Categorical data with inherent spatial dependence: the case of cluster sampling

BERNARD FINGLETON

Senior Lecturer, School of Geography, Cambridgeshire College of Arts and Technology, Cambridge CB1 1PT

Revised MS received 2 June, 1988

ABSTRACT

This paper introduces a statistic appropriate to the analysis of categorical spatial data obtained as a result of cluster sampling. The paper outlines the relations between this and other statistics suggested for complex sample designs, particularly those developed for cluster sampling and for spatial data obtained as a result of systematic sampling.

KEY WORDS: Chi-squared, Log-linear model, Spatial autocorrelation, Categorical spatial data, Complex sample design, Cluster sampling

INTRODUCTION

Geographers who are involved in the design of surveys may opt for a complex sampling scheme because it offers a more practicable and cost effective approach to data collection than is provided by simple random sampling (SRS). A major consequence, however, of the choice of sample design, is the fundamental effect that it has on the type of data analysis that is valid, and whilst there are well established statistical methods for data that have been derived via SRS, valid procedures for data arising from some form of complex sample design are either more difficult to administer, or are unknown. This paper is focused on the methodology appropriate to categorical spatial data that has been obtained using the simplest form of 'complex' sample design, namely cluster sampling. The main intention in the paper is to suggest an appropriate statistic, equation (6), for the analysis of such data. The paper also focuses on the close relation between this new statistic and other methods for handling complex (spatial) sample designs.

CLUSTER SAMPLING—THE PROBLEM OF DEPENDENCE

Since cluster sampling involves the selection of clusters of units of observation, it is potentially more cost effective than SRS. For example, given k clusters each of size C, say, the gathering of a sample of size kC involves travelling to k separate destinations as opposed to kC destinations. Clusters might be a random selection from among the enumeration districts of a city, and the data may comprise all of the households present in these chosen districts. Alternatively, in a spatial sampling setting, a field scientist may choose sample sites by first imposing a number of separate grids (clusters) on the surface of the earth, and then by sampling exhaustively at or near the grid intersection points within each cluster. Indeed, the latter approach has also been suggested in a social survey context by Moser and Kalton (1971) who note that sometimes 'one makes up artificial clusters by imposing grids onto maps'. Figure 1 illustrates this in the form of three representative clusters of 10 *km* squares.

Although, for a given amount of expenditure, one should in principle be able to gather a larger sample by cluster sampling than by simple random sampling, this can be a false economy whenever the data lose the independence that is the cornerstone of classical statistical methodology. With spatial data, dependence may be induced as a result of the too-close proximity of trials which are located within the same cluster, and this means, in effect, that each observation is worth less than one independent

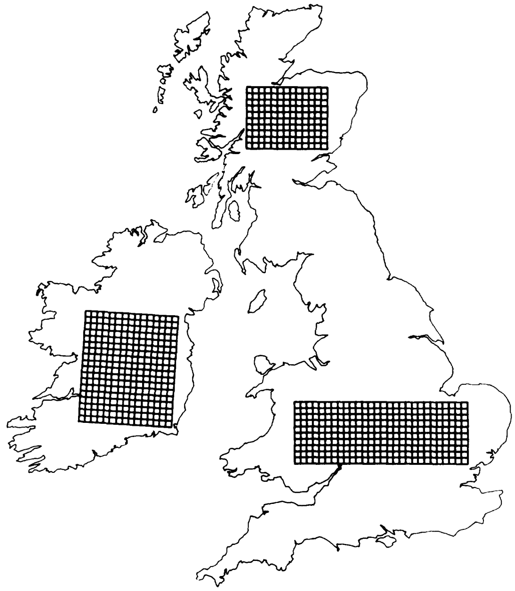

FIGURE 1. Spatial clusters

observation, thus reducing the effective sample size. A conventional chi-squared test of independence (which is a particularly simple form of log-linear model), or the fitting of other standard log-linear models, enhances the possibility that variables will appear to be significantly associated when, in fact, the association may be entirely spurious. A similar phenomenon has been described in detail for systematic sampling by Fingleton (1983a,b, 1986) and there is a close analogy in spatial regression analysis. An important presumption here is that the data are positively autocorrelated.

THE APPROPRIATE METHODOLOGY FOR CLUSTER SAMPLING

In order to set the scene for the new statistic suggested in this paper, assume that the number of units within each of k clusters is equal to C, and that the intention is to carry out a chi-squared test of the association between two categorical variables labelled A and B with I and J categories respectively and with observed frequencies $\{o_{ij}\}$ and with expected frequencies $\{e_{ij}\}$ under the null hypothesis of the independence of A and B. Altham (1976) has noted that the familiar statistic

$$X^2 = \sum_i \sum_j (o_{ij} - e_{ij})^2 / e_{ij} \tag{1}$$

should be corrected and that the appropriate corrected statistic is

$$X^2/[1+(C-1)a] \tag{2}$$

which is distributed as χ^2 with $(I-1)$ $(J-1)$ degrees of freedom under the null hypothesis. In equation (2), the parameter a measures the strength of the correlation between the observations within each cluster. One problem with this approach is, of course, that a is usually unknown and its estimation may be difficult. One simpler option is to assume the maximum amount of positive correlation by setting a equal to 1, in which case the statistic

$$X^2/C \tag{3}$$

is referred to the χ^2 distribution. In fact equation (3) gives the lower bound of the range of possible values for the true, but unknown, statistic with an approximate χ^2 distribution under the null hypothesis. As a result, if the lower bound turns out to be significant when referred to the χ^2 distribution with the appropriate degrees of freedom, one may assume that the true statistic is also (see Altham 1976, 1979; Fingleton 1983a,b). Whilst such an approach is commendable because of its simplicity, it possesses the disadvantage of being highly conservative and therefore of demanding a great deal of evidence before the analyst can infer that variables A and B are associated. In effect, the use of equation (3) amounts to treating each cluster as a single unit of observation, and this can hardly be satisfactory. When the clusters are of unequal size, then the corrected statistic becomes

$$X^2/[1+(N_2-N_1)a/N_1] \tag{4}$$

in which

$$N_1 = \sum_s^k m_s n_s \text{ and } N_2 = \sum_s^k m_s^2 n_s^2$$

and $m_s n_s$ denotes the size of the s'th cluster. Equation (4) is a version of a result given by Cohen (1976) (cf. Rao and Scott, 1981; Brier, 1980) in which the notation has been adapted for the purpose of spatial analysis, as will become apparent in the subsequent section.

The simplest solution to the problem posed by the fact that a is unknown is to again assume the maximum amount of intracluster correlation ($a = 1$) in order to make the rejection of the null hypothesis as difficult to achieve as is possible, and this leads to the highly conservative statistic

$$X^2/\{N_2/N_1\} \tag{5}$$

which reduces to X^2/C with $m_s n_s$ constant. More details of these and related methods are given in Fingleton (1984) and by Rao and Scott (1984, 1985).

A CORRECTED STATISTIC ACCOMMODATING INTRA- AND INTER-CLUSTER VARIATION IN SPATIAL AUTOCORRELATION

The methods described above do not acknowledge the potential influence of location on the level of autocorrelation in the data. Given spatial data, it is likely that the dependence (represented by the parameter a above) will vary with distance. Moreover, the level of dependence at any given distance may differ according to the cluster considered. To accommodate these variations, we introduce the set of autocorrelation parameters $\{a_{rs}\}$, in which r denotes the distance and s denotes the cluster, and this is central to the corrected statistic

$$X^2/[1+2\{\sum_s^k \sum_r^d N(r,\ m_s,\ n_s)\ a_{rs}\}/\sum_s^k m_s n_s] \tag{6}$$

in which $N(r,\ m_s,\ n_s)$ is the exact number of pairs of sites in cluster s which are distance r apart (using the city block metric) on the m_s by n_s lattice. The method for the evaluation of this number is given in Appendix I (cf. Fingleton, 1986). Given that the a_{rs} s faithfully reflect the spatial variation in the level of dependence, then the null distribution of the statistic (6) is approximately as a χ^2 random variable. This is to be preferred to conservative statistics that do not attempt to reflect the autocorrelation with any degree of verisimilitude.

Of course, the operationalisation of the corrected statistic (6) depends on clusters consisting of trials laid out as lattices, for instance as in Figure 1, whereas the corrected X^2 statistics that have been presented above can be applied to a haphazard distribution. Nevertheless, there is a close connection between the statistic (6) and the (aspatial) corrected statistics. In particular, we observe that

$$N_2 - N_1 = 2\sum_s^k \sum_r^d N(r,\ m_s,\ n_s) \tag{7}$$

in which m_s is the number of rows and n_s is the number of columns in cluster s, and

$$d = max\{(m_s - 1) + (n_s - 1);\ s = 1,2, \ldots k\} \tag{8}$$

the maximum possible distance on the largest lattice. The two statistics (4) and (6) are equivalent if $\{a_{rs}\} = a$ for all r and s.

It is important to emphasise that the corrected statistic (6) depends on a particular model being used to represent the within-cluster dependence (cf. Altham, 1976; Cohen, 1976; Tavaré and Altham, 1983). The model is

$$P_{rs}(u,v) = p_u\ \{a_{rs}\, g_{uv} + (1 - a_{rs}) p_v\} \tag{9}$$

in which $P_{rs}(u,v)$ is the probability that category combinations u and v will occur at distance r in cluster s, and $g_{uv} = 1$ if $u = v$ and $g_{uv} = 0$ otherwise. The symbols u and v refer to any two cells or category combinations of the IJ cross-classification table. The category combination u is located at i_1,j_1 within the cluster and v is located at i_2,j_2. In equation (9), the probabilities p_u and p_v are the probabilities of table cells u and v as given, for example, by a log-linear model. The simplest way in which these can be obtained is under the independence model for two categorical variables. Note that $a_{rs} = 1$ is consonant with maximum positive spatial autocorrelation, and this leads to $P_{rs}(u,v) = p_u$ when $u = v$ and to $P_{rs}(u,v) = 0$ when $u \neq v$. An absence of spatial autocorrelation is equivalent to setting $a_{rs} = 0$ for all r and all s; in this case the uncorrected statistic (1) can, with the normal provisos, be referred to the χ^2 distribution. Negative autocorrelation is also possible under the model, but a_{rs} must assume a value at least as large as the maximum of either $-p_u/(1-p_u)$ or $-(1-p_u)/p_u$ in order to avoid negative probabilities. The relation between model (9) and the statistic (6) is outlined in more detail in Appendix II.

SPATIAL DATA, CLUSTERS, AND SYSTEMATIC SAMPLING

The previous two sections relate to cluster sampling and the close connection seen to exist between spatial and aspatial corrected X^2 statistics. In this section we show that a close connection exists between these statistics and those appropriate to systematic sampling.

Consider equation (6) and assume that the clusters are all equal ($m_s = m$ and $n_s = n$) and that the strength of the dependence on the lattice is not cluster specific, though it does still vary with distance, so that $a_{rs} = a_r$. It is then possible to show that the appropriately corrected statistic is

$$X^2/[1+2\{\sum_r^d N(r,m,n)a_r\}/mn] \tag{10}$$

and that precisely the same correction applies to a systematic sample which is equivalent to a single cluster from Figure 1 (see Appendix III). Equation (10) is discussed in the context of systematic sampling in Fingleton (1986). Assume (for the sake of simplicity) that a is equal to one irrespective of distance. It is then the case that the corrected statistic

$$X^2/[1+2\{\sum_r^d N(r,m,n)\}/mn] \tag{11}$$

is appropriate either to k clusters each consisting of m by n lattices or to a single systematic sample consisting of one m by n lattice.

If the lattice distance d is less than $(m+n-2)$ and is, in fact, very small compared with m and n, then equation (10) is approximated by

$$X^2/[1+\sum_r^d 4ra_r] \tag{12}$$

and equation (11) is approximated by the statistic

$$X^2/[1+2d(d+1)] \tag{13}$$

which was derived by Fingleton (1983a, 1983b). Both equations (11) and (13) possess the advantage of fairly minimal assumptions regarding the spatial dependence.

If the assumption that distance r is measured in terms of the city block metric, is replaced by an assumption that diagonally adjacent sites are at distance 1 (rather than distance 2), as in Figure 2b, then equation (12) becomes

$$X^2/[1+\sum_r^d 8ra_r] \tag{14}$$

and if only every alternate cell of the matrix is sampled, in the manner of a chessboard, then a_r equals zero whenever r is an odd number and distance is in terms of the city block metric.

a) city block distances

```
                4
            4   3   4
        4   3   2   3   4
    4   3   2   1   2   3   4
4   3   2   1   *   1   2   3   4
    4   3   2   1   2   3   4
        4   3   2   3   4
            4   3   4
                4
```

b) an alternative distance metric

```
4   4   4   4   4   4   4   4   4
4   3   3   3   3   3   3   3   4
4   3   2   2   2   2   2   3   4
4   3   2   1   1   1   2   3   4
4   3   2   1   *   1   2   3   4
4   3   2   1   1   1   2   3   4
4   3   2   2   2   2   2   3   4
4   3   3   3   3   3   3   3   4
4   4   4   4   4   4   4   4   4
```

FIGURE 2. Two alternative distance measures on the lattice

Some further issues of practical importance

So far little comment has been made regarding the (intracluster) correlation between trials located at pairs of sites at distance r, which we have denoted by a_r. One can either assume that a_r takes some value, or one can strive to estimate its value. We have stated already that estimation is difficult and that some progress can be made by simply assuming that a_r always takes the value 1 irrespective of the distance r. A slightly less conservative statistic is provided if the assumption is made that a_r equals some value less than 1. It is likely in reality that the intersite correlation will vary, and therefore this latter value should be the maximum assumed for this range of values. It is also possible to assume that the strongest intersite correlation will occur for neighbouring sites and that this will decline subsequently. We can make the assumption that this decline in autocorrelation with distance is a smooth process, so that the whole set of correlations over all distances up to d might be the outcome of a simple autoregressive process, in which case the corrected statistic which is comparable to equation (10) is

$$X^2/[1+2\{\sum_{r}^{d} N(r,m,n)\, a_1^r\}/mn] \tag{15}$$

where a_1^r is the r'th power of the autocorrelation at distance $r=1$. An approximation to this is provided by

$$X^2/[1+4\{a_1+2a_1^2+\ \ldots da_1^d\}] \tag{16}$$

so long as d is much less than both of m and n. This simplifies to equation (13) when a_1 takes the value 1. Fingleton (1986) and Fingleton and Porteous (1985) suggest how the values of a_r at different distances on a lattice might be estimated.

An important discovery has recently been made by Porteous (1985) who shows that the conventional chi-squared test statistic is more robust than was hitherto suggested. Working in the context of equations (11) and (13) and systematic sampling, Porteous shows that if the spatial autocorrelation for each of the n categorical variables is considered individually, then at least a pair of variables must be individually autocorrelated for there to be any consequence for the distribution of the test statistic. Therefore, given three variables on a lattice, d need only be equal to the distance beyond which only one of the variables continues to display autocorrelation. One consequence of this is that with only two variables, each of them should be individually autocorrelated for the corrected statistic to be needed. If one of the pair of variables is independent, then there is no need to correct equation (1) for autocorrelation, and if one is autocorrelated to distance d_y on the lattice and the other to distance d_x, where $d_y > d_x$, then the value of d should be set equal to d_x.

CONCLUSION

This paper suggests a method appropriate to the analysis of categorical spatial data obtained as a result of sampling using clusters in which there is variation in spatial autocorrelation by distance between trials within a cluster and by cluster. This is seen to be closely related to methods set within a systematic sampling scheme for categorical spatial data and to the evolving general methodology for categorical data and complex sample designs. The motivation for this work, from the geographer's perspective, is the need for a fuller integration of spatial effects into the family of generalised linear models, following the lead given by the development of spatial regression models (see Cliff and Ord, 1981; Upton and Fingleton, 1985). As Wrigley (1985 p. 309) has observed, in relation to work that has already been published on the analysis of categorical spatial data, 'if this initial research can be built upon, the next decade could see an immensely valuable integration of the spatial dependence, categorical data analysis, and sampling survey literature in which quantitative geographers and environmental scientists can make a distinctive contribution'.

APPENDIX I

The value of $N(r,m_s,n_s)$is given by

$r\{2m_sn_s-(m_s+n_s)r+(r^2-1)/3\}$
  if $r<=\min(m_s,n_s)$; by
$b\{2m_sn_s-(m_s+n_s)b+(b^2-1)/3\}-(r-b)b^2$
  if $r<=\max(m_sn_s)$ and $r>\min(m_sn_s)$; and by
$(b-r+t)\{(b-r+t)^2-1\}/3$
  if $r>\max(m_sn_s)$,

where $b=\min(m_s,n_s)$ and $t=\max(m_s,n_s)$ and r is the distance on the grid measured by the city block metric, hence

$$r=|i_1-i_2|+|j_1-j_2|$$

is the distance between sample sites located at i_1,j_1 and i_2,j_2 within cluster s.

APPENDIX II

In this Appendix it is convenient for notational purposes to work in terms two categorical variables, with I categories and J categories respectively, and the I by J table of probabilities, $\{p_{ij}\}$, where $\sum_i\sum_j p_{ij}=1$.

Assuming that $(p-E(p))(mn)^{\frac{1}{2}}$ is distributed as $N(\mathbf{O},\mathbf{V})$ and that the overall size of the (large) sample is mn, it is known that the null distribution of the generalized Wald statistic

$$X_w^2=mn\,\mathbf{h}(p)'(\mathbf{H}\,\mathbf{V}\,\mathbf{H}')^{-1}\mathbf{h}(p)$$

is approximately as χ^2 with, in this case, $(I-1)$ $(J-1)$ degrees of freedom. The matrix $\mathbf{V}/(mn)$ is a variance-covariance matrix (for the $\{p_{ij}\}$ estimates) which incorporates the spatial dependence in the data and $\mathbf{p}$ is a column vector with elements $p_{11}, p_{12}, \ldots, p_{IJ}$.
In the Wald statistic,

$$\mathbf{h}(p)=(h_{11}(p),h_{12}(p),\ldots,h_{(I-1)(J-1)}(p))$$

is a column vector in which the elements are

$$h_{ij}(p)=p_{ij}-p_{io}\,p_{0j}$$

where

$$p_{io}=\sum_j p_{ij} \text{ and } p_{0j}=\sum_i p_{ij}.$$

These can be thought of as residuals since p_{ij} is the actual probability of categories i and j of the two variables and $p_{i0}p_{0j}$ is the expected probability under the independence model. The matrix **H** is an $(I-1)$ $(J-1)$ by IJ matrix of partial derivatives

$$\mathbf{H}(p) = \partial \mathbf{h}(p)/\partial p.$$

with typical element

$$\partial p_{ij}/\partial p_{ij} - p_{0j}\partial p_{i0}/\partial p_{ij} - p_{i0}\partial p_{0j}/\partial p_{ij}.$$

Unfortunately, the Wald statistic is likely to prove difficult to calculate in practice, and Wald type statistics tend to be unstable due to the inverse of (**HVH′**) (Rao, 1986). It does, however, provide the basis of a more simple to achieve corrected statistic based on the usual chi-squared test statistic X^2. Unlike the Wald statistic, the conventional Pearson chi-squared test statistic X^2 (equation (1) above) requires independent observations as a necessary condition for an approximation to the χ^2 distribution. Its more general distribution has been shown by Holt *et al.* (1980) to be a weighted sum of independent χ^2 random variables given by

$$X^2 = \sum_j L_j z_j^2 \quad j = 1, \ldots, (I-1)(J-1)$$

in which the L_js are the non-zero eigenvalues of the so-called design effect matrix

$$\mathbf{D} = (\mathbf{H\,A\,H'})^{-1}(\mathbf{H\,V\,H'})$$

and the z_js are independent $N(0,1)$ variables. In the design effect matrix, **A** is the matrix consistent with independent observations.

Ideally, we would like to know the values of the L_js to be able to ascertain the true distribution of X^2, though in practice this is likely to be difficult. Various ways of side-stepping this problem are described in Fingleton (1986). In this paper we concentrate on one approach, namely the use of a model (9) for the spatial dependence. The model that is suggested for the dependence among the observations has the property that the L_js are each equal to L and thus a simple correction is possible, so long as the model is true.

Then $X^2 = \sum_j L_j z_j^2$ is distributed as $L\chi^2$ under the null hypothesis and therefore the chi-squared test statistic (1) divided by L approximates to the χ^2 distribution with $(I-1)(J-1)$ degrees of freedom (see Holt *et al.* 1980; Rao and Scott, 1979, 1981, 1984; Bedrick, 1983). When **A** and **V** are identical, then the L_js are equal to 1 and the uncorrected statistic (1) is distributed approximately as χ^2 with $(I-1)$ $(J-1)$ degrees of freedom.

The presence of spatial dependence according to the model (9) ensures that $\mathbf{V} = \mathbf{L\,A}$. To show this, it is useful (see Altham, 1979) to define an indicator variable such that

$$z_{vijs} = 1 \text{ if } v \text{ occurs at row } i\text{, column } j \text{ of cluster } s$$

$$z_{vijs} = 0 \text{ otherwise}$$

and to denote the total number of trials at which category v is observed by $z_v = \sum_{ijs} z_{vijs}$.

Typical element u,v of the matrix **V** is then

$$\begin{aligned}\mathbf{V}_{uv} &= \text{cov}(z_u, z_v) = \text{cov}(\textstyle\sum_{ijs} z_{uijs}, \sum_{ijs} z_{vijs})\\ &= \text{cov}(z_{u111}, z_{v111}) + \text{cov}(z_{u121}, z_{v121}) + \ldots\\ &+ \text{cov}(z_{u112}, z_{v112}) + \text{cov}(z_{u122}, z_{v122}) + \ldots\\ &+ \text{cov}(z_{u111}, z_{v121}) + \text{cov}(z_{u121}, z_{v111}) + \ldots\\ &+ \text{cov}(z_{u112}, z_{v122}) + \text{cov}(z_{u122}, z_{v112}) + \ldots\\ &+ \ldots\end{aligned}$$

A more succinct representation of the above is given by

$$\begin{aligned}\mathbf{V}_{uv} = \sum_s m_s n_s (p_u g_{uv} - p_u p_v)\\ + 2\sum_s N(1, m_s, n_s)(P_{1s}(u,v) - p_u p_v)\\ + 2\sum_s N(2, m_s, n_s)(P_{2s}(u,v) - p_u p_v)\\ + 2\sum_s N(3, m_s, n_s)(P_{3s}(u,v) - p_u p_v)\\ + \ldots \text{etc}\end{aligned}$$

which simplifies to

$$\begin{aligned}\mathbf{V}_{uv} = \sum_s m_s n_s (p_u g_{uv} - p_u p_v)\\ + 2\sum_r \sum_s N(r, m_s, n_s)(P_{rs}(u,v) - p_u p_v)\end{aligned}$$

From the definition of our model (9), this equates to

$$\mathbf{V}_{uv} = \sum_s m_s n_s (p_u g_{uv} - p_u p_v)\{1 + 2\sum_r \sum_s N(r, m_s, n_s) a_{rs} / \sum_s m_s n_s\}$$

which corresponds to equation (6).

The setting of each of the a_{rs}'s to zero gives

$$\mathbf{V}_{uv} = \sum_s m_s n_s (p_u g_{uv} - p_u p_v) = \mathbf{A}_{uv}$$

for all u,v so in this case no correction is required.

APPENDIX III

If $m_s = m$ and $n_s = n$ for all clusters $(s = 1, \ldots, k)$ then

$$\mathbf{V}_{uv} = kmn(p_u g_{uv} - p_u p_v) + 2k\sum_r N(r, m, n)\ (P_{rs}(u,v) - p_u p_v)$$

$$\mathbf{V}_{uv} = kmn(p_u g_{uv} - p_u p_v) + 2k\sum_r N(r, m, n)\ (p_u a_{rs} g_{uv} - a_{rs} p_u p_v)$$

and if $a_{rs} = a_r$ $(r = 1, \ldots, d;\ s = 1, \ldots, k)$

$$\mathbf{V}_{uv} = kmn(p_u g_{uv} - p_u p_v)\{1 + 2\sum_r N(r, m, n) a_r / mn\}$$

If we imagine an extreme case of 'cluster' sampling in which $m = n = 1$ then the 'clusters' amount to independent trials and no correction is required.

REFERENCES

ALTHAM, P. M. E. (1976) 'Discrete variable analysis for individuals grouped into families', *Biometrika* 63: 263–9

ALTHAM, P. M. E. (1979) 'Detecting relationships between categorical variables observed over time: a problem of deflating a chi-squared statistic', *Appl. Statist.* 28: 115–25

BEDRICK, E. J. (1983) 'Adjusted chi-squared tests for cross-classified tables of survey data', *Biometrika* 70: 591–5

·BRIER, S. S. (1980) 'Analysis of contingency tables under cluster sampling', *Biometrika* 67: 591–6

CLIFF, A. D. and ORD, J. K. (1981) *Spatial processes: models and applications* (Pion, London)

COHEN, J. E. (1976) 'The distribution of the chi-squared statistic under clustered sampling from contingency tables', *J. Am. Statist. Ass.* 71: 665–70

FINGLETON, B. (1983a) 'Independence, stationarity, categorical spatial data and the chi-squared test', *Environ. Plann. A.* 15: 483–99

FINGLETON, B. (1983b) 'Log-linear models with dependent spatial data', *Environ. Plann. A.* 15: 801–14

FINGLETON, B. (1984) *Models of category counts* (Cambridge University Press, Cambridge)

FINGLETON, B. (1986) 'Analyzing cross-classified data with inherent spatial dependence', *Geogrl. Analysis* 18: 48–61

FINGLETON, B. and PORTEOUS, B. T. (1985) 'Contribution to the discussion of the paper by Bennett and Haining', *J. R. Statist. Soc. A.* 148: 31–2

HOLT, D., SCOTT, A. J. and EWINGS, P. D. (1980) 'Chi-squared tests with survey data', *J. R. Statist. Soc. A.* 143: 303–20

MOSER, C. A. and KALTON, G. (1971) *Survey methods in social investigation* (Heinemann, London) 2nd edition

PORTEOUS, B. T. (1985) 'Properties of log-linear and covariance selection models', unpubl. PhD thesis, Dept. of Pure Mathematics and Mathematical Statistics, University of Cambridge

RAO, J. N. K. (1986) personal communication

RAO, J. N. K. and SCOTT, A. J. (1979) 'Chi-squared tests for analysis of categorical data from complex surveys', *Proc. Am. Statist. Ass., Section on Survey Research Methods:* 58–66

RAO, J. N. K. and SCOTT, A. J. (1981) 'The analysis of categorical data from complex sample surveys: chi-squared tests of goodness of fit and independence in two-way tables', *J. Am. Statist. Ass.* 76: 221–30

RAO, J. N. K. and SCOTT, A. J. (1984) 'On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data', *Ann. Statist.* 12: 46–60

RAO, J. N. K. and SCOTT, A. J. (1985) 'On simple adjustments to chi-square tests with sample survey data', in *Analysis of categorical data from sample surveys: a collection of five papers* (Technical Report No. 66, Laboratory for Research in Statistics and Probability, Carleton University, Ottawa)

TAVARÉ, S. and ALTHAM, P. M. E. (1983) 'Serial dependence of observations leading to contingency tables, and corrections to chi-squared statistics', *Biometrika* 70: 139–44

UPTON, G. J. G. and FINGLETON, B. (1985) *Spatial data analysis by example* (Wiley, Chichester)

WRIGLEY, N. (1985) *Categorical data analysis for geographers and environmental scientists* (Longman, London)