No.34

78-13 (No.34)

Spatially Constrained Clustering: Parametric and Nonparametric Methods for Testing the Spatially Homogeneous Clusters

Atsuyuki Okabe

November, 1978

(1) INTRODUCTION

In the field of spatial analysis, (such as geography, ecology and urban sociology), one would frequently encounter a problem of dividing a region S into a certain number of homogeneous subregions. To deal with this problem, the conventional method appeared in the related literature[1], (which may be called <u>(spatially) unconstrained clustering)</u>, employs the following procedure: first, the region S is divided into small unit-areas, called <u>cells</u>, s_1, s_2, ..., s_n, ($S = \sum_{i=1}^{n} s_i$, $s_i \cap s_j = \phi$, $i \neq j$). Second, the attribute of cell s_i is represented by x_i in the attribute space X; third, the points $\{x_i\}$ in X are grouped into a certain number of clusters, $\{C_1, C_2, \ldots, C_k\}$, by a certain clustering method; last, cells belonging to cluster C_j are illustrated on a map as the cells of cluster C_j, $j = 1, 2, \ldots, k$. This conventional method, however, will not always yield a required number of subregions, because cells belonging to the same cluster are not always spatially contiguous. To avoid this difficulty, an alternative method introduces spatial constraints in the attribute space X, that is, the distance between x_i and x_j in X is assigned a very large value if cells s_i and s_j are not spatially contiguous[2]. By this modification, a clustering method would yield a required number of spatially contiguous clusters, (or subregions), $\{C_1', C_2', \ldots, C_k'\}$. (This method will be called <u>(spatially) constrained clustering)</u>. It should be noted, however, that since the distance between points in X of the constrained case is always longer than or equal to that of the

unconstrained case, the affinity[3] between cells in the constrained clusters would become weaker than that in the unconstrained clusters. This may imply that the homogeneity of the constrained clusters would be worse than that of the unconstrained clusters. Stated differently, the homogeneity of the unconstrained clusters would be an upper bound for the constrained clusters. It would hence follow that the "most" homogeneous subregions will be observed when the constrained clusters $\{C'_j\}$ are "close" to the unconstrained clusters $\{C_j\}$. If they are "close", we shall say, for convenience, that the constrained clusters are spatially homogeneous clusters[4]. Obviously the spatially homogeneous clusters are of particular importance in the spatial analysis, because they would show a distinct spatial pattern. It may thus be worthwhile to consider a statistical method for testing whether or not the constrained clusters are the spatially homogeneous clusters. The purpose of this paper is to propose parametric (Section 2) and nonparametric (Section 3) methods for that test. With these methods, Section 4 shows a criteria for determining the number of clusters that would yield the "most" spatially homogeneous clusters. Finally a few potential uses of these methods are briefly mentioned.

(2) A PARAMETRIC TEST FOR THE SPATIALLY HOMOGENEOUS CLUSTERS

To make the analysis tractable, this section assumes that attribute value x_i is univariate[6], $x_i \in X = [0, 1]$, and the nearest neighbour method is adoped as a hierarchical clustering technique[7]. (All these restrictions will be relaxed in the nonparametric case). Let $y(i, j) = |x_i - x_j|$, (the Euclid distance), $D = \{y(i, j) \mid i \neq j, i, j = 1, 2, \ldots, n\}$ and $A_k = (a_{ij}^k)$ (a $n \times n$ matrix) be a cluster matrix at the k^{th} step, $(1 \leq k \leq n - 1)$, which is used to identify clusters, that is, $a_{ij}^k = 1$ if cells s_i and s_j belong to the same cluster; $a_{ij}^k = 0$ if otherwise. With this notation, the nearest neighbour clustering technique at the k^{th} step would generally be written as follows: find the minimum distance $y(i_k, j_k)$ in $D \setminus \{d(i, j) \mid a_{ij}^{k-1} = 0\}$; set $a_{ij}^k = a_{ji}^k = 1$ if $(i, j) = (i_k, j_k)$, or if there exist $a_{i\ell}^{k-1}\ a_{\ell m}^{k-1} \ldots a_{nj_k}^{k-1} = 1$ or $a_{j\ell}^{k-1}\ a_{\ell m}^{k-1} \ldots a_{ni_k}^{k-1} = 1$; if otherwise, set $a_{ij}^k = a_{ij}^{k-1}$. (Note that $n - k$ clusters would be obtained at the k^{th} step and that $A^\circ = (0)$). This general procedure, however, could be simplified by the assumption of $x_i \in [0, 1]$. To see it, re-index $\{x_i\}$ in an increasing order, i.e., $x_1 \leq x_2 \leq \ldots \leq x_n$, and let $y_{i-1} = x_i - x_{i-1}$, $i = 2, 3, \ldots, n$. Again re-index $\{y_i\}$ in an increasing order, i.e., $y_1 \leq y_2 \leq \ldots \leq y_{n-1}$. One would then realize that $y(i_k, j_k) = y_k$.

The distinction between the unconstrained and constrained clustering is found in the distance set: the former uses D; the latter uses $D' = \{d(i', j') \mid (i', j') \in K\}$ where $K = \{(i', j') \mid$

cells s_i and s_j share the common boundary line on a map}.
Obviously, $D' \subseteq D$. (Note that the notation of the constrained case is the same as that of the unconstrained case except "'", like D', $y'(i, j)$ and $A'_k = (a^{k'}_{ij})$). It would follow from D' _ D that

$$y'(i_k, j_k) \geq y(i_k, j_k), \qquad k = 1, 2, \ldots, n-1 \tag{2.1}$$

Now let $\Delta_k = y'(i_k, j_k) - y(i_k, j_k)$. Then the closeness between the constrained clusters $\{C'_j\}$ and the unconstrained clusters $\{C_j\}$ may be measured by Δ_i, $i = 1, 2, \ldots, k$. Obviously, if $\{C'_j\}$ and $\{C_j\}$ are completely the same, we have $\Delta_i = 0$, $i = 1, 2, \ldots, k$. With this distance, we shall next construct a probabilistic model in which $\Delta_i \approx 0$ would statistically be tested.

First, we regard a set of the attribute values $\{x_i\}$ as a sample of n random variables that are independently drawn from the uniform distribution $f(x) = 1 \quad 0 \leq x \leq 1$. Recall that y_k is a function of $\{x_i\}$, it is obvious that y_k is also a random variable. We next consider a situation in which one sample $\{\bar{x}_{i1}\}$ (and hence $\bar{y}_{k1}$) is provided and we know the information that y_{k2} which will be obtained from the second sample $\{x_{i2}\}$ has the property of $y_{k2} \geq \bar{y}_{k1}$. In this context, let us consider the conditional probability $P\{y_{k2} \geq y^*_k \mid y_{k2} \geq \bar{y}_{k1}\}$ of a random variable y_{k2} being greater than y^*_k provided that y_{k2} is greater than a given $\bar{y}_{k1}$. If this probability distribution function is known, we may obtain $y^*_{k\alpha}$ that satisfies $P\{y_{k2} \geq y^*_k \mid y_{k2} \geq \bar{y}_{k1}\} = \alpha$, where α is a level of significance. We could then infur that if the

observed value $\bar{y}_{k2}$ of the second sample is less than $y^*_{k\alpha}$, the difference $\bar{y}_{k2} - \bar{y}_{k1}$ is regarded as a chance variation.

The probability $P\{ y_{k2} \geq y^*_k \mid y_{k2} \geq \bar{y}_{k1} \}$ would readily be obtained if one recalls Kendall and Moran's (1963) result[8]. They show that the probability of exactly ℓ intervals among $\{ x_i - x_{i-1} \mid i = 2, 3, \ldots, n \}$ (n - 1 intervals) being greater than y is given by

$$F_\ell(y) = \binom{n-1}{\ell} \sum_{i=0}^{n-\ell-1} \delta_i(y)(-1)^i \binom{n-\ell-1}{i} \{ 1 - (\ell + i)y \}^{n-2}, \tag{2.2}$$

$$\delta_i(y) = \begin{cases} 1 & \text{if } 0 \leq y \leq 1/(\ell + i) \\ 0 & \text{if otherwise.} \end{cases}$$

Hence the probability of the $(\ell + 1)^{th}$ largest interval or the $k = (n - \ell - 1)^{th}$ smallest interval being less than y would be provided by $F_\ell(y)$ or

$$G_k(y) = \binom{n-1}{n-k-1} \sum_{i=0}^{k} \delta_i(y)(-1)^i \binom{k}{i} \{ 1 - (n - k - 1 + i)y \}^{n-2} \tag{2.3}$$

Note that $P\{ y \geq y^* \} = 1 - G_k(y^*)$, one would obtain

$$P\{ y_{2k} \geq y^*_k \mid y_{2k} \geq \bar{y}_{k1} \} = (1 - G_k(y^*_k))/(1 - G_k(\bar{y}_{k1})) \tag{2.4}$$

Upon solving $1 - G_k(y^*_{k\alpha}) = \alpha(1 - G_k(\bar{y}_{k1}))$, the critical value $y^*_{k\alpha}$ may be obtained. Alternatively, if $\alpha_k = \alpha(\bar{y}_{k2} \ \bar{y}_{k1}) \geq \alpha$ where

$$\alpha(\bar{y}_{k2} \mid \bar{y}_{k1}) = (1 - G_k(\bar{y}_{k2}))/(1 - G_k(\bar{y}_{k1})) \tag{2.5}$$

we could infur that $\bar{y}_{k1} \leq \bar{y}_{k2} \leq y_k^*$.

With the above understanding, we may test whether or not the constrained clusters $\{C_j'\}$ are close to the unconstrained clusters $\{C_j\}$, i.e., $\Delta_m \approx 0$. To be explicit, if the relation

$$\alpha(y'(i_m, j_m) \mid y(i_m, j_m)) > \alpha, \qquad m = 1, 2, \ldots, k, \tag{2.6}$$

holds, $\Delta_m \approx 0$, $m = 1, 2, \ldots, k$ would be accepted at significance level α. Stated differently, the constrained clusters $\{C_j'\}$ are the spatially homogeneous clusters if equation (2.6) holds.

(3) A NONPARAMETRIC TEST FOR THE SPATIALLY HOMOGENEOUS CLUSTERS

Let us now consider an alternative method without assuming the distribution of $\{x_i\}$. In this section, it is not necessary to assume the univariate attribute value $\{x_i\}$, or the Euclide distance in the attribute space X, or the nearest neighbour clustering technique. We only need two sets of clusters, $\{C_j\}$ (the unconstrained clusters) and $\{C_j'\}$ (the constrained clusters) which are respectively obtained from a certain clustering technique (whether it may be) in X (generally a vector space) with a distance set $D = \{ y(i, j)\}$ and in X with a distance set $D' = \{ y'(i, j) \mid y'(i, j) = y(i, j)$ if $(i, j) \in K$; $y'(i, j) = \infty$ if $(i, j) \notin K\}$, where the set K is the same as was defined in Section 2. Once the constrained and unconstrained clusters are provided, the attention should be focused upon the "closeness" between $\{C_j\}$ and $\{C_j'\}$. To state the "closeness" explicitly, we define <u>j-connected</u> pairs: if cells s_ℓ and s_m belong to the same cluster C_j, these two cells are called j-connected pairs. Obviously there are $|C_j|\ (|C_j| - 1)/2$ j-connected pairs in C_j, where $|C_j|$ indicates the number of cells in C_j. If two clusters $\{C_j\}$ and $\{C_j'\}$ are almost the same, the most of j-connected pairs in C_j, $j = 1, 2, \ldots, k$ would be preserved in a certain C_i' or some C_i's. (That is, if s_ℓ and s_m are connected pairs in $\{C_j\}$, they would not become unconnected pairs in $\{C_j'\}$). Therefore, the closeness or distance between the constrained and unconstrained clusters may be defined by the number of j-connected pairs, $j = 1, 2, \ldots, k$, of $\{C_j\}$ being lost in $\{C_j'\}$. To be

precise, let $\bar{n}_{ij}$ be the number of cells in C'_j that belong to C_i. Then, noticing that $\bar{n}_{ij}\ (\bar{n}_{ij} - 1)/2$ is the number of j-connected pairs preserved in C'_i, the distance Δ_k between the unconstained and constrained clusters may be defined by

$$\Delta_k = \sum_{j=1}^{k} |C_j| (|C_j| - 1)/2 - \sum_{i=1}^{k} \sum_{j=1}^{k} \bar{n}_{ij} (\bar{n}_{ij} - 1)/2. \tag{3.1}$$

(Obviously if $\{C_j\}$ and $\{C'_j\}$ are completely the same, $\Delta_k = 0$). With this distance, we shall next consider a probabilistic model in which $\Delta_k \approx 0$ will statistically be tested.

Suppose that there are $|C_j|$ j-colored balls, $j = 1, 2, \ldots, k$, which are randomly mixed in box C, $(n = \sum_{j=1}^{k} |C_j|)$, and that there are k boxies, say box j, $j = 1, 2, \ldots, k$, whose capacity is given by $|C'_j|$, $(n = \sum_{j=1}^{k} |C'_j|)$. Consider next that n colored balls in box C are randomly placed in k boxies without replacement, and let u_{ij} be the number of j-colored balls in box i, which is a random variable. It should be noted that

$$\begin{cases} \sum_{i=1}^{n} u_{ij} = |C_j| , \\ \sum_{j=1}^{n} u_{ij} = |C'_i| , \\ \sum_{j=1}^{k} |C_j| = \sum_{i=1}^{k} |C'_i| = n. \end{cases} \tag{3.2}$$

We now define

$$U_k(\{u_{ij}\}) = \sum_{j=1}^{k} |C_j| \, (|C_j| - 1)/2 - \sum_{i=1}^{k} \sum_{j=1}^{k} u_{ij} \, (u_{ij} - 1)/2. \tag{3.3}$$

(One would readily notice that $\sum_{i=1}^{k} u_{ij}(u_{ij} - 1)/2$ corresponds to the number of j-connected pairs in $\{C_j\}$ being preserved in $\{C'_j\}$). Obviously U_k is a random variable and hence U_k has a probability distribution. If this distribution function is obtained, we could calculate the probability of U_k being less than U_k^*. Let $U_{k\alpha}^*$ be the critical value of this probability being α, and $\bar{U}_k$ be an observed value of U_k. We may then accept, at significance level α, that the clusters $\{C'_j\}$ are not randomly constructed clusters, if $\bar{U}_k \geq U_{k\alpha}^*$. Stated differently, two sets of clusters $\{C_j\}$ and $\{C'_j\}$ are more close than random.

Let us now obtain the probability of $u_{ij} = n_{ij}$, i, j = 1, 2, ..., k, denoted by $P\{u_{ij} = n_{ij}\}$. The way of choosing n_{1j} j-colored balls j = 1, 2, ..., k, ($\sum_{j=1}^{k} n_{1j} = |C'_1|$), from n balls that consist of $|C_j|$ j-colored balls, j = 1, 2, ..., k, and placing them in box 1 is given by[9]

$$\prod_{j=1}^{k} \binom{|C_j|}{n_{1j}}$$

The way of choosing n_{2j} j-colored balls, j = 1, 2, ..., k,

$(\sum_{j=1}^{k} n_{2j} = |C_2'|)$, from $n - |C_1'|$ balls that consist of $|C_j| - n_{1j}$ j-colored balls, j = 1, 2, ..., k, and placing them in box 2 is given by

$$\prod_{j=1}^{k} \binom{|C_j| - n_{1j}}{n_{2j}}$$

... The total way of placing n balls in k boxies is given by $n! / \prod_{j=1}^{k} |C_j'|!$. Hence the probability of u_{ij} being equal to n_{ij} would be given by

$$P\{u_{ij} = n_{ij}\} = \prod_{m=1}^{k} \left\{ \prod_{j=1}^{k} \binom{|C_j| - \sum_{\ell=1}^{m-1} n_{\ell j}}{n_{mj}} \right\} \prod_{j=1}^{k} |C_j'|!/n! \qquad (3.4)$$

Let $\mathcal{N}(U_k^*) = \{\{n_{ij}\} \mid U_k(\{n_{ij}\}) \le U_k^*\}$ be a set of $\{n_{ij}\}$ that satisfies $U_k(\{n_{ij}\}) \le U_k^*$. Then the probability of the distance between two clusters $\{C_j\}$ and $\{C_j'\}$ being less than U_k^* would be given by

$$P\{U_k(u_{ij}) \le U_k^*\} = \sum_{\{n_{ij}\} \in \mathcal{N}(U_k^*)} P\{u_{ij} = n_{ij}\}. \qquad (3.5)$$

By solving $P\{U_k(u_{ij}) \le U_k^*\} = \alpha$, the critical value U_k^* may be obtained. With this value, we may infur that if the relation $0 \le \bar{U}_k(\{\bar{n}_{ij}\}) < U_k^*$ holds, where $\{\bar{n}_{ij}\}$ is the observed value, the hypothesis that $\{C_j'\}$ are not randomly constructed clusters will be accepted at significance level α. Alternatively, the same conclusion would be obtained if the relation

$$\alpha_k(\{\bar{n}_{ij}\}) = \sum_{\{n_{ij}\}\in\mathcal{N}(\bar{U}_k)} P\{u_{ij} = \bar{n}_{ij}\} \leq \alpha \tag{3.6}$$

holds.

Although the above method would provide the exact probability, calculating equation (3.5) may be time-consuming when the number of clusters, k, is large. To avoid this difficulty, the following less ambitious but more practical method might be useful. A clue would readily be obtained if $\{n_{ij}\}$ is tabulated in Table 1. One would then realize that this table is regarded as the contingency table. It should be noted, however, that clusters consisting of zero or one cell is out of concern, because the connected pair cannot exist in such clusters. Hence the incomplete contingency table should be used here[10]. To be explicit, let $\delta_{ij} = 1$ if $n_{ij} \geq 2$; $\delta_{ij} = 0$ if $n_{ij} = 0$ or 1. The incomplete contingency table is then written as a subtable of (n_{ij}) where elements of $\delta_{ij} = 0$ are excluded. The test statistic is given by

$$\chi_k^2(\{n_{ij}\}) = \sum_{i=1}^{k}\sum_{j=1}^{k} \delta_{ij}(n_{ij} - \hat{n}_{ij})^2/\hat{n}_{ij}, \tag{3.7}$$

where

$$\hat{n}_{ij} = n|C_i||C'_j|/\sum_{\ell=1}^{k}|C_\ell|\sum_{\ell=1}^{k}|C'_\ell|. \tag{3.8}$$

It is noted that χ_k^2 would follow the chi-square distribution with degrees of freedom $(k-1)^2 - \sum_{i=1}^{k}\sum_{j=1}^{k}(1-\delta_{ij})$. With this statistic and

$$\alpha_k (\{\bar{n}_{ij}\}) = P\{\chi_k \leq \chi_k (\{\bar{n}_{ij}\})\}, \tag{3.9}$$

where $\{\bar{n}_{ij}\}$ is a set of observed values, we may infur that the hypothesis of $\{\bar{n}_{ij}\}$ not being statistically independent is accepted at significance level α if $\alpha_k (\{\bar{n}_{ij}\}) \leq \alpha$. Stated differently, two sets of clusters $\{C_j\}$ and $\{C'_j\}$ are closely related if $\alpha_k \leq \alpha$.

The above model may provide a criteria for judging whether or not the constrained clusters $\{C'_j\}$ are close to the unconstrained clusters $\{C_j\}$. To be explicit, if $\alpha_k(\{\bar{n}_{ij}\})$ of equation (11) or (14) is less than α, we could say that the constrained clusters $\{C'_j\}$ are the spatially homogeneous clusters with significance level α.

(4) THE OPTIMUM NUMBER OF THE SPATIALLY CONSTRAINED CLUSTERS

Generally the number of clusters is predetermined before clustering, but one would sometimes meet a situation in which he should determine this number by observing the results of cluter-ing. In this case, α_k defined in the preceeding sections would provide a useful criteria.

In the parametric case, as was discussed in Section 2, a hypothesis of k clusters being spatially homogeneous was tested in terms of α_i given by equation (2.4), $i = 1, 2, \ldots, n-k$. Hence, the magnitude of the spatially homogeneous state may be measured by

$$\bar{\alpha}_k = \sum_{i=1}^{n-k} \alpha_i / (n - k) . \tag{4.1}$$

Obviously the high value of $\bar{\alpha}_k$ indicates the high spatial homo-geneity. It may thus be reasonable to determine the optimum number k^* of clusters by

$$\alpha_{k^*} = \max \{ \bar{\alpha}_1, \bar{\alpha}_2, \ldots, \bar{\alpha}_{n-1} \} . \tag{4.2}$$

In the nonparametric case, α_k is given by equation (3:6) or (3.9). Since the low value of α_k indicates the high spatial homo-geneity, the optimum number k^* of clusters may be determined by

$$\alpha_{k^*} = \min \{ \alpha_1, \alpha_2, \ldots, \alpha_n \} . \tag{4.3}$$

In both cases, the optimum number k^* would yield the most spatially homogeneous clusters. In other words, k^* clusters would show the most spatially distinct pattern.

(5) CONCLUDING REMARKS

Besides testing the spatially homogeneous clusters, the method proposed here may provide a useful tool for the spatial analysis. The value of α_k defined by (2.4), (3.6) and (3.9) could be used to measure the spatial association of biological, sociological or urban activities in a space. In particular, if α_{k*} of equations (4.2) and (4.3) is adoped, a shortcoming of the quadrat method[11] might be overcome to a certain extent, because the "appropriate" size of areas to which an association statistic[12] is applied is simultaneously determined. Furthermore, α_k^* would reveal the shape and configuration of the most homogeneous areas. Finally it should be noted that this method would provide a useful information for spatial planning, such as land use zoning. We hope to show these applications in another occasion.

Footnotes

1. See, for example, King (1969, Chapter 8), Pielou (1969, Chapter IV), and Herbert and Johnston (1977).
2. See Cliff et al. (1975, Chapter 2).
3. Mathematically, this will be written as equation (2.1).
4. In this paper, the term "spatially homogeneous clusters" has no meaning more than $\{C_j'\}$ being "close" to $\{C_j\}$. The "closeness" will be defined later.
5. Mathematical definition will be provided in Sections 2 and 3.
6. Scores on the principal component may be used if x_i is multivariate.
7. A good review of clustering technique is provided by Cormack (1971).
8. See Section 2.14. Equation (2.36) of Kendall and Moran corresponds to equation (2.2) with a minor correction. (The second term of equation (2.36) should be $\binom{n-k}{1}$.
9. Recall the multivariate hypergeometric distribution.
10. See, for example, Cochran (1954), and Goodman (1969).
11. Ordinary the choice of the quadrat-size is arbitrary. It is noticed, however, that the conclusion sometimes varies according to the quadrat-size.
12. For example, Geary's ratio (1954).

References

Cliff, A. D., Haggett, P., Ord, J. K., Bassett, K., Davies, R. B. (1975), Elements of Spatial Structure, Cambridge: Cambridge University Press.

Cochran, W. G. (1954), "Some Methods for Strengthening the Common χ^2 tests", Biometrics, 10, 417-451.

Cormack, R. M. (1971), "A Review of Classification", (with discussion), Journal of Royal Statistical Society, A, 134, 321-367.

Geary, R. C. (1954), "The Contiguity Ratio and Statistical Mapping", The Incorporated Statistician, 5, 115-141.

Goodman, L. A. (1969), "The Analysis of Cross-Classified Data: Independence, Quasi-Independence, and Interactions in Contingency Tables with or without Missing Entries", Journal of the American Statistical Association, 75, 1-40.

Herbert, D. T., Johnston, R. J. (1977), Social Areas, I, New York: John Wiley & Sons.

Kendall, M. G., Moran, P. A. P. (1963), Geometrical Probability, London: Charles Griffin.

Pielou, E. C. (1969), An Introduction to Mathematical Ecology, New York: John Wiley & Sons.

Table 1: Contingency table of C_j and C'_j

	C_1	C_2		C_k	
C'_1	$\bar{n}_{11}$	$\bar{n}_{12}$	.....	$\bar{n}_{1k}$	$\lvert C'_1 \rvert$
C'_2	$\bar{n}_{21}$	$\bar{n}_{22}$	.....	$\bar{n}_{2k}$	$\lvert C'_2 \rvert$
⋮	⋮	⋮		⋮	⋮
C'_k	$\bar{n}_{k1}$	$\bar{n}_{k2}$	.....	$\bar{n}_{kk}$	$\lvert C'_k \rvert$
	$\lvert C_1 \rvert$	$\lvert C_2 \rvert$		$\lvert C_k \rvert$	n