An Industrial Agglomeration Approach to Central Place and City Size Regularities

Tomoya Mori *and Tony E. Smith †

December 2009

Abstract

An empirical regularity designated as the *Number-Average Size* (*NAS*) *Rule* was first identified for the case of Japan by Mori, Nishikimi and Smith [13], and has since been extended to the US by Hsu [6]. This rule asserts a negative log-linear relation between the number and average population size of cities where a given industry is present, i.e., of industry-choice cities. Hence one of its key features is to focus on the presence or absence of industries in each city, rather than the percentage distribution of industries across cities. But despite the strong empirical regularity of this rule, there still remains the statistical question of whether such location patterns could simply have occurred by chance. In this paper an alternative approach to industry-choice cities is proposed. This approach utilizes the statistical procedure developed in Mori and Smith [15] to identify spatially explicit patterns of agglomeration for each industry. In this context, the desired industry-choice cities are taken to be those (economic) cities that constitute at least part of a significant spatial agglomeration for the industry. These *cluster-based* choice cities are then used to reformulate both the NAS Rule and the closely related *Hierarchy Principle* of Christaller [2]. The key empirical result of the paper is to show that the NAS Rule not only continues to hold under this new definition, but in some respects is even stronger. The Hierarchy Principle is also shown to hold under this new definition. Finally, the present notion of cluster-based choice cities is also used to develop tests of both the *locational diversity* of industries and the *industrial diversity* of cities in Japan.

*Institute of Economic Research, Kyoto University

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

1 Introduction

A remarkable empirical regularity between the (population) size and industrial structure of cities in Japan was reported in our previous paper, Mori, Nishikimi and Smith [13]. This regularity, designated as the *Number-Average Size* (NAS) *Rule*, showed that for a given set of Japanese industrial data[1] there is a strong negative log-linear relationship between the number and average size of *industry-choice cities* in which establishments of each given industry operate.[2] Subsequently, the same regularity was reported by Hsu [6] for the US, using comparable definitions of both industries and cities.

The validity of this rule, however, depends critically on how "industry-choice cities" are defined. In both of the above papers, such cities for a given industry were taken to be those with a positive share of the industry employment. Hence there remains the question of whether such an industrial presence could simply have occurred by chance. Indeed, if cities with only a single establishment of the industry are included, then such chance occurrences would seem to be quite likely.

Hence the central purpose of the present paper is to develop a more meaningful definition of industry-choice cities, and to reconfirm the NAS Rule for Japan in these terms. In particular, we seek to identify for each industry those cities with a *substantial presence* of that industry. While it is possible to simply strengthen the above definition in terms of some minimal threshold share of establishments or employment (say 5% of national totals),[3] the choice of such a threshold is necessarily ad hoc. Hence the approach adopted here is to characterize *substantial presence* in terms of "significant industrial agglomerations". This approach draws on the statistical procedure recently developed by Mori and Smith [15] to identify spatially explicit patterns of significant clustering (agglomeration)[4] for any given industry. In this context, the desired choice cities for an industry are taken to be those which share at least part of a significant cluster for that industry, and are here designated as *cluster-based choice cities*.

The key empirical result of this paper is to show that the NAS Rule not only continues to hold under this new definition, but in some respects is even stronger. In particular, the few outlier industries found for Japan (2001) in Mori et al. [13] turn out to be precisely those industries for which *no significant agglomeration* can be identified. Hence this finding serves to suggest that there may indeed be a strong underlying connection between this NAS Rule and phenomenon of industrial agglomeration itself.

As was also shown in Mori et al. [13, Section 5], there is a strong connection between this Rule and two classical regularities: the *Rank-Size Rule* for cities, and the *Hierarchy Principle* for industries. The former asserts a log-linear relationship between the (population) size and the rank in terms of size of cities. The latter, which is an essential feature of the *Central Place*

[1] In particular, this data was for two time points, 1980 and 2000 (where 1981 establishment location data was associated with the 1980 population data and similarly, 1999 establishment location data was associated with the 2000 population data).

[2] Our present notion of a "city" is taken to be an "urban employment area" as discussed in Section 2.5 below.

[3] Such an approach was investigated in Mori et al.[13], where it was found that the NAS rule for Japan (2001) is indeed robust up to thresholds of around 5%.

[4] We shall also use the terms "cluster" and "agglomeration" interchangeably. See however the discussion in Section 8.1 of Mori and Smith [15] for a possible distinction between these concepts.

Theory of Christaller [2], asserts that industries found in a city of a given size should also be found in all cities at least as large. In particular, it was shown that in the presence of the Hierarchy Principle, the NAS Rule and Rank-Size Rule are in certain respects equivalent. So evidence for the NAS Rule should in principle have consequences for both of these additional types of empirical regularities. Hence a final objective of this paper is to show that the empirical support for both the Rank-Size Rule and Hierarchy Principle found by Mori et al.[13, Section 5] for Japan continues to hold in terms of cluster-based choice cities.

To establish these results, we begin in Section 2 below with an overview of the cluster-detection procedure developed in Mori and Smith [15]. This forms the basis for our subsequent definition of cluster-based choice cities in Section 3. The natural converse of this concept is the notion of *cluster-based choice industries* for each city, as defined in the same section. This concept in turn provides natural extensions of the tests of the Hierarchy Principle in Mori et al. [13, Section 5]. Such extensions are developed in Section 4, and include tests of both the *locational diversity* of an industry as determined by the number of its cluster-based choice cities, and the *industrial diversity* of a city as determined by the number of its cluster-based choice industries. Finally, similar extensions with respect to the NAS Rule are presented in Section 5. The paper concludes in Section 6 with a brief discussion of some directions for further research.

2 Industrial Cluster Analysis

As mentioned above, the present paper draws heavily on the cluster-detection procedure developed in Mori and Smith [15]. This approach to identifying clusters of regions (municipalities) for a given industry is closely related to the statistical clustering procedures proposed by Besag and Newell [1], Kulldorff and Nagarwalla [11], and Kulldorff [10]. To test for the presence of clusters, these procedures start by postulating an appropriate null hypothesis of "no clustering". In the present case, this hypothesis is characterized by a uniform distribution of industrial locations across regions (as discussed further in Section 2.3 below). Such clustering procedures then seek to determine the single "most significant" cluster of regions with respect to this hypothesis. Candidate clusters are typically defined to be approximately circular areas containing all regions having centroids within some specified distance of a given reference point (such as the centroid of a "central" region).

The approach developed in Mori and Smith [15] extends these procedures in two ways. First, the notion of a "circular" cluster of regions is extended to the (metric based) notion of *convex solids* which is meaningful for more general distance structures such as road networks. Second, individual (convex solid) clusters are extended to the more global concept of *cluster schemes*. Hence it is appropriate to begin by sketching these basic concepts in Sections 2.1 and 2.2, respectively. This is followed in Section 2.3 with a brief outline of the cluster-detection procedure based on these concepts. In addition, the test of significance for the resulting cluster schemes is reviewed in Section 2.4. Finally, we briefly describe the industrial and city data sets that will be used here.

2.1 Clusters

We begin with a set, R, of relevant *regions* (municipalities), r, within which each industry can locate. An *industrial cluster* is then taken roughly to be a spatially coherent subset of regions within which the density of industrial establishments is unusually high. Since the explicit construction of such clusters will have consequences for our present definition of cluster-based cities, it is appropriate to outline this construct more explicitly. Here we begin by noting that "spatial coherence" is taken to include the requirement that such regions be contiguous, and as close to one another as possible – where "closeness" is defined with respect to the relevant underlying road network. Using network distances between regional centers, we define *shortest paths* between each pair of regions, r_i and r_j, to be sequences of intermediate regions, $(r_i, r_1, .., r_k, r_j)$ reflecting minimum travel distances with respect to the road network.[5] Hence the key requirement here is that a cluster of regions be *convex* in the sense that it includes all shortest paths between its member regions. But unlike the usual notion of planar convexity with respect to Euclidean distance, the convex clusters may have "holes" in them. An illustrative example is given in the first two panels of Figure 2.1 below.

Figure 2.1 here

Here a stylized system of regions, R, is represented by a grid of square regions. The portion shown in Figure 2.1 is taken to be a small part of R. The set, S, of four black regions in Figure 2.1(a) depicts a grouping of regions where industry density is unusually high (as discussed further below). But while these four regions are close enough to each other to be considered as a single "cluster", they are not contiguous. Hence one would like to "convexify" this set to obtain a more coherent cluster. Here it is assumed that the road network in R has a system of major roads, part of which is shown by the four heavy lines in Figure 2.1(b). Hence the industry concentrations in Figure 2.1(a) are seen to be at crossroads of the major network (possibly to minimize shipping costs). In addition, there is also a finer network of minor roads indicated schematically by the dashed lines in Figure 2.1(b). But these local roads are in fact more circuitous in nature, and hence are effectively much longer. Hence if the travel distance, t, between adjacent regions on the major network is set as $t = 1$, then it is assumed that travel distance between adjacent regions on minor roads is $t = 3$.[6] With respect to this network it is easily seen that all the shortest paths between the members of S consist of the regions on major roads connecting them, as shown by gray in Figure 2.1(c). But in fact, this ring of regions also contains all shortest paths between each pair of its regions. For example, the shortest path in the ring between regions r_1 and r_2 shown in Figure 2.1(c) is seen to be $t = 7$, while the straight-line path between them on minor roads has distance $t = 9$. Hence this ring constitutes the desired *convexification* of S.[7]

[5] Technically these shortest paths may in many cases be longer than actual shortest routes on the network. For additional details see Mori and Smith [15, Section 4.1].

[6] This differences may also be interpreted in terms of effective travel times.

[7] More generally, convexification is an iterative process that requires successively adding the minimal paths of new points until no further new points are added. See Mori and Smith [15, Section 4.2].

But since the six regions inside the ring are not on any shortest path, this convex set contains a large "hole". Hence to obtain a more coherent cluster, one would like to "fill in" this hole. The only complication here is defining the "inside" versus the "outside" of a set, so that holes can be identified and eliminated. The details of this procedure (which defines "outside" with respect to the boundary of the full regional system, R) are given in Mori and Smith [15, Section 4.3]. This process of "solidifying" a convex set is called *convex solidification*, and is detailed more fully in Mori and Smith [15, Section 4.4]. The resulting *convex solids* then constitute the desired class of candidate *clusters* for our purposes. A particular set of cluster examples (for the "livestock products" industry in Japan) are illustrated and discussed in Section 2.3 below.

2.2 Cluster Schemes

Industrial agglomeration patterns generally consist of multiple clusters that are necessarily related to one another. In fact, the spacing between such clusters is a topic of considerable economic interest.[8] Hence it is essential to model such patterns as explicit spatial arrangements of multiple clusters. The simple model proposed in Mori and Smith [15, Section 2] is that of a *cluster scheme*, $\mathbf{C} = (R_0, C_1, .., C_{k_\mathbf{C}})$, that partitions R into one or more disjoint clusters (convex solids), $C_1, .., C_{k_\mathbf{C}}$, together with the residual set, R_0, of all non-cluster regions in R. The individual clusters are implicitly taken to be areas in R where industry density is unusually high. But for modeling purposes, all that is assumed is that inside each cluster, C_j, the location probabilities for randomly sampled industrial establishments is uniform across all locations. Hence if the feasible area[9] for locations in each region, $r \in R$, is denoted by a_r, so that the total area of C_j is $a_{C_j} = \sum_{r \in C_j} a_r$, then the conditional probability of an establishment locating in $r \in C_j$ given that it is located in C_j is simply a_r / a_{C_j}. With this assumption, the only unknown probalities are the marginal location probabilities, $p_\mathbf{C}(j)$, for clusters C_j in $\mathbf{C}$. Hence each cluster scheme, $\mathbf{C}$, generates a candidate *cluster probability model*, $p_\mathbf{C} = [p_\mathbf{C}(j) : j = 1, .., k_\mathbf{C}]$, of establishment locations for the industry.[10] These cluster probability models, $p_\mathbf{C}$, thus amount formally to multinomial sampling models on their underlying cluster schemes, $\mathbf{C}$, with respect to the n establishments for a given industry.[11] Finally, since the observed relative frequencies, $f_\mathbf{C} = [f_C(j) = n_j/n : j = 1, .., k_\mathbf{C}]$, of establishments in each cluster are natural maximum-likelihood estimates of these (multinomial) probabilities, these estimates yield a family of well-defined candidate probability models for describing the agglomeration patterns of each industry.

2.3 Cluster-Detection Procedure

The only question remaining is how to compare these models to find "best" representative model. While many goodness-of-fit criteria are possible, it is argued in Mori and Smith [15, Section 3] that the Bayes Information Criterion (BIC) offers a number of distinct advantages. If the (multinomial) log-likelihood of each cluster scheme, $\mathbf{C}$, given $f_\mathbf{C}$ is denoted by $L_\mathbf{C}(f_\mathbf{C})$,

[8] See, for example, the discussion in Mori and Smith [15, Section 8.2].
[9] Feasible area is here taken to be *economic area* as defined in Section 2.5.4 below.
[10] This probability model is completed by the condition that $p_C(R_0) = 1 - \Sigma_j p_\mathbf{C}(j)$.
[11] See Mori and Smith [15, footnote 12] for related model-based clustering approaches.

then the BIC value for $\mathbf{C}$ is given by

$$BIC_{\mathbf{C}} = L_{\mathbf{C}}(f_{\mathbf{C}}) - \frac{k_{\mathbf{C}}}{2}\ln(n) \tag{2.1}$$

Hence BIC is essentially a penalized goodness-of-measure. Here "goodness-of-fit" is identified with the log-likelihood, $L_{\mathbf{C}}(f_{\mathbf{C}})$, which will assign higher values to those cluster schemes, $\mathbf{C}$, in which the relative frequencies in $f_{\mathbf{C}}$ are indeed "unusually high" relative to those in other cluster schemes. The second term then penalizes those cluster schemes, $\mathbf{C}$, with higher numbers of clusters ($k_{\mathbf{C}}$) relative to the total number of establishments, n (to avoid "over fitting" the data).

Given this criterion function, the *cluster-detection procedure* developed in Mori and Smith [15, Section 5] amounts to a systematic way of searching the space of possible cluster probabity models above to find a cluster scheme, $\mathbf{C}^*$, with a maximum value of $BIC_{\mathbf{C}^*}$.[12] While the details of this search procedure will play no role in the present analysis, the results of this procedure for Japanese industries will play a crucial role. Hence it is appropriate to illustrate these results in terms of the "livestock products" industry in Japan, shown in Figure 2.2 below.

Figure 2.2 here

Here Figure 2.2(a) shows the relative density of "livestock products" establishments in each municipality of Japan,[13] where darker patchs correspond to higher densities. Figure 2.2(b) shows the cluster scheme, $\mathbf{C}^*$, that was produced for the "livestock products" industry by this cluster-detection procedure. Here it is seen that not all isolated patches of density are clusters. But the highest density areas do indeed yield significant clusters. Notice also that while these clusters are by no means circular, the convex solidification procedure above has produced easily recognizable clusters that do seem to reflect the shapes of these high density areas.

2.4 A Test of Significant Clustering

Finally it should be emphasized that even random locational patterns are not perfectly uniform, and hence will tend to exhibit some degree of clustering. So there remains the statistical question of whether the "locally best" cluster scheme, $\mathbf{C}^*$, found for an industry by the above procedure is significantly better (in terms of BIC values) than would be expected in a random location pattern. This can be tested in a straightforward way by (i) generating N random location patterns for the establishments of a given industry, (ii) determining the locally optimal values, say BIC^*_s, for each simulated pattern, $s = 1, .., N$, and (iii) comparing the value, $BIC_{\mathbf{C}^*}$, with this sampling distribution of BIC values. If $BIC_{\mathbf{C}^*}$ is sufficiently large (say in the top 5% of these values), then one may conclude that the clustering captured by $\mathbf{C}^*$ is significantly higher

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

[13] These municipalities are mapped in Figure 2.3 below

than what would be expected under randomness. Otherwise, $\mathbf{C}^*$ is said to involve *spurious clustering*.[14]

2.5 Data for Analysis

In this section, we decribe the data sets to be used in this paper. The regional data, industrial data and spatial network data are the same as those used in Mori and Smith [15], and are summarized in Sections 2.5.1, 2.5.3, and 2.5.4, respectively. The new element here is data for cities, which is summarized in Section 2.5.2 (and which in part overlaps that used in Mori and Smith [14]).

2.5.1 Basic Regions

The basic regions, $r \in R$, in the present study are taken to be *municipalities* in Japan[15] [including cities,[16] wards, towns and villages] as of October 1, 2001.[17] While there are a total of 3,363 municipalities in Japan, we take R to include only 3,207 of these (as shown in Figure 2.3), namely those that are geographically connected to the major islands of Japan (Honshu, Hokkaido, Kyushu and Shikoku). This is convenient for the identification of clusters, as discussed further in Mori and Smith [15, Section 7.1.1].

Figure 2.3 here

2.5.2 City Data

In terms of these basic regional units, an (economic) *city* is formally defined to be an *Urban Employment Area* (UEA), as proposed originally by Kanemoto and Tokuoka [9]. Each UEA is designed to be an urban area of Japan that is comparable to a Core Based Statistical Area (CBSA) in the US.[18] Hence each UEA consists of a core set of municipalities designated as its business district (BD) together with a set of suburban municipalities from which workers commute toward the BD. Following Kanemoto and Tokuoka [9], UEAs are constructed as aggregations of municipalities by a recursive procedure that is detailed in Mori et al.[13].[19]

Using the municipality population and commuting data from the Population Census of Japan in 2000 (Japan Statistics Bureau [7]), 258 cities are identified (see Figure 2.4) which account, respectively, for 92% of the national population, 92% of total employment, and 55% of total area in 2000. As is typically the case, the population distribution among these cities is quite

[14] For additional details, see Mori and Smith [15, Section 5.3].

[15] In Japan, the "municipality"category is designated as *shi-ku-cho-son*.

[16] It is important to note here that "cities" in this municipality catergory are defined in terms of political boundaries, and are not to be confused with "cities" as Urban Employment Areas in Section 2.5.2 below.

[17] The data source for the definition of "municipalities" is the Statistical Information Institute for Consulting and Analysis [18, 19].

[18] See the US Office of Management and Budget [17] for the definition of a CBSA.

[19] Basically this construction starts with a large "seed" municipality, designated as the central municipality of the UEA. This in turn is extended to a BD and an appropropriate set of suburban municipalities.

skewed, with city populations ranging from 31.8 million in Tokyo down to 19,689 in Kucchan (while the average population size is 445,088). Here it should be noted that the present set of cities is larger than that used in the original NAS analysis of Mori et al.[13]. In particular, we here include all UEAs as defined by Kanemoto and Tokuoka [9], i.e., those with a central municipality population of at least 10,000.[20]

Figure 2.4 here

2.5.3 Industry Data

The industry and establishments data used for this analysis is based on the Japanese Standard Industrial Classification (JSIC) in 2001. In particular, we focus on three-digit manufacturing industries, of which 163 industrial types are present in the set of basic regions chosen for this analysis.[21] The establishment counts across these 163 industries is taken from the Establishment and Enterprise Census of Japan [8] in 2001. Such counts range from 1 to 38,643 within the present regional system, R (with a mean and median of 3,958 and 1,825, respectively).[22] Here it should be noted that the original NAS analysis of Mori et al.[13] used a much larger set of 264 industries, including services, wholesale, and retail, as well as manufacturing. However, since manufacturing exhibits a wider and more interesting variety of location patterns at the three-digit level, we choose to focus on these industries.[23]

In this context, the test of cluster significance in Section 2.4 above revealed that the clustering found in nine of these industries was in fact *spurious* (at the 5% level). The main reason for rejection in these cases [which include seven arms-related industries (JSIC331-337), together with "tobacco manufacturing" (JSIC135) and "coke" (JSIC213)], appears to be the small size of these industries.[24] But these industries are special in other ways. For example, both tobacco manufacturing and arms-related industries are highly regulated in Japan, with location patterns influenced by many non-economic factors. Further discussion of these "outlier" industries is given in Section 5 below (where these industries are labeled explicitly in Figure 5.1).[25] Hence, for the present, it suffices to say that all subsequent analyses in this paper are based on the 154 industries which exhibit some significant degree of clustering.

2.5.4 Spatial Data

The notion of "feasible area", a_r, for each basic region (municipality), $r \in R$, employed in Section 2.2 above is here taken to be the *economic area* of r, as defined by the Statistical Information

[20] Mori et al. [13] used only Metropolitan Employment Areas (MEA), i.e., UEAs with central municpality populations of at least 50,000.

[21] More precisely, out of total 164 industrial types in the data, all but one has establishments in R.

[22] In addition, 147 (90%) of these industries have more than 100 establishments, and 125 (77%) have more than 500 establishments.

[23] See Mori and Smith [14, Section 2.2].

[24] The average number of establishments for these industries is 7.89 (in contrast to an average of 4189 establishments for all other industries).

[25] See also discussions in Mori and Smith [14, p.108].

Institute for Consulting and Analysis [18, 19]. This definition of area essentially excludes forests, lakes, marshes and undeveloped areas in r.[26]

In addition, recall from the discussion of shortest-path distances in Section 2.1 above that such distances are derived from an underlying road network. In the present application, distances between adjacent municipalities, $r_1, r_2 \in R$, are defined in terms of the shortest-route distance between their municipality offices on the public road network in Japan. The relevant road-network data is taken from Hokkaido-chizu Co. Lit.[5]. From the computed shortest-route distances between neighboring municipalities, the corresponding shortest-path distances and shortest-path sequences of municipalities between each pair of municipalities are then obtained.[27]

3 Cluster-Based Choice Cities and Industries

In this section we use the clusters identified by the detection procedure above to strengthen the notion of industry-choice cities utilized in Mori et al.[13]. This sharper cluster-based version is developed in Section 3.1 below. The parallel city-oriented notion of cluster-based choice industries is then developed in Section 3.2.

3.1 Cluster-Based Choice Cities

Here we start in Section 3.1.1 by reviewing the original concept of industry-choice cities used in Mori et al.[13]. The extended cluster-based version is then developed in Section 3.1.2. Finally these two definitions are compared empirically in Section 3.1.3 with respect their relative industrial concentrations.

3.1.1 PB-Choice Cities

As mentioned in the Introduction, an industry-choice city was defined in Mori et al.[13] to be any city with a positive share of the employment in that industry. To be more precise, we now denote the set of all *cities* (UEA's) in the regional system R by $\mathcal{U}$, and denote the set of all relevant *industries* by I. Then if the total number of establishments in each industry, $i \in I$, in city $U \in \mathcal{U}$ is denoted by n_{iU}, the set of cities with *positive i-employment* is given by

$$\mathcal{U}_i^+ = \{U \in \mathcal{U} : n_{iU} > 0\} \tag{3.1}$$

Equivalently, $\mathcal{U}_i^+$ is the set of cities where i is present. Hence in this context, it is convenient to designate each city $U \in \mathcal{U}_i^+$ as a *presence-based* (*pb*) choice city for i. A possible shortcoming of this concept (also noted in the Introduction) is that the presence of a few establishments in a city isolated from the rest of the industry may have little significance in terms of the overall

[26] The economic area of Japan as as a whole is 120,205km^2, which amounts to 31.8% of the total area in Japan. Among individual municipalities the proportions of total area that constitute economic area range from 2.1% to 100%, with a mean of 48.5%. For a detailed justification of the use of economic area here, see the discussion in Mori and Smith [15, Section 7.1.2].

[27] Based on this data, the resulting shortest-path distances between (non-adjacent) pairs of municipalities appear to approximate their corresponding shortest-route distances quite well. See Mori and Smith [15, Section 7.1.3] for a further detail.

spatial structure of that industry. While it is difficult to be precise here, this shortcoming can nonetheless be illustrated by examples. For this purpose, we again use the "livestock products" industry in Figure 2.2 above and now show an enlargement of the northern island of Hokkaido in Figure 3.1 below.

Figure 3.1 here

Here the enclosed gray areas in the figure again correspond to the Hokkaido clusters for this industry in Figure 2.2(b). In addition we have now included those *pb*-choice cities for Hokkaido *that do not coincide with clusters* as enclosed dotted areas (the hatched areas can be ignored for the moment). Notice again from a comparison of Figures 2.2(a) and 3.1 that the major concentrations of livestock production include the largest city, Sapporo, together with the cities of Asahikawa, Tomakomai, Obihiro and Hakodate. Moreover, it is also clear (from the gray areas in Figure 3.1) that these concentrations have all been identified as significant "livestock products" clusters. But while there are some *pb*-choice cities near the edges of these clusters, there are also others which are far away from these major concentrations. For example, there is evidently a small number of "livestock products" establishments in the northern tip of Hokkaido around the city of Wakkanai, and also in the eastern tip of Hokkaido around Nemuro. But relative to the concentrations above, these are clearly "outlier" areas. A less clear example is provided by the ring of four small cities around Lake Saroma. But since there are not sufficiently many establishments here to constitute even a small cluster, the significance of this grouping is nonetheless questionable.

3.1.2 CB-Choice Cities

In view of these shortcomings of *pb*-choice cities, the main objective of this paper is to strengthen this concept in a way that does indeed reflect the essential spatial structure of each industry. In particular, we focus on those cities that share at least part of a significant cluster for that industry. To do so, observe first that cities are by definition collections of basic regions (municipalities) in R, so that each city, $U \in \mathcal{U}$, is formally a subset, $U \subset R$. Hence if the cluster scheme identified for each industry $i \in I$ is now denoted by $\mathbf{C}_i = (R_{i0}, C_{i1}, .., C_{ik_{\mathbf{C}_i}})$, then it would seem appropriate to focus on those cities, U, that share at least one basic region with some cluster in $\mathbf{C}_i$, i.e., which satisfy

$$U \cap C_{ij} \neq \varnothing \tag{3.2}$$

for some $j = 1, .., k_{\mathbf{C}_i}$. However, recall that our construction of clusters in terms of convex solidification will often include "empty spaces", i.e., basic regions with no establishments in the given industry. This can be illustrated by the schematic cluster constructed in Figure 2.1(d) above. This cluster is reproduced in Figure 3.2 below, where two specific cities, U_1 and U_2, have also been added, where each consists of five basic regions (shown as hatched, with the central region partially hidden by the city label).

Figure 3.2 here

Here both cities are seen to intersect this cluster. But while the black regions in Figure 2.1 were assumed to contain industry establishments, it may well be that the gray regions do not. In particular the gray region shared with city U_1 may in fact contains no establishments of this industry whatsoever. While this will usually not be the case, condition (3.2) formally allows this possibility.[28] Hence to ensure that the desired industry-choice cities actually share establishments with the given industry cluster, it is appropriate to strengthen condition (3.2) as follows. If n_{ir} denotes the number of i-establishments in region $r \in R$, and if for each cluster, $C_{ij} \in \mathbf{C}_i$ we now let

$$C_{ij}^{+} = \{r \in C_{ij} : n_{ir} > 0\} \tag{3.3}$$

denote the set of *i-employment regions* in cluster C_{ij}, i.e., basic regions with at least one i-establishment, then we now designate a city, $U \in \mathcal{U}$, as a *cluster-based* (*cb*) *choice city* for industry i iff

$$U \cap C_{ij}^{+} \neq \varnothing \tag{3.4}$$

for some $C_{ij} \in \mathbf{C}_i$, i.e., if and only if U shares an i-employment region with some cluster in $\mathbf{C}_i$.[29] In addition, if we let

$$\mathcal{U}_i = \{U \in \mathcal{U} : U \cap C_{ij}^{+} \neq \varnothing \text{ for some } C_{ij} \in \mathbf{C}_i\} \tag{3.5}$$

denote the set of *cb*-choice cities for industry i, then by definition we must have $\mathcal{U}_i \subseteq \mathcal{U}_i^{+}$, so that *cb*-choice cities are seen to be a formal strengthening of *pb*-choice cities.

This stronger definition can be illustrated schematically by city U_2 in Figure 3.2, which is in fact centered on one of the original (crossroad) regions of establishment concentrations. Hence U_2 constitutes an integral part of this cluster, and is clearly a *cb*-choice city for the industry. Empirical examples of *cb*-choice cities for the "livestock products" industry are provided by the five Hokkaido cities mentioned above. The boundaries of these cities are denoted by the enclosed hatched areas in Figure 3.2, and in all cases actually contain at least one significant "livestock products" cluster.

A comparison of the numbers of *cb*-choice cities versus *pb*-choice cities for each of the 154 industries in I is shown in Figure 3.3 below.

Figure 3.3 here

[28] For the case of Japan, where the overall density of industry establishments is very high, there were actually no such cities with respect to the cluster schemes constructed in Mori and Smith [15]. But since the present framework is intended for general use, it is important to exclude such cities explicitly [as in condition (3.4) below].

[29] Here it should be noted that this definition differs slightly from that in Mori and Smith [14] where cities were required to satisfy condition (3.2) and to have a positive employment share. In the present paper this is strengthened to require that the intersection in condition (3.2) itself have a positive employment share. These two definitions are equivalent in the case of our present Japanese data, but are not so in general.

Notice in particular that for industries with smaller numbers of *cb*-choice cities, many are on the 45-degree line. For these industries (42 in number) every *pb*-choice city is also a *cb*-choice city. So the latter concept is seen to be more important for more ubiquitous industries.

Finally we note that these numbers of *cb*-choice cities for industries have spatial consequences, and in particular, reflect the spatial diversity of their location patterns. Hence for each industry, $i \in I$, we now designate this number[30] as the (cluster-based) *locational diversity*[31]

$$d_i = |\mathcal{U}_i| \tag{3.6}$$

of industry i with respect to city system, $\mathcal{U}$. A more general definition with respect to arbitrary locational patterns of industries is given in expression (4.7) below.

3.1.3 Relative Industrial Concentration

Next recall that the primary motivation for introducing *cb*-choice cities was to capture the notion of substantial industry presence in a city. Hence it is important to ask whether industries are indeed more concentrated in *cb*-choice cities than in *pb*-choice cities. Concentration can of course be defined either in terms of establishment numbers or total employment. But as we shall see for the Japan data, industries exhibit higher concentrations in their *cb*-choice cities than *pb*-choice cities regardless of how concentration is defined.

If we first let the *employment* of industry i in city U be denoted by e_{iU}, then we may define the *employment-concentration ratio*, R_i^{emp}, of average i-employment in *cb*-choice cities ($\mathcal{U}_i$) relative to all other *pb*-choice cities ($\mathcal{U}_i^+ - \mathcal{U}_i$) by:

$$R_i^{emp} \equiv \frac{\frac{1}{|\mathcal{U}_i|}\sum_{U\in\mathcal{U}_i} e_{iU}}{\frac{1}{|\mathcal{U}_i^+|-|\mathcal{U}_i|}\sum_{U\in\mathcal{U}_i^+-\mathcal{U}_i} e_{iU}} \quad , \quad i \in I_+ \tag{3.7}$$

where $I_+ = \{i \in I : \left|\mathcal{U}_i^+\right| > |\mathcal{U}_i|\}$. As pointed out in the discussion of Figure 3.3 above, $\mathcal{U}_i^+ = \mathcal{U}_i$ for 42 of the 154 industries with significant clustering. Hence in the present case, this set I_+ consists of the remaining 112 industries for which the employment-concentration ratio is meaningful. For these industries, the values of this ratio range from 2.37 to 120.97 (with an average value of 16.13). In particular, since all values are above one, this shows that *all* industries in I_+ are relatively more concentrated in their *cb*-choice cities than in their other *pb*-choice cities.[32] The full histogram of such values is displayed in Figure 3.4(a) below, where the vertical dashed line denote the critical *unit ratio* value.

Figure 3.4 here

[30] We shall denote the *cardinality* of each set A by $|A|$.

[31] This essentially replaces the term "degree of localization" used in Mori et al. [13] for numbers of *pb*-choice cities. Our present terminology is designed to reflect the parallel between *locational diversity* of industries and *industrial diversity* of cites, as seen more clearly in expressions (4.7) and (4.8) below.

[32] Here it should be noted that similar ratios are calculated in Mori and Smith [14]. However, the set of industries used for that analysis were required to be compatible across two time periods (1981 and 2001), and hence are somewhat different.

In a similar manner, recalling that n_{iU} denotes the number of i-establishement in U, one can define the corresponding *establishment-concentration ratio*, R_i^{est}, by:

$$R_i^{est} \equiv \frac{\frac{1}{|\mathcal{U}_i|}\sum_{U\in\mathcal{U}_i} n_{iU}}{\frac{1}{|\mathcal{U}_i^+|-|\mathcal{U}_i|}\sum_{U\in\mathcal{U}_i^+-\mathcal{U}_i} n_{iU}} \quad , \quad i \in I_+ \tag{3.8}$$

where I_+ has the same meaning as above. For the 112 industries in I_+, these values range from 2.52 to 71.74 (with an average value of 15.05), and hence are again all above unity, as shown in Figure 3.4(b).

So regardless of how industry concentration is measured, it should be clear that the restriction to *cb*-choice cities versus *pb*-choice cities does indeed capture "substantial industry presence" in a structural manner, without imposing ad hoc conditions such as industry-share thresholds.

3.2 CB-Choice Industries

As a parallel to *cb*-choice cities, $\mathcal{U}_i$, for each industry, $i \in I$, one can also identify for each city, $U \in \mathcal{U}$, the set of industries in I for which U is a *cb*-choice city. More formally, it is natural to designate each industry in the set

$$I_U = \{i \in I : U \in \mathcal{U}_i\} \tag{3.9}$$

as a *cluster-based* (*cb*) *choice industry* for city $U \in \mathcal{U}$. Similarly, as a parallel to *pb*-choice cities, we may designate each industry in

$$I_U^+ = \{i \in I : n_{iU} > 0\} \tag{3.10}$$

as a *presence-based* (*pb*) *choice industry* for city $U \in \mathcal{U}$.

In a manner similar to Figure 3.3 above, the numbers of *cb*-choice industries and *pb*-choice industries are plotted in Figure 3.5 below for each of the 258 cities in $\mathcal{U}$.

Figure 3.5 here

Notice that in contrast to Figure 3.3, all cities have more *pb*-choice industries than *cb*-choice industries, except for a few at the very highest end. But since this high end is seen to involve nearly all 154 industries, these numbers are necessarily almost the same. In the three largest cities (Tokyo, Osaka and Nagoya) they are in fact identical.

While this alternative "slice" through the data is of course closely related to *cb*-choice cities, the emphasis here is slightly different. For example, the notion of *locational diversity* for industries in Section 3.1.2 above now has a clear parallel with respect to cities. In particular, the number of *cb*-choice industries for each city is a clear reflection of its industrial diversity. Hence, as a parallel to expression (3.6) above, we now designate the number of *cb*-choice industries for each city, $U \in \mathcal{U}$, as its (cluster-based) *industrial diversity*,

$$d_U = |I_U| \tag{3.11}$$

with respect to the family of industries in I. A more general definition in terms of arbitrary spatial patterns of industries is given in expression (4.8). This concept will play a central role in our analysis of the Hierarchy Principle in Section 4.4 below.

In addition to this parallel between diversity measures, we can now construct concentration ratios for cities paralleling those of industries in expressions (3.7) and (3.8) above. To do so, it is important to note that while the employment levels, e_{iU}, and establishment numbers, n_{iU}, for a given industry i are directly comparable across cities, they are not comparable across industries for a given city U. In particular, these values are only meaningful relative to the *size* of each industry. Hence to develop comparable concentration ratios for cities, it seems more appropriate to use shares rather than counts. Hence. if we now let e_i denote the *total employment* in each industry $i \in I$, so that its *employment share* in city U is given by e_{iU}/e_i, then an *employment-concentration ratio*, R_U^{emp}, for city U paralleling R_i^{emp}above can be defined as,

$$R_U^{emp} \equiv \frac{\frac{1}{|I_U|}\sum_{U\in\mathcal{U}_i}(e_{iU}/e_i)}{\frac{1}{|I_U^+|-|I_U|}\sum_{U\in\mathcal{U}_i^+-\mathcal{U}_i}(e_{iU}/e_i)} \quad , \quad U \in \mathcal{U}_+ \tag{3.12}$$

where $\mathcal{U}_+ = \{U \in \mathcal{U} : \left|I_U^+\right| > |I_U|\}$. As mentioned in the discussion of Figure 3.5 above, $I_U^+ = I_U$ for the three largest cities in Japan. Hence for our present data, $\mathcal{U}_+$ consists of the remaining 255 cities for which this employment-concentration ratio is meaningful. For these industries, the values of this ratio range from 0.35 to 37.58 (with an average value of 6.51). The full histogram of values is given in Figure 3.6(a) below.

Figure 3.6 here

In particular, there are six (out of 255) cities for which this value is *less than one*, as reflected by the position of the unit-ratio line in this figure. These few outliers are small cities with clusters mainly in ubiquitous industries. Since employment in such industries tends to be proportional to population, the industrial employment shares in these towns is very small.

Turning finally to establishment concentrations for cities, if we now let n_i denote the total number of *establishments* in industry i, so that its *establisment share* in each city U is given by n_{iU}/n_i, then an *establishment-concentration ratio*, R_U^{est}, for city U paralleling R_U^{emp}above can be defined as,

$$R_U^{est} \equiv \frac{\frac{1}{|I_U|}\sum_{U\in\mathcal{U}_i}(n_{iU}/n_i)}{\frac{1}{|I_U^+|-|I_U|}\sum_{U\in\mathcal{U}_i^+-\mathcal{U}_i}(n_{iU}/n_i)} \quad , \quad U \in \mathcal{U}_+ \tag{3.13}$$

Here the range of R_U^{est} is from 1.06 to 65.36 (with a mean of 5.10). Hence, in contrast to R_U^{emp}, this ratio is everywhere *above one*, as shown by the position of the unit-ratio line in Figure 3.6(b). In particular, the six outliers for employment concentration above now all have establishment-concentration ratios above one. Here it is of interest to note that if our cluster-detection procedure were based on employment densities (rather than establishment densities), then these six cities would be likely to exhibit no significant clustering at all.

4 Hierarchy Principle

The central purpose of this section is to reformulate the Hierarchy Principle of Christaller [2] in terms of our present notion of industrial diversity, and to develop a test of this Principle. Recall that the original version of the Hierarchy Principle asserted that industries found in a city with a given population should also be found in all cities with populations at least as large. In Mori at al.[13] it was argued that rather than population, a more appropriate measure of "city size" would be to use levels of *industrial diversity*. The notion of industrial diversity used there was defined in terms of *pb*-choice industries for cities. With respect to our present notation, this (*presence-based*) *Hierarchy Principle* asserted formally that for any cities, $U, V \in \mathcal{U}$ with $\left|I_U^+\right| \leq \left|I_V^+\right|$ and any industry, $i \in I$, if $i \in I_U^+$ then $i \in I_V^+$. Hence our main objective is to replace this definition with industrial diversity based on *cb*-choice industries for cities. Again in terms of our present notation, this amount to replacing the set of *pb*-choice industries, I_U^+, for each city U with the corresponding set of *cb*-choice industries, I_U ($\subseteq I_U^+$). More formally, this (*cluster-based*) *Hierarchy Principle* now asserts that for any cities, $U, V \in \mathcal{U}$ and industry, $i \in I$,

$$(i \in I_U) \ \& \ (|I_U| \leq |I_V|) \ \Rightarrow \ i \in I_V \tag{4.1}$$

As in Mori at al.[13], it should be emphasized that while this modification has certain advantages, both in terms of interpretation and testing, it is nonetheless very similar in the spirit to the original Hierarchy Principle. In particular, the rankings of Japanese cities in terms of their populations and cluster-based industrial diversities are quite similar [with a (highly significant) Spearman's rank correlation of 0.742].

To do so, we begin in Section 4.1 by reformulating both *industrial diversity* and *locational diversity* [expressions (3.6) and (3.11) above] within a common framework that is more useful for testing purposes. This will yield tests of these two diversity concepts in Sections 4.2 and 4.3, respectively. The parallel test of the Hierarchy Principle is then developed in Section 4.4. Finally the relation between this Principle and the notion of "specialized cities" popularized by Henderson [4] is developed in Section 4.5.

4.1 Industrial and Locational Diversity

To develop a common framework for industrial and locational diversity, it is convenient to begin by defining a family of indicator functions, $x_{iU} : I \times \mathcal{U} \ \rightarrow \{0, 1\}$, for each industry, $i \in I$, and city, $U \in \mathcal{U}$, as follows[33]

[33] Here it should be noted that the following framwork is closely related to that in Mori et al. [13] (starting on p.185). The key difference is with respect to these indicator functions. In Mori et al. [13] the set R consisted not of a partition of basic regions, but rather a set of municipalities corresponding to Metropolitan Employments Areas (MEAs) [as mentioned in footnote 20 above]. In the present paper we distinguish between basic regions (used for cluster identification) and cities, $U \in \mathcal{U}$, here defined to be Urban Employment Areas (UEAs), as in Section 2.5.2 above. More importantly, the notion of *pb*-choice cities used to define indicator functions in Mori et al. [13] is here replaced by*cb*-choice cities. Hence to avoid confusion, it is convenient to restate this formal framework explicitly in terms of the present definitions.

$$x_{iU} = \begin{cases} 1 & , \ U \in \mathcal{U}_i \\ 0 & , \ \text{otherwise} \end{cases} \tag{4.2}$$

The resulting vector of indicator values,

$$x = (x_{iU} : i \in I, U \in \mathcal{U}) \in \{0,1\}^{I \times \mathcal{U}} \equiv \mathbf{X} \tag{4.3}$$

then constitutes an *industrial location pattern* identifying both the *cb*-choice cities for each industry, $i \in I$, and the *cb*-choice industries for each city, $U \in \mathcal{U}$. In particular, for each location pattern, $x \in \mathbf{X}$, we now denote the set of *cb*-choice cities for industry i in x by,

$$\mathcal{U}_i(x) = \{U \in \mathcal{U} : x_{iU} = 1\} \tag{4.4}$$

and, similarly, denote the set of *pb*-choice industries for city U in x by

$$I_U(x) = \{i \in I : x_{iU} = 1\} \tag{4.5}$$

If the given set of industrial location data is now represented by the *observed industrial location pattern*,

$$x^0 = (x^0_{iU} : i \in I, U \in \mathcal{U}) \tag{4.6}$$

then expressions (3.5) and (3.9) above are related to the present framework by $\mathcal{U}_i \equiv \mathcal{U}_i(x^0)$ and $I_U \equiv I_U(x^0)$, respectively.

Within this more general setting, the *locational diversity* of industry, $i \in I$, in each location pattern, $x \in \mathbf{X}$, is now defined by

$$d_i(x) = \sum_{U \in \mathcal{U}} x_{iU} \ = |\mathcal{U}_i(x)| \tag{4.7}$$

The associated vector, $d_I(x) = [d_i(x) : i \in I\ ]$, then summarizes the *locational diversity structure* for all industries with respect to x. Similarly, the *industrial diversity* of each city, $U \in \mathcal{U}$, in pattern x is defined by

$$d_U(x) = \sum_{i \in I} x_{iU} = |I_U(x)| \tag{4.8}$$

with associated vector, $d_{\mathcal{U}}(x) = [d_U(x) : U \in \mathcal{U}\ ]$, summarizing the *industrial diversity structure* for all cities with respect to x.

In particular, the *observed locational diversity structure* of industries is given by $d^0_I = (d^0_i : i \in I)$, where:

$$d^0_i = d_i(x^0) = \sum_{U \in \mathcal{U}} x^0_{iU} \ , \quad i \in I \tag{4.9}$$

Similarly, the *observed industrial diversity structure* of cities is given by $d^0_{\mathcal{U}} = (d^0_U : U \in \mathcal{U})$, where:

$$d^0_U = d_U(x^0) = \sum_{i \in I} x^0_{iU} \ , \quad U \in \mathcal{U} \tag{4.10}$$

Finally, it should be noted that expressions (3.6) and (3.11) in Section 3.2 above are related to the present definitions by $|\mathcal{U}_i| = d_i \equiv d_i^0$ and $|I_U| = d_U \equiv d_U^0$, respectively.

4.2 A Test of Industrial Diversity

Before proceeding to the Hierarchy Principle itself, we begin by noting that the above concepts of industrial and locational diversity structures are of interest in their own right. In the present section, we consider the industrial diversity of cities in more detail, and develop a test for the presence of significant diversity. A parallel analysis of the locational diversity of industries is developed in Section 4.3 below. Following Mori et al.[13], we start by taking the observed structure of locational diversity among industries as given, and identify the set of all *industrial location patterns* consistent with this data. More precisely, for any given *observed locational diversity structure*, $d_I^0 = (d_i^0 : i \in I)$, the set of *feasible location patterns*, $\mathbf{X}_I^0$, consistent with d_I^0 is given by,

$$\mathbf{X}_I^0 = \left\{ x = (x_{iU} : i \in I, U \in \mathcal{U}) : \sum_{U \in \mathcal{U}} x_{iU} = d_i^0, i \in I \right\} \subset \mathbf{X} \tag{4.11}$$

By restricting industrial location patterns to those consistent with d_I^0, one is preserving as much of the actual locational diversity structure as possible. For example, ubiquitous industries with high levels of locational diversity will continue to be ubiquitous in all location patterns, $x \in \mathbf{X}_I^0$.

In this context, one may then ask what the industrial diversity structure for cities would look like if for these given levels of locational diversity for industries, the locational pattern of industries was otherwise random. This may be formalized by treating location patterns, $x = (x_{iU} : i \in I, U \in \mathcal{U})$, as possible realizations of a random vector, $X = (X_{iU} : i \in I,\ U \in \mathcal{U})$, and considering the null hypothesis:

$$H_I^0 : X \text{ is } \textit{uniformly distributed} \text{ on } \mathbf{X}_I^0 \tag{4.12}$$

In particular, to test whether the observed industrial diversity structure, $d_{\mathcal{U}}^0$, in (4.10) is more heterogenous than would be expected under H_I^0, one may construct some appropriate statistic, say $S(x)$, reflecting the heterogeneity of industrial diversities among cities and ask whether the observed value, $S(x^0)$, is higher (more heterogeneous) than would be expected under H_I^0. One simple choice for $S(x)$ here is given by the *range*, $\Delta d_{\mathcal{U}}(x)$, of industrial diversity levels in $d_{\mathcal{U}}(x)$, as defined for each $x \in \mathbf{X}_I^0$ by

$$\Delta d_{\mathcal{U}}(x) \equiv \max_{U,V \in \mathcal{U}} |d_U(x) - d_V(x)| \tag{4.13}$$

Given this specification, the desired test can be carried out by simply generating a set of Monte Carlo samples $(x_s : s = 1, .., N)$ of X, and calculating the fraction of simulated range values, $[\Delta d_{\mathcal{U}}(x_s); s = 1, .., N]$, that are at least as large as the observed value, $\Delta d_{\mathcal{U}}(x^0)$. In the present case, such calculations are in fact unnecessary since the observed value is literally "off the chart", as shown by the vertical dashed line to the right of the histogram of simulated range values with $N = 1000$ in Figure 4.1(a) below.

Figure 4.1 here

Here the observed value, $\Delta d_{\mathcal{U}}(x^0) = 153$, is vastly higher than the maximum simulated value of $\Delta d_{\mathcal{U}}(x) = 43$. Note that since there are only 154 industries in I, the observed range is almost as large as possible (with an industrial diversity of 154 for Tokyo and an industrial diversity of 1 for the two cities in $\mathcal{U}$ with smallest populations, namely Ashibetsu and Kucchan[34]). Hence it should be clear that even for simulated samples much larger than $N = 1000$, the same results would obtain. So with respect to this range measure, the observed pattern of industrial diverstiy in Japan is vastly larger than what would be expected under randomness.

One alternative to the range would be to focus simply on the largest industrial diversity among cities, namely to replace the range of values in $d_{\mathcal{U}}(x)$ with the maximum value:

$$d_{\mathcal{U}}^{\max}(x) \equiv \max_{U \in \mathcal{U}} d_U(x) \tag{4.14}$$

Exactly the same testing procedure with respect to this statistic (and $N = 1000$) yields the results shown in Figure 4.1(b). Here (as mentioned above) the highest observed value of 154 corresponds to Tokyo, while the highest simulated maximum value is only 89. So these results again confirm the dramatic departure of the observed structure, $d_{\mathcal{U}}(x^0)$, of industrial diversity versus those simulated under the randomness hypothesis in (4.12).

To interpret these results, note that heterogeneity of industrial diversity suggests that many cities tend to exhibit higher levels of industrial diversity than would be expected under randomness. But since the number of cb-choice cities for each industry is being held constant (by the construction of $\mathbf{X}_I^0$) this in turn implies more of these locational choices are coincident with other industries than would be expected. Hence these results suggest that there is significant *spatial coordination* of agglomerations across industries, as implied by the work of Christaller [2] (together with more recent formalizations of this work by Fujita et al. [3], Tabuchi and Thisse [20, 21] and Hsu [6]). Indeed, the test of Christaller's Hierarchy Principle developed in Section 4.4 below will provide an even more direct test of this spatial coordination among industries.

4.3 A Test of Locational Diversity

In a manner completely paralleling the procedure in Section 4.2 above, one may also test for the presence of significant locational diversity among industries given the observed level of industrial diversity among cities, as summarized by the observed industrial diversity structure, $d_{\mathcal{U}}^0 = (d_U^0 : U \in \mathcal{U})$, defined by (4.10). Here we simply sketch the main elements of this test. First, let the set of *feasible location patterns* consistent with $d_{\mathcal{U}}^0$ be denoted by,

$$\mathbf{X}_{\mathcal{U}}^0 = \left\{ x = (x_{iU} : i \in I,\ U \in \mathcal{U}) : \sum_{i \in I} x_{iU} = d_U^0\ ,\ U \in \mathcal{U} \right\} \subset \mathbf{X} \tag{4.15}$$

[34] The singly *cb*_choice industry for Ashibetsu (population = 21,026) is "newspaper industries" (JSIC191) and that for Kucchan (population = 19,689) is "sugar processing" (JSIC125). Note also that the number of *cb*-choice cities for "newspaper industries" and "suger processing" are 153 and 49, respectively. The former is a typical ubiquitous industry which is found in most cities, while the latter is relatively localized industry. Thus, Kucchan can be considered as a typical instance of a "specialized-industry" town.

Next, as a parallel to (4.12) above, consider the null hypothesis,

$$H^0_{\mathcal{U}} : X \text{ is } \textit{uniformly distributed} \text{ on } \mathbf{X}^0_{\mathcal{U}} \tag{4.16}$$

that except for consistency with $d^0_{\mathcal{U}}$, industrial location patterns are otherwise random. Here, the restriction to industrial diversity patterns consistent with $d^0_{\mathcal{U}}$ ensures the preservation of as much of the actual city structure as possible. For example, Tokyo will continue to be a *cb*-choice city for every industry, and all smaller cities will continue to have the same number of *cb*-choice industries as observed in actuality. To measure the heterogeneity of locational diversity levels among industries, we shall here only consider the *range* of such diversity levels, as defined for each locational pattern, $x \in \mathbf{X}^0_U$, by:

$$\Delta d_I(x) \equiv \max_{i,j \in I} |d_i(x) - d_j(x)| \tag{4.17}$$

In these terms, we now wish to test whether the range of observed locational diversity levels, $\Delta d_I(x^0)$, is significantly larger than would be expected under $H^0_{\mathcal{U}}$. The results of a Monte Carlo test (again with $N = 1000$ simulated samples of $\Delta d_I(x)$ under $H^0_{\mathcal{U}}$) are shown in Figure 4.2 below:

Figure 4.2 here

Here the results are in some ways even more dramatic than those in Figure 4.1(a) above. Out of the 258 possible cities in $\mathcal{U}$, the observed range is 212 while the maximum range of the 1000 random location patterns simulated is only 53. Here the most ubiquitous industry (with 224 *cb*-choice cities out of 258) happens to be the industry manufacturing "printing plates" (JSIC194). More generally, printing-related activities often require direct interaction with customers, and are very market oriented. At the other extreme, the most localized industries (each with only 12 *cb*-choice cities) are the "leather glove and mittens" industry (JSIC245) and the "briquettes and briquette balls" industry (JSIC214). The former is an example of a highly specialized industry that is concentrated almost entirely in a group of three small villages accounting for over 90% of the national market share (see Section 4.5 below for further discussion of this industry).[35] The latter is a good example of a resource-oriented ("first-nature") industry with establishments located primarily in the vicinity of briquette mines. Given the locations of such mines in Japan, this industry turns out to be highly localized as well.

Two final points here relate to the interpretation of these results. First, it should be clear that industries with high locational diversity must by definition have many establishments, and correspondingly large levels of employment. Hence it can be argued that such test results essentially reflect a diversity in the *size* of industries. Moreover, from an economic viewpoint, such results in part reflect underlying variations in *scale economies* among industries [as analyzed for example in the city-system models of Fujita, Krugman and Mori [3] and Hsu [6]].

[35] More generally, it is of interest to note that most leather/fur-related industries tend to be similarly specialized with small locational diversities. In fact, five of the ten industries with smallest locational diversities in Japan are in this category.

4.4 A Test of the Hierarchy Principle

Given these initial results, we now turn to the Hierarchy Principle itself. In a manner similar to the diversity measures above, it is convenient to restate this Principle in terms of industrial location patterns. As an extension of the definition in (4.1), we now say that an industrial location pattern, $x = (x_{iU} : i \in I, U \in \mathcal{U}) \in \mathbf{X}$, satisfies the (*cluster-based*) *Hierarchy Principle* if and only if for each pair of cities, $U, V \in \mathcal{U}$ and industry, $i \in I$,

$$[\ i \in I_U(x)\ \&\ d_U(x) \leq d_V(x)\ ]\ \Rightarrow\ i \in I_V(x) \tag{4.18}$$

To test this Principle, we follow the basic approach developed in Mori et al.[13]. In particular, we start by representing the observed industrial location pattern, $x^0 = (x^0_{iU} : i \in I, U \in \mathcal{U})$, as in Figure 4.3 below:

Figure 4.3 here

Here cities, $U \in \mathcal{U}$, are ordered on the horizontal axis from lowest to highest in terms of their observed industrial diversities, d^0_U. Similarly, industries, $i \in I$, are ordered in terms of their observed locational diversities, d^0_i. With respect to this coordinate system, a "plus" symbol (+) in position (U, i) indicates that U is a *cb*-choice city for industry i (and equivalently, that i is a *cb*-choice industry for city U). If we distinguish such positions as *positive*, then the Hierarchy Principle asserts that for each positive position (U, i) there must also be a (+) in every row position $(\cdot, i)$ to the right of (U, i), indicating that all cities with industrial diversities greater than or equal to city U are also *cb*-choice cities for industry i. It is evident from the figure that while the Hierarchy Principle does not hold perfectly, the row density of (+) values increases from left to right in virtually every row.[36] Hence this data is seen to exhibit a strong level of agreement with the Hierarchy Principle that could not have occurred by chance.[37]

In this context, one may regard each occurrence of a full row of (+) values to the right of a positive position (U, i) as a "full hierarchy event" in the sense that it is fully consistent with the Hierarchy Principle. However, if only small fraction of (+) values are missing, then it is natural to consider such cases as being "closer" to a full hierarchy event than if all (+) values were missing. To formalize these ideas for arbitrary industrial location patterns, x, we first observe that such hierarchy events are only meaningful for the *positive* positions in x (i.e., the pairs, iU, for which U is a *cb*-choice city for industry i in x). Hence if for each industrial location pattern, $x \in \mathbf{X}$, we now denote this set of *positive pairs* by

$$\mathcal{P}_x = \{iU \in I \times \mathcal{U} : x_{iU} = 1\}\quad ,\quad x \in \mathbf{X} \tag{4.19}$$

[36] It should also be noted that the SIC classification system for industries is by no means exact. Hence some level of disagreement in such hierarchical relations is unavoidable.

[37] Note that this figure bares a strong resemblance to Figure 7 in Mori et al. [13], as well as Figure 9 in Mori and Smith [14] The key difference from Mori et al. [13] is our present definition of *cb*-choice cities versus *pb*-choice cities. In addition, a larger set of cities is used here (as described in Section 2.5.2 above). The difference from Mori and Smith [14] is mainly in term of industries. In that paper, industries were required to be consistently defined over a twenty-year span, thus resulting in a smaller set of 139 industries. But in spite of these differences, the resulting figures are seen to be qualitatively very similar.

and for each city, $U \in \mathcal{U}$, let

$$S_U(x) = \{V \in \mathcal{U} : d_V(x) \geq d_U(x)\} \tag{4.20}$$

denote the set of cities with industrial diversities in x at least as large as that of U, then the desired *fractional hierarchy event* for each positive pair, $iU \in \mathcal{P}_x$, is defined to be[38]

$$H_{iU}(x) = \frac{1}{|S_U(x)|} \sum_{V \in S_U(x)} x_{iV} \tag{4.21}$$

By definition, $0 < H_{iU}(x) \leq 1$,[39] with the extreme case, $H_{iU}(x) = 1$, constituting a *full heirarchy event* at $iU \in \mathcal{P}_x$.

In these terms, a simple summary measure of the overall consistency of pattern, $x \in \mathbf{X}$, with the Hierarchy Principle is given by the mean of these fractional hierarchy events, which we now designate as the *hierarchy share*

$$H(x) = \frac{1}{|\mathcal{P}_x|} \sum_{iU \in \mathcal{P}_x} H_{iU}(x) \tag{4.22}$$

for pattern x. As a parallel to the underlying fractional hierarchy events, these hierarchy shares must also satisfy $0 < H(x) \leq 1$.[40] Moreover, the full equality condition, $H(x) = 1$, implies that all fractional hierarchy events must be *full*, and hence from (4.18) that x must *satisfy the Hierarchy Principle*. Thus, these hierarchy shares are seen to provide a natural test statistic for the Hierarchy Principle itself.

In this context, it was argued in Mori et al.[13] that the most appropriate null hypothesis for testing this Principle is precisely $H^0_{\mathcal{U}}$ in (4.16) above, namely that except for consistency with the given industrial diversity structure, $d^0_{\mathcal{U}} = (d^0_U : U \in \mathcal{U})$, industrial locations are otherwise random. The advantage of this approach is that it allows industrial location patterns to be "as random as possible" while maintaining the underlying city structure in terms of industrial diversity. So, for example, major cities like Tokyo and Osaka will continue to have high levels of industrial diversity under $H^0_{\mathcal{U}}$.[41]

Given this null hypothesis, our test of the Hierarchy Principle is thus very similar to that in Section 4.3 above. In particular, the observed industrial location pattern, x^0, is again hypothesized to be a typical realization of a uniform random variable, X, on the set of feasible patterns, $\mathbf{X}^0_{\mathcal{U}}$ in (4.15). The only difference here is that the relevant test statistic is now taken to be the *random hierarchy share variable*, $H(X)$. Hence under $H^0_{\mathcal{U}}$, the *observed hierarchy share*, $H(x^0)$ [based on the data represented in Figure 4.3] should be a typical realization of $H(X)$. To test this, we again simulate $N = 1000$ draws $\{x_s : s = 1, .., N\}$ from $\mathbf{X}^0_{\mathcal{U}}$ and calculate their

[38] Note that $U \in S_U(x) \Rightarrow |S_U(x)| > 0$ for all U.

[39] Note that $iU \in \mathcal{P}_x$ implies $x_{iU} = 1$, so that $H_{iU}(x) \geq 1/|S_U(x)| > 0$.

[40] Note that from a technical viewpoint, H_x is not defined for *null pattern*, $x^{null} \in \mathbf{X}$, with $x^{null}_{iU} = 0$ for all iU. Indeed, the Hierarchy Principle is satisfied *vacuously* for this pattern since $\mathcal{P}_{x^{null}} = \varnothing$. Hence, for convenience, we simply ignore this degenerate case in all subsequent analyses.

[41] Recall from the introductory discussion to Section 4 that these levels of industrial diversity are indeed highly correlated with their city sizes.

associated *hierarchy shares*, $\{H(x_s) : s = 1, .., N\}$. Using this simulated data, one may estimate [as in Mori et al. [13]] the cumulative frequency distribution, $F(h) = \Pr(H < h)$, of H under $H^0_{\mathcal{U}}$ by

$$\widehat{F}(h) = \frac{1}{N}|\{s : H(x_s) < h\}| \tag{4.23}$$

and hence estimate the associated *p-value* for a one-sided test of $H^0_{\mathcal{U}}$ by

$$\widehat{\Pr}\left[H \geq H(x^0)\right] = 1 - \widehat{F}[H(x^0)] \tag{4.24}$$

For example, if $H(x^0)$ were larger than 99% of the simulated $H(x_s)$ values [so that $\widehat{F}\left[H(x^0)\right] > 0.99$] then $\widehat{\Pr}\left[H \geq H(x^0)\right] < 0.01$ would imply that the (estimated) chance of observing a value as large as $H(x^0)$ under $H^0_{\mathcal{U}}$ is less than 0.01, and thus that this null hypothesis could be rejected at the 0.01 level.[42]

In fact, the evidence against $H^0_{\mathcal{U}}$ is far stronger than this, as can be seen in Figure 4.4 below. Here the realized values are plotted (in a manner similar to Figure 4.2) as a histogram, with the observed value, $H(x^0) = 0.771$, again represented by a vertical dashed line. As in Figure 4.2, this value is again well above the range of simulated values $[0.634, 0.636]$, and here provides strong evidence for the Hierarchy Principle.

Figure 4.4 here

In summary, these results serve to reconfirm the findings of Mori et al.[13] under the present more stringent definition of industrial diversity in terms of *cb*-choice cities. In particular they show that even after controlling for relative industrial diversities among cities, the location pattern of Japanese (three-digit) manufacturing industries in 2001 shows very significant hierarchical structure.

4.5 Specialization and Agglomeration

It should be noted however that in spite of its statistical significance, the observed hierarchy share, $H(x^0) = 0.771$, is still well below unity. Moreover, since $H(x^0)$ is only an average value over all industries, it should be clear that certain industries may in fact exhibit large deviations from the Hierarchy Principle. To examine this question further, we now let

$$H_i = \frac{1}{|\mathcal{U}_i|} \sum_{U \in \mathcal{U}_i} H_{iU}(x^0) \tag{4.25}$$

denote the (*observed*) *hierarchy share* for each industry $i \in I$. The histogram of these values over the 154 industries in I is shown in Figure 4.5 below.

[42] It should be noted that since $H(x^0)$ is formally postulated to be an additional sample of $H(X)$ under $H^0_{\mathcal{U}}$, one could also estimate $F(h)$ using the larger sample, $\{H(x_s) : s = 0, 1, .., N\}$, of size $N + 1$. But for large N this will make little difference in the results.

Figure 4.5 here

While the mean value, 0.697, is very close to that of the overall hierarchy share, 0.771,[43] the individual values range from 0.213 to 0.969. Of particular interest for our present purposes are those industries on the low end, that deviate quite dramatically from the Hierarchy Principle. The ten industries with smallest hierarchy shares, H_i, are listed in Table 4.1 below.

Table 4.1 here

These industries can be roughly classified into three groups. The first group of industries ["Fur skins" (JSIC248), "leather gloves and mittens" (JSIC245), "leather tanning and finishing" (JSIC241), and "ophthalmic goods, including frames" (JSIC326)] are all examples of industries that are subject to *industry-specific localization economies*. When production externalities are industry specific (such as those related to knowledge shared among workers with specialized skills), the specific locations of industrial concentrations may be largely determined by historical circumstances. For instance (as mentioned in Section 4.3 above), the "leather glove and mittens" industry is almost entirely concentrated in a cluster of three remote municipalities (Hikita, Shiratori and Ohuchi) on Shikoku island (refer to Figure 2.3). While these municipalities have a total population of only 38,000, they account for more than 90% of all leather glove manufacturing in Japan. Similarly, the "opthalmic goods, including frames" industry is highly concentrated in the small town of Sabae (population 65,000) on the northern coast of Honshu (refer to Figure 2.3). This town also accounts for more than 90% of all eye glass frames manufactured in Japan (and in fact, 20% of all eye glass manufacturing in the world). In both of these cases, there are no strong reasons other than historic why such dramatic concentrations should be found at these locations.

The second group of industries ["iron smelting, without blast furnaces" (JSIC262), "petroleum refining" (JSIC211), and "iron industries, with blast furnaces" (JSIC261)] are all subject to *large plant-level scale economies in production.* Since their production processes are relatively self-contained, these industries have little incentive to co-locate with other industries. In particular, since most of their (weight/bulk intensive) inputs are imported by sea, such industries must often compete for suitable coastal locations.

The final group of industries ["briquettes and briquette balls" (JSIC214) and "lacquer ware" (JSIC346)] are examples of *resource-oriented* ("first-nature") industries constrained by their input-supply locations. For example (as mentioned in Section 4.3 above) the "briquettes and briquette balls" industry is primarily located in the vicinity of briquette mines.

What all of these groups have in common is a high degree of specialization in some aspect of their production processes. This suggests that the degree of specialization among industries may in fact help to explain deviations from the Hierarchy Principle. To test this idea, one must

[43] These two mean values are only guaranteed to be the same when the number of choice industries, $|\mathcal{U}_i|$, is the same for each industry $i \in I$.

construct some appropriate measure of "specialization". Here it is of interest to note that while our present version of the Hierarchy Principle focuses on "substantial presence" of industries in given cities, there is no explicit consideration of their actual employment shares in these cities. So one way to measure the "degree of specialization" for industry i is to focus on its employment shares across cities, and to quantify the deviations of these shares from those of the manufacturing sector as a whole. To be more precise, we first recall from Section 3.1.3 that e_{iU} denotes the total employment of industry, $i \in I$, in city, $U \in \mathcal{U}$. With this notation, it follows that for any given industry, $i \in I$, the *within-industry employment share* of i in city U is given by

$$s_{U|i} = \frac{e_{iU}}{\sum_{V \in \mathcal{U}} e_{iV}} \tag{4.26}$$

Similarly, by letting $e_U = \sum_{i \in I} e_{iU}$ denote total manufacturing employment in city U, it follows that the corresponding *total employment share*, s_U, in city U of all manufacturing is given by

$$s_U = \frac{e_U}{\sum_{V \in \mathcal{U}} e_V} \tag{4.27}$$

In this context, it is natural to regard *equality* between these two distributions as representing the extreme case of "no specialization" for industry i. If this is formalized as a null hypothesis

$$H_0^i : (s_{U|i} = s_U : U \in \mathcal{U}) \tag{4.28}$$

for industry i, then an appropriate statistic for testing this hypothesis is the *Kullback-Leibler* (KL) *divergence* of distribution $(s_{U|i} : U \in \mathcal{U})$ from $(s_U : U \in \mathcal{U})$, as defined by (see Kullback [16]):

$$D_i = \sum_{U \in \mathcal{U}} s_{U|i} \ln \left(\frac{s_{U|i}}{s_U} \right) \tag{4.29}$$

As is well known, $D_i \geq 0$, and $D_i = 0$ if and only if H_0^i in (4.28) is satisfied. Hence larger values represent greater "deviations" from the distribution of total employment shares, which in our present context, suggests that D_i can be interpreted as the *degree of specialization* for each industry, $i \in I$.[44]

Given this measure, the above observations suggest that those industries, i, with greater deviations from the Hierarchy Principle (i.e., with lower hierarchy shares, H_i) might in fact be those with higher degrees of specialization, as measured by D_i. A plot of D_i against H_i for the 154 industries in I is given in Figure 4.6 (where the ten industries in Table 4.1 are labeled explicitly), and shows that there is indeed a strong negative relation between these values. In particular, the Spearman's rank correlation between the two is -0.850, and is of course highly significant.

Figure 4.6 here

[44] For a similar application of KL-divergence to measure the *degree of localization* of industries, see Mori et al.[12].

For completeness, the associated histogram of D_i values is given in Figure 4.7 below. As expected from the inverse relation between the two, this histogram is essentially the reverse of that for H_i in Figure 4.5.

Figure 4.7 here

Given this inverse relationship, it is of interest to observe that from a theoretical viewpoint, perhaps the most prominent competitor to the Hierarchy Principle in the economic geography literature is the "system of cities model" (first introduced by Henderson [4]) in which each city is specialized in a single industry (due to industry-specific externalities/scale economies). In this model, *cities that are more specialized in a given industry are expected to exhibit a larger presence of that industry than other cities.* More precisely, if for any given city, $U \in \mathcal{U}$, the *within-city employment share* of industry i in U is defined by

$$s_{i|U} = \frac{e_{iU}}{\sum_{j \in I} e_{jU}} = \frac{e_{iU}}{e_U} \tag{4.30}$$

then those cities U that are more specialized in industry i are expected to exhibit higher within-industry employment shares, $s_{U|i}$, than other cities.

This *specialization-concentration hypothesis* is indeed supported by our Japanese data. In particular, if for each industry $i \in I$ one calculates the Spearman's rank correlation between these *within-city* employment shares, $(s_{i|U} : U \in \mathcal{U})$, and the corresponding *within-industry* employment shares, $(s_{U|i} : U \in \mathcal{U})$, across cities, then the mean of these correlations is 0.697. Moreover, there is a strong concentration around this mean, as shown by the histogram of rank correlation values for all industries in Figure 4.8 below. Hence while these correlations are by no means perfect, they do suggest that elements of this "system of cities model" are exhibited by manufacturing industries in Japan.

Figure 4.8 here

As a possible synthesis of these ideas, we note first that our present Hierarchy Principle makes no assertion whatsoever about this specialization-concentration hypothesis. For example, consider the extreme case in which a city system, $\mathcal{U}$, satisfies the Hierarchy Principle for all industries, but that for each industry, $i \in I$, (i) all cb-choice cities, $U \in \mathcal{U}_i$, have the same within-city employment shares, $s_{i|U} \equiv s_i > 0$, and (ii) all other cities have zero i-employment.[45] Then, assuming that some industries are more specialized than others (i.e., that hypothesis H_0^i does not hold identically for all industries i), it is clear that there can be no correlation between specialization and within-industry employment shares. Hence such relationships are formally independent of the presence or absence of industrial hierarchies.

[45] Note that in this extreme case, $\mathcal{U}$ also satisfies the *presence-based* Hierarchy Principle.

In view of this independence, the inverse relationship in Figure 4.6 suggests that the structure of manufacturing in Japan *exhibits both hierarchical and specialization-concentration structure*. Moreover, these two concepts appear to be complementary in that specialization-concentration tends to be strongest in those industries where hierarchies are the weakest. This suggests that perhaps a more satisfactory theory of urban industrial structure should involve a synthesis of these two ideas.

5 NAS Rule

In addition to the Hierarchy Principle itself, it was also shown in Mori et al. [13, Theorems 1 and 2] that this Principle has consequences for both the Number-Average Size (NAS) Rule for industries and the Rank-Size Rule for cities. In particular, it was shown that in the presence of the Hierarchy Principle, these two rules are essentially equivalent. While these analytical results require that the classical (population based) Hierarchy Principle hold *exactly*, they still suggest that in the presence of a strong hierarchical industrial structure, these two rules should continue to exhibit a close relationship. In this regard, it was shown empirically in Mori et al. [13] that for the *presence-based* version of the Hierarchy Principle, both of these rules indeed exhibit strong statistical significance. For the present *cluster-based* version of this Principle, it was also shown in Mori and Smith [14] that both of these rules not only exhibit strong statistical significance, but also remarkable stablity over a twenty-year time span.

With respect to the Rank-Size Rule in particular, the regression for 2000/2001 in expression (13) of Mori and Smith [14] confirms the significance of this relation for our present set of city data.[46] However, since the NAS Rule involves both industry and city data, and since our combined industry-city data differs from both these previous papers (as discussed in Section 2.5.2 above), it is of interest to reconsider the NAS Rule within the present setting. Hence the main objective of this section is to reconfirm the NAS Rule using the *cluster-based* choice cities generated by our present sets of industries, I, and cities, $\mathcal{U}$.

To do so, we start by recalling that the NAS Rule formulated in Mori et al. [13] asserts that there is a log-linear relationship between the number and average size of *pb*-choice cities for industries. This rule was motivated by a remarkably strong log-linear regression obtained between these variables. In particular, if we let $\mathcal{U}^*$ denote the set of 113 *Metropolitan Employment Areas* (MEAs) for Japan in 2000, and let I^* denote the larger set of 261 Japanese industries in 2000 including services, wholesale, and retail, together with manufacturing,[47] then this regression was based on the *pb*-choice cities in $\mathcal{U}^*$ for all industries in I^*.[48] For these data sets, if we now denote

[46] See Mori and Smith [14, pp.197-202] for a complete discussion.

[47] The full set of such industries is 264 in number. But to maintain a parallel with the regression in expression (2) of Mori et al. [13], the three obvious outliers in Figure 1 of Mori et al. [13], namely, "coke" (JSIC213), "small arms (rifles)" (JSIC331) and "small arms ammunition (bullets)" (JSIC333), are excluded from the regression (5.1) below. Here it should be noted that these three industries are among the nine with spurious clustering, and hence are also excluded from the regressions in (5.2) and (5.3) below. Finally, it should also be noted that the "rifles" industry (JSIC331) no longer appears to be an outlier in Figure 5.1 below. This is a consequence of the additon of new establishments in this (very small) industry between the 1999 establishment-location data used in Mori et al. [13] and the 2001 establishment-location data used here.

[48] Similar results were reported for 1980 data. But for purposes of comparability with the present data, we

the *average size* of *pb*-choice cities in $\mathcal{U}^*$ for a generic industry in I^* by $\overline{SIZE}$, and similarly, denote the *number* of such cities for this industry by $\#CITY$, then the regression obtained was as follows (where standard deviations of estimates are in parentheses):[49]

$$\log(\overline{SIZE}) = \underset{(0.0097)}{17.101} - \underset{(0.0022)}{0.712}\ \log(\#CITY), \quad R^2 = 0.998 \tag{5.1}$$

As noted in that paper (and elsewhere) the usual independent-random-sampling assumptions underlying linear regression are questionable here. But the goodness-of-fit in terms of R^2 is so strong that this relation in fact appears to be almost deterministic. It was this observation that inspired the NAS Rule.

To extend this analysis to the present setting, we now employ the larger set, $\mathcal{U}$, of all 258 UEAs in Japan and the (more comparable) set, I, of 154 manufacturing industries in Japan exhibiting significant clustering. For the sake of comparability with (5.1), we again denote the *average size* of *pb*-choice cities in $\mathcal{U}$ for a generic industry in I by $\overline{SIZE}$, and similarly, denote the *number* of such cities for this industry by $\#CITY$. In these terms, the results of the new regression yield:

$$\log(\overline{SIZE}) = \underset{(0.0200)}{17.030} - \underset{(0.0042)}{0.718}\ \log(\#CITY), \quad R^2 = 0.995 \tag{5.2}$$

The similarity between (5.1) and (5.2) is apparent. Of special importance are the slope and goodness-of-fit, which are essentially the same. Hence the inclusion of all UEAs on the city side, and the restriction to clustered manufacturing on the industry side, has not altered the nature of this NAS regularity.

But as emphasized above, when industries are restricted to those exhibiting significant clustering, it is more appropriate to examine this NAS relationship in terms of *cluster-based* choice cities. Hence if this regression is re-run using the smaller set of *cb*-choice cities for each industry, and if again for the sake of comparison we denote the *average size* of *cb*-choice cities in $\mathcal{U}$ for a generic industry in I by $\overline{SIZE}$, and denote the *number* of such cities for this industry by $\#CITY$, then the results of this new regression yield:

$$\log(\overline{SIZE}) = \underset{(0.0278)}{17.011} - \underset{(0.0062)}{0.717}\ \log(\#CITY), \quad R^2 = 0.989 \tag{5.3}$$

It is the relation between (5.2) and (5.3) which is of primary interest for our present purposes. Here again it is clear that these results are almost indistinguishable. So even when all non cluster-based *pb*-choice cities are eliminated (such as those illustrated for the "livestock products" industry in Figure 3.1 above), this NAS relationship remains strong. Indeed it is our belief that this relationship among the choice cities for each industry is most meaningful when restricted to those cities exhibiting a substantial industry presence in terms of clustering.

A visual comparison of (5.2) and (5.3) can also be made by examining panels (a) and (b)

consider only the results for 2000.

[49] Note also that the intercept, 7.427, in expression (2) of Mori et al. [13] was based on a regression using logs to the base 10, whereas the present results use natural logs. This affects the intercept but not the slope. Hence the intercept (and standard error) reported here have been rescaled to natural logs [i.e., multiplied by ln(10)].

of Figure 5.1, respectively, where these regressions correspond to the solid lines in each panel. It should also be noted that the data points for the 154 industries in I are represented by the (+) symbols in both panels. The additional points shown by ($\odot$) symbols correspond to the remaining nine industries with spurious clustering (as discussed in Section 2.5.3 above). Given this distinction, notice first that the five dramatic outliers in these regressions are all among the nine industries with spurious clustering. In our view, this adds further credence to the hypothesis that industrial clustering plays a significant role in the NAS Rule itself.

Figure 5.1 here

The two dashed curves in each panel represent the upper and lower bounds for the average size of any given number of choice cities. In particular, for each number, n, the upper [resp., lower] bound of the average population size of n choice cities is given by that of the n largest [resp., smallest] cities. Recall that under original (population based) Hierarchy Principle in Section 4 above, the number of choice cities for each industry should achieve these upper bounds *exactly*. Hence, in the presence of a strong hierachical structure of industries, it is reasonable to expect that these average sizes of choice cities will be close to their upper bounds. As seen in both panels of Figure 5.1, this is indeed the case.

Notice also that the upper-bound curve is nearly log linear. It is shown by Mori et al. [13, Theorem 2] that the log linearity of this upper bound is essentially equivalent to that of the rank size distribution for a large number of cities.

Next observe that the data points for those nine industries with spurious clustering, and indeed all industries with less than about 30 choice cities, are identical in these two scatter plots. The reason for this can be seen in Figure 3.3 where these industries all appear on the 45° line, indicating that every *pb*-choice city for these industries is also a *cb*-choice city. Indeed, when the number of *pb*-choice cities for an industry is small, it is reasonable to expect that even cities with only a few of its establishments will constitute a substantial contribution to BIC (in our cluster-detection algorithm of Section 2.3 above), and hence will qualify as *cb*-choice cities. Additional evidence for this is provided by the fact that the number of clusters per establishment is strongly negatively correlated with the number of establishments across industries (Spearman's rank correlation = -0.971).

Note finally that this NAS relation appears to be the strongest among those industries with large numbers of choice cities. This suggests that there may indeed be some "threshold" level of locational diversity required for industries to exhibit this type of regulartity.[50]

6 Concluding Remarks

In this paper, we have introduced the concept of *cluster-based* choice cities for an industry as a means of identifying those cities with a substantial industry presence. This concept was in turn

[50] This is somewhat analogous to the Rank Size Rule presented in Mori and Smith [14, Figure 10], where large cities seem to exhibit special "outlier" features. Hence for case of the NAS Rule, it would appear that industries with small numbers of choice cities (either *pb* or *cb*) play a similar role.

used to develop modified forms of both the classical Hierarchy Principle of Christaller [2] and the NAS Rule of Mori et al. [13]. Finally, these modified regularities were shown to exhibit a significant presence with respect to Japanese manufacturing and city data from 2000/2001.

But this industrial agglomeration approach to central place and city-size regularities also raises a number of additional issues that are appropriate to touch on in these concluding remarks.[51]

6.1 Level of Industrial Aggregation

It should be clear that the notion of industrial clustering itself depends critically on the level of industrial aggregation employed. Indeed, for the completely disaggregated case in which each establishment constitutes a single industry category, there can be no meaningful notion of clustering at all. This is equally true for the notion of cluster-based choice cities. Even at intermediate levels of aggregation, the set of choice cities for industrial categories may change drastically. For example, recall from Figure 2.2 that at the JSIC three-digit level used in this paper, the "livestock products" industry in Japan consists of a large number of small clusters spread throughout the nation. But, it is not clear that all types of livestock (e.g., poultry, cattle, hogs) are equally represented by each cluster. In particular, some types of livestock may be confined to specific sub-regions of the nation.

These aggregation effects in turn have consequences for the validity of both the Hierarchy Principle and the NAS Rule. Indeed neither regularity is even meaningful for completely disaggregated (or completely aggregated) industries. Hence it is clearly of interest to examine the sensivity of these regularities to alternative levels of aggregation, and in particular, to identify the level of aggregation (industrial classification) at which these regularities are most pronounced.

To obtain data at a finer level of disaggregation, observe that since the present analysis requires only the number of industry establishments in each municipality, it is possible to extract such data from the telephone directory. For Japan, we have recently been able to obtain industrial location data for municipalities in 2006 based on the four-digit Nippon Telegraph and Telephone Business Classification System (NTTBCS). This more detailed data contains 539 manufacturing categories with positive employment, versus the 163 categories at the JSIC three-digit level used in the present analysis. By applying this analysis at the NTTBCS four-digit level, we should at least be able to identify differences between these regularities for two important levels of aggregation. Such comparisons will be reported in subsequent work.

6.2 Comparison with the US City System

While this cluster-based approach to central place and city-size regularities has been shown to be successful for the case of Japan, it is important to ask whether such regularities hold more generally. For the US case, Hsu [6] has shown that the NAS Rule (defined with respect to *pb*-choice cities) exhibits a significant presence in both the three- and four-digit industry

[51] See the companion papers, Mori et al. [13, Section 6] and Mori and Smith [14, pp.202-204] for further discussion of our research agenda.

classifications based on the North American Industry Classification System (NAICS).[52] This suggests that such regularities should continue to hold for definitions based on *cb*-choice cities, and will be examined in subsequent work.

In addition, County Business Pattern Data for the US provides establishment locations (at the county level) for industries up to the six-digit level going back as far as 1998. In particular, this data set includes 473 manufacturing categories with positive employment in 2007, which is roughly comparable to the four-digit NTTBCS data for Japan mentioned above. Hence by using these two data sets, it should be possible to conduct *comparative* studies of the US and Japan – at a level of aggregation that is much finer than that used in the present paper.

6.3 The Role of Spatial Structure

Finally, while the success of these cluster-based formulations suggest that both the Hierarchy Principle and NAS Rule reflect underlying spatial coordinations between population and industrial agglomerations, there is no explicit mention of *spatial structure* whatsoever. However, the theoretical models of urban hierarchies mentioned above (Fujita et al. [3], Tabuchi and Thisse [20, 21] and Hsu [6]) indicate that transports costs, scale economies and externalities may influence the spacing of agglomerations within each industry, and thus implicitly determine the spacing of their *cb*-choice cities. If so, then by studying the spatial relationships of *cb*-choice cities both within and between industries, one may hope to gain further insight into the underlying causes of these regularities. Initial efforts to quantify both the spacing of clusters within industries and the spatial coordination of clusters between industries were reported in Mori and Smith [15, Sections 8.2 and 8.3]. Such tools will be employed in subsequent work to examine these spatial questions.

[52] More precisely, the analysis of Hsu [6] includes all three- and four-digit NAICS industries, which are eqivalent to the set of industries considered in Mori et al. [13], i.e., excluding agriculture, forestory, fishing and hunting, mining, and public administration.

References

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

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

[3] Fujita, M., Krugman, P., and Mori, T., "On the evolution of hierarchical urban systems", *European Economic Review* 43, 209-251 (1999).

[4] Henderson, J.V., "Size and types of cities", *American Economic Review* 64, 640-656 (1974).

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

[6] Hsu, W., "Central place theory and city size distribution," mimeograph, Chinese University of Hong Kong (2009).

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

[8] ————, *Establishments and Enterprise Census* (in Japanese) (2001).

[9] Kanemoto, Y. and Tokuoka, K., "The proposal for the standard definition of the metropolitan area in Japan", *Journal of Applied Regional Science* 7, 1-15, in Japanese (2002).

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

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

[12] Mori, T., Nishikimi, K. and Smith, T.E., "A divergence statistic for industrial localization"*Review of Economics and Statistics* 87(4), 635-651 (2005).

[13] ————, "The number-average size rule: a new empirical relationship between industrial location and city size", *Journal of Regional Science* 48, pp.165-211 (2008).

[14] Mori, T. and Smith, T.E., "A Reconsideration of the NAS Rule from an industrial agglomeration perspective", in Burtless, G. and Pack, J.R. (eds.), *The Brookings-Wharton Papers on Urban Affairs: 2009*, Washington, D.C.: Brookings Institution Press.

[15] ————, "A probabilistic modeling approach to the detection of industrial agglomerations", Discussion Paper, No.682, Institute of Economic Research, Kyoto University (2009).

[16] Kullback, S., *Information Theory and Statistics*, New York: John Wiley & Sons, Inc. (1959).

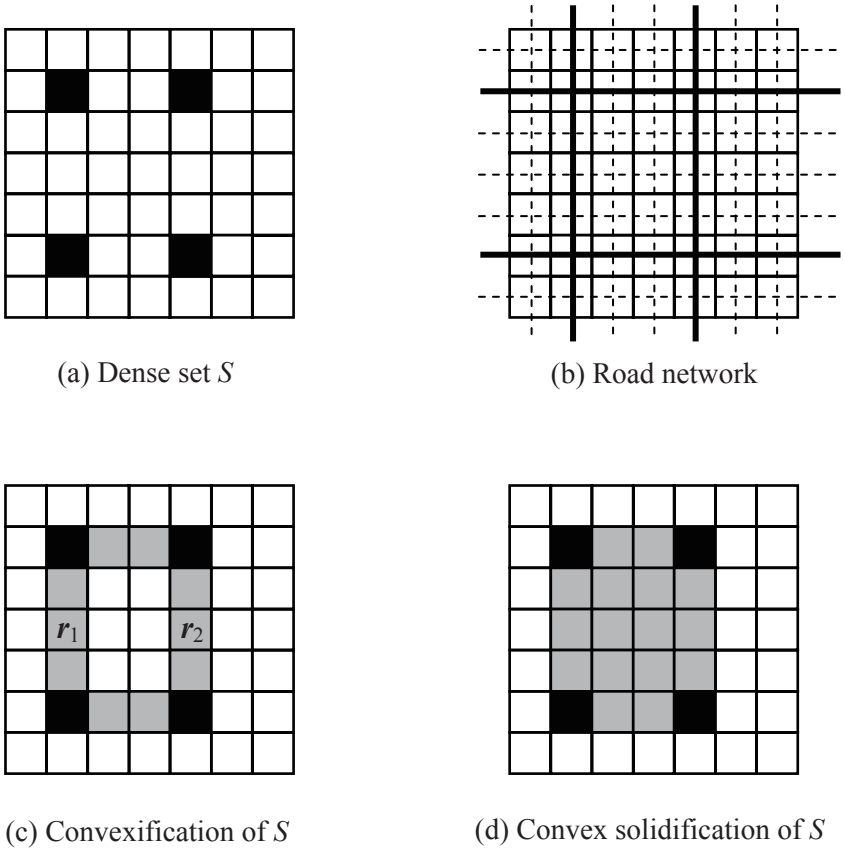
[17] Office of Management and Budget, "Standards for defining metropolitan and micropolitan statistical areas", *Federal Register* Vol.65, No.249 (2000).

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

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

[20] Tabuchi, T. and J.-F. Thisse, "Regional specialization, urban hierarchy, and commuting costs", *International Economic Review* 47, 1295-1317 (2006).

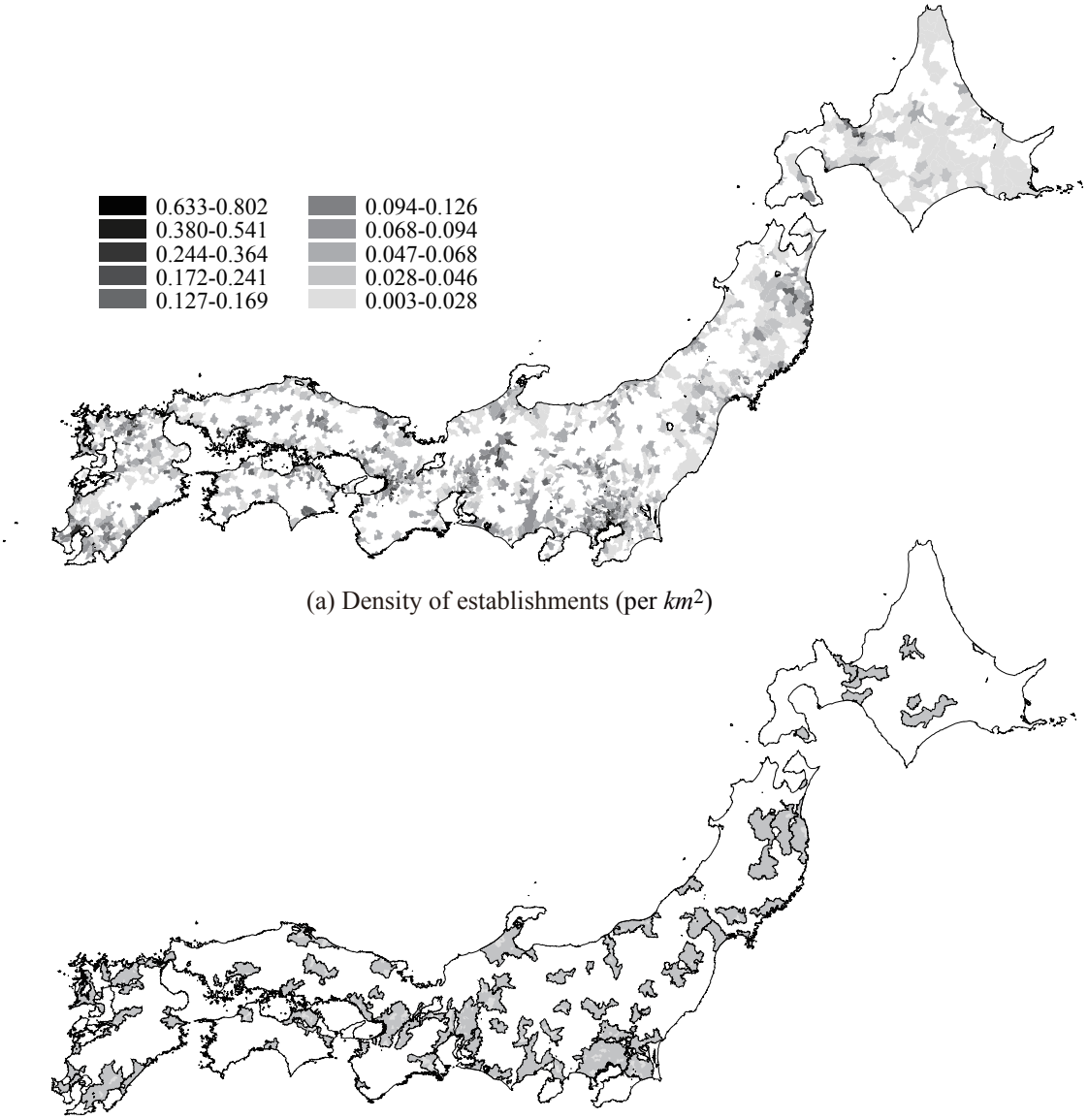
[21] ————, "Self-organizing urban hierarchy, "Discussion paper, No. F-414, Center for International Research on the Japanese Economy, Faculty of Economics, the University of Tokyo (2009).



(a) Dense set S (b) Road network

(c) Convexification of S (d) Convex solidification of S

Figure 2.1. Formation of clusters



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

(b) Clusters

Figure 2.2. Spatial pattern of "livestock products" industry (JSIC121)

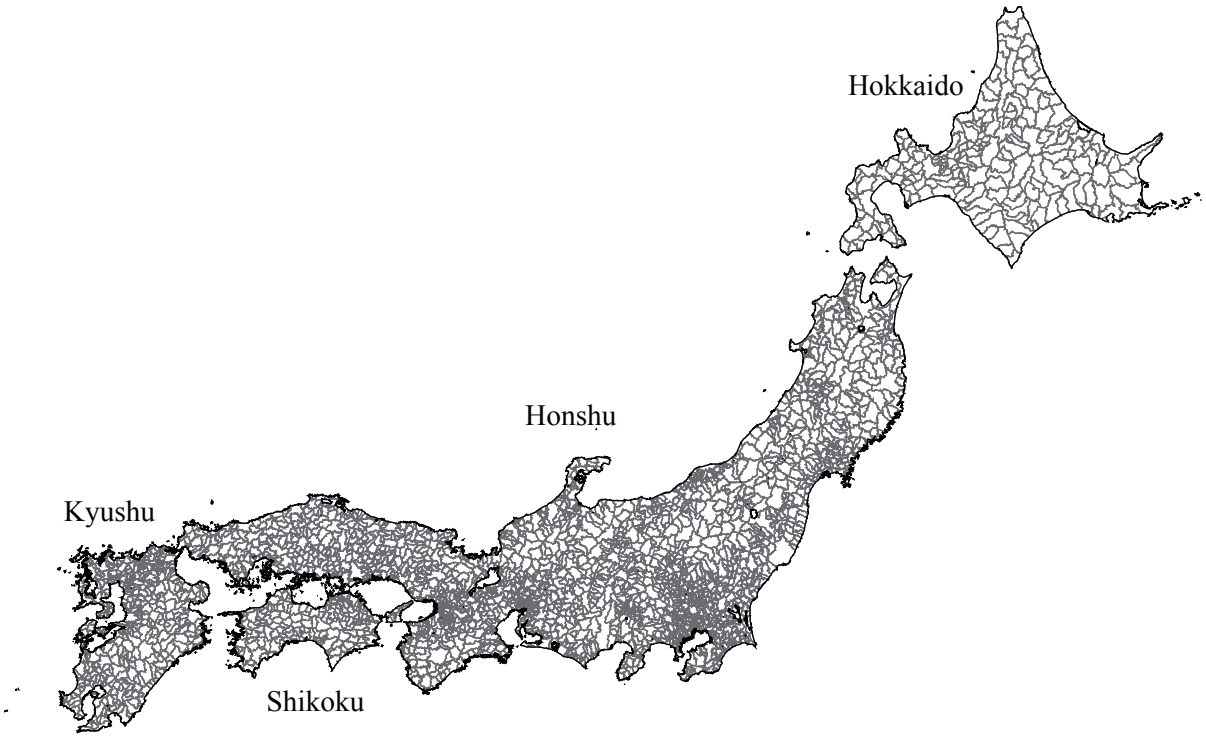


Figure 2.3. The regional system of Japan

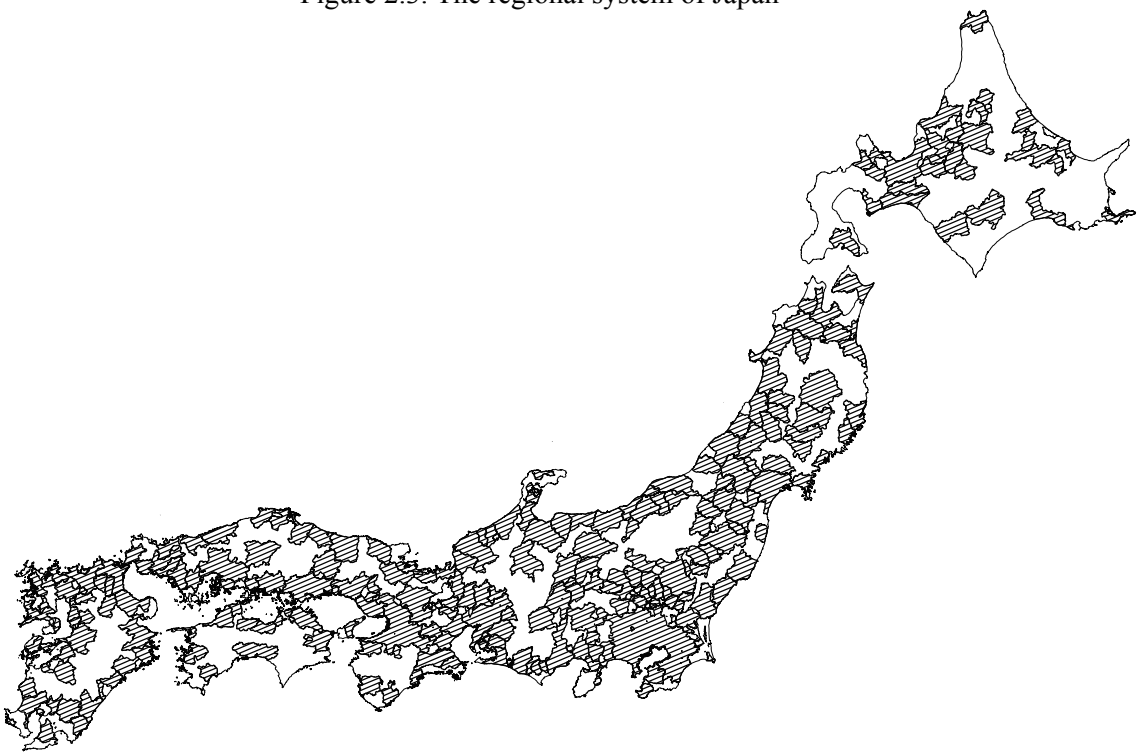


Figure 2.4. Cities in Japan

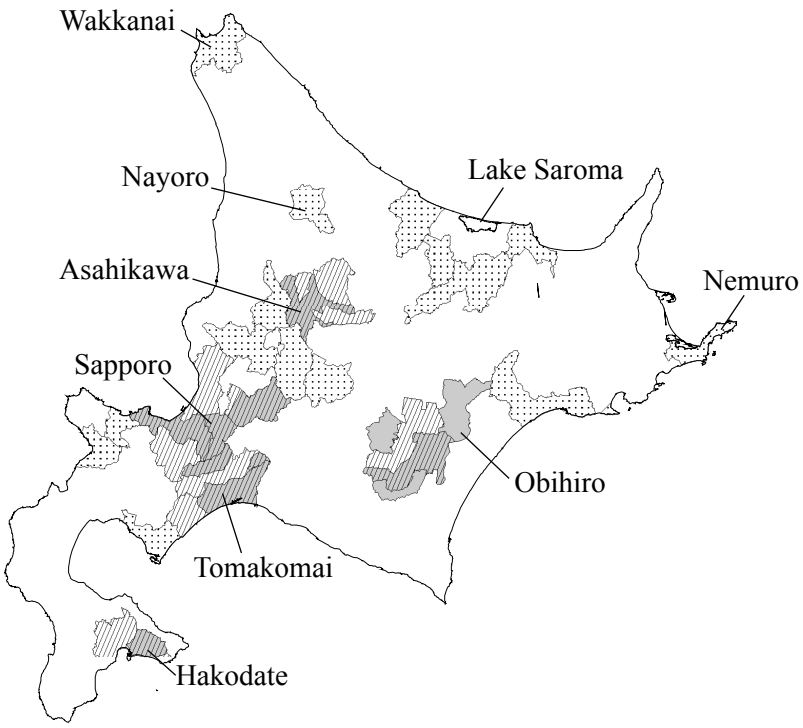


Figure 3.1. Choice cities for “livestock products” industry

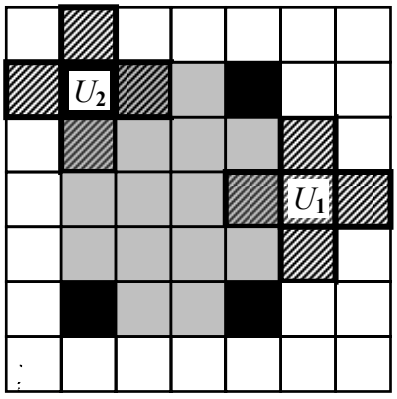


Figure 3.2. Cb-choice cities

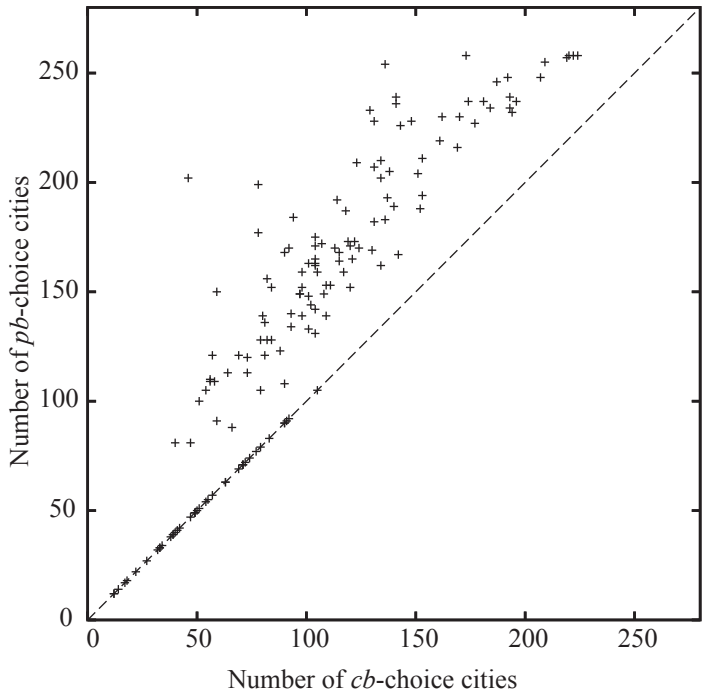


Figure 3.3. Number of industry-choice cities under two approaches

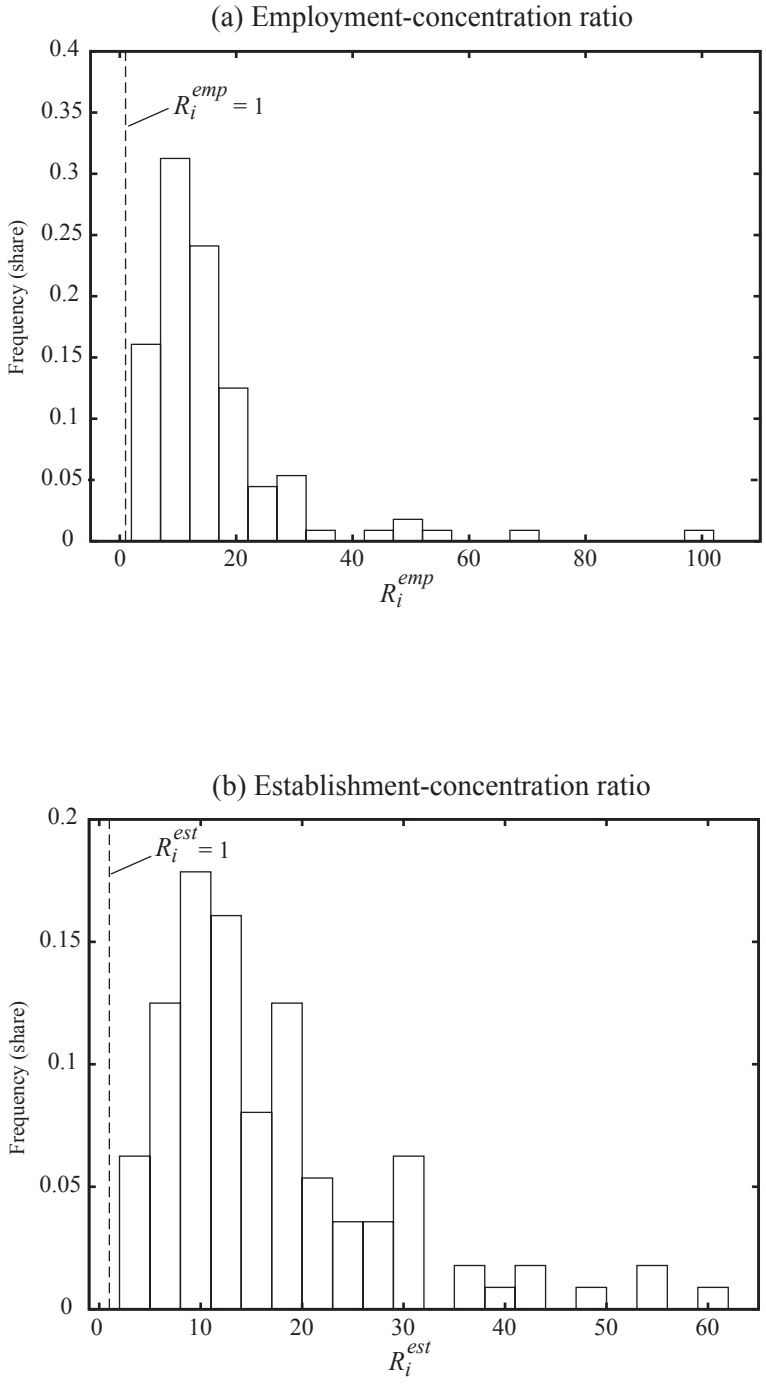


Figure 3.4. Average concentration in *cb*- versus *pb*-choice cities

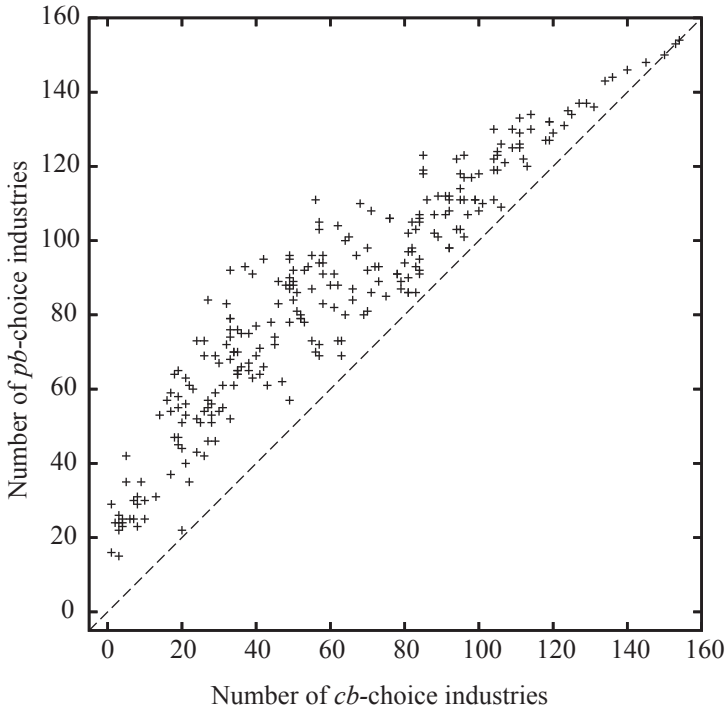


Figure 3.5. Number of choice industries in cities under two approaches

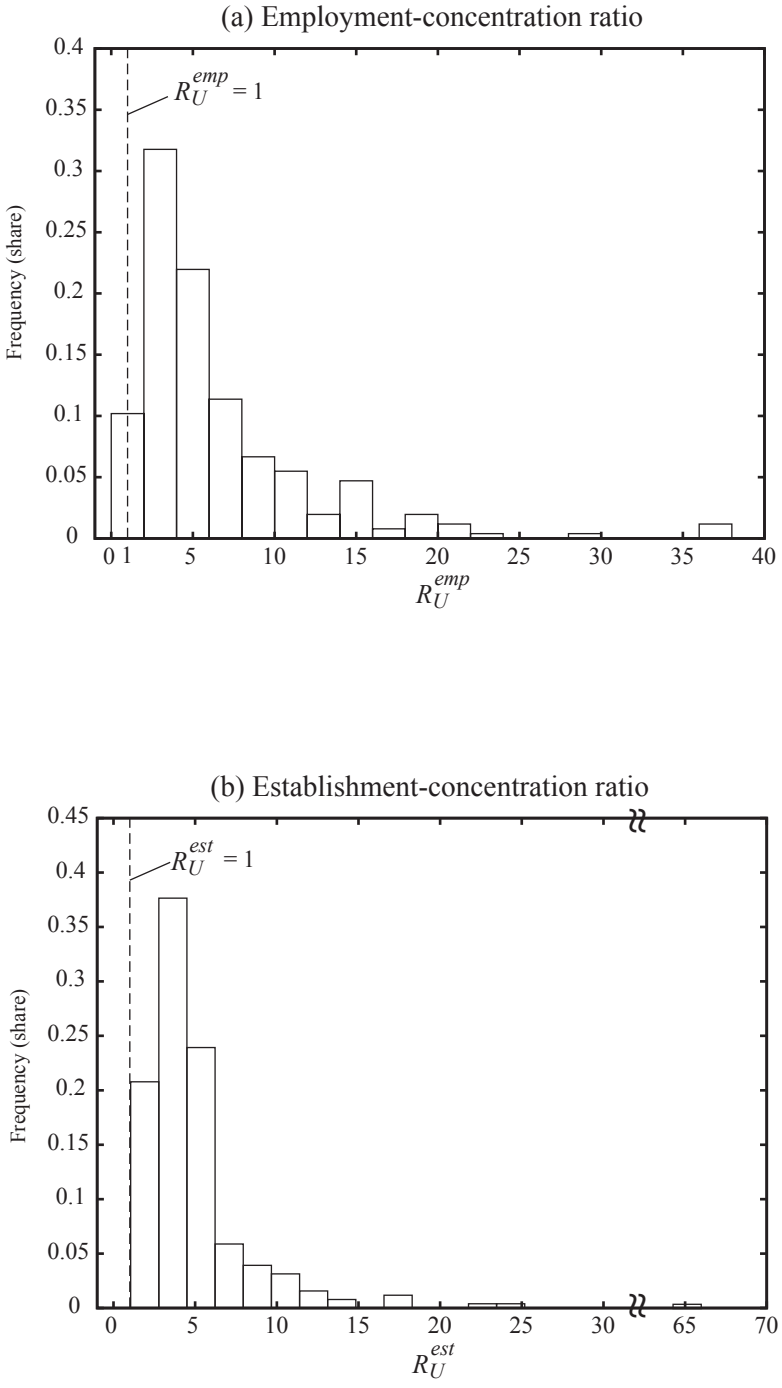


Figure 3.6. Average concentration of *cb*- versus *pb*-choice industries

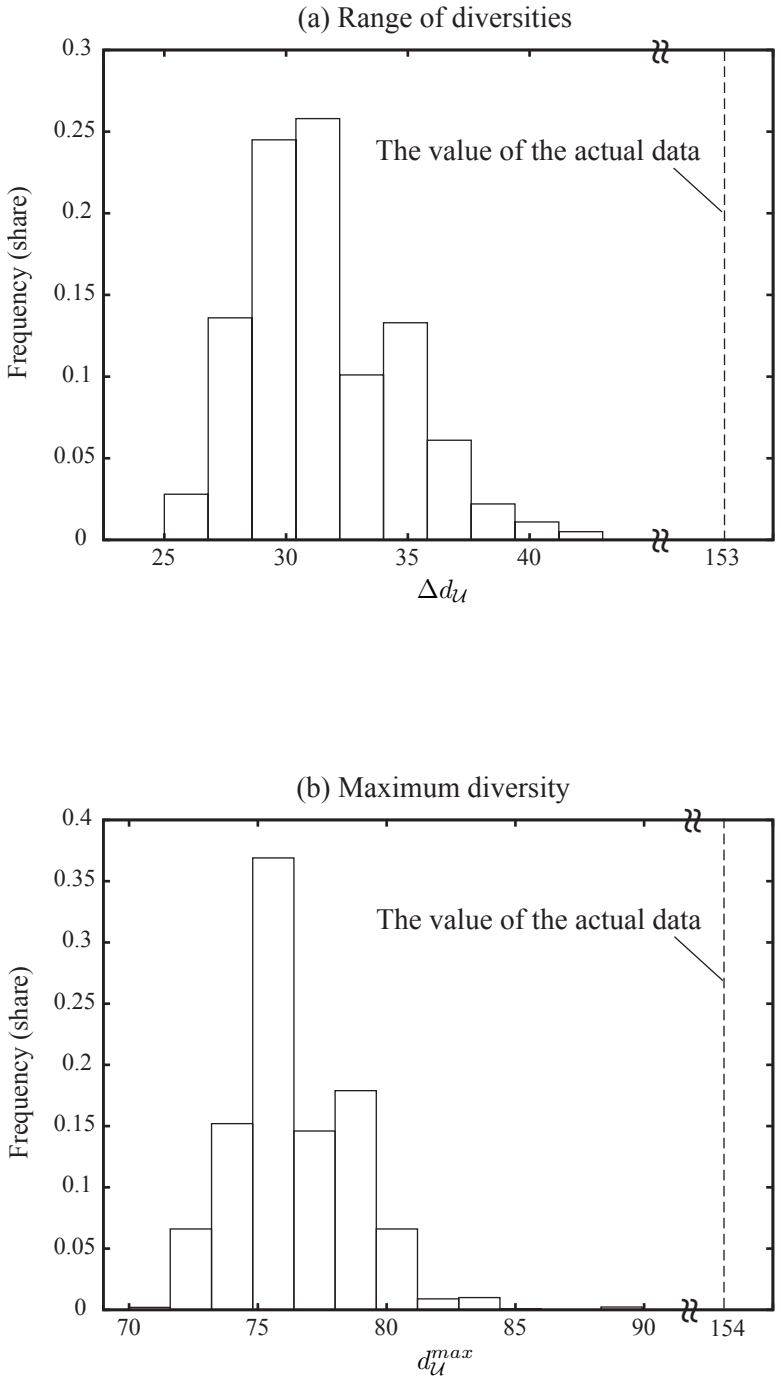


Figure 4.1. Industrial diversity of a city in random samples

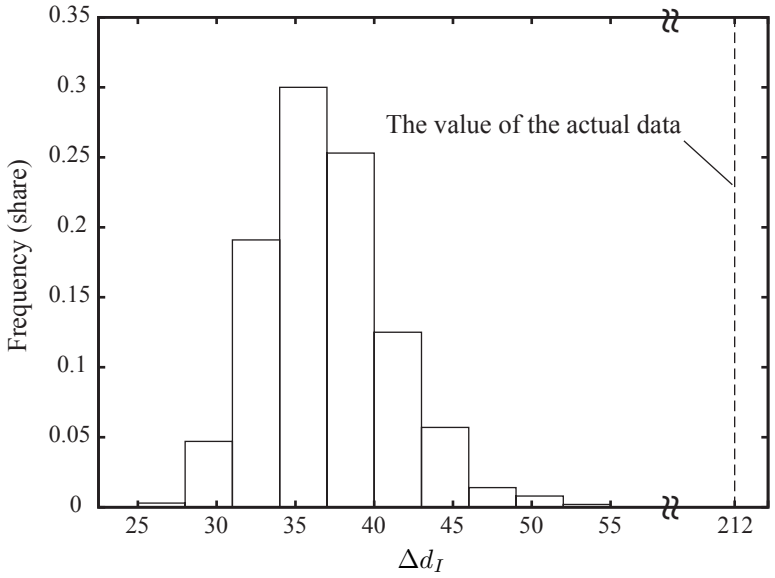


Figure 4.2. Locational diversities of industries in random samples

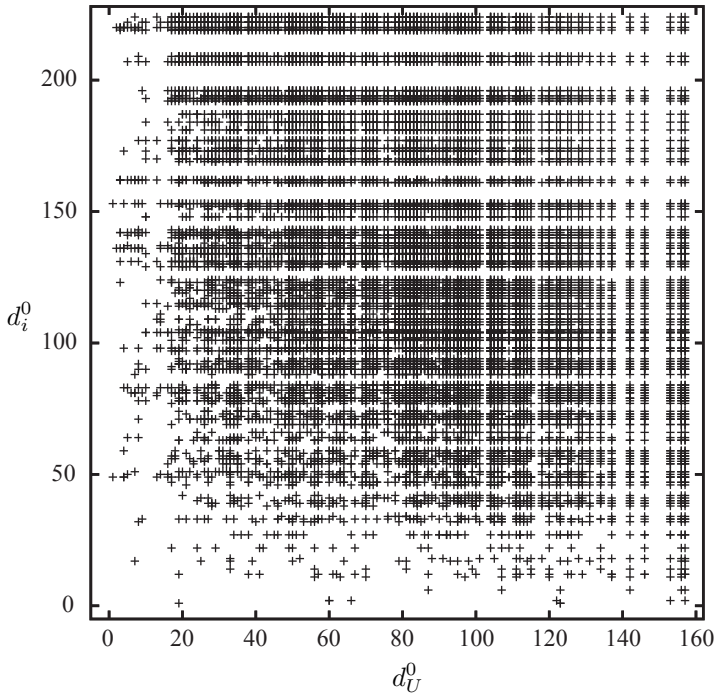


Figure 4.3. Industry-location events

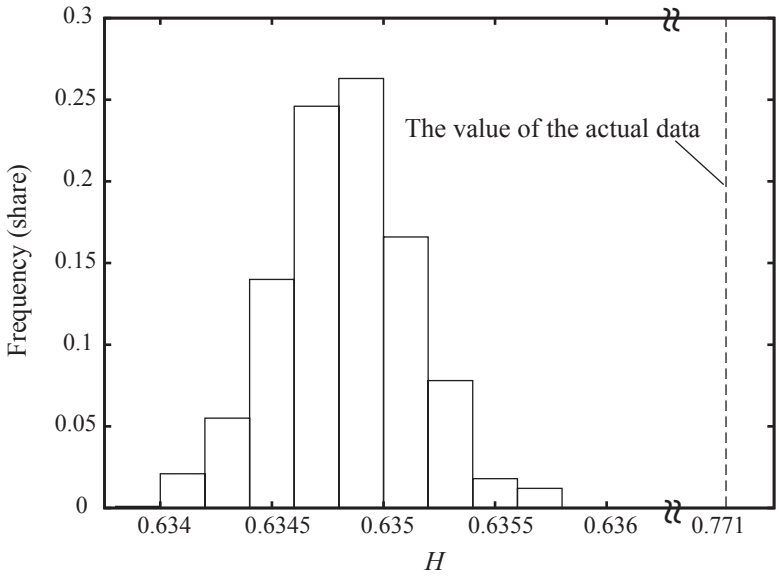


Figure 4.4. Hierarchy shares of random samples

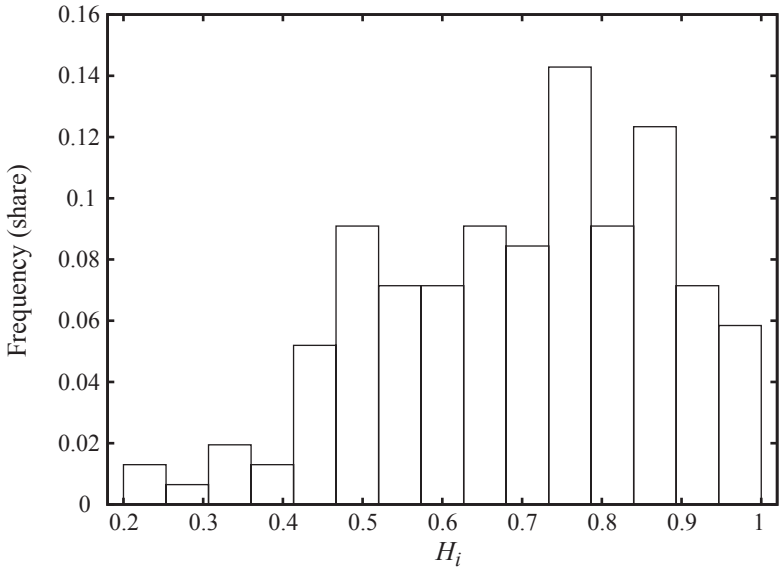


Figure 4.5. Hierarchy share for individual industries

JSIC	Industries with the smallest hierarchy shares	P_i
248	Fur skins	0.213
262	Iron smelting, without blast furnaces	0.252
214	Briquettes and briquette balls	0.270
211	Petroleum refining	0.315
245	Leather gloves and mittens	0.318
261	Iron industries, with blast furnaces	0.364
326	Ophthalmic goods, including frames	0.406
346	Lacquer ware	0.417
147	Rope and netting	0.434
244	Leather footwear	0.441

Table 4.1. Industries deviating from the Hierarchy Principle

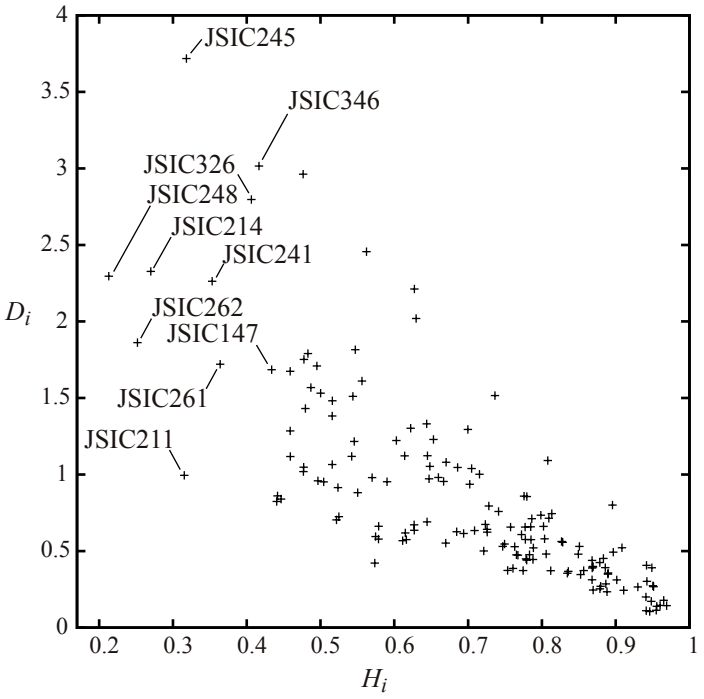


Figure 4.6. Hierarchy share and specialization index

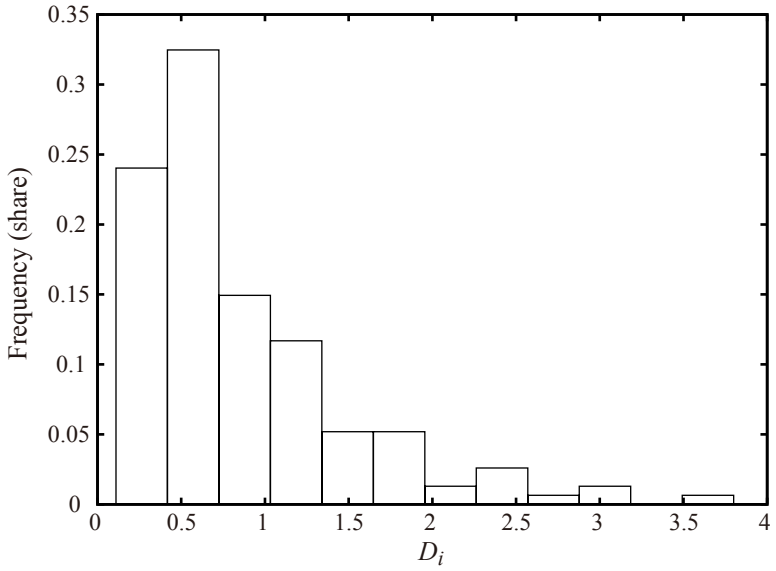


Figure 4.7. Specialization indices for individual industries

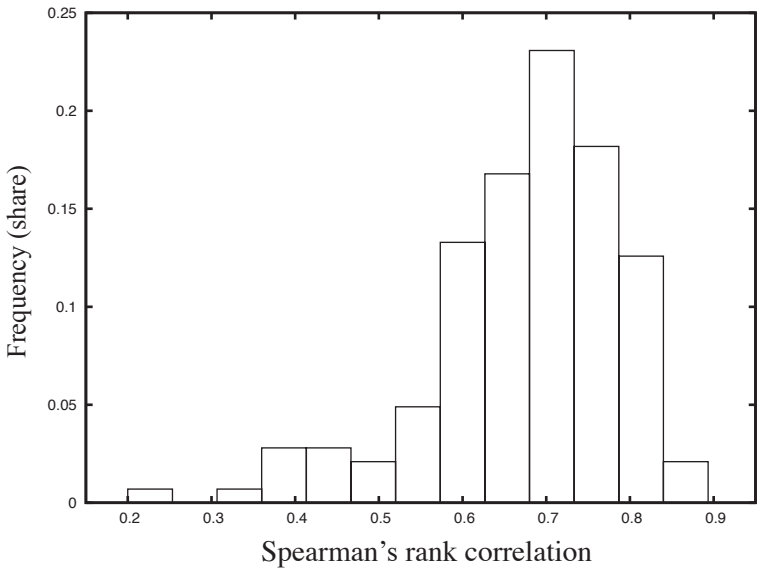


Figure 4.8. Correlation between employment shares within a city ($s_{i|U}$) versus within an industry ($s_{U|i}$)

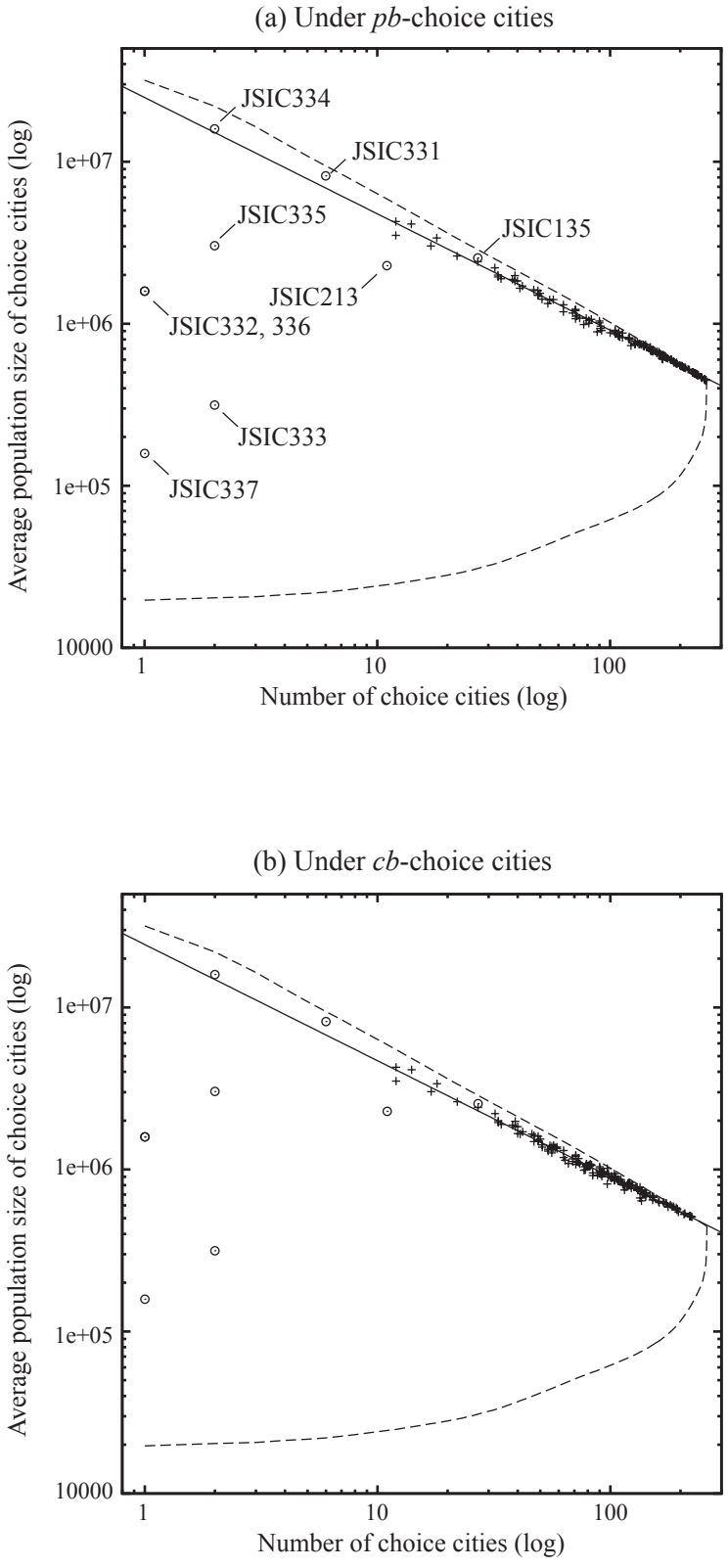


Figure 5.1. The Number-Average Size Rule