Space-Time Clustering Revisited

Gilbert MacKenzie\textsuperscript{1} and, Jing Xu\textsuperscript{1}

\textsuperscript{1} Centre of Biostatistics, Department of Mathematics & Statistics, University of Limerick, Ireland

Abstract: The history of space-time clustering concepts and methods are reviewed briefly. The space-time clustering model of Ederer \textit{et al} is investigated in detail. This method has been used extensively in the epidemiological literature, but we show that the distribution of main test statistic involved does not follow the distribution proposed by the authors. We note, too, that the two indices proposed are not statistically independent, leading to potential over-reporting in the epidemiological literature. We obtain the correlation between the original clustering indices and suggest a new combined test statistic which has the correct null distribution. We develop a fuller spatial model and illustrate the methodology using data from a study of the incidence of childhood leukaemia in Northern Ireland.

Keywords: space-time clustering, exact distribution, correlated outcomes, spatial data, covariance model

1 Introduction

We review, briefly, the history of space-time clustering concepts and methods, ranging from Karl Pearson’s now celebrated 1913 paper to modern times.

In particular we focus on a space-time clustering technique first introduced by Ederer \textit{et al} who studied, \textit{inter alia}, the spatial-temporal distribution of childhood leukaemia. Typically, this method, which has been used extensively in the epidemiological literature, relies on the creation of two indices of clustering $Y_1$, the maximum number of observed cases in any single basic space-time unit, and $Y_2$ the maximum number of observed cases in any two time-contiguous basic units.

First, we demonstrate that that distribution of $Y_1$, on the null hypothesis of no clustering, does not follow the null distribution proposed by the original authors. Second, we note that the two indices are not statistically independent. Consequentially, reports in the epidemiological literature to date may have been subject to over reporting. Third, using the exact distribution we compute the correlation between $Y_1$ and $Y_2$ for important cases which have appeared in the literature and discuss the problem of selecting a suitable bivariate test statistic. A new index, $U$, which is asymptotically
distributed as $\chi^2_2$ is proposed. Its distribution in small samples is investigated by means of a simulation study and a table of critical values is provided. Lastly, we consider how to develop a spatial model for the observed bivariate discrepancy in $Y_1$ & $Y_2$ from the null distribution. The new methodology is illustrated using data from the Northern Ireland Study of Patterns of Disease and Radiation (PODAR, 1989).

2 Basic Model Formulation

2.1 Space-Time Units

The basic idea is to detect unusual clusters of cases of a disease in space and time. A cluster occurs when some cases are distributed more closely in space and time than expected on the basis of some chance rule. The method of Ederer et al requires the creation of space-time units. A basic space-time unit is a defined spatial region studied for a unit of time. The individual spatial regions are referred to as spatial units and the units of time are referred to as temporal units. A study space-time unit is constructed by studying a spatial unit for several units of time, eg a region for say five years. It is assumed that the number of cases of disease can be determined in each year. Accordingly, let $m=$ the total number of spatial units, $k=$ the total number of temporal units and $N = m \times k$ be the total number of basic space-time units. With these arrangements we shall have $m$ study space-time units each studied for $k$ years. Ashitey & MacKenzie (1970) present a simple application.

2.2 Clustering Indices

Now, let $n_{ij}$ be the number of cases of disease in the $i$th spatial unit in temporal unit $j$: $i = 1, \ldots, m$ and $j = 1, \ldots, k$. And finally let $n_{i+} = \sum_j n_{ij}$ be the marginal total number of cases for the $i$th space-time unit, whence we have an occupancy distribution $(n_{i1}, n_{i2}, \ldots, n_{ik})$ with exactly $k$ compartments. Ederer et al define two observed indices of clustering within the $i$th space-time unit, viz:

$$y_{i1} = \text{maximum no. of cases in any 1 temporal unit}$$

$$y_{i2} = \text{maximum no. of case in any 2 adjacent temporal units}$$

These are just two possibilities.

2.3 Probability Functions

The distribution of the corresponding random variables, $(Y_{i1}, Y_{i2})$ and moments can be found in principle from:

$$Pr(n_{i1}, n_{i2}, \ldots, n_{ik}) = n_{i+}! \prod_j \theta_j^{n_{ij}} / \prod_j n_{ij}!$$

(2)
(Feller, 1957) where $\theta_j$ probability of falling into the $j$th compartment. On the null hypothesis of no clustering $\theta_j = \theta = 1/k, \forall j$ and, of course, there are many potential generalizations.

3 Test Statistics

Ederer et al also define two continuity-corrected test statistics based on their indices, viz:

$$U_1 = \left| \sum_i Y_{i1} - E\left(\sum_i Y_{i1}\right) \right| - 0.5^2 / V\left(\sum_i Y_{i1}\right)$$

$$U_2 = \left| \sum_i Y_{i2} - E\left(\sum_i Y_{i2}\right) \right| - 0.5^2 / V\left(\sum_i Y_{i2}\right) \quad (3)$$

and claim that they are distributed asymptotically as $\chi^2$ random variables. We show, via an extensive simulation study, that this claim, in relation to $U_1$, the most extensively quoted index in the epidemiological literature, is untenable. Moreover, in the main paper we propose a relatively simple solution based on a test statistic, $U$, which respects the bivariate distribution of $(Y_{i1}, Y_{i2})$.

4 Presumed Independence

To date, we note, that routine usage of $(U_1, U_2)$ has tacitly assumed that $Pr[(Y_{i1} \cap Y_{i2})] = Pr(Y_{i1}) \times Pr(Y_{i2})$. However, we shall establish that $(Y_{i1}, Y_{i2})$ is approximately $BVN(\mu, \Sigma)$ and that $\Sigma$ is not a diagonal. In particular, below, we present in Table 1 the hitherto unpublished correlation between $(Y_{i1}, Y_{i2})$, for a range of marginal case numbers ($y_{i+}$) and different numbers of temporal units ($k$).

<table>
<thead>
<tr>
<th>$y_{i+}$</th>
<th>$k=3$</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.38</td>
<td>0.44</td>
<td>0.48</td>
<td>0.50</td>
<td>0.51</td>
<td>0.52</td>
<td>0.53</td>
</tr>
<tr>
<td>3</td>
<td>0.61</td>
<td>0.61</td>
<td>0.60</td>
<td>0.58</td>
<td>0.57</td>
<td>0.57</td>
<td>0.56</td>
</tr>
<tr>
<td>4</td>
<td>0.47</td>
<td>0.61</td>
<td>0.64</td>
<td>0.64</td>
<td>0.63</td>
<td>0.63</td>
<td>0.62</td>
</tr>
<tr>
<td>5</td>
<td>0.51</td>
<td>0.57</td>
<td>0.61</td>
<td>0.61</td>
<td>0.65</td>
<td>0.65</td>
<td>0.65</td>
</tr>
<tr>
<td>6</td>
<td>0.52</td>
<td>0.58</td>
<td>0.60</td>
<td>0.60</td>
<td>0.64</td>
<td>0.65</td>
<td>0.65</td>
</tr>
</tbody>
</table>
Broadly speaking, the correlation is non-trivial and this suggests that it should be taken into account when conducting tests using $(U_1, U_2)$.

5 Extensions

The move to a bivariate representation also affords us the possibility of defining useful residuals with which to explore departures from the null hypothesis of no clustering. When this hypothesis is rejected, the residuals may be exploited further to examine, in detail, the spatial distribution of this discrepancy, thereby extending, considerably, the scope of Ederer’s original scheme. These ideas will be developed fully in the main paper.

6 Final Remarks

In summary, we have revisited the space-time clustering technique proposed by Ederer et al (1964), identified some methodological flaws and suggested a number of remedies which generalize the method.

Acknowledgments: This work was supported by Science Foundation Ireland (SFI) as part of their funded BIO-SI programme (www.ul.ie/bio-si), Mathematics Initiative, II. Dr. Jing Xu is a SFI funded Post-Doctoral Research Fellow

References


