A probabilistic modeling approach to the detection of industrial agglomerations

Tomoya Mori*,**,† *and Tony E. Smith****

*Institute of Economic Research, Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501 Japan
**Research Institute of Economy, Trade and Industry (RIETI), 11th floor, Annex, Ministry of Economy, Trade and Industry (METI), 1-3-1, Kasumigaseki Chiyoda-ku, Tokyo 100-8901 Japan
***Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA
†Corresponding author: Tomoya Mori, Institute of Economic Research, Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501 Japan. *email*: <mori@kier.kyoto-u.ac.jp>

Abstract

Dating from the seminal work of Ellison and Glaeser 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 article 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.

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 10 times that of the USA, land is generally considered to be extremely scarce. Yet, more than 60% of the total population and more than 80% of total employment are concentrated in less than 3% of total area. Similar observations can be made for any other developed country. The extent of this concentration phenomenon explains why economic agglomeration is now a major topic in urban and regional economics (see, e.g. Henderson and Thisse, 2004). Industrial agglomeration has also gained increasing interest in the management literature, dating from the seminal work of Porter (1990) on 'industrial cluster theory.'

In terms of empirical work, a substantial number of studies on industrial agglomeration have been published in the recent decades. Some of them have proposed indices of industrial agglomeration that allow testable comparisons of the degree of agglomeration among industries (Brülhart and Traeger, 2005; Duranton and Overman, 2005; Mori et al., 2005; Marcon and Puech, 2010). 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 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.'[1] 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. The reason for this is that such patterns are basically *multidimensional* in nature and are not easily compared with any single index.

This can be illustrated by a sample of our results for Japanese manufacturing industries (developed in more detail in Section 5, and in our companion paper, Mori and Smith, 2011b). Here, we consider two industries that are virtually indistinguishable in terms of their overall degree of spatial concentration (as measured by the Kulback–Leibler measure of concentration sketched in Section 5). But the actual *patterns* of agglomeration for these two industries are quite different. The agglomeration pattern of the first industry, classified as 'plastic compounds and reclaimed plastics', is seen in Figure 13(b). (For now, the area marked in gray can be considered as industrial agglomerations.) The concentration of this industry lies mainly 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. Our second industry, classified as 'soft drinks and carbonated water', exhibits a very different pattern of agglomeration. As seen in Figure 14(b), this industry is spread throughout the nation, but exhibits a large number of local agglomerations. A closer inspection of these industries reveals the nature of these differences. On the one hand, plastic components constitute essential inputs to a variety of manufactured goods, from automobiles to TV sets. Hence, the concentration of this industry along the industrial belt forms a series of intermediate markets for other manufacturing industries using these components. On the other hand, soft drinks are more directly oriented to final markets serving consumers. So while there are still sufficient scale economies to warrant industrial agglomerations, these agglomerations are widely scattered and essentially follow patterns of population density.

---

1 Examples of such reference distributions are the regional distribution of all-industry employment or establishments (e.g. Ellison and Glaeser, 1997; Duranton and Overman, 2005), and that of economic area (e.g. Mori et al., 2005).

Thus, while summary measures of spatial concentration (or dispersion) are unquestionably useful for a wide range of global comparisons, the above illustration suggests that more detailed representations of spatial agglomeration patterns can in principle allow much richer types of comparisons. With this in mind, our central objective is to propose a methodology for representing and identifying such agglomeration patterns.

Before doing so, it is important to note that there have been other attempts to develop statistical measures that are more multidimensional in nature. Most notably, the *K*-density approach of Duranton and Overman (2005) utilizes pairwise distances between individual establishments and is capable of indicating the spatial extent of an agglomeration. In a similar vein, Mori et al. (2005) 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 (1991), Kulldorff and Nagarwalla (1995) and Kulldorff (1997) that have been used for the detection of disease clusters in epidemiology.[2] 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 (e.g. 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.[3]

Hence, our strategy is essentially 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.[4] Our probability model then amounts to a multinomial sampling model on this partition. These candidate cluster schemes can in principle be compared by means of standard model-selection criteria, including *Akaike's* (1973) *information criterion*, Schwarz's (1978) *Bayesian information criterion* (*BIC*) and the *Normalized maximum likelihood* of Kontkanen and Myllymäki (2005).

---

2 We shall use 'clusters' and 'agglomerations' interchangeably throughout the analysis to follow. However, one possible distinction between these terms is suggested in Section 5.3.

3 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 tend to identify too many clusters as being significant. For a further discussion of this 'false discovery' problem, see Castro and Singer (2005) together with the references cited therein.

4 An alternative approach would be to characterize spatial distributions of establishments as smooth surfaces, by utilizing the density estimation methods in Billings and Johnson (2012). However, a primary advantage of our present discrete characterization of agglomerations as spatially disjoint clusters is to allow systematic identification of the location, spatial extent and size of each individual agglomeration. This derived data can in turn lead to more detailed analyses of industrial agglomerations (as discussed in Section 6).

To find a best model (cluster scheme) with respect to such 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 article is to develop a reasonable algorithm for searching the space of possible cluster schemes. Our approach can be considered as an elaboration of the basic ideas proposed by Besag and Newell (1991) 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. In particular, we generalize the Besag–Newell concept of clusters by imposing only *convexity* rather than circularity. Although searching over possible convex sets of regions is computationally impractical when the number of regions is large, the procedure reduces to be reasonably simple if the (continuous) location space is approximated by a (discrete) regional network. Accordingly, we develop the notion of *convex solid*, representing the convexity in the regional network.

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. Although the criteria listed above are conceptually different, it turns out that the cluster schemes found are in high agreement across different criteria. Thus, in this article, we will focus on BIC, which turns out to be the most parsimonious criterion in terms of the number of clusters found (Mori and Smith, 2009, Section 3).

The rest of the article 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. Our criterion for model selection in terms of BIC is also developed. In Section 3, we introduce the notion of convex solids and then in Section 4 present a practical procedure for cluster detection which searches for the best cluster scheme consisting of a set of distinct 'convex' clusters. The results of this procedure are then illustrated in Section 5 in terms of the selected pair of Japanese industries discussed above. Here, we sketch a classification scheme for agglomeration patterns in terms of 'global extent' (GE) and 'local density' (LD) that can be employed to quantify the spatial scale of industrial agglomeration and dispersion. A possible refinement and the results of sensitivity analyses for our cluster detection are also presented. Finally, in Section 6, we briefly discuss a number of directions for further research.

2. A probability model of agglomeration patterns

To motivate our approach to cluster detection, we begin by observing that recent theoretical results on equilibrium location patterns in continuous space (e.g. Tabuchi and Thisse, 2011; Ikeda et al., 2012; Hsu, 2012) suggest that there is remarkable commonality among possible equilibrium patterns of agglomeration within each industry. In particular, the number, size and spacing of agglomerations are shown to be well preserved under a variety of stable equilibria. From this perspective, our objective is to identify these common features. To do so, we treat such equilibria as *stationary states* and develop a probabilistic model of location behavior within such stationary

states. In particular, while individual location decisions may be based on the prevailing steady-state distribution, they can nonetheless be treated as statisitically *independent* events, i.e. as random samples from this distribution.[5] This simplification of course precludes any questions about the process of cluster formation, or even the economic rationale for clustering. Rather, our goal here is to provide a simple statistical framework within which the most salient features of these equilibrium cluster patterns can be identified.[6]

To this end, 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 (e.g. 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 are here assumed to be only in terms of establishment counts for each of a set of disjoint *basic regions* (e.g. municipalities), $\Omega_r \subseteq \Omega$, indexed by $R = \{1, \ldots, k_R\}$. These regions are assumed to partition Ω, so that $\cup_{r \in R} \Omega_r = \Omega$. Hence, the only relevant features of the location probability distribution, P, for our purposes are the location probabilities for each basic region:

$$P = [P(r) \equiv P(\Omega_r) : r \in R] . \tag{2.1}$$

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 establishments are more densely located. For the present, such clusters are left unspecified. A more detailed model of individual clusters is developed in Section 3.

If the full extent of cluster C_j in Ω is denoted by $\Omega_{C_j} = \cup_{r \in C_j} \Omega_r, j = 1, \ldots, k_\mathbf{C}$, then the corresponding location probabilities, $p_\mathbf{C}(j) \equiv P_\mathbf{C}(\Omega_{C_j})$, are implicitly taken to define areas of concentration.[7] To complete these probability models, let the set of *residual regions* be denoted by $R_0 (\equiv C_0) = R \setminus \cup_{j=1}^{k_\mathbf{C}} C_j$, and let $\Omega_{R_0} = \Omega \setminus \cup_{j=1}^{k_\mathbf{C}} \Omega_{C_j}$, with corresponding location probability, $p_\mathbf{C}(0) = P_\mathbf{C}(\Omega_{R_0}) = 1 - \sum_{j=1}^{k_\mathbf{C}} p_\mathbf{C}(j)$.

Each *cluster scheme*, $\mathbf{C} = (R_0, C_1, \ldots, C_{k_\mathbf{C}})$, then constitutes a partition of the regional index set, R, and the location probabilities $[p_\mathbf{C}(j) : j = 0, 1, \ldots, k_\mathbf{C}]$ yield a probability distribution on $\mathbf{C}$.[8] 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*.[9] 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, e.g. bodies of water). Such locations

---

5   Here, all firms within each industry are implicitly treated as identical single-establishment firms.
6   A complementary clustering approach has recently been proposed by Kerr and Kominers (2012) which identifies establishment clusters based on maximal interaction distances. This distance approach is particularly useful when relevant interactions can be documented, as in the case of patent citations within research-intensive industries.
7   An implicit assumption here is that the regions $\{\Omega_r : r \in C_j\}$ in each cluster are contiguous. This assumption is not crucial at present, but will play a central role in the construction of clusters below.
8   A formal definition of cluster schemes is given in Definition 4.1.
9   This implicitly assumes that the regions within a given cluster not only have high densities of establishments but also that these densities are similar.

are here designated as the *economic area* of each region.[10] 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

$$a_{C_j} = \sum_{r \in C_j} a_r, \tag{2.2}$$

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, i.e. that

$$P_{\mathbf{C}}(\Omega_r | \Omega_{C_j}) = \frac{a_r}{a_{C_j}}, \ r \in C_j \ , \ j = 0, 1, \ldots, k_{\mathbf{C}} \ . \tag{2.3}$$

But since $\Omega_r \subseteq \Omega_{C_j}$ implies that $P_{\mathbf{C}}(\Omega_r | \Omega_{C_j}) = P_{\mathbf{C}}(\Omega_r)/P_{\mathbf{C}}(\Omega_{C_j})$, if we let $P_{\mathbf{C}}(r) \equiv P_{\mathbf{C}}(\Omega_r)$, it then follows that for all $r \in R$

$$P_{\mathbf{C}}(r) = p_{\mathbf{C}}(j) \frac{a_r}{a_{C_j}} \ , \ \ r \in C_j \ . \tag{2.4}$$

Hence, for each cluster scheme, **C**, Expression (2.4) yields a well-defined *cluster probability model*, $P_{\mathbf{C}} = [P_{\mathbf{C}}(r) : r \in R]$, which is comparable with the unknown true model (2.1). Note moreover that since all area values are known, it follows that for each given cluster scheme, $\mathbf{C} = (R_0, C_1, \ldots, C_{k_{\mathbf{C}}})$, the only unknown parameters are given by the $k_{\mathbf{C}}$-dimensional vector of *cluster probabilities*, $p_{\mathbf{C}} = [p_{\mathbf{C}}(j) : j = 1, \ldots, k_{\mathbf{C}}]$.[11]

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

$$\Delta_R(n) = \left\{ x = (x_r^{(i)} : r \in R, i = 1, \ldots, n) \in \{0, 1\}^{n \times k_R} : \sum_{r \in R} x_r^{(i)} = 1, \ i = 1, \ldots, n \right\}. \tag{2.5}$$

By independence, the probability distribution of X under the unknown true distribution in (2.1) is given for each location pattern, $x \in \Delta_R(n)$, by

$$P(x) = \prod_{i=1}^{n} \prod_{r \in R} P(r)^{x_r^{(i)}} = \prod_{r \in R} P(r)^{n_r}, \tag{2.6}$$

---

10 As pointed out by a referee, 'economic area' is at best a crude approximation to actual usable area for firms. But without more detailed information, we believe that it provides the best approximation currently available.

11 Note that $p_{\mathbf{C}}(0)$ is constructable from $p_{\mathbf{C}}$ as shown above.

where the total number of estabishments locating in region r is denoted by

$$n_r = \sum_{i=1}^{n} x_r^{(i)} \tag{2.7}$$

[see Expression (2.5)]. 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

$$P_{\mathbf{C}}(x|p_{\mathbf{C}}) = \prod_{r \in R} P_{\mathbf{C}}(r)^{n_r} = \prod_{j=0}^{k_{\mathbf{C}}} \prod_{r \in C_j} \left(p_{\mathbf{C}}(j) \frac{a_r}{a_{C_j}} \right)^{n_r}, \tag{2.8}$$

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 $n_j(x) = \sum_{r \in C_j} n_r$, $j = 0, 1, \ldots, k_{\mathbf{C}}$, are sufficient statistics, since by definition

$$P_{\mathbf{C}}(x|p_{\mathbf{C}}) = \prod_{j=0}^{k_{\mathbf{C}}} \left[p_{\mathbf{C}}(j)^{\sum_{r \in C_j} n_r} \prod_{r \in C_j} \left(\frac{a_r}{a_{C_j}} \right)^{n_r} \right] = a_{\mathbf{C}}(x) \prod_{j=0}^{k_{\mathbf{C}}} p_{\mathbf{C}}(j)^{n_j(x)}, \tag{2.9}$$

where the factor, $a_{\mathbf{C}}(x) = \prod_{j=0}^{k_{\mathbf{C}}} \prod_{r \in C_j} (a_r/a_{C_j})^{n_r}$, is independent of parameter vector, $p_{\mathbf{C}}$.

This likelihood function will form the central element in our comparisons among candidate cluster schemes. As mentioned in Section 1, the specific model-selection criterion to be used here is the BIC of Schwarz (1978). As with a number of other criteria, BIC is essentially a 'penalized likelihood' measure. To state this criterion precisely, we first recall from Expression (2.9), that for any given cluster scheme, **C**, the *log likelihood* of parameter vector, $p_{\mathbf{C}}$, given an observed location pattern, x, is of the form

$$L(p_{\mathbf{C}}|x) = \sum_{j=0}^{k_{\mathbf{C}}} n_j(x) \ln p_{\mathbf{C}}(j) + \ln a_{\mathbf{C}}(x) . \tag{2.10}$$

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

$$\hat{p}_{\mathbf{C}}(j) = \frac{n_j(x)}{n} . \tag{2.11}$$

By substituting (2.11) into (2.10), we obtain a corresponding estimate of the *maximum log-likelihood value* for model $P_{\mathbf{C}}$,

$$L_{\mathbf{C}}(x) = L(\hat{p}_{\mathbf{C}}|x) = \sum_{j=0}^{k_{\mathbf{C}}} n_j(x) \ln \left[\frac{n_j(x)}{n} \right] + \ln a_{\mathbf{C}}(x). \tag{2.12}$$

But since likelihood values are non-decreasing in the number of parameters estimated, it follows in particular that values of $L_{\mathbf{C}}(x)$ will almost always increase as more clusters are introduced. 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, BIC penalizes those cluster schemes with larger numbers of clusters, $k_{\mathbf{C}}$, and for any given sample size, n, is of the

form

$$\text{BIC}_{\mathbf{C}}(x) = L_{\mathbf{C}}(x) - \frac{k_{\mathbf{C}}}{2}\ln(n) . \qquad (2.13)$$

In the actual computations involved in cluster detection (to be described in Section 4), it turned out to be convenient to evaluate the cluster scheme, $P_{\mathbf{C}}$, relative to *the uniform probability distribution model* as a benchmark in which individual establishment location follows uniform probability density over economic area. If the BIC value for the uniform probability distribution model is denoted by $\text{BIC}_0 = \sum_{j=0}^{k_{\mathbf{C}}} n_j \ln(a_{\mathbf{C}_j}/a)(\equiv L_0)$, where $a \equiv \sum_{r \in R} a_r$ represents the total area, then we may reformulate this measure in terms of *BIC-differences* from this benchmark model as

$$\Delta_{\mathbf{C}} = \text{BIC}_{\mathbf{C}}(x) - \text{BIC}_0 \equiv T_{\mathbf{C}}(x) - \frac{k_{\mathbf{C}}}{2}\ln(n) , \qquad (2.14)$$

where $T_{\mathbf{C}}(x)$ is the log-likelihood ratio between the cluster and benchmark models:

$$T_{\mathbf{C}}(x) \equiv L_{\mathbf{C}}(x) - L_0 = \sum_{j=0}^{k_{\mathbf{C}}} n_j(x) \ln\left[\frac{n_j(x)/n}{a_{\mathbf{C}_j}/a}\right] . \qquad (2.15)$$

Since the sample size (number of establishments) for each industry is fixed, it plays no direct role in model selection for that industry. But when comparing cluster patterns for different industries, this penalty term will be more severe in industries with larger numbers of establishments. So, all else being equal, BIC tends to yield more parsimonious cluster schemes for larger industries. Moreover, it tends to yield more parsimonious cluster schemes for *all* industries than the other model-selection criteria mentioned above. It is for this reason that we choose to focus on BIC in the present application.

3. A model of clusters as convex solids

Given the set of basic regions, R, it might 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.

$$\mathbf{C}^* = \arg\max_{\mathbf{C}} \Delta_{\mathbf{C}} . \qquad (3.1)$$

But from a practical viewpoint, the number of possible partitions can be enormous for even modest numbers of basic regions.[12] Moreover, without further restrictions, the components of such partitions can be 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, 1991; Kulldorff and Nagarwalla, 1995; Kulldorff, 1997). Here, we propose a more flexible class of clusters that preserve spatial compactness by requiring only that they be 'approximately convex'. We further simplify the identification of convex clusters by representing the location space in terms of a discrete regional network, since from a practical viewpoint, searching over candidate convex clusters is much simpler on networks than in Euclidian space

---

12 For instance, the numbers of counties in the USA and municipalities in Japan are both over 3000.

(especially when the space is large). This network-based (as opposed to Euclidian space-based) approach is particularly useful when economically meaningful distances are adopted (such as travel distance and time), rather than simplistic straight-line distances between regions. Before developing the details of this approach, it is useful to begin with a brief overview.

To define clusters of basic regions, we first require that they be *convex sets* with respect to the underlying network. This means simply that clusters must include all regions on shortest paths between their members (in the same way, planar convex sets include all lines between their points). But unlike straight-line planar paths, shortest paths on discrete networks can sometimes exclude regions that are obviously interior to the desired clusters, thus leaving 'holes' (as shown in Figures 5 and 6).[13] It is thus appropriate to 'fill' these holes by requiring that regional clusters be *convex solid sets* with respect to the underlying network. The formal procedures for developing these convex solid sets will in fact be utilized in the cluster detection algorithm itself, as detailed in Section 4.2.

3.1. 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 $R = \{1, \ldots, k_R\}$. For our present purposes, it is convenient to consider a larger *world region*, W, in which Ω resides, so that $W \backslash \Omega$ denotes the 'rest of the world', as shown schematically in Figure 1. 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, $\bar{R}$, that share boundary points (i.e. the edges of a basic region cell) with $W \backslash \Omega$. This distinguished set of boundary regions (shown in gray) will play an important role in Section 3.3.

Within this basic continuous geographical framework, we next develop a discrete network representation of the regional system that contains all the 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 regional network. Here, it is assumed that data are available on minimal *travel distances*, $t(r, s)$, between each pair of regions, $r, s \in R$, say between their designated administrative centers. 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. 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 2. Here, it is clear that the shortest route (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).

---

13 In Section 3.2, it is shown that such holes persist for even straight-line approximations to travel networks.
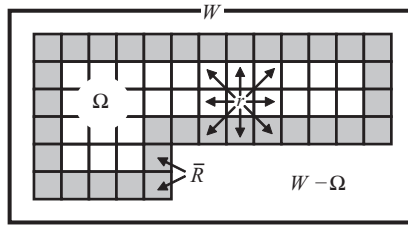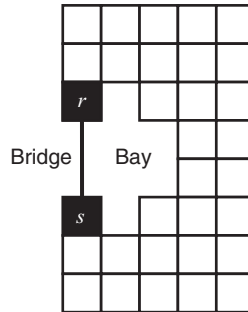
Figure 1. Geographical framework.

Figure 2. Bridge example.

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 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 $L = \cup_{r \in R} \cup_{s \in N(r)} (r, s) \subseteq R^2$, then this defines the relevant set of *links* for our discrete network representation, (R, L), of the regional system. A simple example of such a regional network, (R, L), is shown in Figure 3. Here, R consists of 25 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 40 links (to be discussed later) displayed on the right. Hence, L in this case consists of the 40 distinct regional pairs associated with each of these links, together with the 25 identity pairs (r, r).

Next, we employ travel distances between neighbors to approximate the entire regional network by a shortest path metric on network (R, L). To do so, let each sequence, $\rho = (r_1, r_2, \ldots, r_n)$, of linked neighbors [i.e. with $(r_i, r_{i+1}) \in L$ for $i = 1, \ldots, n-1$] be designated as a *path* in (R, L), and let the set of all paths in (R, L) be denoted by $\mathcal{P} = \{\rho = (r_1, \ldots, r_n) : n > 1, (r_i, r_{i+1}) \in L, i = 1, \ldots, n-1\}$. 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 $\mathcal{P}(r, s) = \{\rho = (r_1, \ldots, r_n) \in \mathcal{P} : r_1 = r, r_n = s\}$, then to ensure that shortest paths between all pairs of regions are meaningful, we henceforth assume that $\mathcal{P}(r, s) \neq \varnothing$ for all $r, s \in R$, i.e. that the given regional network (R, L) is *connected*.[14] In this context, if the *length*, $l(\rho)$, of path, $\rho = (r_1, r_2, \ldots, r_n)$, is now taken to be the sum of travel distances on each of its links, i.e. $l(\rho) = \sum_{i=1}^{n-1} t(r_i, r_{i+1})$, then for any pair of regions,

---

14   See Mori and Smith (2011b, Section 4.2.1) for the treatment of major off-shore islands.
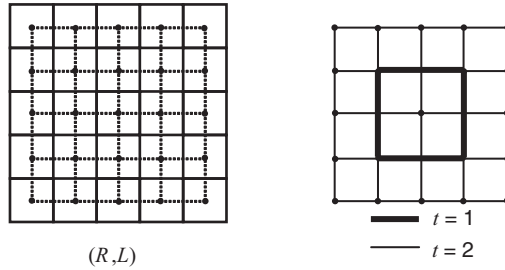
Figure 3. Regional network example.

$r, s \in R$, 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:

$$d(r, s) = \min\{l(\rho) : \rho \in \mathcal{P}(r, s)\} . \tag{3.2}$$

The set of all shortest paths in $\mathcal{P}(r, s)$ is then denoted by $\mathcal{P}_d(r, s) = \{\rho \in \mathcal{P}(r, s) : l(\rho) = d(r, s)\}$. The shortest path distances in (3.2) are 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 $r, s, v \in R$. 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 network may not pass through any intermediate regional centers). Hence, these shortest path distances are only an approximation to shortest route distances.[15] 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 network, so that shortest paths in $\mathcal{P}_d(r, s)$ are more easily identified.

3.2. 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 Section 1, 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 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

---

15 This approximation appears to be good for the municipality network in Japan considered in Section 5. For the ratios of short-path over shortest route distances (d/t) across all 4,491,991 relevant pairs of municipalities, the mean and the 99.5 percentile point are 1.14 and 1.28, respectively.
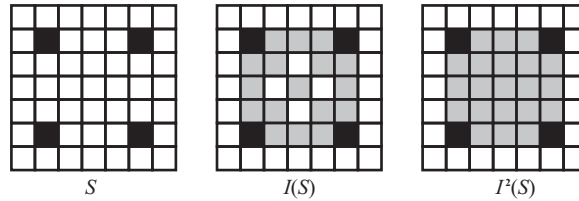
Figure 4. *d*-Convexification of sets.

also in S.[16] More formally, if for any path, $\rho = (r_1, \ldots, r_n) \in \mathcal{P}$, we now denote the *set* of distinct points in ρ by $\langle \rho \rangle = \{r_1, \ldots, r_n\} \subseteq R$, and if the family of all nonempty subsets of R is denoted by $\mathcal{R} = \{S \subseteq R : S \neq \varnothing\}$, then

Definition 3.1 (*d*-Convexity): *(i) A subset of basic regions, $S \subseteq R$, is said to be d-convex iff for all $s, r \in S$, $\rho \in \mathcal{P}_d(r, s) \Rightarrow \langle \rho \rangle \subseteq S$. (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, 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, since 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 included.

This process of adding shortest paths actually yields a well-defined constructive procedure for 'convexifying' a given set, which can be formalized as follows. Let

$$I(r, s) = \bigcup_{\rho \in \mathcal{P}_d(r, s)} \langle \rho \rangle \tag{3.3}$$

denote the (r, s)-*interval* of all points on shortest paths from r to s, and let the mapping, $I : \mathcal{R} \to \mathcal{R}$, defined for all $S \in \mathcal{R}$ by

$$I(S) = \bigcup_{r, s \in S} I(r, s) \tag{3.4}$$

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 all $m > 1$ and $S \in \mathcal{R}$. Since $\{r, s\} \subseteq I(r, s)$ for all $r, s \in R$, it follows from (3.4) that for each set, $S \in \mathcal{R}$,

$$S \subseteq I(S) . \tag{3.5}$$

---

16 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 (1983).

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 $S \in \mathcal{R}$,

$$I^m(S) \subseteq I^{m+1}(S), \quad m \geq 0 \qquad (3.6)$$

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 $(\leq |R \backslash S|)$,[17] such that $I^m(S) = I^{m+1}(S)$.[18] The smallest such integer:

$$m(S) = \min\{m : I^m(S) = I^{m+1}(S)\} \qquad (3.7)$$

is called the *geodesic iteration number of set*, S.[19] With these definitions, it is well known that the unique smallest *d*-convex set containing a given set $S \in \mathcal{R}$ is given by the *d*-convex *hull* (see Proposition A.2 in the appendix for a proof of this assertion),

$$c_d(S) = I^{m(S)}(S) . \qquad (3.8)$$

The mapping, $c_d : \mathcal{R} \to \mathcal{R}$, defined by (3.8) 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

$$\mathcal{R}_d = \{S \in \mathcal{R} : c_d(S) = S\} . \qquad (3.9)$$

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*, $I : \mathcal{R} \to \mathcal{R}$ (as shown in the Corollary to Proposition A.3). Hence, $\mathcal{R}_d$ can also be written as

$$\mathcal{R}_d = \{S \in \mathcal{R} : I(S) = S\} . \qquad (3.10)$$

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, 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 5. Given the regional network, (R, L), in Figure 3, suppose that S consists of the four regions shown in black on the left in Figure 5. These regions are assumed to be connected by major highways as shown by the heavy lines on the right in Figure 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 5), since shortest paths between such regions are always on these highways. But since the

---

17 Throughout this article, we denote *cardinality* of a set A by $|A|$.
18 Since $I^m(S) \neq I^{m+1}(S)$ implies from (3.6) that $|I^{m+1}(S) \backslash I^m(S)| \geq 1$, and since $I^m(S) \subseteq R$ for all m, it follows that this expansion process can involve at most $|R \backslash S|$ steps.
19 In our present application, this iteration number is typically small.
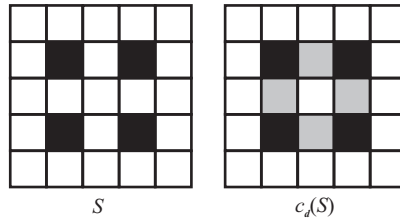
Figure 5. *d*-Convex set with a hole.

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.

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 5 was 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. In particular, since the present notion of *d*-convexity is intended to approximate convexity in the original location space, it is appropriate to fill these holes.

Finally, it is of interest to note that even with simpler approximations to travel distances, such holes can still exist. For example, if shortest paths between adjacent regions are approximated by straight-line paths between their geometric centroids, then this same convexification procedure can still yield holes. This is illustrated by the simple four-region example in Figure 6, where the three exterior regions are seen to form a convex set containing all shortest paths between them. Hence, the central region is not part of this convex set and constitutes an obvious hole.

3.3. 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 1 that relative to the rest of the world, there is a distinguished collection of *boundary regions*, $\bar{R}$, that are essentially 'external' to all subsets of R. If for any subset, $S \subseteq R$, and boundary region, $\bar{r} \in \bar{R}$, it is true that $\bar{r} \notin S$, then it is reasonable to assert that $\bar{r}$ is *outside* of S.[20] This set of boundary regions, $\bar{R}$, thus defines a natural reference set for distinguishing regions in complement, $R \backslash S$, of S that are 'inside' or 'outside' of S. In particular, we now say that a complementary region, $r \in R \backslash S$, is *inside* S if and only if every path joining r to a boundary region in $\bar{R}$ must pass through at least one region of S. For example, given the set, S, of black squares in Figure 7, the complementary region r is seen to be inside of S since every path to the boundary, $\bar{R}$, must intersect S. Similarly, the complementary region s is not inside S, since there is a path from s to $\bar{R}$ that does not intersect S. To formalize this concept, we now let the set of all paths from

---

20   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.
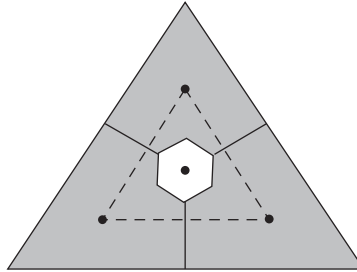
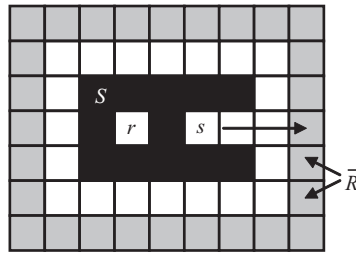**Figure 6.** Non-solid *d*-convex set.

Figure 7. Inside versus outside.

any region, $r \in R$, to $\bar{R}$ be denoted by $\mathcal{P}(r, \bar{R}) = \cup_{\bar{r} \in \bar{R}} \mathcal{P}(r, \bar{r})$. Then, for any nonempty set, $S \in \mathcal{R}$, the set of all complementary regions *inside* S is given by

$$S_0 = \{ r \in R \backslash S : \rho \in \mathcal{P}(r, \bar{R}) \Rightarrow \langle \rho \rangle \cap S \neq \varnothing \} \tag{3.11}$$

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

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 3.2 (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,*

$$\sigma(S) = S \cup S_0 \tag{3.12}$$

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, $\sigma : \mathcal{R} \to \mathcal{R}$, induced by (3.12) 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 (see Lemma A.2 in the appendix).

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 them as follows:

Definition 3.3 (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,

$$\sigma c_d(S) = \sigma[c_d(S)] \tag{3.13}$$

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 3.1–3.3 that

$$\mathcal{R}_{\sigma d} = \mathcal{R}_\sigma \cap \mathcal{R}_d. \tag{3.14}$$

3.4. Convex solidification of sets

As with (3.11) and (3.12), Expression (3.13) 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 3 and 5. If the exterior squares are taken to define the relevant boundary set, $\bar{R}$, in Figure 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 $\bar{R}$ that do not intersect S.[21] 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 3.3 is justified, i.e. that:

Property 3.4 (*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 to construct a new cluster containing $C \cup S$, one need only *d*-convexify this set by the algorithm

$$C \cup S \to I(C \cup S) \to I^2(C \cup S) \cdots \to c_d(C \cup S) \tag{3.15}$$

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

$$\sigma c_d(C \cup S) = c_d(C \cup S) \cup [c_d(C \cup S)]_0 . \tag{3.16}$$

This algorithm has already been illustrated by the simple case in Figure 4, where no solidification was required. A somewhat more detailed illustration is given in Figures 8 and 9. Figure 8 exhibits a subsystem of 19 (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, 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

---

21 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'.
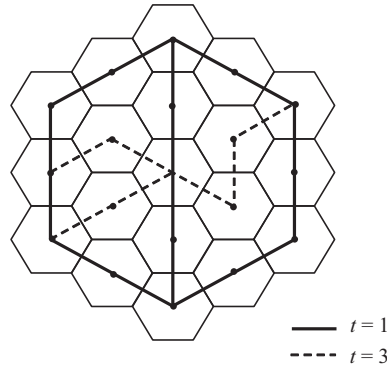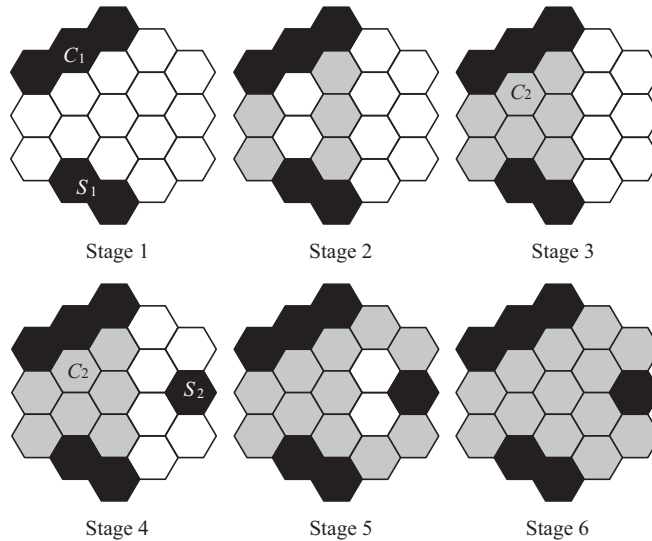
Figure 8. Regional subsystem.

Figure 9. Formation of composite clusters.

distance (or time) values of $t = 1$. Other regions are connected by secondary roads denoted by dashed network links with higher values of $t = 3$.

A possible sequence of steps in the formation of a composite cluster in this subsystem is depicted in Figure 9. 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 8, 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 as 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 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 cluster, C_2. Again, Stage 5 shows that all regions on the freeway corridors from S_2 to C_2 should be added in 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, $C_3 = \sigma c_d(C_2 \cup S_2)$.

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 (3.8), 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 3.5 (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 covexified version of this set.

Recall that our network representation of space is mainly for the computational efficiency, and the *d*-convexity aims for approximating convexity in the original location space. Property 3.5 indicates that *d*-convex solid in the network corresponds to the convex hull in Euclidian space. Thus, as desired, it is conceptually consistent to adopt *d*-convex solid as convex approximation of the spatial coverage of a given cluster.

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 (3.15) and (3.16) always yields a fixed point of the composite mapping, $\sigma c_d : \mathcal{R} \rightarrow \mathcal{R}$:

Property 3.6 (*d*-Convex solid fixed points): *A set, $S \in \mathcal{R}$, is a d-convex solid iff $\sigma c_d(S) = S$.*

Hence, the family, $\mathcal{R}_{\sigma d}$, of all *d*-convex solids in (3.14) can equivalently be written as $\mathcal{R}_{\sigma d} = \{S \in \mathcal{R} : \sigma c_d(S) = S\}$. In this form, each new cluster is seen to be a natural 'stopping point' of the combined *d*-convexification and solidification procedure above.

4. 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 4.1 (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 $i = 1, \ldots, k_\mathbf{C}$ and (ii) (disjointness) $C_i \cap C_j = \varnothing$ for all i,j with $1 \leq i < j$. 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 (or clusters) 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 3. Before completing the description of this procedure (in Section 4.2), we specify the details of option (iii) above in the next section.

4.1. 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 10. Given the subsystem of a regional network shown in Figure 8, suppose that the current cluster scheme includes the clusters C_1 and C_2 shown in Stage 1 of Figure 10. 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 10.

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 $S \in \mathcal{R}$ by

$$U_\mathbf{C}(S) = \sigma c_d(S) \cup \{C_i \in \mathbf{C} : C_i \cap \sigma c_d(S) \neq \varnothing\}, \tag{4.1}$$

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 that 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 3.2 by letting the iterates of $U_\mathbf{C}$ be defined for each $S \in R$ by $U_\mathbf{C}^0(S) = S$, $U_\mathbf{C}^1(S) = U_\mathbf{C}(S)$ and $U_\mathbf{C}^m(S) = U_\mathbf{C}[U_\mathbf{C}^{m-1}(S)]$ for all $m > 1$. Since it is clear by definition that $U_\mathbf{C}^m(S) \subseteq U_\mathbf{C}^{m+1}(S)$ for all $m \geq 0$, this yields a monotone nondecreasing sequence of sets in R. Hence, by the same arguments leading to (3.7), it again follows that there must be an integer, m ($\leq |R \backslash S|$), such that $U_\mathbf{C}^m(S) = U_\mathbf{C}^{m+1}(S)$. As a parallel to (3.7), we may thus designate the smallest integer, $m(S|\mathbf{C}) = \min\{m : U_\mathbf{C}^m(S) = U_\mathbf{C}^{m+1}(S)\}$, 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 $U_\mathbf{C}$,

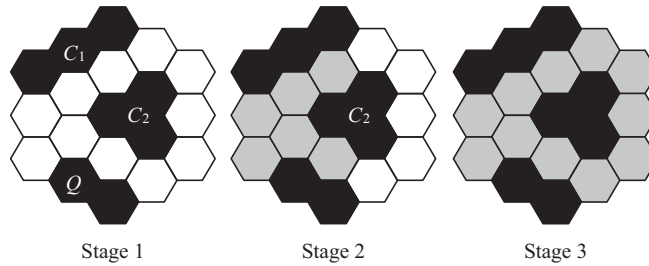$$u_\mathbf{C}(S) = U_\mathbf{C}^{m(S|\mathbf{C})}(S) \tag{4.2}$$

Figure 10. Formation of composite clusters.

as the **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 (4.1) that for all clusters $C_i \in \mathbf{C}$ with $C_i \cap u_\mathbf{C}(S) \neq \varnothing$ we must have $C_i \subseteq u_\mathbf{C}(S)$. Thus, $u_\mathbf{C}(S)$ is always disjoint from any clusters, $C_i \in \mathbf{C}$, that have not already been absorbed into $u_\mathbf{C}(S)$. Moreover, this in turn implies from (4.1) 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.

4.2. Cluster-detection procedure

In terms of Definition 4.1, 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,

$$\mathbf{C}^* = \arg \max_{\mathbf{C} \in \mathcal{C}(R,L)} \Delta_\mathbf{C} . \tag{4.3}$$

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 (4.3) 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 $\mathbf{C}_t = (R_{t,0}, C_{t,1}, \ldots, C_{t, k_{\mathbf{C}_t}})$ denote the current cluster scheme in $\mathcal{C}(R, L)$. The procedure then starts at stage $t = 0$ with the *null cluster scheme*, $\mathbf{C}_0 = \{R_{0,0}\} = \{R\}$, containing no clusters. By Expressions (2.14) and (2.15), it follows that the corresponding initial value of the objective function in (4.3) must be 0. Given data, $[\mathbf{C}_t, \Delta_{\mathbf{C}_t}]$, 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 $\Delta_{\mathbf{C}_{t+1}}$. 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.

4.2.1. 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 $R_0(\mathbf{C}_t) = R_{t,0}$. For each $r \in R_0(\mathbf{C}_t)$, the corresponding expanded cluster scheme is then given by $\mathbf{C}_t^0(r) = (R_{t,0}^0(r), C_{t,1}^0(r), C_{t,2}^0, \ldots, C_{t, k_{\mathbf{C}_t^0(r)}}^0)$, where $R_{t,0}^0(r) = R_{t,0} \backslash \{r\}$, $k_{\mathbf{C}_t^0(r)} = k_{\mathbf{C}_t} + 1$,

$C_{t,1}^0(r) = \{r\}$ and $C_{t,i}^0 = C_{t,i-1}$ for $i = 2, \ldots, k_{\mathbf{C}_t^0(r)}$. 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 $r_0^* = \arg\max_{r \in R_0(\mathbf{C}_t)} \Delta_{\mathbf{C}_t^0(r)}$. For purposes of comparison with other possible modifications of $\mathbf{C}_t$, we now set

$$\mathbf{C}_t^0 \equiv \mathbf{C}_t^0(r_0^*) . \tag{4.4}$$

4.2.2. 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. While the basic mechanics of this expansion procedure were developed in Section 4.1, 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 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 $R \backslash C_{t,j}$ 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.[22] 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

$$d(r, Q) = \min\{d(r, s) : s \in Q\} \tag{4.5}$$

for $r \in R$ and $Q \in \mathcal{R}$. Since *d* is a metric on *R*, it is well known that for each set, $Q \in \mathcal{R}$, (4.5) yields a well-defined distance function that preserves the usual continuity properties of *d* on *R* (e.g. Berge, 1963, Ch. 5). 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 $\delta(Q) = \{r \in R : d(r, Q) < \delta\}$. Hence, the

---

22 The inclusion of large undeveloped regions (e.g. mountains and inland sea) 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.

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.[23]

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, it follows from Expression (4.2) that the corresponding $\mathbf{C}_t$-compatible expansion of $C_{t,j} \cup \{r\}$ is given by

$$C_{t,1}^j(r) = u_{\mathbf{C}_t}(C_{t,j} \cup \{r\}) . \qquad (4.6)$$

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

$$\mathbf{C}_t^j(r) = \left(R_{t,0}^j(r), C_{t,1}^j(r), C_{t,2}^j(r), \ldots, C_{t,k_{\mathbf{C}_t^j(r)}}^j(r) \right), \qquad (4.7)$$

where by Expression (4.1), the set of all other clusters in $\mathbf{C}_t^j(r)$ is given by

$$\left\{ C_{t,2}^j(r), \ldots, C_{t,k_{\mathbf{C}_t^j(r)}}^j(r) \right\} = \{ C_{t,i} \in \mathbf{C}_t : C_{t,i} \cap C_{t,1}^j(r) = \varnothing \} \qquad (4.8)$$

and where the corresponding residual region has the form:

$$R_{t,0}^j(r) = R \setminus \bigcup_{i=1}^{k_{\mathbf{C}_t^j(r)}} C_{t,i}^j(r) . \qquad (4.9)$$

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 $r_j^* = \arg\max_{r \in R_{t,0} \cap \delta(C_{t,j})} \Delta_{\mathbf{C}_t^j(r)}$, 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, $I_j(\mathbf{C}_t) = \{ i \neq j : C_{t,i} \cap \delta(C_{t,j}) \neq \varnothing \}$ [not to be confused with interval sets $I(\cdot)$ in Section 3.2], and for each $i \in I_j(\mathbf{C}_j)$ replace (4.6) with the $\mathbf{C}_t$-compatible expansion $C_{t,1}^j(i) = u_{\mathbf{C}_t}(C_{t,j} \cup C_{t,i})$, then as a parallel to (4.7)–(4.9), the cluster scheme, $\mathbf{C}_t^j(i)$, resulting from this expansion now has the form

$$\mathbf{C}_t^j(i) = \left(R_{t,0}^j(i), C_{t,1}^j(i), C_{t,2}^j(i), \ldots, C_{t,k_{\mathbf{C}_t^j(i)}}^j(i) \right) \qquad (4.10)$$

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

$$\left\{ C_{t,2}^j(i), \ldots, C_{t,k_{\mathbf{C}_t^j(i)}}^j(i) \right\} = \left\{ C_{t,i} \in \mathbf{C}_t : C_{t,i} \cap C_{t,1}^j(i) = \varnothing \right\} \qquad (4.11)$$

---

23 In our application in Section 5, the value used is $\delta = 36.0$ km, which was chosen so that any single expansion of a cluster cannot include a large section without economic area (e.g. inland sea and lakes). This δ value covers about 90% of the shortest path distances between neighboring basic regions (municipalities) in our application. 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.

and with corresponding residual region:

$$R_{t,0}^j(i) = R \setminus \bigcup_{k=1}^{k_{C_t^j(i)}} C_{t,k}^j(i) . \tag{4.12}$$

If i_j^* denotes the cluster in $I_j(\mathbf{C}_t)$ that yields the highest value of the objective function for which $i_j^* = \arg\max_{i \in I_j(\mathbf{C}_t)} \Delta_{\mathbf{C}_t^j(i)}$, 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

$$\mathbf{C}_t^j \equiv \arg\max_{\mathbf{C} \in \{\mathbf{C}_t^j(r_j^*), \mathbf{C}_t^j(i_j^*)\}} \Delta_{\mathbf{C}}, j = 1, \ldots, k_{\mathbf{C}_t}. \tag{4.13}$$

4.2.3. Revision of the cluster scheme

Finally, given these candidate modifications, $\mathbf{C}_t^0, \mathbf{C}_t^1, \ldots, \mathbf{C}_t^{k_{C_t}}$, of $\mathbf{C}_t$ in $\mathcal{C}(R, L)$ [as defined by (4.4) together with (4.13)], let $\mathbf{C}_t^*$ be the best candidate, as defined by

$$\mathbf{C}_t^* = \arg\max_{\mathbf{C} \in \{\mathbf{C}_t^j; j=0, 1, \ldots, k_{\mathbf{C}_t}\}} \Delta_{\mathbf{C}} . \tag{4.14}$$

There are then two possibilities left to consider: If $\Delta_{\mathbf{C}_t^*} > \Delta_{\mathbf{C}_t}$, then set $[\mathbf{C}_{t+1}, \Delta_{\mathbf{C}_{t+1}}] = [\mathbf{C}_t^*, \Delta_{\mathbf{C}_t^*}]$ and proceed to stage $t + 1$. On the other hand, if $\Delta_{\mathbf{C}_t^*} \leq \Delta_{\mathbf{C}_t}$, then no (local) improvement can be made, and the cluster-detection procedure terminates with the (locally) *optimal cluster scheme*, $\mathbf{C}^* = \mathbf{C}_t$.

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 produced results similar to the present search procedure, but proved to be far more computationally demanding.

4.3. A test of spurious clustering

Although the cluster-detection procedure developed above will always find a (locally) best cluster scheme, $\mathbf{C}^*$, with respect to BIC used, there is still the statistical question of whether such clustering could simply have occurred by chance. Hence, one can ask how the optimal criterion value, $\Delta_{\mathbf{C}^*}$, 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*, 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

$$p_r = \frac{a_r}{a} . \tag{4.15}$$

While the sampling distribution of $\Delta_{\mathbf{C}}$ 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 (4.15) 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 $\Delta_{\mathbf{C}}$, say $\Delta_1, \ldots, \Delta_{1000}$. If the value for the actual cluster scheme, $\Delta_0 = \Delta_{\mathbf{C}^*}$, is say bigger than all but five of these in the ordering of values, $\{\Delta_1, \ldots, \Delta_{1000}\}$, then the chance, p, of getting a value as large as this (under the hypothesis that Δ_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 Δ_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, $\mathbf{C}^*$, is not sufficiently significant to warrant further investigation. This procedure was used in the following illustrative application [as well as in the more extensive applications in Mori and Smith (2011a, 2011b, 2012) and Hsu et al. (2011)].

4.4. Essential clusters

The identified clusters, $C \in \mathbf{C}^*$, vary in terms of their contribution to the value of $\Delta_{\mathbf{C}^*}$. While the clusters with larger contributions are often insensitive to small perturbations of the original regional distribution of establishments, those with smaller contributions may be sensitive. Thus, to obtain more robust results, it may be useful to focus on those *essential clusters* which account for a large shares of $\Delta_{\mathbf{C}^*}$.

To formalize this idea, we start by assuming that an optimal cluster scheme, $\mathbf{C} = \mathbf{C}^*$, has been found for the industry. To identify the essential clusters in $\mathbf{C}$, we proceed recursively by successively adding those clusters in $\mathbf{C}$ with maximum incremental contributions to $\Delta_{\mathbf{C}}$.[24] This recursion starts with the 'empty' cluster scheme represented by $\mathbf{C}_0 \equiv \{R_{0,0}\}$, where $R_{0,0}$ denotes the full set of regions, R. If the set of (non-residual) *clusters* in $\mathbf{C}$ is denoted by $\mathbf{C}^+ \equiv \mathbf{C} \backslash \{R_0\}$, then we next consider each possible 'one-cluster' scheme created by choosing a cluster, $C \in \mathbf{C}^+$, and forming $\mathbf{C}_0(C) = \{R_{0,0}(C), C\}$, with $R_{0,0}(C) = R_{0,0} \backslash C$. The 'most significant' of these, denoted by $\mathbf{C}_1 = \{R_{1,0}(C), C_{1,1}\}$, is then taken to be the cluster scheme with the *maximum BIC value* (defined below). If this is called *stage* $t = 1$, and if the *essential cluster scheme* found at each stage $t \geq 1$ is denoted by $\mathbf{C}_t \equiv \{R_{t,0}, C_{t,1}, \ldots, C_{t,t}\}$, then the recursive construction of these schemes can be defined more precisely as follows.

For each $t \geq 1$, let $\mathbf{C}_{t-1}^+$ denote the (non-residual) clusters in $\mathbf{C}_{t-1}$ (so that for $t = 1$ we have $\mathbf{C}_{t-1}^+ = \mathbf{C}_0^+ = \varnothing$), and for each cluster not yet included in $\mathbf{C}_{t-1}$, i.e. each $C \in \mathbf{C}^+ \backslash \mathbf{C}_{t-1}^+$, let $\mathbf{C}_{t-1}(C)$ be defined by, $\mathbf{C}_{t-1}(C) = (R_{t-1,0}(C), C_{t-1,1}, \ldots, C_{t-1,t-1}, C)$, where $R_{t-1,0}(C) = R_{t-1,0} \backslash C$. Then, the *additional essential cluster*, $C_t (\equiv C_{t,t})$ $(\in \mathbf{C}^+ \backslash \mathbf{C}_{t-1}^+)$, at stage $t \geq 1$ is defined by

$$C_t \equiv \arg \max_{C \in \mathbf{C}^+ \backslash \mathbf{C}_{t-1}^+} T_{\mathbf{C}_{t-1}(C)} \, , \tag{4.16}$$

---

24 The procedure for identifying essential clusters in $\mathbf{C}$ is different from the one used to indentify $\mathbf{C}$ in Section 4.2. Here, candidate clusters considered are only those in $\mathbf{C}$ itself.

where $T_{\mathbf{C}_{t-1}(C)}$ is the *estimated maximum log-likelihood ratio* for model $p_{\mathbf{C}_{t-1}(C)}$ given [in a manner paralleling expression (2.15)] by

$$T_{\mathbf{C}_{t-1}(C)} = \sum_{C' \in \mathbf{C}_{t-1}(C)} n_{C'} \ln\left(\frac{n_{C'}/n}{a_{C'}/a}\right), \qquad (4.17)$$

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

$$\Delta_{\mathbf{C}_t} = T_{\mathbf{C}_t} - \frac{t}{2}\ln(n), \qquad (4.18)$$

with

$$T_{\mathbf{C}_t} = \sum_{C \in \mathbf{C}_t} n_C \ln\left(\frac{n_C/n}{a_C/a}\right). \qquad (4.19)$$

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

$$\Delta_t(\mathbf{C}) \equiv \Delta_{\mathbf{C}_t} - \Delta_{\mathbf{C}_{t-1}}. \qquad (4.20)$$

To identify the relevant set of the essential clusters in $\mathbf{C}$, one simple criterion would be to require that each has a BIC contribution at least some specified fraction, μ, of $\Delta_1(\mathbf{C})$. 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 $\Delta_{t^e+1} < \mu\Delta_1$. Refer to Mori and Smith (2011b, Section 3) for an application of these essential clusters.

5. An illustrative application

In this section, we illustrate the above procedure in terms of the two Japanese industries discussed in Section 1, which for convenience we refer to here as simply 'plastics' and 'soft drinks', respectively. These two industries are part of the larger study in Mori and Smith (2011b) that applies the present methodology to 163 manufacturing industries in Japan. As discussed in Section 4.2 of that article, the test of spurious clustering above identified nine industries with spurious clustering, so that only 154 industries were used in the final analysis. The appropriate notion of a 'basic region', r, for purposes of this study was taken to be the municipality category equivalent to a city-ward-town-village in Japan. The relevant set R was then taken to be the 3207 municipalities geographically connected to the major islands of Japan, as shown in Figure 11.[25]

---

25 The establishment counts across these industries are taken from the *Establishment and Enterprise Census* of Japan in 2001. Economic area of each municipality is obtained by subtracting forests, lakes, marshes and undeveloped area from the total area of the municipality. The data are available from the *Toukei de Miru Shi-Ku-Cho-Son no Sugata* in 2002 and 2003 (in Japanese) by the Statistical Information Institute for Consulting and Analysis of Japan.
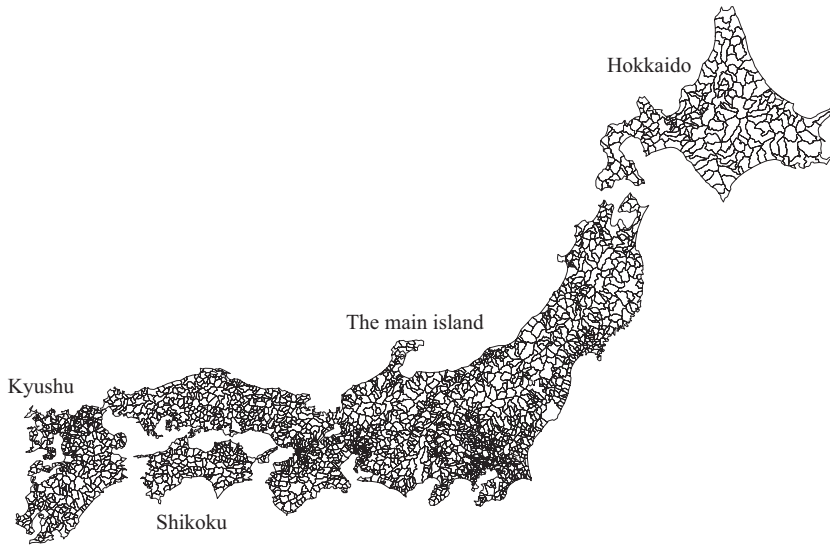
Figure 11. Basic regions (shi-ku-cho-son) of Japan.

5.1. Comparison with a scalar measure of agglomeration

The choice of these two industries is motivated by their similarity in terms of overall degree of agglomeration. This can be illustrated in terms of the *D-index* developed in Mori et al. (2005), which for a given industry i is defined as the Kullback–Leibler (1951) divergence of its establishment location probability distribution, $P_i \equiv [P_i(r) : r \in R]$, [as in expression (2.1)] 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 $P_0(r) = a_r / \sum_{j \in R} a_j$ [as in expression (4.15)]. 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 (2.7)], a corresponding estimate of this *D*-index is given by

$$D(\widehat{P}_i | P_0) = \sum_{r \in R} \widehat{P}_i(r) \ln \left[\frac{\widehat{P}_i(r)}{P_0(r)} \right]. \tag{5.1}$$

The intuition behind this particular index is that it provides a natural measure of distance between probability distributions. So by taking uniformity to represent the complete absence of clustering, it is reasonable to assume that those distributions 'more distant' from the uniform distribution should involve more clustering. Note also that since both D and $\Delta_{\mathbf{C}}$ are based on similar log-likelihood measures of 'distance from uniformity', our cluster detection procedure is closer in spirit to this scalar measure than other possible choices such as the index by Ellison and Glaeser (1997).[26] Hence D provides a natural candidate for comparing the advantages of this approach over scalar measures in general. The histogram of divergence values, D, for the 154 industries in Japan is shown in Figure 12 and is seen to range from $D = 0.47$ up to 5.98. With respect

---

26 In fact, Ellison–Glaeser index is highly correlated with D (refer to Mori et al., 2005, Section D). So, the arguments in this section would remain essentially the same.
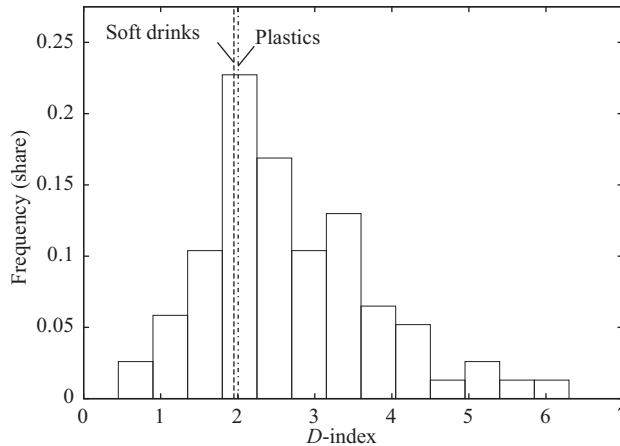
Figure 12. Frequency distribution of *D*-values of Japanese manufacturing industries.

to this overall range, the *D* values, 1.95 and 2.05, for soft drinks and plastics, respectively, are seen to be virtually identical.

But in spite of this overall similarity, the agglomeration patterns obtained for these two industries are substantially different, as seen in Figures 13 and 14.

Panel (a) of each figure displays the establishment densities for the corresponding industry, where those basic regions with higher densities are shown as darker. In Panel (b), the individual clusters in the derived cluster 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 convex solidification).

Before examining these patterns in detail, it is of interest to consider the results of the cluster-detection procedure itself. By comparing the establishment densities and cluster schemes in Panels (a) and (b) of each figure, respectively, it is clear that these cluster schemes closely reflect the underlying densities from which they were obtained. Notice also that individual clusters are by no means 'circular' in shape. Rather each consists of an easily recognizable set of contiguous basic regions (municipalities) in *R* that approximates the area of higher establishment density in Panel (a) of the figure. Notice also that certain clusters in each pattern are themselves contiguous. We shall return to this point below.

To compare these two agglomeration patterns in more detail, we begin by observing that while the plastics industry is more than twice as large as soft drinks in terms of the number of establishments (1555 versus 777), its agglomeration pattern contains only 43 clusters versus 55 clusters for soft drinks. This illustrates the relative parsimoniousness of our cluster-detection procedure with respect to larger industries, as mentioned following the definition of BIC in expression (2.13). Notice also that clustering is indeed much stronger in the plastics industry than in soft drinks. This can be seen in several ways. First, the share of plastics establishments in clusters is much larger than for soft drinks (93.9% versus 64.6%). Second, the average size of these clusters is greater not only in terms of establishments per cluster (as implied by the statistics above), they are also more than three time larger in terms of average areal extent.
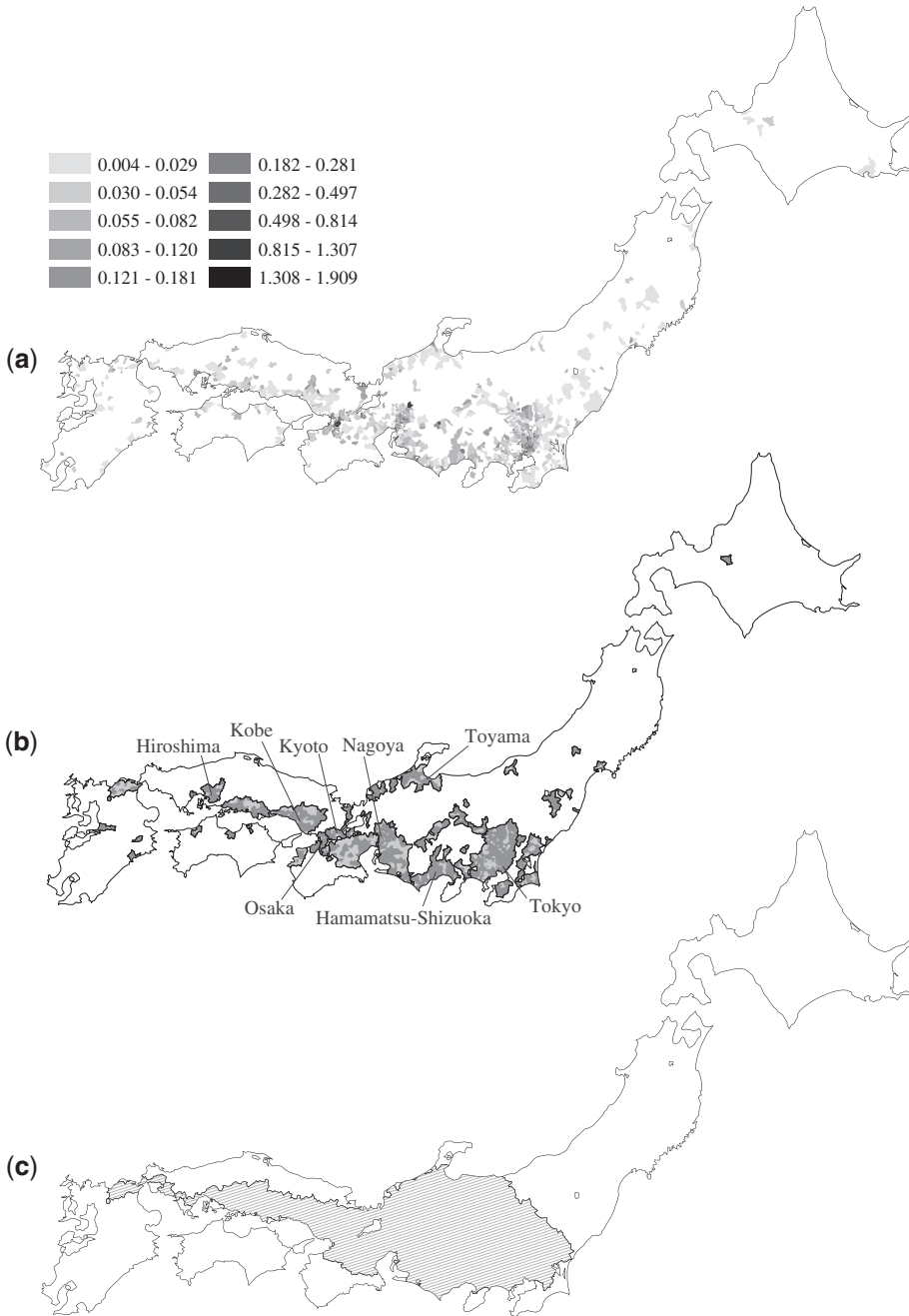
Figure 13. Spatial distributions of establishments and clusters (plastics industry). (a) Density of establishments (per km^2), (b) clusters and (c) essential containment.
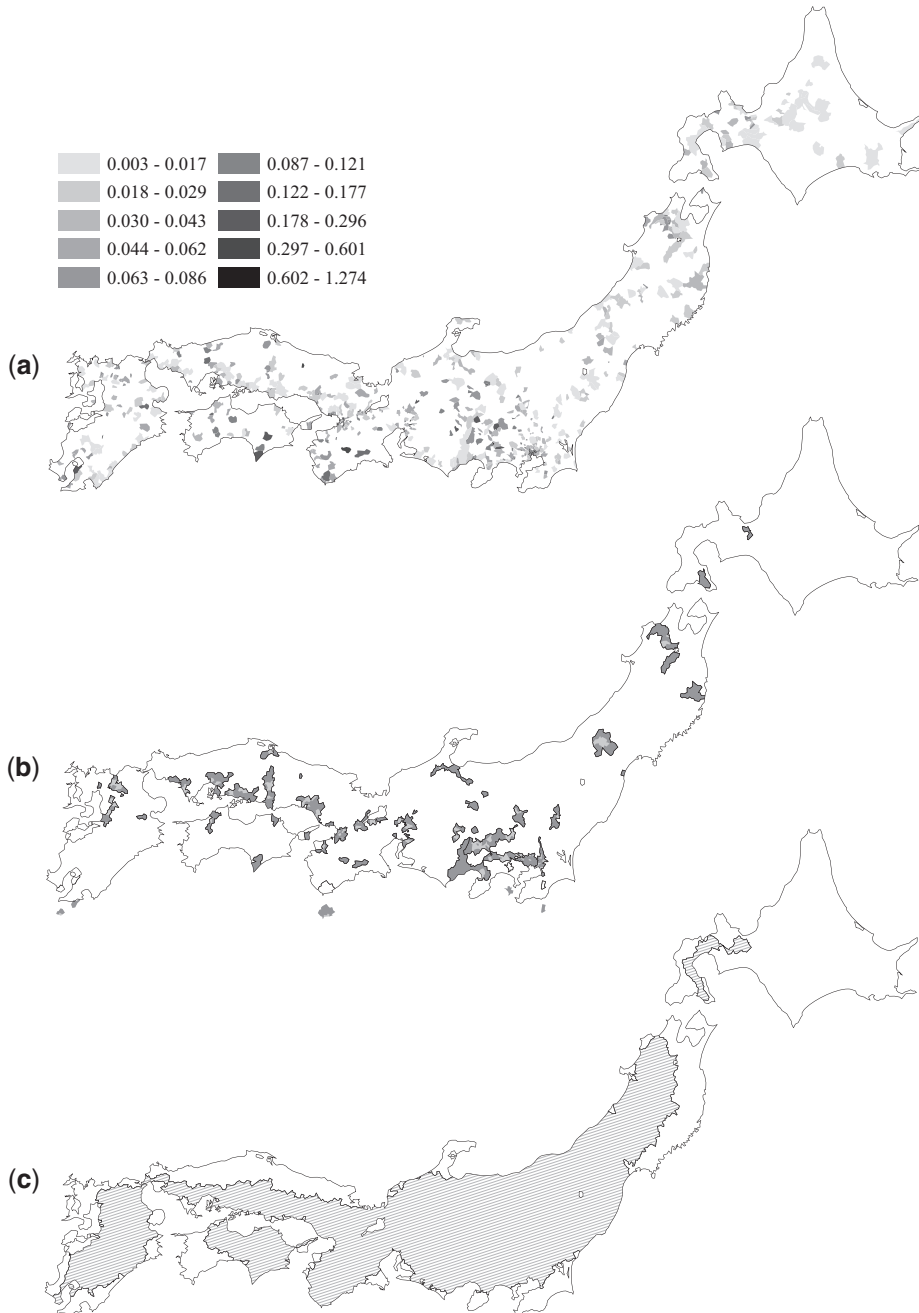
0.003 - 0.017		0.087 - 0.121	
0.018 - 0.029		0.122 - 0.177	
0.030 - 0.043		0.178 - 0.296	
0.044 - 0.062		0.297 - 0.601	
0.063 - 0.086		0.602 - 1.274	

(a)

(b)

(c)

Figure 14. Spatial distributions of establishments and clusters (soft drinks industry). (a) Density of establishments (per km^2), (b) clusters and (c) essential containment.

5.2. Global extent versus local density of agglomerations

Aside from these general comparisons in terms of summary statistics, the level of spatial detail in each of these agglomeration patterns allows a much broader range of comparative measures. While such measures are developed in more detail in Mori and Smith (2011b), their essential elements are well illustrated in terms of the present pair of industries. As mentioned in Section 1, the plastics industry is primarily concentrated along the industrial belt of Japan as in Figure 13(b). More generally, industries often tend to concentrate within specific subregions of the nation, i.e. are themselves 'spatially contained'. To make this precise in terms of our present model of cluster schemes, we adopt a two-stage approach. First, we identify the essential clusters with $\mu = 0.05$ (as defined in Section 4.4) in the optimal cluster scheme, $\mathbf{C}^*$, for a given industry. We then define the *essential containment* (*e-containment*) for that industry to be the convex solidification of these essential clusters, in other words, the smallest convex solid[27] containing all these essential clusters for the industry. The e-containment for the plastics industry is indicated by the hatched area in Figure 13(c) which clearly distinguishes the 'industrial belt' portion of this industry. In contrast, the e-containment for soft drinks shown in Figure 14(c) appears to be much larger and reflects the wide scattering of essential clusters for this industry.

While these visual summaries of 'containment' can be very informative, it is often more useful to quantify such relations for purposes of analysis. One possibility here is to define the *global extent* (GE) of an industry to be the fraction of area in its e-containment relative to the nation as a whole.[28] In the present case, the GE values for plastics and soft drinks are 0.298 and 0.589, respectively. So in terms of this measure, it is clear that the clusters of the plastics industry are much more localized than those of soft drinks.

Next observe that while the GE of the plastics industry is much smaller than that of soft drinks, the average size of its essential clusters is actually much larger. As is clear from Figures 13 and 14, these clusters are thus more densely packed inside the e-containment of the plastics industry. To capture this additional dimension of agglomeration patterns, we now designate the fraction of e-containment area represented by these essential clusters as the *local density* (LD) of the industry. Since the LD values for plastics and soft drinks are given, respectively, by 0.465 and 0.133, it is also clear that the agglomeration pattern for plastics is much more locally dense than that of soft drinks.

5.3. Refinements of cluster schemes

Recall that in terms of our basic probability model of cluster schemes, $\mathbf{C}$, individual clusters, C_j, are implicitly assumed to constitute sets of basic regions with similar (and unusually high) establishment density. But the relations between these clusters is left unspecified. In this regard, it was observed above that the opimal cluster schemes, $\mathbf{C}^*$, for both plastics and soft drinks contain clusters that are mutually *contiguous*. Here, it is natural to ask why such clusters were not 'joined' at some stage during the

---

27 Recall Property 3.3 of convex solidification.
28 Here, we use the full geographic areas of basic regions rather than economic area, to give a better representation of 'extent'. See further discussions in Mori and Smith (2011b, Section 3.2).

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 (2.3)], so that maximum-likelihood estimates for cluster probabilities, $p_C(j)$, are simply proportional to the number of establishments, n_j, in that cluster. 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. As one illustration, there is a contiguous chain of clusters for the plastics industry extending from Tokyo toward west as far as Osaka [Figure 13(b)]. Here, the establishment densities in these contiguous areas are sufficiently different so that by treating each as a different cluster, one obtains a better overall fit in terms of BIC—even though the resulting scheme is penalized for this larger number of clusters.

It is often the case, however, that there are not only very different establishment densities among contiguous clusters, but also strong 'central' clusters: Tokyo, Nagoya and Osaka in this case. More generally, this suggests that there is often more spatial structure in cluster schemes than is captured by a simple listing of their clusters. In particular, this example suggests that a grouping of contiguous clusters around each central cluster (i.e. with the highest establishment density) might best be treated as single *agglomerations* for an industry.

To formalize these ideas, we begin with a given cluster scheme, **C**, that has been identified for an industry. For each individual cluster, $C \in \mathbf{C}$, let $N_{\mathbf{C}}(C)(\subseteq \mathbf{C})$ be the set of *contiguous neighbors of C* in **C** (including C itself), so that by definition there exists for each $C' \in N_{\mathbf{C}}(C)$ a basic region, $r \in C$, which is adjacent to cluster C', i.e. with $N(r) \cap C' \neq \varnothing$.[29] For each $C \in \mathbf{C}$, the maximal-density cluster in its immediate neighborhood, $N_{\mathbf{C}}(C)$, can then be identified by a *hill climbing function*, $H_{\mathbf{C}}(C) \equiv \arg\max_{C' \in N_{\mathbf{C}}(C)} n_{C'}/a_{C'}$.[30] In particular, if $H_{\mathbf{C}}(C) = C$, then cluster C is a *local peak* of establishment density with respect to its contiguous neighbors, $N_{\mathbf{C}}(C)$, and hence can be considered as a *central cluster* in its vicinity. More generally, we can generate a unique central cluster for each $C \in \mathbf{C}$ by recursive applications of this hill climbing function. To do so, we begin by setting $H_{\mathbf{C}}^0(C) \equiv H_{\mathbf{C}}(C)$ and constructing m th-iterates of $H_{\mathbf{C}}$ by $H_{\mathbf{C}}^m(C) = H_{\mathbf{C}}(H_{\mathbf{C}}^{m-1}(C))$ for all integers, $m \geq 1$.[31] It can easily be verified that this recursive mapping reaches a fixed point after a finite number of iterations. If the smallest such number is denoted by $m(C|\mathbf{C}) \equiv \min\{m : H_{\mathbf{C}}^m(C) = H_{\mathbf{C}}^{m+1}(C)\}(< \infty)$, then the fixed point of this mapping, say, $h_{\mathbf{C}}(C)[\equiv H_{\mathbf{C}}^{m(C|\mathbf{C})}(C)]$, identifies the unique central cluster generated by each cluster $C \in \mathbf{C}$. Accordingly, we now define the corresponding *agglomeration*, $\mathcal{A}(C)$, generated by C to be the solidifiation of *all* clusters leading to the *same* central cluster, $h_{\mathbf{C}}(C)$, i.e. $\mathcal{A}(C) \equiv \sigma(\{C' \in \mathbf{C} : h_{\mathbf{C}}(C') = h_{\mathbf{C}}(C)\})$. Note that if C is an isolated cluster, i.e. if $N_{\mathbf{C}}(C) = C$, then by definition, $\mathcal{A}(C) = C$. Moreover, for all clusters, $C, C' \in \mathbf{C}$, either $\mathcal{A}(C) = \mathcal{A}(C')$ or $\mathcal{A}(C) \cap \mathcal{A}(C') = \varnothing$. So this procedure essentially transforms the cluster scheme, **C**, by grouping its contiguous clusters into distinct agglomerations, each with a central cluster.

---

29   Recall from Section 3.1 that $N(r)$ is the set of adjacent neighbors for basic region $r \in R$.
30   In practice, this solution is almost always unique. But if not, then additional conditions must be imposed to ensure uniqueness (such as choosing the maximal-density cluster with the largest number of establishments).
31   For example, $H_{\mathbf{C}}^1(C) = H_{\mathbf{C}}(H_{\mathbf{C}}(C)) = \arg\max_{C' \in N_{\mathbf{C}}(H_{\mathbf{C}}(C))} n_{C'}/a_{C'}$.
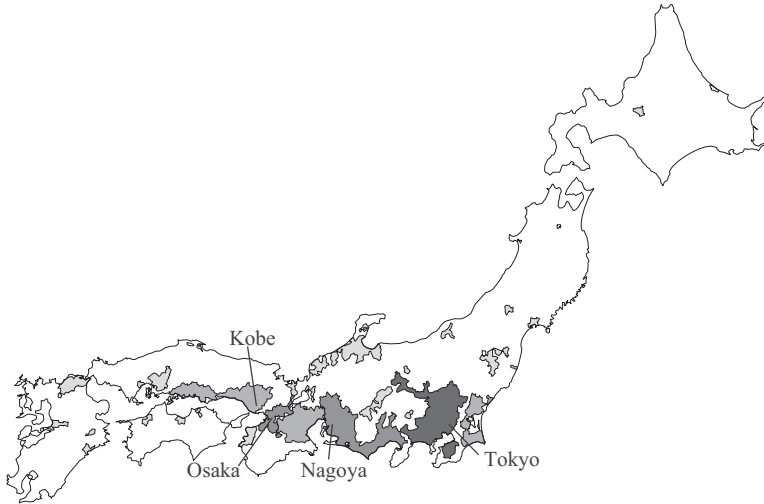
Figure 15. Agglomerations of plastics industry.

Agglomerations identified for the plastic industry are shown in Figure 15, where 43 clusters reduced to 30 agglomerations, where darker colors indicate larger concentrations of establishments. Notice in particular that certain individual clusters in the Tokyo, Nagoya and Osaka areas have now been joined to larger agglomerations in these respective areas.

5.4. Sensitivity analysis

Finally, we report on the sensitivity of identified cluster schemes with respect to small perturbations to both the search algorithm and to regional boundaries. In doing so, it should be stressed that our main objective has been to propose the first practical framework for identifying industrial clusters on a map using regional data. So many refinements of the present search procedure are yet to be made, such as optimizing its computational efficiency. But even at this preliminary stage, it is nonetheless informative to consider the robustness of this procedure with respect to possible perturbations.

5.4.1. Alternative initial clusters

We first investigate the sensitivity of the results with respect to alternative starting points. To do so, we now re-initialize the cluster search procedure for a given industry by taking the initial cluster to be a randomly chosen municipality (with a strictly positive number of establishments of the industry in question). In particular, we have generated 10 such samples for each of $i = 1, \ldots, 154$ industries with non-spurious clusters.

To compare the overlap between cluster schemes identified for industry i in each of these samples with the original cluster scheme, we focus on their agreement in terms of establishments belonging to clusters. To do so, let R_{i0} and R_{ij} denote, respectively, the sets of municipalities in R belonging to clusters identified for industry i using (i) the

original initial cluster and (ii) the jth sampled initial cluster. If for any set of municipalities, $R' \subseteq R$ we let $n_i(R')$ denote the total number of i establishments in these municipalities, then the agreement between these cluster schemes can be measured in terms of the *share*, S_{ij}, of cluster establishments common to both, as defined by

$$S_{ij} = \frac{n_i(R_{i0} \cap R_{ij})}{n_i(R_{i0} \cup R_{ij})},\tag{5.2}$$

where $0 \leq S_{ij} \leq 1$ with $S_{ij} = 0 \Leftrightarrow R_{i0} \cap R_{ij} = \varnothing$ and $S_{ij} = 1 \Leftrightarrow R_{i0} = R_{ij}$.[32] Over the full set of samples, $\{S_{ij} : i = 1, \ldots, 154, j = 1, \ldots, 10\}$, the minimum value of S_{ij} observed was 0.952 (with a mean of 0.999). On this basis, we conclude that the cluster schemes identified are highly robust against the perturbation of initial clusters.

5.4.2. Perturbation of regional divisions

Next, we employ simulation methods to determine whether the identified cluster schemes are sensitive to small perturbations of municipality boundaries. To construct each perturbation, we first randomly partition the set of all municipalities, R, into mutually exclusive adjacent pairs. In particular, if $\mathbf{N} = \{(r, s) \in R : r \in N(s)\}$ denotes the set of all adjacent municipality pairs, then this partition is given by a randomly selected maximal subset, $\mathbf{N}'$, of mutually exclusive pairs in $\mathbf{N}$ [which by definition satisfies the two conditions, that (i) $\{r, s\} \cap \{r', s'\} = \varnothing$ for all $(r, s)(r', s') \in \mathbf{N}'$, and that (ii) for each $(r, s) \notin \mathbf{N}'$ there is some $(r', s') \in \mathbf{N}'$ with $\{r, s\} \cap \{r', s'\} \neq \varnothing$]. Second, for each pair, $(r, s) \in \mathbf{N}'$, we reallocate 5% of both establishments and economic area from one municipality to the other,[33] where the direction of the reallocation is determined randomly.

Using this procedure, we again generated 10 randomly perturbed samples for each of 154 industries. Within each industry, we focus only on the set of essential clusters in the original cluster scheme (using $\mu = 0.05$ as defined in Section 4.4). With respect to these clusters, we then compute the *industry share* (in terms of the number of establishments) that continues to appear in these essential clusters identified for the perturbation. In all but three industries, these industry shares exceeded 95% in each of the 10 sample perturbations. A common property of these three exceptional industries ('alcoholic beverages', 'paving materials' and 'cement and its products') is that they are relatively ubiquitous. As for paving materials, the original clusters account for only 39.1% of all establishments, which is the smallest among all industries (where the average value is 93.6%). Thus, the majority of establishments of this industry are located outside clusters. As for the other two, while original clusters do account for a substantial percentage of industry establishments (more than 70%), these clusters are spread over more than 40% of the national economic area, as opposed to an average share of 22.7% for all industries. Thus, for extremely ubiquitous industries of this type, the change in lumpiness of establishment distributions across municipalities induced by such perturbations in the regional allocation of establishments and economic area can in

---

32 This measure is an instance of the standard Jaccard measure of similarity between sets.
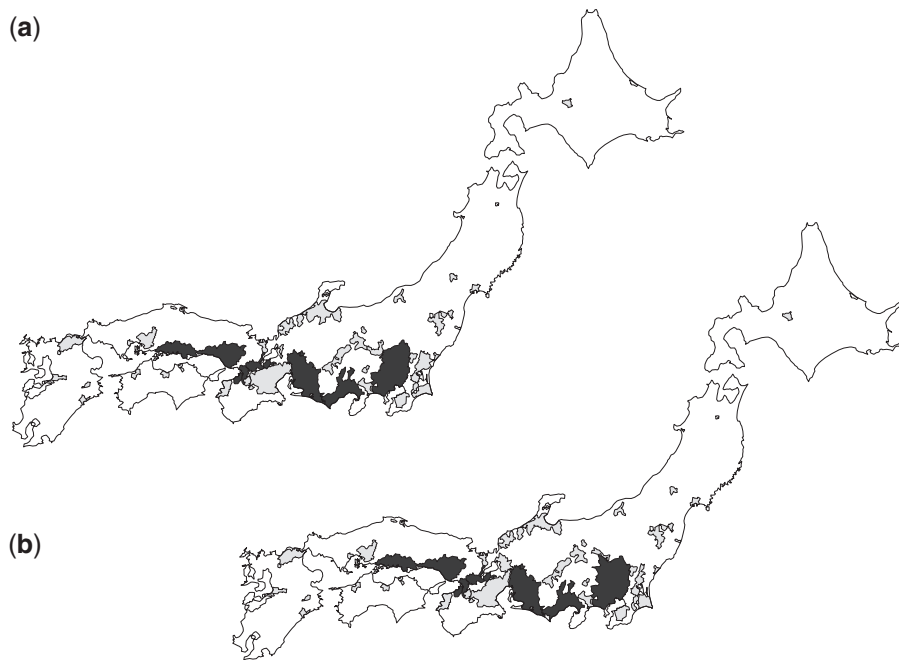33 The number of establishments to be reallocated is rounded to the nearest integer.

(a)

(b)

Figure 16. Clusters under the actual and perturbed municipalities boundaries (plastics industry). (a) Clusters under the actual municipality boundaries and (b) clusters under perturbed municipality boundaries.

principle produce quite different clustering patterns. But for the majority of industries, our cluster-identification procedure does appear to produce robust results.[34]

Figures 16 and 17 show for the cases of our two example industries, plastic and soft drinks industries, respectively. The top and bottom panels in each figure show the clusters identified under the actual and perturbed municipality boundaries, respectively, where for the latter, we chose the random sample for which the industry share deviates the most from the actual pattern, i.e. the worst case. In each panel, the darker clusters represent the essential clusters. These two examples indicate that not only the essential clusters are robust but also the entire cluster distributions are quite similar between the actual and perturbed sample. Basically, this same property holds for 151 of the 154 three-digit industries.

6. Concluding remarks

In this article, we have developed a simple *cluster-scheme* model of agglomeration patterns and have constructed an information-based algorithm for identifying such patterns. To the best of our knowledge, this constitutes the first systematic framework

---

34 It is to be noted that since municipality sizes vary significantly for the case of Japan, with geographic areas of municipalities ranging from 1.64 to 1408.1 km^2, these 5% reallocations of establishments and economic area are actually not 'small' perturbations for relatively large municipalities. Hence, the results here indicate strong robustness of our approach for the case of localized industries.
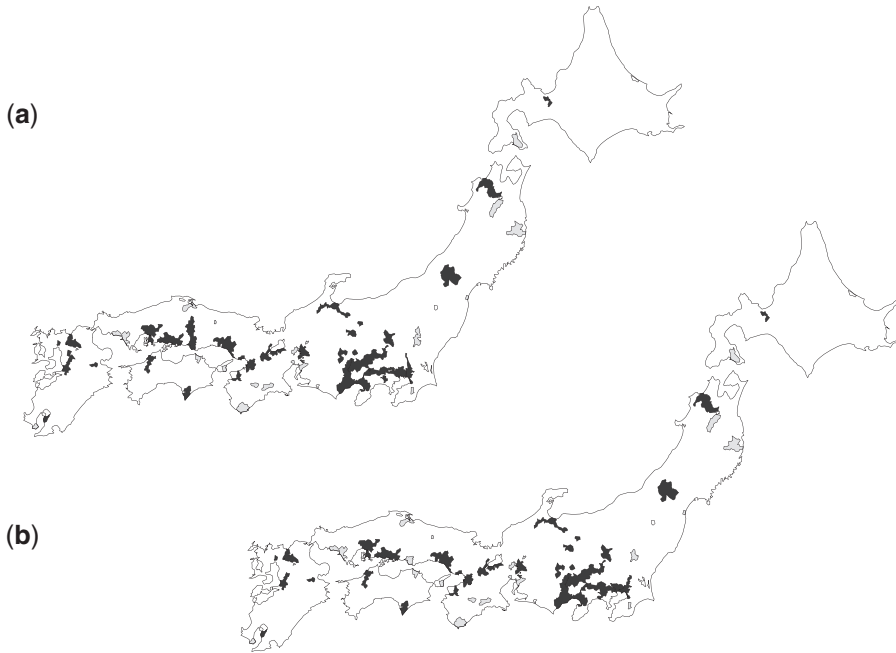
Figure 17. Clusters under the actual and perturbed municipalities boundaries (soft drinks industry). (a) Clusters under the actual municipality boundaries and (b) clusters under perturbed municipality boundaries.

for doing so. In addition, this formal framework opens up a number of possible directions for further research. In particular, by utilizing clusters identified, it becomes possible for the first time to directly identify the spatial patterns of industrial agglomerations on a map, and test the hypotheses implied by the recent theoretical developments on economic agglomerations under many-region/continuous location space (e.g. Fujita et al., 1999; Tabuchi and Thisse, 2011; Ikeda et al., 2011; Hsu, 2012). Below, we touch on two areas where initial investigations are already under way.

6.1. Cluster-based choice cities for industries

In our previous work (Mori et al., 2008), we reported on an empirical regularity between the (population) size and industrial structure of cities in Japan, designated as the *number-average size* (*NAS*) *rule*. This regularity (also established for the USA by Hsu, 2012) asserts a negative log-linear relation between the number and average population size of those cities where a given industry is present. Hence, the validity of the NAS rule depends critically on how such 'industrial presence' is defined. In its follow-up paper (Mori and Smith, 2011a), we have employed the present cluster-detection procedure to identify cities where given industries exhibit a 'substantial' presence with respect to their agglomeration patterns. In particular, if $\mathcal{U}$ denotes the relevant set of cities in R, and if $\mathbf{C}_i$ is the cluster scheme identified for industry i, then each city $U \in \mathcal{U}$ containing establishments from at least one of the clusters in $\mathbf{C}_i$ is designated as a *cluster-based* (*cb*) *choice city* for industry i. This *cb*-approach to industrial presence yields a sharper version of the NAS rule for the case of Japan. In

addition, by identifying those *cb*-choice cities shared by different industries, this also provides one approach to analyzing *spatial coordination* between industries. In ongoing work (Hsu et al., 2012), we are examining the consequences of such industrial coordination for city size distributions, and in particular for the *Rank Size Rule*. In addition, by examining the *spacing* between cb cities for industries, one can also formulate a range of testable propositons about the spatial structure of urban hierarchies.

6.2. Regional agglomeration analysis

As emphasized in Section 1, most analyses of industrial agglomeration have relied on overall indices of agglomeration, and hence have necessarily been aggregate in nature. However, the present identification of local cluster patterns for industries allows the possibility for more disaggregate spatial analyses. Of particular interest is the question of *why* industries agglomerate in certain regions and not others. While this question has of course been addressed by a variety of theoretical models, there has been little empirical work done to date. This is in large part due to the conspicuous absence of 'local agglomeration' measures. While the present cluster-scheme model is not itself numerical, it nonetheless suggests a number of possibilities for such measures.

The simplest are of course binary variables indicating the 'presence' or 'absence' of agglomeration. Indeed, the above definition of cb choice cities yields precisely a binary variable of this type on the set of cities, $\mathcal{U}$. Hence, given appropriate socio-economic data for cities, $U \in \mathcal{U}$, one could in principle test for significant predictors of industrial presence in these cities by employing standard logit or probit models.

Alternatively, one may focus directly on the individual clusters for each industry. Here, one might characterize the degree of local agglomeration for each industry in terms of the contribution of these clusters to the industry as a whole. Natural candidates include the fraction of industry establishments or employment in each cluster. Given the availability of data at the municipality level, one could in principle aggregate such data to the cluster level and use this to identify predictors of local agglomeration by more standard types of linear regression models. As one illustration, in Japan, data on education levels (among others) are available at the municipality level. Thus, by employing appropriate summary measures, 'education accessibility' across cluster municipalities can be defined. Then, by treating 'industry' as a categorical variable, one can attempt to compare the relative importance of these local accessibilities in attracting various industries. Regression analyses of this type will be presented in subsequent work (Mori and Smith, 2012).

Acknowledgments

Funding

References

Akaike, H. (1973) Information theory as an extension of the maximum likelihood principle. In B. N. Petrov, F. Csaki (eds) *Second International Symposium on Information Theory*, pp. 267–281. Budapest: Akademiai Kiado.

Berge, C. (1963) *Topological Spaces*. New York: MacMillan.

Besag, J., Newell, J. (1991) The detection of clusters in rare diseases. *Journal of the Royal Statistical Society, Series A*, 154: 143–155.

Billings, S. B., Johnson, E. B. (2012) A non-parametric test for industrial specialization. *Journal of Urban Economics*, 71: 312–331.

Brülhart, M., Traeger, R. (2005) An account of geographic concentration patterns in Europe. *Regional Science and Urban Economics*, 35: 597–624.

Castro, M. C., Singer, B. 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.

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

Ellison, G., Glaeser, E. L. (1997) Geographic concentration in US manufacturing industries: a dartboard approach. *Journal of Political Economy*, 105: 889–927.

Fujita, M., Krugman, P., Venables, A. J. (1999) *The Spatial Economy: Cities, Regions, and International Trade*. Cambridge, MA: MIT Press.

Henderson, J. V., Thisse, J.-F. (eds) (2004) *Handbook of Regional and Urban Economics, vol. 4*. Amsterdam: North-Holland.

Hsu, W. (2012) Central place theory and the city size Distribution. *Economic Journal*, 122: 903–932.

Hsu, W., Mori, T., Smith, T. E. (2011) Industrial location and city size: does space matter? Unpublished data

Ikeda, K., Akamatsu, T., Kono, T. (2012) Spatial period doubling agglomeration of a core-periphery model with a system of cities. *Journal of Economic Dynamics and Control*, 36: 754–778.

Kerr, W. R., Kominers, S. D. (2012) Agglomerative forces and cluster shapes. Working Paper 12-09, Center for Economic Studies, U.S. Census Bureau.

Kontkanen, P., Buntine, W., Myllymäki, P., Rissanen, J., Tirri, H. (2003) Efficient computation of stochastic complexity. In C. M. Bishop, B. J. Frey (eds) *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, pp. 181–188. Society for Artificial Intelligence and Statistics.

Kontkanen, P., Myllymäki, P. (2005) *Analyzing the Stochastic Complexity via Tree Polynomials*. Technical Report, 2005-4. Helsinki Institute for Information Technology.

Kullback, S., Leibler, R. A. (1951) On information and sufficiency. *Annals of Mathematical Statistics*, 22: 79–86.

Kulldorff, M. (1997) A spatial scan statistic. *Communications in Statistics—Theory and Methods*, 26: 1481–1496.

Kulldorff, M., Nagarwalla, N. (1995) Spatial disease clusters: detection and inference. *Statistics in Medicine*, 14: 799–810.

Marcon, E., Puech, F. (2010) Measures of the geographic concentration of industries: improving distance-based methods. *Journal of Economic Geography*, 10: 745–762.

Mori, T., Nishikimi, K., Smith, T. E. (2005) A divergence statistic for industrial localization. *Review of Economics and Statistics*, 87: 635–651.

Mori, T., Nishikimi, K., Smith, T. E. (2008) The number-average size rule: a new empirical relationship between industrial location and city size. *Journal of Regional Science*, 48: 165–211.

Mori, T., Smith, T. E. (2009) A probabilistic modeling approach to the detection of industrial agglomerations. Discussion Paper No.682, Institute of Economic Research, Kyoto University.

Mori, T., Smith, T. E. (2011a) An industrial agglomeration approach to central place and city size regularities. *Journal of Regional Science*, 51: 694–731.

Mori, T., Smith, T. E. (2011b) Analysis of industrial agglomeration patterns: an application to manufacturing industries in Japan. Discussion Paper No. 794, Institute of Economic Research, Kyoto University.

Mori, T., Smith, T. E. (2012) Spatial approach to identifying agglomeration determinants. In progress. Unpublished data.

Porter, M. E. (1990) *The Competitive Advantage of Nations*. New York: The Free Press.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, 6: 461–464.

Soltan, V. P. (1983) D-convexity in graphs. *Soviet Mathematics-Doklady*, 28: 419–421.

Tabuchi, T., Thisse, J.-F. (2011) A new economic geography model of central places. *Journal of Urban Economics*, 69: 240–252.

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 $\tilde{\rho} = (r_n, r_{n-1}, \ldots, r_2, r_1) \in \mathcal{P}(r_n, r_1)$ denote the *reverse path* in $\mathcal{P}$. Next, for any two paths, $\rho = (r_1, \ldots, r_n), \rho' = (r'_1, \ldots, r'_m) \in \mathcal{P}$, with $r_n = r'_1$, the combined path, $\rho \circ \rho' = (r_1, \ldots, r_n, r'_2, \ldots, r'_m) \in \mathcal{P}$ 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 $l(\rho \circ \rho') = \sum_{i=1}^{n-1} d(r_i, r_{i+1}) + d(r_n, r'_2) + \sum_{i=2}^{m-1} d(r'_i, r'_{i+1}) = \sum_{i=1}^{n-1} d(r_i, r_{i+1}) + \sum_{i=1}^{m-1} d(r'_i, r'_{i+1}) = l(\rho) + l(\rho')$. Using this and (3.5)–(3.8), it is convenient to establish the following well-known properties of *d*-convex sets, as in Definition 3.1 of the text. First, we show that for the *d*-convex*ification function*, $c_d : \mathcal{R} \to \mathcal{R}$, in (3.8), 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 (3.6), it follows that $\{r_1, r_2\} \subseteq I^{k_1 + k_2}(S)$. Thus, $\langle \rho \rangle \subseteq I(I^{k_1 + k_2}(S)) = I^{k_1 + k_2 + 1}(S) \subseteq c_d(S)$. ∎

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 $S \in \mathcal{R}$,*

$$c_d(S) = \cap\{C \in \mathcal{R}_d : S \subseteq C\} . \tag{A.1}$$

Proof: By Proposition A.1, $c_d(S) \in \mathcal{R}_d$, and by (3.5)

$$S \subseteq I(S) \subseteq c_d(S) \tag{A.2}$$

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 (3.4),

$$S \subseteq C \Rightarrow \bigcup_{r,s \in S} I(r,s) \subseteq \bigcup_{r,s \in C} I(r,s) \Rightarrow I(S) \subseteq I(C) . \quad (A.3)$$

Moreover, by (3.3) and (3.4) together with the definition of d-convexity, it follows that

$$C \in \mathcal{R}_d \Rightarrow I(C) = \bigcup_{r,s \in C} I(r,s) \subseteq C . \quad (A.4)$$

Hence, we may conclude from (A.3) and (A.4) 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 $S \in \mathcal{R}$,*

$$S \in \mathcal{R}_d \Leftrightarrow c_d(S) = S . \quad (A.5)$$

Proof: If $c_d(S) = S$ then $S \in \mathcal{R}_d$ by Proposition A.1. Conversely, if $S \in \mathcal{R}_d$ then $S \subseteq c_d(S)$ by (A.2), and $c_d(S) \subseteq S$ by Proposition A.2, hence $c_d(S) = S$. ∎

This in turn implies that the family, $\mathcal{R}_d$, of d-convex sets can be equivalently defined as in Expression (3.9) 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 (3.10) of the text, can easily be obtained from Proposition A.3 as follows:

Corollary (interval fixed points): *For all $S \in \mathcal{R}$,*

$$S \in \mathcal{R}_d \Leftrightarrow I(S) = S . \quad (A.6)$$

Proof: Since $S \in \mathcal{R}_d \Rightarrow I(S) \subseteq S$ by (A.4) (with $C = S$), and since $S \subseteq I(S)$ holds for all S [by (3.5)], 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 (3.8) 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 3.2 of the text. First, in a manner paralleling Proposition A.1, we show for the *solidification function*, $\sigma : \mathcal{R} \to \mathcal{R}$, defined by (3.12), 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 $r \in R \backslash V$ there is some path, $\rho \in \mathcal{P}(r, \bar{R})$ with $\langle \rho \rangle \cap V = \varnothing$. But for any $r \in R \backslash V = R \backslash (S \cup S_0)$, it follows that

$r \in R \backslash S$ and $r \notin S_0$, so that by the definition of S_0 in (3.11), it must be true that there is some boundary region, $\bar{r} \in \bar{R}$, 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 $r_0 \in S_0$, $\rho = (r, \dots, r_0, \dots, \bar{r}) = \rho_1 \circ \rho_2$ 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 $S \in \mathcal{R}$,*

$$S \in \mathcal{R}_\sigma \;\Leftrightarrow\; \sigma(S) = S . \tag{A.7}$$

Proof: If $S \in \mathcal{R}_\sigma$ then $S_0 = \varnothing$, so that $\sigma(S) = S$ by (3.12). Conversely, if $\sigma(S) = S$, then by Lemma A.1, $S \in \mathcal{R}_\sigma$. ∎

As a parallel to (A.6), this in turn implies that the family of solid sets in $\mathcal{R}$ can be equivalently defined as follows:

$$\mathcal{R}_\sigma = \{S \in \mathcal{R} : \sigma(S) = S\} . \tag{A.8}$$

Finally, solid sets also exhibit the following nesting property:

Lemma A.3 (solid nesting): *For all $S, V \in \mathcal{R}$,*

$$S \subseteq V \;\Rightarrow\; \sigma(S) \subseteq \sigma(V) . \tag{A.9}$$

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 $r \in S_0 \backslash V$. Here, we show that r must be in V_0. To do so, observe first that $r \notin V \Rightarrow r \in R \backslash V$. Moreover, $r \in S_0$ implies that for any path, $\rho \in \mathcal{P}(r, \bar{R})$ 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 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 $\langle \rho \rangle \cap [R \backslash \sigma(S)] \neq \varnothing$. 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 $\langle \rho \rangle \cap [R \backslash \sigma(S)]$, so that $\rho = (r_1, \dots, r, \dots, r_2) = \rho_1 \circ \rho_2$ 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 $r \notin \sigma(S) \Rightarrow [r \in R \backslash S$ and $r \notin S_0]$, there must be some path,
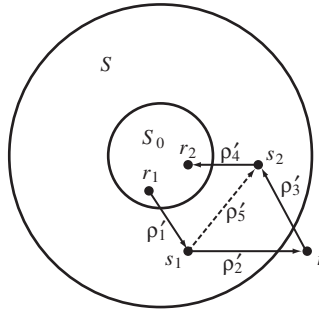
Figure 18. Example (*i*).

$\rho_3 \in \mathcal{P}(r, \bar{R})$ with $\langle \rho_3 \rangle \cap S = \varnothing$. Hence, the combined path, $\rho_1 \circ \rho_3 \in \mathcal{P}(r_1, \bar{R})$, 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 $\tilde{\rho}_2 \in \mathcal{P}(r_2, r)$ 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 $\rho = (r_1, \ldots, s_1, \ldots, r, \ldots, s_2, \ldots, r_2) = \rho_1' \circ \rho_2' \circ \rho_3' \circ \rho_4'$ 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 18.

But if we choose any shortest path, $\rho_5' \in \mathcal{P}_d(s_1, s_2)$ (as in Figure 18), 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 $\langle \rho_2' \circ \rho_3' \rangle \not\subseteq S$]. Hence, for the path, $\rho' = \rho_1' \circ \rho_5' \circ \rho_4' \in \mathcal{P}(r_1, r_2)$, we must have $l(\rho') = l(\rho_1') + l(\rho_5') + l(\rho_4') < l(\rho_1') + [l(\rho_2') + l(\rho_3')] + l(\rho_4') = l(\rho_1' \circ \rho_2' \circ \rho_3' \circ \rho_4') = l(\rho)$ 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 $\rho = (r_1, \ldots, s_1, \ldots, r, \ldots, r_2) = \rho_1' \circ \rho_2' \circ \rho_2$ 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 19.

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 19), 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 $l(\rho'') = l(\rho_1') + l(\rho_6') < l(\rho_1') + [l(\rho_2') + l(\rho_2)] = l(\rho_1' \circ \rho_2' \circ \rho_2) = l(\rho)$ 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 $\langle \rho \rangle \cap [R \backslash \sigma(S)] \neq \varnothing$, 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 3.3. First, we show that for the *d*-convex *solidification function*, $\sigma c_d : \mathcal{R} \to \mathcal{R}$, in (3.13), 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.*
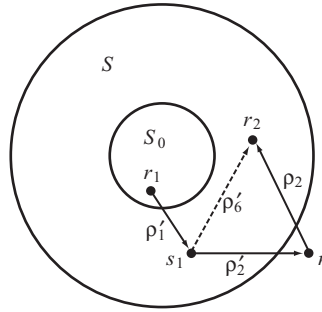
Figure 19. Example (*ii*).

Proof: First observe from Definition 3.3 that we may use Expressions (A.6) and (A.7) to define the family of all *d*-convex solids in equivalent terms as

$$\mathcal{R}_{\sigma d} = \mathcal{R}_\sigma \cap \mathcal{R}_d \,. \tag{A.10}$$

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, $S \in \mathcal{R}$,*

$$\sigma c_d(S) = \cap \{C \in \mathcal{R}_{\sigma d} : S \subseteq C\} \,. \tag{A.11}$$

Proof: First observe from Theorem A.2 that $\sigma c_d(S) \in \mathcal{R}_{\sigma d}$ and from Expression (A.2) 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 obtain the following conclusion from Lemma A.3 together with Lemma A.2, and the result is established.

$$c_d(S) \subseteq C \Rightarrow \sigma[c_d(S)] \subseteq \sigma(C) = C \Rightarrow \sigma c_d(S) \subseteq C∎ \tag{A.12}$$

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 $S \in \mathcal{R}$,*

$$S \in \mathcal{R}_{\sigma d} \iff c_{\sigma d}(S) = S \,. \tag{A.13}$$

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 $c_{\sigma d}(S) = \sigma[c_d(S)] = \sigma(S) = S$, and the result is established.∎