KIER DISCUSSION PAPER SERIES

KYOTO INSTITUTE
OF
ECONOMIC RESEARCH

"A Probabilistic Modeling Approach to the Detection of
Industrial Agglomerations"

Tomoya Mori and Tony E. Smith

KYOTO UNIVERSITY

KYOTO, JAPAN

A Probabilistic Modeling Approach to the Detection of Industrial Agglomerations

Tomoya Mori*and Tony E. Smith†

September 9, 2009

Abstract

Dating from the seminal work of Ellison and Glaeser [17] in 1997, a wealth of evidence for the ubiquity of industrial agglomerations has been published. However, most of these results are based on analyses of single (scalar) indices of agglomeration. Hence it is not surprising that industries deemed to be similar by such indices can often exhibit very different patterns of agglomeration – with respect to the number, size, and spatial extent of individual agglomerations. The purpose of this paper is thus to propose a more detailed spatial analysis of agglomeration in terms of multiple-cluster patterns, where each cluster represents a (roughly) convex set of contiguous regions within which the density of establishments is relatively uniform. The key idea is to develop a simple probability model of multiple clusters, called *cluster schemes*, and then to seek a "best" cluster scheme for each industry by employing a standard model-selection criterion. Our ultimate objective is to provide a richer characterization of spatial agglomeration patterns that will allow more meaningful comparisons of these patterns across industries.

JEL Classifications : C49, L60, R12, R14

Keywords : Industrial Agglomeration, Cluster Analysis, Geodesic Convexity, Bayesian Information Criterion.

---

*Institute of Economic Research, Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501 Japan. Email: mori@kier.kyoto-u.ac.jp. Phone: +81-75-753-7121. Fax: +81-75-753-7198.

†Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA. Email: tesmith@seas.upenn.edu. Phone: +1-215-898-9647. Fax: +1-215-898-5020.

1 Introduction

Economic agglomeration is the single most dominant feature of industrial location patterns throughout the modern world. In Japan, with a population density more than ten times that of the US, land is generally considered to be extremely scarce. Yet, 65% of the total population and 86% of total employment are concentrated in so-called *densely inhabited districts* accounting for only 10% of total economic area (3% of total area).[1] Essentially similar observations can be made for any other developed country.[2] The extent of this concentration phenomenon explains why economic agglomeration is now a major area of research in urban and regional economics. This is underscored by the fact that the majority of material in the latest Handbook of Regional and Urban Economics [31] is devoted to this topic. This handbook also indicates that economic agglomeration plays a key role in a broader range of fields including economic growth, international trade and economic development. Industrial agglomeration has also gained increasing interest in the management literature, dating from the seminal work of Porter [51] on "industrial cluster theory."

In terms of empirical work, a substantial number of industrial agglomeration studies have been published during the last decade. Some of these studies have provided indices of industrial agglomeration that allow testable comparisons of the degree of agglomeration among industries (Duranton and Overman [15], Brülhart and Traeger [6], and Mori, Nishikimi and Smith [47]). The results of these works suggest that industrial agglomeration is far more ubiquitous than previously believed, and extends well beyond the traditional types of industrial agglomeration (such as information technology industries in Silicon Valley[3] and automobile manufacturing in Detroit). Moreover, the degree of such agglomeration has been shown to vary widely across industries.

But while these studies provide ample evidence for the ubiquity of industrial agglomerations, they tell us very little about the actual *spatial structure* of agglomerations. In particular (to our knowledge), there have been no systematic efforts to determine the number, location and spatial extent of agglomerations within individual industries. Most indices of agglomeration currently in use measure the discrepancy between industry-specific regional distributions of establishments/employment and some hypothetical reference distribution representing "complete dispersion."[4] But even if industries are judged to be similar with respect to these indices, their spatial patterns of agglomeration may appear to be quite different. Such patterns are basically *multidimensional* in nature, and are not easily compared by any single index.

Historically, these scalar indices have been largely motivated by simple two-region models

---

[1]Data source: Population Census of Japan [35] for employment and population data, and Statistical Information Institute for Consulting and Analysis [60, 61] for economic area data. For a definition of economic area see Section 7.1.2 below.

[2]In France, the Île-de-France (metropolitan area of Paris), produces 30% of total GDP while accounting for only 2.2% of the area of France and 18.9% of its population. Even within the Île-de-France, only 12% of the available land is used for urban purposes, and the remaining area is devoted to agriculture, or is undeveloped (Fujita [19]). In the US, 75% of the population is concentrated in 2% of the land area (Rosenthal and Strange[54]).

[3]See for example the well-known study by Saxenian [55].

[4]Examples of such reference distributions are (1) the regional distribution of all-industry employment, used by Ellison and Glaeser [17], (2) the regional distribution all-industry establishments, used by Duranton and Overman [15], and (3) the regional distribution of economic area used by Mori et al. [47].

1

of industrial location, where "agglomeration" is typically the extreme case of complete concentration in one of the two regions, and "dispersion" is the other extreme involving uniform spread across both regions. However, this simple dichotomy has been called into question by the results of the "new economic geography"[5] where industrial location is modeled in continuous space.[6] Here it has been shown that the spatial structure of agglomeration and dispersion can change at different scales of analysis. In particular, such variations in spatial structure arise from specific types of interactions among plant-level increasing returns, product differentiation and transport costs. But within a two-region world, the relative spatial scales of agglomeration and dispersion cannot be distinguished. Hence it is difficult to extend the results of these highly aggregated models to more complex disaggregated regional systems.[7]

However, it is shown in the present paper that this diversity of patterns can in fact be identified empirically. This can be illustrated by a brief preview of our results for Japanese manufacturing industries (developed in more detail in Section 7 below). First, there are industries which clearly exhibit strong spatial concentration, such as the "plastic compounds and reclaimed plastics" industry shown in Figure 7.13(b) [For now, the area marked in grey can be considered as industrial agglomerations.] While some establishments of this industry are attracted to port cities along the northern coast, the main industrial concentration lies along the inland Industrial Belt extending westward from Tokyo to Hiroshima. 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). But, even much more dispersed industries often form small agglomerations at local scales. For example, the "livestock products" industry shown in Figure 7.6(b) is spread throughout the nation, but exhibits a large number of local agglomerations. We describe this type of spatial pattern as "globally dispersed" and "locally sparse." In addition to these extremes, a variety of other patterns can be identified, as discussed more fully in Sections 6 and 7.3 below. Finally, it is important to emphasize that the range of patterns identified here actually bears a close relation to those identified in the new economic geography.[8]

However, it should also be stressed that the continuous-space models of the new economic geography have thus far been limited to one-dimensional worlds, or at best, very stylized two-dimensional worlds that can be modelled in tractable ways. Hence the strategy of the present paper is to start from the empirical side, and to develop statistical cluster models of industrial location patterns that are sufficiently rich to allow a broader range of comparisons between different industries. The immediate goal of this research is to apply these statistical tools to identify such patterns. But the longer range goal is to identify structural properties of location patterns that may contribute to our theoretical understanding of location behavior in more general spatial settings.

It should also be noted that there have been other attempts to develop statistical measures

---

[5]See, e.g., Fujita, Krugman and Venables [22] and Combes, Mayer and Thisse [10] for an overview of the literature.
[6]See Fujita and Mori [24, 25] for a survey.
[7]See again the discussion in Fujita and Mori [24, 25].
[8]See Fujita and Mori [23] and Fujita, Krugman and Mori [21] for the case of "globally dispersed" and "locally sparse" agglomeration patterns, and Mori [46] for the case of "globally confined" and "locally dense" agglomeration patterns.

2

that are more multidimensional in nature. Most notably, the *K*-density approach of Duranton and Overman [15] utilizes pairwise distances between individual establishments, and is capable of indicating the spatial extent of an agglomeration. In a similar vein, Mori et al. [47] proposed a spatially decomposable index of regional localization that yields some information about the most relevant geographic scales of agglomeration within individual industries. However, neither of these approaches is designed to identify specific (map) locations of industrial agglomerations, from which spatial patterns of agglomerations can be characterized.

Methodologically, our approach is closely related to cluster-identification methods proposed by Besag and Newell [4], Kulldorff and Nagarwalla [43], and Kulldorff [42], that have been used for the detection of disease clusters in epidemiology.[9] As with the agglomeration indices mentioned above, these methods start by postulating a null hypothesis of "no clustering" (in terms of a uniform distribution of industrial locations across regions), and then seek to test this hypothesis by finding a single "most significant" cluster of regions with respect to this hypothesis. Candidate clusters are typically defined to be approximately circular areas containing all regions with centroids within some specified distance from a reference point (which may be the centroid of a "central" region). While this approach is in principle extendable to multiple clusters by recursion (i.e., by removing the cluster found, and repeating the procedure) such extensions are piecemeal at best.[10]

Hence our central objective is to generalize their approach by finding the single most significant "cluster scheme" rather than "cluster." We do so by formalizing these schemes as probability models to which appropriate statistical model-selection criteria can be applied for finding a "best cluster scheme." Here a *cluster scheme* is simply a partition of space in which it is postulated that firms are more likely to locate in "cluster" partitions than elsewhere.[11] Our probability model then amounts to a multinomial sampling model on this partition.[12]

---

[9]While "agglomeration" can in principle be viewed as a special type of "clustering," we shall use these two term interchangeably throughout the analysis to follow. (However, see the discussion in Section 8.1 of the Concluding Remarks.)

[10]In particular, the recursive application of such procedures gives rise to the notorius "multiple testing" problem that these procedures were originally designed to overcome. In essence, multiple applications of this procedure will tend to identify too many clusters as being significant. For a recent discussion of this "false discovery" problem in the context of spatial clustering, see Castro and Singer [7] together with the references cited therein.

[11]An alternative approach might be to characterize spatial distributions of establishments by smooth surfaces, utilizing recent advances in density estimation methods (e.g., Silverman [59]). However, our present discrete characterization of agglomerations in terms of spatially disjoint clusters was motivated by the following two considerations. First, an examination of the data shows that spatial distributions of industrial establishments are typically spiky, i.e., concentrations take place in a small set of municipalities. Indeed, there are usually a large number of municipalities with no establishments whatsoever. In our present study of the 163 3-digit manufacturing industries in Japan (Section 7 below), the average percent of all 3207 municipalities in Japan having any establishments in a given industry was only 22.6%. Moreover, 89.5% of these 163 industries have establishments in fewer than one half of all municipalities. Our second motivation for the present discrete approach is the observation that a certain percent of the land area in most regions is unsuitable for industrial location (such as woods, lakes, and marshes). While such constraints are difficult to capture with continuous densities, they can be easily handled within the present discrete framework. For example, to construct uniform distributions for testing null hypotheses of "no clusters," it is a simple matter to replace the total area of each region by its total feasible area, designated here as its "economic" area.

[12]It should be noted that other probability models of multiple clusters have been proposed in the literature. The most well-known of these is the model-based formulation of Dasgupta and Raferty [11] in which multiple clusters are modeled as Bayesian mixture distributions. An alternative Bayesian model which is closer in spirit to the present approach is that of Gangnon and Clayton [26, 27]. Here multiple clusters are modeled as a hierachical Poisson process with gamma priors on cluster intensities. However the present approach is much simpler, and in our view, is

These candidate cluster schemes can then be compared by means of standard model-selection criteria such as *Akaike's* [1] *Information Criterion* (*AIC*), *Normalized Maximum Likelihood* (*NML*) by Konkatnen and Myllymäki [38], and Schwarz's [56] *Bayesian Information Criterion* (*BIC*).

To find a best model (cluster scheme) with respect to these criteria, it would of course be ideal to compare all possible cluster schemes constructible from the given system of regions. But even for modest numbers of regions this is a practical impossibility. Hence a second major objective of this paper is to develop a reasonable algorithm for searching the space of possible cluster schemes. Our approach here is essentially an elaboration of the basic ideas proposed by Besag and Newell [4] in which one starts with an individual region and then adds contiguous regions within a given distance from this initial region to identify the single most significant cluster. Here we find it useful to extend the Besag-Newell concept of clusters by introducing a more flexible class of spatially coherent sets which we designate as *convex solids*. The relevant notion of "convexity" for our purposes is based on minimal travel distances between regional centers (rather than straight-line distances) and hence is somewhat more meaningful economically. This particular cluster definition is useful for growing larger clusters, since arbitrary sets can be "convexly solidified" in a natural way.

In this context, cluster schemes are grown by (*i*) adding new disjoint clusters, or by (*ii*) either expanding or combining existing clusters until no further improvement in the given model-selection criterion is possible. The final result is thus a "locally best cluster scheme" with respect to this criterion. While the criteria listed above are conceptually quite different, it turns out that the locally best cluster schemes found are in high agreement across different criteria. Thus, in the present paper, we will focus on *BIC*, which turns out to be the most parsimonious criterion in terms of the number of clusters found.[13]

The paper is organized as follows. We begin in Section 2 by defining a probabilistic location model for an establishment, where location probabilities are assumed to be industry-specific, and independent for each establishment within a given industry as well as across industries. In Section 3, we briefly develop the three standard model-selection criteria mentioned above. In Section 4, we introduce the notion of convex solids, and then in Section 5, present a practical procedure for cluster detection which searches for the best cluster scheme consisting of a set of distinct "convex" clusters. In the context of this cluster detection framework, we also introduce in Section 6 the notion of "global extent" and "local density" of clusters in order to quantify the spatial scale of industrial agglomeration and dispersion. This procedure is applied in Section 7 to the case of Japanese manufacturing industries and, as previewed above, is illustrated by typical cluster patterns corresponding to theoretical patterns derived in the new economic geography. Finally in Section 8, we briefly discuss a number of directions for further research.

---

more appropriate for the analysis of industrial agglomeration.

[13]A more detailed theoretical comparison of all these criteria will be given in Smith and Mori [62].

2   A Probability Model of Agglomeration Patterns

We start by assuming that the location behavior of individual establishments in a given industry can be treated as independent random samples from an unknown industry-specific *locational probability distribution*, P, over a continuous *location space*, Ω (which represents, for example, a national location space). Hence for any (measurable) subregion, $S \subseteq \Omega$, the probability that a randomly sampled establishment locates in S is denoted by $P(S)$. In this context, the class of all possible location models corresponds to the set of probability measures on Ω.

However, observable location data is here assumed to be only in terms of establishment counts for each of a set of disjoint *basic regions*, $\Omega_r \subseteq \Omega$, indexed by $R = \{1, \ldots, k_R\}$.[14] These regions are assumed to partition Ω, so that

$$\bigcup_{r=1}^{k_R} \Omega_r = \Omega \tag{1}$$

Hence the only relevant features of the location probability distribution, P, for our purposes are the location probabilities for each basic region:

$$P = [P(r) \equiv P(\Omega_r) : r \in R] \tag{2}$$

We now consider an approximation of P by probability models, $P_{\mathbf{C}}$, that postulate areas of relatively intense locational activity. Each model is characterized by a "cluster scheme," $\mathbf{C}$, consisting of disjoint *clusters* of basic regions, $C_j \subset R$, $j = 1, \ldots, k_{\mathbf{C}}$, within which locational activity is postulated to be more intense. For the present, such clusters are left unspecified. A more detailed model of individual clusters is developed in Section 4 below.[15]

If the full extent of cluster C_j in Ω is denoted by

$$\Omega_{C_j} = \bigcup_{r \in C_j} \Omega_r \ , \ \ j = 1, .., k_{\mathbf{C}} \tag{3}$$

then the corresponding location probabilities

$$p_{\mathbf{C}}(j) \equiv P_{\mathbf{C}} \left(\Omega_{C_j} \right) \ , \ \ j = 1, .., k_{\mathbf{C}} \tag{4}$$

are implicitly taken to define areas of concentration. To complete these probability models, let the set of *residual regions* be denoted by

$$R_0 = R - \bigcup_{j=1}^{k_{\mathbf{C}}} C_j \ , \ \ \Omega_{R_0} = \Omega - \bigcup_{j=1}^{k_{\mathbf{C}}} \Omega_{C_j} \tag{5}$$

with corresponding location probability

$$p_{\mathbf{C}}(0) = P_{\mathbf{C}} \left(\Omega_{R_0} \right) = 1 - \sum_{j=1}^{k_{\mathbf{C}}} p_{\mathbf{C}}(j) \tag{6}$$

---

[14]In our application in Section 7 below, basic regions are municipalities.

[15]In particular, it is implicitly assumed here that the regions $\{\Omega_r : r \in C_j\}$ in each cluster are contiguous, so that Ω_{C_j} [defined in (3) below] is a connected set. This assumption is not crucial for the present section, but will play a central role in the construction of clusters below.

Each *cluster scheme*, $\mathbf{C} = (R_0, C_1, .., C_{k_\mathbf{C}})$, then constitutes a partition of the regional index set, R, and the location probabilities $[p_\mathbf{C}(j) : j = 0, 1, .., k_\mathbf{C}]$ yield a probability distribution on $\mathbf{C}$.[16] Finally, to specify location probabilities for basic regions, it is assumed that within each cluster, C_j, the location behavior of individual establishments is *completely random*.[17] To define "complete randomness" in the present setting, it is important to focus on those locations within each basic region where establishments could potentially locate (excluding bodies of water, etc.) Such locations are here taken to correspond to the *economic area* of each region (as discussed further in Section 7.1.2 below). Hence, if for each basic region $r \in R$, we let a_r denote the (economic) *area* of Ω_r, so that the *total area* of cluster C_j is given by

$$a_{C_j} = \sum_{r \in C_j} a_r \qquad (7)$$

then for each establishment locating in C_j, it is postulated that the conditional probability of locating in basic region, $r \in C_j$, is proportional to the area of region r,[18] i.e., that

$$P_\mathbf{C}(\Omega_r \mid \Omega_{C_j}) = \frac{a_r}{a_{C_j}} \ , \ \ r \in C_j \ , \ \ j = 0, 1, .., k_\mathbf{C} \qquad (8)$$

But since $\Omega_r \subseteq \Omega_{C_j}$ implies that

$$P_\mathbf{C}(\Omega_r | \Omega_{C_j}) = \frac{P_\mathbf{C}(\Omega_r)}{P_\mathbf{C}(\Omega_{C_j})} = \frac{P_\mathbf{C}(r)}{p_\mathbf{C}(j)} \qquad (9)$$

it then follows that for all $r \in R$

$$P_\mathbf{C}(r) = p_\mathbf{C}(j) \frac{a_r}{a_{C_j}} \ , \ \ r \in C_j \qquad (10)$$

Hence for each cluster scheme, $\mathbf{C}$, expression (10) yields a well-defined *cluster probability model*,

$$P_\mathbf{C} = [P_\mathbf{C}(r) : r \in R] \qquad (11)$$

which is comparable with the unknown true model (2). Note moreover that since all area values are known, it follows that for each given cluster scheme, $\mathbf{C} = (R_0, C_1, .., C_{k_C})$, the only unknown parameters are given by the $k_\mathbf{C}$-dimensional vector of *cluster probabilities*, $p_\mathbf{C} = [p_\mathbf{C}(j) : j = 1, .., k_\mathbf{C}]$.[19]

Within this modeling framework, we now consider a sequence of n independent location decisions by individual establishments. For each establishment, $i = 1, .., n$, let the location choice of establishment i be modeled by a random (indicator) vector, $X^{(i)} = \left(X_r^{(i)} : r \in R \right)$, with

---

[16]A more complete definition of cluster schemes is given in Definition 5.1 below.

[17]This implicitly assumes that the regions within a given cluster not only have high densities of establishments but also that these densities are similar. Moreover, since we require (in Section 5 below) that clusters be disjoint, the low-density peripheries of clusters will in many cases be ignored.

[18]In the theory of spatial point processes, this hypothesis is referred to as *complete spatial randomness* (see for example Diggle [12]). See also Section 5.3 below.

[19]Note that $p_\mathbf{C}(0)$ is constructable from $p_\mathbf{C}$ in terms of (6).

$X_r^{(i)} = 1$ if establishment i locates in region r, and $X_r^{(i)} = 0$, otherwise. This set of location decisions is then representable by a random matrix of indicators, $X = (X^{(i)} : i = 1, \ldots, n)$, with the following finite set of possible realizations (*location patterns*):

$$\Delta_R(n) = \left\{ x = (x_r^{(i)} : r \in R, i = 1, .., n) \in \{0, 1\}^{n \times k_R} : \sum_{r \in R} x_r^{(i)} = 1, \ i = 1, .., n \right\} \quad (12)$$

By independence, the probability distribution of X under the unknown true distribution in (2) is given for each location pattern, $x \in \Delta_R(n)$, by

$$P(x) = \prod_{i=1}^{n} \prod_{r \in R} P(r)^{x_r^{(i)}} = \prod_{r \in R} P(r)^{n_r} \quad (13)$$

where

$$n_r = \sum_{i=1}^{n} x_r^{(i)} \quad (14)$$

denotes the total number of estabishments locating in region r [see expression (12)]. Similarly, for each cluster probability model, $P_\mathbf{C}$, the postulated distribution of X is given for each pattern, $x \in \Delta_R(n)$, by:

$$P_\mathbf{C}(x|p_\mathbf{C}) = \prod_{r \in R} P_\mathbf{C}(r)^{n_r} = \prod_{j=0}^{k_C} \prod_{r \in C_j} \left(p_\mathbf{C}(j) \frac{a_r}{a_{C_j}} \right)^{n_r} \quad (15)$$

where the relevant parameter vector, $p_\mathbf{C}$, for each such model has been made explicit. In most contexts, it will turn out that the locational frequencies

$$n_j(x) = \sum_{r \in C_j} n_r, \quad j = 0, 1, .., k_\mathbf{C} \quad (16)$$

are sufficient statistics since by definition

$$P_\mathbf{C}(x|p_\mathbf{C}) = \prod_{j=0}^{k_C} \left[p_\mathbf{C}(j)^{\sum_{r \in C_j} n_r} \prod_{r \in C_j} \left(\frac{a_r}{a_{C_j}} \right)^{n_r} \right] = a_\mathbf{C}(x) \prod_{j=0}^{k_C} p_\mathbf{C}(j)^{n_j(x)} \quad (17)$$

where the factor

$$a_\mathbf{C}(x) = \prod_{j=0}^{k_C} \prod_{r \in C_j} \left(\frac{a_r}{a_{C_j}} \right)^{n_r} \quad (18)$$

is completely independent of parameter vector, $p_\mathbf{C}$.

3 Model Selection Criteria

Within this general framework, there are at least three standard approaches to evaluating the relative adequacy of alternative cluster probability models, designated respectively as the *Akaike Information Criterion* (*AIC*), the *Bayes Information Criterion* (*BIC*), and the *Normalized Maximum Likelihood* (*NML*) criterion. The *AIC* and *BIC* measures were introduced, respectively, by Akaike [1] and Schwarz [56]. The *NML* criterion was first introduced by Shtarkov [58], but was given its strongest optimality characterization by Rissanen [53].

Each of these three criteria essentially involves a trade-off between "model fit" and "model complexity." Here the measure of model fit is the same in all cases, namely log likelihood. So to state these criteria explicitly, it is natural to begin with log likelihood. By (17) above, it follows that for any given cluster scheme, $\mathbf{C}$, the *log likelihood* of parameter vector, $p_{\mathbf{C}}$, given an observed location pattern, x, is of the form

$$L(p_{\mathbf{C}}|x) = \sum_{j=0}^{k_{\mathbf{C}}} n_j(x) \ln p_{\mathbf{C}}(j) \ + \ \ln a_{\mathbf{C}}(x) \tag{19}$$

But since the second term is independent of $p_{\mathbf{C}}$, it follows at once (by differentiation) that the *maximum-likelihood estimate*, $\widehat{p}_{\mathbf{C}} = [\widehat{p}_{\mathbf{C}}(j) : j = 1, .., k_{\mathbf{C}}]$, of $p_{\mathbf{C}}$ is given for each $j = 1, \ldots, k_{\mathbf{C}}$ by

$$\widehat{p}_{\mathbf{C}}(j) = \frac{n_j(x)}{n} \tag{20}$$

Hence, by substituting (20) into (19) we obtain a corresponding estimate of the *maximum log-likelihood value* for model $P_{\mathbf{C}}$,

$$L_{\mathbf{C}}(x) = L(\widehat{p}_{\mathbf{C}}|x) = \sum_{j=0}^{k_{\mathbf{C}}} n_j(x) \ln \left(\frac{n_j(x)}{n} \right) \ + \ \ln a_{\mathbf{C}}(x) \tag{21}$$

It is this value that constitutes the common measure of *model fit* in most model-selection criteria.

The main difficultly with this concept is that (like the R-squared measure of model fit in regression) maximum log-likelihood must by definition *increase* as more clusters (i.e., parameters) are introduced.[20] Hence the "best" cluster scheme with respect to model fit alone is the completely disaggregated scheme in which every basic region constitutes its own cluster. To avoid this obvious "over fitting" problem, each of the model selection criteria above (*AIC,BIC,NML*) subtracts a penalty term from (21) which effectively penalizes models with larger numbers of clusters. Hence each criterion involves a trade-off between *model fit* and *model complexity*, and is of the general form

$$Crit_{\mathbf{C}}(x) = L_{\mathbf{C}}(x) - Pen_{\mathbf{C}}(k_{\mathbf{C}}, n) \tag{22}$$

where $Pen(k_{\mathbf{C}}, n)$ is a penalty function that increases with $k_{\mathbf{C}}$ and possibly with n as well. So in all cases those models, $\mathbf{C}$, with *higher* values of the criterion measure, $Crit_{\mathbf{C}}(\cdot)$, are taken to be "better" models.[21] In terms of (22), *AIC* is the simplest criterion, and is defined by:

$$AIC_{\mathbf{C}}(x) = L_{\mathbf{C}}(x) - k_{\mathbf{C}} \tag{23}$$

But in spite of its simplicity, the actual derivation of this measure is quite complex. Akaike [1] derived this measure as an asymptotic approximation to the Kullback-Leibler [41] distance

---

[20]To be more precise, maximum log-likelihood can never decrease as more clusters are added, and will almost always increase.

[21]Note that since these criterion values will almost always be negative, it is also very common to reverse signs, and seek models with mimumum (nonnegative) values of $-Crit$, rather than maximum values of $Crit$. In this form, the *negative* log-likelihood, $-L_{\mathbf{C}}(x)$, is often referred to a *lack-of-fit* measure. Note also for the *AIC* and *BIC* criteria defined below, we have (for simplicity) dropped the factor 2 which is often present in these measures.

from $P_\mathbf{C}$ in (15) to the true distribution P in (13).[22] In contrast to *AIC*, the *BIC* measure depends explicitly on sample size, n, and is given by

$$BIC_\mathbf{C}(x) = L_\mathbf{C}(x) - \frac{k_\mathbf{C}}{2} \ln(n) \tag{24}$$

Again, the derivation is complex and, as the name suggests, is Bayesian in origin. In particular, Schwarz [56] derived this measure as an asymptotic approximation to Bayes factors, which constitute the most fundamental Bayesian criterion for model selection.[23] It should also be noted that unlike *AIC*, this measure focuses directly on comparisons between candidate models (cluster schemes) rather than comparisons with the "true" model (which need not even exist in this framework).

The final measure, *NML*, offers a more intuitive model selection criteria in that it involves no asymptotic approximations whatsoever. Rather it penalizes models directly in terms of relative likelihoods. Recall that for any observed location pattern, x, and candidate cluster schemes, $\mathbf{C}$ and $\mathbf{C}'$, the likelihood ratio, $P(x|\widehat{p}_\mathbf{C}(x))/P(x|\widehat{p}_{\mathbf{C}'}(x))$, can be viewed as the reflecting the evidence in favor of $\mathbf{C}$ versus $\mathbf{C}'$ given data x. In a similar manner, for any possible location patterns x and y of size n, one can view the likelihood ratio, $P(x|\widehat{p}_\mathbf{C}(x))/P(y|\widehat{p}_\mathbf{C}(y))$, as reflecting the relative evidence for scheme $\mathbf{C}$ in pattern x versus pattern y. Hence if one considers all possible pattern realizations, $y \in \triangle_R(n)$, of size n [as in expression (12) above], then it is natural to regard the ratio,

$$\frac{P(x|\widehat{p}_\mathbf{C}(x))}{\sum_{y \in \triangle_R(n)} P(y|\widehat{p}_\mathbf{C}(y))} \tag{25}$$

as reflecting the relative evidence in pattern x for cluster scheme $\mathbf{C}$ compared to *all* other possible patterns of size n. To gain further insight into (25), observe that schemes $\mathbf{C}$ with large numbers of clusters will tend to fit any location pattern reasonably well, and hence will assign high likelihoods, $P(y|\widehat{p}_\mathbf{C}(y))$, to a large set of patterns, y. So even if pattern x has a high likelihood, this need not provide strong evidence unless it is high *relative* to other pattern likelihoods for scheme $\mathbf{C}$. This is precisely the *normalized likelihood* approach to model selection. To put this criterion in the form of expression (22), simply take logs and use the identity, $L_\mathbf{C}(x) = \ln P(x|\widehat{p}_\mathbf{C}(x))$ to rewrite (25) as

$$NML_\mathbf{C}(x) = L_\mathbf{C}(x) - \ln\left[\sum_{y \in \triangle_R(n)} P(y|\widehat{p}_\mathbf{C}(y))\right] \tag{26}$$

where the penalty term in this case is given by

$$Pen_\mathbf{C}(k_\mathbf{C}, n) = \ln\left[\sum_{y \in \triangle_R(n)} P(y|\widehat{p}_\mathbf{C}(y))\right] \tag{27}$$

and is seen to depend explicitly on the structure of cluster scheme, $\mathbf{C}$, as well as parameters $k_\mathbf{C}$ and n.

---

[22]The clearest derivation of this results is given in Burnham and Anderson [5, Section 7.2].

[23]Again, a simple derivation can be found in Burnham and Anderson [5, Section 6.4.1]. It is also worth noting that in spite of its Bayesian origins, this asymptotic approximation is entirely independent of the prior distributions used in Bayes models.

To compare these criteria it is convenient to start with *AIC* and *BIC*, which are by far the most popular for model selection – due in large part to their simplicity and ease of calculation. While there is no general consensus as to which is better,[24] it is clear from (26) and (27) that even for moderate sample sizes ($n \geq 8$) *BIC* always penalizes larger cluster schemes more heavily than does *AIC*. Hence *BIC* is more *parsimonious* than *AIC* in that it tends to favor smaller cluster schemes. This is especially evident when there are many establishments, n. Since the penalty term in *AIC* is independent of sample size, it follows that log likelihood, $L_{\mathbf{C}}(x)$, will be the dominant term for large n, and hence that *AIC* will suffer from the same "over-fitting" problem as likelihood itself.

Turning next to *NML*, it should be clear that while this measure is conceptually the simplest of the three, its penalty term is by far the most complex. Indeed, the exact calculation of this penalty term remains largely an open problem.[25] Hence the asymptotic approximation of *NML* established by Rissanen [52] is used in almost all applications. In our case, this approximation takes the form

$$\ln\left[\sum_{y\in\triangle_R(n)} P(y|\widehat{p}_{\mathbf{C}}(y))\right] = \frac{k_{\mathbf{C}}}{2}\ln(n) + \left[\ln\left(\frac{\pi^{(k_{\mathbf{C}}+1)/2}}{\Gamma[(k_{\mathbf{C}}+1)/2]}\right) - \frac{k_{\mathbf{C}}}{2}\ln(2\pi)\right] + o(1) \quad (28)$$

where the first term on the right hand side is precisely the *BIC* penalty term. Equally important is the fact that even for small numbers of clusters ($k_{\mathbf{C}} \geq 2$) the second term is *negative*, and that for large n this term is completely dominated in size by the first term (since the second term is independent of n). Hence it follows from (24) and (26) that for $k_C \geq 2$, *BIC* is *more parsimonious* than *NML* (i.e., penalizes large clusters more heavily), but is eventually *indistinguishable* from *NML* as n becomes large.[26] Hence from a practical viewpoint it is natural to regard *BIC* as a simple and easily computable approximation to the more intuitive *NML* criteria.

For these reasons, we shall focus entirely on *BIC* in the applications given here. More systematic comparisons of these criteria in terms of simulated firm location patterns will be given in Smith and Mori [62].[27]

4   A Model of Clusters as Convex Solids

Given the set of basic regions, R, it would in principle seem desirable to treat cluster schemes, $\mathbf{C}$, as arbitrary partitions of R, and then to identify the *best cluster scheme* from this class, i.e.,

$$\mathbf{C}^* = \arg\max_{\mathbf{C}} BIC_{\mathbf{C}} \quad (29)$$

---

[24] However, from a *consistency* viewpoint, it has been shown by Haughton [30] that for a fixed model size ($k_{\mathbf{C}}$ in our case) *BIC* will eventually choose the correct model with probability one as n becomes large. But in the present setting, it is not clear that the assumption of fixed model size is viable: As the number of firms, n, increases it is reasonable to expect that larger numbers of clusters, $k_{\mathbf{C}}$, will become more likely. Under these conditions, the results of Shibata [57] for regression variable selection suggest that *AIC* may actually have better consistency properties than *BIC*.

[25] Fortunately for our present multinomial model, a very efficient algorithm for calculating (27) has been developed by Kontkanen and Myllymäki [38] (see also Kontkanen et al. [39]).

[26] A sharper comparison will be given in Smith and Mori [62].

[27] In addition to the three criteria above, the comparison in Smith and Mori [62] will also include standard *likelihood-ratio tests*, which constitute meaningful model selection criteria within the present nested-model framework.

But from a practical viewpoint, the number of possible partitions can be enormous for even modest numbers of basic regions.[28] Moreover, without further restrictions, the components of such partitions can be quite bizarre, and difficult to interpret as "clusters." This has long been recognized by cluster analysts, who have typically proposed that clusters be roughly circular in shape (as in Besag and Newell [4], Kulldorff and Nagarwalla [43], and Kulldorff [42]). Hence our first objective is to develop a more flexible class of candidate clusters, designated as convex solids, which requires approximate convexity on cluster shape. To this end, we begin by representing our regional system in terms of a discrete network over the set of basic regions on which these convex solids are defined.

4.1 A Discrete Network Representation of the Regional System

Recall in Section 2 that the relevant location space, Ω, is partitioned into a set of basic regions, $\Omega_r \subseteq \Omega$, indexed by $R = \{1, .., k_R\}$. For our present purposes it is convenient to consider a larger *world region*, W, in which Ω resides, so that $W - \Omega$ denotes the "rest of the world," as shown schematically in Figure 4.1 below. As in Section 2 we identify Ω with the set of regional labels for R. In this framework, the *boundary* of the given location space consists of the subset of basic regions, $\overline{R}$, that share boundary points with $W - \Omega$ (where "boundary points" correspond to the edges of each basic region cell in the figure[29]). This distinguished set of boundary regions (shown in gray) will play an important role in Section 4.3 below.

<div align="center">

Figure 4.1 here

</div>

Within this basic continuous geographical framework, we next develop a discrete network representation of the regional system that contains all relevant information needed for our cluster model. 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 road network. Here it is assumed that data is available on minimal *travel distances*, $t(r, s)$, between each pair of regions, $r, s \in R$, say between their designated administrative centers.[30] These neighbors should of course include regional pairs (r, s) for which the shortest route from r to s passes through no regions other than r and s. But for computational convenience, we choose to approximate this relation by the standard "contiguity" relation that takes each pair of basic regions sharing some common boundary to be neighbors.[31] While this approximation is reasonable in most cases, there are exceptions. Consider for example the coastal regions, r and s, joined by a bridge, as shown in Figure 4.2 below. Here it is clear that the shortest route

---

[28]In our Japanese data, the number of basic regions is over 3000.

[29]More generally, a *boundary point* of Ω is any point $\omega \in \Omega$ for which there exist points outside of Ω that are arbitrarily close to ω (in Euclidean distance). We suppress topological details here in order to avoid confusion with similar graph-theoretic topological concepts to be developed below.

[30]In the application below (Section 7) for the case of Japan, we use road-network distances as travel distances between municipality offices.

[31]In the terminology introduced by Cliff and Ord [9] these are known as "queen" contiguities, rather than "rook" contiguities, where only regions sharing a full boundary face are considered neighbors. Such contiguity relations are easily calculated in most standard Geographical Information System (GIS) software.

(path) between regions r and s passes through no other regions, even though r and s share no common boundary. Hence to maintain a reasonable notion of "closeness" among neighbors, it is appropriate to include such regional pairs as neighbors. Finally, it is mathematically convenient to include r as a neighbor of itself (since r is always "closer" to itself than to any other region).

Figure 4.2 here

If this set of *neighbors* for region $r \in R$ is denoted by $N(r)$, then for the region r shown in the schematic regional system of Figure 4.1, $N(r)$ is seen to consist of eight neighbors other than r itself. Our only formal requirement is that neighbors be symmetric, i.e., that $r \in N(s)$ if and only if $s \in N(r)$. If we now denote the full set of neighbor pairs by

$$L = \bigcup_{r \in R} \bigcup_{s \in N(r)} (r,s) \subseteq R^2 \tag{30}$$

then this defines the relevant set of *links* for our discrete network representation, (R, L), of the regional system.[32] A simple example of such a regional network, (R, L), is shown in Figure 4.3 below. Here R consists of twenty five square regions shown on the left. These regions are connected by the road network shown by dotted lines on the left, with travel distances on each of the forty links (to be discussed later) displayed on the right. Hence L in this case consists of the forty distinct regional pairs associated with each of these links, together with the twenty five identity pairs (r,r).

Figure 4.3 here

Next we employ travel distances between neighbors to approximate the entire road network by a shortest-path metric on network (R, L). To do so, we note that minimum travel distances naturally satisfy the metric conditions (*i*) $t(r,r) = 0$, and (*ii*) $t(r,s) = t(s,r)$, for all neighbor pairs (r,s). In addition, for every triad of mutual neighbors, $r, v, s \in R$ [i.e., with $r \in N(s)$ and $v \in N(r) \cup N(s)$] these distances must also satisfy the metric triangle-inequality condition (*iii*) $t(r,s) \leq t(r,v) + t(v,s)$.[33] Given these metric conditions, one can extend t to a shortest-path metric on (R, L) in the following way. Let each sequence, $\rho = (r_1, r_2, .., r_n)$, of linked neighbors [i.e., with $(r_i, r_{i+1}) \in L$ for $i = 1, .., n-1$] be designated as a *path* in (R, L), and let the set of all paths in (R, L) be denoted by

$$\mathcal{P} = \{\rho = (r_1, .., r_n) : n > 1, \ (r_i, r_{i+1}) \in L, \ i = 1, .., n-1\} \tag{31}$$

---

[32]Equivalently, the network (R, L) can be viewed as a *graph* with *vertices*, R, and *edges*, L. Note also that both L and the individual neighborhoods, $N(r)$, depend on travel distance, t. But for notational simplicity we leave this dependency implicit.

[33]Since travel from r to s can always be accomplished by taking shortest routes from r to v, and then for v to s, it must be true that the minimum travel distance, $t(r,s)$ cannot exceed the combined distance, $t(r,v) + t(v,s)$, of these two trips.

12

If for each pair of regions, $r, s \in R$, we denote the subset of all paths from r to s in $\mathcal{P}$ by

$$\mathcal{P}(r,s) = \{\rho = (r_1, .., r_n) \in \mathcal{P} : r_1 = r, r_n = s\} \quad (32)$$

then to ensure that shortest paths between all pairs of regions are meaningful, we henceforth assume that that $\mathcal{P}(r,s) \neq \varnothing$ for all $r, s \in R$, i.e., that the given regional network (R, L) is *connected*.[34] In this context, if the *length*, $l(\rho)$, of path, $\rho = (r_1, r_2, .., r_n)$, is now taken to be the sum of travel distances on each of its links, i.e.,

$$l(\rho) = \sum_{i=1}^{n-1} t(r_i, r_{i+1}) \quad (33)$$

then for any pair of regions, $r, s \in R$, the *shortest-path distance*, $d(r,s)$, from r to s is taken to be the length of the (possibly nonunique) shortest path from r to s:

$$d(r,s) = \min\{l(\rho) : \rho \in \mathcal{P}(r,s)\} \quad (34)$$

The set of all shortest paths in $\mathcal{P}(r,s)$ (also called "geodesics" from r to s) is then denoted by

$$\mathcal{P}_d(r,s) = \{\rho \in \mathcal{P}(r,s) : l(\rho) = d(r,s)\} \quad (35)$$

The shortest-path distances in (34) are then easily seen to define a *metric* on R, i.e., to satisfy (i) $d(r,r) = 0$, (ii) $d(r,s) = d(s,r)$, and (iii) $d(r,s) \leq d(r,v) + d(v,s)$ for all $r, s, v \in R$.[35] Moreover, these distances always agree with travel distances between neighbors [i.e., $d(r,s) = t(r,s)$ for all $(r,s) \in L$], but for non-neighbors, $(r,s) \notin L$, it will generally be true that $d(r,s) > t(r,s)$ (since the shortest route from r to s on the actual road network may not pass through any intermediate regional centers). Hence these shortest-path distances are only an approximation to shortest-route distances. The advantage of this approximation for our present purposes is that for any r and s, the number of paths in $\mathcal{P}(r,s)$ is generally much smaller than the number of routes from r to s on the road network, so that shortest paths in $\mathcal{P}_d(r,s)$ are more easily identified.

4.2 Convexity in Networks

Within this network framework we now return to the question of defining candidate clusters as spatially coherent groups of basic regions. As mentioned in the Introduction, the standard approach to this problem is to require that clusters be as close to "circular" as possible. To broaden this class, we begin by observing that a key property of circular sets in the plane is their convexity. More generally, a set, S, in the plane is *convex* if and only if for every pair of points, $s, v \in S$, the set S also contains the line segment joining s and v. But since lines are shortest paths with respect to Euclidean distance, an equivalent definition of convexity would be to say that S contains all shortest paths between points in S. Since shortest paths are equally well defined for

---

[34]See however the discussion in Section 7.1.1 regarding major off-shore islands.
[35]As in footnote 33 above, the triangle inequality follows directly from the additivity of path lengths together with the fact that any path from r to v to s is necessarily a path from r to s.

the network model above, it then follows that we can identify convex sets in the same way.

In particular, a set of basic regions, S, is now said to be *d-convex* if and only if for every pair of regions r and s in S, the set of regions on every shortest path from r to s is also in S.[36] More formally, if for any path, $\rho = (r_1,..,r_n) \in \mathcal{P}$, we now denote the *set* of distinct points in ρ by $\langle \rho \rangle = \{r_1,..,r_n\} \subseteq R$, and if the family of all nonempty subsets of R is denoted by $\mathcal{R} = \{S \subseteq R : S \neq \varnothing\}$, then

Definition 4.1 (*d*-Convexity) (*i*) *A subset of basic regions, $S \subseteq R$, is said to be d-convex iff for all $s, r \in S$,*

$$\rho \in \mathcal{P}_d(r,s) \Rightarrow \langle \rho \rangle \subseteq S \tag{36}$$

(*ii*) *The family of all d-convex sets in $\mathcal{R}$ is denoted by $\mathcal{R}_d$.*

For example, suppose that in the schematic regional system of Figure 4.4 below it is assumed that regional squares sharing boundary points (faces or corners) are always neighbors, and that travel distance, t, between neighbors is simply the Euclidean distance between their centers. Then with respect to the induced shortest-path distance, d, it is clear that the set, S, on the left consisting of four black squares is not d-convex, since the gray squares in the middle figure belong to shortest paths between the black squares. But even if these gray squares are added to S, the resulting set is still not d-convex because the four white squares remaining in the middle belong to shortest paths between the gray squares. However, if these four squares are added, then the resulting set on the right is seen to be d-convex since all squares on every shortest path between squares in the set are already included.

Figure 4.4 here

This process of adding shortest paths actually yields a well-defined constructive procedure for "convexifying" a given set, which can be formalized as follows. Define the (*r*, *s*)-*interval*, $I(r,s)$, to be the set of all points on shortest paths from r to s, i.e.,

$$I(r,s) = \bigcup_{\rho \in \mathcal{P}_d(r,s)} \langle \rho \rangle \tag{37}$$

and let the mapping, $I : \mathcal{R} \to \mathcal{R}$, defined for all $S \in \mathcal{R}$ by

$$I(S) = \bigcup_{r,s \in S} I(r,s) \tag{38}$$

be designated as the *interval function* generated by d. For notational convenience, we set $I^0(S) = S$, $I^1(S) = I(S)$, and construct the mth-*iterate* of I recursively by $I^m(S) = I(I^{m-1}(S))$ for

---

[36]Our present notion of d-convexity is an instance of the more general notion of geodesic convexity applied to graphs, and appears to have first been introduced by Soltan [63]. For more explicit minimal-path (geodesic) treatments of d-convexity, see for example Farber and Jamison [18] and Duchet [14] .

all $m > 1$ and $S \in \mathcal{R}$. Since $\{r, s\} \subseteq I(r, s)$ for all $r, s \in R$, it follows from (38) that for each set, $S \in \mathcal{R}$,

$$S \subseteq I(S) \tag{39}$$

By the same argument, it follows that for any $S \in \mathcal{R}$ and $r \in I^m(S)$ with $m > 0$, we must have $r \in I[I^m(S)] = I^{m+1}(S)$. Hence these interval iterates satisfy the following nesting property for all $S \in \mathcal{R}$,

$$I^m(S) \subseteq I^{m+1}(S), \quad m \geq 0 \tag{40}$$

and thus constitute a *monotone nondecreasing* sequence of sets. It then follows that for any subset, $S \subseteq R$, of nodes in the *finite* network, (R, L), there must be an integer, $m \ (\leq |R - S|)$,[37] such that $I^m(S) = I^{m+1}(S)$.[38] The smallest such integer:

$$m(S) = \min\{m : I^m(S) = I^{m+1}(S)\} \tag{41}$$

is called the *geodesic iteration number of set, S*.[39] With these definitions, it is well known that the unique smallest d-convex set containing a given set $S \in R$ is given by the *d-convex hull*,[40]

$$c_d(S) = I^{m(S)}(S) \tag{42}$$

The mapping, $c_d : \mathcal{R} \rightarrow \mathcal{R}$, defined by (42) is designated as the *d-convexification function*. With this definition, it is shown in Proposition A.3 of the Appendix that d-convex sets are equivalently characterized as the *fixed points* of this mapping, i.e, a set $S \in \mathcal{R}$ is d-convex if and only if $c_d(S) = S$. So the family of all d-convex sets can be equivalently defined as

$$\mathcal{R}_d = \{S \in \mathcal{R} : c_d(S) = S\} \tag{43}$$

However, for purposes of *constructing* d-convex sets, it is more useful to note that they are equivalently characterized as the fixed points of the *interval function*, $I : \mathcal{R} \rightarrow \mathcal{R}$ (as shown in the Corollary to Proposition A.3). Hence $\mathcal{R}_d$ can also be written as

$$\mathcal{R}_d = \{S \in \mathcal{R} : I(S) = S\} \tag{44}$$

This in turn implies that a simple constructive algorithm for obtaining $c_d(S)$ is to iterate I until the iteration number, $m(S)$, is found. This procedure is in fact illustrated by Figure 4.4 above, where $m(S) = 2$.

But while this particular set, $I^2(S)$, does indeed look reasonably compact (and close to circular), this is not always the case. One simple counterexample is shown in Figure 4.5 below.

---

[37]Throughout this paper we denote *cardinality* of a set A by $|A|$.

[38]Since $I^m(S) \neq I^{m+1}(S)$ implies from (40) that $|I^{m+1}(S) - I^m(S)| \geq 1$, and since $I^m(S) \subseteq R$ for all m, it follows that this expansion process can involve at most $|R - S|$ steps.

[39]This concept was first introduced by Harary and Neiminen [29], who showed that without further assumptions, the bound $m(S) \leq |R - S|$ cannot be significantly reduced. However, in our present application this iteration number is typically small.

[40]For a proof of this assertion, see Proposition A.2 in the Appendix. For further properties of interval functions and d-convex hulls, see for example Duchet [14].

Given the regional network, (R, L), in Figure 4.3 above, suppose that S consists of the four regions shown in black on the left in Figure 4.5. These regions are assumed to be connected by major highways as shown by the heavy lines on the right in Figure 4.3, with travel distances, $t = 1$, on each link. All other road links are assumed to be circuitous secondary roads, as represented by a travel distance of $t = 3$ on each link. Here it is clear that the d-convexification, $c_d(S)$, of S is obtained by adding all other regions connected by the ring of major highways (as shown in gray on the right in Figure 4.5), since shortest paths between such regions are always on these highways. But since the central region shown in white is not on any of these paths, we see that $c_d(S)$ is a d-convex set with a "hole" in the middle.

Figure 4.5 here

This is very different from convex sets in the plane, which are always "solid." But in more general metric spaces this need not be true. Indeed, for the present case of a network (or graph) structure, the notion of a "hole" itself is not even meaningful. For example, if the central node in Figure 4.5 were pulled "outside" the coastal regions (leaving all links in tact) then the *network*, (R, L), would remain the same. So it is clear that the above notion of a "hole" depends on additional spatial structure, including the *positions* of regions relative to one another.

4.3   Convex Solids in Networks

These observations motivate the spatial structure that we now impose in order to characterize "solid" subsets of R in (R, L). The key idea here is to recall from Figure 4.1 that relative to the rest of the world, there is a distinguished collection of *boundary regions*, $\overline{R}$, that are essentially "external" to all subsets of R. If for any subset, $S \subseteq R$, and boundary region, $\bar{r} \in \overline{R}$, it is true that $\bar{r} \notin S$, then it is reasonable to assert that $\bar{r}$ is *outside* of S.[41] This set of boundary regions, $\overline{R}$, thus define a natural reference set for distinguishing regions in complement, $R - S$, of S that are "inside" or "outside" of S. In particular, we now say that a complementary region, $r \in R - S$, is *inside S* if and only if every path joining r to a boundary region in $\overline{R}$ must pass through at least one region of S. For example, given the set, S, of black squares in Figure 4.6, the complementary region r is seen to be inside of S since every path to the boundary, $\overline{R}$, must intersect S. Similarly, the complementary region s is not inside S, since there is a path from s to $\overline{R}$ that does not intersect S. To formalize this concept, we now let the set of all paths from any region, $r \in R$, to $\overline{R}$ be denoted by

$$\mathcal{P}(r, \overline{R}) = \bigcup\nolimits_{\bar{r} \in \overline{R}} \mathcal{P}(r, \bar{r}) \tag{45}$$

Then for any nonempty set, $S \in \mathcal{R}$, the set of all complementary regions *inside S* is given by,

$$S_0 = \{r \in R - S : \rho \in \mathcal{P}(r, \overline{R}) \Rightarrow \langle \rho \rangle \cap S \neq \varnothing\} \tag{46}$$

---

[41]Even if $\bar{r}$ is an element of S, it must always be part of the boundary of S. Hence it is still reasonable to assert that $\bar{r}$ is "on the outside" of S.

and is designated as the *interior complement* of *S*.

Figure 4.6 here

With this concept, we now say that a set, $S \in \mathcal{R}$, is *solid* if and only if its interior complement is empty. In addition, we can now *solidify* a set *S* by simply adjoining its interior complement. More formally, we now say that:

Definition 4.2 (Solidity) *For any nonempty subset, $S \in \mathcal{R}$,*
  (*i*) *S is said to be* solid *iff $S_0 = \varnothing$.*
  (*ii*) *The set formed by adding S_0 to S,*

$$\sigma(S) = S \cup S_0 \tag{47}$$

  is designated as the solidification *of S.*
  (*iii*) *The family of all solid sets in $\mathcal{R}$ is denoted by $\mathcal{R}_\sigma$.*

The justification for the terminology in (*ii*) is given by Lemma A.1 in the Appendix, where it is shown that for any set, $S \in \mathcal{R}$, the set, $\sigma(S)$, is solid in the sense of (*i*) above. The mapping, $\sigma : \mathcal{R} \rightarrow \mathcal{R}$, induced by (47) is designated as the *solidification function*. As with the *d*-convexification function above, it also follows that solid sets are precisely the fixed points of the solidification function.[42]

With these definitions, the two properties of *d*-convexity and solidity are taken to constitute our desired model of clusters in *R*. Hence we now combine these properties as follows:

Definition 4.3 (d-Convex Solids) *For any nonempty subset, $S \in \mathcal{R}$,*
  (*i*) *If S is both d-convex and solid, then S is designated as a d-convex solid in $\mathcal{R}$.*
  (*ii*) *The composite image set,*

$$\sigma c_d(S) = \sigma[c_d(S)] \tag{48}$$

  is designated as the d-convex solidification of S.

If we now let $\mathcal{R}_{\sigma d}$ denote the family of all *d*-convex solids in $\mathcal{R}$, then it follows at once from Definitions 4.1 through 4.3 that

$$\mathcal{R}_{\sigma d} = \mathcal{R}_\sigma \cap \mathcal{R}_d \tag{49}$$

---

[42]See Lemma A.2 in the Appendix.

17

4.4 Convex Solidification of Sets

As with (46) and (47) above, expression (48) induces a composite mapping, $\sigma c_d : \mathcal{R} \to \mathcal{R}$, designated as the *d-convex solidification function*. We now examine this function in more detail. To do so, it is instructive to begin by observing that the *order* in which these two maps are composed is critical. In particular it is *not* true that the *d*-convexification of a solid set is necessarily a *d*-convex solid. This can be illustrated by the example in Figures 4.3 and 4.5 above. If the exterior squares are taken to define the relevant boundary set, $\overline{R}$, in Figure 4.3, then it is clear that the original set, S, of four black squares is solid, since there are paths from every complementary region to $\overline{R}$ that do not intersect S.[43] But, the *d*-convexification, $c_d(S)$, of S is precisely the *non-solid* set that was used to motivate solidification. So in this case, the composite image, $c_d[\sigma(S)] = c_d(S)$ is not solid (and hence not a *d*-convex solid).

With this in mind, the key result of this section, established in Theorem A.1 of the Appendix, is to show that the terminology in Definition 4.3 is justified, i.e., that:

Property 4.4 (*d*-Convex Solidification) *For any set, $S \in \mathcal{R}$, the image set, $\sigma c_d(S)$, is a d-convex solid.*

Hence if one is enlarging a given cluster, C, by adding a set, S, of new regions, i.e., $C \to C \cup S$, then to construct a new cluster containing $C \cup S$, one need only *d*-convexify this set by the algorithm

$$
\begin{aligned}
C \cup S &\to I(C \cup S) \\
&\to I^2(C \cup S) \\
&\vdots \\
&\to c_d(C \cup S)
\end{aligned}
\tag{50}
$$

and then solidify the resulting set by identifying all regions in the interior complement $[c_d(C \cup S)]_0$ of $c_d(C \cup S)$ and forming

$$
\sigma c_d(C \cup S) = c_d(C \cup S) \cup [c_d(C \cup S)]_0
\tag{51}
$$

This algorithm has already been illustrated by the simple case in Figure 4.4, where no solidification was required. A somewhat more detailed illustration is given in Figures 4.7 and 4.8 below. Figure 4.7 exhibits a subsystem of nineteen (hexagonal) basic regions in R, along with the major road network (solid and dashed lines) connecting the centers of these regions. As in Figure 4.4, it is assumed that there are primary roads (freeways) and secondary roads. Some regions lie along freeway corridors, as denoted by solid network links with travel distance (or time) values of $t = 1$. Other regions are connected by secondary roads denoted by dashed network links with higher values of $t = 3$.

---

[43]Note also from this example that the notion of "solidity" by itself is rather weak. However, when applied to *d*-convex sets, this turns out to be exaclty what is needed for "filling holes."

Figure 4.7 here

A possible sequence of steps in the formation of a composite cluster in this subsystem is depicted in Figure 4.8. Stage 1 begins at the point where it has been determined that an existing cluster (d-convex solid), C_1, of three regions (shown in black) should be expanded to include a secondary set, S_1, of two regions (also shown in black). Given the shortest-path distances, d, generated by the t-values in Figure 4.7, it is clear that the d-convexification, $c_d(C_1 \cup S_1)$, of this composite set, $C_1 \cup S_1$, is given by adding the gray regions shown in Stage 2. This larger ring of regions lies entirely on freeway corridors, and thus includes all shortest paths joining its members (in a manner similar to the ring of regions in Figure 4.5). Hence the two regions in the center of this ring lie in the internal complement of $c_d(C_1 \cup S_1)$, and are thus added in Stage 3 to form an new cluster (d-convex solid), $C_2 = \sigma c_d(C_1 \cup S_1)$, containing $C_1 \cup S_1$. In Stage 4 it is determined that one additional singleton set, S_2, should also be added to the existing composite cluster, C_2. Again, Stage 5 shows that all regions on the freeway corridors from S_2 to C_2 should be added to form a new d-convexification, $c_d(C_2 \cup S_2)$. Finally, this d-convex set is again seen to have two regions in its interior complement, which are thus added to achieve the final d-convex solid cluster, $C_3 = \sigma c_d(C_2 \cup S_2)$.

Figure 4.8 here

Before proceeding, it is appropriate to note several additional features of this d-convex solidification procedure that parallel the basic procedure of d-convexification itself. First, as a parallel to d-convex hulls in (42), it is shown in Theorem A.3 of the Appendix that for any given set of regions, S, the d-convex solidification, $\sigma c_d(S)$, yields a "best d-convex solid approximation" to S in the sense that:

Property 4.5 (Minimality of d-Convex Solidifications) *For any set, $S \in \mathcal{R}$, the d-convex solidification, $\sigma c_d(S)$, of S is the smallest d-convex solid containing S.*

Hence this process of cluster formation can be regarded as a *smoothing procedure* that approximates each candidate set of high-density regions by a more spatially coherent version of this set.

Next, as a parallel to the fixed-point property of d-convexifications, it is shown in Theorem A.4 of the Appendix that the procedure in (50) and (51) always yields a fixed point of the composite mapping, $\sigma c_d : \mathcal{R} \to \mathcal{R}$:

Property 4.6 (d-Convex Solid Fixed Points) *A set, $S \in \mathcal{R}$, is a d-convex solid if and only if $\sigma c_d(S) = S$.*

19

Hence the family, $\mathcal{R}_{\sigma d}$, of all d-convex solids in (49) can equivalently be written as

$$\mathcal{R}_{\sigma d} = \{S \in \mathcal{R} : \sigma c_d(S) = S\} \tag{52}$$

In this form, each new cluster is seen to be a natural "stopping point" of the combined d-convexification and solidification procedure above.

Finally, it should be noted that while this process of d-convex solidification tends to produce reasonably cohesive clusters in many cases, there are exceptions. For example, as with many spatial constructions, this procedure is prone to "edge effects." In the present case of Japan, where the coastline is often highly irregular, the d-convex solidification of regional groups near the coast can in some cases require the annexation of large vacant regions. More generally, when the entire regional network, (R, L), is itself highly irregular in space, the basic notion of d-convex solids in (R, L) can become somewhat problematic.

5 A Cluster-Detection Procedure

Given the cluster model developed above, the set of relevant cluster schemes for regional network (R, L) can now be formalized as follows:

Definition 5.1 (Cluster Schemes) *A finite partition,* $\mathbf{C} = (R_0, C_1, \ldots, C_{k_\mathbf{C}})$, *of R is designated as a cluster scheme for* (R, L) *iff*
    (i) [d-convex solidity] $C_i \in \mathcal{R}_{\sigma d}$ *for all* $i = 1, .., k_\mathbf{C}$,
    (ii) [disjointness] $C_i \cap C_j = \varnothing$ *for all i, j with* $1 \le i < j$.
Let $\mathcal{C}(R, L)$ *denote the class of admissible cluster schemes for* (R, L).

Below, we develop our search procedure to identify the best cluster scheme. Before developing the details of this procedure, however, it is useful to begin with an overview.

For any given industry, we start with the single best cluster consisting of a single basic region. Then at each subsequent step, we decide whether we should (*i*) stay with the current cluster scheme; (*ii*) expand one of the existing clusters; or (*iii*) start a new cluster. In alternative (*ii*), we compare potential expansions of all the existing clusters. Such expansions involve annexations of nearby regions which are then further enlarged to maintain d-convex solidity. A new cluster in alternative (*iii*) consists of the best basic region in the current set of residual regions, R_0. At each step, the best option among these three is selected, and the system of clusters continues growing until option (*i*) is evaluated as the best among the three.

Before completing the description of this procedure (in Section 5.2), we specify the details of option (*ii*) above in the next section.

5.1 Operational Rules for Cluster Expansion

At each step of the search procedure outlined above, option (*ii*) involves the expansion of an existing cluster by first annexing certain nearby regions and then further enlarging this set to

maintain "spatial cohesiveness." In view of the above definition of a cluster scheme, this requires that such annexations be enlarged so as to maintain both d-convex solidity and disjointness with respect to other existing clusters. This procedure can sometimes require the annexation of other existing clusters, as illustrated by Figure 5.1 below. Given the subsystem of a regional network shown in Figure 4.7 above, suppose that the current cluster scheme includes the clusters C_1 and C_2 shown in Stage 1 of Figure 5.1. Suppose also that it has been determined that the next step of the search procedure should be an expansion of cluster C_1 to include the set Q shown in Stage 1. The composite cluster, $\sigma c_d(C_1 \cup Q)$, resulting from d-convex solidification of $C_1 \cup Q$, includes $C_1 \cup Q$ together with the gray region shown in Stage 2. But since cluster C_2 is seen to overlap this composite cluster, it is clear that disjointness between clusters can only be maintained by annexing cluster C_2 as well. This results in the larger composite cluster, $\sigma c_d[\sigma c_d(C_1 \cup Q) \cup C_2]$, shown by the combined black and gray region of Stage 3 in Figure 5.1.

$$\boxed{\text{Figure 5.1 here}}$$

More generally, if some current cluster, $C_j \in \mathbf{C} = (R_0, C_1, \ldots, C_{k_{\mathbf{C}}})$, is to be expanded by annexing a set $Q \subseteq R_0$, then the d-convex solidification, $\sigma_{C_d}(C_j \cup Q)$, must be further enlarged to include all clusters, $C_i \in \mathbf{C}$, intersecting $\sigma_{C_d}(C_j \cup Q)$. For any given current cluster scheme $\mathbf{C} = (R_0, C_1, \ldots, C_{k_{\mathbf{C}}})$, this procedure can be formalized in terms of the following operator, $U_{\mathbf{C}} : R \to R$, defined for all $S \in R$ by

$$U_{\mathbf{C}}(S) = \sigma c_d(S) \cup \{C_i \in \mathbf{C} : C_i \cap \sigma c_d(S) \neq \varnothing\} \tag{53}$$

where the relevant sets, S, of interest will be of the form, $S = C_j \cup Q$, with $C_j \in \mathbf{C}$ and $Q \subseteq R_0$. Observe next this single operation is not sufficient, since the resulting image sets, $U_{\mathbf{C}}(S)$, may fail to be d-convex solids. Moreover, the d-convex solidification, $\sigma c_d[U_{\mathbf{C}}(S)]$, may again fail to be disjoint from other existing clusters in $\mathbf{C}$. So it should be clear that what is needed here is an iteration of this operator until both conditions are met. To formalize such iterations, we proceed as in Section 4.2 above by letting the iterates of $U_{\mathbf{C}}$ be defined for each $S \in R$ by

$$U_{\mathbf{C}}^0(S) = S, \quad U_{\mathbf{C}}^1(S) = U_{\mathbf{C}}(S), \quad \text{and } U_{\mathbf{C}}^m(S) = U_{\mathbf{C}}[U_{\mathbf{C}}^{m-1}(S)] \quad \text{for all } m > 1 \tag{54}$$

Since it is clear by definition that

$$U_{\mathbf{C}}^m(S) \subseteq U_{\mathbf{C}}^{m+1}(S), \quad m \geq 0 \tag{55}$$

this yields a monotone nondecreasing sequence of sets in R. Hence by the same arguments leading to (41) above, it again follows that there must be an integer, $m \ (\leq |R - S|)$, such that $U_{\mathbf{C}}^m(S) = U_{\mathbf{C}}^{m+1}(S)$. As a parallel to (41) we may thus designate the smallest integer,

$$m(S|\mathbf{C}) = \min\{m : U_{\mathbf{C}}^m(S) = U_{\mathbf{C}}^{m+1}(S)\} \tag{56}$$

21

satisfying this condition as the *expansion iteration number* of S given $\mathbf{C}$. Finally, if (as a parallel to d-convex hulls) we now designate the resulting fixed point of $U_{\mathbf{C}}$,

$$u_{\mathbf{C}}(S) = U_{\mathbf{C}}^{m(S|\mathbf{C})}(S) \tag{57}$$

as the $\mathbf{C}$-*compatible expansion* of S, then it is this set that satisfies the expansion properties we need. First observe that the fixed point property, $U_{\mathbf{C}}[u_{\mathbf{C}}(S)] = u_{\mathbf{C}}(S)$, of this expanded set implies at once from (53) that for all clusters $C_i \in \mathbf{C}$,

$$C_i \cap u_{\mathbf{C}}(S) \neq \varnothing \Rightarrow C_i \subseteq u_{\mathbf{C}}(S) \tag{58}$$

and hence that $u_{\mathbf{C}}(S)$ is always disjoint with any clusters, $C_i \in \mathbf{C}$, that have not already been absorbed into $u_{\mathbf{C}}(S)$. Moreover, this in turn implies from (53) that $u_{\mathbf{C}}(S) = \sigma c_d[u_{\mathbf{C}}(S)]$, and hence that $u_{\mathbf{C}}(S)$ must be a d-convex solid.

5.2 Cluster-Detection Procedure

In terms of Definition 5.1, the objective of this procedure, which we now designate as the *cluster-detection procedure*, is to find a cluster scheme, $\mathbf{C}^* \in \mathcal{C}(R, L)$, satisfying,

$$\mathbf{C}^* = \arg \max_{\mathbf{C} \in \mathcal{C}(R,L)} BIC_{\mathbf{C}} \tag{59}$$

From a practical viewpoint, it should be stressed that the following search procedure will only guarantee that the cluster scheme found is a "local maximum" of (59) with respect to the class of admissible "perturbations" in $\mathcal{C}(R, L)$ defined by the procedure itself.

To specify these perturbations in more detail, we begin with the following notational conventions. At each stage, $t = 0, 1, 2, ...$, of this procedure, let $\mathbf{C}_t = (R_{t,0}, C_{t,1}, ..., C_{t,k_{\mathbf{C}_t}})$, denote the current cluster scheme in $\mathcal{C}(R, L)$. The procedure then starts at stage $t = 0$ with the *null cluster scheme*

$$\mathbf{C}_0 = \{R_{0,0}\} = \{R\} \tag{60}$$

which contains no clusters. By expressions (7), (21) and (24), it then follows that the corresponding initial value of the objective function in (59) must be

$$BIC_{\mathbf{C}_0} = L_0 \equiv \ln \sum_{r \in R} a_r \tag{61}$$

Given data, $[\mathbf{C}_t , BIC_{\mathbf{C}_t}]$, at stage t, we then seek the modification (perturbation), $\mathbf{C}_{t+1}$, of $\mathbf{C}_t$ in $\mathcal{C}(R, L)$ which yields the highest value of $BIC_{\mathbf{C}_{t+1}}$. As outlined above, these modifications are of two types: (i) the formation of a new cluster in scheme $\mathbf{C}_t$, or (ii) the expansion of an existing cluster in scheme $\mathbf{C}_t$. We now develop each of these steps in turn.

22

5.2.1 New Cluster Formation

Given the current cluster scheme, $\mathbf{C}_t = (R_{t,0}, C_{t,1}, \ldots, C_{t,k_{\mathbf{C}_t}})$, at stage t, one can start a new cluster, $\{r\}$, by choosing some residual region, $r \in R_{t,0}$, which is disjoint with all existing clusters. Hence the set of feasible choices for r is given by

$$R_0(\mathbf{C}_t) = R_{t,0} \tag{62}$$

For each $r \in R_0(\mathbf{C}_t)$, the corresponding expanded cluster scheme is then given by

$$\mathbf{C}_t^0(r) = \left(R_{t,0}^0(r), C_{t,1}^0(r), C_{t,2}^0, \ldots, C_{t,k_{\mathbf{C}_t^0(r)}}^0 \right) \tag{63}$$

where

$$k_{\mathbf{C}_t^0(r)} = k_{\mathbf{C}_t} + 1 \tag{64}$$

$$C_{t,1}^0(r) = \{r\} \tag{65}$$

$$C_{t,i}^0 = C_{t,i-1} \quad \text{for } i = 2, \ldots, k_{\mathbf{C}_t^0(r)}, \tag{66}$$

and

$$R_{t,0}^0(r) = R_{t,0} - \{r\} \tag{67}$$

The superscript "0" in cluster scheme, $\mathbf{C}_t^0(r)$, indicates that a change is made to the residual region, $R_{t,0}$, rather than to one of the clusters in $\mathbf{C}_t$. Note that since $\{r\}$ is automatically a d-convex solid, and since $r \in R_0(\mathbf{C}_t)$ guarantees that disjointness of all clusters is maintained, it follows that $\mathbf{C}_t^0(r) \in \mathcal{C}(R, L)$, and hence that $\mathbf{C}_t^0(r)$ is an admissible modification of $\mathbf{C}_t$.

The best candidate for new cluster formation is of course the region, $r_0^* \in R_0(\mathbf{C}_t)$, that yields the highest value of the objective function, i.e., for which

$$r_0^* = \arg \max_{r \in R_0(\mathbf{C}_t)} BIC_{\mathbf{C}_t^0(r)} \tag{68}$$

For purposes of comparison with other possible modifications of $\mathbf{C}_t$, we now set

$$\mathbf{C}_t^0 \equiv \mathbf{C}_t^0(r_0^*) \tag{69}$$

5.2.2 Expansion of an Existing Cluster

Next, we consider a potential expansion of each cluster, $C_{t,j} \in \mathbf{C}_t$, by annexing a set Q of nearby regions in $R_{t,0}$. While the basic mechanics of this expansion procedure were developed in Section 5.1 above, the specific choice of Q was not. Recall that such annexations can potentially result in large expansions of $C_{t,j}$, given the need to preserve both d-convex solidity and disjointness. Hence to maintain reasonably "small increments" in our search process, it is appropriate to restrict initial annexations to single regions whenever possible. Of course, when such regions

are already part of another cluster, it will be necessary to annex the whole cluster in order to preserve disjointness. But to motivate our basic approach, it is convenient to start by considering the annexation of a single region not in any other cluster, i.e., to set $Q = \{r\}$ for some $r \in R_{t,0}$. Here it would seem natural to consider only regions in the immediate neighborhood of $C_{t,j}$. However, this often turns out to be too restrictive, since there may exist much better choices that are not direct neighbors of $C_{t,j}$.

In fact, it might seem more reasonable to consider all possible regions in $R - C_{t,j}$, and simply let our model-selection criterion determine the best choice. But if one allows choices of r "far away" from $C_{t,j}$, then our d-convex solidity and disjointness criteria can lead to the formation of very large clusters that violate any notion of spatial cohesiveness.[44] So it is convenient at this point to introduce a new set of neighborhoods which strike a compromise between these two extremes. To do so, we first extend *shortest-path distances*, d, between points to corresponding distances between points and sets by letting

$$d(r, Q) = \min \{ d(r, s) : s \in Q \} \tag{70}$$

for $r \in R$ and $Q \in \mathcal{R}$. Since d is a metric on R, it is well known that for each set, $Q \in \mathcal{R}$, (70) yields a well-defined distance function that preserves the usual continuity properties of d on R.[45] Hence one can define well-behaved neighborhoods of Q in terms of this distance function as follows. For each $Q \in \mathcal{R}$, the δ-*neighborhood* of Q in R is defined to be

$$\delta(Q) = \{ r \in R : d(r, Q) < \delta \} \tag{71}$$

Hence the appropriate choices for expansions of $C_{t,j}$ are taken to be regions in $\delta(C_{t,j})$ for some pre-specified choice of parameter δ.[46]

As mentioned above, there are two cases that need to be distinguished here. First suppose that for some given cluster $C_{t,j}$ we consider the annexation of a region not in any other cluster, i.e., a region $r \in R_{t,0} \cap \delta(C_{t,j})$. Then follows from expression (57) that the corresponding $\mathbf{C}_t$-compatible expansion of $C_{t,j} \cup \{r\}$ is given by

$$C_{t,1}^j(r) = u_{\mathbf{C}_t}(C_{t,j} \cup \{r\}). \tag{72}$$

---

[44]This is particularly evident in our application below, where an unconstrained choice of regions can in some cases lead to the inclusion of regions r separated from $C_{t,j}$ by undeveloped mountain regions, or even the inland sea of Japan. More generally, the inclusion of large less developed regions of the nation can lead to an exaggerated depiction of agglomeration involving areas that are mostly devoid of establishments. It should be noted that this is in part due to our use of economic area (rather than total area), which effectively ignores such undeveloped land when expanding clusters.

[45]See for example in Berge [3, Chapter 5].

[46]In the application below, the value used was $\delta = 36.0$ km, which was chosen so that any single expansion of a cluster cannot include large sections without economic area (e.g., inland sea and lakes). This particular neighborhood size covers about 90% of the shortest-path distances between neighboring jurisdictional offices. It is also worth noting from a practical viewpoint that this use of uniform δ-neighborhoods has the added advantage of controlling (at least in part) for size differences among basic regions.

Thus the cluster scheme, $\mathbf{C}_t^j(r)$, resulting from this expansion has the form

$$\mathbf{C}_t^j(r) = \left(R_{t,0}^j(r), C_{t,1}^j(r), C_{t,2}^j(r), \ldots, C_{t,k_{\mathbf{C}_t^j(r)}}^j(r) \right) \tag{73}$$

where, by expression (53), the set of all other clusters in $\mathbf{C}_t^j(r)$ is given by

$$\left\{ C_{t,2}^j(r), \ldots, C_{t,k_{\mathbf{C}_t^j(r)}}^j(r) \right\} = \left\{ C_{t,i} \in \mathbf{C}_t : C_{t,i} \cap C_{t,1}^j(r) = \varnothing \right\} \tag{74}$$

and where the corresponding residual region has the form:

$$R_{t,0}^j(r) = R - \bigcup_{i=1}^{k_{\mathbf{C}_t^j(r)}} C_{t,i}^j(r) \tag{75}$$

As above, if r_j^* now denotes the region in $R_{t,0} \cap \delta(C_{t,j})$ that yields the highest value of the objective function, i.e., for which

$$r_j^* = \arg \max_{r \in R_{t,0} \cap \delta(C_{t,j})} BIC_{\mathbf{C}_t^j(r)} \tag{76}$$

then the best cluster expansion for $C_{t,j}$ in $\mathbf{C}_t$ starting with regions in $R_{t,0} \cap \delta(C_{t,j})$ is given by $\mathbf{C}_t^j(r_j^*)$.

Next recall that it is possible that another cluster, $C_{t,i}$ in $\mathbf{C}_t$, intersects $\delta(C_{t,j})$ so that the annexation of $C_{t,i}$ is a possible expansion of $C_{t,j}$. For this case it is necessary to annex the entire cluster $C_{t,i}$ in order to preserve disjointness. So if we now define the index set,

$$I_j(\mathbf{C}_t) = \{ i \neq j : C_{t,i} \cap \delta(C_{t,j}) \neq \varnothing \} \tag{77}$$

[not to be confused with interval sets $I(\cdot)$ in Section 4.2 above] and for each $i \in I_j(\mathbf{C}_j)$ replace (72) with the $\mathbf{C}_t$-compatible expansion

$$C_{t,1}^j(i) = u_{\mathbf{C}_t}(C_{t,j} \cup C_{t,i}). \tag{78}$$

then as a parallel to (73) through (75), the cluster scheme, $\mathbf{C}_t^j(i)$, resulting from this expansion now has the form

$$\mathbf{C}_t^j(i) = \left(R_{t,0}^j(i), C_{t,1}^j(i), C_{t,2}^j(i), \ldots, C_{t,k_{\mathbf{C}_t^j(i)}}^j(i) \right) \tag{79}$$

with the set of all other clusters in $\mathbf{C}_t^j(i)$ given by

$$\left\{ C_{t,2}^j(i), \ldots, C_{t,k_{\mathbf{C}_t^j(i)}}^j(i) \right\} = \left\{ C_{t,i} \in \mathbf{C}_t : C_{t,i} \cap C_{t,1}^j(i) = \varnothing \right\} \tag{80}$$

25

and with corresponding residual region:

$$R_{t,0}^j(i) = R - \bigcup_{k=1}^{k_{\mathbf{C}_t^j(i)}} C_{t,k}^j(i) \tag{81}$$

If i_j^* now denotes the cluster in $I_j(\mathbf{C}_t)$ that yields the highest value of the objective function, i.e., for which

$$i_j^* = \arg\max_{i \in I_j(\mathbf{C}_t)} BIC_{\mathbf{C}_t^j(i)} \tag{82}$$

then the best cluster expansion for $C_{t,j}$ in $\mathbf{C}_t$ is given by $\mathbf{C}_t^j(i_j^*)$. Hence the best cluster expansion, $\mathbf{C}_t^j$, of $\mathbf{C}_t$ starting with cluster $C_{t,j}$ is given by

$$\mathbf{C}_t^j \equiv \arg\max_{\mathbf{C} \in \{\mathbf{C}_t^j(r_j^*), \mathbf{C}_t^j(i_j^*)\}} BIC_{\mathbf{C}} \ , \quad j = 1,..,k_{\mathbf{C}_t} \tag{83}$$

5.2.3 Revision of the Cluster Scheme

Finally, given these candidate modifications, $\mathbf{C}_t^0, \mathbf{C}_t^1, \ldots, \mathbf{C}_t^{k_{\mathbf{C}_t}}$, of $\mathbf{C}_t$ in $\mathcal{C}(R, L)$ [as defined by (69) together with (83)], let $\mathbf{C}_t^*$ be the best candidate, as defined by

$$\mathbf{C}_t^* = \arg\max_{\mathbf{C} \in \{\mathbf{C}_t^j : j = 0,1,...,k_{\mathbf{C}_t}\}} BIC_{\mathbf{C}} \tag{84}$$

There are then two possibilities left to consider: If $BIC_{\mathbf{C}_t^*} > BIC_{\mathbf{C}_t}$, then set

$$[\mathbf{C}_{t+1}, BIC_{\mathbf{C}_{t+1}}] = [\mathbf{C}_t^*, BIC_{\mathbf{C}_t^*}] \tag{85}$$

and proceed to stage $t + 1$. On the other hand, if $BIC_{\mathbf{C}_t^*} \leq BIC_{\mathbf{C}_t}$, then no (local) improvement can be made, and the cluster-detection procedure terminates with the (locally) *optimal cluster scheme*:

$$\mathbf{C}^* = \mathbf{C}_t^* \tag{86}$$

Finally, it is of interest to note that this cluster-detection procedure is roughly analogous to "mixed forward search" procedure in stepwise regression, where in the present case we add new clusters or merge existing ones until some locally optimal stopping point is found. With this analogy in mind, it is in principle possible to consider "mixed backward search" procedures as well. For example, one could start with a maximal number of singleton clusters, and proceed by either eliminating or merging clusters until a stopping point is reached. Some experiments with this approach in our application below produced results similar to the present search procedure, but proved to be far more computationally demanding.

5.3 A Test of Spurious Clustering

While the cluster-detection procedure developed above will always find a (locally) best cluster scheme, $\mathbf{C}^*$, with respect to the *BIC* criterion used, there is still a statistical question of whether

such clustering could simply have occurred by chance. Hence one can ask how the optimal criterion value, BIC_{C^*}, obtained compares with typical values obtainable by applying the same cluster-detection procedure to randomly generated spatial data. This can be formalized in terms of the hypothesis of *complete spatial randomness* (see footnote 18), which in this present context asserts that the probability, p_r, that any given establishment will locate in region, $r \in R$, is proportional to the areal size, a_r, of that region, i.e., that

$$p_r = \frac{a_r}{\sum_{r \in R} a_j} \tag{87}$$

While the sampling distribution of BIC_C under this hypothesis is complex, it can easily be estimated by Monte Carlo simulation. More precisely, for any given industrial location pattern of n establishments, one can use (87) to generate, say, 1000 random location patterns of n establishments, and apply the cluster-detection procedure to each pattern. This will yield 1000 values of BIC_C, say $BIC_1, \ldots, BIC_{1000}$. If the value for the actual cluster scheme, $BIC_0 = BIC_{C^*}$, is say bigger than all but five of these in the ordering of values, $\{BIC_1, \ldots, BIC_{1000}\}$, then the chance, p, of getting a value as large as this (under the hypothesis that BIC_0 is coming from the same population of random patterns) is, $p = (5 + 1)/(1000 + 1) \sim 0.005$. This would indicate very "significant clustering." On the other hand, if BIC_0 were only bigger than say 800 of these values, then the p-value, $p = (200 + 1)/(1000 + 1) \sim 0.20$, would suggest that the observed cluster scheme, C^*, is not sufficiently significant to warrant further investigation (as discussed further in Section 7.2 below).

6 Measures for Classifying Agglomeration Patterns

As emphasized in the Introduction, the main strength of our cluster detection approach is to indentify cluster schemes in a manner that preserves the two-dimensional spatial aspects of agglomerations. By so doing, it is possible to consider the spatial patterns of industrial agglomerations themselves. As we will see for the case of Japanese manufacturing industries in Section 7 below, agglomerations of given industries often tend to concentrate within specific subregions of the nation, 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 meaure of the relative size of these e-containments, designated as the *global extent* of the industry. Industries with small global extents can be regarded as relatively "confined," and those with large global extents 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 second measure of the *local density* of agglomerations within each e-containment. This will yield a useful classification of agglomeration patterns ranging from *maximally concentrated* patterns with agglomerations densely distributed in confined e-containments to *minimally concentrated* patterns with agglomerations sparsely distributed in dispersed e-containments.

6.1 Essential Containment

To make these ideas precise, we start by defining the essential containment for a given industry, where it is assumed that an optimal cluster scheme, $\mathbf{C}$, has been indentified for the industry. 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 basic regions in R. To identify a set of "most significant" clusters, we proceed recursively by successively adding those clusters in $\mathbf{C}$ with maximum incremental contributions to *BIC*.[47] 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 $\mathbf{C}^+ \equiv \mathbf{C} - \{R_0\}$, 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 $R_{0,0}(C) = R_{0,0} - C$. 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

$$\mathbf{C}_t \equiv \{R_{t,0}, C_{t,1}, \ldots, C_{t,t}\} \tag{88}$$

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 $C \in \mathbf{C}^+ - \mathbf{C}_{t-1}^+$, let $\mathbf{C}_{t-1}(C)$ be defined by,

$$\mathbf{C}_{t-1}(C) = (R_{t-1,0}(C), C_{t-1,1}, \ldots, C_{t-1,t-1}, C) \tag{89}$$

where

$$R_{t-1,0}(C) = R_{t-1,0} - C \tag{90}$$

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

$$C_t \equiv \arg \max_{C \in \mathbf{C}^+ - \mathbf{C}_{t-1}^+} L\left(\widehat{p}_{\mathbf{C}_{t-1}(C)} | \mathbf{C}_{t-1}\right) \tag{91}$$

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 expressions (18) through (21)] by

$$L\left(\widehat{p}_{\mathbf{C}_{t-1}(C)} | \mathbf{C}_{t-1}\right) = \sum_{C' \in \mathbf{C}_{t-1}(C)} n_{C'} \ln\left(\frac{n_{C'}}{n}\right) + \sum_{C' \in \mathbf{C}_{t-1}(C)} \sum_{r \in C'} n_r \ln\left(\frac{a_r}{a_{C'}}\right) \tag{92}$$

---

[47] At this point it should be emphasized that the following procedure for identifying "significant clusters" in $\mathbf{C}$ is different from the recursive scheme used to indentify $\mathbf{C}$ in Section 5.2 above. In particular, the only candidate clusters now being considered are those in $\mathbf{C}$ itself.

with

$$n_{C'} \equiv \sum_{r \in C'} n_r \tag{93}$$

$$n \equiv \sum_{r \in R} n_r \tag{94}$$

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

$$BIC_{\mathbf{C}_t} = L_{\mathbf{C}_t} - \frac{t}{2} \ln(n) \tag{95}$$

where [as a parallel to (92)] we now have

$$L_{\mathbf{C}_t} = \sum_{C \in \mathbf{C}_t} n_C \ln \left(\frac{n_C}{n} \right) + \sum_{C \in \mathbf{C}_t} \sum_{r \in C} n_r \ln \left(\frac{a_r}{a_C} \right) \tag{96}$$

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:

$$\triangle BIC_t \equiv BIC_{\mathbf{C}_t} - BIC_{\mathbf{C}_{t-1}} \tag{97}$$

To identify the relevant set of "significant clusters" in $\mathbf{C}$, it would thus seem 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* $t = 1, .., k_{\mathbf{C}}$. Hence our first requirement for significance of cluster C_t is that it yield a "substantial" increment to *BIC*. One hypothetical illustration with $k_{\mathbf{C}} = 7$ is given in Figure 6.1(a) below, where each successive increment to *BIC* is seen to be positive [and where the values on the horizontal axis can be ignored for the moment]. By the nature of our recursive procedure, it can be expected that the first increment ($t = 1$) will be the largest, and that successive increments will continue to diminish in size.[48] In the example shown, it appears that the increments for $t = 2, 3$ are comparable to $t = 1$, but that there is a noticeable decrease at $t = 4$ and beyond. Hence one simple criteria for a "substantial increment," $\triangle BIC_t$, would be to require that it be at least some specified fraction, μ, of $\triangle BIC_1$.[49] In terms of this criterion, the procedure would stop at the first stage, t^e, where additional increments fail to satisfy this condition, i.e., where $\triangle BIC_{t^e+1} < \mu \triangle BIC_1$.

Figure 6.1 here

But while this *substantial-increment condition* provides a reasonable criterion for identifying the set of most significant clusters with respect to *BIC*, such clusters may in some cases represent

---

[48]This situation is somewhat analogous to successive increments in adjusted R-square resulting from a foward stepwise regression procedure.

[49]The values $\mu = .03$ and $\mu = .05$ were selected for our application in Section 7.3 below..

only a small subset of all clusters in $\mathbf{C}$. More importantly, they may represent only a small portion of all *establishments* in such clusters. Hence, if the "essential containment" for the industry is to include a substantial portion of these *agglomeration establishments*, then it is desirable to impose an additional condition on the stopping rule above. In particular, if the *share* of agglomeration establishments in each cluster scheme, $\mathbf{C}_t$, of expression (88) is denoted by

$$s\left(\mathbf{C}_t\right) = \frac{\sum_{C \in \mathbf{C}_t^+} n_C}{\sum_{C \in \mathbf{C}^+} n_C} \tag{98}$$

then it is reasonable to require that the above recursive procedure continue until this share has reached some specified fraction, ζ, of all agglomeration establishments.[50] If the desired *stopping point* is again denoted by $t^e \in \{1,..,k_\mathbf{C}\}$, then this modified stopping rule can be formalized as follows: (i) if $k_\mathbf{C} = 1$, set $t^e = 1$; (ii) if $k_\mathbf{C} \geq 2$, and if for the given pair of threshold fractions, $\mu, \zeta \in (0,1)$, there is at least one stage, $t \in \{2,3,\ldots,k_\mathbf{C}-1\}$ satisfying the following two conditions,

$$\triangle BIC_{t+1} \quad < \quad \mu \triangle BIC_1 \qquad [\textit{substantial-increment condition}] \tag{99}$$
$$s\left(\mathbf{C}_t\right) \quad \geq \quad \zeta \qquad\qquad [\textit{substantial-establishments condition}] \tag{100}$$

then choose t^e to be the *smallest* of these; and otherwise, (iii) set $t^e = k_\mathbf{C}$. This stopping rule is again illustrated by Figure 6.1 above where hypothetical shares of agglomeration establishments, $s\left(\mathbf{C}_t\right)$, are shown at each stage, $t = 1,..,k_\mathbf{C}(=7)$, on the horizontal axis. Hence if $\zeta = .80$ and if $\triangle BIC_t / \triangle BIC_1$ first falls below the specified value of μ at $t = 4$ in Figure 6.1(a), then $t^e = 3$. However, if the shares if agglomeration establishments are as shown in Figure 6.1(b) [which uses the same *BIC* increments as Figure 6.1(a)], then the procedure will not terminate until stage $t^e = 5$.

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*) for an industry with cluster scheme $\mathbf{C}$ is taken to be the smallest solid d-convex set in R containing $\mathbf{C}^e$, i.e., the d-convex solidification of $\mathbf{C}^e$:

$$ec(\mathbf{C}) = \sigma_{c_d}\left(\mathbf{C}^e\right) \tag{101}$$

These concepts can be illustrated by the stylized location patterns in Figure 6.2 below. For example, if the relevant cluster scheme, $\mathbf{C}$, for a given industry corresponds to the five clusters (shown in black) in Figure 6.2(a), and if the subset of essential clusters, $\mathbf{C}^e$, consists of the three largest clusters on the left, then the essential containment, $ec(\mathbf{C})$, for this industry is given by the filled square containing these three clusters. Similar interpretations can be given to the filled rectangles of Figures 6.2(b,c,d).

Figure 6.2 here

---

[50]Note that this condition could also be formulated in terms of *agglomeration employment*.

6.2 Global Extent and Local Density

With these definitions we next seek to compare e-containments for different industries in terms of their relative sizes. To do so, it is convenient to employ total *geographic area* rather than *economic area* (used for modeling the potential locations of individual establishments as discussed in Sections 2 and 7.1.2 above).[51] 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 nation* (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 nation, i.e.,

$$GE(\mathbf{C}) = \frac{\sum_{r \in ec(\mathbf{C})} A_r}{A} \in (0, 1] \tag{102}$$

Industries with small global extents (say, $GE < 0.50$) might be classified as "globally confined" industries [illustrated by the industries in Figures 6.2(a,c)]. Similarly, industries with large global extents (say, $GE > 0.50$) might be classified as "globally dispersed" industries [illustrated by those in Figures 6.2(b,d)].

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 *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.,

$$LD(\mathbf{C}) = \frac{\sum_{r \in \mathbf{C}^e} A_r}{\sum_{r \in ec(\mathbf{C})} A_r} \in (0, 1] \tag{103}$$

Industries with a high density of agglomerations in their e-containments (say, $LD > 0.50$) might be classified as "locally dense" industries [illustrated by the industries in Figures 6.2(a,b)]. Similarly, industries with a low density of agglomerations in their e-containments (say, $LD < 0.50$) might be classified as "locally sparse" industries [illustrated by those in Figures 6.2(c,d)].

More generally, Figure 6.2 is intended to summarize the main features of this classification system. First, the concept of *essential containment* is designed to capture the region of most significant agglomeration for an industry, while at the same time including most of its establishments. This is illustrated in each of the figures by filled regions containing the largest agglomerations for the cluster schemes shown. In each case, the "outlier" agglomerations excluded from this region are implicitly assumed to be less significant, both in terms of their contributions to BIC and their overall share of establishments for the industry.

In addition, Figure 6.2 illustrates the four possible extreme cases in this classification system. As already mentioned, *maximal spatial concentration* in this system corresponds to the case of globally confined and locally dense agglomeration patterns, such as Figure 6.2(a). The opposite extreme of *minimal spatial concentration* is characterized most naturally by globally dispersed and locally sparse agglomeration patterns, such as Figure 6.2(d).[52] The two "intermediate" extremes

---

[51]The main motivation for *geographic area* in the present context is that it tends to be a more accurate reflection of "spatial extent" than the more limited notion of economic area.

[52]However, it should be borne in mind that "minimal spatial concentration" in our present framework is not the same as "complete spatial randomness." In particular, since *all* spatial patterns are assumed to have passed the "spurious cluster" test developed above, even globally dispersed and locally sparse patterns must contain some

are somewhat more difficult to interpret, but do indeed occur (as will be seen in Section 7.3 below). Here it should be noted that these intermediate extremes do have implications for the overall *size* of the industries involved. In particular, only industries with many establishments can exhibit dense patterns of significant agglomerations over large areas [such as Figure 6.2(b)], and only industries with small numbers of establishments can exhibit sparse patterns of agglomerations in confined areas [such as Figure 6.2(c)]. Additional features and examples of this classification system will be developed in Section 7.3 below.

6.3 Comparison with A Scalar Measure

As stressed in the Introduction, it is not possible to characterize spatial patterns by any single numerical index. So while the above classification scheme in terms of paired measures (*GE* and *LD*) is still necessarily limited, it does provide a richer picture than any single summary measures of the "degree of aggomeration." This can be illustrated by comparing the present classification scheme with one such measure, namely the *D*-index developed in Mori, Nishikimi and Smith [47].[53] The *D-index* for a given industry i is defined as the Kullback-Leibler [41] divergence of its establishment location probability distribution, $P_i \equiv [P_i(r) : r \in R]$, [as in espression (2)] from purely random establishment locations. Here the latter is characterized by the uniform probability distribution, $P_0 \equiv [P_0(r) : r \in R]$, with $P_0(r) = a_r / \sum_{r \in R} a_j$ [as in expression (87)]. By using the sample estimate of P_i, namely, $\widehat{P}_i = [\widehat{P}_i(r) : r \in R]$ with $\widehat{P}_i(r) \equiv n_r / n$ [as in expression (14)], a corresponding estimate of this *D*-index is given by

$$D(\widehat{P}_i | P_0) = \sum_{r \in R} \widehat{P}_i(r) \ln \left(\frac{\widehat{P}_i(r)}{P_0(r)} \right). \tag{104}$$

The intuition behind this particular index is simply that Kullback-Leibler divergence provides a natural measure of distance between probability distributions. So by taking uniformity to represent the complete absence of clustering, it is reasonable to assume that those distributions "more distant" from the uniform distribution should involve more clustering.

But the difficulty with this measure (or in fact any continuous measure of distance between distributions) is that many distributions must necessarily be equidistant from any given distribution. So with respect to the uniform distribution particular, there are a multitude of different distributions with identical D values. As one illustration, consider the simple variant of Figure 4.5 above, involving two clustering patterns for two different industries depicted in Figure 6.3 below (say industry i on the left and industry j on the right) within the same (square-grid) system, R, of 144 basic regions.

Figure 6.3 here

---

significant degree of local clustering.

[53]Other scalar indices could be used here, such as the well known index of Ellison and Glaeser [17]. But in fact, such indices tend to be highly correlated with D (refer to Mori et al. [47, Sec.D]). So, the arguments in this section would remain essentially the same.

Here it also assumed for simplicity that within each industry, the number of establishments (or workers), n_r, is a positive constant for all black regions, r, and is zero for all other regions. Under these assumptions it should be clear that both of these industrial agglomeration patterns must necessarily exhibit the same D value. In particular, each involves 16 black regions, r, with $\widehat{P}_i(r) \equiv \widehat{P}_j(r) \equiv 1/16$, so that the distributions, $\widehat{P}_i$ and $\widehat{P}_j$, differ only by the labeling of regions. Hence in terms of the D-index it must be concluded that the "degrees of agglomeration" within industries i and j are *identical*.

But it should be clear by inspection that these two agglomeration patterns are in fact quite different. In particular, industry i is seen to be highly concentrated in one large cluster involving the 16 central regions of R. Here it is possible that i may enjoy large scale economies in production, and hence may serve world markets as well as the local market in R. Moreover, since the e-containment for industry i is seen to be identical with this single large cluster, the spatial concentration of i is readily captured by our paired classification scheme as a "globally confined" and "locally dense" pattern of agglomeration (with $GE = 16/144 \ll .5$ and $LD = 16/16 \gg .5$).

Alternatively, industry j is seen to be much more dispersed, with four separate clusters apparently each serving a local market within R. Hence, noting that the e-containment for j now includes the entire gray area in the figure, we see that this agglomeration pattern for j is distinguished from that of i as being "globally dispersed" and "locally sparse" (with $GE = 100/144 \gg .5$ and $LD = 16/100 \ll .5$).

Finally it should be noted that while the above illustration is rather extreme by design, such failures of single measures to distinguish substantially different pattern types do occur in practice. For example in the application below, this D-index fails to distinguish between many different concentration/dispersion patterns. One example is provided by the "soft drinks and carbonated water" industry and "plastic compounds and reclaimed plastics" industry discussed in Sections 7.4.1 and 7.4.4, respectively. These two industries have respective D-values of 1.95 and 2.06, and thus are very close in terms of this summary measure of "degree of agglomeration."[54] However, their actual spatial patterns of clusters are quite different, as shown in Figures 7.5 and 7.13, respectively.

7 Application

In this section, we present some preliminary results from the application of our cluster-detection approach to Japanese manufacturing industries. We begin with a description of the data in Section 7.1, and then present the results of spurious-cluster tests for this application in Section 7.2 (all subsequent analyses focus on industries with non-spurious clusters). The classification scheme developed in Section 7.3 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 7.4.

---

[54]The term "close" is here interpreted relative to the range of the sampling distribution of D values which, for the Japanese industries studied in Section 7, is from 0.471 to 5.984.

7.1 Data

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.

7.1.1 Basic Regions

The relevant notion of a "basic region" for this analysis is taken to be the *shi-ku-cho-son* (SKCS), which is a municipality category equivalent to a city-ward-town-village. The specific SKCS boundaries are taken to be those of October 1, 2001.[55] While there are a total of 3363 SKCS's in Japan, we only consider 3207 of these (as shown in Figure 7.1), 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. The only exception here is Hokkaido, which is one of the four major islands (refer to Figure 7.1), but is disconnected from the road network covering the other three. But given its size (217 SKCS's), as seen Figure 7.1, we still include Hokkaido as a potential location for establishments. Hence for 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 clusters including basic regions in both Hokkaido and other major islands. In terms of our δ-neighborhood definition in expression (71) [and footnote 46], the distances between Hokkaido regions and those of the major islands are implicitly assumed to exceed δ. Second, e-containments for each industry are obtained as the union of the two d-convex solidified subsets of essential clusters within and without Hokkaido [see, for example, the cases of "soft drinks, and carbonated water," "livestock products," and "sliding doors and screens," shown in Figures 7.5(c), 7.6(c) and 7.7(c), respectively, in Section 7.4 below].

Figure 7.1 here

7.1.2 Economic Area of Regions

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[60, 61]).[56] The economic area of Japan as a whole (120,205km^2) amounts to only 31.8% of total area in Japan.

---

[55]The data source for these SKCS boundaries is the Statistical Information Institute for Consulting and Analysis [60, 61].

[56]There is of course a certain degree of interdependence between the size of economic areas and the presence of industries in those areas. In particular, 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) can be easily be made available for industrial location in the short run.

Among individual SKCS's this percentage ranges from 2.1% to 100%, with a mean of 48.5%. Not surprisingly, those SKCS's 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.[57]

7.1.3 Interregional Distances

The shortest-route distance between each pair of neighboring SKCS's is computed as the distance between their municipality offices along the road network [using ArcGIS version 9.1 (ESRI Inc.) based on Dijkstra's [13] shortest-path algorithm]. This road network data is taken from Hokkaido-chizu Co. Lit. [32], 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. From the computed shortest-route distances between neighboring SKCS's, the corresponding shortest-path distances and shortest-path sequences of SKCS's between each pair of SKCS's are then obtained as in (34) and (35) [again using Dijkstra's algorithm].[58] 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. With respect to topography, it should also be noted that since Japan is quite mountainous with very irregular coastlines (along which the majority of industrial sites are found), shortest-route distances are generally much longer than straight-line distances. Hence the use of shortest-route distances is particularly important for countries like Japan.

7.1.4 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.[59] The establishment counts (n) across these 163 industries is taken from the Establishment and Enterprise Census of Japan [36] in 2001, and the frequency distribution of these counts is shown in Figure 7.2. 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.

Figure 7.2 here

---

[57]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).

[58]As noted in Section 4.1, 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 (d/t) 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.

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

7.2 Tests of Spuriousness of Cluster Schemes

Using the cluster-detection procedure developed in Section 5.2 above, optimal cluster schemes, C_i^*, were identified for each industry, $i = 1, .., 163$.[60] Each cluster scheme, C_i^*, was then tested for spuriousness using the testing procedure developed in Section 5.3.[61] Among the 163 industries studied, the null hypothesis of complete spatial randomness (Section 5.3 above) was strongly rejected for 154 (95%) of these industries, with p-values virtually zero. For the remaining nine industries, this null hypothesis could not be rejected at the .05 level. The main reason for non-rejection in these cases [which include seven arms-related industries (JSIC331-337), together with tobacco manufacturing (JSIC135) and coke (JSIC213)], appears to be the small size of these industries, with $n < 40$ in all cases.[62] In view of these findings, we chose to drop the nine industries in question and focus our subsequent analyses on the 154 industries exhibiting significant clustering.

For these 154 industries, Figure 7.3 shows the frequency distribution of the share of establishments for each industry i that are included in the clusters of it cluster scheme, C_i^*. These shares range from 39.1% to 100% with a median [resp., mean] share of 95.2% [resp., 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 (JSIC215) and "sawing, planning mills and wood products" industry (JSIC161) 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 7.3 here

7.3 On the Classification of Cluster Patterns

Figure 7.4 plots LD versus GE for each of 154 industries (with non-spurious clusters) under four different sets of threshold levels, μ and ζ [refer to (99) and (100), respectively]. The patterns are essentially the same for a reasonable range of μ and ζ values, although the range of (GE, LD) pairs tends to become more diverse for smaller values of ζ. In particular, there is seen to be wide

---

[60]The computation time required to identify the best cluster scheme for a given industry varies depends on the number and the spatial distribution of establishments of this industry, and of course, computational environment. Other things being equal, an industry with a smaller number of establishments requires a smaller amount of time. Computation takes more time for an industry with spatially larger clusters, e.g., in the case of industrial belt (refer to Section 6.4.3). In our computational environment (Intel C++ version 9.1 on a computer with quadratic core Xeon 2.8GHz with 32GB random access memory), the computational time for detecting the best cluster scheme ranges from less than a minute to about a week. However, it should be noted that computational time depends strongly on the implementation of the algorithms. Since the computational efficiency is not the main theme of the present paper, there should be a large room for improvement on the actual implementation of the algorithms.

[61]These tests of spuriousness were based on the BIC values for a sample of 100 completely random location patterns for each industry.

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

variation in both measures, i.e., in both the global extent and local density of cluster schemes across industries. Note also that their is no clear relationship between them, indicating that all four extremes in Figure 6.2 do in fact occur.[63] However, the overall dispersion of (GE, LD) pairs appears to be relatively more sensitive to values of ζ than μ. For example, under $\zeta = 0.8$ [Diagrams (a) and (c)], there are a few industries in the northwest section of the diagram, but not under the larger value, $\zeta = 0.9$ [Diagrams (b) and (d)]. Because these industries exhibit a high degree of spatial concentration (i.e., limited global extent and high local density), they tend to have only a few significant clusters. Thus the inclusion of an only single additional cluster can dramatically affect the size of their e-containment, and hence their global extent. For example, in Section 7.4.4 below, Figures 7.9(c) and 7.10 show the essential containment of "leather gloves and mittens" (JSIC245) under $\zeta = 0.8$ and $\zeta = 0.9$ (with $\mu = 0.03$), respectively. In the latter case, the essential containment contains a large vacant area since it includes a remote cluster in Tokyo, while the former captures a more compact and highly specialized region around Hikita-Ohuchi-Shiratori. Note also that a visual comparison of JSIC245 in Figure 7.9(c) with "motor vehicle, parts and accessories" (JSIC311) in Figure 7.12(c) suggests that the former is more "spatially concentrated," even though the latter appears to be "closer" to the maximally-concentrated northwest corner of Figure 7.4(a). Hence it should also be clear that even these two measures, GE and LD, taken together can be expected to provide only a rough classification of spatial-concentration types.

Figure 7.4 here

7.4   Examples of Cluster Schemes of Individual Industries

In this section we present a more detailed discussion of representative industries with cluster schemes exhibiting a variety of (GE, LD) combinations. Here we focus mainly on the case of Figure 7.4(a) [$\mu = 0.03$ and $\zeta = 0.8$] which is seen to exhibit the widest variation of GE and LD values. Figures 7.5 through 7.13 each focus on a different industry. For each industry i, the associated figure displays its density of establishments in each basic region (Diagram, a), the spatial pattern of clusters in its cluster scheme, $\mathbf{C}_i^*$ (Diagram, b), and the essential containment, $ec(\mathbf{C}_i^*)$, of this cluster scheme (Diagram, c). In Diagram (a), basic regions with higher densities of establishments are shown as darker. In Diagram (b), the individual clusters in scheme $\mathbf{C}_i^*$ are represented by enclosed gray areas. The portion of each cluster in lighter gray shows those basic regions which contain no establishments (but are included in $\mathbf{C}_i^*$ by the process of d-convex solidification). Finally, the hatched area in Diagram (c) depicts the e-containment, $ec(\mathbf{C}_i^*)$, of this cluster scheme.

---

[63]The relative positions of Diagrams (a) through (d) in Figure 6.2 are arranged to match the relative positions in each diagram of Figure 7.4. In particular, globally confined patterns in Figures 6.2(a,c) [resp., locally dense patterns in Figures 6.2(a,b)] are found in western [resp., northern] part of each diagram in Figure 7.4.

7.4.1 Globally Dispersed and Locally Sparse Patterns

Industries with relatively high values of GE and low values of LD [near the southeast corner of Figure 7.4(a)] can be described as exhibiting patterns of agglomeration that are both "globally dispersed and locally sparse." A clear example is provided by the "soft drinks and carbonated water" industry (JSIC131) shown in Figure 7.5 [with $GE = 0.589$ and $LD = 0.133$]. Bottled/packed soft drinks are weight/bulk-gaining products. Thus to minimize transport costs, establishments in this industry are naturally attracted to individual market locations, resulting in a pattern of global dispersion. In addition, the individual clusters shown in Figure 7.5(b) appear to be locally concentrated, perhaps due to scale economies of production combined with only modest needs for land. Thus in terms of total area occupied, this pattern of clusters is relatively sparse.

Figure 7.5 here

A second example is provided by the "livestock products" industry (JSIC121) depicted in Figure 7.6 [with $GE = 0.771$ and $LD = 0.281$]. Here the perishable nature of livestock products again leads to market-oriented location behavior, and hence to global dispersion. But in this case, the extensive land requirements for livestock production produce higher local densities in terms of area occupied, and thus result in larger clusters than JSIC131 [as seen in Figure 7.6(b)].

Figure 7.6 here

7.4.2 Globally Dispersed and Locally Dense Patterns

Industries with both high values of GE and LD [near the northeast corner of Figure 7.4(a)] 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*. While there are no extreme examples in Figure 7.4(a), one relatively ubiquitous example is provided by the "sliding doors and screens" (JSIC173) [with $GE = 0.777$, $LD = 0.473$]. As seen in Figure 7.7(a), the establishments of this industry are indeed found almost everywhere, with clusters densely distributed throughout the nation [Figure 7.7(b)]. Such products are often custom made and require face-to-face contact with customers. Hence there are strong market-attraction forces that contribute to the ubiquity of this industry. In such cases, the local density of clusters tends to correspond roughly to that of population.

Figure 7.7 here

It is also of interest to note (as mentioned at the end of Section 6) that such ubiquitous industries are by their very nature quite large in terms of establishment numbers. In the present case, industry JSIC173 has 15,363 establishments, which is well above the mean of 4189 for all industries (with no spurious clusters, i.e., exhibiting significant agglomeration). In terms of establishments in clusters, JSIC173 has 13,565 establishments relative to a mean of only 3896 for all industries.

7.4.3 Globally Confined and Locally Sparse Patterns

The opposite extreme of "globally confined and locally sparse" agglomeration patterns [in the southwest corner of Figure 7.4(a)] is well illustrated by the "ophthalmic goods" (JSIC326) [with $GE = 0.166$ and $LD = 0.139$]. As seen in Figure 7.8(a) this industry has only a small number of establishments (located mainly between Tokyo and Osaka), with a disproportionate concentration in the small town of Sabae (population 65,000). In fact, this single remote town accounts for more than 90% of the national market share in ophthalmic goods. As with many specialized industries, the location pattern of this industry is governed more by historical circumstances than economic factors. In terms of establishment numbers, such industries are necessarily small in size. In the present case, JSIC326 has only 1139 establishments, which is well below the mean of 4188 for all industries (as above). Even given the fact that all 1139 establishments are in clusters, this number is still well below the mean of 3896 for all industries (as above).

Figure 7.8 here

A similar example of this pattern is the "leather gloves and mittens" industry (JSIC245) depicted in Figure 7.9 [with $GE = 0.019$ and $LD = 0.418$]. Like JSIC326, this industry is not concentrated in large cities. Rather, its major concentration (accounting for 90% of the leather glove market in Japan) is confined to a cluster of three small towns, Hikita-Ohuchi-Shiratori, shown in Figure 7.9(b).

Figure 7.9 here

Here it is of interest to note that while the value of LD for JSIC245 seems relatively large compared to JSIC326 above, this is mostly due to its small e-containment, as reflected by its low level of GE relative to JSIC326 [compare Figures 7.8(c) and 7.9(c)]. When GE is very small for an industry, its value of LD is necessarily sensitive to the number of clusters in its e-containment.

In addition, it is also important to note that for globally confined industries with few clusters (such as JSIC245 and JSIC326), the values of GE and LD are both quite sensitive to the cut-off criteria, μ and ζ, in (99) and (100), respectively. As one illustration, Figure 7.10 shows the essential containment of JSIC245 obtained with $\zeta = 0.9$ rather than $\zeta = 0.8$ as in Figure 7.9(c).

While this higher value of ζ allows the inclusion of only one additional cluster, the location of this cluster in Tokyo leads to the inclusion of a large vacant area between Osaka and Tokyo in the resulting d-convex solidification of these clusters.

Figure 7.10 here

A final example is provided by the larger "publishing industry" (JSIC192) depicted in Figure 7.11 [with $GE = 0.342$ and $LD = 0.232$]. Unlike JSIC326 and JSIC245, publishing is a typical "urban-oriented" industry with a location pattern generally reflecting urban density. As seen in Figure 7.11(b) this pattern is more concentrated toward 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 7.11 here

7.4.4 Globally Confined and Locally Dense Patterns

Finally, as mentioned in Section 6 above, those industries with agglomeration patterns that are both "globally confined and locally dense" [i.e., in the northwest corner of Figure 7.4(a)] constitute the single most spatially concentrated class of industries. Such industries are well illustrated by the "motor vehicles, parts and accessories" (JSIC311) in Figure 7.12 [with $GE = 0.221$ and $LD = 0.751$]. A comparison of the e-containment for this industry in Figure 7.12(c) with that of the urban-oriented publishing industry in Figure 7.11(c) shows that JSIC311 again follows the chain of large metro areas extending westward from Tokyo through Osaka to Hiroshima. But here the containment is even more concentrated along this chain, and coincides with the so-called Industrial Belt that constitutes the manufacturing core of Japan. This manufacturing core is in fact dominated by the major auto assembly plants in this industry, which by definition produce weight/bulk-gaining products requiring proximity to consumers in the metro centers. Moroever, the chain of contiguous clusters seen in Figure 7.12(b) essentially fills in the gaps between these metro centers, creating the effect of a single "megalopolis." The outputs of JSIC311 provide an important clue to the nature of this "filling-in" process. In particular, "parts and accessories" are basically factor inputs to the auto assembly process ("motor vehicles"). Moreover, since parts suppliers tend to sell to more than one car assembler,[64] the intermediate locations between these assemblers provide natural market economies for such suppliers.[65]

Figure 7.12 here

---

[64]In 1999, parts suppliers on average sold to 3.05 of the 9 auto assemblers in Japan, while auto assemblers on average bought the same parts from 2.46 different suppliers (Kinnou [37]).

[65]For a theoretical development of this "filling-in" process in the context of the new economic geography model see Mori [46].

As mentioned in Section 6.3, a second example is provided by the "plastic compounds and reclaimed plastics" industry (JSIC225) [with $GE = 0.298$ and $LD = 0.465$]. From Figure 7.13(b) it is clear that most clusters for this industry also follow the Industrial Belt.[66] Moreover, the outputs of this industry are again primarily intermediate inputs for a variety of manufactured goods, and in particular for motor vehicles (such as the molded plastic parts for seats, fenders, and instrument panels). Thus the intermediate locations between these manufacturers again constitute natural market-oriented locations for this industry. Hence the filling-in process that created this industrial belt is largely a consequence of the fact that typical automobiles consist of as many as 20,000 to 30,000 separate parts.

Figure 7.13 here

8 Concluding Remarks

In this paper we have developed a simple *cluster-scheme* model of agglomeration patterns and have constructed an information-based algorithm for identifying such patterns. In addition, we have proposed a simple classification of pattern types based on measures of *global extent* and *local density* derived from cluster schemes. But the ultimate utility of this approach will of course depend on how it can be applied in practical situations.

Here it should be noted that the distinction between local and global properties of agglomeration patterns implicit in our classification scheme has already served to sharpen certain concepts in the literature. For example it was pointed out in Section 7.4.4 of our application that the Japanese Industrial Belt is an instance of the more general notion of a "megalopolis," first proposed by Gottman [28] to describe the continuum of cities along the US Atlantic seaboard (stretching from Boston to Washington, DC, via New York). But to date, no formal methods have been developed for identifying such agglomeration structures statistically. In this light, the analysis of Section 7.4.4 shows that such structures can be regarded as natural instances of "globally confined and local dense" agglomeration patterns.

More generally, there appear to be a number of questions raised in the literature which can potentially be addressed by our present approach. Hence it is appropriate to mention two possible research directions involving, respectively, the spacing of agglomertions within industries and the coordination of agglomerations between industries. But before doing so, it is useful to begin by observing that certain consequences of simple cluster-scheme model proposed here need to be made more explicit.

8.1 Refinements of Cluster Schemes

Recall that each cluster within a given cluster scheme implicitly defines a set of basic regions with similar (and unusually high) establishment density. But the relations between these individual

---

[66]The lower density for this industry is due mainly to the fact that the e-containment in Figure 7.13(c) also includes clusters on the Sea of Japan coast around Toyama (refer to Figure 7.13(b)).

41

clusters is left unspecified. In this regard it is important to observe that in many of the cluster schemes we have identified for industries in Japan, there are notable groupings of *contiguous* clusters. As one example, consider "publishing industry " (JSIC192) in Figure 7.11. For this urban oriented industry there seem to be significant contiguous clusters around Tokyo, Nagoya and Osaka.

Here is it natural to ask why such clusters were not "joined" at some stage during the cluster-detection procedure. The reason is that our basic cluster probability model assumes that location probabilities are essentially *uniform* within each cluster [as in expression (8)]. Hence with respect to the *BIC* measure underlying this procedure, contiguous clusters with very different uniform densities often yield a better fit to establishment data than does their union with its associated uniform density. This is well illustrated by the contiguous clusters for the publishing industry in the Tokyo area, as shown in the enlargement in Figure 8.1(a) below.

Figure 8.1 here

This example shows not only that the establishment densities in these contiguous clusters are quite different, but also that such variations exhibit clear *spatial structure*. In particular, the single darkest (most dense) cluster corresponds precisely to the heart of downtown Tokyo, with adjacent clusters gradually diminishing in density. Hence this density pattern might be well described as a *hill structure* with "peak" in downtown Tokyo and "foot" consisting of the ring of outer-most contiguous clusters. As seen in Figure 8.1(b), a similar structure can be obtained by plotting the *BIC* increments associated with the essential-cluster construction in Section 6.1 above.

More generally, this example shows that there is often more spatial structure in given cluster schemes than is captured by a simple listing of their clusters. In particular it seems clear that groupings of contiguous clusters are best treated a single *agglomerations* for an industry. So while we have implicitly used the terms "clusters" and "agglomerations" interchangeably in this paper, it would seem that latter term is best reserved for *maximal contiguous sets* of clusters. Under this definition, each cluster scheme, $\mathbf{C} = (R_0, C_1, .., C_{k_C})$, then generates a unique *agglomeration scheme*, $\boldsymbol{\mathcal{A}} = (R_0, \mathcal{A}_1, .., \mathcal{A}_{k_{\mathcal{A}}})$ [which is identical with $\mathbf{C}$ if there are no contiguous clusters]. Such refinements of our basic cluster-scheme model will be considered more explicitly in subsequent work. But for the present, it is convenient to use this broader definition of agglomerations in discussing the additional extensions below.

8.2 Agglomeration Spacing within Industries

Within the new economic geography, a class of models have been developed to explain the spacing between individual agglomerations for a given industry (e.g., Krugman [40], Fujita and Krugman [20], Fujita and Mori [23], Fujita et al. [22, Ch.6]). From the view point of general equilibrium theory, these models predict whether an agglomeration of industrial firms will be viable at a given location, depending on how other agglomerations 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 [20] have shown that each agglomeration casts a so-called *agglomeration shadow* in which firms have no incentive to relocate from the existing agglomerations. For within this "shadow" firms are too close to existing agglomerations (i.e., competitors) to realize sufficient local monopoly advantages. Hence the presence of such shadows serves to limit the number of viable agglomerations 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. Hence the presence of such shadows has a number of observable spatial consequences.

But while there has been empirical work to study the spacing between urban centers (as for example in Chapter 7 of Marshall [45] and in Ioannides and Overman [34]), there have to our knowledge been no systematic efforts to study the spacing between industrial agglomerations – 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 agglomerations for each industry, it is a simple matter to construct explicit measures of the spacing between them. In the present setting, the most natural measure of spacing between any pair of agglomerations, $\mathcal{A}_i$ and $\mathcal{A}_j$, is the road-network distance between their closest basic regions, which [as an extension of expression (70)] is given by

$$d(\mathcal{A}_i, \mathcal{A}_j) = \min\{d(r,s) : r \in \mathcal{A}_i, s \in \mathcal{A}_j\} \tag{105}$$

Moreover, for any agglomeration scheme, $\mathcal{A} = (R_0, \mathcal{A}_1, .., \mathcal{A}_{k_\mathcal{A}})$, the size of the shadow around each agglomeration i is best reflected by the distance from i to its *nearest neighbor* in $\mathcal{A}$:

$$d_i(\mathcal{A}) = \min_{j \neq i} d(\mathcal{A}_i, \mathcal{A}_j) \tag{106}$$

The average of these nearest-neighbor distances thus yields a natural *mean-spacing measure*

$$d(\mathcal{A}) = \frac{1}{k_\mathcal{A}} \sum_{i=1}^{k_\mathcal{A}} d_i(\mathcal{A}) \tag{107}$$

for $\mathcal{A}$. This summary measure can then be employed for testing purposes. In particular, one can test for the presence of significant agglomeration-shadow effects by asking whether the mean spacing, $d(\mathcal{A})$, for an observed agglomeration scheme, $\mathcal{A}$, is significantly larger than would be expected if such agglomerations were randomly located.

Given the spatially extensive nature of agglomerations, this task is more complex than for random relocations of points. But one simple approach to constructing *random versions*, $\mathcal{A}'$, of $\mathcal{A}$ is to reorder the individual agglomerations, $\mathcal{A}_1, .., \mathcal{A}_{k_\mathcal{A}}$, in $\mathcal{A}$ by size and regenerate them sequentially from largest to smallest. For the largest agglomeration, $\mathcal{A}_1$, one can choose a random starting region, $r \in R$, and "spiral out" until a set of contiguous regions, $\mathcal{A}'_1 \subset R$, is

achieved that approximates the size of $\mathcal{A}_1$. By removing this set of regions from R, the same procedure can then be repeated for constructing a random version, $\mathcal{A}_2' \subset R - \mathcal{A}_1'$, of the second largest agglomeration, $\mathcal{A}_2$, and so on. In this way, many random versions, $\mathcal{A}' = (R_0', \mathcal{A}_1', .., \mathcal{A}_{k_\mathcal{A}}')$, of $\mathcal{A}$ can be constructed for testing purposes. The appropriate null hypothesis of "random spacing" for this test is then that $d(\mathcal{A})$ is a typical realization from the sampling distribution of mean spacings, $d(\mathcal{A}')$, generated by many random versions $\mathcal{A}'$ of $\mathcal{A}$. Applications of this procedure will be reported in subsequent work.[67]

8.3 Agglomeration Coordination between Industries

Within the context of Christaller's [8] 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.[68] Theoretical efforts to explain this phenomenon have focused mainly on the role of demand externalities in determining industrial locations (see Fujita, Krugman and Mori [21], Tabuchi and Thisse [64] and Hsu [33]).[69] In particular, the types of demand externalities which induce industrial agglomerations 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 agglomerations 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. The results of the present paper suggest one direct test of co-location using the randomization procedure outlined above. In particular, if we associate larger market sizes with smaller numbers of agglomerations,[70] and consider any pair of industries, i and j, with different market sizes ($|\mathcal{A}_i| < |\mathcal{A}_j|$), then one could test whether the agglomerations of industry i are closer to those of industry j than would be expected in random configurations. If the observed agglomeration patterns of these industries are denoted respectively by $\mathcal{A}_i = (R_{i0}, \mathcal{A}_{i1}, .., \mathcal{A}_{ik_{\mathcal{A}_i}})$ and $\mathcal{A}_j = (R_{j0}, \mathcal{A}_{j1}, .., \mathcal{A}_{jk_{\mathcal{A}_j}})$, then one could start by identifying the *nearest-neighbor distance* from each agglomeration, $\mathcal{A}_{ih} \in \mathcal{A}_i$, to those in $\mathcal{A}_j$:

$$d(\mathcal{A}_{ih}, \mathcal{A}_j) = \min\{d(\mathcal{A}_{ih}, \mathcal{A}_{jm}) : \mathcal{A}_{jm} \in \mathcal{A}_j\} \tag{108}$$

[67]Here it is of interest to note that initial investigations of such spacing comparsions suggest that further restrictions need to be imposed. In particular, for those industries with small e-containments, it is clear that random versions located throughout all of Japan will necessarily tend to exhibit larger mean spacing for rather spurious reasons. One possibility here is to preserve the e-containment of each industry, and to restrict random versions to these e-containments. This should provide more meaingful tests of the presence of agglomeration shadows in which the overall spatial scale of each industry is preserved.

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

[69]There were earlier attempts by, e.g., Christaller [8], Lösch[44], Beckmann [2] and Eaton and Lipsey [16]. But, all lacked formal microeconomic foundations leading to the Hierarchy Principle.

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

and then defining the *mean distance* between $\mathcal{A}_i$ and $\mathcal{A}_j$ to be the average of these:

$$d(\mathcal{A}_i, \mathcal{A}_j) = \frac{1}{k_{\mathcal{A}_i}} \sum_{h=1}^{k_{\mathcal{A}_i}} d(\mathcal{A}_{ih}, \mathcal{A}_j) \tag{109}$$

To employ this summary measure as a test statistic, one could then use the procedure above to generate many random versions, $\mathcal{A}'_i$, of $\mathcal{A}_i$, and test whether $d(\mathcal{A}_i, \mathcal{A}_j)$ is significantly smaller than would be expected from the sampling distribution of mean-distance values, $d(\mathcal{A}'_i, \mathcal{A}_j)$. Applications of this testing procedure will also be reported in subsequent work.

Finally, it should also be noted that the present cluster methodology has already been applied by Mori and Smith [49] to test the Hierarchy Principle in a different way. This test was originally developed in Mori, Nishikimi and Smith [48] using the criteria that an industry is present in a city if at least one of its establishments is located in that city. But later work revealed that such a definition was too broad in that a single establishment may locate in a given city by chance alone. To develop a stronger definition, the present cluster-dectection procedure was employed to identify those cities containing establishments that are actually part of a cluster for the industry. Such cities are designated as *cluster-based choice cities* for that industry. By extending the testing procedures of Mori et al. [48] to cluster-based choice cities, it was found that even stronger evidence for the Hierarchy Principle could be demonstrated. The specifics of this testing procedure will be detailed more fully in a forthcoming paper, Mori and Smith [50].

9   APPENDIX. Formal Analysis of *d*-Convex Solids

To develop formal properties of *d*-convex solids, we require a few additional definitions. First, for any path, $\rho = (r_1, r_2, .., r_{n-1}, r_n) \in \mathcal{P}(r_1, r_n)$, let $\tilde{\rho} = (r_n, r_{n-1}, .., r_2, r_1) \in \mathcal{P}(r_n, r_1)$ denote the *reverse path* in $\mathcal{P}$. Next, for any two paths, $\rho = (r_1, .., r_n)$, $\rho' = (r'_1, .., r'_m) \in \mathcal{P}$, with $r_n = r'_1$, the combined path, $\rho \circ \rho' = (r_1, .., r_n, r'_2, .., r'_m) \in \mathcal{P}$ is designated as the *concatenation* of ρ and ρ'. It then follows by definition that the length of any concatenated path, $\rho \circ \rho'$, is simply the sum of the lengths of ρ and ρ', i.e., that

$$
\begin{aligned}
l(\rho \circ \rho') &= \sum_{i=1}^{n-1} d(r_i, r_{i+1}) + d(r_n, r'_2) + \sum_{i=2}^{m-1} d(r'_i, r'_{i+1}) \\
&= \sum_{i=1}^{n-1} d(r_i, r_{i+1}) + \sum_{i=1}^{m-1} d(r'_i, r'_{i+1}) \\
&= l(\rho) + l(\rho')
\end{aligned}
\tag{110}
$$

Using (110), as well as (39) through (42), it is convenient to establish the following well-known properties of *d*-convex sets, as in Definition 4.1 of the text. First, we show that for the *d-convexification function*, $c_d : \mathcal{R} \to \mathcal{R}$, in (42), the naming of this function is justified by the fact that:

Proposition A.1 (*d*-**Convexification**) *For all $S \in \mathcal{R}$, the image set, $c_d(S)$, is d-convex.*

Proof: For any $r_1, r_2 \in c_d(S)$ and shortest path, $\rho \in \mathcal{P}_d(r_1, r_2)$, it must be shown that $\langle \rho \rangle \subseteq c_d(S)$. But by definition, $r_i \in c_d(S) \Rightarrow r_i \in I^{k_i}(S)$ for some k_i, $i = 1, 2$. Hence by (40) it follows that $\{r_1, r_2\} \subseteq I^{k_1 + k_2}(S)$, and thus that $\langle \rho \rangle \subseteq I(I^{k_1 + k_2}(S)) = I^{k_1 + k_2 + 1}(S) \subseteq c_d(S)$.∎

Next we show that the *d*-convex hull, $c_d(S)$, can be characterized as the unique smallest *d*-convex superset of S. More precisely, if $\mathcal{R}_d$ denotes the family of all *d*-convex sets in $\mathcal{R}$, then we have:

Proposition A.2 (**Minimality of *d*-Convexifications**) *For all $S \in \mathcal{R}$,*

$$
c_d(S) = \cap \{C \in \mathcal{R}_d : S \subseteq C\}
\tag{111}
$$

Proof: By Proposition A.1, $c_d(S) \in \mathcal{R}_d$, and by (39)

$$
S \subseteq I(S) \subseteq c_d(S)
\tag{112}
$$

Hence it suffices to show that for all sets, C, with $C \in \mathcal{R}_d$ and $S \subseteq C$, we must have $c_d(S) \subseteq C$. By the definition of $c_d(S)$ this in turn is equivalent to showing that $I^k(S) \subseteq C$ for all $k \geq 1$. But by (38),

$$
S \subseteq C \Rightarrow \bigcup_{r,s \in S} I(r,s) \subseteq \bigcup_{r,s \in C} I(r,s) \Rightarrow I(S) \subseteq I(C)
\tag{113}
$$

Moreover, by (37) and (38) together with the definition of *d*-convexity it follows that

$$
C \in \mathcal{R}_d \Rightarrow I(C) = \bigcup_{r,s \in C} I(r,s) \subseteq C
\tag{114}
$$

Hence we may conclude from (113) and (114) that $I(S) \subseteq C$. Finally, since the same argument shows that $I^k(S) \subseteq C \in \mathcal{R}_d \Rightarrow I^{k+1}(S) = I[I^k(S)] \subseteq C$, the result follows by induction on k.∎

Finally, using these two results, we show that d-convex sets can be equivalently characterized as the *fixed points* of the d-convexification mapping, $c_d : \mathcal{R} \to \mathcal{R}$:

Proposition A.3 (d-Convex Fixed Points) *For all $S \in \mathcal{R}$,*

$$S \in \mathcal{R}_d \Leftrightarrow c_d(S) = S \tag{115}$$

Proof: If $c_d(S) = S$ then by Proposition A.1 above, $S \in \mathcal{R}_d$. Conversely, if $S \in \mathcal{R}_d$ then by (112), $S \subseteq c_d(S)$, and by Proposition A.2, $c_d(S) \subseteq S$. Thus $c_d(S) = S$, and the result is established.∎

This in turn implies that the family, $\mathcal{R}_d$, of d-convex sets can be equivalently defined as in expression (43) of the text. But while this definition provides a natural parallel to the case of d-convex solids developed below, the more useful *interval* characterization of $\mathcal{R}_d$ in expression (44) of the text, can easily be obtained from Proposition A.3 as follows:

Corollary (Interval Fixed Points) *For all $S \in \mathcal{R}$,*

$$S \in \mathcal{R}_d \Leftrightarrow I(S) = S \tag{116}$$

Proof: Since $S \in \mathcal{R}_d \Rightarrow I(S) \subseteq S$ by (114) [with $C = S$], and since $S \subseteq I(S)$ holds for all S [by (39)], it follows on the one hand that $S \in \mathcal{R}_d \Rightarrow I(S) = S$. Conversely, since $I(S) = S \Rightarrow I^k(S) = S$ for all $k \geq 1$ [by recursion on k], it follows from (42) and Proposition A.3 that $I(S) = S \Rightarrow c_d(S) = S \Rightarrow S \in \mathcal{R}_d$. ∎

Given these properties of d-convex sets, one objective of this Appendix is to show that each of these properties is inherited by d-convex solids. To do so, we begin with an analysis of *solid* sets as in Definition 4.2 of the text. First, in a manner paralleling Proposition A.1 above, we show for the *solidification function*, $\sigma : \mathcal{R} \to \mathcal{R}$, defined by (47), the naming of this function is justified by the fact that:

Lemma A.1 (Solidification) *For all $S \in \mathcal{R}$, the image set, $\sigma(S)$, is solid.*

Proof: If $V = \sigma(S) = S \cup S_0$, then it must be shown that for all $r \in R - V$ there is some path, $\rho \in \mathcal{P}(r, \overline{R})$ with $\langle \rho \rangle \cap V = \varnothing$. But for any $r \in R - V = R - (S \cup S_0)$, it follows that $r \in R - S$ and $r \notin S_0$, so that by the definition of S_0 in (46) it must be true that there is some boundary region, $\bar{r} \in \overline{R}$, and path, $\rho \in \mathcal{P}(r, \bar{r})$ with $\langle \rho \rangle \cap S = \varnothing$. Next we show that $\langle \rho \rangle \cap S_0 = \varnothing$ as well. To do so, suppose to the contrary that $\langle \rho \rangle \cap S_0 \neq \varnothing$, so that for some $r_0 \in S_0, \rho = (r, .., r_0, .., \bar{r}) = \rho_1 \circ \rho_2$ with $\rho_1 \in \mathcal{P}(r, r_0)$ and $\rho_2 \in \mathcal{P}(r_0, \bar{r})$. Then again by the definition of S_0 it must be true that $\langle \rho_2 \rangle \cap S \neq \varnothing$, which contracts the fact that $\langle \rho_2 \rangle \subseteq \langle \rho \rangle$ and $\langle \rho \rangle \cap S = \varnothing$. Hence $\varnothing = (\langle \rho \rangle \cap S) \cup (\langle \rho \rangle \cap S_0) = \rho \cap (S \cup S_0) = \langle \rho \rangle \cap V$, and the result is established.∎

If the family of all solid sets in $\mathcal{R}$ is denoted by $\mathcal{R}_\sigma = \{S \in \mathcal{R} : S_0 = \varnothing\}$, then we next show that these sets are precisely the fixed points of the solidification function:

Lemma A.2 (Solid Fixed Points) *For all $S \in \mathcal{R}$,*

$$S \in \mathcal{R}_\sigma \iff \sigma(S) = S \tag{117}$$

Proof: If $S \in \mathcal{R}_\sigma$ then $S_0 = \varnothing$, so that $\sigma(S) = S$ by (47). Conversely, if $\sigma(S) = S$, then by Lemma A.1, $S \in \mathcal{R}_\sigma$.∎

As a parallel to (116), this in turn implies that the family of solid sets in $\mathcal{R}$ can be equivalently defined as follows:

$$\mathcal{R}_\sigma = \{S \in \mathcal{R} : \sigma(S) = S\} \tag{118}$$

Finally, solid sets also exhibit the following nesting property:

Lemma A.3 (Solid Nesting) *For all $S, V \in \mathcal{R}$,*

$$S \subseteq V \implies \sigma(S) \subseteq \sigma(V) \tag{119}$$

Proof: Since $S \subseteq V \subseteq V \cup V_0 = \sigma(V)$, it suffices to show that $S_0 \subseteq \sigma(V)$. Hence consider any $r \in S_0$, and observe from the above that $r \in V \Rightarrow r \in \sigma(V)$. Hence it remains to consider $r \in S_0 - V$. Here we show that r must be in V_0. To do so, observe first that $r \notin V \Rightarrow r \in R - V$. Moreover, $r \in S_0$ implies that for any path, $\rho \in \mathcal{P}(r, \overline{R})$ we must have $\langle \rho \rangle \cap S \neq \varnothing$. But $S \subseteq V$ then implies $\langle \rho \rangle \cap V \neq \varnothing$. Hence $r \in V_0 \subseteq \sigma(V)$, and the result is established.∎

With these properties of solid sets, we are now ready to analyze d-convex solids in $\mathcal{R}$. As asserted in the text, our key result is to show that d-convexity is preserved under solidifications:

Theorem A.1 (Solidification Invariance of d-Convexity) *For all d-convex sets, $S \in \mathcal{R}$, the image set, $\sigma(S)$, is also d-convex.*

Proof: Suppose to the contrary that for some d-convex set, S, the image set $\sigma(S)$ is not d-convex. Then there must exist some pair of elements, $r_1, r_2 \in \sigma(S) = S \cup S_0$, and some shortest path, $\rho \in \mathcal{P}_d(r_1, r_2)$, with $\langle \rho \rangle \cap [R - \sigma(S)] \neq \varnothing$. But if $\{r_1, r_2\} \subseteq S$ then by the d-convexity of S we would have $\langle \rho \rangle \subseteq S \subseteq \sigma(S)$. So at least one of these elements must be in S_0. Without loss of generality, we may suppose that $r_1 \in S_0$ and that r is some element of $\langle \rho \rangle \cap [R - \sigma(S)]$, so that $\rho = (r_1, .., r, .., r_2) = \rho_1 \circ \rho_2$ with $\rho_1 \in \mathcal{P}(r_1, r)$ and $\rho_2 \in \mathcal{P}(r, r_2)$. But then we must have $S \cap \langle \rho_1 \rangle \neq \varnothing$. For if not then we obtain a contradiction as follows. Since $r \notin \sigma(S) \Rightarrow [r \in R - S$ and $r \notin S_0]$, there must be some path, $\rho_3 \in \mathcal{P}(r, \overline{R})$ with $\langle \rho_3 \rangle \cap S = \varnothing$. Hence the combined path, $\rho_1 \circ \rho_3 \in \mathcal{P}(r_1, \overline{R})$, then satisfies $\langle \rho_1 \circ \rho_3 \rangle \cap S = \varnothing$, which contradicts the hypothesis that $r_1 \in S_0$. Thus we may assume that there is some $s_1 \in S \cap \langle \rho_1 \rangle$, and consider the following two cases:

(i) Suppose first that r_2 is also an element of S_0. We then show that this contradicts the hypothesized shortest-path property of ρ as follows. Observe first that if $\widetilde{\rho}_2 \in \mathcal{P}(r_2, r)$ denotes the

reverse path for $\rho_2 \in \mathcal{P}(r, r_2)$ above, then the same argument used for $\rho_1 \in \mathcal{P}(r_1, r)$ above now shows that there must be some $s_2 \in S \cap \langle \tilde{\rho}_2 \rangle = S \cap \langle \rho_2 \rangle$, so that $\rho = (r_1, .., s_1, .., r, .., s_2, .., r_2) = \rho_1' \circ \rho_2' \circ \rho_3' \circ \rho_4'$ with $\rho_1' \in \mathcal{P}(r_1, s_1)$, $\rho_2' \in \mathcal{P}(s_1, r)$, $\rho_3' \in \mathcal{P}(r, s_2)$, and $\rho_4' \in \mathcal{P}(s_2, r_2)$. These paths are shown in Figure A.1 below.

$$\boxed{\text{Figure A.1 here}}$$

But if we choose any shortest path, $\rho_5' \in \mathcal{P}_d(s_1, s_2)$ [as in Figure A.1], then it follows from the d-convexity of S, together with $s_1, s_2 \in S$ and $r \notin S$ that $l(\rho_5') < l(\rho_2' \circ \rho_3')$ [since every shortest path in $\mathcal{P}_d(s_1, s_2)$ lies in S, and $\langle \rho_2' \circ \rho_3' \rangle \nsubseteq S$]. Hence for the path, $\rho' = \rho_1' \circ \rho_5' \circ \rho_4' \in \mathcal{P}(r_1, r_2)$, we must have

$$
\begin{aligned}
l(\rho') &= l(\rho_1') + l(\rho_5') + l(\rho_4') \\
&< l(\rho_1') + [l(\rho_2') + l(\rho_3')] + l(\rho_4') \\
&= l(\rho_1' \circ \rho_2' \circ \rho_3' \circ \rho_4') \\
&= l(\rho)
\end{aligned}
\tag{120}
$$

which contradicts the shortest-path property of ρ.

(ii) Finally, suppose that $r_2 \in S$, and for the point $s_1 \in S \cap \langle \rho_1 \rangle$ above, consider the representation of ρ as $\rho = (r_1, .., s_1, .., r, .., r_2) = \rho_1' \circ \rho_2' \circ \rho_2$ with $\rho_1' \in \mathcal{P}(r_1, s_1)$, $\rho_2' \in \mathcal{P}(s_1, r)$, and $\rho_2 \in \mathcal{P}(r, r_2)$, as shown in Figure A.2 below.

$$\boxed{\text{Figure A.2 here}}$$

Then we again show that this contradicts the shortest-path property of ρ as follows. For any shortest path, $\rho_6' \in \mathcal{P}_d(s_1, r_2)$ [as in Figure A.2], the d-convexity of S, together with $s_1, r_2 \in S$ and $r \notin S$, now implies that $l(\rho_6') < l(\rho_2' \circ \rho_2)$. Thus for the path, $\rho'' = \rho_1' \circ \rho_6' \in \mathcal{P}(r_1, r_2)$, we must have

$$
\begin{aligned}
l(\rho'') &= l(\rho_1') + l(\rho_6') \\
&< l(\rho_1') + [l(\rho_2') + l(\rho_2)] \\
&= l(\rho_1' \circ \rho_2' \circ \rho_2) \\
&= l(\rho)
\end{aligned}
\tag{121}
$$

which again contradicts the shortest-path property of ρ. Hence for each pair of elements, $r_1, r_2 \in \sigma(S) = S \cup S_0$, there can be no shortest path, $\rho \in \mathcal{P}_d(r_1, r_2)$, with $\langle \rho \rangle \cap [R - \sigma(S)] \neq \varnothing$, so that $\sigma(S)$ is d-convex. ∎

With this result, we can now establish parallels to Propositions A.1, A.2, and A.3 above for d-convex solids, as in Definition 4.3. First, we show that for the *d-convex solidification function*, $\sigma c_d : \mathcal{R} \to \mathcal{R}$, in (48), the naming of this function is justified by the fact that:

Theorem A.2 (*d*-**Convex Solidification**) *For each set, $S \in \mathcal{R}$, the image set, $\sigma c_d(S)$, is a d-convex solid.*

Proof: First observe from Definition 4.3 that we may use expressions (116) and (117) to define the family of all *d*-convex solids in equivalent terms as

$$\mathcal{R}_{\sigma d} = \mathcal{R}_\sigma \cap \mathcal{R}_d \tag{122}$$

Hence it suffices to show that $\sigma c_d(S) \in \mathcal{R}_d \cap \mathcal{R}_\sigma$. But by Proposition A.1, it follows that $c_d(S) \in \mathcal{R}_d$, and hence as a direct consequence of Theorem A.1 that $\sigma c_d(S) = \sigma[c_d(S)] \in \mathcal{R}_d$. Moreover, since $c_d(S) \in \mathcal{R}$ also implies from Lemma A.1 that $\sigma[c_d(S)] \in \mathcal{R}_\sigma$, it then follows that $\sigma c_d(S) \in \mathcal{R}_{\sigma d}$.∎

Next, as a parallel to Proposition A.2 we now have:

Theorem A.3 (**Minimality of *d*-Convex Solidifications**) *For each set, $S \in \mathcal{R}$,*

$$\sigma c_d(S) = \cap \{C \in \mathcal{R}_{\sigma d} : S \subseteq C\} \tag{123}$$

Proof: First observe from Theorem A.2 that $\sigma c_d(S) \in \mathcal{R}_{\sigma d}$ and from expression (112) that $S \subseteq c_d(S) \subseteq \sigma[c_d(S)] = \sigma c_d(S)$ [since by definition, $V \in \sigma(V)$ for all V]. Hence, it suffices to show that $\sigma c_d(S) \subseteq C$ whenever $S \subseteq C \in \mathcal{R}_{\sigma d}$. But by Proposition A.2, $C \in \mathcal{R}_{\sigma d} \subseteq \mathcal{R}_d$ and $S \subseteq C$ imply that $c_d(S) \subseteq C$. Moreover, since $C \in \mathcal{R}_{\sigma d} \subseteq \mathcal{R}_\sigma$, we may then conclude from Lemma A.3 together with Lemma A.2 that

$$c_d(S) \subseteq C \Rightarrow \sigma[c_d(S)] \subseteq \sigma(C) = C \tag{124}$$
$$\Rightarrow \sigma c_d(S) \subseteq C$$

and the result is established.∎

Finally, we may use these results to show that *d*-convex sets are equivalently characterized as *fixed points* of the *d*-convex solidification function, $c_{\sigma d} : \mathcal{R} \to \mathcal{R}$:

Theorem A.4 (*d*-**Convex Solid Fixed Points**) *For all $S \in \mathcal{R}$,*

$$S \in \mathcal{R}_{\sigma d} \Leftrightarrow c_{\sigma d}(S) = S \tag{125}$$

Proof: If $c_{\sigma d}(S) = S$ then by Theorem A.2, $S \in \mathcal{R}_{\sigma d}$. Conversely, if $S \in \mathcal{R}_{\sigma d}$ then since $S \in \mathcal{R}_{\sigma d} \subseteq \mathcal{R}_d$ implies from Proposition A.3 that $c_d(S) = S$, we may conclude from Lemma A.2 that

$$c_{\sigma d}(S) = \sigma[c_d(S)] = \sigma(S) = S \tag{126}$$

and the result is established.∎

References

[1] Akaike, Hirotsugu (1973), "Information Theory as An Extension of the Maximum Likelihood Principle," in Petrov, B.N., and Csaki, F. (eds.) *Second International Symposium on Information Theory* (Budapest: Akademiai Kiado), 267-281.

[2] Beckmann, Martin J. (1958), "City Hierarchies and the Distribution of City Size," *Economic Development and Cultural Change*, **6**, 243-248.

[3] Berge, Claude (1963), *Topological Spaces* (New York: MacMillan).

[4] Besag, Julian and Newell, James (1991), "The Detection of Clusters in Rare Diseases," *Journal of the Royal Statistical Society*, *Series A*, **154**, 143-155.

[5] Burnham, Kenneth P. and Anderson, David A. (2002), *Model Selection and Multimodel Inference*, Second Edition (New York: Springer-Varlag).

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

[7] Castro, Marcia. C. and Singer, Burton H. (2005), "Controlling the False Discovery Rate: A New Application to Account for Multiple and Dependent Tests in Local Statistics of Spatial Association," *Geographical Analysis*, **38**, 180-208.

[8] 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).

[9] Cliff, Andrew D. and Ord, John K. (1973), *Spatial Autocorrelation* (London: Pion).

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

[11] Dasgupta, Abhijit, and Raftery, Adrian E. (1998), "Detecting Features in Spatial Point Processes with Clutter via Model-Based Clustering," *Journal of the American Statistical Association*, **93**, 294-302.

[12] Diggle, Peter J. (2003), *Statistical Analysis of Spatial Point Patterns* (London: Oxford University Press).

[13] Dijkstra, Edsger W. (1959),"A Note on Two Problems in Connexion with Graphs," *Numerische Mathematik*, **1**, S.269-271.

[14] Duchet, Pierre. (1988), "Convex Sets in Graphs II: Minimal Path Convexity," *Journal of Combinatorial Theory*, *Series B*, **44(3)**, 307-316.

[15] Duranton, Gilles and Overman, Henry G. (2005), "Testing for Localization Using Micro-Geographic Data,"*Review of Economic Studies*, **72**, 1077-1106.

[16] Eaton, B. Curtis and Lipsey, Richard G. (1982), "An Economic Theory of Central Places," *Economic Journal*, **92(365)**, 56-72.

[17] Ellison, Glenn and Glaeser, Edward L. (1997), "Geographic Concentration in US Manufacturing Industries: A Dartboard Approach," *Journal of Political Economy*, **105(5)**, 889-927.

[18] Farber, Martin, and Jamison, Robert E. (1986), "Convexity in Graphs and Hypergraphs, *SIAM Journal of Algebraic Discrete Methods*, **7**, 433-444.

[19] Fujita, Masahisa (ed.) (2005), *Spatial Economics* (Cheltenham: Edward Elger).

[20] Fujita, Masahisa and Krugman, Paul R. (1995), "When Is the Economy Monocentric?: von Thüen and Chamberlin Unified," *Regional Science and Urban Economics* **25**, 505-528.

[21] Fujita, Masahisa, Krugman, Paul R. and Mori Tomoya (1999) "On the Evolution of Hierarchical Urban Systems," *European Economic Review*, **43**, 209-251.

[22] Fujita, Masahisa, Krugman, Paul R. and Venables, Anthony J. (1999), *The Spatial Economy: Cities, Regions and International Trade* (Cambridge, MA: The MIT Press).

[23] Fujita, Masahisa and Mori, Tomoya (1997), "Structural Stability and Evolution of Urban systems," *Regional Science and Urban Economics*, **27**, 399-442.

[24] ———— (2005), "Frontiers of the New Economic Geography," *Papers in Regional Science*, **84(3)**, 307-405.

[25] ———— (2005), "Transport Development and the Evolution of Economic Geography," *Portuguese Economic Journal*, **4(2)**, 129-156.

[26] Gangnon, Ronald E. and Clayton, Murray K. (2000), "Bayesian Detections and Modeling of Spatial Disease Clustering," *Biometrics*, **56(3)**, 922-935.

[27] ————(2004), "Likelihood-Based Tests For Localized Spatial Clustering of Disease," *Environmetrics*, **15**: 797-810.

[28] Gottman, Jean (1961), *Megalopolis: The Urbanized Northern Seaboad of the United States* (New York: The Twentieth Century Fund).

[29] Harrary, Frank, and Nieminen, John (1981), "Convexity in Graphs," *Journal of Differential Geometry*, **16**, 185-190.

[30] Haughton, Dominique M.A. (1988), "On the Choice of a Model to Fit Data From an Exponential Family," *The Annals of Statistics*, **16(1)**, 342-355.

[31] Henderson, J. Vernon and Thisse, Jacques-Francçis (eds.) (2004), *Handbook of Regional and Urban Economics*, Vol.4 (Amsterdam: North-Holland).

[32] Hokkaido-chizu, Co. Ltd. (2002), *GIS Map for Road*.

[33] Hsu, Wen-Tai (2009), "Central Place Theory and the City Size Distribution" (Manuscript, Department of Economics, Chinese University of Hong Kong).

[34] Ioannides, Yannis M. and Overman, Henry G. (2004), "Spatial Evolution of the US Urban System," *Journal of Economic Geography*, **4(2)**, 131-156.

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

[36] Japan Statistics Bureau (2001), *Establishments and Enterprise Census of Japan* (in Japanese).

[37] Konnou, Yoshinori (2003), "Jidosha Buhin Torihiki No Open-ka No Kensho," *Keizaigaku Ronshu* **68(4)**, 58-86 (in Japanese).

[38] Kontkanen, Petri, and Myllymäki, Petri (2005), "Analyzing the Stochastic Complexity via Tree Polynomials" (Technical Report, 2005-4, Helsinki Institute for Information Technology).

[39] Kontkanen, Petri, Buntine, Wray, Myllymäki, Petri, Rissanen, Jorma, and Tirri, Henry (2003), "Efficient Computation of Stochastic Complexity," in Bishop, C.M. and Frey, B.J. (eds.) *Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics*, 233-238.

[40] Krugman, Paul R. (1993), "On the Number and Location of Cities," *European Economic Review*, **37**, 293-298.

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

[42] Kulldorff, Martin (1997), "A Spatial Scan Statistic," *Communications in Statistics-Theory and Methods*, **26**, 1481-1496.

[43] Kulldorff, Martin and Nagarwalla, Neville (1995), "Spatial Disease Clusters: Detection and Inference," *Statistics in Medicine*, **14**, 799-810.

[44] Lösch, August (1940), *The Economics of Location* (Jena, Germany: Fischer). English translation by Stolper, Wolfgang F. and Woglom, William H. (1954) (New Haven, CT: Yale University Press).

[45] Marshall, John U. (1989), *The Structure of Urban Systems* (Toronto: University of Toronto Press).

[46] Mori, Tomoya (1997), "A Modeling of Megalopolis Formatin: the Maturing of City Systems," *Journal of Urban Economics* **42**, 133-157.

[47] Mori, Tomoya, Nishikimi, Koji and Smith, Tony E. (2005), "A Divergence Statistic for Industrial Localization," *Review of Economics and Statistics*, **87(4)**, 635-651.

[48] ———— (2008), "The Number-Average Size Rule: A New Empirical Relationship between Industrial Location and City Size," *Journal of Regional Science*, **48(1)**, 165-211.

[49] Mori, Tomoya and Smith, Tony E. (2009a), "A Reconsideration of the NAS Rule from An Industrial Agglomeartion Perspective," *the Brookings-Wharton Papers on Urban Affairs* (Washington, D.C.: The Brookings Institution Press), 175-205.

[50] ———— (2009b), "The NAS Rule as A Consequence of Spatial Coordination between Population and Industrial Agglomeration" (Manuscript, Institute of Economic Research, Kyoto Universituy).

[51] Porter, Michael E. (1990), *The Competitive Advantage of Nations* (New York: The Free Press).

[52] Rissanen, Jorma (1996), "Fisher Information and Stochastic Complexity," *IEEE Transactions on Information Theory*, **42**, 40-47.

[53] ———— (2001), "Strong Optimality of the Normalized ML Models as Universal Codes and Information in Data," *IEEE Transactions on Information Theory*, **47**, 1712-1717.

[54] Rosentahl, Stuart S. and Strange, William C. (2004), "Evidence on the Nature and Sources of Agglomeration Economies," in Henderson, J. Vernon and Thisse, Jacques-Françŏis (eds.), *Handbook of Regional and Urban Economics*, Vol.4 (Amsterdam: North-Holland), Ch.49.

[55] Saxenian, Annalee (1994), *Regional Advantage: Culture and Competition* (Cambridge, MA: Harvard University Press).

[56] Schwarz, Gideon (1978), "Estimating the Dimension of A Model," *Annals of Statistics*, **6(2)**, 461-464.

[57] Shibata, Ritei (1981), "An Optimal Selection of Regression Variables," *Biometrika*, **68(1)**, 45-54.

[58] Shtarkov, Y.M. (1987), "Universal Sequential Coding of Single Messages," *Problems of Information Transmission*, **23**, 3-17.

[59] Silverman, Bernard W. (1986), *Density Estimation for Statistics and Data Analysis* (Boca Raton, FL: Chapman & Hall).

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

[61] Statistical Information Institute for Consulting and Analysis (2003), *Toukei de Miru Shi-Ku-Cho-Son no Sugata* (in Japanese).

[62] Smith, Tony E. and Mori, Tomoya (2007), "On the Model-Based Clustering for the Detection of Industrial Agglomerations," in progress.

[63] Soltan, V.P. (1983), "D-Convexity in Graphs," *Soviet Mathematics-Doklady*, **28**, 419-421.

[64] Tabuchi, Takatoshi and Thisse, Jacques-Françŏis (2006), "Regional Specialization, Urban Hierarchy, and Commuting Costs," *International Economic Review*, **47**, 1295-1317.
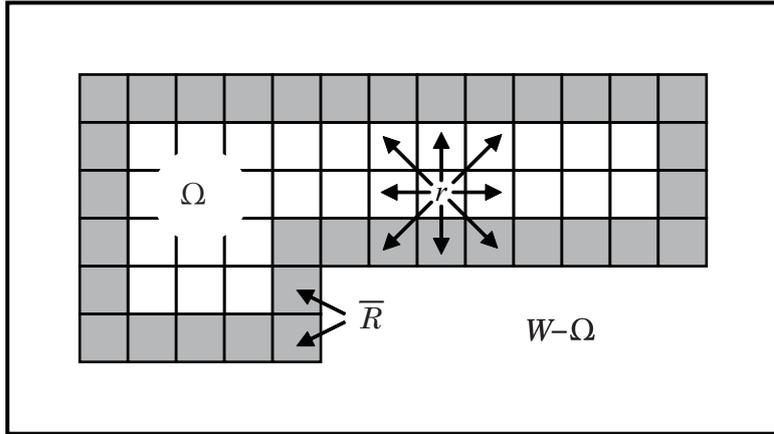
Figure 4.1. Geographical framework
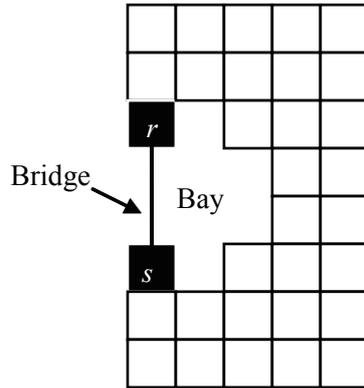
Figure 4.2. Bridge example

(R, L)

$t = 1$

$t = 3$
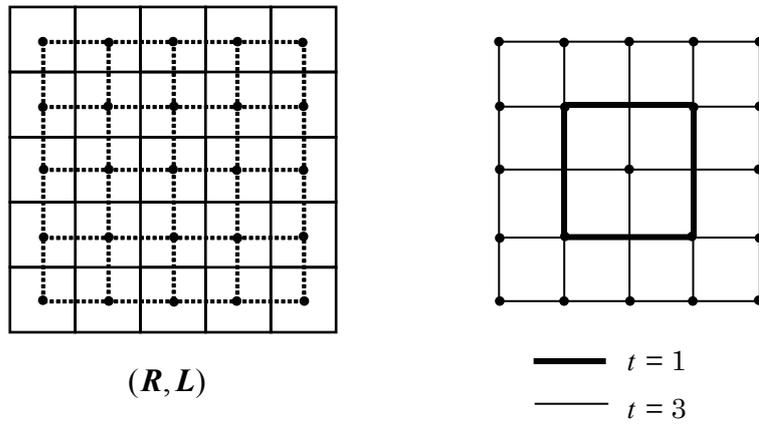
Figure 4.3. Regional network example
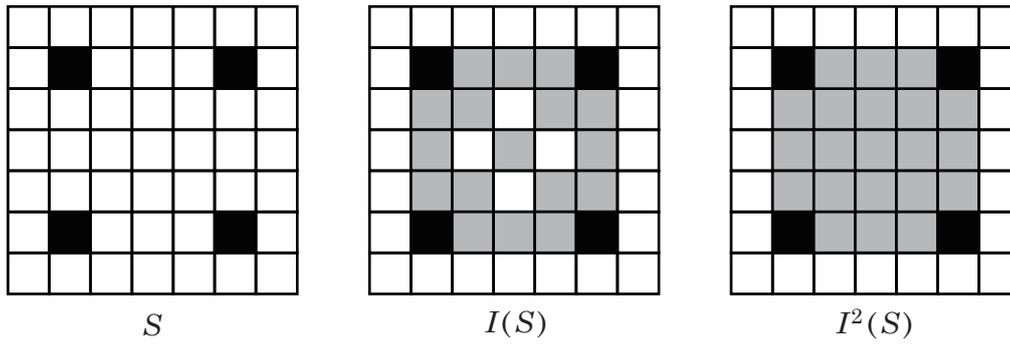
Figure 4.4. *d*-convexification of sets
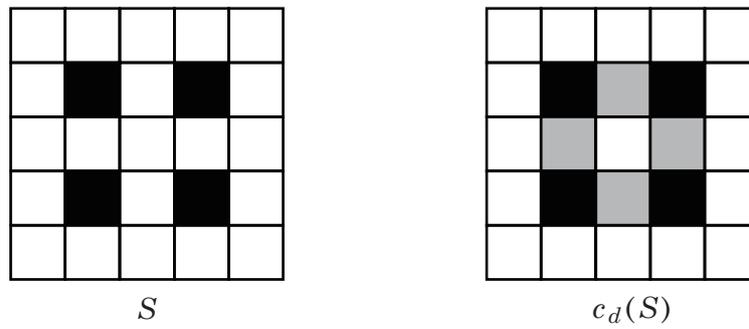
Figure 4.5. *d*-convex set with a hole
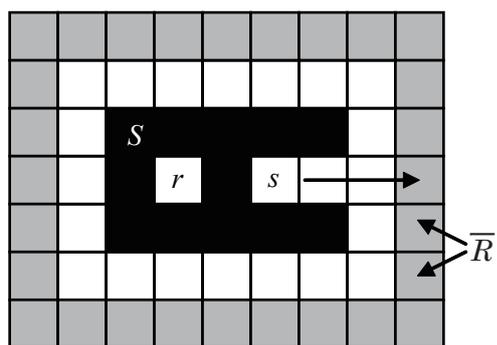
Figure 4.6. Inside versus outside
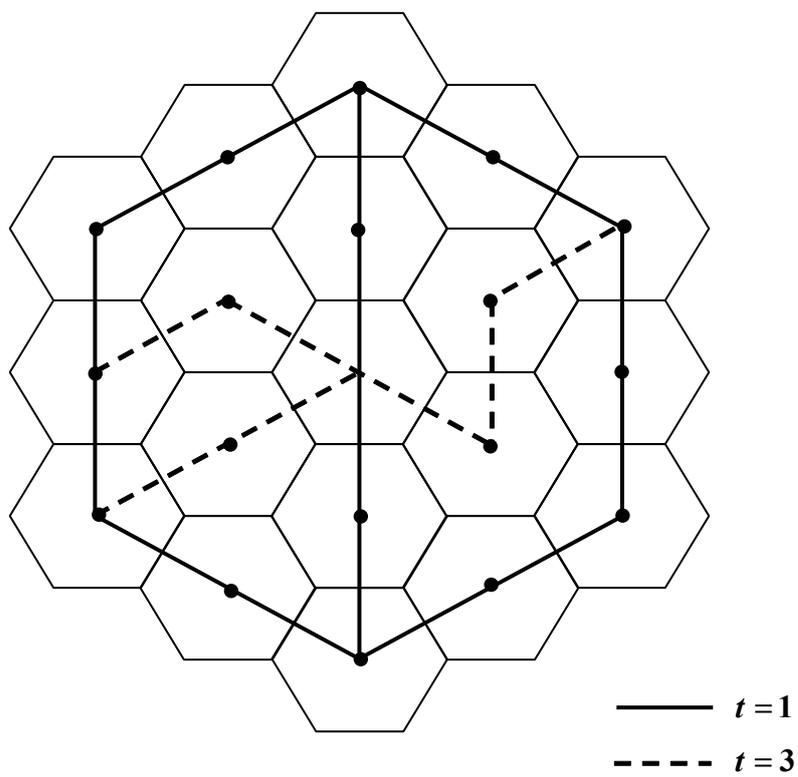
Figure 4.7. Regional subsystem
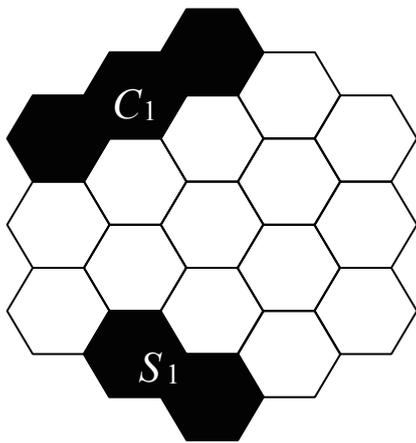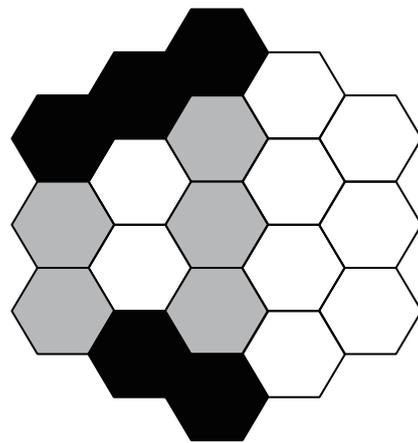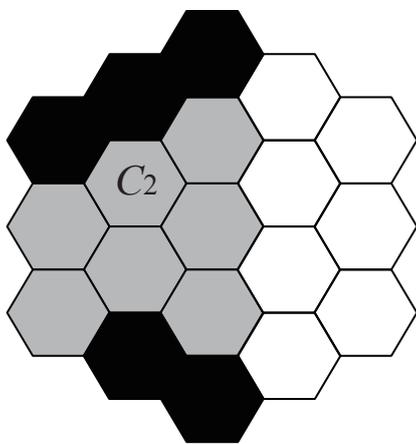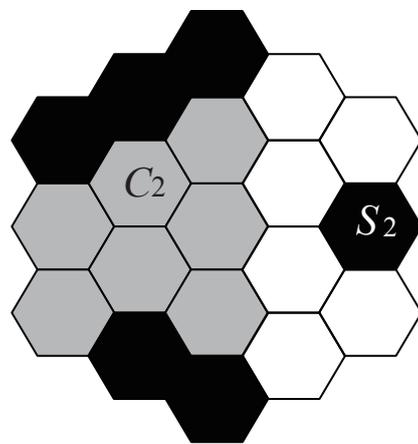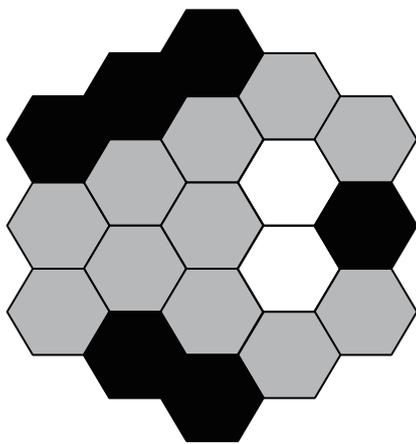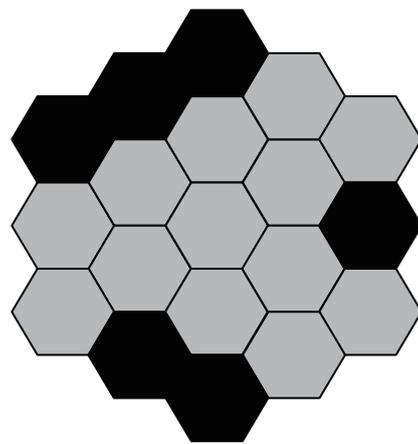
Figure 4.8. Formation of composite clusters
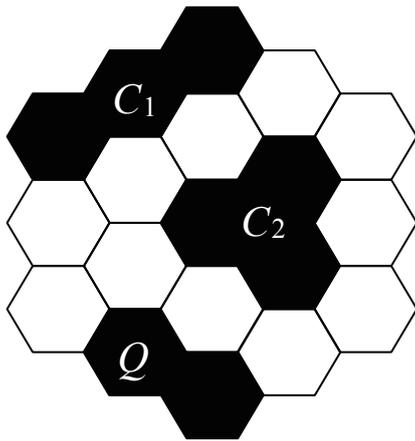
Stage 1
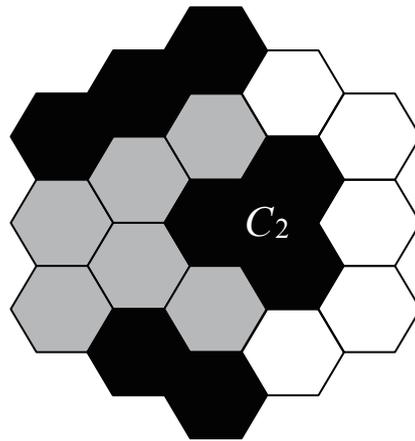
Stage 2
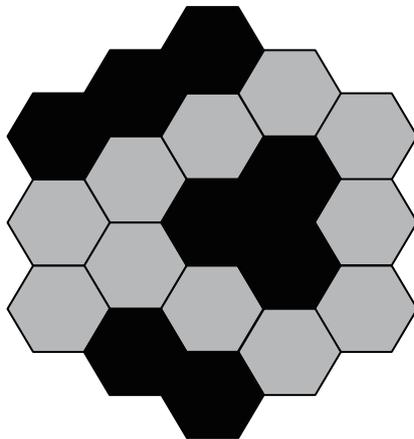
Stage 3

Figure 5.1. Formation of composite clusters

(a) BIC-increment threshold

(b) Employment-share threshold
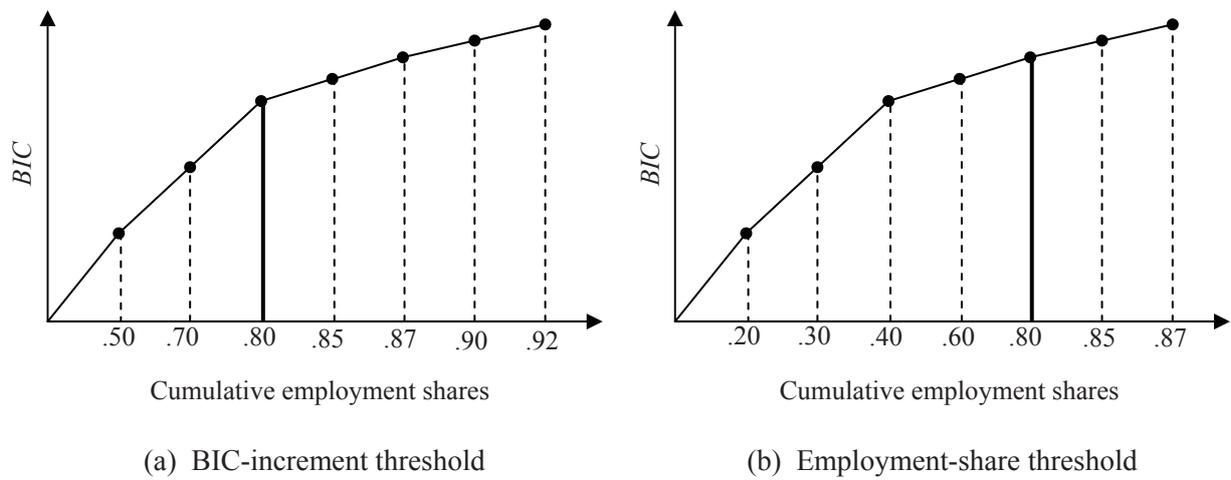
Figure 6.1. Thresholds for essential containment

(a) Globally confined and locally dense

(b) Globally dispersed and locally dense

(c) Globally confined and locally sparse

(d) Globally dispersed and locally sparse
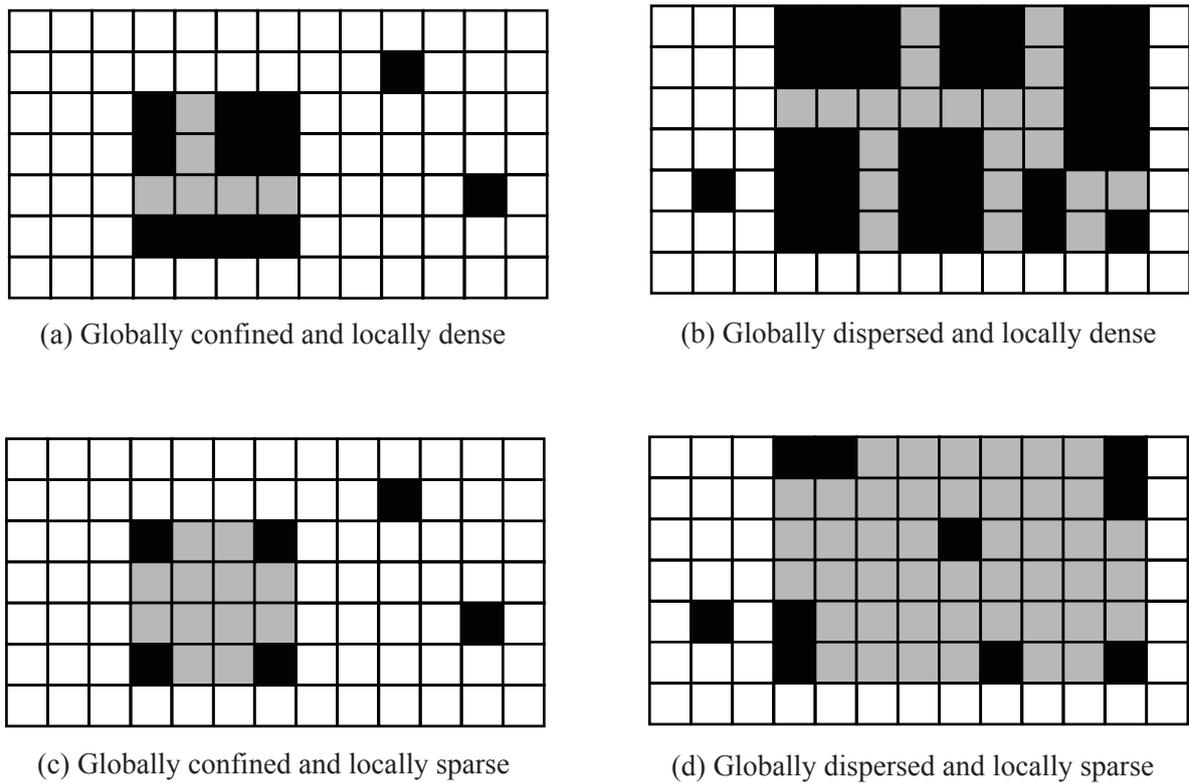
Figure 6.2. Classifications of agglomeration patterns
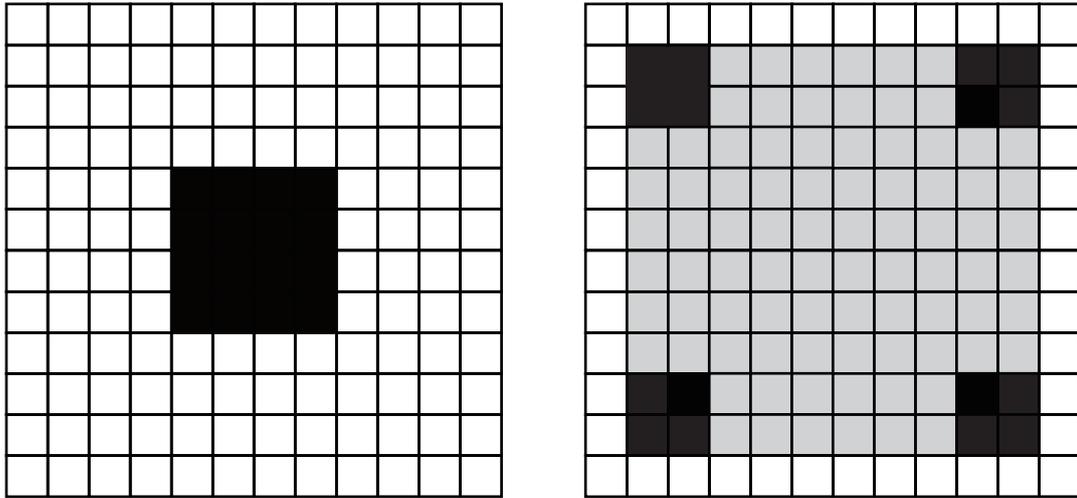
Figure 6.3. Cluster patterns and the degree of agglomeration
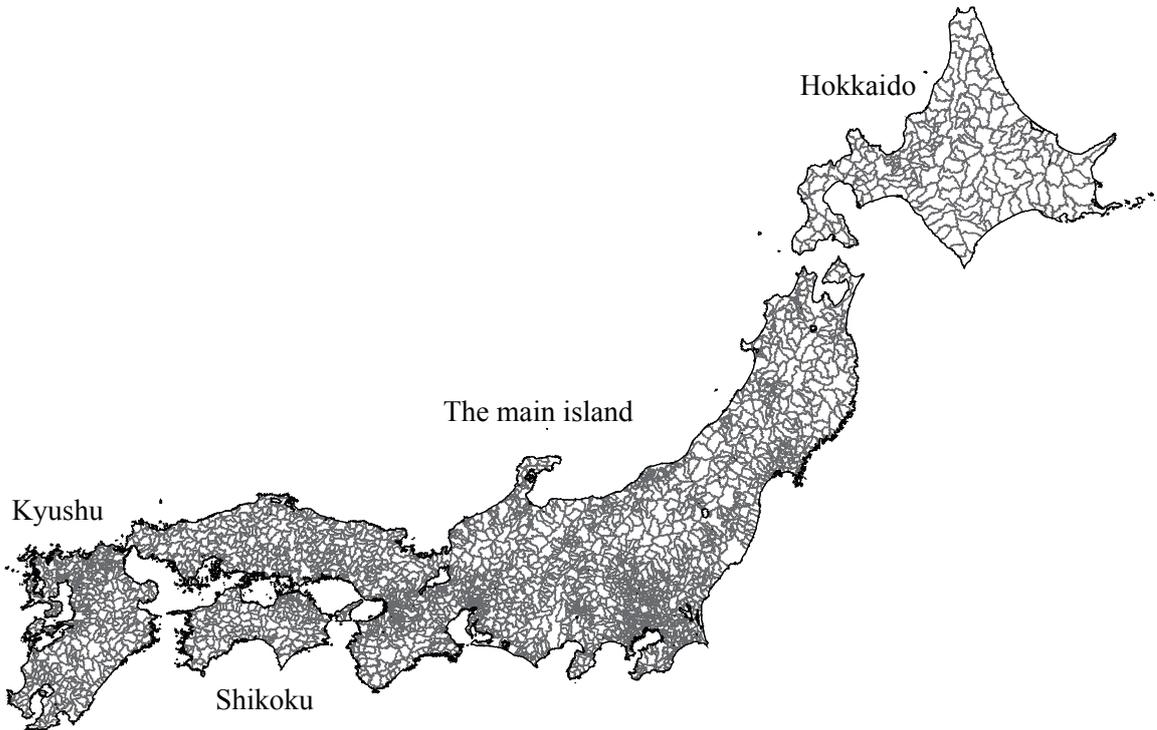
Figure 7.1. Basic regions (shi-ku-cho-son) of Japan
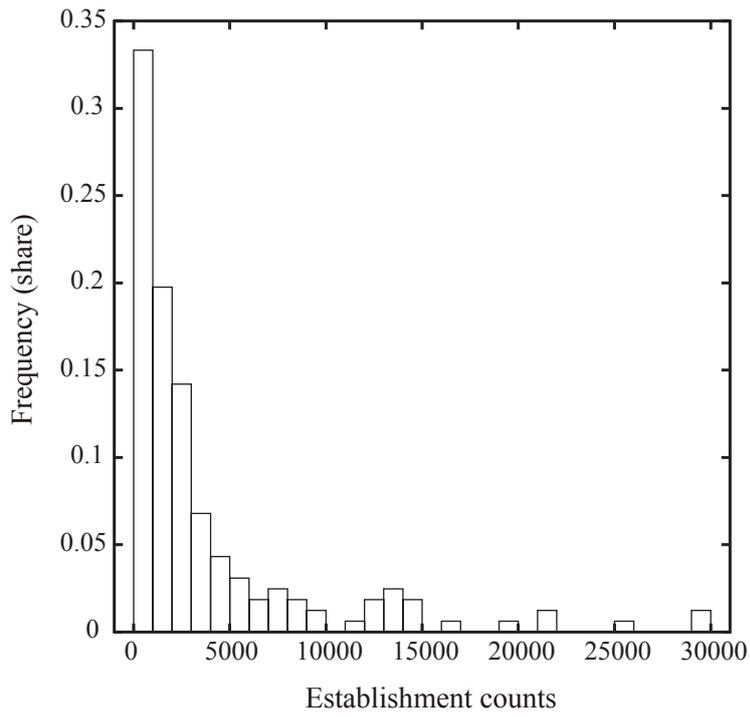
Figure 7.2. Frequency distribution of establishment counts
in Japanese three-digit manufacturing industries
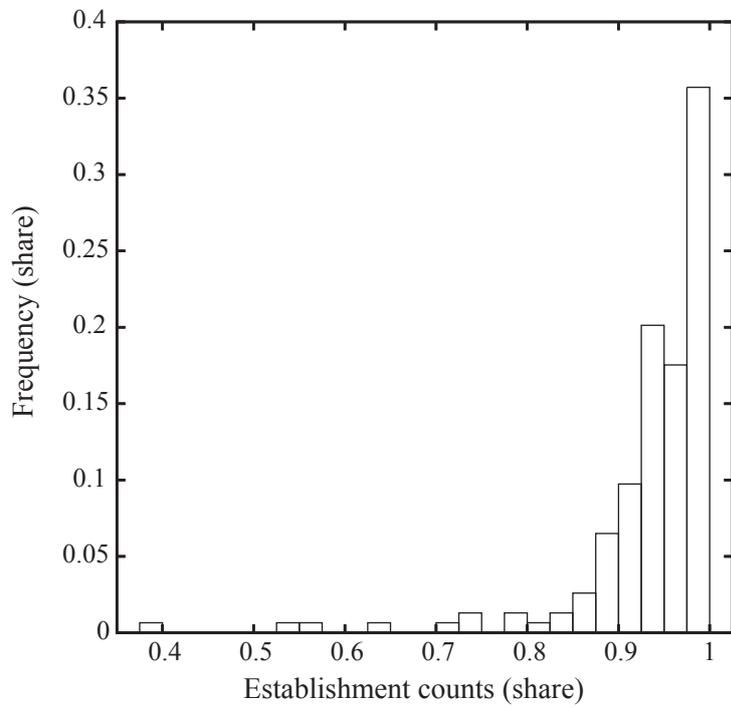
Figure 7.3. Share of establishment counts included in clusters
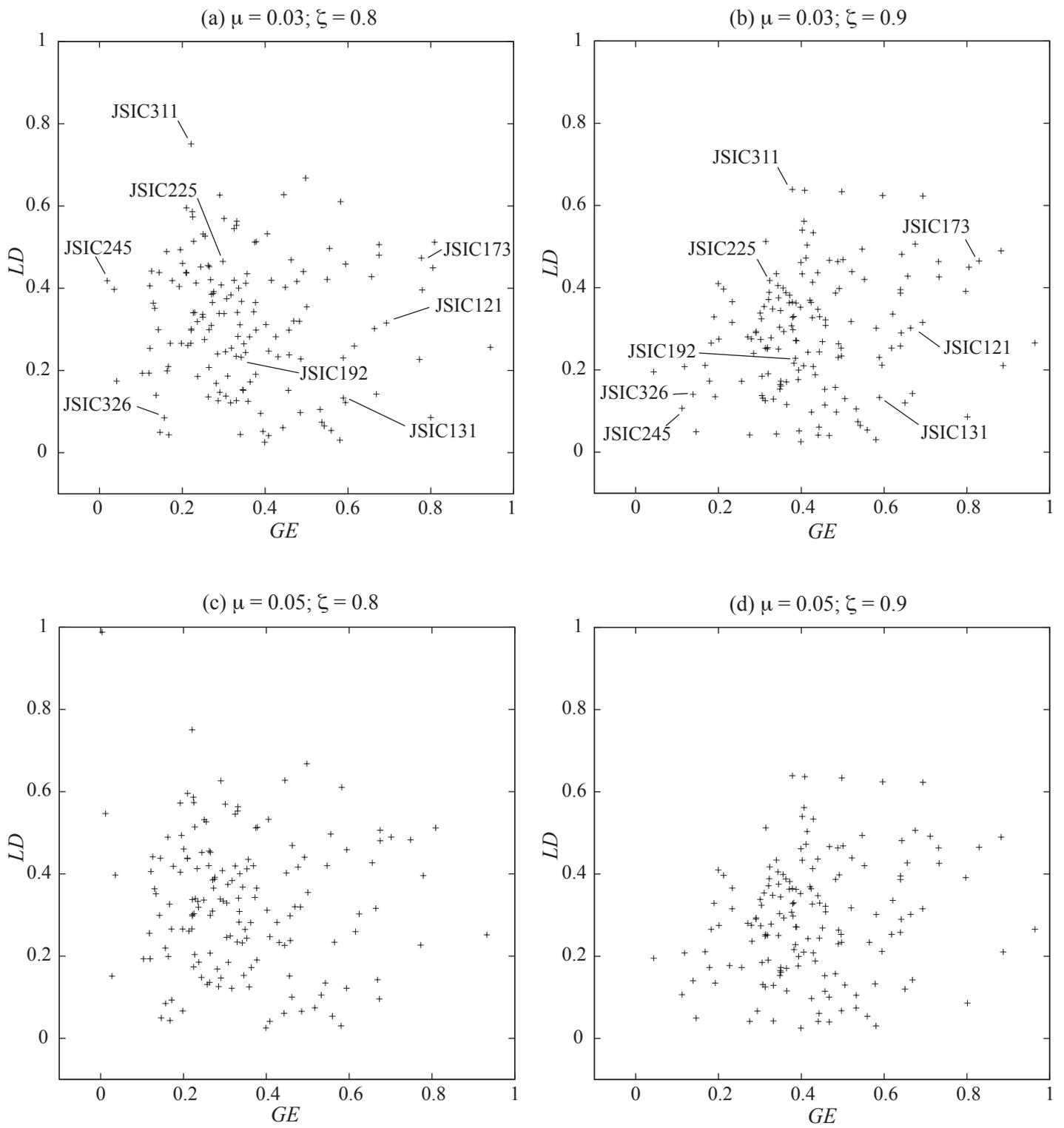
Figure 7.4. Global extent and local dispersion of clusters

(a) Density of establishments (per km^2)

■	0.601-1.274	▨	0.060-0.082
■	0.296-0.601	▨	0.043-0.060
■	0.177-0.296	▨	0.029-0.043
▨	0.117-0.177	▨	0.018-0.029
▨	0.082-0.117	▨	0.000-0.018
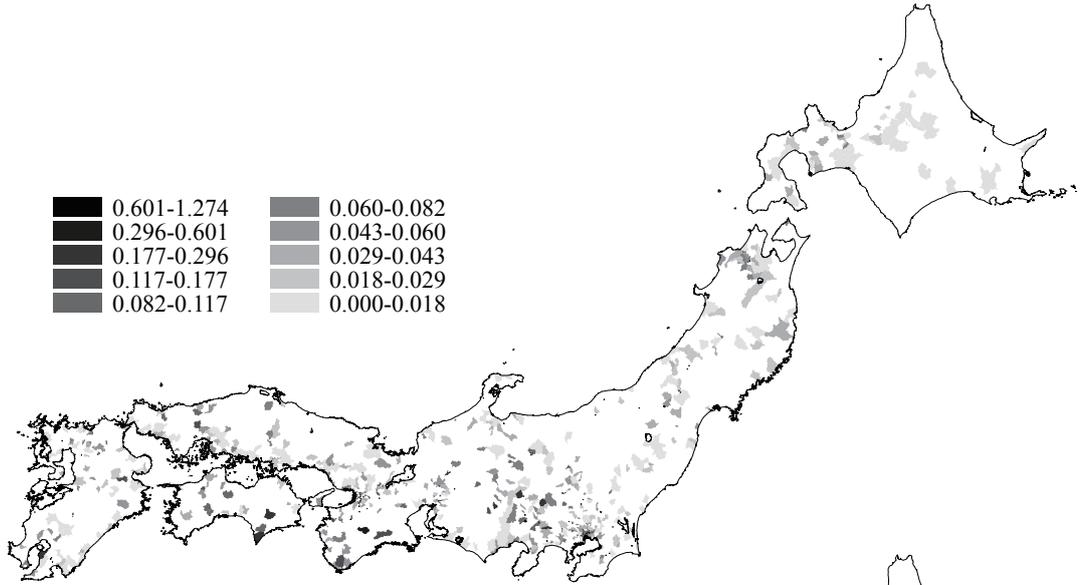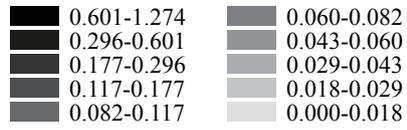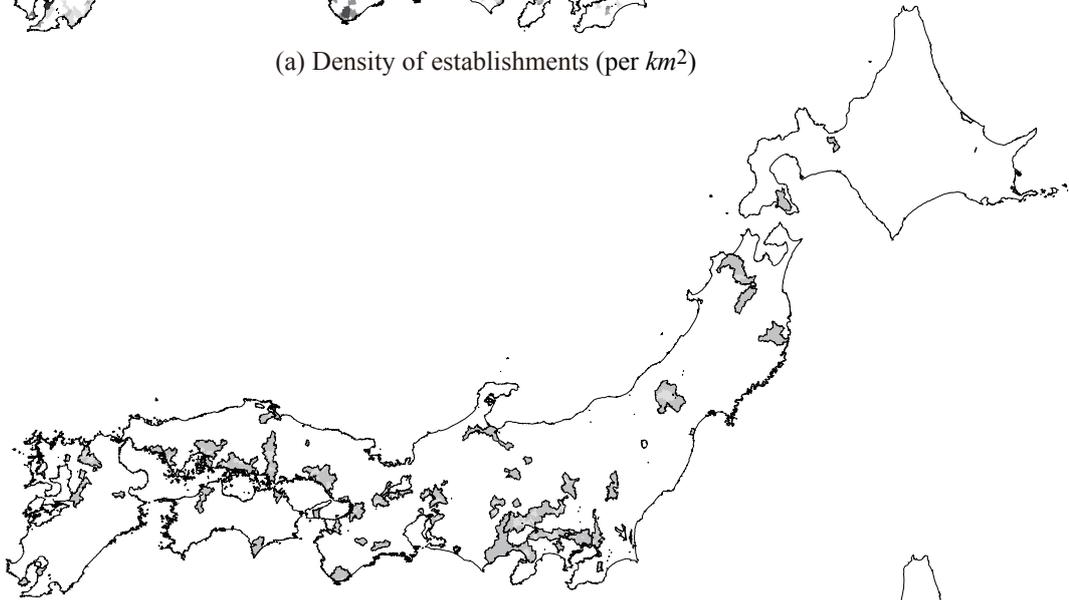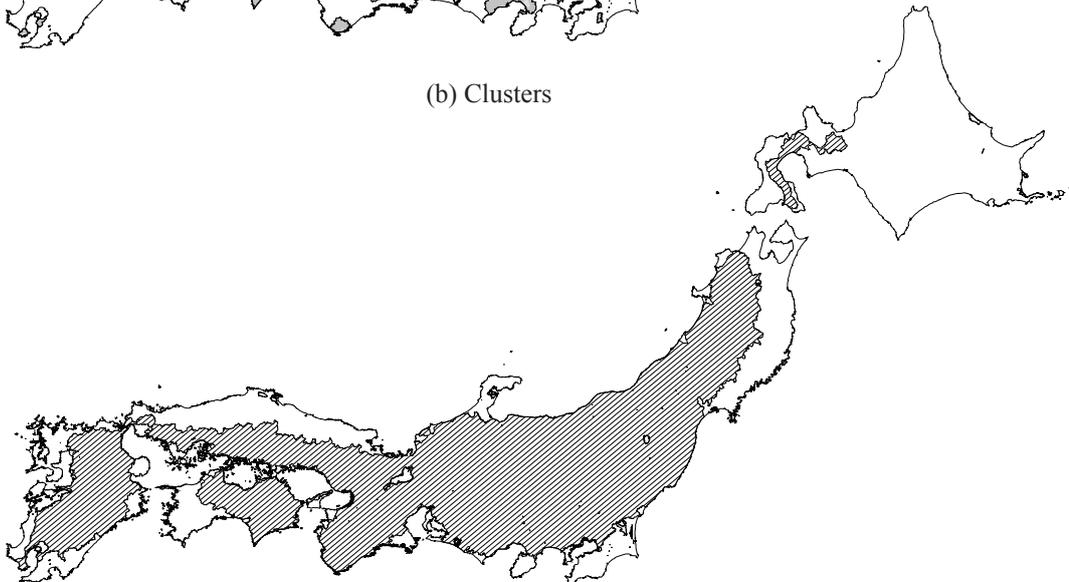
(b) Clusters

(c) Essential containment

Figure 7.5. Global dispersed and locally sparse pattern: soft drinks and carbonated water (JSIC131)

■ 0.633-0.802	▦ 0.094-0.126
■ 0.380-0.541	▦ 0.068-0.094
■ 0.244-0.364	▦ 0.047-0.068
▦ 0.172-0.241	▦ 0.028-0.046
▦ 0.127-0.169	▦ 0.003-0.028
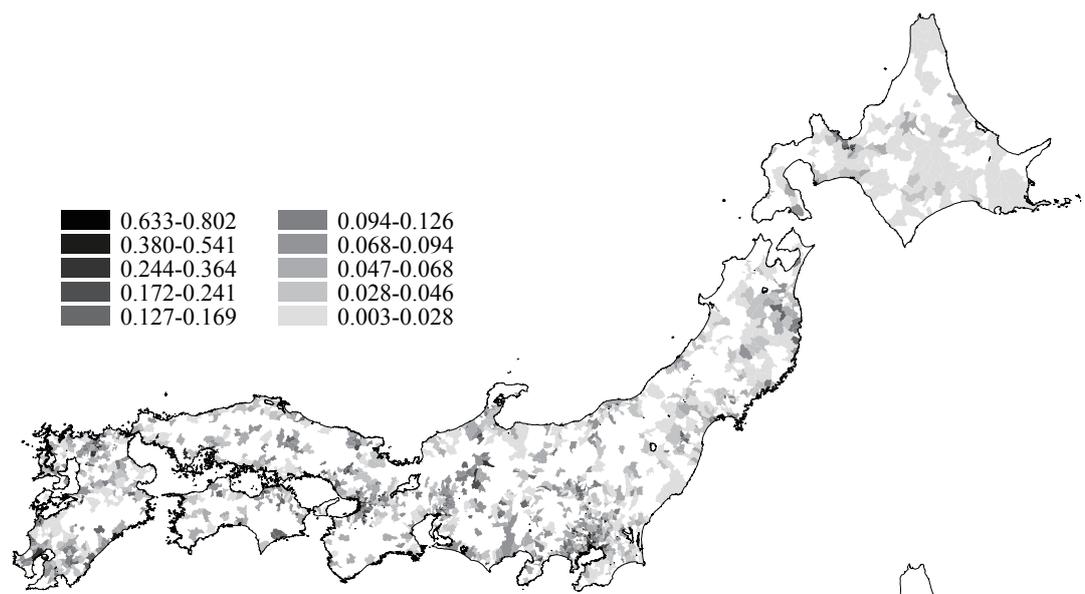
(a) Density of establishments (per km^2)
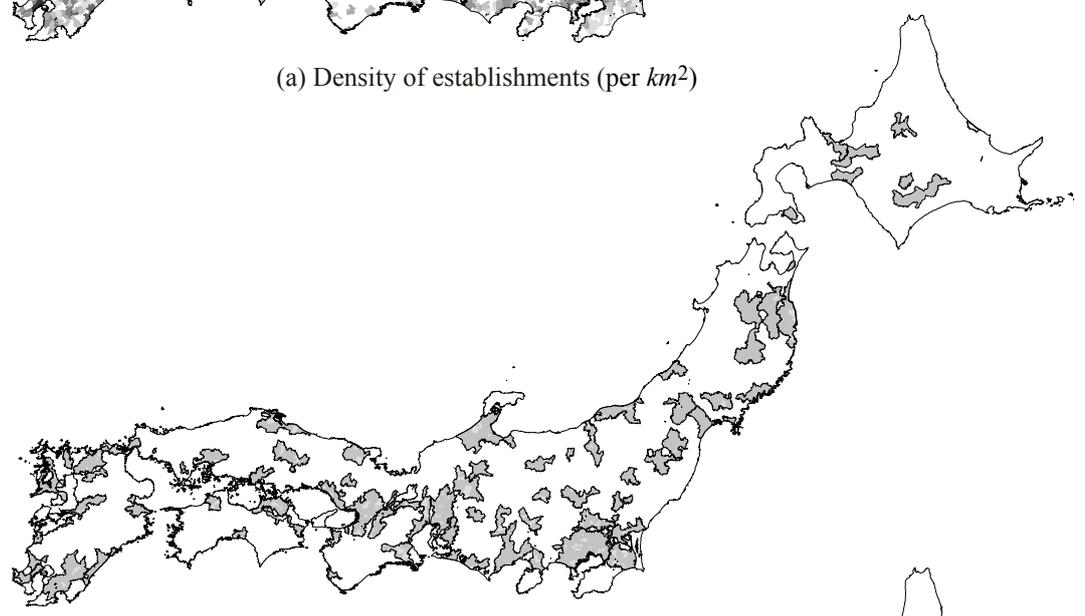
(b) Clusters
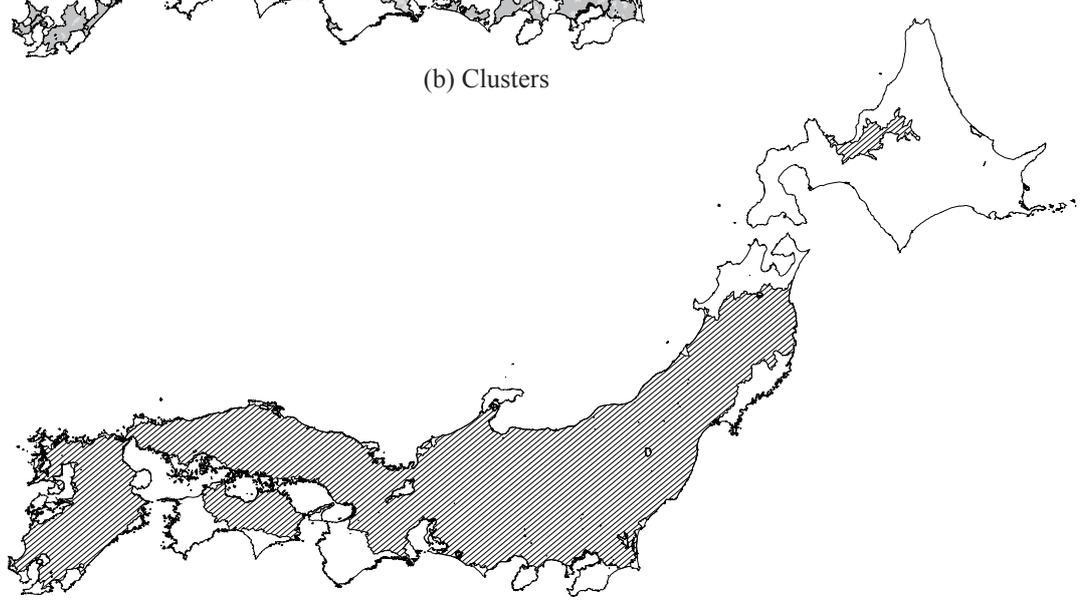
(c) Essential containment

Figure 7.6. Global dispersed and local sparse pattern: livestock products (JSIC121)

■	3.079-8.325	▨	0.292-0.421
■	1.599-3.079	▨	0.203-0.292
■	0.974-1.599	▨	0.137-0.203
■	0.616-0.974	▨	0.080-0.137
■	0.421-0.616	▨	0.000-0.796

(a) Density of establishments (per *km²*)

(b) Clusters

(c) Essential containment
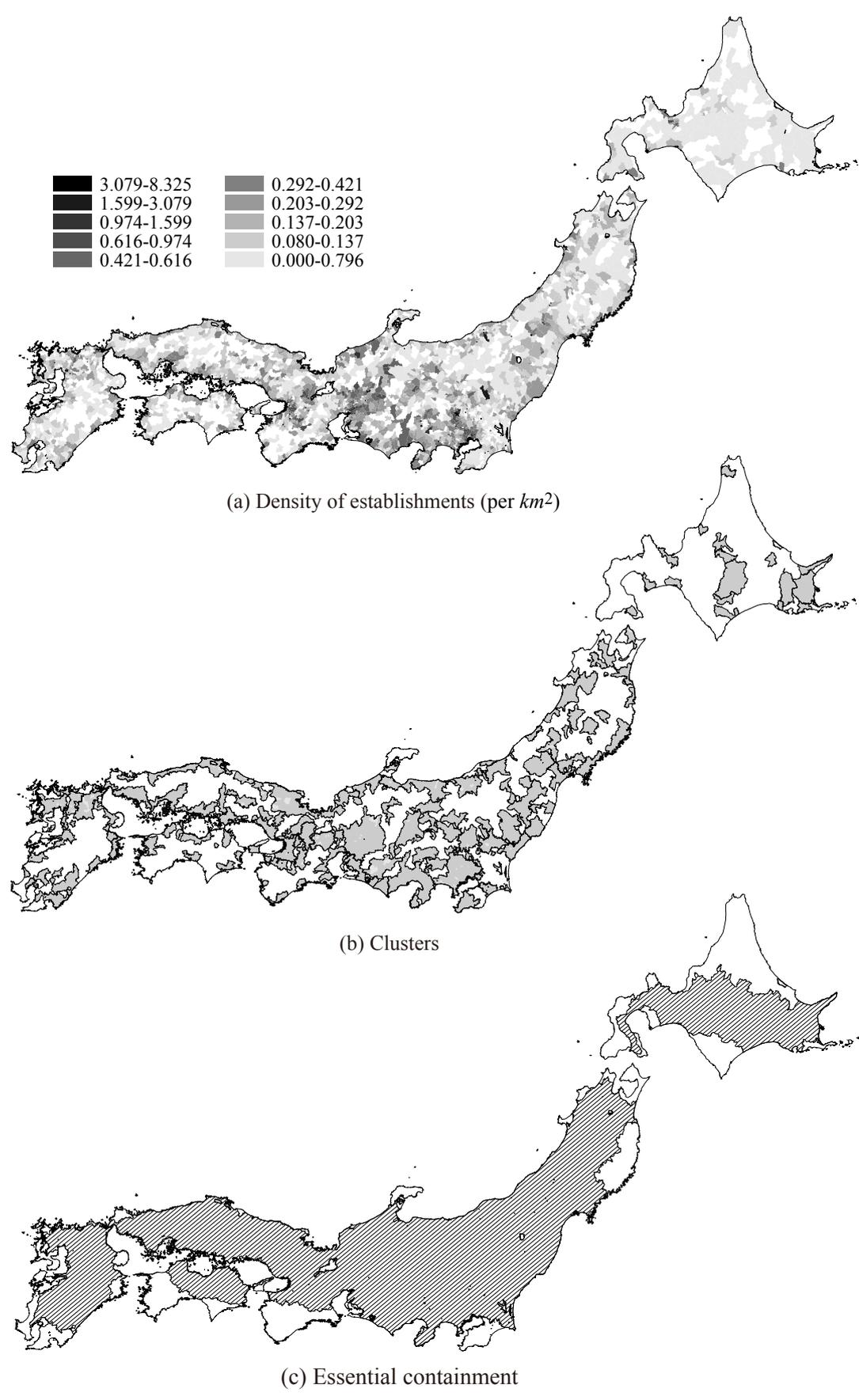
Figure 7.7. Globally dispersed and locally sparse pattern: sliding doors and screens (JSIC173)

■	12.93-12.93	▓ 0.075-0.113
■	2.553-4.745	▓ 0.051-0.066
▓	0.356-0.647	▒ 0.031-0.047
▓	0.192-0.308	░ 0.017-0.029
▓	0.123-0.173	░ 0.004-0.016
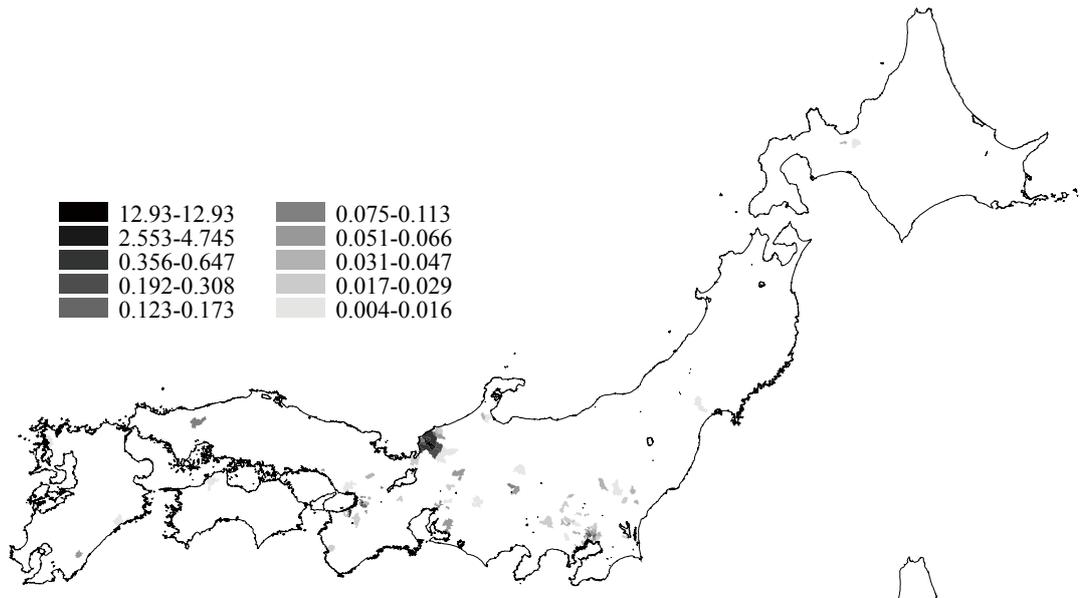
(a) Density of establishments (per km^2)
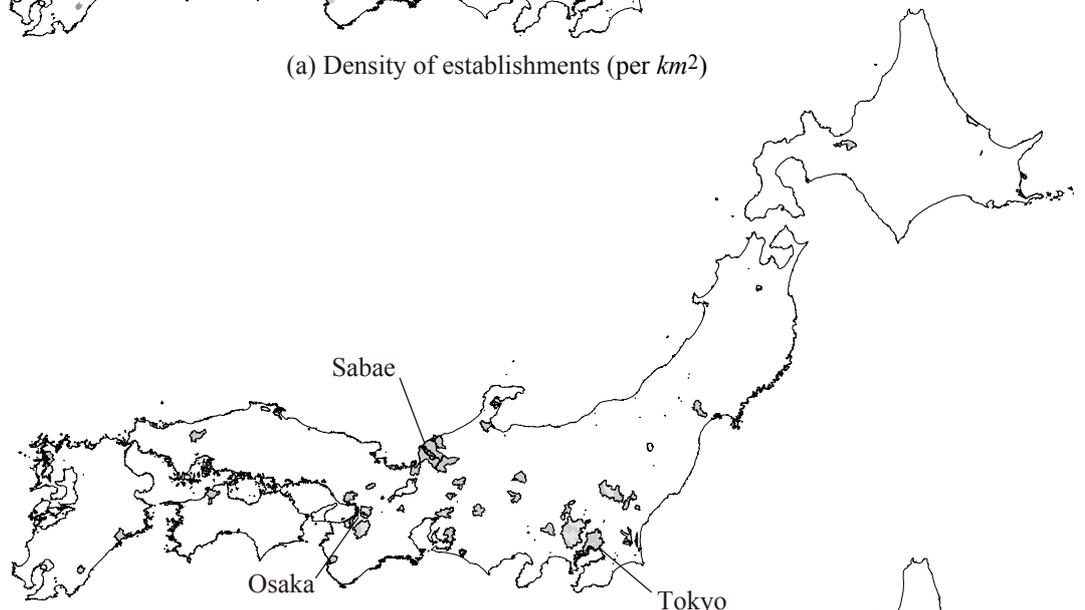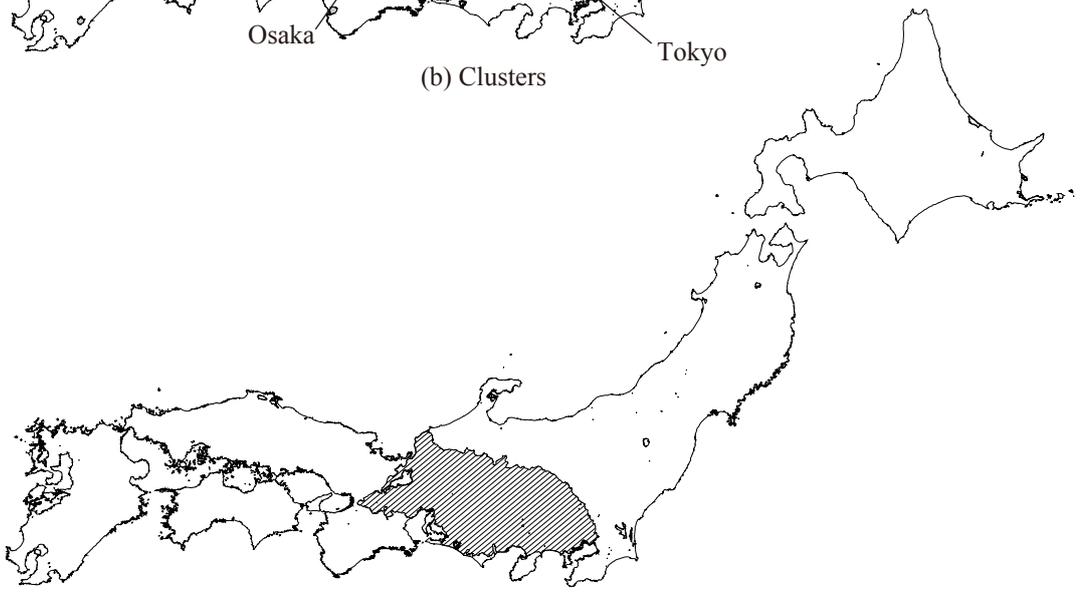
(b) Clusters

(c) Essential containment

Figure 7.8. Globally confined and locally sparse pattern: ophthalmic goods, including frames (JSIC326)

(a) Density of establishments (per km^2)

(b) Clusters

Hikita-Ohuchi-Shiratori

Tokyo
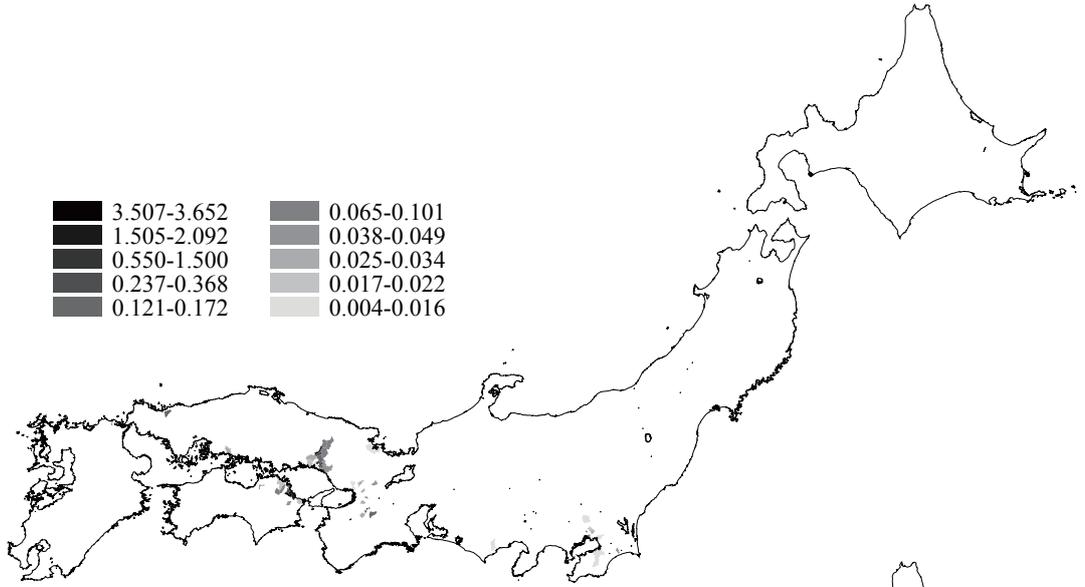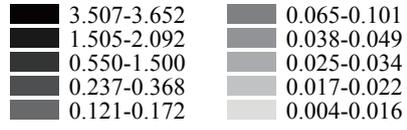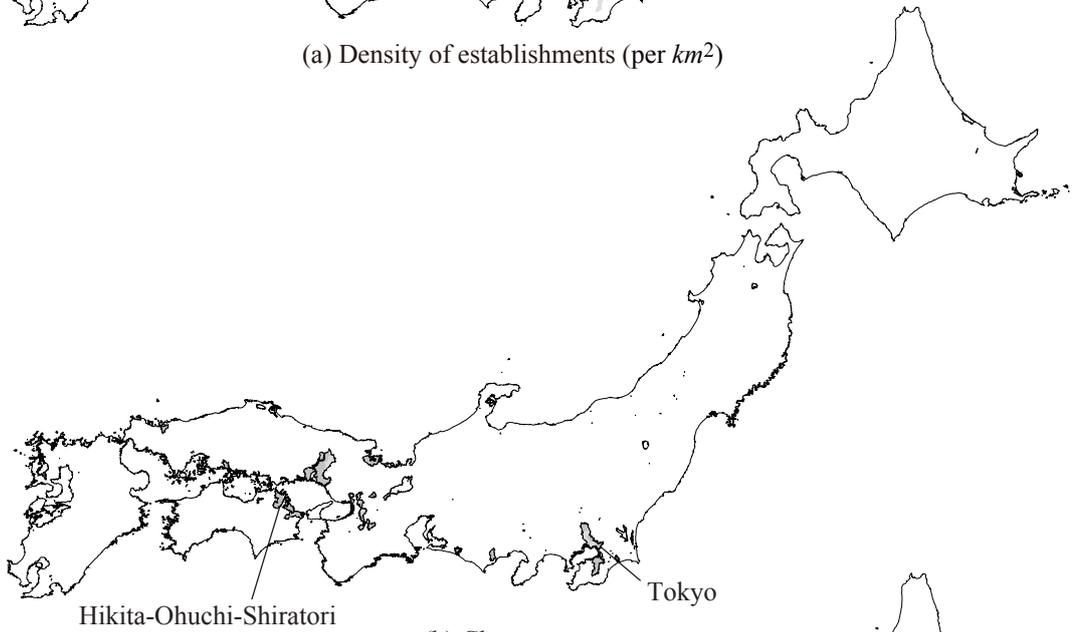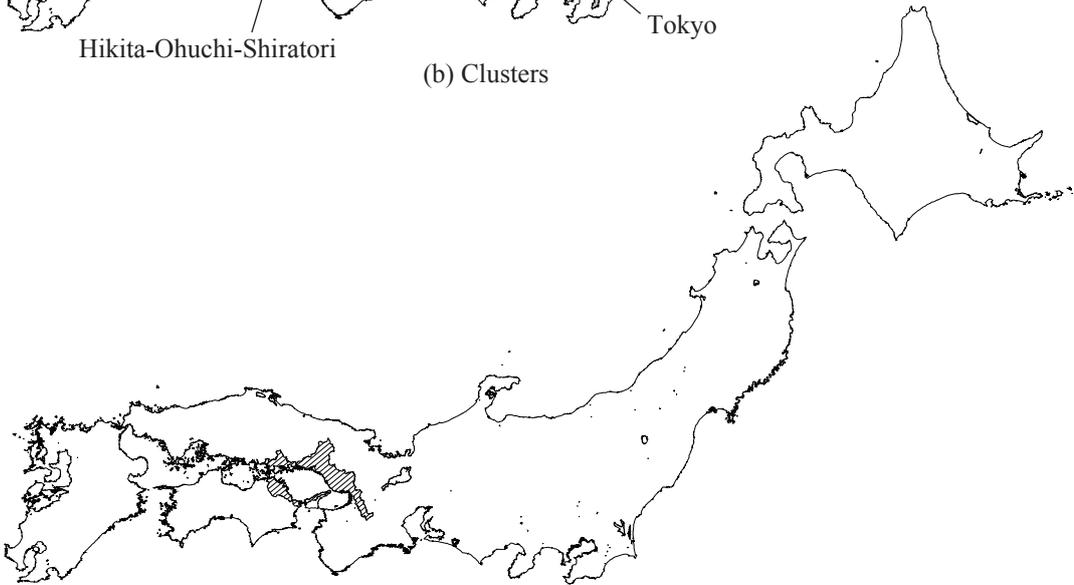
(c) Essential containment

Figure 7.9. Globally confined and locally sparse pattern: leather gloves and mittens (JSIC245)

Legend:
3.507-3.652
1.505-2.092
0.550-1.500
0.237-0.368
0.121-0.172
0.065-0.101
0.038-0.049
0.025-0.034
0.017-0.022
0.004-0.016
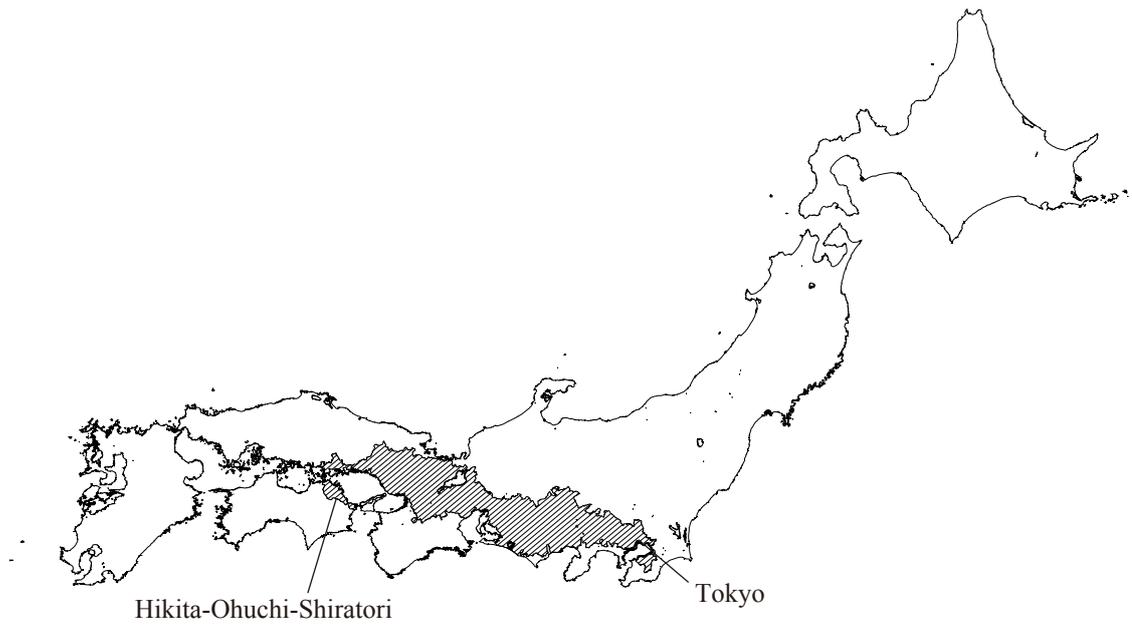
Figure 7.10. Essential containment of leather gloves and mittens (JSIC245) with $\delta = 0.03$ and $\zeta = 0.9$

(a) Density of establishments (per *km*^{2})

Fukuoka Hiroshima Kobe Nagoya

Osaka Tokyo

(b) Clusters

(c) Essential containment
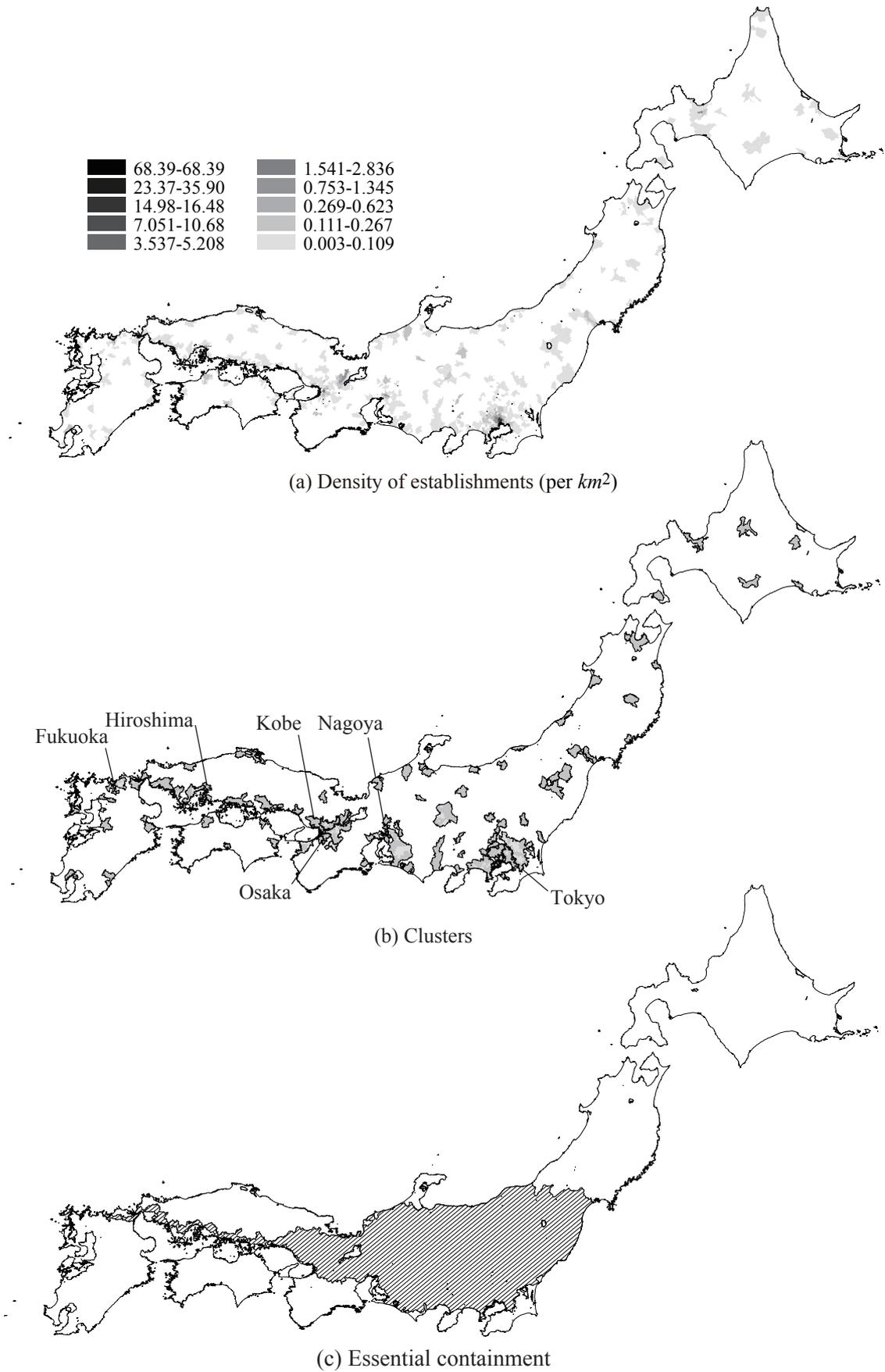
Figure 7.11. Globally confined and locally sparse pattern: publishing industries (JSIC192)

(a) Density of establishments (per *km²*)

6.683-8.554
3.983-5.455
2.679-3.810
1.717-2.549
1.231-1.709
0.772-1.199
0.453-0.768
0.248-0.448
0.118-0.247
0.003-0.117

Hiroshima   Kobe   Nagoya

Osaka   Tokyo
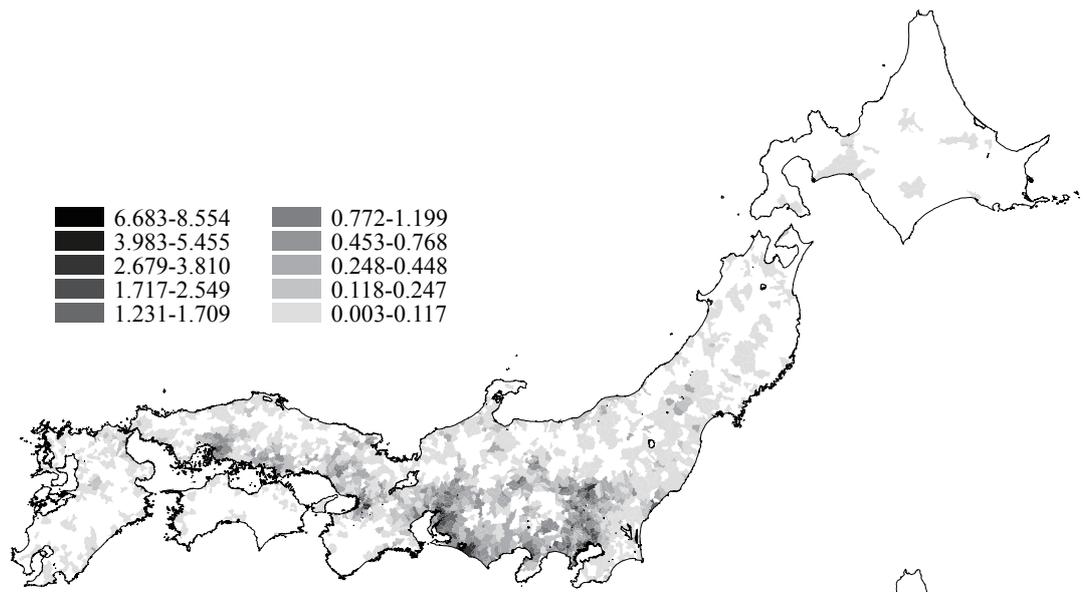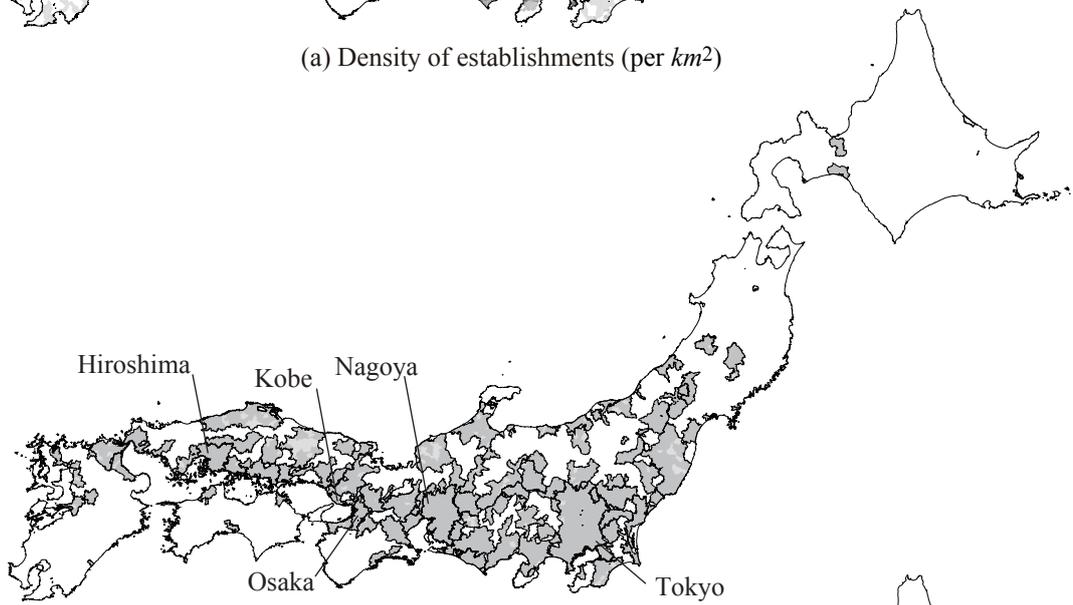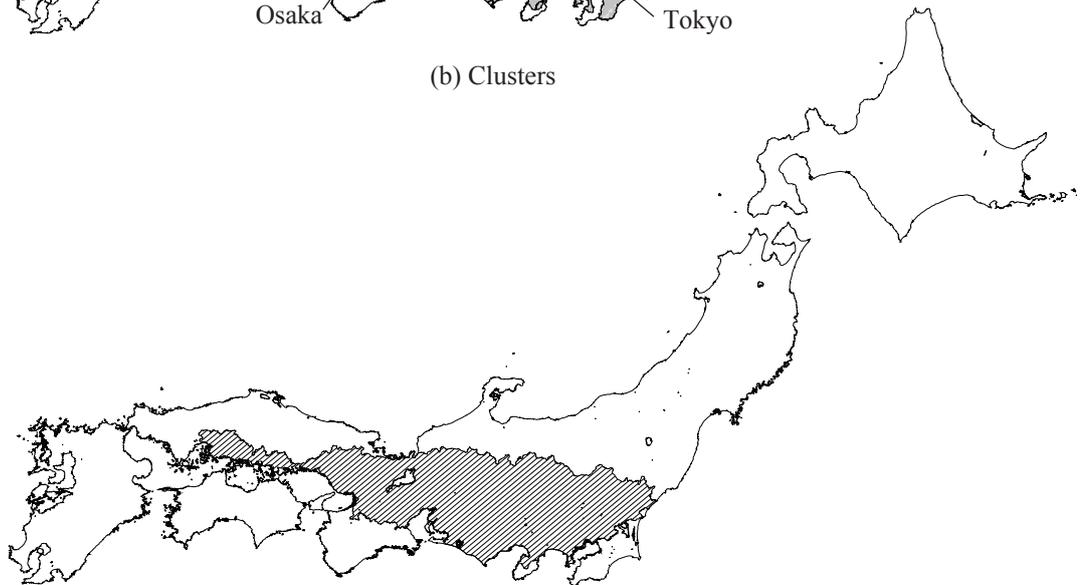
(b) Clusters

(c) Essential containment

Figure 7.12. Globally confined and locally dense pattern: motor vehicle, parts and accessories (JSIC311)

■ 1.307-1.909	▦ 0.094-1.374
■ 0.655-1.121	▦ 0.065-0.094
■ 0.360-0.568	▦ 0.044-0.065
▦ 0.219-0.336	▦ 0.026-0.043
▦ 0.139-0.214	▦ 0.004-0.025

(a) Density of establishments (per km^2)

Nagoya
Kyoto
Toyama
Hiroshima
Kobe
Osaka
Tokyo
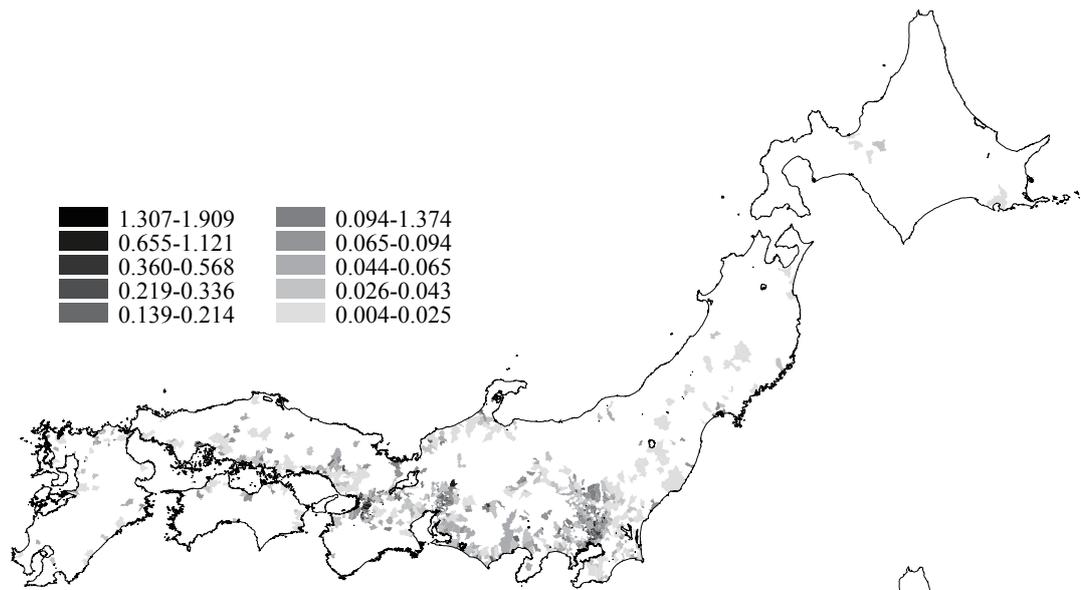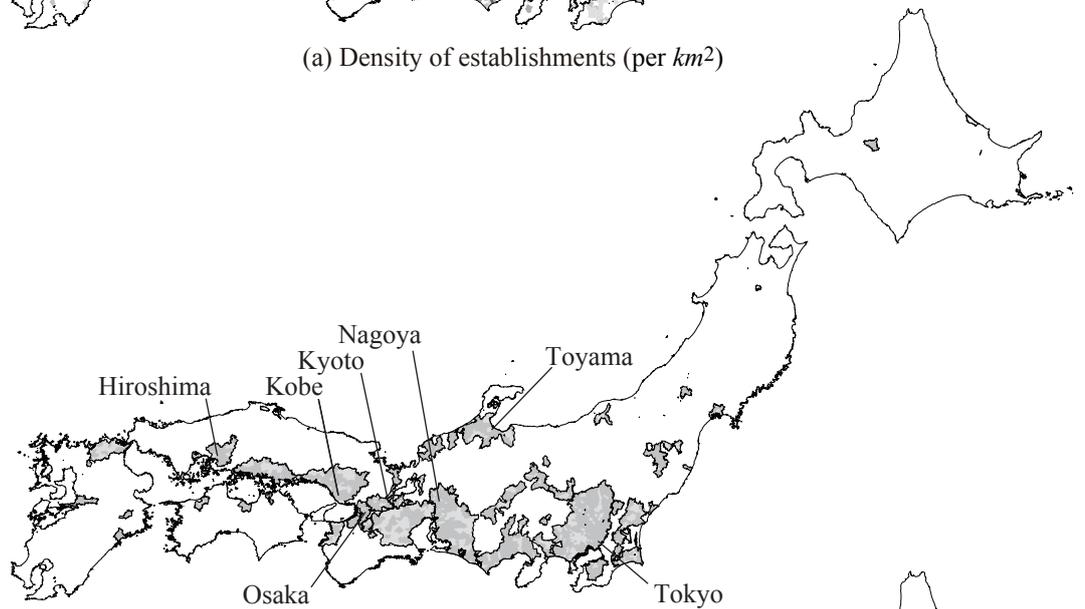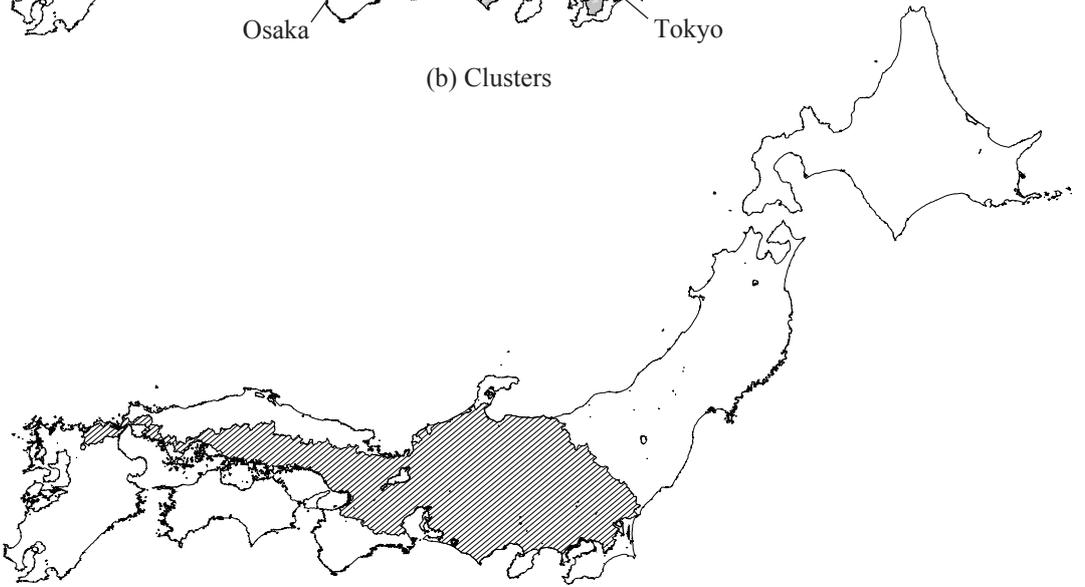
(b) Clusters

(c) Essential containment

Figure 7.13. Globally confined and local dispersed pattern: compounding plastic materials, including reclaimed plastics (JSIC225)

■ 68.39-68.39	■ 1.541-2.836
■ 23.37-35.90	■ 0.753-1.345
■ 14.98-16.48	■ 0.269-0.623
■ 7.051-10.68	□ 0.111-0.267
■ 3.537-5.208	□ 0.003-0.109

■ 1559.4-14550	■ 49.07-94.70
■ 551.37-1559.4	■ 28.05-49.07
■ 336.09-551.37	■ 16.01-28.05
■ 200.44-336.09	□ 11.27-16.01
■ 94.70-200.44	□ 0.052-11.27

(a) Density of establishments (per km^2)   (b) BIC contribution
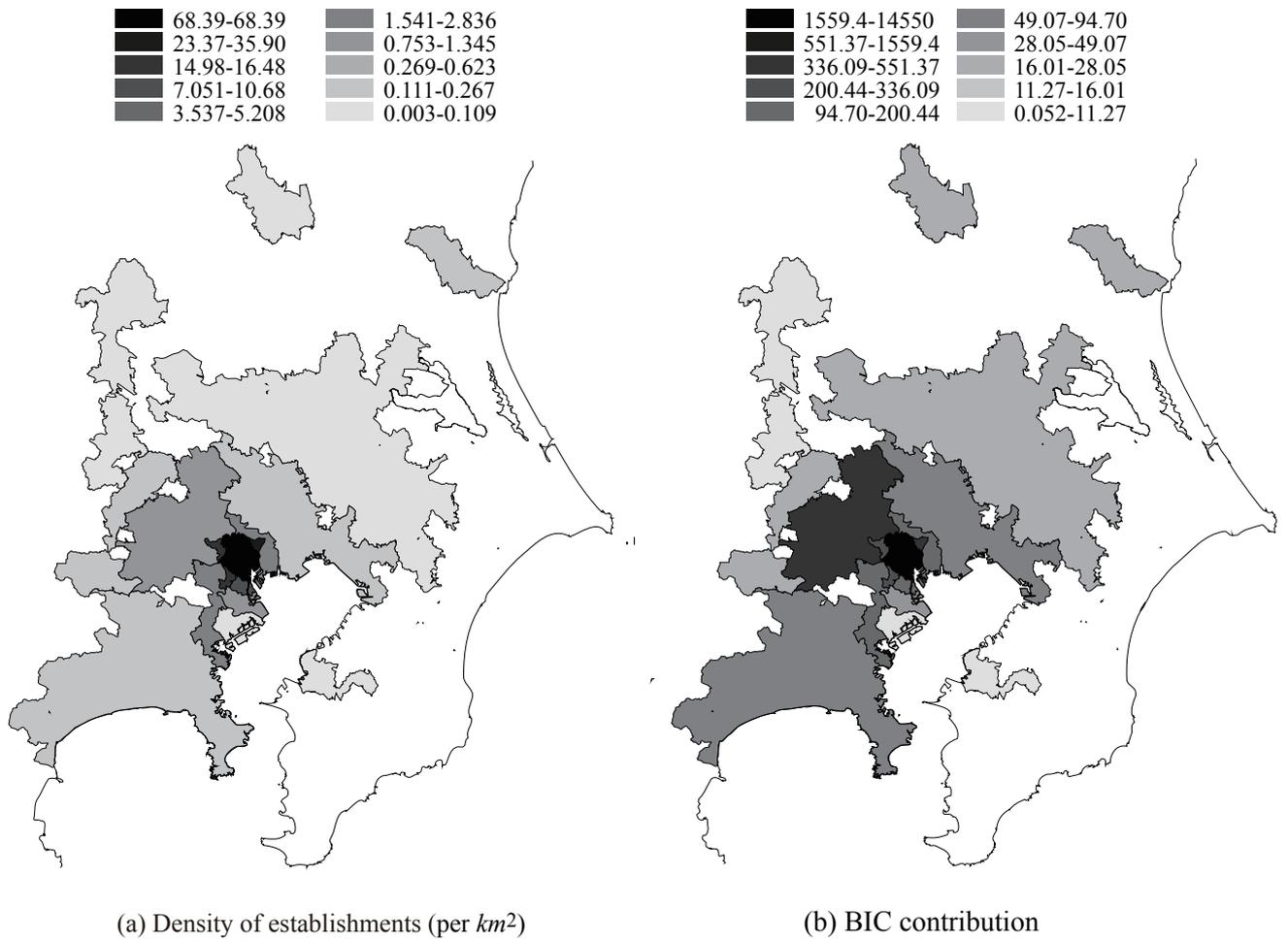
Figure 8.1. Clusters of "publishing industry" (JSIC192) around Tokyo

Figure A.1

Figure A.2