Environment and Planning B: Planning and Design 2011, volume 38, pages 216 – 230

doi:10.1068/b36093

An assessment of the effectiveness of multiple hypothesis testing for geographical anomaly detection

Chris Brunsdon
Department of Geography, University of Leicester, Leicester LE1 7RH, England;
e-mail: cb179@le.ac.uk

Martin Charlton
National Centre for Geocomputation, National University of Ireland Maynooth, Maynooth,
Co. Kildare, Ireland; e-mail: martin.charlton@nuim.ie

Received 30 July 2009; in revised form 9 April 2010

Abstract. The practice of multiple significance testing is reviewed, and an alternative to the frequently used Bonferroni correction is considered. Rather than controlling the family-wise error rate (FWER)—the probability of a false positive in *any* of the significance tests—this alternative due to Benjamini and Hochberg controls the false discovery rate (FDR). This is the proportion of tests reporting a significant result that are actually 'false alarms'. The methods (and some variants) are demonstrated on a procedure to detect clusters of full-time unpaid carers based on UK census data, and are also assessed using simulation. Simulation results show that the FDR-based corrections are typically more powerful than FWER-based ones, and also that the degree of conservatism in FWER-based procedures is quite extreme, to the extent that the standard Bonferroni procedure intended to constrain the FWER to be below 0.05 actually has a FWER of around 6×10^{-5}. We conclude that in situations where one is scanning for anomalies, the extreme conservatism of FWER-based approaches results in a lack of power, and that FDR-based approaches are more appropriate.

1 Introduction

In a number of geographical problems, there is a need to detect 'clusters' or spatial 'hotspots' in which the prevalence of some phenomena is unusually concentrated in some particular place. Here, we will term such phenomena anomalies. Often methods to identify anomalies involve the repeated application of statistical significance tests over a number of places. However, unmodified multiple testing leads to unacceptably high false positive rates, that is, the chance of falsely labelling a place as anomalous is unacceptably high. Traditionally this has been dealt with by the use of Benferroni's adjustment (Bonferroni, 1935), but this results in a greatly reduced power of the detection procedure; in other words, the chances of finding anomalies when they actually exist is unreasonably low. Recent advances, such as that proposed by Benjamini and Hochberg (1995) provide an alternative strategy for reducing the false positive rate by limiting the false discovery rate (FDR). The FDR is the proportion of places reported as anomalies that, in reality, are not. This differs from the false positive rate, as it is a proportion only of reported anomalies not of all tests. Benjamini and Hochberg's (BH) approach provides higher power than that of Bonferroni. A recent two-step modification of this (BH2S; Benjamini et al, 2006) is claimed to improve slightly on the performance of the original. These tests are valid when they are independent or positively correlated. Another adjustment proposed by Benjamini and Yekutieli (2001) (BY) is similar to BH, but is valid even when the individual tests are not independent or positively correlated.

Despite a citation count in excess of 5000 for the Benjamini and Hochberg (1995) paper, the BH approach has received little attention in the geographical literature. Yamada et al (2009) refer to the paper, but opt for a Bonferroni approach in their worked example involving multiple tests. Chung et al (2005) apply the approach to the problem of detecting multiple signals embedded in noisy observations from a sensor array and find, using simulations, that it does increase the power of signal detection when regarded as a multiple-significance detection procedure. Greve et al (2008) also consider the BH procedure in an investigation of spatial variation in bird body mass in South Africa, but it is used to investigate tests partitioned by species-richness categories, rather than to detect spatial pattern directly. In Caldas de Castro and Singer (2006) the FDR approach is applied to the detection of clusters of high malaria rates using a number of local indicators of spatial association (LISAs): Getis G^* and G statistics (Getis and Ord, 1992; Ord and Getis, 1995) and local Moran's I (Anselin, 1995). Using this approach a number of important clusters were identified that went undetected using the traditional Bonferroni approach. Caldas de Castro and Singer go on to test the effectiveness of the approach with a clustered dataset simulated on a regular grid. Inspired by this work, the aim of this study is to extend these ideas and apply a similar approach to a new situation. In particular, the application here involves UK census data from 2001 about voluntary carers, that is, people who spend in excess of twenty hours per week providing unpaid care for another person. Clearly all such people face a number of difficulties, but various spending authorities may need to identify areas with high concentrations of such people as in these areas increased demands on resources such as doctors surgeries present a unique set of problems. Identification of such areas could be a useful guide to providing extra local resources in the hope of overcoming such problems. Concentrations of such carers are likely to occur in relatively small clusters, typically within neighbourhoods; for example, within community sheltered-housing projects. The size of a cluster may well be smaller than the size of the geographical reporting units, and would therefore be more likely to manifest itself as a raised-rates incidence within an individual unit, rather than as a locally high level of autocorrelation between neighbouring units. For this reason, we compare area incidence against a standard binomial distribution (based on national rates) rather than using the LISA-based approach of Caldas de Castro and Singer (2006).

In this study, counts of carers and baseline populations at the lower super output area (LSOA) level for Leicestershire, England are considered. As suggested above, a significance test based on a binomial-distribution model is the basis for detecting anomalous LSOAs, with the Bonferroni, BH, BH2S, and BY adjustments applied in turn. Also, the proposed method of comparing local incidence discussed above is also the basis for anomaly detection in the geographical analysis machine (GAM) of Openshaw et al (1987), a technique for detecting raised incidence rates of geographical phenomena which, although highly influential, received some criticism due to the way in which it dealt with multiple hypothesis tests. In this paper we will demonstrate how the ideas above can be used to modify the GAM and address this criticism. It will also be shown that these techniques can be used in conjunction with a technique proposed by Schweder and Spjøtvoll (1982) to provide a graphical assessment of the rate of violation of a given null hypothesis.

A number of simulations are run using real data with different levels of concentration in a set of anomalous LSOAs, the others being set to have the proportion of carers equal to the English average (around 6%). As an alternative to considering the effectiveness of the method in detecting clusters in a single simulated dataset as in Caldas de Castro and Singer (2006), here it is intended to measure the *power* of the technique, that is, the probability of detecting genuine effects in the long run. A number of simulations are run (in this case 1000) and the number of times anomalous LSOAs are detected is counted, and from this the probability of successful detection is estimated. The motivation for this is that power calculated in this way is a widely adopted convention as a measure of the effectiveness of statistical tests. For example, funding agencies, ethics boards, and research review panels often request estimates of power expressed in this way when considering proposals for research involving statistical procedures. Here, power will be estimated via simulation over a range of degrees of clustering, where the simulated levels of unpaid carers in anomalous LSOAs will vary over a range of values.

Next we consider the methods of adjustment for multiple testing in more detail. Following this the Leicestershire data will be introduced, together with a detailed exposition of the anomaly detection technique used here. After this the simulation to estimate power will be described in detail, and the results of this will be presented. Following this, the graphical exploration will be considered, followed by a concluding discussion.

2 The multiple-testing issue in detail

As stated in the introduction, many standard spatial statistical procedures provide formal tests for clustering in general but do not provide information about the location of any specific anomalies. These are described by Openshaw et al (1987) as 'whole map statistics'. In response to this, a number of approaches have been proposed, and a commonly adopted procedure in these is to carry out a number of local significance tests and to flag those localities for which a null hypothesis of no clustering, or of a 'normal' level of incidence, is rejected.

However, a major concern with this approach is that of *multiple hypothesis testing*. If there are m locations and each test is carried out at a significance level α (typically, $\alpha = 0.05$) when the null hypothesis is true at all locations then on average $m\alpha$ of these will test positive. Despite this, there is a temptation even if just one test proves positive to claim to have 'found a cluster', implicitly rejecting a null hypothesis of 'no clustering' or 'no anomalies'. Clearly the probability of detecting at least one cluster in this process under the null hypothesis is much higher than α. Under this hypothesis, and also assuming that each test is independent, then the probability of *not* rejecting at every locality is $(1-\alpha)^m$, and so α^*, the probability of finding one or more significant locations, is related to α, the location-wise significance level, by $\alpha^* = 1-(1-\alpha)^m$. This quantity is effectively the significance level of the overall—or whole map—test for clustering if one works on an 'any significant result rejects the null hypothesis' basis. The quantity is often referred to in the literature as the family-wise error rate (FWER) as it is the error rate based on a whole family of tests, rather than any individual test. For example, if $m = 100$ and $\alpha = 0.05$ then α^* is around 0.99 so it is almost certain that at least one location will have a significant result. Clearly, then, simply reporting every significant location is problematic in its unmodified form. One can control for this by choosing the location-wise significance level α so that α^* has a desirable value, typically 0.05. This may be done by noting that

$$\alpha = 1-(1-\alpha^*)^{1/m} , \tag{1}$$

or, approximately,

$$\alpha = \frac{\alpha^*}{m} . \tag{2}$$

Equation (1) is due to Šidàk (1967), and equation (2) stems from the work of Bonferroni (1935), first used in this context by Dunn (1961), is often referred to as 'Bonferroni adjustment'. Returning to the example with $m = 100$, if we required α^* to be 0.05, then the individual α's should be set at around 0.0005.

However, this approach is not without its own problems. Noting the very low values of α needed for testing in each locality, the level of evidence needs to be extremely high for an individual cluster to be identified. As demonstrated (eg Rogerson and Yamada, 2009, page 121) this in turn leads to a reduction in the power of the test; that is, the probability that the null hypothesis is rejected when clustering actually occurs. Thus, although the Bonferroni adjustment is risk averse in the sense of limiting the probability that any clustering is found when the null hypothesis is true, this degree of caution results in a reduced chance of finding clusters when they really exist. A further issue is that the adjustment is based on the assumption that the location-based tests are independent and often this may not be the case. In fact, it turns out that, in general,

$$\frac{\alpha^*}{m} \leqslant \alpha \tag{3}$$

even when the test are dependent (Games, 1977; Šidàk, 1967). This suggests that in all cases one can find an α for any given *conservative* false positive rate α^*. Here, a conservative false positive rate is defined to be a rate that the actual false positive rate does not exceed. Thus, under the null hypothesis, the map-wide false positive rate lies in the interval $[0, \alpha^*]$. Typically, if there is a high degree of positive dependency in the tests, the actual false positive rate may be considerably lower than α^*. This frequently implies a further loss of power compounding the effect outlined and demonstrated by Rogerson and Yamada (2009). A more recent procedure by Holm (1979) improves on this situation by providing an alternative procedure that dominates the Bonferroni method in terms of power, but this improvement is marginal in many situations.

In this type of practical situation described at the beginning of this section this may be unsatisfactory. For example, in the situation relating to carers discussed in the introduction perhaps it is worse to fail to detect locations requiring special attention, or an increase in locally targeted resources, than it is to incorrectly flag a small number of areas having no such requirements.

BH is an alternative approach proposed by Benjamini and Hochberg (1995) which limits the FDR rather than α^*. The FDR is defined as the proportion of localities in which a test is significant for which the null hypothesis is true. In basic terms, it is the proportion of 'false alarms' in the localities tagged as 'significant'. The method will be described in detail later in the paper. For example, one might work with an FDR of 0.05, rather than an α^* of the same value. Typically, procedures that limit the FDR rather than the probability of one or more false alarms are less stringent, but they increase the power of the testing procedure, so the chances of detecting genuine clusters are improved provided we are willing to accept that there may be more false alarms in our set of 'significant' localities. Using this approach is more in tune with the practical situation described above. We may be willing to pay a higher price in false positives for a gain in identifying genuinely anomalous localities. This procedure is carried out as follows: suppose that for each of the m locations a p-value is computed based on the null hypothesis. Suppose we denote the ith ranking p-value as $p_{(i)}$, where $p_{(1)}$ is the smallest value and $p_{(m)}$ is the largest, and that q^* is a desired FDR, then let

$$k = \max \left\{ i : p_{(i)} < q^* \frac{i}{m} \right\} . \tag{4}$$

Then rejecting the hypothesis corresponding to $\{p_{(1)}, p_{(2)}, \ldots, p_{(k)}\}$ will result in an FDR less than or equal to q^*. If $p_{(1)} > q^*/m$ then no hypotheses are rejected. This generally proves to be more powerful than the Bonferroni approach. However in the first instance the approach was only proven for the case when all tests are independent. In fact, the method is still conservative in terms of the FDR under certain types of model dependency. Benjamini and Yekutieli (2001) discuss a number of possible situations where this is the case. For example, if the test statistics are multivariate normal, with all entries in their variance–covariance matrix being nonnegative, then the adjustment procedure associated with equation (4) still holds.

More recently, this approach has been improved using a two-step approach (Benjamini et al, 2006). In this case, the procedure is as follows (based on Benjamini et al, 2006, definition 6)
(1) Carry out the BH procedure replacing q^* by $q' = q^*/(1+q^*)$.
(2) Define $\hat{m}_0$ as the number of cases rejected in step 1.
(3) Carry out the BH procedure again, with $q'' = q'n/(n - \hat{m}_0)$.

The entire approach using all the above steps is shown still to be conservative with respect to the FDR [in section (5) of the same paper] and the test is less conservative than the original method. This procedure is the BH2S adjustment.

Finally, a more generally applicable approach was also described in Benjamini and Yekutieli (2001). It applies for all sets of tests regardless of the degree or nature of any dependence between the tests. In this case k in equation (4) is replaced by

$$k = \max\left\{ i : p_{(i)} < \left[\frac{q^*}{\sum_{i=1}^{m} i^{-1}} \right] \frac{i}{m} \right\}, \tag{5}$$

and then the same procedure is carried out. This approach, the BY adjustment, is therefore generally applicable although it is more conservative and less powerful than the BH or BH2S procedures in situations where the latter are valid.

3 A practical example

As stated in the introduction, a key aim of this study was to consider these ideas in a 'real-world' context, examining levels of unpaid full-time carers in Leicestershire. For the first time, in 2001, the UK census contained a question relating to the provision of unpaid care:

> "Do you look after or give any help or support to family members, friends, neighbours, or other because of
> - long-term physical or mental ill-health or disability, or
> - problems related to old age?"

(General Register Office for Scotland, 2001; Northern Ireland Statistics and Research Agency, 2001; ONS, 2003). In addition a further question asked:

> "Over the last twelve months would you say your health has on the whole been
> - Good?
> - Fairly good?
> - Not good?"

The combination of these two questions highlights an important issue—that of unpaid carers who are themselves not in good health—discussed, for example, in Doran et al (2003). In particular, they argue that elderly or young carers in this situation are particularly vulnerable. In this example we focus on elderly carers (aged 65 years and over at the time of the census) who do not report their health as being 'good'. Count data was obtained for each LSOA in the county of Leicestershire and the Leicester Unitary Authority, for the total population aged over 65 years who do not consider their health to be 'good', together with counts for the subset of this group who also provide 20 hours or more of unpaid care each week. The rates for the carers (expressed as percentages) are mapped in figure 1 together with the names of some major places in the study area.

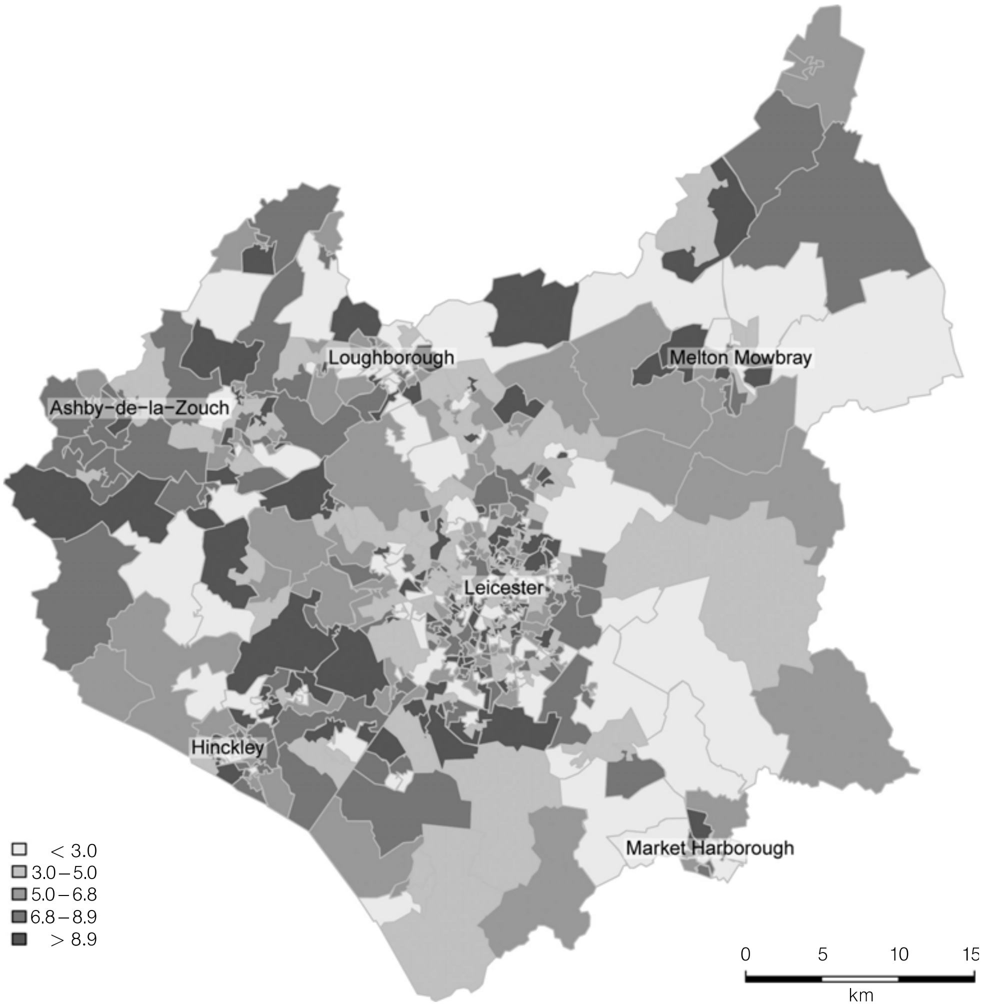

Figure 1. Proportion of unpaid full-time carers 65 years of age and over and in poor health by lower super output area in Leicester Unitary Authority and Leicestershire County.

In order to detect anomalous LSOAs, a model is needed for a null distribution of the number of carers aged over 65 years and not in good health in each LSOA. For England as a whole, according to the 2001 UK census there are 295 059 such unpaid carers out of a denominator population of 4 796 773, giving an average proportion of $\theta_E = 0.0615$ (to three significant digits). Assuming that any member of the baseline population is equally likely to be a carer, and if the population of any given LSOA (with index number i) is N_i, then one would expect the number of carers C_i counted in this LSOA to follow a binomial distribution with parameters (N_i, θ_E), that is,

$$P(C_i = c) = \binom{c}{N_i} \theta_E^c (1 - \theta_E)^{N_i - c} . \tag{6}$$

Using this model, one can test the null hypothesis that the C_i associated with any given LSOA comes from this distribution. More formally the test is for

$$\begin{aligned} H_0 &: \theta = \theta_E \text{ against} , \\ H_1 &: \theta > \theta_r . \end{aligned} \tag{7}$$

Here, the test is one-sided as there is only interest in evidence that the proportion of carers in an LSOA *exceeds* the English average. The count of carers in an LSOA can be used as a test statistic for this hypothesis. The p-value associated with a count C_i at a given LSOA is given by

$$p_i = P(c \geqslant C_i) = \sum_{c=C_i}^{N_i} \binom{c}{N_i} \theta_{\mathrm{E}}^{c} (1-\theta_{\mathrm{E}})^{N_i - c} \ . \qquad (8)$$

3.1 Basic results

Thus, it is possible to test for anomalous values of C_i using equation (8), and, of course, to control for multiple hypothesis testing using one of the methods discussed in the previous section. Note that under the null hypothesis each of the tests is independent, so the BH adjustment can be justified here. In figure 2 the results of applying both Bonferroni-adjusted and BH-adjusted tests are shown. Under Bonferroni, four anomalous LSOAs are detected, and with BH a further twelve are added. Using BH2S no further anomalous LSOAs are found. Finally, the BY adjustment identified a set of

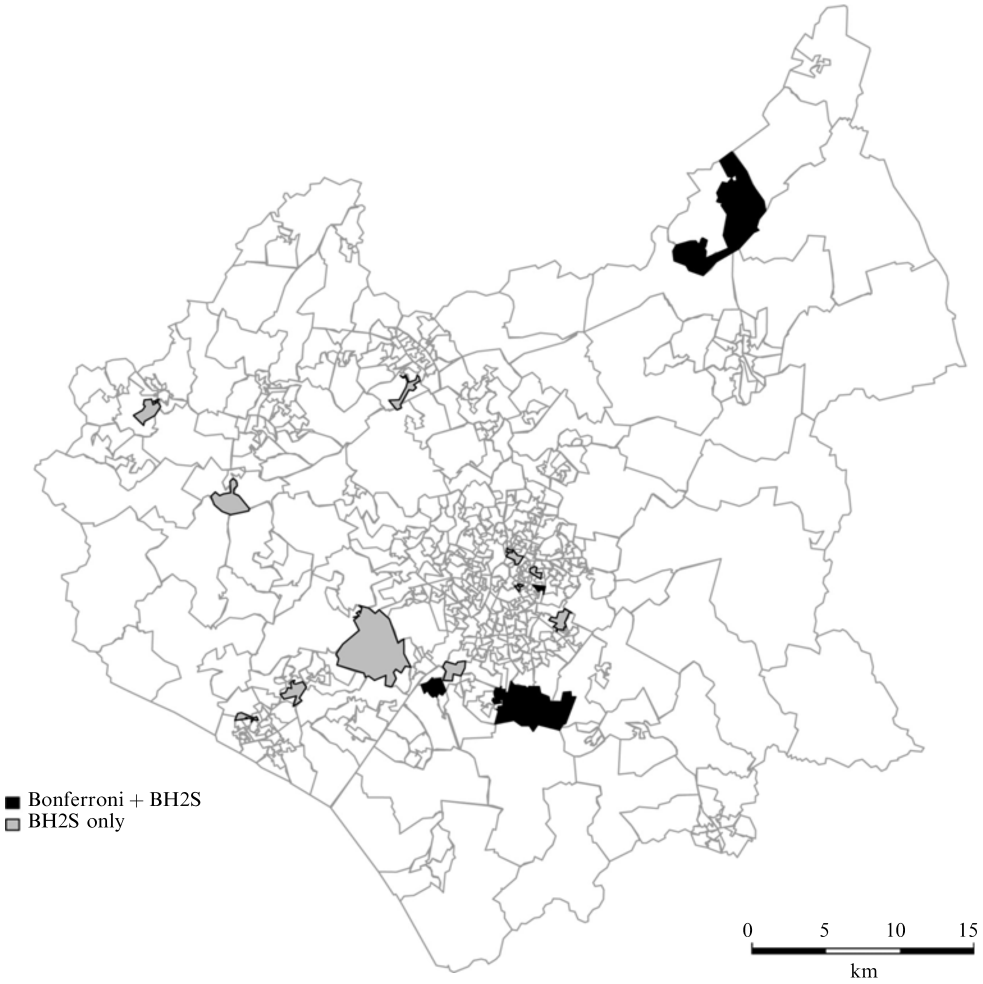

Figure 2. Anomalies of unpaid full-time carers 65 years of age and over and in poor health in Leicester Unitary Authority and Leicestershire County. (See text for details of Bonferroni and BH2S procedures.)

anomalous LSOAs identical to those found by the Bonferroni procedure. Here, since under H_0 the hypothesis tests are independent, use of the BH and BH2S procedures is justified. In some cases the anomalous LSOAs appear as small groups of adjacent regions, suggesting that the high concentration of carers extends beyond one single LSOA. Note also that a number of the 'extra' LSOAs flagged by the BH and BH2S adjustments are in areas to the east of the map and that with the Bonferroni adjustment only one urban anomaly (in Leicester city centre) is detected. Thus, the phenomenon of east Leicestershire 'clusters of carers' and a small group of clusters in Leicester city centre is in danger of going largely undetected if only the Bonferroni adjustment is used.

3.2 The GAM revisited

It is also possible to apply this technique to a GAM style of cluster detection (Openshaw et al, 1987). In this approach, a number of locations with associated search radii are considered as prospective 'cluster' centres. For each cluster every spatial unit within the search radius is identified, and, on the basis of these, the total population and total number of incidences of the phenomenon under study (in this case elderly unpaid carers) are found. A statistical test is then carried out—for example, based on the p-values computed using equation (8)—and any of the prospective clusters with a significant result are then mapped. Typically, mapping is carried out by drawing a circle of the associated search radius centred on each of the locations having a significant result. The method is seminal in that it draws attention to the need not only to test for the existence of clusters but also to identify their locations.

In the original GAM, published before the work of Benjamini and Hochberg, significance testing was based on unadjusted p-values. However, the method may be modified to adjust the p-values using techniques such as those reviewed here. Here, we apply a GAM-type strategy to detect potential 'clusters' of elderly carers centred around general practitioner (GP) practices (or health centres) in Leicester and Leicestershire. The aim here is to provide a 'primary service provision' view of geographical clusters of carers by identifying those practices having significantly large proportions of elderly carers in their vicinity. In practical terms this could identify those practices needing extra resources to address the specific needs of carers and their charges.

Thus, in this study the cluster centres are the locations of GP practices. These may be obtained via the NHS Choices website,[1] which provides post codes for all GP practices in England. The search radius for each GP practice is dependent on the nearby population density. It is likely that in rural areas GP practices will have registered patients living further away than in urban areas. Recent reports suggest that a 'large' practice has a size of 8000 patients or more (NHS: The Information Centre, 2007). The approach here is to define a search radius around a GP practice to be the largest radius to contain a population less than 18 000 people. This allows for some crossover between practices and for a potentially large practice in all cases, and could also identify regions in which a cluster of practices are subject to a cluster of carers. Clearly this choice may require further consideration, but it serves as a reasonable initial example.

Carrying out the GAM analysis on this basis gives the results shown in figure 3. Both Bonferroni [figure 3(a)] and BH2S adjustments [figure 3(b)] are applied before identifying and mapping significant clusters. Again, the results from both show clusters in Leicester city centre. However, a western cluster is evident in the BH2S corrected analysis, but detected by Bonferroni. In this case none of the eastern rural clusters show, but this may be because the GP practices associated with some of

[1] http://www.nhs.uk/servicedirectories/Pages/PrimaryCareTrustListing.aspx

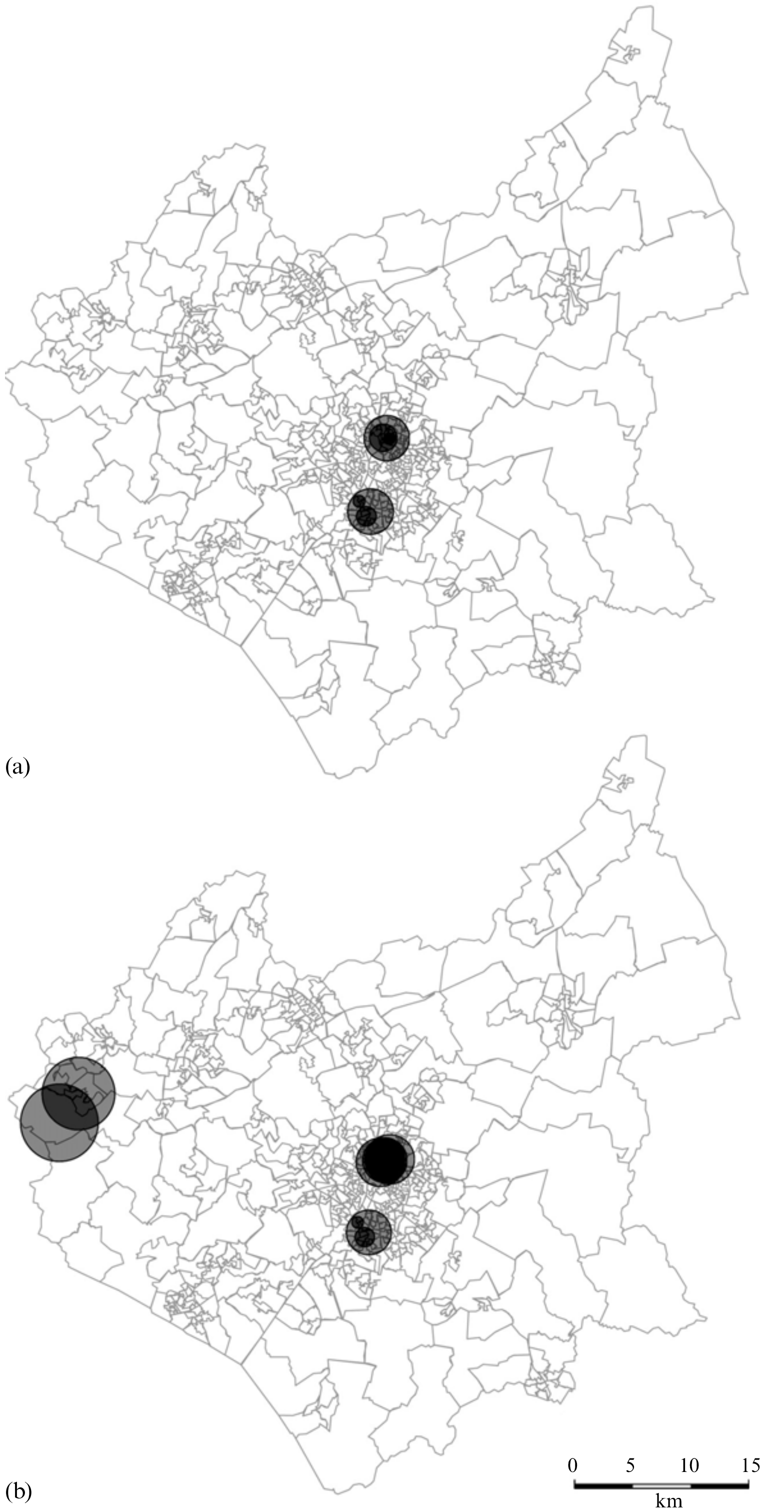

Figure 3. Geographical analysis machine analysis of General Practitioner (GP) practice-based clusters of carers 65 years of age and over. Anomalous GP zones: (a) Bonferroni; (b) false discovery rate.

these rural clusters are situated in town. However, similar to the previous example, some potential GP practices would go 'unflagged' if only the Bonferroni adjustments were used.

4 Evaluation through simulation

It appeared that there may be thirty-three anomalous LSOAs. However, to fully appraise this analysis it is helpful to have some idea of the levels of anomaly that the above procedure is capable of detecting. To investigate this a number of simulations were carried out. The model used in the simulations was as set out in equation (6) for the nonanomalous (background) LSOAs, that is, the count of carers had a binomial (N_i, θ_E) distribution. For the anomalous observations, the distribution was binomial (N_i, θ_a), where $\theta_a > \theta_\mathrm{E}$. The LSOAs that were tagged as anomalous in the BH-adjusted testing of the previous section are assumed to be anomalous in the simulation, and counts are simulated with θ_a, whilst the remainder are simulated with θ_E. Choosing the simulated anomalous LSOAs in this way is intended to present a realistic spatial arrangement of anomalous LSOAs, reflecting the level of clustering of adjacent anomalous zones, and the urban–rural split likely to be found in reality.

For a given θ_a 1000 simulations were run. For each simulation, BH-adjusted and BH2S anomaly detected procedures were applied, as was a Bonferroni adjusted procedure. Thus, for each simulation, as in Benjamini and Hochberg (1995) a table (table 1) can be drawn up comparing the actual status of each LSOA (as prescribed in the simulation), with that reported using either the Bonferroni, BH, or BH2S adjusted detection procedure. In the simulations described $m_0 = 33$ and $m = 583$—these are known a priori; U, V, T, S and R (see table 1) will vary depending on each simulation. Also, although m_0 was known in the simulations, for BH2S the estimated value was used as the intention was to assess the performance of this procedure in a real-world situation where this quantity must be estimated.

Table 1. Outcomes of anomaly detection.

	Anomaly reported		Total
	no	yes	
Not anomalous	U	V	$m - m_0$
Anomalous	T	S	m_0
Total	$m - R$	R	m

The power of the test when applied to the whole map can then be defined as the proportion of genuinely anomalous LSOAs that are detected—in terms of table 1 this is $S/(S+T)$. To estimate the power for a given θ_a, this quantity is computed for each of the 1000 simulations, and the average is recorded. By carrying out this procedure for a range of values of θ_a, starting at just over the English average value θ_E (around 0.02), the increase in power of a testing as the degree of anomaly increases may be estimated. This is carried out for the Bonferroni, BH, and BH2S adjustment procedures. In fact, here it is also done for the BY adjustment. Although BY is not required here, it is useful to give an indication of the power of this procedure relative to the others. Here, simulations were run for all three adjustments, for $\theta_a = \{0.025, 0.03, 0.035, \ldots, 0.06\}$, that is, from just above θ_E to around three times this value. The results are displayed in graph form in figure 4.

The figure suggests that the powers of the BH and BH2S adjusted procedures always exceeds that of the Bonferroni, and when $\theta_a = 0.03$, around 50% larger than the English national rate, BH and BH2S are around three times as powerful. There is

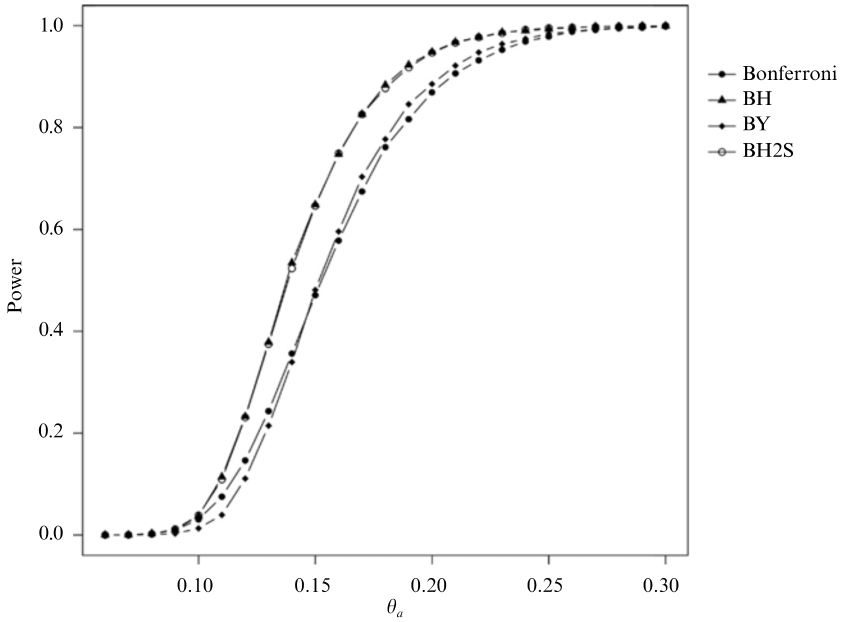

Figure 4. Powers of anomaly detection procedures. (See text for details of Bonferroni, BH, BY, and BH2S procedures.)

little difference here between BH and BH2S. Similar results (using BH) were found by Caldas de Castro and Singer (2006) in an investigation applying multiple hypothesis tests to significance levels of Getis's *G*-statistics (Getis and Ord, 1992; Ord and Getis, 1995) and local Moran's *I*-statistics (Anselin, 1995; Anselin et al, 1996).

The BY procedure is generally less powerful than BH, although it is of course generally applicable, unlike BH and BH2S. It is more powerful than Bonferroni (which is also generally applicable), except in the situation where θ_a is only just larger than θ_{E}.

5 Conservatism

For all the techniques outlined here, the multiple testing criteria are claimed to be conservative in some way—so that the true FWER (Bonferroni) or the FDR (BH, BH2S, BY) are below a value set as the 'control' value in the multiple testing procedure. In this short section, the experimental values of these quantities are considered based on the simulation results, to obtain some idea of *how* conservative these approaches are, and the relationship between this and the degree of clustering characterised by θ_a. On the basis of table 1 the expressions for these quantities are the average value of V/R for the FDR, and $\mathrm{pr}(V > 0|\mathrm{H}_0)$ for the FWER. These quantities are shown in graph form in figure 5.

From these it may be seen empirically that all the tests are conservative with respect to the quantities that theory suggests. The FWER for the Bonferroni procedure is typically around 0.02 which falls below the control value of 0.05 by a factor of around 2.5. The BY procedure has a FWER of around 0.05, despite actually controlling for the FDR. BH and BH2S have notably lager values, a FWER of around 0.5 suggesting that there is an odds-on chance of falsely rejecting at least one hypothesis out of the 583. However, this demonstrates the essence of the FDR approach; although there is a notable chance of a small number of false rejections the chances of detecting genuine anomalies is increased. Note also that both the BH and BH2S procedures are conservative with respect to the FDR, at worst having values of around 0.035 when the intention is to constrain the worst case to below 0.05.

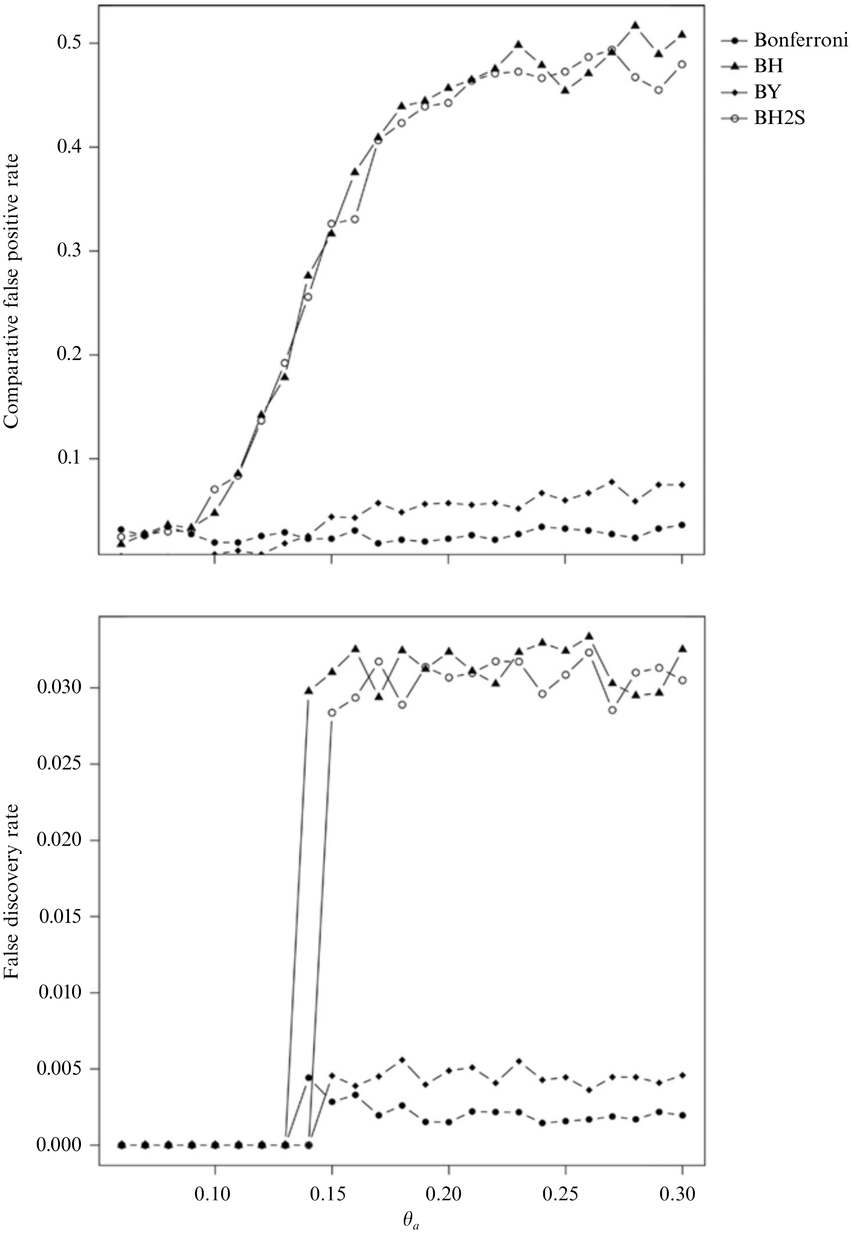

Figure 5. Family-wise error rate and false discovery rate of anomaly procedures. (See text for details of Bonferroni, BH, BY, and BH2S procedures.)

6 Graphical evaluation

In addition to the more formal method outlined in the previous sections, it is also possible to evaluate the above model graphically, using a method outlined by Schweder and Spjøtvoll (1982). Here, we use a slightly modified version of this method. To explain the approach, first note that when a p-value is computed for a given region (in this case an LSOA), this is itself a random quantity. If the null hypothesis of equation (7) is true, then p is in fact uniformly distributed on $[0, 1]$. In cases where the null hypothesis is not true—in particular for H_1 in equation (7)—p will not follow this distribution. Typically we would expect p to be notably lower than under H_0. Alternatively, in cases where there is actually a *lower* level of concentration than H_0 we would expect usually *high* p-values. If we consider the distribution of p for all areas where H_0 is untrue, we might reasonably expect this to be a u-shaped distribution, with much higher values close to the extremes of zero and one, and near-zero values away from these extremes. Suppose this distribution has the density function $U(p)$. When H_0 is true, the p-values have a distribution equal to the constant value of one. Then, using the notation in table 1, the probability density of p over all regions, $f(p)$, is given by

$$f(p) = \frac{m_0}{m} U(p) + \frac{m - m_0}{m} . \tag{9}$$

When p is not close to zero or one, $U(p)$ is close to zero, and in this case the density is approximated by the constant value $(m - m_0)/m$. Integrating this gives the cumulative distribution function $F(p)$, which is approximately

$$\text{constant} + \frac{m - m_0}{m} p , \tag{10}$$

where p is not close to zero or one. The constant term arises due to the integral of $U(p)$ for values of p close to zero.

Thus if we plot the empirical distribution function for an observed set of p-values, we should expect a central region in which the function is more or less linear. The slope of the line is $(m - m_0)/m$, the proportion of places for which the null hypothesis is true. In practice, the entire empirical distribution function is not plotted, just the set of points $\{[p_i, \hat{F}(p_i)], i = 1, \ldots, m\}$, where

$$\hat{F}(p_i) = \frac{\text{rank}(p_i)}{m} , \tag{11}$$

where the lowest p-value has rank 1, and the largest has rank m, and p_i is the p-value for location i. Inspecting this plot of points gives a number of useful diagnostic indications. Firstly, as stated above, the slope of the line in the centre of the graph gives an estimate of the proportion of areas for which H_0 is true. Secondly, those points that do not lie on this line give some suggestions that the areas corresponding to these points are ones in which H_0 is untrue. Finally, if no straight line is apparent in the plot, this suggests that the null hypothesis seems to hold only for a very small number of areas, and that perhaps the underlying model for H_0 should be reconsidered.

For the LSOA-based carers data, a plot is shown in figure 6. Note that this plot does have a central linear region; experimentally, the slope of this is around 0.68. Residuals around the region do show some correlation, but given that they are based on cumulative p-values this is unsurprising. This suggests that around 68% of areas are well modeled by the null hypothesis so roughly 187 areas do not follow the model. Note also that deviations from linearity occur at both ends of the range of p-values, but that there seem to be more at the upper end than at the lower end, so the number of LSOAs at the upper end of the observed levels of carers form a larger proportion of these 187 areas. However, in this application concern is focused mostly on the lower end of the distribution, as it is low p-values that lead to the rejection of the null hypothesis in the example. Close inspection suggests there are around thirty p-values outside the usual range of variation from the straight line in the lower end of the graph. This is somewhat larger than the number identified by the methods in the previous sections (however, it should be noted that this technique is essentially exploratory) and does seem to give a similar order of magnitude for the number of anomalous zones.

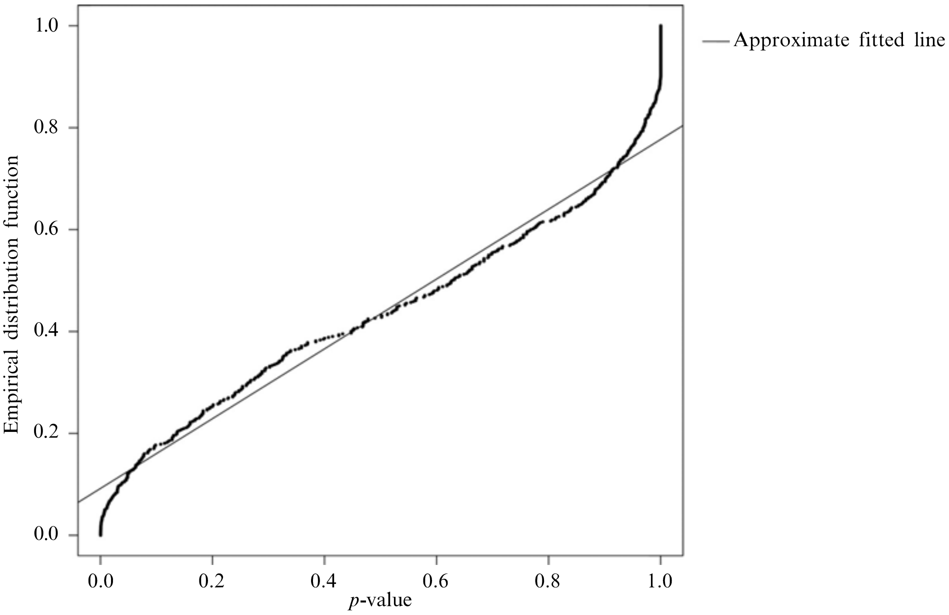

Figure 6. Exploratory plot of p-values for carers data.

7 Conclusions

In this study a number of methods to identify anomalous 'clusters' of elderly unpaid carers have been applied to real-world data, and the procedures were further investigated using simulations. A number of potential clusters have been identified, with those found by the BH and BH2S procedures outnumbering those found by the Bonferroni and BY methods, the latter being a subset of the former. The simulations demonstrated the BH and BH2S procedures to be more powerful than the Bonferroni or BY procedures. Also, in all cases the FDR was well below the 0.05 upper limit set by the BH, BH2S, and BY techniques. However, typically, the BH and BH2S procedures were considerably less conservative (in terms of FDR) than the Bonferroni approach, and this reduction in stringency seemed to result in a notable increase in power for clusters where the unpaid carer rates were simulated at around 1.5 – 3.5 times the English average rate.

When the rates were increased beyond this level, all the techniques had powers very close to one. The effect of anomalous zones was sufficiently extreme that all methods rarely failed to detect them. In such circumstances there may be an argument for continuing to use the Bonferroni approach as this is likely to detect all 'true' clusters and have the lowest false positive rate. Benjamini and Hochberg (2000) stated that in cases where most of the hypotheses are far from being true there is hardly any penalty due to the simultaneous testing of many hypotheses, and it seems that in our study even when a relatively small number of hypothesis are 'far from being true' this may still be the case. However, one could argue that such situations could be identified by exploratory approaches, where a group of LSOAs could be noted to have outlying rates of unpaid carers using, for example, a stem-and-leaf plot. In more subtle situations where anomalies are less marked, it could be argued that the cost of failing to detect an anomalous LSOA may be higher in social terms than that of falsely flagging an LSOA that is not anomalous, and that therefore the more powerful approaches of BH and BH2S should be adopted.

Acknowledgements. Research presented in this paper was funded by a Strategic Research Cluster Grant (07/SRC/I1168) by the Science Foundation Ireland under the National Development Plan. The authors gratefully acknowledge this support.

References

Anselin L, 1995, "Local indicators of spatial association—LISA" *Geographical Analysis* **27** 93 – 115

Anselin L, Bera A, Florax R, Yoon M, 1996, "Simple diagnostic tests for spatial dependence" *Regional Science and Urban Economics* **26** 77 – 104

Benjamini Y, Hochberg Y, 1995, "Controlling the false discovery rate: a practical and powerful approach to multiple testing" *Journal of the Royal Statistical Society Series B* **57** 289 – 300

Benjamini Y, Hochberg Y, 2000, "On the adaptive control of the false discovery rate in multiple testing with independent statistics" *Journal of Educational and Behavioral Statistics* **25** 60 – 83

Benjamini Y, Yekutieli D, 2001, "The control of the false discovery rate in multiple testing under dependency" *The Annals of Statistics* **29** 1165 – 1188

Benjamini Y, Kreiger A, Yekutieli D, 2006, "Adaptive linear step-up procedures that control the false discovery rate" *Biometrika* **93** 491 – 507

Bonferroni C E, 1935, "Il calcolo delle assicurazioni su gruppi di teste", in *Studi in Onore del Professore Salvatore Ortu Carboni* (Rome) pp 13 – 60

Caldas de Castro M, Singer B, 2006, "Controlling the false discovery rate: a new application to account for multiple and dependent tests in local statistics of spatial association" *Geographical Analysis* **38** 180 – 208

Chung P, Bohme J, Mecklenbrauker C, Hero A, 2005, "Multiple signal detection using the Benjamini – Hochberg procedure", in *Computational Advances in Multi-sensor Adaptive Processing 2005* (IEEE, New York) pp 209 – 212

Doran T, Drever F, Whitehead M, 2003, "Health of young and elderly informal carers: analysis of UK census data" *British Medical Journal* **327** 1388

Dunn O, 1961, "Multiple comparisons among means" *Journal of the American Statistical Association* **56** 52 – 64

Games P A, 1977, "An improved t table for simultaneous control on g contrasts" *Journal of the American Statistical Association* **72** 531 – 534

General Register Office for Scotland, 2001, "Scotland's Census. Edinburgh: Gros, 2001", http://www.gro-scotland.gov.uk/census

Getis A, Ord J, 1992, "The analysis of spatial association by use of distance statistics" *Geographical Analysis* **24** 189 – 206

Greve M, Gaston K J, van Rensburg B J, Chown S L, 2008, "Environmental factors, regional body size distributions and spatial variation in body size of local avian assemblages" *Global Ecology and Biogeography* **17** 514 – 523

Holm S, 1979, "A simple sequentially rejective multiple test procedure" *Scandinavian Journal of Statistics* **6** 65 – 70

NHS: The Information Centre, 2007, "2006/2007 general practice workload survey", http://www.ic.nhs.uk/webfiles/publications/gp/GPWorkloadReport.pdf

Northern Ireland Statistics and Research Agency, 2001, "Northern Ireland Census 2001 Output. Belfast, NISRA 2001", http://www.nisranew.nisra.gov.uk/census/start.html

ONS, 2003, "Census 2001: [cd supplement to the national report for England and Wales and key statistics for local authorities in England and Wales]", Office for National Statistics, http://www.ons.gov.uk/census/index.html

Openshaw S, Charlton M, Wymer C, Craft A, 1987, "A mark 1 geographical analysis machine for the automated analysis of point data sets" *International Journal of Geographical Information Systems* **1** 335 – 358

Ord J K, Getis A, 1995, "Local spatial autocorrelation statistics: distributional issues and an application" *Geographical Analysis* **27** 286 – 306

Rogerson P, Yamada I, 2009 *Statistical Detection and Surveillance of Geographical Clusters* (Chapman and Hall/CRC, Boca Raton, FL)

Schweder T, Spjøtvoll E, 1982, "Plots of p-values to evaluate many tests simultaneously" *Biometrika* **69** 492 – 502

Šidàk Z, 1967, "Rectangular confidence region for the means of multivariate normal distributions" *Journal of the American Statistical Association* **62** 626 – 633

Yamada I, Rogerson P A, Lee G, 2009, "GeoSurveillance: a GIS-based system for the detection and monitoring of spatial clusters" *Journal of Geographical Systems* **11** 155 – 173