis.

As in Figure 9 above for inter-industry linkages, these distinctions are seen more clearly at the municipality level. Hence if the log of *relative e-population access* for municipality $r$ is denoted by $\ln F_r^e \equiv \ln(A_r^e / A_r^P)$ [as defined by (14) above], then the distributions of these values for $e =$ junior highschool ($J$), highschool ($H$), and university ($U$), are shown, respectively, in Panels (a), (b) and (c) of Figure 10.

[Figure 10]

Here one can see that populations with the highest education levels [Panel (c)] are heavily concentrated in the major urban centers shown. In contrast, those with the lowest education levels [Panel (a)] exhibit almost exactly the inverse relation, with lowest concentrations in urban centers. Finally, populations with intermediate education levels [Panel (b)] exhibit a spatial distribution qualitively different from either of these extremes. This distribution is mostly concentrated in suburban areas surrounding major population centers, as discussed in more detail in Section 3.3 below. The key point here is that the differences between these distributions are seen most clearly in terms of relative access measures.

[20] Note that since populations with education levels below $J$ are not included, the total of these shares $(P_r^j + P_r^H + P_r^U)$ is less than $P_r$.

### 3.1.4 Labor Cost Advantage

In addition to labor access, it should also be clear that labor costs, in terms of nominal wage rates, are a relevant determinant of establishment locations. For example, in industries where proximity to consumers is not essential but where production is highly labor intensive (such as textile products), it is reasonable to suppose that establishments are strongly attracted to areas with relatively low wages. As a proxy for the nominal wage, we use per-capita nominal (pre-tax) income, $y_r$, of salaried population (who are here interpreted as potential "workers") in municipality, $r$.[21] If the probability $p(r|s)$ that a worker in municipality $s$ chooses to work in municipality $r$ is assumed to be proportional to the relative access between $s$ and $r$, i.e., $\alpha_{sr} / \sum_{v \in R} \alpha_{sv}$, then the expected nominal wage in $r$ is given by

$$W_r = \textstyle\sum_{s \in R} y_s w_s \frac{\alpha_{sr}}{\sum_{v \in R} \alpha_{sv}} \, , \tag{15}$$

where $w_s$ the share of workers in municipality $s$. By using the same notational conventions in (4), it follows that for each cluster $ik$ the appropriate *expected nominal wage* in cluster $ik$ is then given by

$$W_{ik} = \textstyle\sum_{r \in R_{ik}} p_r^{ik} W_{ir} \, . \tag{16}$$

### 3.1.5 Natural Conditions

As potential "first nature" determinants of agglomeration, we consider coastal access, and three climate conditions: annual precipitation, mean hours of sunshine, and mean temperature.[22]

If the set of municipalities on *coastal boundaries* facing the ocean is denoted by $B \subset R$, then *coastal access*, $A_r^C$, of a given region $r \in R$ can then be defined by

$$A_r^C = \max_{s \in B} \alpha_{rs} \, . \tag{17}$$

Given the large share of domestic freight transportation accounted for by coastal shipping (42.1% in ton-km in 2001),[23] the most important measure of coastal access for a cluster appears to be the total amount of its land accessible to the coast.[24] Thus, rather than using

[21] Both the total before-tax annual income and the number of tax payers in 2000 are available for each municipality in the report, *Shi-cho-son Zei Kazei Jokyo Tou No Shirabe* (2001) from the Local Tax Bureau in the Ministry of Internal Affairs and Communications of Japan. This data distinguishes between salaried and non-salaried populations, and allows us to use per-capita nominal income per salaried worker as the proxy for average nominal wage of workers.

[22] All climatic data are obtained from the *Mesh Climatic Data 2000* provided by Japan Meteorological Agency. Each data set involves 1km mesh data based on mean values from 1971 to 2000. The value in each municipality thus represents the mean value of all data points within the municipality.

[23] This share is computed from Table 1-1 of the report, *Rikuun Toukei Yoran* (in Japanese), provided by Ministry of Land, Infrastructure, Transport and Tourism of Japan (2006).

[24] This is also consistent with our preliminary analyses, which indicated that total coastal access was the

the average accessibilities defined for clusters so far, the coastal access of a given cluster *ik* is here defined in terms of *total* access,

$$A_{ik}^{C} = \sum\nolimits_{r \in R_{ik}} a_r A_r^C \ . \tag{18}$$

Finally, we consider three possibly relevant *climate conditions*, denoted respectively $Pcp =$ annual parcipitation, $Sun =$ mean hours of sunshine, and $Temp =$ mean temperature. If the level of each climate condition, $m \in Cl \equiv \{Pcp, Sun, Temp\}$ in municipality, $r \in R$, is denoted by $A_r^m$, then we assume that the corresponding climate condition for each cluster, *ik*, is given by

$$A_{ik}^{m} = \sum\nolimits_{r \in R_{ik}} \overline{p}_r^{ik} A_r^m \ , \quad m \in Cl \tag{19}$$

where $\overline{p}_r^{ik}$ is the share of *total area* (rather than economic area) of municipality *r* in cluster *ik* (i.e., $\sum_{r \in R_{ik}} \overline{p}_r^{ik} = 1$).

## 3.2 Regression Methodology

In this section, we propose a basic regression framework for analysis of industrial clusters in relation to the local regional attributes defined in the previous section. In Section 3.2.1, we first formulate the basic regression model and discuss some of its properties. In Section 3.2.2 we then summarize the results of a number of exploratory analyses using this model. In particular, we pare down the set of final explanatory variables based on these results. The final model specification is then presented in Section 3.2.3.

### 3.2.1 A Regression Model for Comparative Analyses of Cluster Size

In the analyses to follow, we characterize the size of each individual cluster for an industry in terms of employment levels. So for each sample, *ik*, the dependent variable of interest is taken to be the *employment level*, $E_{ik}$, of cluster *k* in industry *i*.[25] As is often true of nonnegative dependent variables, log transformations tend to yield more reliable parameter inference with respect to normality assumptions. This is indeed the case here, so that the dependent variable of interest is taken to be *log employment*, $\ln E$. As mentioned above, our primary objective is to identify a set of local regional attributes, *Q*, such that the corresponding attribute variables, $\{X_q : q \in Q\}$, are potentially significant predictors of

most effective explanatory variable for use in the regressions of Section 3.3 below.

[25] Though one could also use the number of industry establishments in each cluster, the results are essentially the same. In terms of our ultimate goal in this analysis, the best choice for a dependent variable would of course be the industrial productivity of each cluster. But as discussed further in Section 5 below, this data is not currently available at the desired level of aggregation.

log employment. In addition, we are interested in comparing differences in the significance of predictors across industries. The simplest way to do so is to treat industries as "fixed effects" by introducing industry dummies, $(d_i : i \in I)$, [with $d_i(j) = 1$ if $i = j$ and $d_i(j) = 0$ otherwise]. In this way, we can distinguish the relevant sets of local regional predictors for each industry $i$ by interacting $d_i$ with each predictor variable. This yields a multiple regression model of the following form:

$$\ln E_{ik} = \beta_0 + \sum_{j \in I} \beta_j d_j(i) + \sum_{q \in Q} \beta_q X_{ik,q} + \sum_{j \in I} \sum_{q \in Q} \beta_{jq} d_j(i) X_{ik,q} + \varepsilon_{ik} \,, \tag{20}$$

where the residuals, $\varepsilon_{ik}$, are assumed to be *iid* normal.[26] But to interpret this model in the most natural way, it is convenient to renormalize these dummy variables to sum to zero. While this procedure is quite standard, it is worthwhile making it explicit here in order to facilitate the subsequent interpretation of all parameters. To do so, we start by selecting a reference industry, say $i = 1$, and defining a new set of indicator variables $(\delta_i : i \in I)$ as follows:

$$\begin{aligned} \delta_i(j) &= d_i(j) - d_1(j) \,, \ \ i \neq 1 \,, \qquad (21) \\ \delta_1(j) &= -\sum_{i \neq 1} d_i(j) \,. \qquad (22) \end{aligned}$$

In these terms, expression (20) is then simply rewritten as

$$\ln E_{ik} = \beta_0 + \sum_{j \in I} \beta_j \delta_j(i) + \sum_{q \in Q} \beta_q X_{ik,q} + \sum_{j \in I} \sum_{q \in Q} \beta_{jq} \delta_j(i) X_{ik,q} + \varepsilon_{ik} \,. \tag{23}$$

To see the advantage of this renormalization, consider any given industry $i \neq 1$ and observe that since $\delta_i(i) = 1$ and $\delta_i(j) = 0$ for $j \neq i$ [by expresssion (21)], it follows that for all clusters, $k$, of industry $i$, expression (23) reduces to:

$$\begin{aligned} \ln E_{ik} &= \beta_0 + \beta_i + \sum_{q \in Q} \beta_q X_{ik,q} + \sum_{q \in Q} \beta_{jq} X_{ik,q} + \varepsilon_{ik} \\ &= (\beta_0 + \beta_i) + \sum_{q \in Q} (\beta_q + \beta_{iq}) \, X_{ik,q} + \varepsilon_{ik} \,. \qquad (24) \end{aligned}$$

Similarly, since $\delta_1(1) = 0$ and $\delta_1(j) = -\sum_{i \neq 1} d_i(j) = -1$ for all $j \neq 1$ [by expresssion (22)], the corresponding regression for the reference industry, $i = 1$, takes the form

$$\begin{aligned} \ln E_{ik} &= \beta_0 - \sum_{i \neq 1} \beta_i + \sum_{q \in Q} \beta_q X_{ik,q} + \sum_{q \in Q} \left( -\sum_{i \neq 1} \beta_{jq} \right) X_{ik,q} + \varepsilon_{ik} \\ &= \left( \beta_0 - \sum_{i \neq 1} \beta_i \right) + \sum_{q \in Q} \left( \beta_q - \sum_{i \neq 1} \beta_{jq} \right) \, X_{ik,q} + \varepsilon_{ik} \,. \qquad (25) \end{aligned}$$

[26] The possibility of spatially autocorrelated residuals is discussed in detail in Section 4 below.

Expressions (24) and (25) amount to a set of separate regressions for each industry, with the added feature that the usual intercept and slope parameters have been decomposed into two parts. By focusing on the slope parameters for a given explanatory variable, $q$, we see that these parameters for each industry have a common part, $\beta_q$, together with individual parts that *sum to zero*. Hence the *mean slope* across all industries is precisely $\beta_q$, and the individual parts represent *deviations* from this common mean. Notice also that the choice of reference industry here makes no difference. In particular, if we now set $\beta_1 = -\sum_{i\neq 1} \beta_i$ and $\beta_{1q} = -\sum_{i\neq 1} \beta_{jq}$, so that expression (24) holds for all industries, then the resulting parameter values are easily seen to be invariant with respect to the choice of reference industry. Hence this renormalization provides a natural intepretation of all parameters.

The *main effect*, $\beta_q$, for each explanatory variable in (24) is the average slope coefficient of regional variable, $X_q$, across all industries, and the *interaction effect*, $\beta_{iq}$, in (24) represent the slope deviation for each industry $i$. So for example, a significantly negative value of $\beta_{iq}$ does not necessarily imply a negative influence of $X_q$ on the (log) employment size of clusters in industry $i$. Rather it indicates that this influence is significantly below the average for all industries. To determine the actual sign of this influence, one must test the sign of the *whole effect*, $\beta_q + \beta_{iq}$, for industry $i$. We shall return to these points when discussing regression results in Section 3.3 below.

Finally it should be noted that the same argument applies to the intercept term, $\beta_0 + \beta_i$, in (24) as well. So here $\beta_i$ is not a "main effect" for industry $i$. Rather it is again a deviation, with significant positive (negative) values denoting intercepts significantly above (below) average. So a significantly negative value of $\beta_i$, for example, would indicate that employment levels in the clusters of industry $i$ tend to be smaller than average, other things being equal.

### 3.2.2 Model Specifications and Preliminary Findings

A number of exploratory analyses were run using the regression framework above, together with the full set of candidate explanatory variables in Section 3.1. Here it was found that, as with the dependent variable, more stable results were obtained by using log tranformations of all (nonnegative) explanatory variables. But even with these transformations it was found that certain variables provided little information about variations in employment size among clusters.

**Climate Variables** This was particularly evident for the three climate variables, $\ln A^{Pcp}$, $\ln A^{Sun}$ and $\ln A^{Temp}$. Here it was found that none of these variables were even weakly significant in any regressions. Moreover, as can be seen in Table 1, none of these three variables exhibit substantial correlation, either with each other or with any other candidate explanatory variables. So collinearities do not appear to be an issue here.

[Table 1]

However, further investigation suggested that climate does indeed matter at a more fundamental level. For most industries, it appears that climatic conditions influence almost all location decisions in a similar way. For example the "communication equipment" industry (JSIC304) tends to concentrate most of its production in the relatively dry Tohoku region of Japan, as seen in Figure 4(c). A similar example is provided by the "watches and clocks"industry (JSIC327), not shown.

While these examples are of course anecdotal, one can actually test such hypotheses by looking for significant climatic differences between municipalities in clusters for a given industry and those municipalities not in clusters. More formally (recalling the notation in Section 2.1 above), if for each industry $i$ we let $R_i$ denote the set of all municipalities in the clusters of scheme, $\mathbf{C}_i$, so that by definition the set of all municipalities not in clusters is given by, $R_{0i} = R - R_i$, then for each climatic variable, $m \in \{Pcp,Sun,Temp\}$, one can test whether mean values of $m$ are significantly different between municipalities in $R_i$ and $R_{0i}$. While $t$-tests of such mean differences can easily be constructed using indicator variables similar to the regression framework above, non-parametric tests seem to provide a more robust alternative here. In the present case we employ the Mann-Whitney $U$-Test. Since this test is developed in detail is Section 5 below, it suffices to simply sketch its application here. In the present case, the null hypothesis of "no difference"can be modeled by simulating randomly reassigned values of $m$ between municipalities in $R_i$ and $R_{0i}$. If the $U$-statistic for testing mean difference in variable $m$ is denoted by $U_m$, then one can simply compare the observed $U_m$-value with the corresponding frequency distribution of simulated $U_m$-values.[27] For example, if the observed value is significantly small (in the top 5%) then one may conclude that mean levels of $m$ are significantly higher inside clusters than outside. We have conducted 1000 simulated reassignments for each climatic variable, $m \in \{Pcp, Sun, Temp\}$, and the test results are very clear. For 86%, 95%, and 99% of industries, respectively, municipalities inside clusters have significantly less precipitation, more sunshine, and higher mean temperature.

Hence, while these climatic conditions appear to be important determinants of industrial location in general, they do not help to explain size variations of clusters in terms of employment. For this reason, the corresponding variables, $\ln A^{Pcp}$, $\ln A^{Sun}$ and $\ln A^{Temp}$, were not used in the final regressions.

[27] To construct $U_m$ in the present case, one starts with any municipality, $r \in R_{0i}$, and counts the number, $n_r$, of municipalities in $R_i$ with less favorable values of climatic condition, $m$ (such as "more rainfall"in the case of communication equipment). Then $U_m$ is simply the sum of these values, i.e., $U_m = \sum_{r \in R_{0i}} n_r$.

**Labor-Related Variables** While collinearties were not an issue with these climatic variables, the opposite is true for labor-related variables. This is not surprising in view of the fact that (*i*) population access, $\ln A^P$, is a crucial variable for predicting employment sizes, and (*ii*) many labor-related variables are closely tied to population. This is particularly true for labor costs, $\ln W$, which exhibited the single highest correlation (0.932) with $\ln A^P$among all pairs of variables studied. The inevitable conlusion here seems to be that this income variable is too closely tied to market access to provide any additional information. So in the absence of more detailed information on labor costs, it was decided to drop $\ln W$ from the final analysis.

This is also true of the education variables that reflect access to different labor skills. This can be seen in Table 1, where the three highest absolute correlations (shown in bold) all involve eduction variables. In particular, there is a strong negative correlation (-0.9233) between the variables, $\ln F^J$ and $\ln F^U$, reflecting relative access to junior-high and university educated workers, respectively. It is also clear that both these variables are strongly correlated with population access, $\ln A^P$. This adds further confirmation to the map comparisons in Section 3.1.3 above, where it was seen that the highest educated workers are concentrated in population centers, while the lowest educated workers exhibit the opposite tendency. Not surprisingly, the results of the full regression model (23) showed considerable variation, depending on which subset of these three variables ($\ln A^P, \ln F^J, \ln F^U$) was present. Since tri-variate collinearities of this type are not fully captured by pairwise correlations, it was decided to examine this issue further by regressing both $\ln F^U$ and $\ln F^J$ on all other explanatory variables. In both cases, $R^2$ was close to 90% percent. Even when $\ln F^J$ was omitted from the $\ln F^U$ regression, and visa versa, $R^2$ was still above 70% in both cases. So in spite of the potential relevance of all these variables for predicting cluster employment sizes, it was necessary to omit some from the final regression. Since population access, $\ln A^P$, was again deemed to be the most important of these three variables (both with respect to labor-demand and final-demand considerations), it was decided to omit both $\ln F^U$ and $\ln F^J$ from the final regression. It is also worth noting that that by omitting $\ln F^U$ and $\ln F^J$, the resulting variance inflation factor on $\ln A^P$ was drastically reduced, thus ensuring more reliable estimate of its effect on cluster employment size.

Finally, as noted in Section 3.1.3 above, relative access to highschool-educated workers, $\ln A^H$, appears to behave quite differently from the two extremes above. This is again evident in the correlations of Table 1. More importantly, our exploratory regressions suggest that this variable does indeed provide useful additional information about cluster employment sizes. As will be discussed further in Section 3.3 below, workers with this intermediate level of education are mostly concentrated in suburban area around major population centers. These areas also appear to offer strong locational advantages for certain types of industries. In particular, they combine access to urban centers with relatively low costs of both labor and land.

### 3.2.3 Final Model Specification

These preliminary analyses have led to a certain paring down of our intial list of candidate variables. So the final set of explanatory variables $(X_q : q \in Q)$ in model (23) consist of the four variables:

$$\begin{aligned}
\ln A^P &= \textit{Population Access} \\
\ln A^C &= \textit{Coastal Access} \\
\ln F^T &= \textit{Relative Transactions Access} \\
\ln F^H &= \textit{Relative Highschool-Education Access}
\end{aligned}$$

which we shall occasionally denote by $Q = \{P, C, T, H\}$. For sake of completeness, it is convenient to specify this model for a typical industry-cluster sample, *ik*, as is done in model (24) above:

$$\begin{aligned}
\ln E_{ik} &= (\beta_0 + \beta_i) + (\beta_P + \beta_{iP}) \ln A^P_{ik} + (\beta_C + \beta_{iC}) \ln A^C_{ik} \\
&\qquad + (\beta_T + \beta_{iT}) \ln F^T_{ik} + (\beta_H + \beta_{iH}) \ln F^H_{ik} + \varepsilon_{ik} \,. \qquad (26)
\end{aligned}$$

It is this model that we shall analyze in detail below.

## 3.3 Regression Results

First recall from the discussion in Section 3.1.1 that the distance decay parameter for this analysis is set to $\tau = 0.10$.[28] Hence all access variables are defined in terms of this value. Next observe that for the 154 industries with non-spurious clusters, there are a total of 12,350 clusters. So the number of distinct samples, *ik*, in this regression is 12,350. With these preliminary observations, the summary results of this regression are displayed in Table 2. Note first that the low value of adjusted $R^2$ (0.3447) reported here serves to underscore the data limitations of the present study.[29] As more data becomes available at the municipality level, we anticipate that a more complete account of agglomeration determinants can be given. Hence the present regression is best viewed as a demonstration of the types of local spatial analyses that can be carried out with cluster information obtained in stage one of our proposed approach.

[Table 2]

[28] The basic results remain essentially the same under $\tau = 0.05, 0.20$.

[29] However, it is worth noting that such low values are not uncommon in the industrial agglomeration literature. For example, while the many regressions in Rosenthal and Strange [21] are not directly comparable with our framework, the maximum values of adjusted $R^2$ for their regressions are between 0.3 and 0.4.

To interpret the additional summary results in Table 2, note first that the variable, $\delta$, here represents the full set of indictator variables, $\delta_i$, for all industries, $i \in I$, each with corresponding coefficient, $\beta_i$, in expression (26). Since these coefficients must sum to zero by contruction, there are only 153 $(= 154 - 1)$ independent parameters. So the first line of this summary table reports the results of testing the null hypothesis that $\beta_1 = \cdots = \beta_{153} = 0$. The $F$ value (8.8434) shows that there is indeed a great deal of variation in average cluster employment levels among these industries.[30] Similarly, the interaction variables, such as $\delta \times \ln A^C$ in the last row of the table, represent the full set of interactions, $\delta_i \times \ln A^C$, between industries, $i \in I$, and in this case, coastal access. Each interaction variable has associated coefficient, $\beta_{iC}$, in expression (26), where again only 153 are independent. So the last row of the table summarizes the results of testing the null hypothesis, $\beta_{1C} = \cdots = \beta_{153,C} = 0$. Here the $F$ value (1.7443) again indicates that (for this large sample size) there are significant differences between industries in terms of the importance of coastal access as a predictor of cluster employment size. Similar conclusions hold with respect to all interaction variables in the table. The rows corresponding to individual explanatory variables, such as $\ln A^C$, simply report the significance of "main effects" (average slopes), such as $\beta_C$ in expression (26). All are significantly different from zero except for $\ln A^H$, indicating that the average slope, $\beta_H$, on access to highschool-educated labor is not significantly different from zero. As will be evident below, this does *not* imply that this variable is insignificant across industries, but rather the positive and negative effects tend to balance out on average.

Finally, we turn to the results of most interest, namely the significance results for predictors of cluster-employment size by industry. These are reported in Table 3 below. Since there is a great deal of information conveyed in this table, it is instructive to take industry, $i =$ "seafood products", in the first row as an illustrative example. Note that we have included the divergence values, $D$, for each industry in the first column (which for seafood products is $D_i = 1.6464$, indicating a farely ubiquitous industry as mentioned in Section 2.2 above). Turning next to population access, $\ln A^P$, the actual coefficient value shown is the regression estimate, $\widehat{\beta}_P + \widehat{\beta}_{iP} = 0.2674$, of the *whole effect*, $\beta_P + \beta_{iP}$, for $\ln A^P$ in expression (26). Similarly, all coefficient values shown in each column denote whole effects, $\widehat{\beta}_q + \widehat{\beta}_{iq}$, $q \in \{P, C, T, H\}$. The number in parentheses below this coefficient is the associated $t$-value (3.98), indicating that this effect is significantly positive.[31] Such significance is also indicated by $^{**}$ in the usual star system [where $^{**}$ denotes significance at the 0.05 level and $^{*}$ denotes (weak) significance at the 0.10 level]. The value of the *interaction*

[30] For ease of exposition, we drop all references to "logged"values in the discussion to follow.

[31] Here it should be noted that while standard software reports significance levels for individual coefficients, this is not true for sums of coefficients. Such customized tests (especially large numbers of such tests as in the present case) must usually be done either off-line or with customized scripts. We have written a Matlab program for this purpose.

*effect*, $\beta_{iP}$, is not shown. But its sign and level of significance are shown by the double minus ($--$) placed under $^{**}$, indicating that $\beta_{iP}$ is significantly negative. So for this example, even though population access is a significant predictor of cluster employment for the seafood industry, its effect is significantly smaller than average for all industries. More generally a single minus ($-$) denotes a weakly signficant interaction effect, while ($++$) and ($+$) denote significant and weakly significant positive interaction effects, respectively. For example, the last entry in this row shows that the estimated whole effect ($\widehat{\beta}_C + \widehat{\beta}_{iC} = 0.2274$) of coastal access for the seafood industry is not only very significant, but is also (not surprisingly) way above average ($++$).

[Table 3]

Given this general description of Table 3, we now discuss its substantive results in more detail. In doing so, our main objective is to illustrate the kinds of information provided by this two-stage procedure that is not obtainable by any scalar measure of agglomeration degree. For convenience, we again focus on the examples presented in Section 2.2 above, starting with less concetrated industries (in terms of $D$ index) and proceeding to more concentrated industries.

### 3.3.1 Less "Concentrated" Industries

Recall from Section 2.2 that industries with very low degrees of agglomeration (in terms of $D$) are by definition present almost everywhere (ubiquitous), and in this sense exhibit similar spatial distributions. Typical examples of such industries are "bakeries and confectionery products" (JSIC127) and "sliding doors and screens" (JSIC173) as discussed in Section 2.2 above (refer to Figure 3). Since these ubiquitous industries tend to be strongly consumer oriented, it is not surprising that their main agglomeration determinants also tend to be quite similar. As indicated in Table 3, the coefficient of market access for both of these industries is positive and significantly larger than average.[32] So in such extreme cases, scalar measures of agglomeration do seem to work rather well, i.e., industries with low degrees of agglomeration are necessarily similar in many other important ways.

However, as the degree of agglomeration increases, industries with similar $D$ values start to exhibit subtantially different patterns of spatial agglomeration. As discussed in Section 2.2, "textile outer garments and shirts" (JSIC151) and "seafood products" (JSIC122) have very different cluster patterns [refer to Panels (a) and (b) in Figure 4 above, respectively], although they have statistically indistinguishable values of $D$. On the one hand, textile industries are typically labor cost sensitive, concentrating more in low-wage

[32] Ubiquitous industries whose clusters are densely dispersed over the nation all exhibit similar tendencies. This includes most food products, as well as weight/bulk gaining industries such as "beverages and feed" (JSIC13), "lumber and wood products" (JSIC16), and "furniture and fixtures" (JSIC17) .

peripheral regions such as Kyushu, San-in and Tohoku as indicated in Figure 4(a). As mentioned in Section 3.2.2, this effect is picked up by the strongly positive coefficient for highschool labor access in the "textile outer garments and shirts" industry, reflecting suburban concentration of clusters around major metro areas (e.g., Tokyo, Osaka and Nagoya) as shown in Figure 4(a). On the other hand (as pointed out in the illustration above), the dominant agglomeration determinant for "seafood products" is coastal access.

A second example in this range of $D$ values is "communication equipment" (JSIC304) versus "paper containers" (JSIC185), which are also statistically indistinguishable in terms of $D$. But, again these two industries have very different cluster patterns as depicted in Panels (c) and (d) in Figure 4 above, respectively. The dominant effect for "communication equipment" is clearly transactions access, which is significantly positive and way above average. This reflects the heavy concentration of this industry near its transaction partners in the northeast region of the main island. On the other hand, the "paper containers" industry is more consumer oriented, with clusters found mainly in the suburbs of major metro areas along the Pacific Coast between Tokyo and Fukuoka. In the present regression results, this pattern is consistent with the above-average positive effects of both market and highschool labor access.[33]

### 3.3.2 More "Concentrated" Industries

At higher levels of concentration, these differences are well illustrated by "musical instruments" (JSIC342) and "steel with rolling facilities" (JSIC263), which have statistically indistinguishable $D$ values [refer to Panels (a) and (b) in Figure 5 above, respectively]. But, again, both their cluster locations and significant agglomeration determinants are quite different. With respect to the "musical instruments" industry, the major clusters (Tokyo, Hamamatsu, and Nagano) are all within a one-day trip of each other. There is also a strong overlap in the range of modern musical instruments produced in each cluster, suggesting strong production linkages between them.[34] This is quite consistent with our regression results, where the dominant factor for this industry is clearly seen to be transactions access.

In contrast, the dominant factor for "steel with rolling facilities" is seen to be access to population centers, which in Japan coincide with major port locations along the Pacific Coast. Given the weight/bulk gaining nature of its outputs, final outputs are highly sensitive to transport costs, so that access to major ports is important for steel exports. Moreover, since most of its final markets inside Japan are also concentrated in population centers, this creates an added locational incentive (as can be seen by the positive significance of transactions access). So while coastal access does exhibit some degree of positive

[33] Recall that highschool labor access acts as a surrogate for suburban attraction.

[34] Here it should be noted that most of the other clusters in Figure 5(a), including Fukuyama, Morioka, and Minamiaizu, specialize in more traditional Japanese musical intruments.

signficance for this industry, such effects are heavily outweighed by access to population centers.

Our final example, at even higher levels of concentration, is provided by the "publishing" (JSIC192) industry versus the "iron smelting without blast furnaces" (JSIC262) industry. While their degrees of concentration are virtually indistinguishable, their major agglomeration determinants are again quite different [refer to Panels (a) and (b) in Figure 6 above, respectively]. On the one hand, "publishing" is a typical market-oriented urban industry for which population access is clearly the major determinant (together with transactions access). But for "iron smelting without blast furnaces", the major determinant is seen to be coastal access. In contrast to "steel with rolling facilities" above, this industry is strongly weight/bulk reducing in terms of final outputs. So coastal access is far more critical for its inputs than its outputs. This is also evident in the below-average significance of transactions access for this industry, again reflecting the locational importance of inputs versus outputs.

In addition to these specific examples illustrating the limitations of scalar agglomeration measures, Table 3 contains a wealth of information about the agglomeration determinants of manufacturing industries in Japan. Just to mention a few, coastal access is seen to be particularly important for industries like "spinning mills" (JSIC142), where inputs are mostly imported and outputs are mostly exported. Another interesting class of industries are those typically involving vertically integrated firms, such as those in the two-digit categories, "chemical and allied products" (JSIC20), "leather products and fur skins" (JSIC24), "ceramic, stone and clay products" (JSIC25), "non-ferrous metals and products" (JSIC27), and "general machinery" (JSIC29). As seen from the table, these industries naturally tend to be quite sensitive to transactions access. Finally, industries requiring both large land inputs and access to consumer markets, like those in the two-digit catergory "pulp and paper products" (JSIC18), are generally attracted to suburban areas which satisfy both these requirements. Consequently, their coefficients for both market and highschool labor accessibilities are significantly positive.

# 4 Spatial Autocorrelation Issues

As is well known, regressions involving spatial units often exhibit spatial autocorrelation effects which violate the assumption of independent residuals [as in (20) above]. But, as mentioned in the Introduction, industrial clusters tend to exhibit far less spatial autocorrelation than other spatial units in common use. To justify this claim, we now compare the spatial autocorrelation of regression residuals at various levels of spatial aggregation, by essentially repeating the analysis above at both the municipality level and the prefecture level in Japan. At the municipality level, recall that the 3207 municipalities of Japan were shown in Figure 1 above. At the prefecture level, the 46 prefectures of Japan used in the present study are shown in Figure 11 below.[35]

[Figure 11]

Since clusters of different industries overlap, systematic comparisons can only be made on an industry-by-industry basis. So here the full regression model with interaction effects in (26), is replaced by a single-industry version defined for each industry $i \in I$. To do so, it is convenient to denote the set of prefectures by *Pref*, and to denote the set of clusters for each industry $i$ by $K_i$. With this notation, the general regression model for this section can be defined in terms of spatial units, $s \in S$, which are taken to mean, $r \in R$, at the *municipality* level, $k \in K_i$, at the *cluster* level, and $p \in Pref$, at the *prefecture* level. In these terms, our regressions at each level for industry $i$ can now be written in terms of sample units, $is$, as follows:

$$\ln E_{is} = \beta_0 + \beta_P \ln A_{is}^P + \beta_T \ln F_{is}^T + \beta_H \ln F_{is}^H + \beta_C \ln A_{is}^C + \varepsilon_{is}, \quad s \in S\ . \tag{27}$$

To illustrate the appropriate construction of explanatory variables in (27) at each level, $s \in \{r, k, p\}$, we focus on population access. First, for $s = r$, the quantity, $A_{is}^P$, is precisely expression (3) above, namely

$$A_{ir}^P = \sum\nolimits_{s \in R} \alpha_{rs} P_s\ ,\ \ r \in R\ . \tag{28}$$

Similarly, for $s = k$, the quantity, $A_{is}^P$, is given by (4) above, i.e.,

$$A_{ik}^P = \sum\nolimits_{r \in R_{ik}} p_r^{ik} A_r^P\ ,\ \ k \in K_i\ . \tag{29}$$

Finally, if the set of municipalities in prefecture, $p$, is denoted by $R_p \subset R$, then for $s = p$, the quantity, $A_{is}^P$, is given by

$$A_{ip}^P = \sum\nolimits_{r \in R_p} p_r^{ip} A_r^P\ ,\ \ p \in Pref \tag{30}$$

where $p_r^{ip} = a_r / \Sigma_{v \in R_p} a_v$. All other explanatory variables are defined at each level in a similar manner. With respect to these explanatory variables, the coefficients in (27) are simply standard beta coefficients. So for $s = k$ in particular, coefficients like $\beta_P$ in (27) now correspond to *whole effects* rather than to *main effects*, as in (26).

As with any multiple regression, the residuals, $\varepsilon_{is}$, in (27) are implicitly assumed to be *iid* normal at each given level of aggregation. If this assumption is treated as a null hypothesis, then this hypothesis can be tested using Moran's **I** statistic. As with all such

[35] One prefecture, Okinawa, is excluded as it is disconnected from other major islands.

spatial tests, this statistic is constructed with respect to some pre-specified measure of spatial proximity. In the present case, since all access variables are defined for each pair of spatial units, $s, v \in S$ [as in (2)] by the distance decay function[36]

$$\alpha_\tau(s,v) = \exp[-\tau d(s,v)]\,, \tag{31}$$

we use this as our basic measure of proximity. But to add further flexibility to the test statistic, we also employ $m$-nearest-neighbor truncations of this decay function.[37] So if for each spatial unit, $s \in S$, we denote the set of $m$ nearest neighbors of $s$ by $N_m(s) \subset S$, then the relevant set of *spatial proximity weights* are here defined for each pair of spatial units $(s,v) \in S^2$ by:

$$w_{m\tau}(s,v) = \begin{cases} \alpha_\tau(s,v) & , \ v \in N_m(s) \\ 0 & , \ \text{otherwise} \end{cases} \tag{32}$$

With these definitions, if we now denote the estimated residuals in regression (27) by $\widehat{\varepsilon}_{is}$, then for each industry, $i \in I$, the relevant family of Moran's **I** statistics for our purposes takes the form:

$$\mathbf{I}_{m\tau}(i) = \frac{|S|}{\sum_s \sum_v w_{m\tau}(s,v)} \frac{\sum_s \sum_v w_{m\tau}(s,v)\widehat{\varepsilon}_{is}\widehat{\varepsilon}_{iv}}{\sum_s \widehat{\varepsilon}_{is}^2}\,, \tag{33}$$

where $|S|$ denotes the appropriate number of spatial units (i.e., municipalities, clusters, or prefectures).

Here, rather than appealing to the asymptotic normality of $\mathbf{I}_{m\tau}(i)$ under the null hypothesis of independent residuals, we choose to use random permutation tests based on $\mathbf{I}_{m\tau}(i)$. In these terms, the appropriate null hypothesis is taken to be that the estimated residuals $(\widehat{\varepsilon}_{is})$ are not statistically distinguishable from random spatial permutations of these values.[38] To test this hypothesis, the given set of regression residuals $(\widehat{\varepsilon}_{is})$ are randomly permuted (reassigned) among spatial units in $S$ a number of times ($t = 1, .., T$). For each permutation, $t$, Moran's **I** is recomputed as $\mathbf{I}^{(t)}_{m\tau}(i),\ t = 1, .., T$, and the histogram of these values is used to estimate the sampling distribution of $\mathbf{I}_{m\tau}(i)$ under the null hypothesis. One can then construct $p$-values for a standard two-sided tests of this hypothesis.

Here a number of combinations of $(m, \tau)$ were tried for all industries, using $m \in \{3, 5\}$ and $\tau \in \{0.05, 0.10, 0.20\}$. For each combination, permutation tests were run using $T = 1000$ random permutations. Since there was little difference between values of $m$, only

[36] Here distances, $d(s, v)$, between both cluster pairs and prefecture pairs are computed as median distances between their constituent municipalities.

[37] As is well known, such tests of spatial autocorrelation in terms of distance decay (kernel) functions are often sensitive to truncation points (kernel bandwidths).

[38] This is actually a somewhat weaker version of the above independence hypothesis, in the sense that spatial residuals are only hypothesized to be "exchangeable"with respect to locations. But given the present sample sizes (expecially at the prefecture level), such permutation tests are typically more robust that tests based on asymptotic normality.

results for $m = 3$ are reported.

Table 4 summarizes the share of industries (out of 154) for which spatial autocorrelation appears to be significant (i.e., for which the randomness hypothesis is rejected). These shares are reported for both 0.01 and 0.05 levels of significance. These results are organized with rows denoting selected distance decay levels, $\tau = 0.05, 0.10$, 0.20, and columns denoting significance levels, grouped by levels of aggregation (municapilities, clusters, prefectures).

[Table 4]

Note first that with respect to distance decay, there is little qualitative difference between the three rows of Table 4. As expected, there is a general tendency for the fractions of significant values to decrease as $\tau$ increases (since all distance effects are diminishing toward zero). But even these effects are minimal within the given range of $\tau$ values.

Far more interesting for our purposes are the differences between levels of spatial aggregation. In this regard, it should be clear from the first two columns that spatial autocorrelation of regression residuals is strongly evident at the municipality level. This is not surprising in view of our clustering results, which show that industry clusters almost always consist of many contiguous municipalities. Hence the residuals for these municipalities must surely include many common unobserved properties of both the firms in these clusters and the agglomeration determinants attracting these firms. A more surprising fact is that significant *negative* levels of autocorrelation are quite common here, and in fact are present about 45% (70 of 154) industries at the municipality level for the case of $\tau = 0.10$. We shall return to this issue in the discussion of autocorrelation at the cluster level below. But the key conclustion at the municipality level (as suggested in the Introduction), is that these spatial units are simply too small to capture meaningful relations between industrial employment and potential determinants of agglomeration.

## 4.1 Cluster-Level Autocorrelation

Turning next to the cluster level, it is clear that spatial correlation effects are dramatically reduced. While rejection frequencies are still somewhat larger than the theoretical “size”of these tests,[39] the assumption of independent residuals appears to be much less of a problem at the cluster level.

Nonetheless, one can gain further insight here by noting that in cases where autocorrelation is significant at the cluster level, this appears to be largely due to the presence of *contiguous* clusters. But even here there appear to be two types of effects. These are well

[39] Under “perfect randomness”, one would expect to reject about 5% of these industries at the 0.05 level.

illustrated by the two industries shown in Figure 12, where dark clusters correspond to positive regression residuals and light gray clusters to negative residuals.

[Figure 12]

Here the industry, "fabricated constructional and architectural metal products" (JSIC284), in Panel (a) exhibits extreme *positive* autorcorrelation, and the "newspaper" (JSIC191) industry in Panel (b) exhibits extreme *negative* autocorrelation. Notice in particular the differences between residual patterns in the enlarged Tokyo region for each panel. Here it is clear that residuals around Tokyo in Panel (a) are almost all positive, indicating a general degree of underestimation of employment in this area. Such positive correlation between nearby residuals is rather intuitive, and suggests that there must be strong linkages between firms in these clusters that are not fully captured by our limited set of explanatory variables.

But in Panel (b) we see quite a different picture, where a darker cluster (positive residual) is surrounded by lighter clusters (negative residuals) yielding strong negative correlations among these contiguous neighbors. Closer inspection reveals that there is a sharp peak of employment in the central cluster, which falls off in surrounding clusters. This central cluster, Chiyoda, contains 440 establishments with a total of 25,002 workers. In fact, this single cluster accounts for 8% of all establishments and 34% of all employment in the industry. In contrast, its seven contiguous neighbors account for less than 1% of all establishments and less than 3% of employment. So not only is concentration of establishments in this central cluster much denser, but also its establishments are on average much larger. As often occurs with such strong nonlinear data patterns, our linear regression model is underestimating central employment and overestimating peripheral employment. It is this spatial configuration of errors that produces negative autocorrelation.

More importantly, this type of pattern is not uncommon, and occurs in about 15% (22 of 154) industries at the cluster level. One might think that this is perhaps an artifact of our clustering procedure in stage one. But as mentioned above, such negative autocorrelation is even more common at the municipality level. Indeed, 18 of the above 22 industries also exhibit significant negative autocorrelation at the municipality level. So while these effects are occurring at somewhat different scales, they appear to be qualitatively similar. More generally, this suggests that at smaller spatial scales, the distribution of employment for many industries becomes more "spiky", leading to more frequent occurrences of locally negative autocorrelation.

## 4.2 Comparison of Cluster- and Prefecture-Level Regressions

Finally we consider regression results at prefecture level in relation to those at the cluster level. Notice first from Table 4 that there is even less spatial autocorrelation at the prefecture level. Indeed, rejection levels here are quite consistent with the size of each test, and suggest that independence of residuals is quite a reasonable assumption. So at first glance it might appear that our regression model exhibits more desirable properties at the prefecture level. But while the residuals are consistent with independence, the key questions of interest must be in terms of what the model *tells* us about industrial employment patterns. To anticipate the more detailed results below, recall from our introductory discussion that prefectures are simply too large to capture the kinds of industrial agglomeration behavior that we are after. Indeed, the average size of prefectures in terms of economic areas is about eight times that of the industrial clusters identified in stage one. This of course helps to explain why autocorrelation is not much of a problem at this level. But more important is the fact that at this scale, different types of relationships are being picked than those we are interested in.

To substantiate this claim, we now compare results for the cluster and prefectue level in more detail. To do so, it is important to note that while regression model (27) was useful for comparing spatial residual autocorrelation at each level of aggregation, it is far less useful for actual inter-industry analyses of agglomeration determinants. Hence, in order to develop a fair comparison between cluster and prefecture levels, it is appropriate to construct a parallel version of model (26) for spatial units defined by prefectures. This can be accomplished by simply replacing the cluster-sample units, *ik*, in (26) with prefecture-sample units, $ip$, as follows:

$$\begin{aligned}\ln E_{ip} &= (\beta_0+\beta_i)+(\beta_P+\beta_{iP})\ln A_{ip}^P+(\beta_C+\beta_{iC})\ln A_{ip}^C \\ &\quad +(\beta_T+\beta_{iT})\ln F_{ip}^T+(\beta_H+\beta_{iH})\ln F_{ip}^H+\varepsilon_{ip}\,, \qquad (34)\end{aligned}$$

where the definitions of explanatory variable used here are the same as those in model (27) above with $s=p$. Hence the key difference between these two formulations is that we can now compare estimates of industrial *interaction effects* (such as $\widehat{\beta}_{iP}$ for population access) between the cluster-level regression [model (26)] and prefecture-level regression [model (34)]. In Table 5 we present a summary comparision between these interaction effects.

[Table 5]

By way of illustration, the first row of this table lists the number of industries in each model for which interaction effects with population access (denoted by $\delta \times \ln A^P$) are significant (either $++$ or $--$). Here there are seen to be 33 industries with significant interaction effects at the cluster level (26) and 22 significant at the prefecture level (34). While these numbers would seem to be comparable in magnitude, the final column shows that only *two* industries are common to both lists. So it should be clear that quite different information is being provided by each model about the relative importance of population access for predicting employment levels across industries. Note that a similar pattern appears in all four rows, suggesting that this is not simply a coincidence. So the major questions remaining are exactly how and why these results are so different. As stated above, the key element appears to be the relative *sizes* of clusters and prefectures.

This is best seen by simply illustrating the results for a typical example in the first row. Here we have already seen that "bakeries and confectionery products" (JSIC127) is strongly consumer oriented, and hence exhibits a population-access effect at the cluster level that is way above average (++). Yet surprisingly this industry fails to exhibit even a weak interaction with population access at the prefecture level, suggesting that location sensitivity of "bakeries" to consumer markets is no more than average across industries. Why is this happening? Part of the story can be seen from the plot of "bakeries" employment by prefecture in Figure 13, where darker color indicates larger concentration of employment.

[Figure 13]

Here it is enough to consider the northern most prefecture, Hokkaido. As seen in the figure, a substantial portion (4%) of all such employment is in Hokkaido. Moreover, a comparison of clusters for "bakeries" in Figure 3(a) with population access at the municipality level in Figure 8 shows that in Hokkaido this industry is heavily concentrated around the large population centers of Sapporo and Asahikawa (where Sapporo alone is the seventh largest city in Japan). So this employment concentration in Hokkaido is completely consistent with a strong population-access effect. But at the prefecture level, population access looks dramatically different. As seen by a comparison of Figures 1 and 8, most of the municipalities in Hokkaido have very low levels of population access. So when $A^P$ is calculated at the prefecture level by averaging over municipalities, as in (30), it is not surprising that a very low level of population access is obtained. This leads to an extreme underestimate of "bakeries" employment in Hokkaido, based only on population access. It is this type of error that reduces the overall estimated sensitivity of "bakeries" to population access at the prefecture level.[40]

Similar examples can be illustrated for the opposite case in which interaction effects are significant at the prefecture level but not at the cluster level. But in all cases the stories are similar. In general, such differences arise from coarse aggregations of regional attributes at the prefecture level that tend to distort the desired determinants of local industrial employment.

[40] A similar story can be given for Niigata prefecture indicated in Figure 13.

# 5 Productivity of Clusters

In this section, we consider the relation between industrial agglomeration and productivity. Here, we use data from the *Census of Manufactures of Japan* to estimate total factor productivity (TFP) for individual establishments.[41] To obtain a sufficiently large number of observations for the comparison of productivities inside and outside clusters, our present strategy is to group the 154 three-digit industries into their corresponding 22 two-digit categories, and then to test each category for differences in average productivity between establishments that belong to clusters (at the three-digit level) and those that do not.

To formalize this approach, let $G$ denote the set of all 22 two-digit industry categories, and for convenience, refer to each member of $G$ as an industry *category*, $g \in G$, to distinguish them from three-digit *industries*, $i \in I$. Next, if $J$ denotes the set of all *establishments*, $j$, then $J_i$ and $J_g$ denote the sets of esblishments in each industry, $i$ , and category, $g$, respectively. In particular, if $I_g$ denotes the set of industries in category $g$, then these sets are related by $J_g = \cup_{i \in I_g} J_i$.

To measure TFP for individual establishments, $j \in J_g$, we adopt the standard approach of postulating Cobb-Douglas production technologies for each industry category, $g$, and treating the log of TFP as the residual term in a log linear regression of value added, $V$, on capital, $K$, and labor, $L$. More specifically, for each category, $g \in G$, we consider a regression of the form:

$$\ln V_j = \beta_0 + \beta_K \ln(K_j) + \beta_L \ln(L_j) + \phi_j, \;\; j \in J_g \tag{35}$$

where $\beta_0$, $\beta_K$ and $\beta_L$ are category-specific constants to be estimated, and where the residual, $\phi_j$ is taken to be $\ln(\mathrm{TFP}_j)$ for each establishment $j \in J_g$. Our baseline estimates, $\hat{\phi}$ , of these residuals, $\phi$ , are obtained by OLS. (Such estimates are later checked for robustness, as discussed below.)

Using these estimates, we next identify those establishments inside clusters and outside clusters for each category, $g \in G$. In particular, the set of establishments, $J_g$, is now partitioned into the set, $C_g$, of those establishments located in at least one cluster of some industry, $i \in I_g$, and the complementary set, $\overline{C}_g = J_g - C_g$, of establishments outside the clusters of these industries. We are then interested in whether the average productivity of establishments belonging to $C_g$ is significantly greater than that of establishments in $\overline{C}_g$.

To test this difference, we employ a (one-sided) Mann-Whitney $U$-Test, in a manner similar to the tests of climatic conditions in Section 3.2.2 above.[42] The appropriate $U$-statistic

[41] This data is provided by the Research Institute of Economy, Trade and Industry of Japan, and covers the subset of establishments included in Section 2 above to identify industrial clusters which actually produce outputs and employ at least four workers.

[42] Here it should be noted that while there are many tests of mean differences (such as the standard Welch

for this non-parametric test can be constructed as follows. First, for each establishment, $j \in \overline{C}_g$, let $n_j$ denote the number of establishments in $C_g$ with productivity levels *lower* than that of $j$.[43] Then the value of the $U$-statistic for category $g$ is obtained by simply adding these counts for all $j \in \overline{C}_g$, i.e.,

$$U_g = \sum\nolimits_{j \in \overline{C}_g} n_j \, . \tag{36}$$

Notice that by construction, relatively *small* values of $U_g$ indicate that productivity levels in $C_g$ tend to be *higher* than those in $\overline{C}_g$.

As with the Moran tests of spatial autocorrelation above, rather than using the standard asymptotic normal approximation for this $U$-test, we again choose to employ a permutation-test version.[44] In this case, the null hypothesis of "no productivity difference" is operationalized by hypothesizing that the observed sets of productivity levels in $C_g$ and $\overline{C}_g$ are not statistically distinguishable from random permutations (reassignments) of these levels between $C_g$ and $\overline{C}_g$. To test this hypothesis, we again generate a number of such random permutations, $t = 1, .., T$, and construct $U$-values, $U_g^{(t)}$, for each. The histogram of these values is then used to estimate the sampling distribution of $U_g$ under the null hypothesis. Using this distribution, one can then construct $p$-values for a one-sided test of this hypothesis. In particular, if the value of $U_g$ is significantly *small*, then we may conclude that the clustered establishments have significantly higher average productivity levels than non-clustered ones.

Columns (1) through (3) in Table 6 below summarize the results of these tests, based on 10,000 random permutations for each industry category, $g \in G$. As indicated by the $p$-values of Mann-Whitney test listed in column (2), the average productivity of clustered establishments is seen to be significantly higher than that of non-clustered establishments in all industrial categories[45] except for "petroleum and coal products" (JSIC21), as discussed further below. This is also seen in terms of relative median TFP levels in column (3), where values greater than one (except for the "petroleum and coal products" category) show that median productivity is higher inside clusters than outside clusters.[46]

test used in most software), the Mann-Whitney test depends only on ordinal properties of the data, and hence is completely insensitive to outliers. In the present data, there are a number of industrial categories, $g$, with relatively few non-cluster establishments, so that mean productivities in $\overline{C}_g$ are particularly sensitive to outliers.

[43] Note that since the Mann-Whitney test depends only on the ordering of data values, it makes no difference whether TFP or ln(TFP) values are used for productivity.

[44] This permutation test is generally much more reliable in cases where the size of one set (in this case $\overline{C}_g$) is much smaller than the other. In the present case, the share of establishments outside clusters, i.e., $|\overline{C}_g| / |N_g|$ is 0.26 at maximum, and 0.06 on average.

[45] While three are only weakly significant (below 0.10), most $p$-values are below 0.001.

[46] If the same tests are conducted for three-digit industries in $I$, the smaller sample sizes (especially for non-cluster establishments) yield results that are somewhat less clear. But even here, productivity differences are significant for 85 out of 154 industries (with 12 of those being weakly significant). Moreover, there is no

As for the exceptional case of “petroleum and coal products”, the higher median TFP value outside clusters appears to be due the unusual industrial composition of this category. Among the six three-digit industries in this category there is one industry, “pavement material” (JSIC215), which is by far the largest. Since this industry is very dispersed, it accounts for 75.6% of all non-cluster establishments, but only 42.7% of clustered establishments. So comparison between non-clustered and clustered establishments is here largely a comparison between this pavement-material industry and all others. Hence, in this case, it is appropriate to separate pavement-materials from the rest of the category, as is done in Table 6. Both subcategories are now seen to exhibit higher TFP in clusters. Part of the reason for this shift is that TFP for the pavment-materials industry is uniformly higher than the rest of the category. So its larger contribution to non-clustered establishments acts to reduce the TFP gap between clustered and non-clustered establishments for the category as a whole.

[Table 6]

Next observe that for all results based on OLS estimates of TFP, it is implicitly assumed that there are no correlations between the explanatory variables (factors of production) and the residuals (productivity). But as is well known, there may often be interdependencies (endogeneities) between productivity and factor usage. Thus, to assess the robustness of these results, we have applied two alternative methods to estimate TFP, as proposed by Olley and Pakes (OP) [20] and by Levinsohn and Petrin (LP) [13] . Both methods utilize lagged instrumental variables (“investment” in the case of OP and “material inputs” in LP) to account for possible endogeneities between productivity and factor inputs.[47] To construct these lagged variables, we use panel data from 1995 to 2000 to estimate TFP of establishments in 2000.[48,49]

Columns (4) through (6) and (7) through (9) in Table 6 summarize results based, respectively, on the OP and LP estimates of TFP for establishments. Note first that all results for both OP and LP are essentially identical. More importantly, these results are also comparable with those of OLS. The only two exceptions are “petroleum and coal products”

instance where the average productivity of non-cluster establishments is significantly higher than that of cluster establishments.

[47] In addition, OP takes into account possible selection bias resulting from certain establishments exiting during the sample period.

[48] Value added and capital inputs are converted to the real terms as of 2000. Investment in year $t$ used under OP is defined as $I_t = K_t / K_{t-1}$, instead of more standard, $I_t = K_t - K_{t-1}$, since more than half of the observations would need to be dropped in the latter case to ensure positivity of investments (as pointed out by Levinsohn and Petrin [13]). The material input used as an instrument for endogeneity correction in LP is defined to be the sum of costs for fuels, electricity costs and other miscellaneous intermediate inputs. The estimation of TFP was conducted using Stata routines, *opreg* and *levpet* for OP and LP, respectively.

[49] We chose not use data more recent than 2000, since such data no longer includes capital inputs for small establishments.

(JSIC21), mentioned for OLS above, and "pulp, paper and paper products" (JSIC18). These two industry categories exhibit rather striking differences, and require further discussion.

For "petroleum and coal products", the explanation in the case of OLS above still applies, and the results appear to be fairly comparable with OLS, once the pavement-material industry is separated from this category. In particular the median TFP ratio for this industry is comparable among all three methods. However, for the rest of the caterory, these ratios are distinctively higher under OP and LP (2.98 and 2.77, respectively) than OLS (1.10). Here factor inputs and productivity are highly correlated (between 0.6 and 0.8 for both factors), suggesting that endogeneities are the problem for OLS in this case.

As for "pulp, paper and paper products" (JSIC18), the difference in TFP levels between inside and outside clusters is insignificant under OP and LP, while it is weakly significant under OLS. But since endogeneities are again evident in this case (with all correlations above 0.5), it would appear that the test results under OP and LP are more reliable. Indeed, there appear to be good reasons why TFP is higher for non-clustered establishments in this category. Since pulp and paper products are bulk/weight loosing in nature, they tend to be attracted to material-input sites (i.e., woodlands). On the other hand, their markets are concentrated in population centers, where the majority of clusters in this category are found. Hence, assuming that establishments in these clusters are more market oriented, it is not surprising that productivity is higher outside clusters in this case.[50]

So aside from these exceptions, our results above do indeed suggest that establishments in clusters tend to exhibit higher levels of productivity than those not in clusters.

# 6 Concluding Remarks

The main objective of this paper has been to develop a spatial approach to identifying possible local determinants of industrial agglomeration. A two-stage approach was developed in which spatially explicit cluster patterns were identified for each industry, and then used to construct a regression framework for identifying and comparing possible determinants of agglomeration across industries. In this final section, we touch on several questions raised in the text.

A key question relates to how one might improve the model in order to capture more of the unexplained variation in cluster employment levels. Aside from the need to obtain additional data on potentially relevant regional variables at the municipality level, there are a number of other possibilities that are worth exploring. The first relates to the inclusion of *industry attributes* themselves. In the present model we distinguish industries only in terms of fixed effects. But there are a host of industry attributes that are potentially relevant here.

[50] It should also be noted that land inputs are less costly outside population centers, which in turn increases value added for the non-cluster establishments.

For example one might include measures of weight/bulk increasing versus decreasing industries, as discussed for specific industries in the illustrations above. A variety of other potentially relevant industry measures appear in Rosenthal and Strange [21, 22]. But while such measures might well provide interesting interpretations in themselves, it is doubtful that they can account for much unexplained variation in the present framework, at least in terms of adjusted $R^2$.[51]

However, there are a number of other interesting possibilities that are worth exploring. One relates to our implicit assumption that the relevant "accessibility" scale is the same for all industries, as characterized by the common choice of distance decay value, $\tau = 0.10$. But while population access may be relevant for both "bakeries"and "steel products", it is doubtful that such distance sensitivities are the same for both. Hence by estimating such decay parameters, $\tau_i$, for each industry, $i \in I$, (say in terms of the distance distributions of their output deliveries), it might be possible to capture more of this unexplained variation between industries.[52]

Another possible extension focuses on changes in cluster patterns over time, which can often be substantial. For example, changes in the number of clusters for three-digit manufacturing industries in Japan between 1981 and 2001 ranged from -63.4% to 114.3%.[53] As suggested by Duranton and Overman [6], locations of new entrants are often influenced by those of the existing establishments. In addition, they may be influenced by the spatial pattern of input-output transactions accessibilities for the industry. Such transactions accessibilities may change in response to changes in production technologies for the industry. Even when production technologies remain the same, transaction accessibilities may change due to relocations of transactions partners. More generally, it is of interest to identify those factors shaping the evolution of spatial cluster patterns for each industry.

One final direction for extending the present framework has implications for both the substantive and statistical properties of our two-stage approach. Recall from the "newspaper" example in Figure 12(b) that set of contiguous clusters in the enlarged Tokyo area exhibits a characteristic of "central peak" structure, in which a central cluster of larger establishments is surrounded by clusters of smaller establishments. In addition, our negative autocorrelation results suggest that this pattern is not uncommon among sets of contiguous clusters. Even more generally, there appears to be a strong relationship

[51] The reason here is simply that unadjusted $R^2$ is necessarily maximized by fixed effects, which effectively add a new parameter for each industry. Moreover, the present large sample size, $n = 12,350$, already far exceeds the number of parameters, $k = 153 + 4 = 157$, in our fixed-effects model. Hence the adjusted $R^2$ penalty factor, $(n-1)/(n-k-1) = 1.029$, is so close to one that it leaves little room for improvement.

[52] For the two-digit industrial categories in Japan, relevant establishment-level shipment data is available from *the Net Freight Flow Census* by the Ministry of Land, Infrastructure, Transport and Tourism (MLIT), where information is drawn from stratified random samples of actual shipments. It may be possible to use this data to approximate shipment-distance distributions.

[53] Here, the industrial classification is based on that in 1981, and includes 146 industries (with non-spurious clusters) rather than 154 in 2001. See Mori and Smith [18] for the detail.

between cluster size and average size of establishments across industries.[54] This finding is consistent with observations made by others that larger firms in manufacturing industries tend to concentrate together (see, e.g., Holmes and Stevens [8]). But our present results reveal a more explicit spatial pattern in which such concentrations of large firms are typically surrounded by significant clusters of smaller firms. These results suggest that such sets of contiguous clusters are best interpreted as single units. In a previous paper (Mori and Smith[17]) we proposed that all maximal sets of contiguous clusters be combined as single *agglomerations*. This not only provides a more meaningful interpretation of such clusters, but may indeed help to account for some of the unexplained variation in cluster employment levels. Finally, as an added bonus, this grouping of clusters should help to mitigate the remaining spatial autocorrelation (both positive and negative) that was observed in the cluster-level residuals above.

[54] In fact, with respect to the industrial clusters we have identified, the simple correlation between the logs of cluster employment size and the average size of cluster establishments across industries is 0.64 (with 25% of industries exhibiting correlations above 0.80).

# References

[1] Abe, Takeshi, Hitomi, Kazuya, Konishi, Yoko, Tomita, Hideaki, Utino, Taisuke. 2012. "Kougyou Toukei Chosa no panel-ka no tameno converter (1993-2009)." RIETI Policy Discussion Paper Series 12-P-007.

[2] Ciccone, Antonio and Robert E. Hall. 1996. "Productivity and the density of economic activity." *American Economic Review* 86(1): 54-70.

[3] Combes, Pierre-Philippe, Overman, Henry G. 2004. "The spatial distribution of economic activities in the European Union" in: J. V. Henderson and J. F. Thisse (eds.) *Handbook of Regional and Urban Economics*, Vol.4, Ch.64: 2845-2909, Elsevier.

[4] ——, Gilles Duranton, Laurent Gobillon, Diego Puga, and Sebastien Roux. 2012. "The productivity advantages of large cities: Distinguishing agglomeration from firm selection." *Econometrica*, forthcoming.

[5] Duranton, Gilles, Overman, Henry G. 2005. "Testing for localization using micro-geographic data." *Review of Economic Studies* 72(4): 1077-1106.

[6] ——, ——. 2008. "Exploring the detailed location patterns of UK manufacturing industries using microgeographic data." *Journal of Regional Science* 48(1): 313-343.

[7] Ellison, Glenn, Glaeser, Edward L. .1997. "Geographic concentration in US manufacturing industries: A dartboard approach." *Journal of Political Economy* 105(5): 889-927.

[8] Holmes, Thomas J., Stevens, John J. 2004. "Geographic concentration and establishment scale." *Review of Economics and Statistics* 84(4): 682-690.

[9] Hsu, W. 2009. "Central place theory and the city size distribution." *Economic Journal* 122: 903-932.

[10] Ikeda, Kiyohiro, Akamatsu, Takashi, Kono, Tatsuhito. 2012. "Spatial Period Doubling Agglomeration of a core-periphery model with a system of cities." *Journal of Economic Dynamics and Control* 36(5): 754-778.

[11] Kanemoto, Yoshitsugu, Tokuoka, Kazuyuki. 2002. "The proposal for the standard definition of the metropolitan area in Japan." *Journal of Applied Regional Science* 7: 1-15.

[12] Kullback, Solomon, Leibler, Richard A. 1951. "On Information and Sufficiency." *Annals of Mathematical Statistics* 22(1): 79-86.

[13] Levinsohn, James, Petrin, Amil. 2003. "Estimating production functions using inputs to control for unobservables." *The Review of Economic Studies* 70: 317-341.

[14] Marcon, Eric, Puech, Florence. 2010. "Measures of the geographic concentration of industries: improving distance-based methods." *Journal of Economic Geography* 10(5): 745-762.

[15] Melo, Patricia C., Graham, Daniel J., Noland, Robert B. 2009. “A meta-analysis of estimates of urban agglomeration economies.” *Regional Science and Urban Economics* 39(3): 332-342.

[16] Mori, Tomoya, Nishikimi, Koji, Smith, Tony E. 2005. “A divergence statistic for industrial localization.” *Review of Economics and Statistics* 87(4): 635-651.

[17] ——, Smith, Tony E. 2009. “A probabilistic modeling approach to the detection of industrial agglomerations”Metholodogical framework.” Discussion paper No.682, Institute of Economic Research, Kyoto University.

[18] ——, ——. 2009. “A reconsideration of the NAS Rule from an industrial agglomeration perspective.” *Brookings-Wharton Papers on Urban Affairs*. Whashington, D.C.: Brookings Institution Press: 175-214.

[19] ——, ——. 2011. “A probabilistic modeling approach to the detection of industrial agglomerations: Metholodogical framework.” Discussion paper No.777, Institute of Economic Research, Kyoto University.

[20] Olley, Steven, Pakes, Ariel. 1996. “The dynamics of productivity in the telecommunications equipment industry.” *Econometrica* 64(6): 1263-1297.

[21] Rosenthal, Stuart S., Strange, William C. 2001. “The determinants of agglomeration.” *Journal of Urban Economics* 50(2): 191-229.

[22] ——, ——. 2004. “Evidence on the nature and sources of agglomeration economies” in : Henderson, J. Vernon and Thisse, Jacques-Francǫis (eds.). *Handbook of Regional and Urban Economics*, Vol.4 (Amsterdam: North-Holland), Ch.49, Elsevier.

[23] Tabuchi, Takatoshi, Thisse, Jacques-Francǫis. 2011. “A new economic geography model of central places.” *Journal of Urban Economics* 69(2): 240-252.

| | $\ln E$ | $\ln A^P$ | $\ln F^T$ | $\ln F^J$ | $\ln F^H$ | $\ln F^U$ | $\ln A^C$ | $\ln A^{Pcp}$ | $\ln A^{Sun}$ | $\ln A^{Temp}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $\ln E$ | 1.0000 | 0.4030 | 0.2061 | -0.3067 | -0.0956 | 0.3293 | 0.2542 | 0.1362 | 0.0854 | 0.1715 |
| $\ln A^P$ | | 1.0000 | 0.0832 | **-0.8365** | -0.2639 | **0.8309** | 0.2886 | 0.2938 | 0.2166 | 0.4065 |
| $\ln F^T$ | | | 1.0000 | -0.0878 | 0.0213 | 0.0901 | -0.0088 | 0.0646 | 0.0487 | 0.0156 |
| $\ln F^J$ | | | | 1.0000 | 0.1753 | **-0.9233** | -0.2136 | -0.3097 | -0.1830 | -0.3323 |
| $\ln F^H$ | | | | | 1.0000 | -0.3325 | -0.0678 | -0.0249 | -0.0002 | -0.1992 |
| $\ln F^U$ | | | | | | 1.0000 | 0.1783 | 0.2023 | 0.2595 | 0.3914 |
| $\ln A^C$ | | | | | | | 1.0000 | 0.0223 | -0.1444 | 0.4292 |
| $\ln A^{Pcp}$ | | | | | | | | 1.0000 | 0.0569 | -0.1362 |
| $\ln A^{Sun}$ | | | | | | | | | 1.0000 | 0.2037 |
| $\ln A^{Temp}$ | | | | | | | | | | 1.0000 |

Table 1: Correlations among regional attributes and sizes of clusters under $\tau = 0.10$

$R^2 = 0.3855$, Adjusted $R^2 = 0.3447$

Number of observations : 12350

| Variables | # Parameters | $F$ values | $p$ values |
|---|---|---|---|
| $\delta$ | 153 | 8.8434 | $< .0001$ |
| $\ln A^P$ | 1 | 909.5989 | $< .0001$ |
| $\ln F^T$ | 1 | 305.1541 | $< .0001$ |
| $\ln F^H$ | 1 | 0.0159 | 0.8996 |
| $\ln A^C$ | 1 | 447.6792 | $< .0001$ |
| $\delta \times \ln A^P$ | 153 | 2.9873 | $< .0001$ |
| $\delta \times \ln F^T$ | 153 | 2.2103 | $< .0001$ |
| $\delta \times \ln F^H$ | 153 | 1.9255 | $< .0001$ |
| $\delta \times \ln A^C$ | 153 | 1.7443 | $< .0001$ |

Table 2: Summary of cluster-level regression with $\tau = 0.10$

| Industry name (JSIC) | $D$ index | $\ln A^P$ | $\ln F^T$ | $\ln F^H$ | $\ln A^C$ |
|---|---|---|---|---|---|
| **Average of all industries** | — | $0.4577^{**}$<br>(30.16) | $0.7070^{**}$<br>(17.47) | $-0.0262$<br>($-0.13$) | $0.0952^{**}$<br>(0.0045) |
| **12. Food products** | | | | | |
| Seafood products (122) | 1.6464 | $0.2674^{**}_{--}$<br>(3.98) | $-0.2958_{--}$<br>($-1.13$) | $1.4585$<br>(1.38) | $0.2274^{**}_{++}$<br>(4.66) |
| Canned fruit & vegetable products (123) | 1.0782 | $0.5100^{**}$<br>(3.87) | $-0.1409$<br>($-0.31$) | $4.6383^{*}_{+}$<br>(2.29) | $0.0781^{*}$<br>(1.91) |
| Seasonings (124) | 1.0851 | $0.8176^{**}_{+}$<br>(4.79) | $0.0450$<br>(0.11) | $0.3346$<br>(0.15) | $0.0644$<br>(1.5) |
| Sugar processing (125) | 3.2437 | $0.5599^{**}$<br>(3.44) | $0.6958$<br>(1.22) | $-6.8223^{*}_{-}$<br>($-2.24$) | $0.0843$<br>(1.47) |
| Flour & grain mill products (126) | 1.5142 | $1.2274^{**}_{++}$<br>(6.32) | $0.0497$<br>(0.1) | $0.1464$<br>(0.06) | $-0.0070_{-}$<br>($-0.14$) |
| Bakery & confectionery products (127) | 1.0314 | $1.0639^{**}_{++}$<br>(10.96) | $-0.1163_{--}$<br>($-0.37$) | $-0.3041$<br>($-0.21$) | $0.0818^{**}$<br>(2.43) |
| Animal & vegetable oils & fats (128) | 2.8061 | $0.6748^{**}$<br>(4.51) | $-0.4251_{--}$<br>($-1.16$) | $-0.8643$<br>($-0.39$) | $0.1730^{**}$<br>(3.71) |
| Misc. food products (129) | 0.7289 | $0.9862^{**}_{++}$<br>(9.96) | $-0.3355_{--}$<br>($-1.33$) | $1.9728$<br>(1.47) | $0.0575^{*}$<br>(2.03) |
| **13. Beverages &and feed** | | | | | |
| Soft drinks (131) | 1.9478 | $0.8245^{**}_{+}$<br>(5.48) | $1.0525$<br>(1.55) | $-0.9963$<br>($-0.52$) | $-0.0152_{-}$<br>($-0.3$) |
| Tea & coffee (133) | 2.5988 | $0.4539^{**}$<br>(3.36) | $-0.1402_{-}$<br>($-0.41$) | $-2.2464$<br>($-1.02$) | $0.0759$<br>(1.35) |
| **14. Textile mill products** | | | | | |
| Silk reeling plants (141) | 4.7109 | $0.2714$<br>(1.4) | $-0.3226_{--}$<br>($-1.31$) | $-0.5084$<br>($-0.16$) | $-0.0079$<br>($-0.14$) |
| Spinning mills (142) | 3.5945 | $0.1980$<br>(1.39) | $0.2044_{-}$<br>(0.81) | $1.0094$<br>(0.41) | $0.2007^{**}_{+}$<br>(3.81) |
| Knit fabrics mills (145) | 3.2003 | $0.7052^{**}_{+}$<br>(5.84) | $1.0402^{**}$<br>(4.68) | $-0.6081$<br>($-0.28$) | $0.0207$<br>(0.48) |
| Dyed & finished textiles (146) | 3.3886 | $0.6014^{**}$<br>(4.4) | $1.3333^{**}_{++}$<br>(7.8) | $0.9776$<br>(0.54) | $0.0296$<br>(0.81) |
| Rope & netting (147) | 3.2230 | $0.1932$<br>(1.27) | $0.6360$<br>(1.18) | $-0.8011$<br>($-0.37$) | $0.2980^{**}_{++}$<br>(4.4) |
| **15. Apparel & fabrics products** | | | | | |
| Textile outer garments & shirts (151) | 1.6112 | $0.2591^{**}_{-}$<br>(3.13) | $0.1803_{--}$<br>(1.03) | $6.9999^{**}_{++}$<br>(4.88) | $0.1415^{**}$<br>(4.63) |
| Knitted garments & shirts (152) | 2.4498 | $0.2098^{*}_{--}$<br>(2.24) | $0.1842_{-}$<br>(0.72) | $2.5142$<br>(1.5) | $0.0526$<br>(1.38) |
| Underwear (153) | 2.1776 | $0.0838_{--}$<br>(0.83) | $0.3853$<br>(1.53) | $3.4678^{*}$<br>(1.92) | $0.1321^{**}$<br>(3.05) |
| Japanese style apparel (155) | 3.0356 | $0.8007^{**}_{++}$<br>(6.71) | $0.5144^{**}$<br>(2.94) | $0.0069$<br>(0.0) | $0.0630$<br>(1.38) |
| **16. lumber & wood products** | | | | | |
| Misc. wood products (169) | 1.8541 | $0.4712^{**}$<br>(4.06) | $0.9601^{**}$<br>(2.52) | $-1.6069$<br>($-0.92$) | $-0.0146_{--}$<br>($-0.43$) |
| **17. Funiture & fixtures** | | | | | |
| Furniture (171) | 1.5232 | $0.6952^{**}$<br>(5.07) | $1.5403^{**}$<br>(2.85) | $4.1513^{**}_{+}$<br>(2.42) | $0.0795^{*}$<br>(2.2) |
| Furniture for religious purposes (172) | 2.7187 | $0.5990^{**}$<br>(3.87) | $0.9271^{**}$<br>(3.23) | $-2.3112$<br>($-0.97$) | $0.0087_{-}$<br>(0.2) |
| Sliding doors & screens (173) | 0.7600 | $0.9023^{**}_{++}$<br>(9.14) | $0.0846_{--}$<br>(0.35) | $3.4740^{**}_{+}$<br>(2.33) | $0.0599^{*}$<br>(2.07) |
| Misc. furniture & fixtures (179) | 2.2103 | $0.5240^{**}$<br>(3.02) | $1.0596$<br>(1.39) | $-4.6141^{*}_{-}$<br>($-2.1$) | $0.0196$<br>(0.39) |
| **18. Pulp & paper products** | | | | | |
| Pulp (181) | 3.7495 | $0.2193$<br>(0.95) | $0.5943^{*}$<br>(2.2) | $3.6382$<br>(1.27) | $-0.0471_{-}$<br>($-0.76$) |
| Paper (182) | 3.2083 | $0.6668^{**}$<br>(4.15) | $0.8878^{**}$<br>(2.83) | $7.1977^{**}_{++}$<br>(2.86) | $0.1289^{*}$<br>(1.93) |
| Paper containers (185) | 1.8217 | $0.7313^{**}_{+}$<br>(5.46) | $0.2219$<br>(0.45) | $3.8618^{*}_{+}$<br>(2.19) | $0.1074^{**}$<br>(2.87) |

Table 3: Coefficient estimates for selected industries under $\tau = 0.10$

| Industry | $D$ index | $\ln A^P$ | $\ln F^T$ | $\ln F^H$ | $\ln A^C$ |
|---|---|---|---|---|---|
| **19. Publishing & allied industries** | | | | | |
| Newspaper industries (191) | 3.0838 | $0.9697^{**}_{++}$<br>(6.8) | $-0.2786_{--}$<br>($-0.79$) | 2.8796<br>(1.5) | 0.0424<br>(0.98) |
| Publishing industries (192) | 4.2724 | $0.8423^{**}$<br>(3.86) | $0.6381^{*}$<br>(1.67) | $-2.6726$<br>($-1.24$) | $-0.0153_{-}$<br>($-0.34$) |
| Printing (193) | 1.8879 | $0.9015^{**}_{++}$<br>(8.06) | $-0.2350_{--}$<br>($-0.67$) | 2.0926<br>(1.59) | $0.0349_{-}$<br>(1.3) |
| Bookbinding & printed matter (195) | 3.6825 | $0.8222^{**}$<br>(3.45) | $-0.2098_{-}$<br>($-0.52$) | $-3.9392^{*}$<br>($-1.77$) | 0.0141<br>(0.32) |
| Pringing-related services (199) | 4.4519 | $1.0860^{**}_{+}$<br>(3.95) | $-0.0585$<br>($-0.11$) | $-1.3907$<br>($-0.41$) | $-0.0219$<br>($-0.29$) |
| **20. Chemical & allied products** | | | | | |
| Industrial inorganic chemicals (202) | 2.0096 | $0.4749^{**}$<br>(3.48) | $1.6844^{**}_{+}$<br>(4.28) | $-1.7109$<br>($-0.98$) | 0.0425<br>(0.93) |
| Industrial organic chemicals (203) | 2.1858 | $0.5702^{**}$<br>(2.87) | $2.3160^{**}_{++}$<br>(4.03) | 1.9311<br>(0.72) | $0.2038^{**}$<br>(3.37) |
| Chemical fibres (204) | 3.3159 | $0.5155^{**}$<br>(2.43) | 0.1932<br>(0.55) | $9.5643^{**}_{++}$<br>(2.78) | $0.1726^{*}$<br>(2.16) |
| Oil & fat products (205) | 2.2388 | $1.0451^{**}_{++}$<br>(5.01) | 1.0040<br>(1.4) | $-4.4943$<br>($-1.56$) | 0.1003<br>(1.06) |
| **21. Petroleum & coal products** | | | | | |
| Petroleum refining (211) | 3.8309 | 0.1534<br>(0.77) | $2.0670^{**}_{++}$<br>(4.39) | $-5.1775^{*}_{-}$<br>($-1.97$) | $0.4317^{**}_{++}$<br>(4.39) |
| **22. Plastic products** | | | | | |
| Compounding plastic materials (225) | 2.0558 | $1.0477^{**}_{+}$<br>(4.03) | $-0.1809$<br>($-0.3$) | $-0.9833$<br>($-0.23$) | 0.0493<br>(0.88) |
| **24. Leather products & fur skins** | | | | | |
| Leather tanning & finishing (241) | 5.2908 | 0.2933<br>(1.37) | $1.3962^{**}$<br>(3.57) | $-3.8433$<br>($-1.19$) | $-0.0823_{--}$<br>($-1.22$) |
| Cut stock & findings for shoes (243) | 5.2811 | $0.1256_{--}$<br>(1.12) | $0.8720^{**}$<br>(6.14) | $-1.1032$<br>($-0.56$) | $0.0815^{*}$<br>(1.76) |
| Leather footwear (244) | 4.2347 | $0.2036_{-}$<br>(1.59) | $0.5964^{**}$<br>(2.88) | $4.3619^{*}$<br>(1.71) | 0.0752<br>(1.15) |
| **25. Ceramic** | | | | | |
| Pottery & related products (254) | 3.8360 | $0.8272^{**}_{+}$<br>(5.11) | $1.0054^{**}$<br>(3.38) | 2.4085<br>(1.08) | 0.0533<br>(1.2) |
| Clay refractories (255) | 3.8428 | 0.2871<br>(1.44) | $2.0683^{**}_{++}$<br>(4.87) | $-6.7383^{**}_{--}$<br>($-2.8$) | $0.0977^{*}$<br>(1.83) |
| Carbon & graphite products (256) | 3.2076 | $-0.1951_{--}$<br>($-1.07$) | $0.7602^{*}$<br>(1.94) | $-0.6141$<br>($-0.24$) | $0.1882^{**}$<br>(3.71) |
| Abrasive products (257) | 2.7666 | $0.5096^{**}$<br>(3.49) | $-0.1589$<br>($-0.28$) | $-4.0174^{*}$<br>($-1.92$) | $0.0029_{-}$<br>(0.07) |
| **26. Iron & steel products** | | | | | |
| Iron industries, with blast furnaces (261) | 4.9551 | $-1.5425^{**}_{--}$<br>($-2.74$) | $-1.3124_{-}$<br>($-1.41$) | $-12.2945^{**}_{-}$<br>($-2.39$) | $0.8512^{**}_{++}$<br>(5.74) |
| Iron smelting, without blast furnaces (262) | 4.2366 | 0.1963<br>(0.84) | $-0.5631_{--}$<br>($-1.53$) | $7.7245^{*}$<br>(1.78) | $0.4986^{**}_{++}$<br>(4.29) |
| Steel, with rolling facilities (263) | 3.0002 | $1.0619^{**}_{++}$<br>(4.98) | $0.7948^{**}$<br>(2.61) | $5.7294^{*}_{+}$<br>(2.06) | $0.1029^{*}$<br>(2.25) |
| **27. Non-ferrous metals & products** | | | | | |
| Electric wire & cable (274) | 2.1996 | $0.9740^{**}_{++}$<br>(4.92) | $1.0198^{*}$<br>(2.03) | 2.4224<br>(0.96) | 0.0123<br>(0.25) |
| Non-ferrous metal machine parts (275) | 2.1235 | $-0.0469_{--}$<br>($-0.26$) | $1.8947^{**}_{+}$<br>(3.73) | $-0.9107$<br>($-0.36$) | $0.1214^{**}$<br>(2.89) |
| Misc. non-ferrous metal products (279) | 2.8747 | $0.0227_{--}$<br>(0.14) | $0.9688^{*}$<br>(2.12) | $-1.8696$<br>($-0.85$) | $0.1504^{**}$<br>(3.63) |
| **28. Fabricated metal products** | | | | | |
| Fabricated constructional metal products (284) | 1.1529 | $0.8955^{**}_{++}$<br>(8.43) | 0.0771<br>(0.22) | $4.2234^{**}_{++}$<br>(3.11) | $0.1271^{**}$<br>(3.77) |
| Metal machine parts (285) | 2.1985 | 0.1566<br>(0.82) | $2.0197^{**}_{+}$<br>(3.6) | 1.0673<br>(0.39) | $0.1360^{**}$<br>(3.49) |

Table 3: Coefficient estimates for selected industries under $\tau = 0.10$ (continued)

| Industry | $D$ index | $\ln A^P$ | $\ln F^T$ | $\ln F^H$ | $\ln A^C$ |
|---|---|---|---|---|---|
| **29. General machinery** | | | | | |
| Boilers, engines & turbines (291) | 3.0425 | $0.6411^{**}$<br>(3.72) | $1.7289^{**}_{++}$<br>(4.47) | $-7.7648^{**}_{--}$<br>($-3.94$) | $-0.0100_{-}$<br>($-0.19$) |
| Office & household machines (298) | 1.8137 | $0.4938^{**}$<br>(2.8) | $1.8651^{**}_{+}$<br>(3.16) | 0.0948<br>(0.04) | $0.0684^{*}$<br>(1.75) |
| **30. Electrical machineries & equipments** | | | | | |
| Communication equipments (304) | 1.7555 | $0.4553^{**}$<br>(2.86) | $1.8487^{**}_{++}$<br>(4.3) | $-0.6690$<br>($-0.35$) | $0.1056^{*}$<br>(2.25) |
| Electronic parts & devices (308) | 1.2092 | $0.5877^{**}$<br>(4.64) | 0.3556<br>(1.45) | $4.3088^{**}_{++}$<br>(2.9) | $0.0867^{**}$<br>(2.56) |
| Misc. electrical machineries (309) | 2.1762 | $0.0381_{-}$<br>(0.2) | 0.4346<br>(1.13) | $5.1789^{*}$<br>(1.85) | $0.1801^{**}$<br>(3.92) |
| **31. Transportation equipment** | | | | | |
| Shipbuilding & repairing (314) | 2.1348 | $0.4658^{**}$<br>(4.7) | $0.0594_{--}$<br>(0.29) | 1.3998<br>(1.17) | $0.1961^{**}$<br>(3.78) |
| Misc. transportation equipment (319) | 2.3215 | $0.7989^{**}_{+}$<br>(4.97) | $1.8287^{**}_{++}$<br>(4.52) | $-6.3046^{**}_{--}$<br>($-3.42$) | $0.0899^{*}$<br>(2.11) |
| **32. Precision instruments & machineries** | | | | | |
| Physical & chemical instruments (324) | 3.5657 | $-0.2882_{--}$<br>($-1.04$) | $1.3633^{*}$<br>(1.77) | $-8.9366^{**}_{-}$<br>($-2.39$) | $0.1185^{*}$<br>(2.19) |
| Optical instruments & lenses (325) | 2.4246 | $0.4541^{**}$<br>(2.36) | 0.5900<br>(1.2) | 1.5676<br>(0.62) | $0.0013_{-}$<br>(0.03) |
| Watches & clocks (327) | 2.6747 | $0.0578_{--}$<br>(0.37) | $0.8854^{*}$<br>(2.32) | $-0.1951$<br>($-0.09$) | $-0.0128_{-}$<br>($-0.28$) |
| **34. Misc. manufacturing industries** | | | | | |
| Precious metal products (341) | 3.4635 | $0.3087^{*}$<br>(1.82) | $-0.0074$<br>($-0.02$) | $-6.3223^{**}_{--}$<br>($-3.02$) | 0.0415<br>(0.86) |
| Musical instruments (342) | 2.9529 | 0.1179<br>(0.48) | $2.7968^{**}_{++}$<br>(3.52) | $-4.9542$<br>($-1.45$) | $0.1465^{**}$<br>(2.47) |
| Lacquer ware (346) | 4.0946 | $0.1278_{-}$<br>(0.85) | 0.6477<br>(1.21) | 2.7026<br>(1.39) | 0.0331<br>(0.78) |
| Information recording materials (34C) | 4.1611 | 0.0382<br>(0.16) | 0.2836<br>(0.47) | $-10.0208^{**}_{--}$<br>($-3.57$) | 0.0158<br>(0.27) |

Table 3: Coefficient estimates for selected industries under $\tau = 0.10$ (continued)

| | Municipalities | | Clusters | | Prefectures | |
|---|---|---|---|---|---|---|
| $\tau$ | .01 level | .05 level | .01 level | .05 level | .01 level | .05 level |
| 0.05 | 0.6494 | 0.7597 | 0.0584 | 0.1234 | 0.0130 | 0.0455 |
| 0.10 | 0.6234 | 0.7273 | 0.0584 | 0.1494 | 0.0065 | 0.0390 |
| 0.20 | 0.5909 | 0.6753 | 0.0195 | 0.1364 | 0.0065 | 0.0195 |

Table 4: Share of industries with significant Moran's **I** among regression residuals

| Attributes | Cluster | Prefecture | Intersection |
|---|---|---|---|
| $\delta \times \ln A^P$ | 33 | 22 | 2 |
| $\delta \times \ln F^L$ | 28 | 34 | 13 |
| $\delta \times \ln F^H$ | 20 | 10 | 2 |
| $\delta \times \ln A^C$ | 18 | 21 | 6 |

Table 5: Number of industries with significant coefficients (at .05 level) under $\tau = 0.10$

| Industry | OLS |  |  | Olley-Pakes |  |  | Levinsohn-Petrin |  |  |
|---|---|---|---|---|---|---|---|---|---|
|  | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
|  | # Obs. | $p$-value | TFP ratio | # Obs. | $p$-value | TFP ratio | # Obs. | $p$-value | TFP ratio |
| 12. Food | 17,818 | 0.0002 | 1.1081 | 16,915 | 0.0018 | 1.0911 | 17,593 | 0.0012 | 1.1332 |
| 13. Beverages and feed | 2,219 | 0.0000 | 1.2778 | 2,099 | 0.0000 | 1.1935 | 2,191 | 0.0000 | 1.2687 |
| 14. Textile mill products | 3,592 | 0.0002 | 1.2310 | 3,501 | 0.0000 | 1.2244 | 3,570 | 0.0009 | 1.2153 |
| 15. Apparel and other finished products | 8,422 | 0.0000 | 1.1658 | 8,080 | 0.0000 | 1.1480 | 8,354 | 0.0003 | 1.1185 |
| 16. Lumber and wood products | 4,149 | 0.0000 | 1.1341 | 3,972 | 0.0000 | 1.1189 | 4,112 | 0.0000 | 1.0889 |
| 17. Furniture | 3,224 | 0.0002 | 1.2074 | 3,094 | 0.0002 | 1.1643 | 3,201 | 0.0005 | 1.2603 |
| 18. Pulp, paper and paper products | 4,529 | 0.0523 | 1.0286 | 4,341 | 0.2694 | 1.0094 | 4,479 | 0.3031 | 1.0132 |
| 19. Publishing, printing and allied industries | 9,219 | 0.0000 | 1.2824 | 8,780 | 0.0000 | 1.3432 | 9,148 | 0.0000 | 1.4414 |
| 20. Chemical and allied products | 3,811 | 0.0961 | 1.0851 | 3,659 | 0.0096 | 1.1875 | 3,760 | 0.0030 | 1.2149 |
| 21. Petroleum and coal products | 347 | 0.1693 | 0.9380 | 333 | 0.0002 | 1.1478 | 345 | 0.0002 | 1.2975 |
| Without pavement material (JSIC215) | 157 | 0.0622 | 1.1017 | 149 | 0.0013 | 2.7689 | 156 | 0.0020 | 2.9777 |
| Pavement material (JSIC215) | 190 | 0.0187 | 1.2621 | 184 | 0.0087 | 1.1665 | 189 | 0.0468 | 1.1882 |
| 22. Plastic product | 8,263 | 0.0003 | 1.1016 | 7,795 | 0.0103 | 1.0691 | 8,141 | 0.0281 | 1.0397 |
| 23. Rubber products | 1,708 | 0.0000 | 1.2838 | 1,616 | 0.0000 | 1.3317 | 1,689 | 0.0040 | 1.2559 |
| 24. Leather tanning, leather products and fur skins | 938 | 0.0003 | 1.5144 | 897 | 0.0000 | 1.4871 | 930 | 0.0002 | 1.4699 |
| 25. Ceramic, stone and clay products | 8,527 | 0.0995 | 1.0064 | 8,189 | 0.0088 | 1.0369 | 8,435 | 0.0011 | 1.0603 |
| 26. Iron and steel | 2,917 | 0.0039 | 1.1375 | 2,788 | 0.0151 | 1.1002 | 2,876 | 0.0159 | 1.0264 |
| 27. Nonferrous metals and products | 1,692 | 0.0058 | 1.1402 | 1,617 | 0.0098 | 1.0854 | 1,670 | 0.0232 | 1.0518 |
| 28. Fabricated metal products | 15,253 | 0.0000 | 1.2245 | 14,512 | 0.0000 | 1.1873 | 15,074 | 0.0000 | 1.1952 |
| 29. General machinery | 16,101 | 0.0000 | 1.1463 | 15,305 | 0.0000 | 1.1623 | 15,876 | 0.0000 | 1.1815 |
| 30. Household electric appliances | 14,971 | 0.0000 | 1.2282 | 14,161 | 0.0000 | 1.3059 | 14,755 | 0.0000 | 1.2074 |
| 31. Transportation equipment | 6,512 | 0.0002 | 1.1723 | 6,193 | 0.0020 | 1.1293 | 6,412 | 0.0063 | 1.1508 |
| 32. Precision instruments and machinery | 2,374 | 0.0008 | 1.2061 | 2,252 | 0.0007 | 1.2218 | 2,349 | 0.0086 | 1.1369 |
| 34. Miscellaneous manufacturing industries | 3,694 | 0.0052 | 1.1476 | 3,496 | 0.0062 | 1.1736 | 3,651 | 0.0163 | 1.2117 |

Table 6: Comparison of TFP of establishments inside and outside clusters

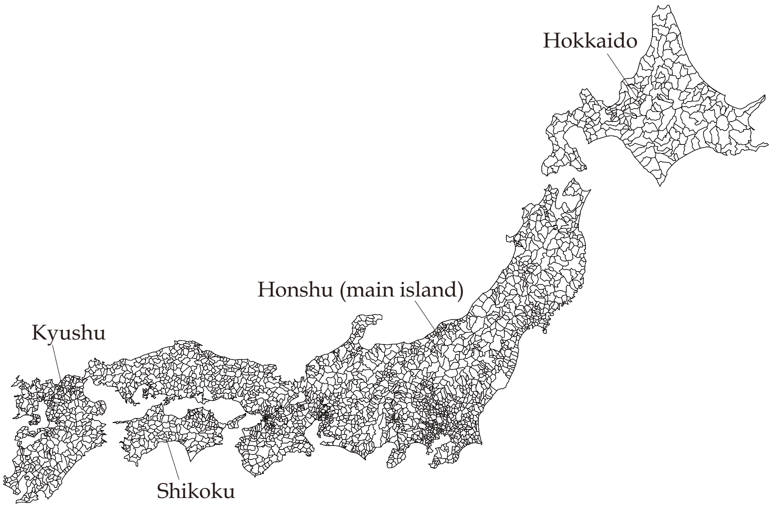


Figure 1: Municipalities of Japan

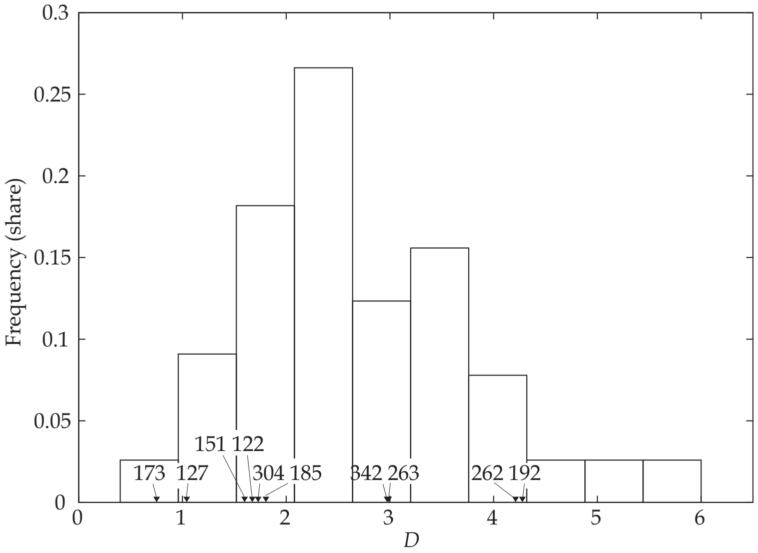
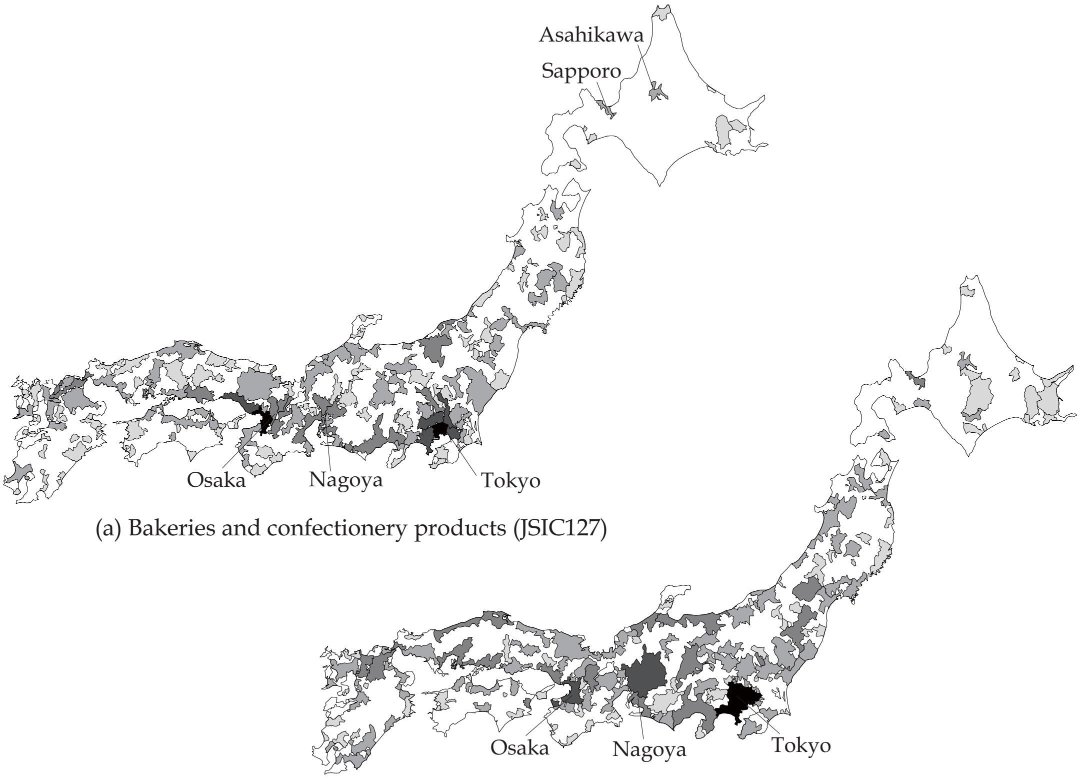


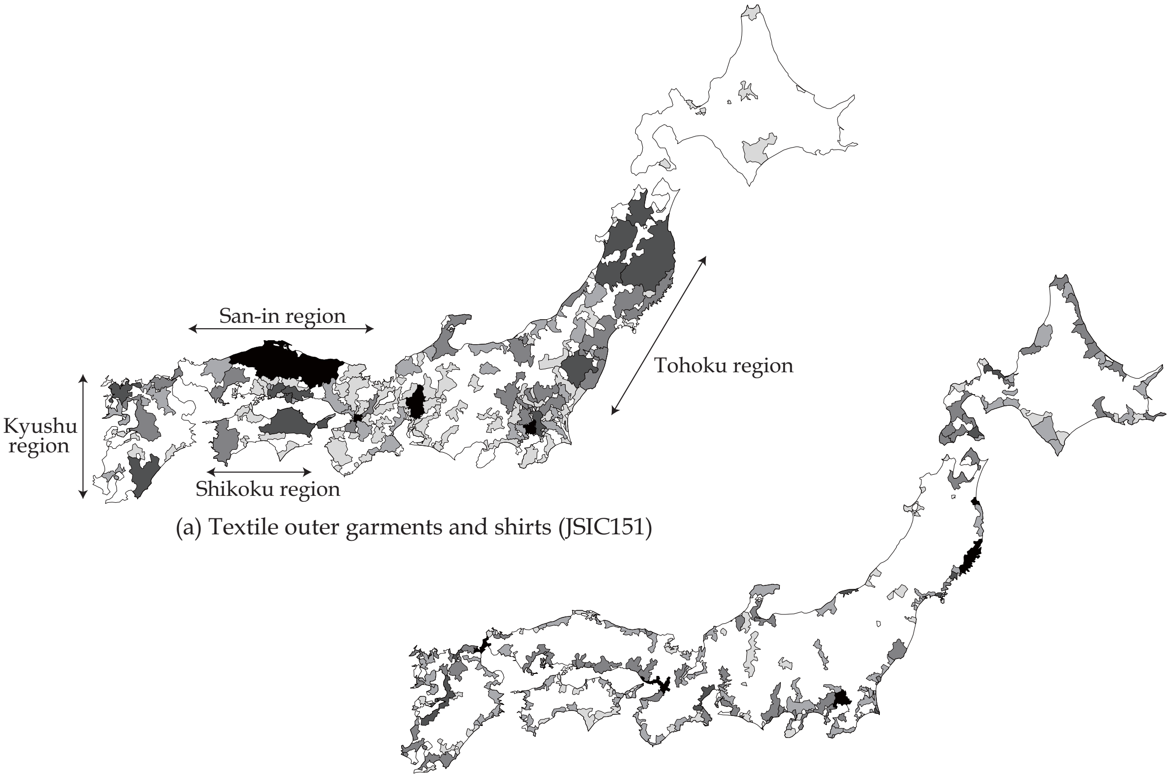
Figure 2: Frequency distribution of *D* values



(a) Bakeries and confectionery products (JSIC127)

(b) Sliding doors and screens (JSIC173)

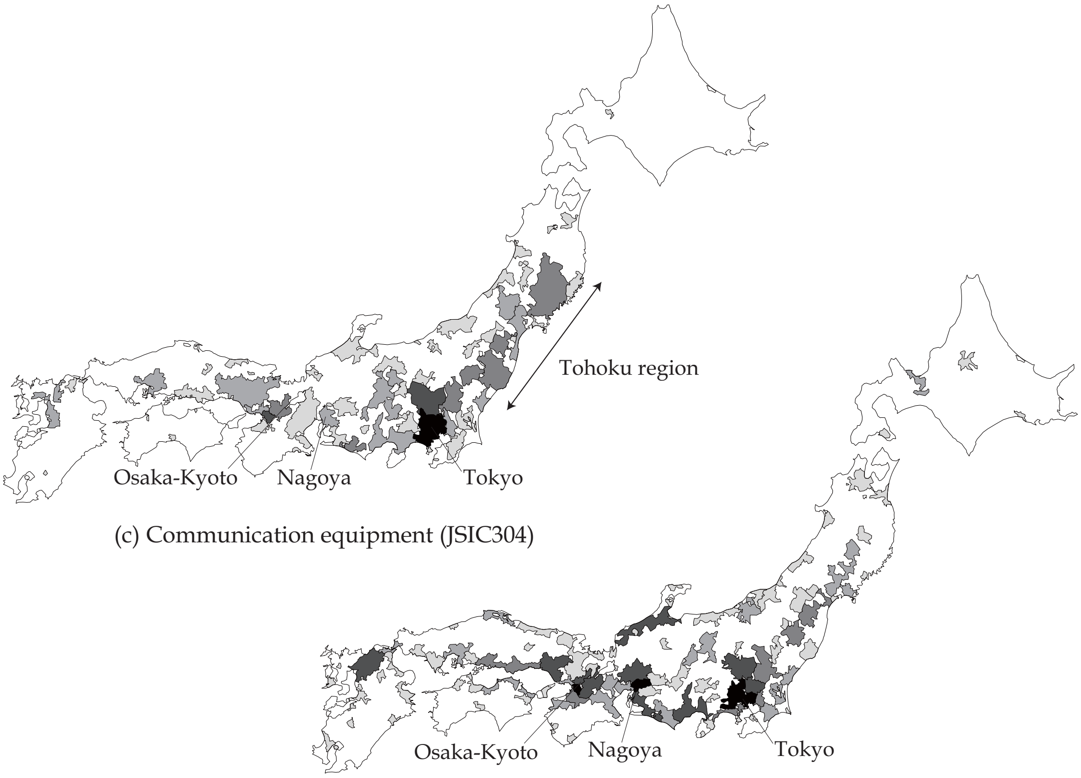
Figure 3: Ubiquitous industries



(a) Textile outer garments and shirts (JSIC151)

(b) Seafood products (JSIC122)

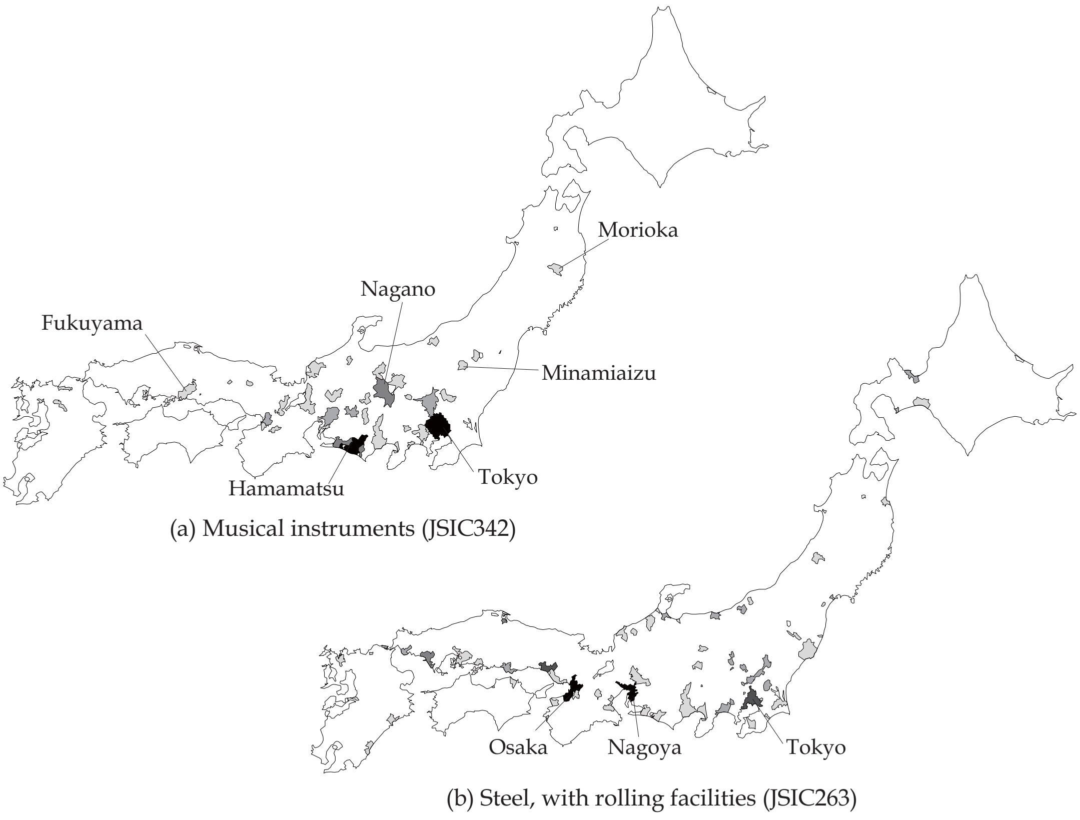
Figure 4: Industries with low levels of *D*



(c) Communication equipment (JSIC304)

(d) Paper containers (JSIC185)

Figure 4: Industries with low levels of *D* (continued)



(a) Musical instruments (JSIC342)

(b) Steel, with rolling facilities (JSIC263)

Figure 5: Industries with intermediate levels of *D*

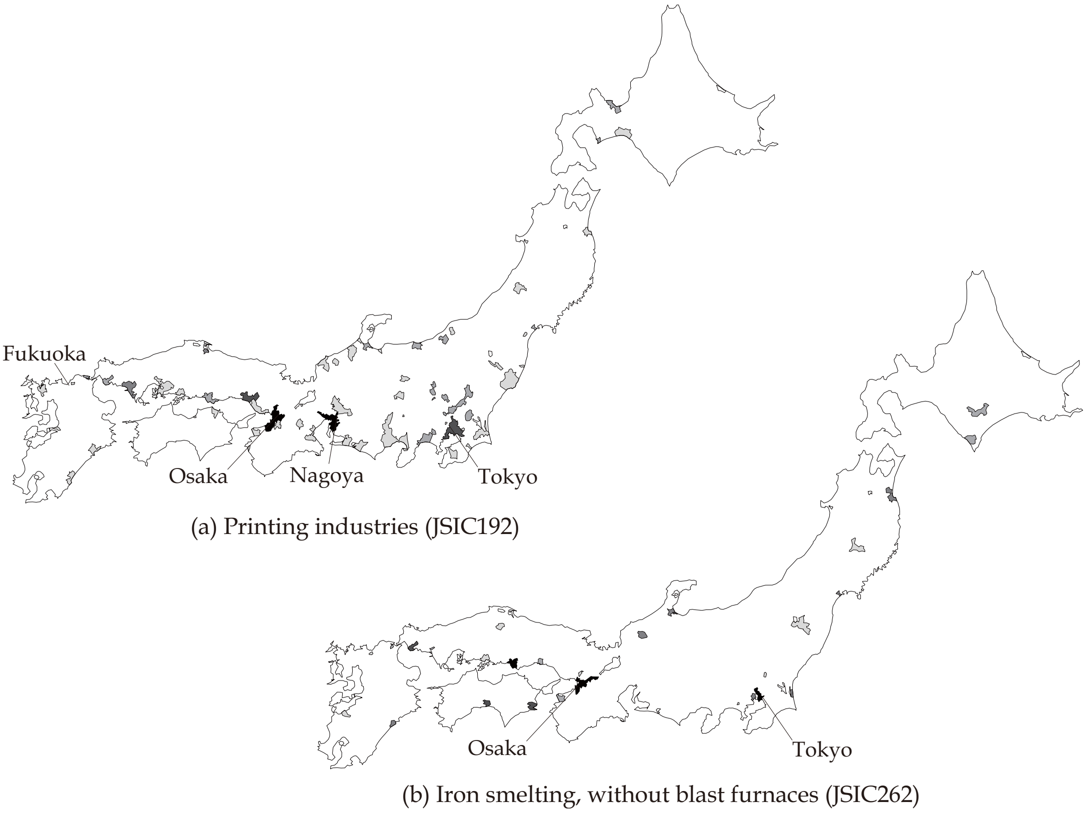

Fukuoka
Osaka
Nagoya
Tokyo
(a) Printing industries (JSIC192)
Osaka
Tokyo
(b) Iron smelting, without blast furnaces (JSIC262)

Figure 6: Industries with high levels of $D$

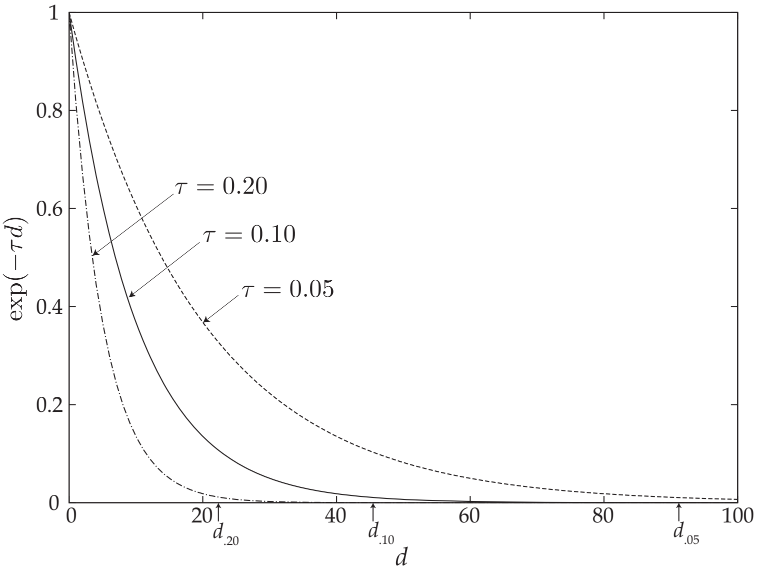


Figure 7: Distance decay functions

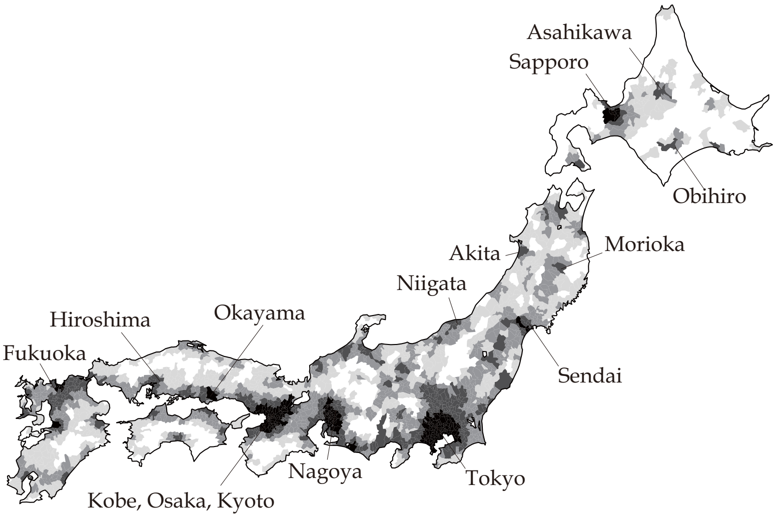
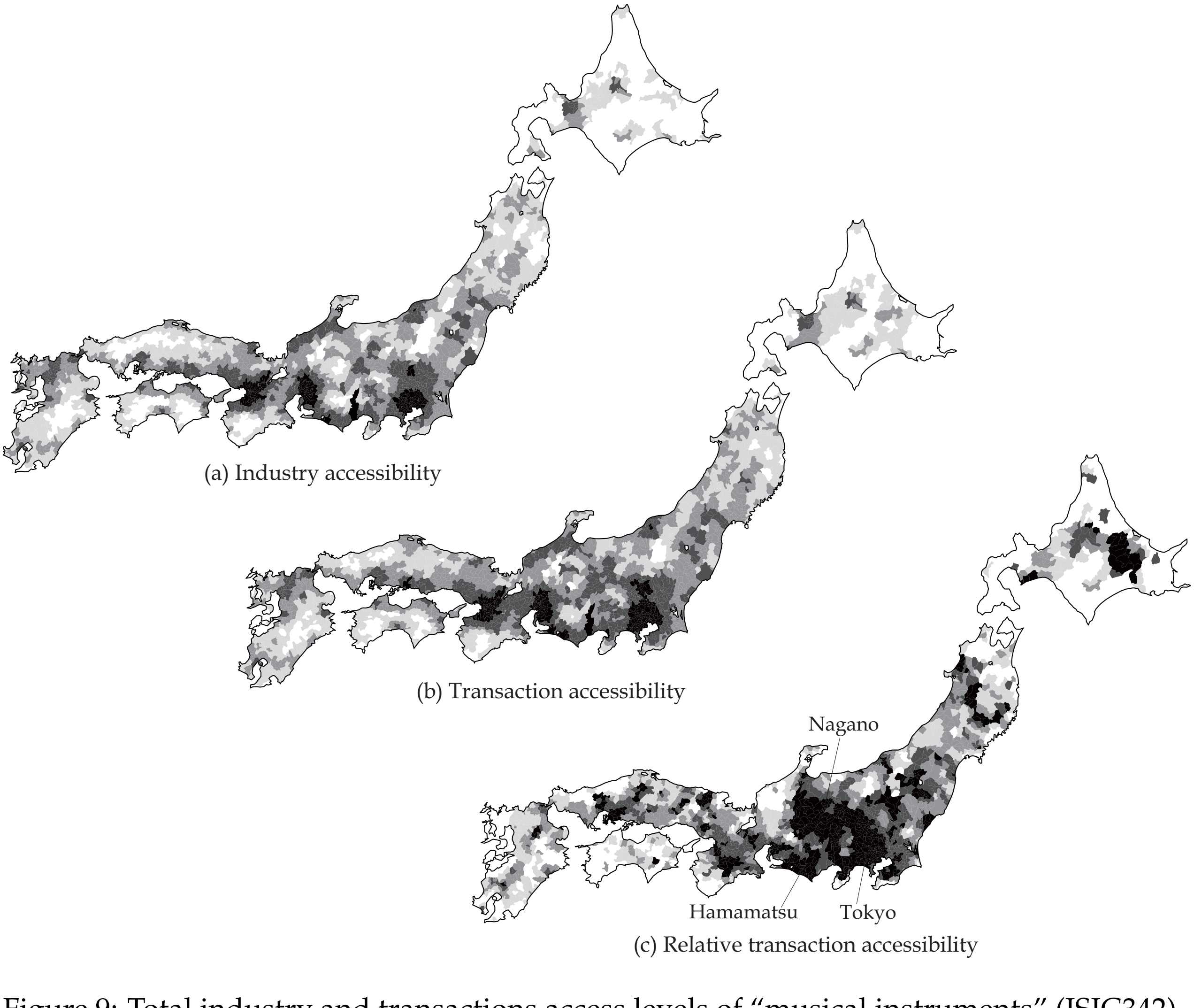


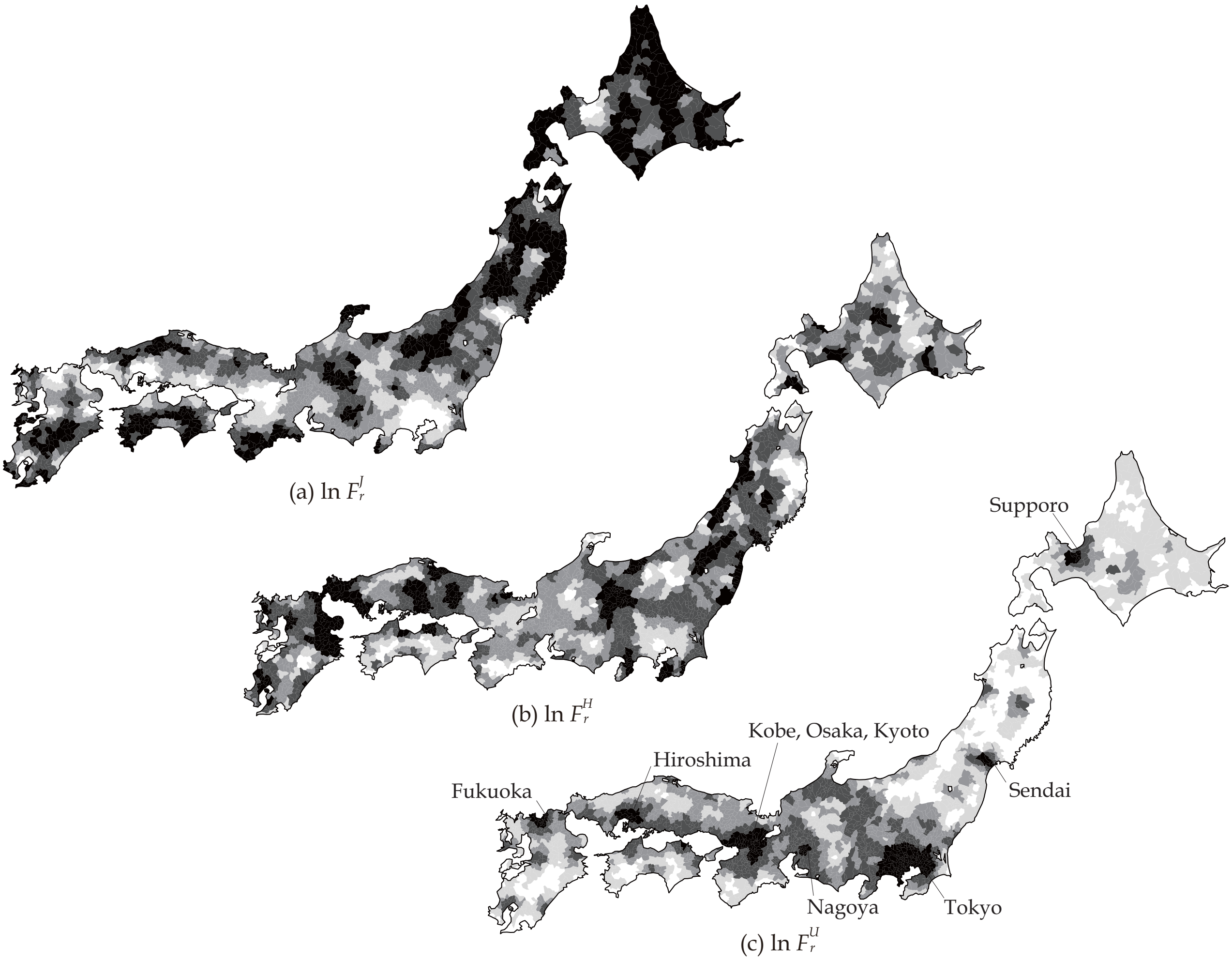
Figure 8: Population access of municipalities with $\tau = 0.10$


(a) Industry accessibility

(b) Transaction accessibility

(c) Relative transaction accessibility

Figure 9: Total industry and transactions access levels of "musical instruments" (JSIC342) with $\tau = 0.10$



(a) ln $F_r^I$

(b) ln $F_r^H$

(c) ln $F_r^U$

Figure 10: Labor accessibilities with $\tau = 0.10$

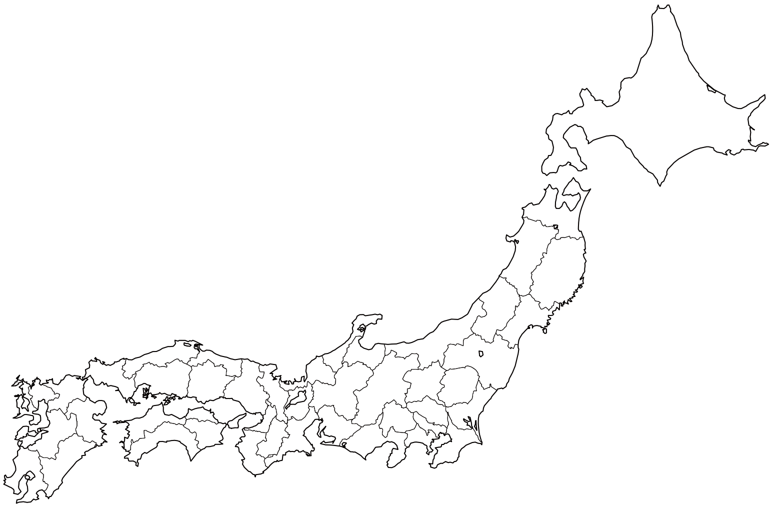
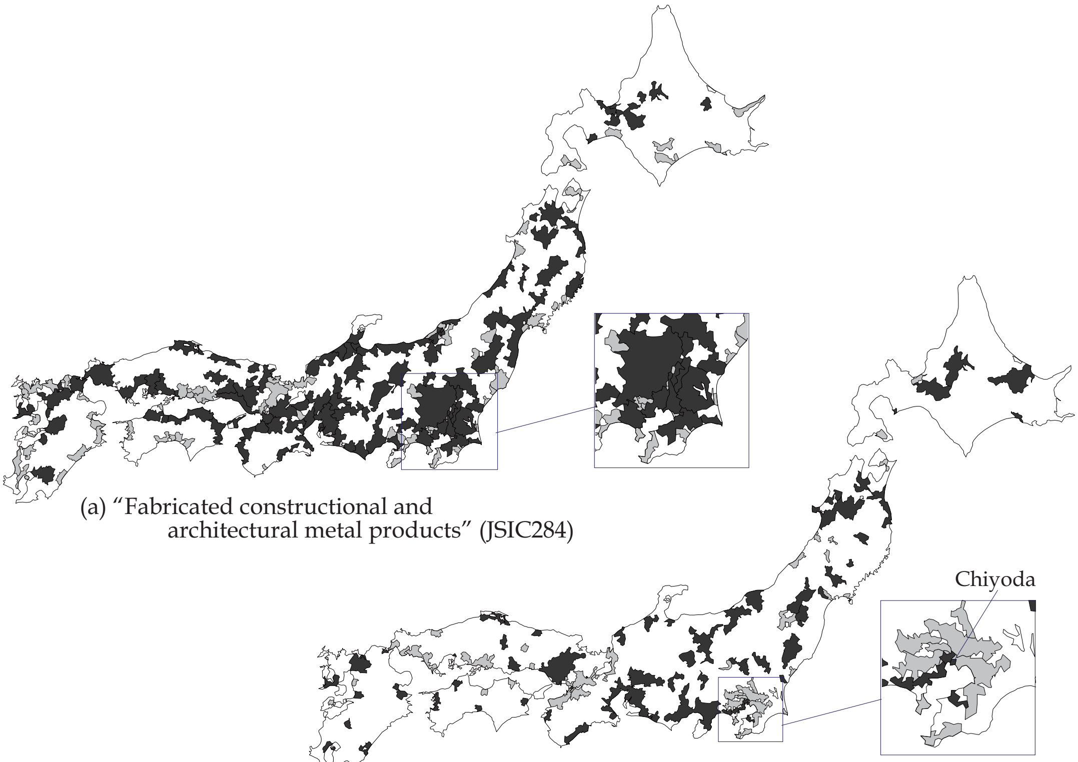


Figure 11: Prefectures of Japan



(a) “Fabricated constructional and architectural metal products” (JSIC284)

(b) “Newspaper industries” (JSIC191)

Figure 12: Spatial autocorrelations for residuals under the cluster-level regression with $\tau = 0.10$

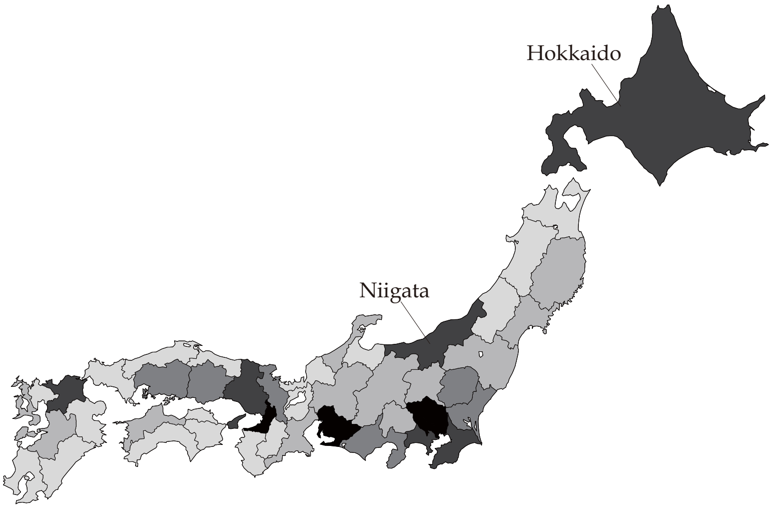


Figure 13: Employment of “bakeries” industry (JSIC127) at the prefecture level