10

Clustering in Time and Space

In this chapter, you learn how to detect clustering in time and space and to validate clustering models. You use the generalized quadratic form in its several guises including Mantel's U and Mielke's multiresponse permutation procedure to work through a series of applications in atmospheric science, epidemiology, ecology, and archeology.

10.1 The Generalized Quadratic Form

10.1.1 Mantel's *U*

Mantel's U [Mantel, 1967] $\sum\sum a_{ij}b_{ij}$ is perhaps the most widely used of all multivariate statistics. In Mantel's original formulation, a_{ij} is a measure of the time or temporal distance between items i and j, while b_{ij} is a measure of the spatial distance. As an example, suppose the pair (t_i, l_i) represents the day t_i on which the ith individual in a study came down with cholera and $l_i = (l_{i1}, l_{i2})$ denotes her position in space. For all i, j, set $a_{ij} = 1/(t_i - t_j)$ and

$$b_{ij} = 1/\sqrt{(l_{i1} - l_{j1})^2 + (l_{i2} - l_2)^2}$$

A large value for U would support the view that cholera spreads by contagion from one household to the next. How large is large? As always, we compare the value of U for the original data with the values obtained when we fix the i's but permute the j's as in $U' = \sum\sum a_{ij}b_{i\pi(j)}$.

10.1.2 An Example

An ongoing fear among many parents is that something in their environment—asbestos or radon in the walls of their house, or toxic chemicals in their air and ground water—will affect their offspring. Table 10.1 is extracted from data collected by Siemiatycki and McDonald [1972] on congenital neural tube defects.

Table 10.1. Incidents of pairs of anencephalic infants by distance and time months apart.

km apart	<1	<2	<4
<1	39	101	235
<5	53	156	364
<25	211	652	1516

Eyeballing the gradient along the diagonal of this table, one might infer that births of anencephalic infants occur in clusters. One could test this hypothesis statistically using the methods of Chapter 8 for ordered categories, but a better approach, since the exact time and location of each event is known, is to use Mantel's U. The question arises as to which measures of distance and time we should employ. Mantel [1967] reports striking differences between one analysis of epidemiologic data in which the coefficients are proportional to the differences in position and a second approach (which he recommends) to the same data in which the coefficients are proportional to the reciprocals of these differences.[1] Using Mantel's approach, a pair of infants born 5 km and three months apart contribute $(1/3)(1/5) = 1/15$ to the correlation. Summing the contribution from all pairs, then repeating the summing process for a series of random rearrangements, Siemiatycki and McDonald conclude that the clustering of anencephalic infants is not statistically significant.

10.2 Applications

By appropriately restricting the values of a_{ij} and b_{ij}, the definition of Mantel's U can be seen to include several of the standard measures of correlation including those usually attributed to Pearson, Pitman, Kendall, and Spearman [Hubert, 1985]. Mantel's U has been rediscovered frequently, often without proper attribution (see Whaley [1983]). In this section we consider three diverse approaches to the problem of assessing the presence of clustering in space and time. In each case, the permutation distribution of the quadratic form is used to provide a baseline against which the behavior of the observations may be assessed.

10.2.1 The MRPP Statistic

One such variant is the MRPP or *multiresponse permutation procedure* [Mielke, 1979], which is used in applications as diverse as weather and the spatial distribution of archaeological artifacts. The MRPP uses the

[1] One further caveat: Mantel's U fails completely if the spatial distribution of the underlying population is also changing with time [Roberson and Fisher, 1986].