Mori, Tomoya; Smith, Tony

Conference Paper

On the spatial scale of industrial agglomerations

55th Congress of the European Regional Science Association: "World Renaissance: Changing roles for people and places", 25-28 August 2015, Lisbon, Portugal

Provided in Cooperation with:
European Regional Science Association (ERSA)

Suggested Citation: Mori, Tomoya; Smith, Tony (2015) : On the spatial scale of industrial agglomerations, 55th Congress of the European Regional Science Association: "World Renaissance: Changing roles for people and places", 25-28 August 2015, Lisbon, Portugal, European Regional Science Association (ERSA), Louvain-la-Neuve

This Version is available at:
https://hdl.handle.net/10419/124562

On the Spatial Scale of Industrial Agglomerations

Tomoya Mori and Tony E. Smith[*,†]

January 2015

Abstract

Standard approaches to studying industrial agglomeration have been in terms of scalar measures of agglomeration within each industry. But such measures often fail to distinguish spatial scales of agglomeration. In a previous paper, Mori and Smith [37] proposed a pair of quantitative measures for distinguishing both the scale and degree of industrial agglomeration based on an explicit method for detecting spatial clusters. The first, designated as the *global extent* of industrial clusters, measures the spatial spread of these clusters in terms of the areal size of their *essential containment*, defined to be the (convex-solid) region containing the most significant subset of these clusters. The second, designated as the *local density* of industrial clusters, measures the spatial extent of individual clusters *within their essential containment* in terms of the areal share of that containment occupied by clusters. The present paper applies this pair of measures to manufacturing industries in Japan, and the results obtained are systematically compared to those of the most prominent scalar measures currently in use. Finally, these measures are shown to support certain predictions of new economic geography models concerning the relationship between shipment distances and spatial scales of agglomeration for individual industries.

JEL Classifications : C49, L60, R12, R14

Keywords : Industrial Agglomeration, Cluster analysis, Spatial patterns of agglomeration, Spatial scale of agglomeration, Shipment distances, New economic geography

[*]Mori: Institute of Economic Research, Kyoto University and Research Institute of Economy, Trade and Industry (RIETI) of Japan. Email: mori@kier.kyoto-u.ac.jp. Smith: Department of Electrical and Systems Engineering, University of Pennsylvania. Email: tesmith@seas.upenn.edu.

[†]We thank Gilles Duranton, Ryo Nakajima and two anonymous referees for their constructive comments, and Kohei Takeda for his excellent research assistance. This research is conducted as part of the project, *the formation of economic regions and its mechanism: theory and evidence*, undertaken at the Research Institute of Economy, Trade and Industry, and has been partially supported by the Grant in Aid for Research (Nos. 25285074, 25380294, 2624503) of MEXT of Japan.

1 Introduction

The standard approach to studying industrial agglomeration has focused on the overall degree of agglomeration for each industry, and typically measures the discrepancy between industry-specific regional distributions of establishments (or employment) and a given hypothetical reference distribution representing "complete dispersion" in terms of some scalar index (e.g., Ellison and Glaeser [10], Duranton and Overman [9], Brülhart and Traeger [4], Mori et al. [34]).[1] But even if industries are judged to be similar with respect to these indices, their spatial patterns of agglomeration may be quite different. In particular, these aggregate measures often fail to distinguish between industries that exhibit substantially different spatial scales of agglomeration. One example (discussed further below) is the contrast between (i) an industry which is found in only one region of a country, but is ubiquitous throughout this region, and (ii) an industry which is found in every region of the country, but is concentrated in a particular district of each region. Such industry pairs are often judged to be similarly concentrated (or dispersed) by aggregate indices.

In a previous paper, Mori and Smith [37] proposed a new pair of quantitative measures for distinguishing both the scale and degree of industrial agglomeration based on an explicit method for detecting spatial clusters. The first, designated as the *global extent* (*GE*) of an industry's clusters within a given country, measures the spatial spread of these clusters in terms of the areal share of their *essential containment* within that country, namely, the smallest "convex-solid" region containing all "significant" clusters (to be formally defined in Section 3.1). Smaller values of *GE* for industries imply that their major clusters are essentially confined to smaller regions of the country, while larger values indicate that these clusters are more dispersed. In contrast to this global measure of spread, the second measure, designated as *local density* (*LD*), focuses solely on clusters *within the essential containment* for that industry, and measures their local density in terms of areal share within this containment. Larger (smaller) values of *LD* for an industry thus imply that its clusters tend to be more (less) spread out within this critical region.

[1] Examples of such reference distributions are (1) the regional distribution of all-industry employment, used by Ellison and Glaeser [10], (2) the regional distribution all-industry establishments, used by Duranton and Overman [9], and (3) the regional distribution of economic area used by Mori et al. [34]. Brülhart and Traeger [4] adopted (1) and (3).

These specific measures are largely inspired by theoretical results from the "new economic geography" (NEG).[2] In this framework, the spatial structure of agglomeration and dispersion is determined through the interactions between global and local dispersion forces, depending on a host of factors including plant-level increasing returns.[3] The basic intuition can be illustrated by considering the spatial effects of transport costs in simple "core-periphery" models of industrial location (e.g., Tabuchi [46]; Murata and Thisse [38]). At very high levels of transport costs, the dispersion of consumers between the "core" and "periphery" regions leads to a corresponding dispersion of manufacturing, where manufacturing firms spread over spatially dispersed local markets in order to minimize their transport costs to final markets. But as transport costs decrease and distance to consumers becomes less critical, manufacturing tends to concentrate (in the core region). Finally, at even lower levels of transport costs, the commuting costs dominate (together with congestion effects) in the core region and can induce a second phase of manufacturing dispersion (popularly referred to as "re-dispersion" or "revival"). Alternatively, the responses of the manufacturing industry to the different levels of transport costs may be interpreted as the responses of different (manufacturing) industries to a given transport cost level. In particular, industries that are very sensitive (resp., insensitive) to transport costs tend to disperse over space when transport costs are very high (resp., low), while those with intermediate sensitivity to transport costs tend to agglomerate.

Although these two dispersion patterns often appear to be exactly the same in the context of the two-region model (i.e., a symmetric distribution of manufacturing between the two regions), the associated shipment patterns are quite different. In the first phase of dispersion with high transport costs, manufacturing firms disperse to be closer to their markets, so that average shipment distances tend to be relatively small. But in the second phase of dispersion with lower transport costs, firms are able to access more extended markets, so that average shipment distances tend to be larger. Moreover, in NEG models involving more general location spaces (e.g., Krugman [27]; Fujita and Mori [15]), these two phases also differ with respect to their spatial scales of dispersion. While the first-phase dispersion of manufacturing to serve peripheral markets is quite global in nature (as in core-periphery models), the second phase of dispersion

[2] See, e.g., Fujita et al. [14] and Combes et al. [7] for an overview of the literature.

[3] See Fujita and Mori [17] for a survey.

tends to involve either expansions of existing core clusters or filling in between existing clusters,[4] both of which are more local in nature.[5]

Such theoretical findings raise important questions as to whether this diversity of agglomeration patterns as well as the corresponding shipment patterns can in fact be identified empirically. Hence the specific measures proposed here are designed to quantify pattern differences both in terms of their global and local properties. While the details of these measures require a more formal definition and construction of agglomeration patterns, the basic ideas can be illustrated by a preview of the types of patterns we have identified for Japanese manufacturing industries in 2001.

First, there are industries which clearly exhibit strong spatial concentration, such as the "compounding plastic materials, including reclaimed plastics" industry. The agglomeration pattern derived for this industry is shown in Figure 12(b), where the areas marked by the enclosed red regions denote industrial clusters.[6] Notice that the main industrial concentration lies clearly in the Industrial Belt along the Pacific coast extending westward from Tokyo to Fukuoka. Moreover, the individual clusters of establishments within this belt are seen to be densely packed from end to end. We describe this type of agglomeration pattern as "globally confined" and "locally dense" (here with respect to the Industrial Belt). In particular, this pattern is reminiscent of the type of "second-phase" dispersion of manufacturing identified in the NEG models described above. But even globally dispersed industries often form small clusters at local scales. For example, the agglomeration pattern for the "manufactured ice" industry shown in Figure 9(b) is spread throughout the country, but exhibits a large number of local clusters. Such patterns, which we describe as "globally dispersed" and "locally sparse", are closer in spirit to the "first-phase" dispersion of manufacturing in the NEG models above.

With respect to the aggregate measures of agglomeration above, it is not clear which of these two industries would be judged as "more agglomerated", since the first industry exhibits agglomeration at the global scale but dispersion at the local scale, while the opposite is true for the second industry. In fact, this

[4] An explicit example of this type of filling-in process is the formation of "industrial belts", as discussed in Mori [33].

[5] See also Behrens [3] for a related discussion on the spatial extent of agglomeration in NEG models.

[6] See Section 4.3 below for a more detailed discussion of these figures.

may not even be an appropriate comparison. Aside from these extremes, there are a variety of other patterns that can be identified, as discussed more fully in Sections 3 and 4 below. As will be shown in Section 5, industries generally exhibit wide variations in *GE* and *LD*. With respect to scalar measures of agglomeration, it will also be shown and that those of Ellison and Glaeser [10] and Mori et al. [34] are roughly equally well represented by these two components, while the scalar measure of Duranton and Overman [9] is more strongly associated with *GE*.

Finally, by using Japanese micro data for the shipments of individual establishments, we show that shipment distances of individual industries are *negatively* correlated with *GE* (i.e., global dispersion) and are *positively* correlated with *LD* (i.e., local dispersion) of clusters. To our knowledge, this provides the first empirical evidence for the theoretical predictions of NEG models described above.

To develop these ideas, the paper is organized as follows. In Section 2 we develop the formal framework for analysis, and briefly sketch the cluster identification procedure developed in Mori and Smith [37]. This is followed in Section 3 with a development of our summary measures for analyzing and classifying the agglomeration patterns obtained. These methods are then applied in Section 4 to (*i*) identify establishment clusters for each manufacturing industry in Japan, and to (*ii*) identify the spatial scales of these agglomeration patterns. In Section 5, the relationship between existing scalar indices of agglomeration and our pair of measures, GE and LD, are discussed. Finally, the relationship between shipment distances and spatial scales of agglomeration is investigated in Section 6. The paper concludes with brief discussions of related research in Section 7.

2 Identification of Industrial Clusters

This section provides an overview of the cluster detection framework developed by Mori and Smith [37].[7] We begin with a set, R, of *basic regions* (municipalities), r, within which each industry can locate. An *industrial cluster* is then taken

[7]All the relevant C++ and Python programs for the cluster detection introduced in this section can be downloaded from the web: http://www.mori.kier.kyoto-u.ac.jp/data/cluster_detection.html. Also, all the input and output data as well as map data for the application to Japanese manufacturing industries in Section 4 can be downloaded from the same site.

roughly to be a spatially coherent subset of regions within which the density of industrial establishments is unusually high. Since the explicit construction of such clusters will have consequences for the summary measures to be developed, it is appropriate to outline this construction more explicitly. The present notion of "spatial coherence" is taken to include the requirement that such regions be contiguous, and as close to one another as possible – where "closeness" is defined with respect to the relevant underlying regional network, where the nodes of this network are represented by the set R of basic regions, and the links are taken to represent pairs of regional "neighbors" in terms of the underlying regional network. By using travel distances between regional centers along this network, we define *shortest paths* between each pair of regions, r_i and r_j, to be sequences of intermediate regions, $(r_i, r_1, \ldots, r_k, r_j)$ reflecting minimum travel distances with respect to the road network. Our key requirement for spatial coherence of a cluster is that it be *convex-solid* in the sense that it includes all shortest paths between its member regions (convexity), and allows no holes (solidity).[8]

2.1 Interregional Distance

Since the underlying interregional network will have direct impact on the industrial clusters to be identified, it is worth discussing our choice of the network structure. In our practical applications in Section 4 and thereafter, we adopt the actual road network, and hence the interregional distances are measured in terms of the travel distance along the road network. Note that it is possible to use more stylized interregional distances, based for example on Great-Circle distances (as is common in the literature[9]). But, the advantage of choosing road-network data is primarily to take into account the underlying topographical heterogeneity, which could hardly be reflected in the Great-Circle distances. In the case of Japan to be studied below, the Pearson's correlation between the Great-Circle distances and road-network distances for all pairs of basic regions (municipalities) is as high as 0.976.[10] But, as suggested by Duranton and Overman [9, §4] the size of most industrial clusters are within the 40km range, and almost

[8]The requirement of solidity is not essential. But, it provides a more cohesive view of clusters as areas of industrial agglomeration.

[9]In particular, Duranton and Overman [9] and all of their followers.

[10]The magnitude of the correlation is comparable to 0.97 for the case of the United Kingdom reported in Duranton and Overman [9, footnote 4].

all are within 100km range. So this broad correlation over all scales is not sufficiently informative to gauge the relevance of the network distance. Namely, as is clear from the frequency distributions of interregional distances shown in Figure 1, the majority in the entire set of municipality pairs are simply too distant from one another to constitute meaningful clusters. More specifically, the municipality pairs within 100km range account for less than 10% of all the municipality pairs in both Great-Circle and road-network distances. In fact, the correlations reduce to 0.711, 0.633, 0.485 and 0.480, for the municipality pairs within 100km, 50km, 20km and 10km ranges (in terms of the Great-Circle distance), respectively. The corresponding correlations reduce even to 0.539, 0.461, 0.338 and 0.249, respectively, for the municipality pairs along the sea coast, where most of the major clusters are to be identified.

[Figure 1 about here.]

These results are not specific to Japan. In the case of 4626 unions of the continental Germany (which has comparable areal size with Japan) in 2008, while the Pearson's correlation for all the 10,697,625 union pairs is 0.911, it reduces to 0.803, 0.777, 0.630 and 0.382 for all the union pairs within 100km, 50km, 20km and 10km ranges, respectively.[11] In the case of 3106 counties in the continental US in 2007, while the same correlation for all the 4,822,065 county pairs is as high as 0.928, it reduces to as low as 0.216, 0.136, 0.073 and 0.019 for all the county pairs within 100km, 50km, 20km and 10km ranges, respectively.[12] Although the values of correlations are not directly comparable between Japan/Germany and the US, since the sizes of these regions greatly differ,[13] the discrepancy in the US case appears to be far more serious than the cases of Japan and Germany.

These evidences suggest that it is important to adopt realistic distance data to obtain reliable results on agglomeration patterns. In addition, it is a simple matter to compute bilateral distances along a given network in *R* by using a GIS software. In ArcGIS (ver.10.2) of ESRI, for instance, all the interregional dis-

[11]The distances between these unions are computed in terms of those between union offices. We thank Wolfgang Dauth and Jens Südekum for sharing Germany data with us.

[12]The distances between the US counties are computed in terms of those between the county courthouses, or some other public facilities if there are no courthouses.

[13]The US counties on average more than twenty times larger in areal size than Japanese municipalities.

tances can be automatically computed by utilizing the "network analyst" extension. Thus, today, there is no strong reason to choose simplistic distance data.[14]

2.2 Cluster Schemes

Most industries consist of multiple clusters in R that together define the agglomeration pattern for that industry. In fact, the spacing between such clusters is a topic of considerable economic interest (as discussed further in Section 7.1 below). Hence it is essential to model such patterns as explicit spatial arrangements of multiple clusters. The model proposed in Mori and Smith [37] is a *cluster scheme*, $\mathbf{C} = (R_0, C_1, \ldots, C_{k_\mathbf{C}})$, that partitions R into one or more disjoint clusters (convex solids), $C_1, \ldots, C_{k_\mathbf{C}}$, together with the residual set, R_0, of all non-cluster regions in R. The individual clusters are implicitly taken to be areas in R where industry density is unusually high. But within each cluster, C_j, all that is assumed for modeling purposes is that location probabilities for randomly sampled industrial establishments are uniform across the feasible locations in C_j. More precisely, if the *feasible area* as defined in Section 4.1.2 below for locations in each region, $r \in R$, is denoted by a_r, so that the total area of C_j is given by $a_{C_j} = \sum_{r \in C_j} a_r$, then location probabilities in C_j are taken to be uniform over a_{C_j}. In particular, this implies that the conditional probability of an establishment locating in $r \in C_j$ given that it is located in C_j is simply a_r/a_{C_j}. With this assumption, the only unknown probabilities are the marginal location probabilities, $p_\mathbf{C}(j)$, for clusters C_j in $\mathbf{C}$. Hence each cluster scheme, $\mathbf{C}$, generates a possible *cluster probability model*, $p_\mathbf{C} = [p_\mathbf{C}(j) : j = 1, \ldots, k_\mathbf{C}]$, of establishment locations for the industry.[15] If there are n establishments in the given industry, then each cluster probability model, $p_\mathbf{C}$, amounts formally to multinomial sampling model with sample size, n, and outcomes given by the $k_\mathbf{C} + 1$ sets in cluster scheme, $\mathbf{C}$, with respect to samples of size n. Finally, since the observed relative frequencies, $\widehat{p}_\mathbf{C} = [\widehat{p}_C(j) = n_j/n : j = 1, \ldots, k_\mathbf{C}]$, of establishments in each cluster are well known to be the maximum-likelihood estimates of these (multinomial) probabilities, such estimates yield a family of well-defined candidate probability

[14] See Combes and Lafourcade [6] for more sophisticated definition of interregional distances which takes into account more general costs for travel such as time and fuel costs. See also Özak [40] for the derivation of the least-cost routes based various topographical and climatic characteristics.

[15] This probability model is completed by the condition that $p_C(R_0) = 1 - \Sigma_j p_\mathbf{C}(j)$.

models for describing the agglomeration patterns of each industry.

2.3 Cluster-Detection Procedure

The key question remaining is how to find a "best" cluster-scheme for capturing the observed distribution of industry establishments. It is argued in Mori and Smith [37] that the *Bayes Information Criterion* (BIC) offers an appropriate measure of model fit in the present setting. In particular, for any given cluster scheme, $\mathbf{C}$, the (multinomial) log-likelihood of $\widehat{p}_{\mathbf{C}}$ is given by

$$L_{\mathbf{C}}(\widehat{p}_{\mathbf{C}}) = \sum_{j=0}^{k_{\mathbf{C}}} n_j(x) \ln\left(\frac{n_j(x)}{n}\right) \; + \; \sum_{j=0}^{k_{\mathbf{C}}} \sum_{r \in C_j} n_r \ln\left(\frac{a_r}{a_{C_j}}\right) \tag{1}$$

and that in terms of $L_{\mathbf{C}}(\widehat{p}_{\mathbf{C}})$, the appropriate value of BIC is given for each candidate cluster scheme, $\mathbf{C}$, by

$$BIC_{\mathbf{C}} = L_{\mathbf{C}}(\widehat{p}_{\mathbf{C}}) - \frac{k_{\mathbf{C}}}{2} \ln(n) \; . \tag{2}$$

Hence BIC is a "penalized likelihood" measure, where the second term in (2) essentially penalizes cluster schemes with a large number of clusters, $k_{\mathbf{C}}$, to avoid "over fitting" the data.

Given this criterion function, the present *cluster-detection procedure* amounts to a systematic way of searching the space of possible cluster probabity models to find a cluster scheme, $\mathbf{C}^*$, with a maximum value of $BIC_{\mathbf{C}^*}$.[16] While the details of this search procedure will play no role in the present analysis, the results of this procedure for Japanese industries will play a crucial role. Hence it is appropriate to illustrate these results in terms of the "livestock products" industry in Japan, shown in Figure 8 in Section 4.3.1 below.

Here Figure 8(a) shows the relative density of "livestock products" establishments in each municipality of Japan, where darker patches correspond to higher densities.[17] The red patches surrounded by a solid curve in Figure 8(b) show the cluster scheme, $\mathbf{C}^*$, that was produced for the "livestock products" industry by

[16] However, it should be emphasized that this space of probability models is very large, and hence that one can only expect to find *local* maxima (with respect to the particular perturbations defined by the search procedure itself).

[17] These municipalities are mapped in Figure 3 in Section 4.1.1.

this cluster-detection procedure.[18] Here it is seen that not all isolated patches of density are clusters. But the highest density areas do indeed yield significant clusters. Notice also that the convex solidification procedure above has produced easily recognizable clusters that do seem to reflect the shapes of these high density areas.[19]

2.4 A Test of Spurious Clusters

When attempting to identify "significant" clusters, it must be emphasized that even random locational patterns will tend to exhibit some degree of clustering. So there remains the statistical question of whether the "locally best" cluster scheme, $\mathbf{C}^*$, found for an industry by the above procedure is significantly better (in terms of BIC values) than would be expected in a random location pattern. A "random" location pattern is taken to be one in which location probabilities in all regions, $r \in R$, are proportional to their feasible areas, a_r. Hence a Monte Carlo test can be constructed by (*i*) generating N random location patterns for the establishments of a given industry, (*ii*) determining the locally optimal values, say BIC^*_s, for each simulated pattern, $s = 1 \ldots, N$, and (*iii*) comparing the value, $BIC_{\mathbf{C}^*}$, with this sampling distribution of BIC values. If $BIC_{\mathbf{C}^*}$ is sufficiently large (say in the top 1% of these values), then one may conclude that the clustering captured by $\mathbf{C}^*$ is significantly higher than what would be expected under randomness. Otherwise, $\mathbf{C}^*$ is said to involve *spurious clustering*. Results of this testing procedure for the application to Japanese manufacturing industries will be discussed in Section 4.2 below.

[18] The red area within each cluster contains establishments of the "livestock products" industry, while there is no establishments in the pink area which has been incorporated into the cluster through convex-solidification. See also for the refinement of cluster scheme proposed by Mori and Smith [37, §5.3] which constructs a set of *agglomerations*, each of which consists of a set of contiguous clusters with a single peak of establishment density.

[19] A complementary clustering approach has recently been proposed by Kerr and Kominers [26] which identifies establishment clusters based on maximal interaction distances. While this distance approach is particularly useful when the actual interactions between agents are known as in the case of patent citations, it is not directly applicable to the detection of establishment/employment clusters in general. Moreover, even when the interactions between agents are known, their approach by itself does not identify spatial patterns of clusters [such as our results in parts (b) of Figures 7 through 14].

2.5 On the Modifiable Areal Unit Problem

Finally, it should be noted that measures of agglomeration based on point data (as in Duranton and Overman [9]) are often considered to be "less biased" than those based on regional data (as in Ellison and Glaeser [10] and Mori et al. [34]). In particular, such measures are not restricted by either the size or shapes of existing regional units. Since our present results are based on Japanese municipalities, it is important to consider the robustness of these results with respect to this choice of regional units. In particular, there is in fact a systematic size difference among municipalities such that they are finer in urban than in rural area. Here it is possible to develop a robustness check by employing the somewhat finer grained equal-sized "secondary mesh" used by the Japanese Census. As detailed in Appendix A.1, the results produced by this mesh system are virtually identical to those using municipalities, and do indeed suggest that our results are robust with respect to the choice of regional units.

3 Spatial Scales of Agglomeration

As emphasized in the Introduction, the main strength of our cluster detection approach is to identify cluster schemes in a manner that preserves their two-dimensional spatial properties. By so doing, it is possible to analyze the spatial patterns of industrial agglomeration in more detail. As we will see for the case of Japanese manufacturing industries in Section 4, agglomerations of given industries often tend to concentrate within specific subregions of the country, i.e., are themselves "spatially contained". Hence our first task below is to construct an operational definition of such containments, designated as the *essential containment* (*e-containment*) for each industry. Our next task is to construct a measure of the relative size of these *e*-containments, designated as *global extent*. Industries with small global extent can be regarded as relatively "confined", and those with large global extent can be regarded as relatively "dispersed". Finally, industries can also differ with respect to their patterns of agglomeration *within* these *e*-containments. Some patterns may be "dense" and others "sparse". To compare such patterns, we construct a measure of the *local density* of clusters within each *e*-containment. This will yield a useful classification of agglomeration patterns in terms of their spatial scales to be discussed in Section 3.2.

3.1 Essential Containment

To formalize the notion of an industry's essential containment, we start by assuming that an optimal cluster scheme, $\mathbf{C} = \mathbf{C}^*$, has been identified for the industry.[20] The main idea is to identify an appropriate subset of "most significant" clusters in $\mathbf{C}$, and then take *essential containment* to be the convex solidification of this set of clusters in R. To identify "most significant" clusters, we proceed recursively by successively adding those clusters in $\mathbf{C}$ with maximum incremental contributions to *BIC*.[21] This recursion starts with the "empty" cluster scheme represented by $\mathbf{C}_0 \equiv \{R_{0,0}\}$ where $R_{0,0}$ denotes the full set of regions, R. If the set of (non-residual) clusters in $\mathbf{C}$ is denoted by $\mathbf{C}^+ \equiv \mathbf{C}\backslash\{R_0\}$, then we next consider each possible "one-cluster" scheme created by choosing a cluster, $C \in \mathbf{C}^+$, and forming $\mathbf{C}_0(C) = \{R_{0,0}(C), C\}$, with $R_{0,0}(C) = R_{0,0}\backslash C$. The "most significant" of these, denoted by $\mathbf{C}_1 = \{R_{1,0}(C), C_{1,1}\}$, is then taken to be the cluster scheme with the *maximum BIC value* (defined below). If this is called *stage* $t = 1$, and if the *most significant cluster scheme* found at each stage $t \geq 1$ is denoted by

$$\mathbf{C}_t \equiv \{R_{t,0}, C_{t,1}, \ldots, C_{t,t}\} \ , \tag{3}$$

then the recursive construction of these schemes can be defined more precisely as follows.

For each $t \geq 1$ let $\mathbf{C}^+_{t-1}$ denote the (non-residual) clusters in $\mathbf{C}_{t-1}$ (so that for $t = 1$ we have $\mathbf{C}^+_{t-1} = \mathbf{C}^+_0 = \varnothing$), and for each cluster not yet included in $\mathbf{C}_{t-1}$, i.e., each $C \in \mathbf{C}^+\backslash\mathbf{C}^+_{t-1}$, let $\mathbf{C}_{t-1}(C)$ be defined by,

$$\mathbf{C}_{t-1}(C) = (R_{t-1,0}(C), C_{t-1,1}, \ldots, C_{t-1,t-1}, C) \ , \tag{4}$$

where

$$R_{t-1,0}(C) = R_{t-1,0}\backslash C \ . \tag{5}$$

Then the *most significant additional cluster*, $C_t (\equiv C_{t,t})$ $(\in \mathbf{C}^+\backslash\mathbf{C}^+_{t-1})$, at stage $t \geq 1$ is defined by

$$C_t \equiv \arg \max_{C \in \mathbf{C}^+\backslash\mathbf{C}^+_{t-1}} L\left(\widehat{p}_{\mathbf{C}_{t-1}(C)} | \mathbf{C}_{t-1}\right) \ , \tag{6}$$

[20] For notational simplicity we drop the asterisk in $\mathbf{C}^*$.

[21] At this point it should be emphasized that the following procedure for identifying "significant clusters" in $\mathbf{C}$ is different from the one used to indentify $\mathbf{C}$ in Section 2.3. In particular, the only candidate clusters now being considered are those in $\mathbf{C}$ itself.

where $L\left(\widehat{p}_{\mathbf{C}_{t-1}(C)}|\mathbf{C}_{t-1}\right)$ is the *estimated maximum log-likelihood value* for model $p_{\mathbf{C}_{t-1}(C)}$ given [in a manner paralleling expression (1) above] by

$$L\left(\widehat{p}_{\mathbf{C}_{t-1}(C)}|\mathbf{C}_{t-1}\right) = \sum_{C'\in\mathbf{C}_{t-1}(C)} n_{C'} \ln\left(\frac{n_{C'}}{n}\right) + \sum_{C'\in\mathbf{C}_{t-1}(C)} \sum_{r\in C'} n_r \ln\left(\frac{a_r}{a_{C'}}\right) , \tag{7}$$

where $n_{C'} \equiv \sum_{r\in C'} n_r$ and $n \equiv \sum_{r\in R} n_r$. Thus, at each stage $t \geq 1$ the likelihood-maximizing cluster, C_t, is removed from the residual region, $R_{t-1,0}$, and added to the set of significant clusters in $\mathbf{C}_{t-1}$. The resulting *BIC* value at each stage t is then given by

$$BIC_{\mathbf{C}_t} = L_{\mathbf{C}_t} - \frac{t}{2}\ln(n) \tag{8}$$

with

$$L_{\mathbf{C}_t} = \sum_{C\in\mathbf{C}_t} n_C \ln\left(\frac{n_C}{n}\right) + \sum_{C\in\mathbf{C}_t} \sum_{r\in C} n_r \ln\left(\frac{a_r}{a_C}\right) . \tag{9}$$

Finally, the *incremental contribution* of each new cluster, C_t, to *BIC* is given by the increment for its associated cluster scheme, $\mathbf{C}_t$, as follows:

$$\triangle BIC_t \equiv BIC_{\mathbf{C}_t} - BIC_{\mathbf{C}_{t-1}} . \tag{10}$$

To identify the relevant set of "significant clusters" in $\mathbf{C}$, relevant requirements would depend on the objectives. For our present purpose of distinguishing the spatial scale of agglomeration, it suffices to impose a simple requirement that the sum of *BIC* contributions by the first $t^e \geq 1$ essential clusters accounts for at least a given *share*, $\lambda \in (0,1]$, in that of $\mathbf{C}$:[22]

$$\sum_{t=1}^{t^e} \Delta BIC_t \geq \lambda BIC_{\mathbf{C}} . \tag{11}$$

If the set of *essential clusters* in $\mathbf{C}$ is now defined to be $\mathbf{C}^e = \mathbf{C}^+_{t^e}$, then the desired *essential containment* (*e-containment*), $ec(\mathbf{C})$, for an industry with cluster scheme $\mathbf{C}$ is taken to be the smallest convex-solid set in R containing $\mathbf{C}^e$, i.e., the convex solidification of $\mathbf{C}^e$.[23]

[22] It would seem the most natural to simply add clusters as long as the increments are positive. But from the original construction of $\mathbf{C}$, it should be clear that these increments may often be positive for *all* $t = 1,\ldots,k_{\mathbf{C}}$. See Mori and Smith [37, §4.4] for alternative requirements.

[23] In terms of the d-convex solidification operator, σc_d, defined in Mori and Smith [37, eq. (26)] (with respect to shortest-path travel distance, d), the formal definition of *e-containment* is given

These concepts can be illustrated by the stylized location patterns in Figure 2 below. For example, if the relevant cluster scheme, $\mathbf{C}$, for a given industry corresponds to the five clusters (shown in black) in Figure 2(a), and if the subset of essential clusters, $\mathbf{C}^e$, consists of the three largest clusters on the left, then the e-containment, $ec(\mathbf{C})$, for this industry is given by the filled square containing these three clusters.[24] Similar interpretations can be given to the filled rectangles of Figures 2(b,c,d).

[Figure 2 about here.]

3.2 Global Extent and Local Density

With these definitions we next seek to compare e-containments for different industries in terms of their relative sizes. In order to reflect the actual spatial extent of such containments, it is now more appropriate to measure "size" in terms of total *geographic area* rather than the more limited notion of *feasible area* (employed for modeling the potential locations of individual establishments, as in Sections 2.2 above). Hence if we now let A to denote *geographic area*, then the economic areas for *basic regions* (a_r), *clusters* (a_C), and the *entire country* (a), are here replaced by A_r, A_C, and A, respectively. With these conventions, the *global extent* (GE) of an industry is now taken to be simply the total area of its e-containment, $ec(\mathbf{C})$, relative to that of the entire country:

$$GE(\mathbf{C}) = \frac{\sum_{r \in ec(\mathbf{C})} A_r}{A} \in (0,1] \; . \tag{12}$$

Industries with relatively small global extents might be classified as "globally confined" industries [illustrated by the industries in Figures 2(a,c)]. Similarly, industries with substantially larger global extents might be classified as "globally dispersed" industries [illustrated by those in Figures 2(b,d)].[25]

Finally, we consider the relative denseness of essential clusters within the e-containment for each industry. As a parallel to global extent, we now define the

by $ec(\mathbf{C}) = \sigma c_d(\mathbf{C}^e)$.

[24] It is assumed that the centroid of each cell is a vertex of the underlying regional network, and that only the neighboring pairs of cells are connected directly (with the relevant distance being Euclidean distance between their centroids).

[25] One might consider more exact classifications, such as $GE < 1/2$ for "globally confined" and $GE \geq 1/2$ for "globally dispersed." But in our view, the appropriate ranges of GE may often be context dependent.

local density (LD) of a given industry to be simply the total area of its essential clusters, $\mathbf{C}^e$, relative to that of its e-containment, $ec(\mathbf{C})$, i.e.,

$$LD(\mathbf{C}) = \frac{\sum_{r \in \mathbf{C}^e} A_r}{\sum_{r \in ec(\mathbf{C})} A_r} \in (0, 1] \, . \tag{13}$$

Industries with a relatively high density of clusters in their e-containments might be classified as "locally dense" industries [illustrated by the industries in Figures 2(a,b)]. Similarly, industries with a substantially lower density of clusters in their e-containments might be classified as "locally sparse" industries [illustrated by those in Figures 2(c,d)].

More generally, Figure 2 is intended to summarize the main features of this classification system. First, the concept of the e-containment is designed to capture the region of most significant agglomeration for an industry. This is illustrated in each of the stylized figure panels by filled regions containing the largest clusters within the cluster schemes shown. In each case, the "outlier" clusters excluded from this region are implicitly assumed to be less significant in terms of their contributions to BIC.

Each of the four panels in this figure depicts a type of extreme case in the present classification system. However, it should be emphasized that there is no unambiguous ordering among these extremes. Indeed, it is a fundamental tenet of this paper that the types of concentration/dispersion continua implied by scalar measures of concentration are simply too limiting. In contrast, Figure 2 can be said to represent the extremes of a *two-dimensional* ordering: For any given level of Local Density, higher values of Global Extent tend to reflect industrial patterns that are more dispersed throughout the country. Similarly, for any given level of Global Extent, higher values of Local Density tend to reflect industrial location patterns that are more dispersed throughout their essential containments. More detailed examples of these extremes will be developed in Section 4 below.[26]

[26]It should also be noted that the extremes in Figure 2 have differing implications for the overall *size* of the industries involved. In particular, only industries with many establishments can possibly exhibit dense patterns of significant clusters over large areas [such as Figure 2(b)], and only industries with small numbers of establishments can exhibit sparse patterns of significant agglomeration in confined areas. [such as Figure 2(c)]. This constrast can also be seen by comparing Figures 7 and 13 in Section 4 below.

4 Detection of Industrial Clusters in Japan

In this section, we apply the above set of cluster-analytic tools to study the agglomeration patterns of manufacturing industries in Japan. We begin in Section 4.1 with a description of the relevant data for analysis. This is followed in Section 4.2 with a summary of results for the spurious-cluster test described in Section 2.4. The classification scheme developed in Section 3 is then given an operational form for the present application. Finally, this classification scheme is illustrated by means of a number of selected examples in Section 4.3.

4.1 Data for Analysis

The data required for this application includes both quantitative descriptions of the relevant system of regions and the class of industries to be studied. We consider each of these data types in turn.

4.1.1 Basic Regions

The relevant notion of a "basic region" for this analysis is taken to be the *shi-ku-cho-son*, which is a municipality category equivalent to a city-ward-town-village in Japan. The specific municipality boundaries are taken to be those of October 1, 2001.[27] While there are a total of 3363 municipalities in Japan, we only consider 3207 of these (as shown in Figure 3), namely those that 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.[28]

The only exception here is Hokkaido, which is one of the four major islands (refer to Figure 3), but is disconnected from the road network covering the other three. Given its size (217 municipalities), as seen in Figure 3, we still include Hokkaido as a potential location for establishments. Aside from this exceptional case, we adopt the following conventions. First, while we allow establishments to locate freely within the 3207 municipalities, we do not allow the formation of any

[27] The data source for these municipality boundaries is the Statistical Information Institute for Consulting and Analysis [44, 45].

[28] See Appendix A.1 for the robustness analysis using equal-sized mesh regions instead of municipalities.

clusters including municipalities in both Hokkaido and other major islands.[29] Second, *e*-containments for each industry are obtained as the union of the two convex solidified subsets of essential clusters within and without Hokkaido [see, e.g., the cases of "sliding doors and screens", "livestock products", and "manufactured ice" shown in Figures 7(b), 8(b) and 9(b), respectively, in Section 4.3 below].

[Figure 3 about here.]

4.1.2 Economic Area

To represent the areal extent of each basic region we adopt the notion of "economic area", obtained by subtracting forests, lakes, marshes and undeveloped area from the total area of the region (available from the Statistical Information Institute for Consulting and Analysis [44, 45]).[30] 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%. Not surprisingly, those municipalities with highest proportions of economic area are concentrated in urban regions. In this respect, our present approach is relatively more sensitive to clustering in rural areas.[31]

4.1.3 Interregional Distances

The travel distance between each pair of neighboring municipalities is computed as the length of the shortest route between their municipality offices along

[29] In terms of our δ-neighborhood definition in Mori and Smith [37, §4.2.2], the distances between Hokkaido regions and those of the major islands are implicitly assumed to exceed δ.

[30] There is of course a certain degree of interdependence between the size of economic areas and the presence of industries in those areas. On the one hand, industrial growth in a region may well lead to a gradual increase in the economic area of that region (say by land fills or deforestation). But to capture agglomeration patterns at a given point in time, we believe that it is more reasonable to adopt economic area than total area as the potential location space for establishments. In Japan, for example, it is doubtful that mountainous forested regions (which account for 98% of non-economic areas after two centuries of history since the beginning of cultivation) can be easily be made available for industrial location in the short run. On the other hand, our economic regions may be overstating the feasible area for establishments at least in the short-run if zoning restrictions are taken into account. See Appendix A.2 for the robustness analysis for alternative feasible area which reflects zoning restrictions more closely.

[31] In other words, for any given number of firms, n_r, in a basic region r, our clustering algorithm implicitly regards n_r as a more significant concentration in regions with smaller economic areas (other things being equal).

the road network.[32] From the computed pairwise distances between neighboring (contiguous) municipalities, the *shortest-path distances* (and associated sequences of neighboring municipalities) are computed in terms of Mori and Smith [37, eq.(15)].[33] While there is of course some degree of interdependency between industrial locations and the road network, the spatial structure of this network is mainly determined by topographical factors.

4.1.4 Industry and Establishments Data

Finally, the industry and establishments data used for this analysis is based on the Japanese Standard Industry Classification (JSIC) in 2001. Here we focus on three-digit manufacturing industries, of which 163 industrial types are present in the set of basic regions chosen for this analysis.[34] The establishment counts (n) across these 163 industries is taken from the Establishment and Enterprise Census of Japan [25] 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.

4.2 Tests of Spuriousness of Cluster Schemes

Using the cluster-detection procedure developed in Section 2.3, optimal cluster schemes, $\mathbf{C}_i^*$, were identified for each industry, $i = 1,\ldots,163$. Each cluster scheme, $\mathbf{C}_i^*$, was then tested for spuriousness using the testing procedure developed in Section 2.3.[35] Among the 163 industries studied, the null hypothesis of complete spatial randomness (Section 2.4 above) was strongly rejected for 155 of these industries. For the remaining eight industries, this null hypothesis could

[32]This road network data is taken from Hokkaido-chizu Co. Lit. [19], and includes both prefectural and municipal roads. However, if a given municipality office is not on one of these roads, then minor roads are also included.

[33]As noted in Mori and Smith [37, §3.1], shortest-path distances are always at least as large as shortest-route distances. But in the present case, shortest-path distance appears to approximate shortest-route distance quite well. For the distribution of ratios of short-path over shortest-route distances across all 4,491,991 relevant pairs of municipalities, the median and mean are both equal to 1.14. In fact, the 99.5 percentile point of this distribution is only 1.28.

[34]More precisely, out of the 164 industrial types in Japan, all but one have establishments in at least one of our basic regions.

[35]These tests of spuriousness were based on the *BIC* values for a sample of 10,000 completely random location patterns for each industry.

not be rejected at the .01 level. The main reason for non-rejection in these cases (which include seven arms-related industries, together with "coke"), appears to be the small size of these industries, with $n < 40$ in all cases.[36] In view of these findings, we chose to drop the eight industries in question and focus our subsequent analyses on the 155 industries exhibiting significant clustering.[37]

For these 155 industries, Figure 4 shows the frequency distribution of the share of establishments for each industry i that are included in the clusters of its cluster scheme, $\mathbf{C}_i^*$. These shares range from 39.1% to 100% with a median (mean) share of 95.2% (93.6%). The industries with the smallest shares of establishments in clusters are typically those which exhibit the weakest tendency for clustering. For instance, "paving materials" industry and "sawing, planning mills and wood products" industry have 39.1% and 54.0% of their establishments in the clusters, respectively. Since both of these industries are typically sensitive to transport costs, their establishment locations tend to reflect population density.

[Figure 4 about here.]

4.3 Classification of Cluster Patterns

To apply our two measures (GE, LD) for classification purposes, we begin by recalling that the key parameter defining e-containments for industries with cluster schemes, $\mathbf{C}$, is the *share*, λ, of the total $BIC_{\mathbf{C}}$ values accounted for by clusters in these e-containments. So it is necessary to specify an appropriate value of λ. In terms of classification, it is useful to consider the consequences of λ for possible correlations between GE and LD. For if these measures are too highly correlated (either positive or negative), then it is doubtful that they can both provide distinct information useful for classification purposes. With this in mind, we first observe if λ is very small, then e-containments will include only a few highly significant clusters. If these clusters are concentrated in a small region for a given

[36] These industries are also rather special in other ways. Arms-related industries are highly regulated industries, so that their location patterns are not determined by market forces, while "coke" is a typical declining industry in Japan (steel industries have gradually replaced coke production by less expensive powder coal after the 1970s).

[37] In the application in Mori and Smith [37, §5] using the same data, 154 instead of 155 industries exhibited significant clustering based on 1000 samples for random establishment locations, instead of 10,000 samples in the present study. Specifically, "tobacco manufacturing" industry turned out to exhibit significant clustering under the present larger Monte Carlo simulation.

industry, then *GE* will be small and *LD* is likely to be large. Conversely, if these clusters are widely separated for a given industry, then *GE* will be large and *LD* is likely to be smaller. So for small λ it seems clear that *GE* and *LD* should be strongly negatively correlated across industries. On the other hand, if λ is very large, then e-containments will tend to include almost all of an industry's clusters. So the question is whether industries that are more spread out (i.e., with higher *GE* values) also tend to have denser cluster patterns (i.e., higher *LD* values). In our data this appears to be the case, so that *GE* and *LD* are in fact positively correlated at high values of λ .

These observations are quantified in Figure 5, where we have plotted the (Pearson) correlations, ρ, between *GE* and *LD* across our 155 industries (with non-spurious cluster schemes) for a the full range of λ values. Here the solid red curve shows correlation values, ρ, and the dashed blue curve shows the corresponding p-values (for a two-sided test of ρ significance). As is seen in the figure, ρ is significantly negative (at the 0.05 level) for λ less than about 0.67, and is significantly positive for λ above 0.92. Moreover, since p-values rise sharply between these two extremes, it can be concluded that *GE* and *LD* are essentially uncorrelated within the range, $\lambda \in [0.67, 0.92]$, so that industries are most differentiated in terms of their agglomeration patterns within this range of λ.

[Figure 5 about here.]

In particular, the correlation between *GE* and *LD* is seen to be least significant at approximately $\lambda = 0.88$. For this value of λ, we have plotted the pairs (GE, LD) for each of the 155 industries in Figure 6. Here it is seen that *GE* and *LD* are essentially unrelated, so that all four extremes in Figure 2 do in fact occur simultaneously. For convenience, the relative positions of panels (a) through (d) in Figure 2 are arranged to match the relative positions in Figure 6. For example, the types of globally confined patterns illustrated in the left panels (a,c) of Figure 2 are typical of industries with (GE, LD) pairs in the left portion of Figure 6. Similarly, the locally dense patterns in the top panels (a,b) of Figure 2 are typical of industries with (GE, LD) pairs toward the top of Figure 6.

[Figure 6 about here.]

Within this general framework, it is of interest to consider more detailed examples of industries with cluster schemes exhibiting a variety of (GE, LD) com-

binations. Here we focus on the case, $\lambda = 0.88$, in Figure 6 which exhibits the widest variation of GE and LD values.[38] Figures 7 through 12 focus on different industries. For each industry i, panel (a) of the figure shows the density of i establishments across municipalities (where darker colors denote municipalities with higher densities). Panel (b) of the figure shows both the spatial pattern of clusters and their e-containment for industry i. Here individual clusters are represented by the enclosed red areas,[39] and the corresponding e-containment (for $\lambda = 0.88$) is shown in yellow.

4.3.1 Globally Dispersed and Locally Dense Patterns

Industries with high values of both GE and LD (located in the upper-right portion of Fig. 6) can be described as exhibiting patterns of agglomeration that are "globally dispersed and locally dense". Such industries are by definition present almost everywhere, and can equivalently be described as *ubiquitous industries*. As discussed in Section 3.2, this pattern is evaluated as the "maximally dispersed" in terms of scalar indices of agglomeration. A typical example is the "sliding doors and screens" (with $GE = 0.749$, $LD = 0.336$; $\odot$ in Fig. 6). As indicated by Figure 7(a), establishments are present in almost all municipalities, and the clusters are found to be densely distributed throughout the country. Their products are often custom made and require face-to-face contact with customers, and hence, there are strong market-attraction forces that contribute to the ubiquity of this industry.

It is also of interest to note (as mentioned in footnote 26) that such ubiquitous industries are by their very nature quite large in terms of establishment numbers. In the present case, "sliding doors and screens" industry has 15,363 establishments, which is well above the mean of 4189 for all industries. In terms of establishments in clusters, this industry has 13,565 establishments relative to the mean of only 3896 for all industries.

[Figure 7 about here.]

[38] In fact, the qualitative results hereafter remain the same for all $\lambda \in [0.10, 1]$, i.e., except for the degenerate case involving a single essential cluster.

[39] The portion of each cluster in pink shows those basic regions which contain no establishments (but are included in the cluster by the process of convex solidification).

Figure 8 shows the location patterns of another ubiquitous industry, "livestock products". The clusters of this industry exhibit slightly smaller global extent and local density ($GE = 0.645$ and $LD = 0.258$; ⊙ in Fig. 6) than does "sliding doors and screens" industry above. But, they are still relatively globally dispersed and locally dense. The reason for ubiquity of clusters in this industry is straightforward, since freshness is critical for most of its products so that market proximity is a major determinant of establishment locations.

[Figure 8 about here.]

4.3.2 Globally Dispersed and Locally Sparse Patterns

Industries with relatively high values of GE and low values of LD (near the lower-right portion of Fig. 6) can be described as exhibiting patterns of agglomeration that are "globally dispersed and locally sparse". A clear example is provided by the "manufactured ice" industry shown in Figure 9 (with $GE = 0.589$ and $LD = 0.133$; ⊙ in Fig. 6). Global dispersion here reflects the high cost of shipping ice, while local sparseness suggests that there are scale economies in production. In fact, the number of establishments in this industry is only 387 which is about one tenth of the mean establishment counts of all the three-digit manufacturing industries.

[Figure 9 about here.]

Another extreme example is provided by the "seafood products" industry depicted in Figure 10 (with $GE = 0.931$ and $LD = 0.116$; ◬ in Fig. 6). The primary location determinant for this industry is obviously proximity to the coast, so that establishment locations are dense along the coast but sparse inland.

[Figure 10 about here.]

4.3.3 Globally Confined and Locally Dense Patterns

Industries with relatively low values of GE and high values of LD (in the upper-left portion of Fig. 6) can be described as exhibiting patterns of agglomeration that are "globally confined and locally dense". An extreme example of such industries is provided by the "ophthalmic goods, including frames" industry in Figure 11 (with $GE = 0.009$ and $LD = 0.988$; ◬ in Fig. 6). This industry is strongly

concentrated in a single town, Sabae, with a population of only 65,000. In fact, this one small town accounts for more than 90% of the national market share in ophthalmic goods. Not surprisingly, the e-containment for this industry consists only of this single town, as shown in Figure 11(b). As with many specialized industries, the location pattern of this industry is governed more by historical circumstances than economic factors at present. In terms of establishment counts, such industries are necessarily small in size. In the present case, there are only 1139 establishments, which is well below the mean of 4188 for all industries. So even though all of its 1139 establishments are in clusters, this number is still well below the mean of 3896 for all industries.

[Figure 11 about here.]

A second example is provided by the "compounding plastics and reclaimed plastics" industry (with $GE = 0.240$ and $LD = 0.478$; △ in Fig. 6). From Figure 12, it is clear that most clusters for this industry, and indeed most of its establishments, lie in the Industrial Belt. The outputs of this industry are primarily intermediate inputs for a variety of manufactured goods produced along the Belt, particularly home electronics appliances and motor vehicles (such as the molded plastic parts for seats, fenders, and instrument panels). Thus the intermediate locations between these manufacturers constitute natural market-oriented locations for this industry. In fact, many industries with (GE, LD) values similar to this industry also exhibit Industrial-Belt type agglomerations.

[Figure 12 about here.]

4.3.4 Globally Confined and Locally Sparse Patterns

Finally, "globally confined and locally sparse" agglomeration patterns (in the lower-left portion of Fig. 6) are mostly exhibited by those industries with establishments concentrated in the major cities along the Pacific coast. A representative case is provided by the "iron industry with blast furnaces" industry (with $GE = 0.068$ and $LD = 0.090$; ⊙ in Fig. 6), where plant-level scale economies are so large that the entire industry consists of only 38 establishments. As seen in Figure 13(a), most establishments are concentrated around the major ports along the Pacific coast (in order to gain access to both their imported inputs and

largest output markets). Since these major ports are widely spaced along the coast from Tokyo to Oita (more than 1000km apart), clustering also appears to be locally sparse in this region.

[Figure 13 about here.]

A final example is provided by the "publishing industry" depicted in Figure 14 (with $GE = 0.354$ and $LD = 0.145$; △ in Fig. 6). Publishing is typical of "urban-oriented" industries with location patterns tending to reflect urban density. While both the establishments and clusters of this industry are spread throughout the country, Figure 14 shows that there is relatively more concentration in the Pacific coast area between Tokyo and Osaka, with a narrow band stretching beyond Osaka to include the major metro areas further west (Kobe, Okayama, Hiroshima, and Fukuoka).

[Figure 14 about here.]

5 Comparisons with Scalar Indices

The 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 [42] 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.[40] But, not surprisingly, such scalar measures often yield similar values for industries with very different spatial patterns of agglomeration.

As will be seen below, the scalar index which is most closely related to our cluster detection approach is the *D-index* developed by Mori et al. [34]. This *D*-index for a given industry i is defined by the Kullback-Leibler [28] divergence of its establishment location probability distribution, $P_i \equiv [P_i(r) : r \in R]$, from a purely random establishment location patterns, $P_0 \equiv [P_0(r) : r \in R]$, as defined in Section 2.3 above. By using the sample estimate of P_i, namely, $\widehat{P}_i = \left[\widehat{P}_i(r) : r \in R\right]$

[40] Refer to footnote 1 for the choice of reference distributions in the existing indices.

with $\widehat{P}_i(r) \equiv n_r/n$, a corresponding estimate of this D-index is given by

$$D\left(\widehat{P}_i|P_0\right) = \sum_{r\in R} \widehat{P}_i(r) \ln\left(\frac{\widehat{P}_i(r)}{P_0(r)}\right). \tag{14}$$

The intuition behind this particular index is that it provides a natural measure of distance between probability distributions. So if uniformity is taken to represent the complete absence of clustering, then it is reasonable to assume that those distributions "more distant" from the uniform distribution should involve more agglomeration. Note that since both D and BIC given by (2) 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.

In terms of popularity, the primary index has been the γ-index developed by Ellison and Glaeser [10]. For a given industry $i \in I$, the *γ-index* is defined by

$$\gamma_i = \frac{G_i - \left(1 - \sum_{r\in R} x_r^2\right) H_i}{\left(1 - \sum_{r\in R} x_r^2\right)(1 - H_i)}. \tag{15}$$

In (15), G_i represents the Herfindahl-Hirschman index of employment concentration of industry i given by $\sum_{r\in R}(x_{ir} - x_r)^2$, where x_{ir} and x_r are the shares of region $r \in R$ in the total employment of industry i and that of the aggregate industry, respectively[41]; H_i is the Herfindahl-Hirschman index of employment distribution across all the establishments in industry i given by $\sum_{j\in E_i} h_j^2$, where E_i is the set of all establishments in the industry, and h_j is the share of establishment $j \in E_i$ in the total employment of industry i. Notice that the definition of "complete dispersion" for this index is different from the D-index above as well as our present cluster detection. Specifically, γ measures the *squared deviation* of the employment distribution of industry in question from that of the aggregate industry (with certain adjustments for heterogeneity in establishment sizes), which means that industries whose establishments are either more spatially concentrated or dispersed than that of the aggregate industry are evaluated as *equally* more concentrated than the aggregate industry. As pointed out below, this raises certain questions about the interpretation of γ .

Another popular index is the one proposed by Duranton and Overman [9]. They start by computing the Euclidean distance between each pair of establish-

[41]The "aggregate industry" in the present case is all manufacturing.

ments in a given industry $i \in I$. Given that there are n_i establishments in this industry, the estimator of the density of bilateral distances, called K-density, at each distance level, d, is defined by

$$\widehat{K}_i(d) = \frac{1}{n_i(n_i-1)h} \sum_{j=1}^{n_i-1} \sum_{k=j+1}^{n_i} f\left(\frac{d-d_{jk}}{h}\right), \tag{16}$$

where d_{jk} is the distance between establishments j and k, f the Gaussian kernel function, and h the bandwidth set according to Silverman [43, §3.4.2]. Roughly speaking, $\widehat{K}_i(d)$ is larger when the distances between many establishment pairs in industry i are approximately d. For each industry i, this K_i-density is then compared with the *counterfactual K-density* estimated from 1000 simulations of bilateral distances between n_i randomly sampled (distinct) establishments in the aggregate industry.

To identify the distance levels at which the industry in question exhibit significant concentration (or dispersion), Duranton and Overman [9] distinguish between "local" and "global" confidence bands. Our interest focuses only on global confidence bands, which are defined in the following way. First, one defines *local p% confidence bands* by identifying the p-percentiles of the simulated counterfactual distributions of $K(d)$ values at each distance $d = 0, 1, \ldots, 296$, and then interpolating these percentile points into continuous bands, where $d = 296$km is the median bilateral distance of all the establishments. Given these local bands, one then defines the *upper* [resp., *lower*] 5% *global confidence* $\overline{K}_i(d)$ [resp.,$\underline{K}_i(d)$] for this industry to be the highest [resp., lowest] local confidence band that is hit by at least 5% of the simulated counterfactual K-densities. In these terms, industry i is said to be *localized* if $\widehat{K}_i(d) > \overline{K}_i(d)$ for at least one distance $d \in [0, 296]$, and similarly, is said to be *dispersed* if it is not localized, and $\widehat{K}_i(d) < \underline{K}_i(d)$ for at least one $d \in [0, 296]$.[42] In these terms, the *degree of localization* at each distance, d , is defined by

$$\Gamma_i(d) \equiv \max\left\{\widehat{K}_i(d) - \overline{K}_i(d), 0\right\}, \tag{17}$$

[42]Duranton and Overman [9] use the respective terms "globally localized" and "globally dispersed".

and the corresponding *degree of dispersion* is defined by

$$\Psi_i(d) \equiv \begin{cases} \max\left\{\underline{K}_i(d) - \widehat{K}_i(d), 0\right\} & \text{, if } \sum_{d=0}^{296} \Gamma_i(d) = 0\,, \\ 0 & \text{, otherwise}\,. \end{cases} \tag{18}$$

While the overall degrees of localization and dispersion for a given industry i are defined separately in Duranton and Overman [9] by $\Gamma_i \equiv \sum_{d=0}^{296} \Gamma_i(d)$ and $\Psi_i \equiv \sum_{d=0}^{296} \Psi_i(d)$, respectively, these can be combined to define a single *localization index* as follows:

$$\Gamma_i^* \equiv \Gamma_i - \Psi_i\,, \tag{19}$$

where industry i is a *localized* (*dispersed*) industry (relative to the aggregate industry) if Γ_i^* is positive (negative).[43]

To relate these indices to GE and LD, the most direct approach is simply to plot their pairwise relations (for $\lambda = 0.88$) as in Figure 15, where these relations are seen most clearly in terms of $\log(GE)$ and $\log(LD)$ [and where γ is has also been transformed to $\log\gamma$].[44]

[Figure 15 about here.]

Here it is clear that both D and γ are significantly (negatively) correlated with both GE and LD, while Γ^* is correlated only with GE.[45] But recall that since GE and LD are uncorrelated (for $\lambda = 0.88$), these visual relations can best be quantified in terms of the following multiple regression model:

$$Y = a + b\log GE + c\log LD + \varepsilon \tag{20}$$

where $Y = D$, $\log\gamma$ or Γ^* and where a, b and c are parameters to be estimated (assuming normal errors, ε). The results of these regressions are shown in Table 1, where all the visual observations above are confirmed. It should be emphasized that the qualitative result here remains essentially the same for all $\lambda \geq 0.10$, i.e., most of the cases in which essential clusters are multiple.

[43] Micro data of the the Establishment and Enterprise Census of Japan [25] obtained via RIETI has been used to compute γ and Γ^*-indices.

[44] Γ^* values were computed by using the R-package, *dbmss*, developed by Marcon et al. [30]. We thank Kohei Takeda for his research assistance.

[45] Spearman's rank correlations between GE and (D, γ, Γ^*) and are respectively $(-.0574, -0.375, -0.773)$, and between LD and (D, γ, Γ^*) are respectively $(-0.681, -0.598, 0.060)$, where only the correlation between LD and Γ^* is not significant.

[Table 1 about here.]

Moreover, the adjusted R^2 results suggest that these scalar indices (D,logγ,Γ^*) are reasonably well approximated by their predicted values ($\widehat{D}$,$\widehat{\log\gamma}$,$\widehat{\Gamma}^*$) as linear combinations of log(GE) and log(LD). This is confirmed by the regression plots shown in Figure 16.

[Figure 16 about here.]

These regressions help to illustrate the more important similarities and distinctions between the three indices in terms GE and LD. With respect to similarities, it should be clear that all these indices tend to agree when industries are unambiguously concentrated in space, i.e., when GE is extremely small. This is well illustrated by the "ophthalmic goods" industry (Fig. 11), which corresponds to the symbol, ◬ , at the extreme end of the clustering spectrum on all three indices. A less extreme example is provided by "publishing industries" (Fig. 14) with GE again quite small and with symbol, ◬ , located toward the extreme clustering end for all three indices.

But aside from these extreme cases, the interpretations of such scalar indices can often be quite ambiguous. In particular, it is difficult for these indices to differentiate between "globally confined and locally dense" and "globally dispersed and locally sparse" patterns – which can be very different. Such differences are most often related to the *spatial scale* of agglomeration in the relevant industries. A good example of the first type of industry is provided by "compounding plastics materials" (refer to Fig. 12) with e-containment confined to the Industrial Belt stretching for more than 1000km along the Pacific coast area, but with e-clusters quite densely packed inside this area. The spatial scale of agglomeration for this industry is thus best described by the Industrial Belt itself. More generally, industries with relatively large LD compared to GE can be said to exhibit agglomeration at *larger* spatial scales. The converse is true for industrial patterns that are globally dispersed but locally sparse, i.e., with relatively large GE compared to LD. A good example is provided by the "manufactured ice" industry (refer to Fig. 9) where agglomeration is seen to occur at a *smaller* spatial scale, in this case extending only over a few adjacent municipalities. But in spite of the differences between these two types of industrial patterns, such industries are often grouped closely together by scalar indices. For "compounding plastics

materials" and "manufactured ice" in particular, this is seen to be true for all three indices (as indicated by the closeness of their respective positions, ◬ and ⊙ on the horizontal axes in Figure 16.[46]

Turning now to a more detailed consideration of these three indices themselves, note first from the adjusted R^2 values in Table 1 (as well as an inspection of Figure 16) that the *D-index* of Mori et al. is most fully captured by model (20). Note in particular that since the estimated coefficients of both $\log(GE)$ and $\log(LD)$ for D are close to one, the relative values of D are well approximated by $-\log(GE \times LD)$. Moreover, since the product, $GE \times LD$, is seen from (12) and (13) to be simply the *areal share* of an industry's *e*-clusters within the nation as a whole, it follows that D itself is essentially a decreasing function of this areal share. This simple relation is due largely to the fact that the uniform reference measure on which D is based is essentially area itself.

Turning next to the γ*-index* of Ellison and Glaeser, note from Table 1 that the estimated coefficients of both $\log(GE)$ and $\log(LD)$ for $\log(\gamma)$ are almost identical, so that γ is again seen to be essentially decreasing in areal share. This accounts for much of the similarity in the behavior of D and γ . But note also that there are important differences, as seen by the much larger degree of unexplained variation in $\log(\gamma)$ [i.e., lower adjusted R^2]. As mentioned above, this is largely due to the *unsigned* nature of squared deviations implicit in γ, which can in principle equate very different types of patterns. This is well illustrated by a comparison of the spatially concentrated "iron industries with blast furnaces" (Fig. 13) with the much more ubiquitous "sliding doors and screens" industry (Fig. 7), as denoted respectively by ⊙ and ⊙ in Fig. 16. This difference is strongly reflected by D in panel (a) where the "sliding doors and screens" industry is seen to be much more uniformly distributed (i.e., smaller D). But the γ-index essentially equates the two, reflecting the fact that these two industries are deviating in opposite directions from the aggregate industry.

Turning finally to Γ^**-index* based on Duranton and Overman, we begin by observing from Table 1 that this index is far more sensitive to GE than to LD. There appear to be at least two factors contributing to this asymmetry. One is the relation of bilateral distances for individual industries to those of the aggregate industry. As documented by Mori et al. [35] and Mori and Smith [36] for Japan

[46]Note however that "closeness" between Γ^* values on either side of zero is somewhat more difficult to gauge.

and Hsu [20, Appendix] and Davis and Dingel [8] for the US, clustering tends to be spatially coordinated *across* industries, so that clusters of many of industries tend to coincide in larger cities. (As an extreme case, Tokyo contains clusters of all industries.) So when sampling counterfactuals from the aggregate industry, there tend to be larger numbers of small distances than would be expected for individual industries. The result is that such frequency comparisons tend to *understate* the significance of local concentrations for individual industries relative to the aggregate industry.[47] The second contributing factor is that bilateral distances underlying the Γ^*-index include *all* establishment pairs, so that no distinction is made between *within*-cluster and *between*-cluster pairs. The consequences of this lack of distinction are most severe for "globally dispersed and locally sparse" industries, where there tend to be many more between-cluster pairs than within-cluster pairs. As a result, the frequencies of larger distances between clusters tend to dominate those of smaller distances within clusters. (A simple example illustrating these effects is given in Appendix B.)

In fact, if one considers all 41 industries, i, that are "globally dispersed and locally sparse" in the sense that GE_i is above the median and LD_i is below the median, then it turns out that *none* of these 41 exhibit significant localization at distances below 100km relative to the aggregate industry. This can be illustrated in more detail by comparing two globally dispersed industries "livestock products" (Fig. 8) and "manufactured ice" (Fig. 9) with similar global extents ($GE = 0.645$ and 0.589) but with very different local densities ($LD = 0.258$ and 0.133) reflecting the more locally concentrated nature of "manufactured ice". While both D and γ reflect this difference, and evaluate "manufactured ice" to be more concentrated, these two industries are essentially indistinguishable in terms of Γ^* [compare the relative locations of ⊙ and ⊙ in panels (a) and (b) with those in panel (c) of Fig. 16]. For as seen by their respective K-densities in Figure 17, all differences between these two patterns are completely overwhelmed by the lack of any discernible localization at small distances under such K-density tests.[48]

[47]The justification for adopting the establishment location pattern of the aggregate industry as counterfactual location pattern for each individual industry according to Duranton and Overman [9, p.1085] is that it roughly reflects the set of feasible locations under zoning restrictions. But, as summarized in Appendix A.2, all of the qualitative results in this section remain the same even after controlling for zoning restrictions in identifying industrial clusters. The main reason for this is that among all feasible locations for manufacturing, the aggregate industry is disproportionately more concentrated in urban areas.

[48]While this underestimate of local concentration by Γ^*-index is partly due to the downward

[Figure 17 about here.]

Finally, Table 2 lists the fifteen most as well as least localized industries in terms of D-index with all the corresponding values of γ, Γ^*, GE and LD, where the industries picked up in Section 4.3 as well as highlighted in Figure 16 are boldfaced, and the numbers in parentheses indicate the ranking of industries in terms of the degree of agglomeration under the corresponding scalar indices.

[Table 2 about here.]

6 Shipment Distances and Spatial Patterns of Agglomeration

Finally in this section, we investigate the predictions of NEG models regarding the relationship between shipment distances and spatial scales of agglomeration for individual industries. Here, interactions between global and local dispersion forces play a key role. On the one hand, the spatial dispersion of consumers (driven mainly by land-intensive production together with a general scarcity of usable land) generates a *global dispersion force* in which industries with higher transport costs tend to spread over spatially dispersed local markets (both cities and towns) in order to minimize their transport costs to final markets. In our terminology, industries with more dispersed cluster patterns (higher GE) should thus be those in which firms tend to ship more locally. On the other hand, there are two types of *local dispersion forces* affecting industries. One is a *crowding-out force* due to congestion and local scarcity of land that motivates some firms (and residents) to expand existing clusters, rather than form new clusters (as in the case of global dispersion above).[49] The other is a *filling-in force* that tends to

bias of K-density function at small distance levels pointed out by Nakajima et al. [39], our example in Appendix B is independent of this bias, hence the above argument remains true even after correcting this bias.

[49] In continuous-location NEG models (such as Fujita and Krugman [12]), where land is neither consumed nor used as inputs in cities, each city initially occupies only a single point in space. But as populations grow and congestion externalities increase, mobile agents in such cities have incentives to relocate just outside the city, where they can avoid congestion costs while still enjoying proximity to the city market. In this sense, cities can be said to expand spatially in equilibrium. Similar crowding-out effects are found in the many-region extension of the model by Helpman [18] in which land scarcity is the primary dispersion force (see Akamatsu and Takayama [1] for more detail).

transform collections of distinct clusters into a continuum of clusters (as in the formation of "industrial belts"). This happens for example when firms in footloose industries with relatively lower transport costs are attracted to locations between existing clusters to gain access to markets in more than one cluster.[50] Under both crowding-out and filling-in forces, local dispersion takes place that tends to leave the degree of global dispersion relatively unaffected. In our terminology, one thus expects industries with relatively lower transport costs to exhibit more locally dispersed cluster patterns (higher *LD*) for any given level of *GE*. Such dispersion in turn implies that these industries should also exhibit longer shipment distances to their extended markets.

The theoretical prediction above can be directly tested by using the shipment distances for individual establishments obtained from the 2000 Net Freight Flow Census [32] for Japanese manufacturing industries.[51] One restriction here is that industrial shipment data is only available at the two-digit level of classification. Thus, we must analyze spatial patterns in terms of average *GE* and *LD* values within each two-digit category. In particular, our 155 three-digit industries are grouped into 22 categories at the two-digit level. So by letting I_i denote the set of three-digit industries in each two-digit category, i, we can summarize the spatial pattern for each category $i = 1,\ldots,22$ by its *average global extent*, $\overline{GE}_i \equiv \frac{1}{|I_i|}\sum_{j\in I_i} GE_j$, and *average local density*, $\overline{LD}_i \equiv \frac{1}{|I_i|}\sum_{j\in I_i} LD_j$. While these average spatial patterns, $\left(\overline{GE}_i, \overline{LD}_i : i = 1,\ldots,22\right)$, for categories are far fewer in number than our original 155 spatial patterns for industries, it can be seen from Figure 18 that they continue to be uncorrelated in a manner similar to Figure 6 [with $\rho(\overline{GE}, \overline{LD}) = 0.10$ and a p-value of 0.64 for a two-sided test of ρ significance].[52] So these two average indices continue to provide distinct spatial information.

[Figure 18 about here.]

[50] See Mori [33] for the theoretical mechanism underlying the formation of a continuum of cities.

[51] This micro data is provided by the Ministry of Land, Infrastructure, Transport and Tourism of Japan and has been obtained via RIETI. Since the origins and destinations of shipments can be identified in terms of municipalities, the corresponding shipment distances are computed as shortest-route distances along the road network between municipality centers, as in Section 2.3.

[52] The three-digit industries indicated in Figure 6 belong to the two-digit categories indicated by using the same symbols, except that "livestock products" and "seafood products" both belong to "manufactured food" category.

Given these average indices, if we now let the average shipment distance for establishments in each industry, $j = 1, \ldots, 155$, be denoted by SD_j, then our objective is to relate these indices to the *average shipment distance*, $\overline{SD}_i \equiv \frac{1}{|I_i|} \sum_{j \in I_i} SD_j$, for each two-digit category i by employing a multiple regression approach paralleling expression (20) above. The results of this regression are shown below:

$$\log \overline{SD}_i = \underset{(23.86)}{5.812} - \underset{(-3.40)}{0.491} \log \overline{GE}_i + \underset{(4.11)}{0.529} \log \overline{LD}_i \,, \quad \text{adj.}\, R^2 = 0.496 \,, \qquad (21)$$

where the numbers in the parentheses are t-values, so that all estimated coefficients are seen to be significant at the 1% level.[53] In view of the low correlation between the two explanatory variables, $\overline{GE}$ and $\overline{LD}$, these relations are well approximated by their corresponding simple regressions, which can be shown graphically as in panels (a) and (b) of Figure 19, respectively.[54]

[Figure 19 about here.]

Here we see that larger values of average global extent, $\overline{GE}$, correspond to lower average shipment distances, $\overline{SD}$ – which is consistent with the *global dispersion force* prediction of NEG above. Similarly, larger values of average local density, $\overline{LD}$, correspond to larger average shipment distances, $\overline{SD}$ – which is consistent with the *local dispersion force* prediction of NEG. Thus, to our knowledge, this regression provides the first empirical support for these theoretical predictions of NEG. In fact, these results constitute perhaps the first systematic empirical evidence relating degrees of agglomeration to the spatial extent of industrial transactions.[55]

However, there are at least two caveats in interpreting eq. (21). One is that these relations involve only average values across rather broad two-digit industry categories. Second, even if this same relation were to hold for individual industries, the average shipment distance, SD, for each industry would only be associated with those values of GE and LD *realized in equilibrium*. So no causal

[53] The signs of the estimated coefficients are the same for all $\lambda \geq 0.20$, and they are significant at the 1% level for all $\lambda \in [0.74, 0.99]$ (and at the 5% level for all $\lambda \geq 0.70$).

[54] The simple-regression coefficients are naturally somewhat different, and in this case are -0.334 for $\log \overline{GE}$ in panel (a) and 0.412 for $\log \overline{LD}$ in panel (b).

[55] As a related piece of evidence, Kerr and Kominers [26] have used patent citation data in the US to show that inventors with higher levels of mutual interaction tend to be more spatially concentrated.

inferences can be drawn from these relations, and even the relative magnitudes of estimated coefficients should be interpreted with caution.

7 Concluding Remarks

In this paper we have applied the cluster-detection procedure developed by Mori and Smith [37] to study the agglomeration patterns of manufacturing industries in Japan. In particular, we have proposed a simple classification of pattern types based on a pair of quantitative measures, global extent (*GE*) and local density (*LD*), for distinguishing both the scale and degree of industrial agglomeration derived from the cluster schemes. But the ultimate utility of this approach will of course depend on how it can be applied in practical situations.

As alluded to in the Introduction, these measures can already help to sharpen certain concepts in the literature. For example, the differences between spatial dispersion of manufacturing at high versus low levels of transport costs, as derived in general NEG models, can be characterized in terms of these measures. In particular, the type of dispersion associated with high levels of transport costs ("first-phase" dispersion) can in principle be quantified empirically in terms of large *GE* values and small *LD* values.[56] In contrast, dispersion patterns associated with low levels of transport costs ("second-phase" dispersion) might be quantified in terms of small *GE* values and large *LD* values. Hence, such differences between dispersion patterns might be quantified in terms of directed distances within a given *GE*-*LD* space. In fact, given appropriate historical data on industrial location patterns at various stages of transportation technology, one might even be able to test the significance of such differences.

But it should also be emphasized that these two measures are by no means the only relevant properties of agglomeration patterns that can be quantified. Indeed, our present construction of such patterns in terms of cluster schemes provides a potentially rich spatial data set for studying a wide range of problems. Along these lines, it is appropriate to mention three possible research

[56] Here it should be noted that since firms have no "area" in such continuous models, our present notion of local density is somewhat ambiguous. But given fixed employment levels for industries, the essence of this type of dispersion is that individual clusters become smaller and more scattered throughout the spatial continuum. So local "employment" density decreases under this type of dispersion.

directions involving, respectively, the spacing of clusters within industries and the coordination of clusters between industries.

7.1 Agglomeration Spacing within Industries

Within the NEG, a number of models have been developed to explain the spacing between individual clusters for a given industry (e.g., Krugman [27], Fujita and Krugman [12], Fujita and Mori [15], Fujita et al. [14, Ch.6], Tabuchi et al. [48], Ikeda et al. [22], Akamatsu et al. [2]). From the viewpoint of general equilibrium theory, these models predict whether an agglomeration of industrial firms will be viable at a given location, depending on how other clusters of the same industry (as well as population) are distributed over the location space. In these models, industrial agglomeration is typically induced by demand externalities arising from the interactions between product differentiation, plant-level scale economies and transport costs. In particular, Fujita and Krugman [12] have shown that each agglomeration casts a so-called *agglomeration shadow* in which firms have no incentive to relocate from the existing clusters, since within this "shadow" firms are too close to existing clusters (i.e., competitors) to realize sufficient local monopoly advantages. Hence the presence of such shadows serves to limit the number of viable clusters within each industry. Note also that since the level of internal competition differs between industries (depending on their degree of product differentiation and transport costs), the size of agglomeration shadows should also be industry specific.

But while there has been empirical work to study the spacing between urban centers (e.g., Marshall [31, Ch.7], Ioannides and Overman [23], Hsu et al. [21]), to our knowledge there have been no systematic efforts to study the spacing between industrial clusters – and in particular, no efforts to identify the presence of actual agglomeration shadows. However, it should be clear that our present approach to cluster identification offers a promising method for doing so. In particular, since our cluster-detection procedure enables one to identify individual clusters for each industry, it is a simple matter to construct explicit measures of the spacing between them. For example, one natural measure of spacing between clusters in our present framework would be the shortest path distance between their closest basic regions. Agglomeration spacing for cluster schemes as a whole might then be summarized by the mean nearest-neighbor distance be-

tween their constituent clusters. To test whether such spacing is larger (or more uniform) than would be expected by chance alone, one could in principle generate appropriate random versions of cluster schemes to serve as counterfactuals. Such spacing analyses will be reported in subsequent work.

7.2 Agglomeration Coordination between Industries

Within the context of Christaller's [5] celebrated theory of *Central Places*, a topic of major interest has long been the spatial coordination of locations across industries. In particular, the "Hierarchy Principle" underlying this theory asserts that the set of industries found in smaller metro areas is always a subset of those found in larger metro areas.[57] Theoretical efforts to explain this phenomenon have focused mainly on the role of demand externalities in determining industrial locations (see Quinzii and Thisse [41], Fujita et al. [13], Tabuchi and Thisse [47] and Hsu [20]). In particular, the types of demand externalities which induce industrial agglomeration are often shared by many different industries, so that their spatial markets overlap. In such cases, it is natural for these industries to co-locate. Moreover, in terms of market sizes, it is also natural for clusters in more concentrated industries (with larger markets) to coincide with those of less concentrated industries (with smaller markets), thus leading to the type of synchronization predicted by the Hierarchy Principle.

But while these theoretical arguments are quite plausible, there has been surprisingly little work done to actually test the empirical validity of the Hierarchy Principle itself.[58] In fact, the detailed spatial structure of cluster schemes permits direct comparisons of spatial coordination between individual industries. In particular, by associating larger market sizes with smaller numbers of clusters for an industry,[59] one may ask whether industries with larger market sizes do in fact tend to coordinate their spatial locations with industries having smaller market sizes. More specifically one may ask whether their cluster schemes are

[57] Obviously, this principle implicitly assumes a certain degree of industry aggregation, since it could not hold if industries are fully disaggregated, i.e., where each industry consists of one establishment.

[58] One approach proposed by Mori and Smith [36] focuses on the hierarchical industrial structure of cities implied by this principle. In particular, the present cluster-detection procedure was used to identify the industrial composition of each city in terms of the set of industries whose clusters overlap with this city.

[59] In fact this relationship underlies the results in the theoretical papers above.

closer to those of industries with smaller market sizes than would be expected by chance alone. By again measuring "closeness" in terms of (shortest-path) nearest-neighbor distances, one could test this hypothesis in a manner similar to Section 7.1 above. Note that this test can also be interpreted as the test of *co-localization* among different industries, which could in principle provide an alternative approach to those of Ellison et al. [11] and Duranton and Overman [9, §6]. Such investigations will be reported in subsequent work.

7.3 Refining Essential Containments

Because our present spatial measures, GE and LD, are defined solely in terms of area, there of course remains a certain degree of ambiguity regarding spatial *patterns* of agglomeration. This is particularly evident when analyzing the nature of "local dispersion" within e-containments, as illustrated by the two e-containment patterns in Figure 20. While both GE and LD are identical for each pattern, it is evident that "local dispersion" is far more ubiquitous in panel (b) than panel (a). In fact panel (a) might be better described as two major agglomerations of clusters concentrated at opposite ends of this e-containment. So it is important to ask how our present set of measures might be extended to capture such distinctions.

[Figure 20 about here.]

One possibility is suggested by our procedure of building e-containments, where clusters are added until some appropriate BIC threshold is achieved. Having done so, one may continue to combine e-clusters in a pairwise manner that "least detracts" from this threshold value, and then analyze the resulting sequence of decreasing values. For example one would expect that an application of this procedure to panel (a) of Figure 20 would first combine clusters on either end of the e-containment until at some point these two agglomerations of clusters would be joined. At this point, a much larger drop in BIC might be expected, reflecting the "loss of fit" resulting when these two agglomerations are combined. In contrast, panel (b) should be expected to yield a more even sequence of decreases, with no major drops. So by studying these respective patterns of decrease, one might be able to detect major changes in spatial patterns

that represent important intermediate levels of agglomeration structure. Further investigations along these lines will be reported in subsequent work.

A Robustness of Identified Cluster Patterns

In this section, we summarize the findings of two robustness checks on our results. We first consider robustness with respect to MAUP issues in Section A.1 below. This is followed in Section A.2 with robustness checks on the feasible location space for counterfactual establishments.

A.1 Modifiable Areal Unit Problem

Mori and Smith [37] have already shown that the cluster patterns identified by our procedure are robust against small perturbations of municipality boundaries. But, since municipalities tend to be smaller in urban areas than in rural areas, this creates a systematic size bias among municipalities. To determine whether this bias has any critical effect on our results, we here employ an alternative system of equal-sized mesh regions. The mesh regions adopted are "the secondary mesh" used in Japanese Census, where the size of each cell in this mesh is 1/12 of a degree in longitude by 1/10 of a degree in latitude. Since the entire set of 3207 municipalities used in our paper is covered by 4103 mesh cells, these cells are on average about 28% smaller than municipalities.

Using the cluster schemes detected under this mesh-regional system, we repeated each of the analyses in the paper, and found that all qualitative results remain the same.[60] In particular, both GE and LD are highly correlated between mesh-based and municipality-based clusters (with respective Pearson correlations of 0.959 and 0.941, for $\lambda = 0.88$), and the same qualitative conclusions were

[60]To construct appropriate mesh-level data, we have employed micro data from the Establishment and Enterprise Census of Japan [25] as provided by RIETI, together with land utilization segmented mesh data from 2006 as provided by the Ministry of Land, Infrastructure, Transport and Tourism of Japan (http://nlftp.mlit.go.jp/ksj-e/gml/datalist/KsjTmplt-L03-b.html). The representative location for each cell in this secondary-mesh was taken to be the tertiary-mesh cell with largest establishment density in that secondary-mesh cell (or that with the largest percent of economic area if the secondary-mesh cell contained no establishments). The size of each tertiary-mesh cell is 1/100 of a secondary-mesh cell. As in the case of municipalities, inter-mesh road-network distances were computed by using the network analyst extension of ArcMap Ver.10.2.2 of ESRI.

drawn with respect to the scalar-measure comparisons developed in Section 5 (for all $\lambda \geq 0.10$). So in this sense, our results appear to be quite robust with respect to possible MAUP issues.

A.2 Counterfactual Location Patterns

The counterfactual location patterns of establishments employed in our cluster analyses are based on the assumption of uniformly distributed locations over "economic area", which is obtained by subtracting forests, lakes, marshes, rivers, and undeveloped areas from the total area in each municipality. However, it might be argued that the feasible area for manufacturing establishments is in reality also restricted by zoning policies. In particular, manufacturing establishments are typically prohibited from locating in agricultural areas (at least in a short-run). Thus, to check for robustness with respect to such restrictions, we excluded all agricultural land from economic area (and thus reduced the feasible location space to less than one-third of the original economic area). Our cluster-detection procedure was then carried out within this more restrictive setting.

Since agriculture is more abundant in rural areas, this modification necessarily leads to higher concentrations of counterfactual locations in urban areas. Thus one might expect these concentrations to understate the significance of clustering in urban areas. But this effect turned out to be very small, and in fact the cluster patterns obtained were very similar to those under the original economic area. In particular, the values of GE and LD across industries were highly correlated between these two specifications of feasible locations (with respective Pearson correlations of 0.925 and 0.929, for $\lambda = 0.88$). On this basis, we conclude that the identified cluster patterns are also quite robust with respect to possible zoning restrictions.

B Underestimation of Local Concentration by Duranton and Overman Index

In this section, we construct a simple example to illustrate why the Γ^* index tends to underestimate local concentrations for globally dispersed and locally sparse industries. Consider an economy consisting of K regions of equal unit area, with

a Port city in region 1 and all other regions forming a linear hinterland as shown in Figure 21. For additional simplicity we assume that all bilateral distances between distinct regions j and $j+m$ are of length m, and are of length $\varepsilon < 1$ within each region .

[Figure 21 about here.]

Suppose that all industries except one are concentrated in region 1 around the Port city [reflecting possible export orientation of these industries, as well possible *co-location* tendencies (as mentioned in Section 5)]. The exceptional industry, say *industry* i, is assumed to be more locally oriented and is located in a central region, k, of the hinterland as well as in region 1. If industry i has n establishments in both of these regions and if the *aggregate industry* consists of N establishments (with $N \gg 2n$), then by construction there are a total of $N-n$ establishments in region 1 and n establishments in region k, as shown in the figure (with shades of gray reflecting relative establishment concentrations).

Industry i exemplifies the type of industry that is spread out to serve local markets, and is thus more globally dispersed and locally sparse than the other industries. To state this more precisely within our present framework, we take the sets of n establishments for industry i in regions 1 and k to constitute the two "essential" clusters for i (in the sense of Section 3.1), so that the *essential containment* for this industry includes regions 1 through k. Similarly, the essential containment for each other industry j is taken to consist only of region 1. Since the area of the full set of regions is K, it then follows that the *global extent* and *local density* for industry i are given respectively by $GE_i = k/K$ and $LD_i = 2/k$. Similarly for each other industry, j, it follows that $GE_j = 1/K$ and $LD_j = 1$. Thus, by assuming that $k > 2$, we see that both $GE_i > GE_j$ and $LD_i < LD_j$ for all $j \neq i$, and may conclude that industry i is indeed "globally dispersed and locally sparse" relative to all other industries.

The object of this example is then to show that this type of local concentration by industry i cannot be detected by Γ^*. To do so, we first note that there are only two possible bilateral distances, namely ε and $k-1$ (where the latter is between regions 1 and k, and where by assumption $k-1 > 1 > \varepsilon$). Moreover, we shall assume that the bandwidth, h, in expression (16) is sufficiently small [$h < (k-1)-\varepsilon$] to ensure that the kernel densities, $f((d-d_{ij})/h)$, only register bilateral *d-frequencies* (i.e., $f = 0$ unless $d_{ij} = d$). In this setting it is clear that $\widehat{K}_i(\varepsilon)$ in (16)

is proportional to the total ε-frequencies in its two regions, i.e., $\widehat{K}_i(\varepsilon) = c\, n(n-1)$ with proportionality constant, $c = [2n(2n-1)h]^{-1}$. But for any sample, say s, of $2n$ establishments from the aggregate industry, this sample almost always contains more establishments in the *same* region (namely region 1), and thus yields higher ε-frequencies. In particular, one may verify that $c\, n(n-1) \leq \widehat{K}_s(\varepsilon) \leq c\, n(2n-1)$, so that $\widehat{K}_s(\varepsilon) \geq \widehat{K}_i(\varepsilon)$ with strict inequality holding in all but the "$s = i$" case. Thus (in a manner similar to the pair of globally dispersed and locally sparse industries shown in Figure 17), $\widehat{K}_i(\varepsilon)$ for industry i is far below the median value of $\widehat{K}_s(\varepsilon)$, and certainly exhibits no local concentration at distance ε. So as suggested in Section 5, the main factors contributing to this failure of Γ^* to detect local concentration in the present example are (i) the high concentration of small distances (ε) in the aggregate industry, resulting at least partially from *co-location* of industries, and (ii) the inclusion of bilateral distances ($k-1$) between establishments in *different* clusters, such as those in regions 1 and k for industry i.

References

[1] ——— and Yuki Takayama. 2014. "Do polycentric patterns emerge in NEG models?" Unpublished manuscript. Graduate School of Information Sciences, Tohoku University.

[2] ———, Yuki Takayama and Kiyohiro Ikeda. 2012. "Spatial discounting, fourier, and racetrack economy: A recipe for the analysis of spatial agglomeration models." *Journal of Economic Dynamics & Control* 36: 1729-1759.

[3] Behrens, Kristian. 2007. "On the location and lock-in of cities: Geography vs transport technology." *Journal of Urban Economics* 37(1): 22-45.

[4] Brülhart, Marius and Rolf Traeger. 2005. "An account of geographic concentration patterns in Europe." *Regional Science and Urban Economics* 35(6): 597-624.

[5] Christaller, William. 1933. *Die Zentralen Orte in Suddeutschland* (Jena, Germany: Gustav Fischer). English translation by Baskin, Carlisle W. (1966), *Central Places in Southern Germany* (London: Prentice Hall).

[6] Combes, Pierre-Phillippe and Miren Lafourcade. 2005. "Transport costs: measures, determinants, and regional policy implications for France." *Journal of Economic Geography* 5(3): 319-349.

[7] Combes, Pierre-Phillippe, Mayer, Thierry, Thisse, Jaques-Francǫis. 2008. *Economic Geography: The Integration of Regions and Nations* (Princeton, NJ: Princeton University Press).

[8] Davis, Donald R. and Jonathan I. Dingel. 2013. "The comparative advantage of cities." Discussion paper, Columbia University.

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

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

[11] ———, ——— and William R. Kerr. 2010. "What causes industry agglomeration? Evidence from coagglomeration patterns." *American Economic Review* 100(3): 1195-1213.

[12] Fujita, Masahisa and Paul R. Krugman. 1995. "When is the economy monocentric?: von Thüen and Chamberlin unified." *Regional Science and Urban Economics* 25: 505-528.

[13] ———, Paul R. Krugman and Tomoya Mori. 1999. "On the evolution of hierarchical urban systems," *European Economic Review* 43: 209-251.

[14] ———, ——— and Anthony J. Venables. 1999. *The Spatial Economy: Cities, Regions and International Trade* (Cambridge, MA: The MIT Press).

[15] ——— and Tomoya Mori. 1997. "Structural stability and evolution of urban systems." *Regional Science and Urban Economics* 27: 399-442.

[16] ———, ———. 2005. "Frontiers of the new economic geography." *Papers in Regional Science* 84(3): 307-405.

[17] ———, ———. 2005. "Transport development and the evolution of economic geography." *Portuguese Economic Journal* 4(2): 129-156.

[18] Helpman, Elhanan. 1998. "The size of regions." In: *Topics in Public Economics: Theoretical and Applied Analysis* (Cambridge: Cambridge University Press): 33-54.

[19] Hokkaido-chizu, Co. Ltd. 2002. *GIS Map for Road*.

[20] Hsu, Wen-Tai. 2012. "Central place theory and the city size distribution." *Economic Journal* 122: 903-932.

[21] ———, Tomoya Mori and Tony E. Smith. 2014. "Spatial patterns and size distributions of cities." Discussion paper No.882. Institute of Economic Research, Kyoto University.

[22] Ikeda, Kiyohiro, Takashi Akamatsu, and Tatsuhito Kono. 2012. "Spatial period doubling agglomeration of a core-periphery model with a system of cities." *Journal of Economic Dynamics & Control* 36: 743-778.

[23] Ioannides, Yannis M. and Henry G. Overman. 2004. "Spatial evolution of the US urban system." *Journal of Economic Geography* 4(2): 131-156.

[24] Japan Statistics Bureau. 2000. *Population Census of Japan* (in Japanese).

[25] ———. 2001. *Establishments and Enterprise Census of Japan* (in Japanese).

[26] Kerr, William R. and Scott Duke Kominers. 2014. "Agglomerative forces and cluster shapes." *Review of Economics and Statistics*, forthcoming.

[27] Krugman, Paul R. 1993. "On the number and location of cities." *European Economic Review* 37: 293-298.

[28] Kullback, Solomon and Richard A. Leibler. 1951. "On information and sufficiency." *Annals of Mathematical Statistics* 22: 79-86.

[29] Limão, Nuno and Anthony J. Venables. 2001. “Infrastructure, geographical disadvantage, transport costs and trade.” *World Bank Economic Review* 15: 451-479.

[30] Marcon, Eric, Stéphane Traissac, Florence Puech and Gabriel Lang. 2013. “dbmss: Distance-based measures of spatial structures.” http://cran.r-project.org/web/packages/dbmss/index.html.

[31] Marshall, John U. 1989. *The Structure of Urban Systems* (Toronto: University of Toronto Press).polis formatin: the maturing of city systems,” *Journal of Urban Economics* 42: 133-157.

[32] Ministry of Land, Infrastructure, Transport and Tourism. 2000. *Net Freight Flow Census*.

[33] Mori, Tomoya. 1997. “A modeling of megalopolis formation: The maturing of city systems.” *Journal of Urban Economics* 42: 133-157.

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

[35] ———, Koji Nishikimi and Tony E. Smith. 2008. “The number-average size rule: A new empirical relationship between industrial location and city size.” *Journal of Regional Science* 48: 165-211.

[36] ——— and Tony E. Smith. 2011. “An industrial agglomeration approach to central place and city size regularities.” *Journal of Regional Science* 51(4): 694-731.

[37] ———, ———. 2014. “A probabilistic modeling approach to the detection of industrial agglomeration.” *Journal of Economic Geography* 14(3): 547-588.

[38] Murata, Yasusada and Jaques-Francǫis Thisse. 2005. “A simple model of economic geography à la Helpman-Tabuchi.” *Journal of Urban Economics* 58(1): 137-155.

[39] Murata, Yasusada, Ryo Nakajima and Ryuichi Tamura. 2014. “Testing for localization using micro-geographic data: A new approach.” Unpublished manuscript.

[40] Özak, Ömer. 2010. “The voyage of homo-œconomicus: Some economic measure of distance.” Unpublished manuscript. Department of Economics, Brown University.

[41] Quinzii, Martine and Jacques-François Thisse. 1990. “On the optimality of central places.” *Econometrica* 58: 1101-1119.

[42] Rosenthal, Stuart S. and William C. Strangelliam. 2004. “Evidence on the nature and sources of agglomeration economies.” In Henderson, J. Vernon and Thisse, Jacques-François (eds.), *Handbook of Regional and Urban Economics*, Vol.4 (Amsterdam: North-Holland), Ch.49.

[43] Silverman, Bernard W. 1986. *Density Estimation for Statistics and Data Analysis*. New York: Chapman and Hall.

[44] Statistical Information Institute for Consulting and Analysis. 2002. *Toukei de Miru Shi-Ku-Cho-Son no Sugata* (in Japanese).

[45] ———. 2003. *Toukei de Miru Shi-Ku-Cho-Son no Sugata* (in Japanese).

[46] Tabuchi, Takatoshi. 1998. “Urban agglomeration and dispersion: A synthesis of Alonso and Krugman.” *Journal of Urban Economics* 44(3): 333-351.

[47] ——— and Jacques-François Thisse. 2006. “Regional specialization, urban hierarchy, and commuting costs.” *International Economic Review* 47: 1295-1317.

[48] ———, ——— and Dao-Zhi Zeng. 2005. “On the number and size of cities.” *Journal of Economic Geography* 5(4): 423-448.