Variable Clumping Method (VCM): An Explanatory Tool for Detecting Spatial Hierarchy in the Distribution of Points

Atsuyuki Okabe* and Shino Funamoto**

January 31, 1999

* Center for Spatial Information Science, University of Tokyo
** Department of Urban Engineering, University of Tokyo
7-3-1 Hong, Bunkyo-ku, Tokyo 113-8656, Japan

Abstract

This paper shows an explanatory tool (a FORTRAN program linked to GIS software), called VCM (Variable Clumping Method), for detecting a geometrical form of spatial hierarchy in the distribution of points. The paper first formulates a clumping method, called the variable clumping method, in which a clump is defined by a set of connected circles centered at given points and the state of connected circles is observed with respect to a variable radius of the circles. Second, the paper develops a computational method for the variable clumping method. Third, the paper shows the procedure for running the program VCM. Last, the paper discusses the performance of VCM.

Acknowledgements

Part of this paper was presented in the session of the Commission on Modelling Geographical Systems of the IGU Conference held in Lisbon, 1998. The authors are thankful for valuable comments from the floor, in particular, comments by Professor Barry Boots. We also acknowledge the use of VORONOI2 developed by Sugihara and Iri (1989).

List of Figures

1 Introduction

This paper shows an explanatory tool (a FORTRAN program linked to GIS software), called VCM, for detecting a geometrical form of spatial hierarchy in the distribution of points.

The idea of exploratory analysis goes back Tukey (1977), but the development of this idea in geographical analysis began in the late 1980's. Openshaw *et al.* (1987) developed GAM (Geographical Analysis Machine) to deal with clustering phenomena in the distribution of points (cancer patients). Haslett *et al.* (1987) developed SPIDER (Spatial Interactive Data Explorer) to provide functions of interactive visual presentation of geographical data. Walker and Moore (1988) developed SIMPLE (Spatial and Inductive Modelling Package for Land Evaluation) to link geographical programs to general programs, such as MINITAB and GLIM. Anselin *et al.* (1993) developed SPACESTAT for spatial econometrics. The importance of these exploratory tools in geographical analysis is discussed in depth by Goodchild (1987), Haslett *et al.* (1990), Haining (1990), Openshaw *et al.* (1990, 91), and Fotheringham and Zhan (1996), among others.

Having noticed the importance of exploratory data analysis in the data rich environment of these days, we develop a user-friendly computer program for detecting significant spatial patters in the distribution of points. Our tool is closely related to Openshaw *et al.* (1987), but it has two different features. First, our tool deals with a more specific cluster patters, i.e., spatial hierarchy; second, the tool employs a statistical method for detecting significant patters of spatial hierarchy.

The concept of spatial hierarchy has been discussed widely in geography, spatial economics, archeology, ecology, and OR since Christallar (1933) and

Figure 1: An ideal pattern of spatial hierarchy.

Figure 2: A distribution of points (a hypothetical test example).

Lösch (1940). As is seen in the papers published in these fields, the concept of spatial hierarchy includes many aspects, such as functional hierarchy, flow hierarchy between nodes, geometrical hierarchy and so forth. In this paper we focus on geometrical hierarchy in the distribution of points where the points represent the locations of point-like objects, such as stores, trees, accidents, and patients. We assume that the points are indifferent expect for their locations (spatial hierarchy with different weights is discussed in Okabe and Sadahiro (1996)). Thus the spatial hierarchy discussed in this paper implies a purely geometric pattern.

To give a clear image of spatial hierarchy, we depict Figure 1 which shows ideal spatial hierarchy, where three levels of spatial hierarchy is indicated by three different sizes of circles. In the real world, however, such clear spatial hierarchy is rarely observed. Rather we meet ambiguous distributions, such as in Figure 2. To examine whether or not there exists spatial hierarchy in this kind of ambiguous distributions, we develop an exploratory tool, called VCM.

The flow chart of the program VCM is shown in Figure 3. The inputs data are: the distribution of points; the shape of a region over which the points are distributed; and a set of parameter values (shown in Section 2). The program VCM runs with PC FORTRAN or Arc/Info AML. The output is a form of significant spatial hierarchy if it exists.

Figure 3: The flow chart of the program VCM.

Figure 4: clumping states for different clump radii.

2 Variable Clumping Method (VCM)

2.1 Clumping states with respect to a clump radius

To detect spatial hierarchy we employ the clumping method (Roach, 1968). The clumping method is a class of methods for finding 'clumps' in the distribution of points, $p_1, \ldots, p_n$, over a bounded region S. Usually a clump is defined in terms of circles centered at given points $p_1, \ldots, p_n$ as in Figure 4. The radius of the circles is called a *clump radius*. A *clump* is then defined as a set of points whose circles are connected. The number of connected circles in a clump is called the *size* of the clump. In the example of Figure 4(b) there are five clumps of size 2, one clump of size 3, and two clumps of size 4 (note that we do not call a clump of size 1 a (proper) clump). We describe the *state of clumping* or *clumping state* of n points in S in terms of a set, $\mathcal{C}(r)$, of the observed numbers, $\hat{n}_i(r)$, of clumps of size $i = 2, 3, \ldots, n$ for a clump radius r, i.e., $\mathcal{C}(r) = \{\hat{n}_i, i = 2, \ldots, n\}$. For example, the clumping state shown in Figure 4(b) is described by $\mathcal{C}(45) = \{\hat{n}_2(45) = 4; \hat{n}_3(45) = 1; \hat{n}_4(45) = 2; \hat{n}_i(45) = 0, i = 5, 6, \ldots\}$. Note that the largest possible clump size occurs when all points form one clump, $i = n$.

In the ordinary clumping method the clump radius r is fixed (we shall call it a *fixed clumping method* to distinguish from the following method). A clumping state, however, varies according to a clump radius r. For example, as is seen in Figure 4, we observe one clump in panel (a): $\mathcal{C}(15) = \{\hat{n}_2(15) = 1; \hat{n}_i(11) = 0, i = 4, \ldots, 30\}$, but seven clumps in panel (b): $\mathcal{C}(45) = \{\hat{n}_2(45) = 4, \hat{n}_3(45) = 1, \hat{n}_4(45) = 2, \hat{n}_i(45) = 0, i = 5, \ldots, 30\}$. If we observe a clumping state for one clump radius r, we see only a lo-

Figure 5: Clumps formed by three different clump radii applied to the ideal pattern of spatial hierarchy shown in Figure 1.

cal pattern. To see a global pattern, we should observe clumping states by varying a clumping radius r. For instance, in the case of the ideal spatial hierarchy shown in Figure 1, we can detect three levels of clumps by applying three different clump radii, $r = 15, 28, 40$, as shown in Figure 5. This fact suggests that a global pattern of the distribution of points can be revealed by a clumping method in which a clumping state $\mathcal{C}(r)$ is observed over a continuum of a clump radius r (from a small clump radius to a large clump radius). We call this clumping method the *variable clumping method* (abbreviated to *VCM*), which should be distinguished from the ordinary clumping method, i.e. the fixed clumping method.

2.2 A computational method for observing clumping states

A clumping state $\mathcal{C}(r)$ with respect to r $(0 < r < \infty)$ can be easily obtained through a Voronoi diagram. To be explicit, let $P = \{p_1, \ldots, p_n\}$ be a set of points distributed over S, and $d(p, p_i)$ be the Euclidean distance between an arbitrary point p in S and p_i. We define a set, $V(p_i)$, of points from which the nearest point in P is p_i, i.e.,

$$V(p_i) = \{p \mid d(p, p_i) \leq d(p, p_j)\, j \neq i, j = 1, \ldots, n\}. \tag{1}$$

We call the set $\mathcal{V}(P) = \{V(p_1), \ldots, V(p_n)\}$ the *Voronoi diagram* generated by P, and $V(p_i)$ the *Voronoi polygon* associated with p_i (a general review of this diagram is provided by Okabe, Boots, Sugihara and Chiu (1999)). Figure 6 shows a Voronoi diagram (note that this is not an ordinary one

Figure 6: A bounded Voronoi.

Figure 7: The Delaunay triangulation obtained from Figure 6.

because it is bounded by a rectangle, i.e., $\{V(p_1) \cap S, \ldots, V(p_n) \cap S\}$; such a Voronoi diagram is called a *bounded Voronoi diagram*).

For $i, j = 1, \ldots, n$ $(i \neq j)$, if $V(p_i)$ and $V(p_j)$ share the common boundary, we join p_i and p_j with a line segment (see, for example, the broken line segment in Figure 6). The line segments generated in this manner for $i, j = 1, \ldots, n (i \neq j)$ form a tessellation as shown in Figure 7. We call this tessellation the *Delaunay triangulation* spanning P, and denoted it by $\mathcal{D}(P)$ (note that since the Voronoi diagram in Figure 6 is a bounded Voronoi diagram, the resulting Delanuany triangulation does not form the convex hull of P).

We now order the edges of $\mathcal{D}(P)$ from the shortest to the longest, and obtain a set, M, of edges through the following procedure.

Step 0. Initialize $M = \emptyset$ and $k = 1$.
Step 1. Choose the kth shortest edge in $\mathcal{D}(P)$ and examine if this edge and the edges in M form a loop. If not, include the kth shortest edge in M and go to Step 2; otherwise, discard the kth shortest edge and go to Step 2.
Step 2. If the kth shortest edge is the longest edge, stop and return M; otherwise, replace k with $k + 1$ and go to Step 1.

The edges in M form a tree, which is well-known as a *minimum spanning tree*. An example is shown in Figure 8 (obtained from Figure 7), where edges

Figure 8: The minimum spanning tree obtained from Figure 7.

Figure 9: Clumping states $\mathcal{C}(r)$ with respect to r obtained from the minimum spanning tree in Figure 8.

are ordered from the shortest to the longest. Let $M(r)$ be a set of edges in M that are shorter than or equal to $r/2$. Then we can completely describe the clumping state $\mathcal{C}(r)$ with respect to r in terms of $M(r)$ for $0 < r < r_{\max}$ where the $r_{\max}$ is the half of the longest edge in $\mathcal{D}(P)$. An example is illustrated in Figure 9.

2.3 Significant clumps

When we apply VCM to the distribution of points, we can always obtain 'insignificant' clumps. Consider, for instance, the case in which we adopt a very long clump radius. Then we always have one clump consisting of all points. Obviously this clump has little implication. Besides this extreme case, it is likely to exist some 'insignificant' clumps in the distribution of points, because clumps may appear even in the distribution of random points. We are not interested in these insignificant clumps; we are interested in 'significant' clumps. The significant clumps may appear when points has the tendency of forming clumps at a certain clumping radius r. If the points has this tendency, we expect that the number of clumps of size i for r is significantly lager than the number of clumps that would appear in the distribution of random points. Thus, if the observed number $\hat{n}_i(r)$ is greater than the number of clumps that would appear in the distribution of random points, we may say that these clumps are *significant clumps*. VCM attempts to find such significant clumps explicitly in the following manner.

Suppose that n points are distributed randomly over a bounded region S. As a result, we obtain the number, $n_i(r)$, of clumps of size i for r.

This number is probabilistic and has a probability function, $f(n_i(r))$. If this function is known, we can obtain a critical number, $n_i^*(r)$, such that the probability of $n_i(r)$ being greater than $n_i^*(r)$ is less than a given significance level α, say $\alpha = 0.05$. If the observed number $\hat{n}_i(r)$ of clumps of size i is greater than $n_i^*(r)$, we may say with significance level α that those clumps are *significant clumps* of size i for r.

To adopt the above statistical test in practice, we have to obtain the function of $f(n_i(r))$ explicitly. In the theory of clumping, the expected value of $n_i(r)$ is approximately obtained (Roach, 1967), although this approximation is not always satisfactory. In the above statistical test, however, the expected value does not help. VCM needs the probability function $f(n_i(r))$. In the related literature few analytical methods are found. We suspect that the analytical method is intractable because we should treat a bounded irregular region. In VCM, hence, Monte Carlo simulation is adopted. VCM places n points randomly over S for 10000 times, and obtain 10000 minimum spanning trees $M(r)$. From these trees, VCM obtains the frequency distribution $f(n_i(r))$ of $n_i(r)$, from which VCM obtains the critical numbers $n_i^*(r)$. Note that in practice, we use $r = r_j = r_o j, j = 1, \ldots n_r (r_o n_r = r_{\max})$. An example is shown in Table 1. This table says, for instance, that when $r = 112.1$, if the observed number $\hat{n}_2$ is more than 3, we may say that such clumps are significant clumps; if the observed number $\hat{n}_i$ $i = 3, 4, 5$ is more than 1, we may say that such clumps are significant clumps.

Table 1: Critical numbers $n_i^*(r_j)$ of clumps with respect to a clump size $i = 1, \ldots, 10$ and a clump raidus r_j, $j = 1, \ldots, 16$ in a 1000m by 1000m square.

r_j	$i = 2$	3	4	5	6	7	8	9	10
0.00	0	0	0	0	0	0	0	0	0
28.02	0	0	0	0	0	0	0	0	0
56.05	2	1	0	0	0	0	0	0	0
84.07	3	1	1	0	0	0	0	0	0
112.10	3	1	1	1	0	0	0	0	0
140.12	3	2	1	1	1	0	0	0	0
168.15	3	2	1	1	1	1	1	0	0
196.17	2	1	1	1	1	1	1	1	0
224.20	2	1	1	1	1	1	1	1	1
252.22	1	1	1	1	1	1	1	1	1
280.24	1	1	0	0	1	1	1	1	1
308.27	1	0	0	0	0	1	1	1	1
336.29	1	0	0	0	0	0	1	1	1
364.32	0	0	0	0	0	0	1	1	1
392.34	0	0	0	0	0	0	0	1	1
$\geq$ 420.37	0	0	0	0	0	0	0	0	1

We should remark one important property. If we observe significant clumps for r_i and r_j $(r_i < r_j)$, then the clumps for r_i are included in clumps for r_j. In this sense, the clumps for r_i and those for r_j form *successively inclusive spatial hierarchy*.

2.4 Procedure for running VCM

Having shown the theory of VCM, we now show a practical procedure for running VCM.

Step 0 (Initial setting). We assume that the region S is represented by a polygon, and the polygon is described by a series of the coordinates of

the vertices of S. The input data are: the coordinates of these vertices, the number, n, of points, the interval, r_o, of a variable clump radius and the level of significance α.

Step 1 (Computation of the observed numbers $\hat{n}_i(r_j)$ of clumps). For a given distribution of n points, VCM counts the observed number, $\hat{n}_i(r_j)$, of clumps of size i with respect to clump size $i = 1, \ldots, n$ and a clump radius, r_j, $j = 1, \ldots, n_r$.

Step 2 (Computation of the critical number $n_i^*(r)$ of clumps by Monte Carlo simulation). VCM generates n random points over S and counts the number, $n_i(r_j)$, of clumps with respect to $i = 1, \ldots, n$ and $r_j, j = 1, \ldots, n_r$. VCM carries out this trial for 10000 times. From the 10000 trials, VCM obtains $f(n_i(r_j))$, which gives the critical number, $n_i^*(r_j)$, of clumps with respect to $i = 1, \ldots, n$ and $r_j, j = 1, \ldots, n_r$. The results are stored in a table, like Table 1.

Step 3 (Statistical detection). By comparing the observed numbers $\hat{n}_i(r_j)$ obtained in Step 1 and the critical numbers $n_i^*(r_j)$ obtained in Step 2, VCM detects which clumps are significant, and the significant clumps are visually shown with GIS.

3 Concluding Discussion

To test the usefulness of VCM, we applied VCM to a number of hypothetical as well as actual examples. The results were almost satisfactory. A typical result is shown in Figure 10. Human eyes hardly notice spatial hierarchy in Figure 2, but VCM reveals significant spatial hierarchy as shown in Figures 10(a), (b) and (c).

Figure 10: Spatial hierarchy detected by VCM.

In testing VCM, however, we noticed some arguable problems. First let us discuss computational time. The major geometrical computation in VCM is the construction of a Voronoi diagram. VCM employs the program developed by Sugihara and Iri (1989), called VORONOI2. The average computational time of the algorithm adopted in this program is of the order, $O(n)$, of the number n of points. In the worst case (i.e. points are locally clustered), the algorithms runs with order $O(n^2)$, but since VCM generates uniform random points, the linear time $O(n)$ is realized. In practice, however, computational time hinges on 10000 Monte Carlo trials, and so the total computational time is of order of $10000n$.

This fairly high order computational time results from that fact that VCM treats an arbitrary irregular region S and an arbitrary number n of points. We can drastically reduce the computational time if the region S is assumed to be a unit square and the critical numbers $n_i^*(r_j)$ are obtained by interpolation from the critical numbers obtained for $n = 10, 20, 50, 100, 200, 500$ (an example of $n = 10$ is shown in Table 1). In this case, we compute the critical values in advance and hence the computation asked by a user is just to interpolate. We are now adding this simplified function to VCM, which will be useful when quick analysis is necessary at the risk of rough analysis.

One might question the boundary effect as is often discussed in this kind of statistical tests. Since VCM generates random points over a given region S, the boundary effect is exactly taken into account. This is an advantage of VCM.

Spatial hierarchy revealed by VCM may be arguable. As noted, significant clumps of r_1 is included in significant clumps of r_2 if $r_2 > r_1$. It should be

note, however, that the converse is not true. A significant clump of r_2 may not include significant clumps of $r_1 < r_2$ (observe clumps near the left boundary in Figure 10).

Although VCM has some limitations to be improved, we consider from our experiments that VCM is a practically useful explanatory tool for detecting spatial hierarchy in the distribution of points.

Last we note that VCM is open to public. At present, however, the manual of VCM is written in Japanese. We are planning to translate it in English in near future.

References

Anselin, L., Dodson, R.F. and Hudak, S. (1993) Linking GIS and spatial data analysis in practice, *Geographical Systems*, **1**, 3-23.

Christaller, W. (1933) *Die Zentralen Orte in Süddeutschland* Jena: Fisher.

Fotheringham, A.S. and Zhan, F.B. (1996) A comparison of three exploratory methods for cluster detection in spatial point patterns, *Geographical analysis*, **28**, 200-218.

Goodchild, M. (1987) A spatial analytical perspective on Geographical Information Systems, *International Journal of Geographical Information Systems*, **1**, 327-334.

Goodchild, M., Haining, R. and Wise, S. (1992) Integrating GIS and spatial data analysis: problems and possibilities, *International Journal of Geographical Information Systems*, **6**, 407-423.

Haslett, J., Wills, G. and Unwin, A. (1990) SPIDER-an interactive statistical tool for the analysis of spatially distributed data, *International Journal of Geographical Information Systems*, **4**, 285-296.

Lösch, A. (1940) *Die Raumliche Ordnung der Wirkschaft* Jena: Fisher.

Okabe, A., Boots, B., Sugihara, K. and Chiu, S. N. (1999) *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams* (2nd edition), Chichester: John Wiley.

Okabe, A. and Sadahiro, Y. (1996) An illusion of spatial hierarchy: spatial hierarchy in a random configuration, *Environment and Planning A*, **28**, 1533-1552.

Openshaw, S., Cross, A. and Charlton, M. (1990) Building a prototype geographical correlates exploration machine, *International Journal of Geographical Information Systems*, **4**, 297-311.

Openshaw, S., Brunsdon, C. and Charlton, M. (1991) A spatial analysis toolkit for GIS, *European Conference on Geographical Information Systems* , 788-796.

Roach, S.A. (1968) *The Theory of Random Clumping*, London: Methuen.

Sugihara, K. and Iri, M. (1989) VORONOI2 Reference Manual. (unpublished)

Tukey, J.W. (1977) *Exploratory Data Analysis*, Massachussetts: Addison-Wesley.

Walker, P.A. and Moore, D.M. (1988) SIMPLE: an inductive modelling and mapping tool for spatially-oriented data, *International Journal of Geographical Information Systems*, **2**, 347-363.