A Local Join Counts Methodology for Spatial Clustering in Disease from Relative Risk Models

Peter Congdon, School of Geography, Queen Mary University of London, Mile End Rd, London E1 4NS, UK
Email: p.congdon@qmul.ac.uk

Abstract.
This paper considers adaptation of hierarchical models for small area disease counts to detect disease clustering. A high risk area may be an outlier (in local terms) if surrounded by low risk areas, whereas a high risk cluster requires that both the focus area and surrounding areas demonstrate common elevated risk. A local join count method is suggested to detect local clustering of high disease risk in a single health outcome, and extends to assessing bivariate spatial clustering in relative risk. Applications include assessing spatial heterogeneity in effects of area predictors according to local clustering configuration, and gauging sensitivity of bivariate clustering to random effect assumptions.

Keywords: Relative Risk. Cluster Centre. Exceedance. Local join-count. Spatial.

1. Introduction

Spatial analyses of health outcomes at a small-area scale are important for assessing geographic heterogeneity in disease risk and detecting areas with elevated risk. However, the small area administrative subdivisions used in such analyses are generally arbitrary and spatial variations in health risk are likely to straddle small area boundaries. Hence evidence of local clustering across sets of neighbouring areas (i.e. a between area perspective) is of relevance to resource allocation and identification of possible causal influences, whether socio-economic or environmental (Han et al, 2005; Bell et al, 2008). Elevated risk identified in a particular area may not necessarily coincide with common elevated risk both in that area and its surrounding locality of nearby areas (close in distance terms to a particular area, or adjacent to that area). Similar issues occur in development of indices measuring area social structure (Rae, 2009). For example Dietz (2002) mentions that "the neighborhood effect may be within or among neighborhoods. In almost all cases, the existing research examines within neighborhood effects.[...]Thus, neighborhoods with identical characteristics but dissimilar neighboring neighborhoods are considered equivalent".

There are several possible global indices for measuring spatial correlation and clustering across a region, but for identifying clustering in subregions, local indices of spatial association are more appropriate. Local spatial correlation analysis has most commonly been applied to known area outcomes such as house prices (Anselin et al, 2006a). By contrast, in spatial analysis of small area health outcomes under a Bayesian perspective, the outcomes, namely relative disease risks, are taken to be unknowns, and inferences about them are based on stochastic modelling. Typically the data on which the modelling is based consist of

actual and expected disease counts $\{y_i, e_i\}$. Especially when e_i is small, maps of maximum likelihood estimates of risk y_i/e_i will be distorted by variance instability (Lawson et al, 2000; Anselin et al, 2006b).

In contrast to maximum likelihood estimation, hierarchical modelling of disease count data using random effects is oriented to borrowing strength between prior and likelihood, and also takes account of the spatial correlation in latent area risks r_i. Hierarchical models generally focus on the posterior probabilities of elevated risks (often called exceedance probabilities) in each area separately, namely the probability that $r_i > \tau_r$, where τ_r is some threshold. The focus on exceedance does not usually pay regard to the broader local clustering pattern around each area. For example, a high risk area may be an outlier (in local terms) if it is surrounded by low risk areas, whereas a high risk cluster centre would occur when both the area and its surrounding locality of nearby areas demonstrate common elevated risk.

The present analysis discusses the detection and measurement of different forms of local clustering in hierarchical models for small area disease counts. It focusses especially on local join-count statistics as these enable discrimination between different types of clustering. As discussed in section 2, the method proposed here permits decomposition of the exceedance probabilities into a cluster member probability and an outlier probability, and allows integrated inferences on both high and low risk clustering. For exploratory analysis with a small number of areas, a visual assessment of clustering in exceedance probabilities may be sufficient, but this does not provide the additional insights possible using the decomposition of exceedance probabilities using local join-counts (see section 2.3 for additional discussion).

The methodology is here applied to the case where risks, and hence risk status, are unknowns. A Bayesian estimation strategy is adopted with prior densities specified on unknown parameters, and posterior inferences based on an MCMC estimation. Boots (2003, 2006) considers local indices of spatial association for binary mapped data with particular focus on raster data, and with known binary values for each spatial unit. Here the focus is on detecting risk clustering in irregular area lattice data, with latent binary status, varying between iterations in terms of an MCMC estimation.

Applications of the methodology include assessing spatial heterogeneity in effects of area predictors (e.g. deprivation) according to local risk configuration, and gauging sensitivity of bivariate clustering to random effect assumptions. Two case studies are carried out. The first considers spatial heterogeneity in effects of area deprivation on emergency hospitalisations for chronic obstructive pulmonary disease (COPD). The spatial framework is 113 small areas (Middle Super Output Areas or MSOAs) in four London boroughs. The second application uses the same areas but considers bivariate clustering in rates of child obesity at different ages (pre-primary and end-primary ages).

2. Measures for Local Clustering in Relative Disease Risks

2.1 *Join Counts for Global Disease Risk Clustering*

Indices to summarise spatial association patterns in geographically close areas depend on the form of indicator for each area (e.g. whether continuous or binary) and how spatial interaction between areas is measured (Stevens and Jenkins, 2000). In small area health applications, a cluster is generally defined as a set of neighboring high-risk areas (i.e. a high risk cluster) or of neighboring low-risk areas (i.e. a low risk cluster), but other association patterns may occur.

Consider binary measures of health risk status b_i for areas $i = 1, ..., n$, namely $b_i = 1$ for elevated risk, $b_i = 0$ otherwise. For example, if area relative risks r_i average 1 over all areas in a region, then one could define $b_i = 1$ for $r_i > \tau_r$, where τ_r is a high risk threshold, most commonly a default value $\tau_r = 1$ (Wakefield and Kim, 2013), and $b_i = 0$ otherwise.

Let symmetric weights $w_{ij} = w_{ji}$ measure spatial interaction (with $w_{ii} = 0$). One approach to detecting *global* spatial clustering for binary mapped data uses join-count statistics, based on concordance in status between pairs of areas. Thus the total of area joins where both b_i and b_j are 1 (elevated risk in both areas i and j) is

$$J_{11} = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} b_i b_j, \qquad (1)$$

with the most commonly used binary joint count statistic being $0.5J_{11}$, also known as the BB statistic (e.g. Bell et al, 2008). The total of joins where b_i and b_j are discrepant (differing health risk status in area i as compared to area j) can be written

$$J_{10} = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} (b_i - b_j)^2, \qquad (2)$$

while total joins where both b_i and b_j are 0 (low risk in both areas) can be written

$$J_{00} = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} (1 - b_i)(1 - b_j). \qquad (3)$$

Letting $S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$, it may be noted that the observed join-counts are interdependent, with $S_0 = J_{11} + J_{10} + J_{00}$, so if two are known the other is obtained by subtraction from S_0 (McKenzie et al, 2008). For binary w_{ij} based on adjacency, the join-counts (J_{11}, J_{10}, J_{00}) are integers and can be regarded as multinomial with unknown probabilities $(\pi_{11}, \pi_{10}, \pi_{00})$ and total sample size S_0.

2.2 *Local Join Counts for Local Clustering*

One can obtain a *local* version of the join-count statistics in (1)-(3), namely join-counts regarding area i as the focus. Such local join counts are relevant to assessing whether area i and the areas surrounding it form a high risk cluster, or demonstrate some other localized risk pattern. For measuring common high risk, when both area i and its neighbouring areas tend to be high risk, one obtains

$$J_{11i} = b_i \sum_{j=1}^{n} w_{ij} b_j.$$

For measuring common low risk (when both area i and the areas close to it tend to be low risk) one obtains

$$J_{00i} = (1 - b_i) \sum_{j=1}^{n} w_{ij} (1 - b_j).$$

In the case of discrepant risk status between area pairs when the focus is on area i, it is potentially important for describing locality risk patterns to distinguish high-low risk pairings ($b_i = 1$, $b_j = 0$) from low-high risk pairings ($b_i = 0$, $b_j = 1$). The two local join count statistics in these cases are respectively

$$J_{10i} = b_i \sum_{j=1}^{n} w_{ij} (1 - b_j),$$

$$J_{01i} = (1 - b_i) \sum_{j=1}^{n} w_{ij} b_j.$$

Commonly spatial interactions w_{ij} are taken to binary, and based on whether areas i and j are adjacent ($w_{ij} = 1$) or not ($w_{ij} = 0$). In this case, let N_i denote the neighbourhood of area i, namely the set of areas adjacent to area i (those with $w_{ij} = 1$), and assume that this neighbourhood contains L_i areas (the total number of areas around i). Then the above local join-count statistics become

$$J_{11i} = b_i \sum_{j \in N_i} b_j,$$

$$J_{10i} = b_i \sum_{j \in N_i} (1 - b_j),$$

$$J_{01i} = (1 - b_i) \sum_{j \in N_i} b_j,$$

$$J_{00i} = (1 - b_i) \sum_{j \in N_i} (1 - b_j).$$

The total number of joins with area i as the focus when w_{ij} is binary is $L_i = S_{0i}$, so that $L_i = J_{11i} + J_{00i} + J_{10i} + J_{01i}$.

2.3 *Probabilities of Exceedance and their Relationship to Local Clustering Indicators*

Let $E_i = E(b_i)$ be the marginal probability of elevated risk in area i. The advantage of the local join count method is that this probability may be disaggregated to reflect risk clustering in the vicinity of area i. Continuing the assumption of binary adjacency, the proportion of joins π_{11i} focussed on area i that are joint high risk, defined by

$$E(J_{11i}) = L_i \pi_{11i},$$

provides a summary probability index that area i is a member of a high risk cluster. By contrast, the proportion of joins π_{10i} focussed on area i that are $(1, 0)$ pairs, defined by

$$E(J_{10i}) = L_i \pi_{10i},$$

provides an index that area i is a high risk local outlier, though this probability may also be relatively high for areas on cluster edges, which may be adjacent to low risk areas as well as areas in the high risk cluster. The join count $J_{11i} = b_i \sum_{j \in N_i} b_j$ can be written as $I(b_i = 1) \sum_{j \in N_i} b_j$, and it follows that

$$I(b_i = 1)L_i = J_{11i} + J_{10i} \qquad (4)$$

and so that

$$E_i = \pi_{11i} + \pi_{10i}.$$

So the marginal high risk probability is the sum of the high risk cluster member probability and the high risk local outlier probability. If all neighbours of area i are classified as high risk, then $\pi_{11i} \simeq E_i$, and area i can be considered as a high risk cluster centre (i.e. an area embedded in a high risk cluster, with all surrounding areas being high risk). If area i is high risk, but not all neighbours of area i are high risk (e.g. for cluster edge areas), then π_{10i} will tend to become relatively larger as a share of E_i (this situation is considered in the simulated data example of section 3.3). For high risk outliers π_{10i} will generally predominate over π_{11i}.

As to low risk clustering, the proportion of joins focussed on area i that are joint low risk, π_{00i}, as defined by $E(J_{00i}) = L_i\pi_{00i}$, provides a summary index that area i is the member of a low risk cluster. The proportion of joins focussed on area i that are (0,1) pairs, π_{01i}, as defined by $E(J_{01i}) = L_i\pi_{01i}$, provides an index that area i is a low risk area, but surrounded by high or medium risk areas (i.e. a low risk outlier). Since the join count $J_{00i} = (1 - b_i) \sum_{j \in N_i} (1 - b_j)$ can be written as $I(b_i = 0) \sum_{j \in N_i} (1 - b_j)$, one has

$$I(b_i = 0)L_i = J_{00i} + J_{01i},$$

and so that

$$1 - E_i = \pi_{01i} + \pi_{00i}.$$

Table 1 contains a glossary of the above parameters.

The parameters $(\pi_{11i}, \pi_{10i}, \pi_{01i}, \pi_{00i})$ are subject to $\pi_{11i} + \pi_{10i} + \pi_{01i} + \pi_{00i} = 1$, and represent probabilities that joins focused on area i show common high risk, discordant risk status between the focus area and its locality (high-low, or low-high), or common low risk. If spatial interaction is based on adjacency, these represent multinomial probabilities of generating the join pattern $(J_{11i}, J_{10i}, J_{01i}, J_{00i})$ among the L_i joins focused on area i.

The information provided by the above decompositions of E_i and $1 - E_i$ allows direct inferences on clustering patterns (both on high risk and low risk clustering). It extends to allowing identification of spatial outliers. This strategy can be adapted to allow for model uncertainty during MCMC estimation, as in methods which allow retention (or not) of sets of random effects or covariates (e.g. Scheipl et al, 2012). One might instead consider analysis of posterior mean estimates of the exceedance rates E_i themselves, for example by cluster analysis of the E_i subject to contiguity constraints (e.g. Recchia, 2010). Such a strategy does not, however, allow for model uncertainty, and raises questions about sensitivity of inferences to the choice of clustering algorithm. One could possibly include a contiguity constrained clustering algorithm within the MCMC sampling, so that at each iteration t the risks $r_i^{(t)}$, or status indicators $b_i^{(t)}$, are subject to clustering, but this may be computationally intensive. Another possible strategy involves simply visual assessment of clustering in exceedances at a set threshold (for example, $E_i > 0.8$, in the case of high risk clustering), and this might

be feasible for a relatively small number of areas. The latter strategy may be sensitive to cutpoints on the E_i, and becomes infeasible for a large number of areas, such as the 3144 US counties, or MSOAs across England.

3. Discrimination between Local Clustering Patterns under Relative Risk Models

3.1 *Modelling and Assessing Relative Risk*

In disease mapping applications the classification of an area as high or low risk typically depends on unknown parameters. Consider disease counts $(y_i, i = 1, .., n)$, with expected values e_i obtained by multiplying area populations by the region-wide disease rate, and with $\sum_i y_i = \sum_i e_i$. Then subject to the necessity to take account of overdispersion, the y_i may be taken as Poisson,

$$y_i \sim Poi(e_i r_i), \qquad (5)$$

where r_i denotes relative disease risk in area i. A number of studies support the utility of the convolution prior of Besag et al (1991) (or BYM prior after the authors) in analyzing spatial health variations and identifying areas with excess risk (Lawson et al, 2000; Richardson et al, 2004). This model involves two sets of random effects for each area, namely s_i representing a relatively smooth underlying spatial effect, and an *iid* random effect u_i to represent overdispersion. Assuming spatial interaction is represented by binary adjacency, one has the intercept only regression

$$log(r_i) = \beta_0 + u_i + s_i, \qquad (6)$$

where the u_i are *iid* Normal errors, $u_i \sim N(0, \sigma_u^2)$, while the s_i follow a conditional spatially autoregressive scheme, $s_i | s_{[i]} \backsim N\left(\sum_{j \in N_i} s_j / L_i, \sigma_s^2 / L_i\right)$, where $s_{[i]}$ represents the collection of s effects excluding s_i, and L_i is the number of areas adjacent to area i.

Disease mapping models often have particular focus on assessing the probability of elevated risk in a particular area. Assuming a Bayesian estimation approach with sampling over MCMC iterations $t = 1, .., T$, let $r_i^{(t)} = \exp(\beta_0^{(t)} + u_i^{(t)} + s_i^{(t)})$ denote the relative risk in area i at iteration t under the BYM prior. Then one may define binary indicators $b_i^{(t)} = I(r_i^{(t)} > \tau_r)$, where τ_r is a high risk threshold (most commonly $\tau_r = 1$), with posterior probabilities of locally elevated risk $Pr(r_i > \tau_r | y) = \Pr(b_i = 1 | y)$ then estimated as $\widehat{E}_i = \frac{1}{T} \sum_{t=1}^{T} b_i^{(t)}$.

Areas with probabilities $\widehat{E}_i > \tau_E$ exceeding some threshold probability (e.g. τ_E=0.8 or 0.9) can be classified as high risk. Decision rules for suitable (τ_r, τ_E) providing optimal trade off between false positive and false negative classifications have been proposed; for example, Richardson et al (2004) propose $(\tau_r = 1, 0.7 < \tau_E < 0.8)$ based on a simulation study.

Detection of *low* risk can be based on binary indicators $d_i^{(t)} = 1 - b_i^{(t)} = I(r_i^{(t)} < \tau_r)$. Posterior probabilities of locally depressed risk $Pr(r_i < \tau_r | y) = \Pr(b_i = 0 | y)$ are estimated

as $\widehat{D}_i = 1 - \widehat{E}_i = \frac{1}{T}\sum_{t=1}^{T} d_i^{(t)}$, and classification of areas as low risk can be made for $\widehat{D}_i$ sufficiently high.

3.2 *Cluster Detection Rules based on Disease Risk Models*

However, significance assessments or detection rules for each area separately detect only "hot spot" exceedance (i.e. single areas with elevated risk) (Lawson, 2013, p. 123), and do not provide evidence on clustered patterns of risk in the neighbourhood encompassing each area and surrounding areas. Building on the decision rules used for detecting single area elevated risk, may be proposed cluster detection rules. From the indicators $b_i^{(t)} = I(r_i^{(t)} > \tau_r)$ at each MCMC iteration, local join counts, namely $J_{11i}^{(t)}, J_{10i}^{(t)}, J_{01i}^{(t)}$ and $J_{00i}^{(t)}$, can be obtained. In turn estimates of π_{11i} and π_{10i} are obtained as $\widehat{\pi}_{11i} = \sum_{t=1}^{T} J_{11i}^{(t)}/(TL_i)$, and $\widehat{\pi}_{10i} = \sum_{t=1}^{T} J_{10i}^{(t)}/(TL_i)$. It follows from relation (4) that

$\widehat{E}_i = \widehat{\pi}_{11i} + \widehat{\pi}_{10i}$,

namely the estimated marginal probability of elevated risk (a frequent focus in relative risk models) can be obtained as the sum of the estimated high risk cluster member probability and the estimated high risk local outlier probability. Similarly from the relation $I(b_i = 0)L_i = J_{00i} + J_{01i}$ one has

$\widehat{D}_i = \widehat{\pi}_{00i} + \widehat{\pi}_{01i}$,

namely that the estimated marginal probability of depressed risk can be obtained as the sum of the low risk cluster probability estimate and the local low risk outlier probability estimate.

Based on these relations, one may categorise areas according to their local cluster configuration, and such configurations may have relevance in analysing heterogeneity in outcomes and predictor effects (see section 4). Clusters of areas according to risk could be obtained by conventional cluster analysis methods but the approach here reflects both varying risk levels and spatial clustering patterns. For example, supposing n_1 areas are classified as high risk (e.g. if $\widehat{E}_i > 0.95$), one may subdivide such areas into n_{11} high risk cluster centres (e.g. when both $\widehat{E}_i > 0.95$ and $\widehat{E}_i > \widehat{\pi}_{11i} > 0.9$), and n_{10} other high risk areas (when $\widehat{E}_i > 0.95$ but $\widehat{\pi}_{11i} < 0.9$). This other high risk group encompasses (a) high risk "cluster edge" areas, where $\widehat{\pi}_{11i}$ is lowered (as compared to cluster centre areas) because the area has both high risk and low risk neighbours, and (b) high risk local outliers with mainly low risk neighbours. Similarly, supposing there are n_0 low risk areas (e.g. areas with $\widehat{D}_i > 0.95$), one may subdivide this total into n_{00} low risk cluster centres (e.g. if both $\widehat{D}_i > 0.95$ and $\widehat{D}_i > \widehat{\pi}_{00i} > 0.9$), and n_{01} other low risk areas (if $\widehat{D}_i > 0.95$ but $\widehat{\pi}_{00i} < 0.9$). There are likely to remain $n_I = n - n_1 - n_0$ residual (intermediate risk) areas with neither high $\widehat{E}_i$ nor high $\widehat{D}_i$.

The above scheme implies clustering into 5 groups based both on varying risk and on the associated spatial clustering patterns. This provides an alternative to conventional cluster analysis in specifically taking account of spatial clustering. More disaggregated clustering

schemes (with more than 5 groups) could be devised. For example, one could distinguish within the "other high risk" group a further sub-group of local high risk outliers (e.g. where $\widehat{E}_i > 0.95$ and $\widehat{\pi}_{10i} > \widehat{\pi}_{11i}$), and within the "other low risk" group a further sub-group of local low risk outliers (where $\widehat{D}_i > 0.95$ and $\widehat{\pi}_{01i} > \widehat{\pi}_{00i}$).

3.3 *Simulations for Known Clustering Scenarios*
Varying frequencies of events, and location of areas within clusters, may both affect classification of areas as high risk and their cluster status, for example as cluster centres or edges. To demonstrate such effects for known clustering patterns, we use the same spatial framework as in the case studies in sections 4 and 5, namely 113 MSOAs (Middle Level Super Output Areas) in four London boroughs (Barking & Dagenham, Havering, Redbridge, Waltham Forest).

Under the first scenario (scenario A), three clusters of MSOAs with elevated risk are defined, with the relative risk set at 1.75 in the 15 areas in the three clusters, and with relative risk (background risk) in remaining 98 areas set at $RR_B = 0.89$. Figure 1 shows the location of the three high risk clusters. It is apparent that cluster centres (high risk areas completely surrounded by other high risk areas) are area 21 in cluster 1 (areas 16,17,18,21), areas 23 and 25 in cluster 2 (areas 22,23,25,27,28), and area 66 in cluster 3 (areas 61, 63, 66, 68, 72, 74).

Three different average event frequencies are considered: average e_i of 20, 60 and 100, with corresponding total expected events of 2260, 6780 and 11300. In practice, because y_i are necessarily integers, there will be minor calibration around the products $e_i RR_B$ to ensure $\sum_i y_i = \sum_i e_i$. Thus when $\sum_i e_i = 6780$, one has $y_i = 1.75 \times 60 = 105$ in the 15 high risk areas, and $y_i = 53$ (the integer part of $0.89 \times 60 + 0.5$) in all remaining areas, except for 11 areas with $y_i = 54$, to ensure that $\sum_i y_i = 6780$.

We adopt the likelihood and model as in equations (5)-(6). The risk threshold to decide risk status b_i is set at $\tau_r = 1$. Appendix 1 contains a listing of the Winbugs code for this analysis. Inferences are based on the second halves of two chain runs of 50,000, with convergence assessed according to BGR statistics (Brooks & Gelman, 1998). For each of the 15 high risk areas, Table 2 reports the hot spot probabilities $\widehat{E}_i = Pr(b_i = 1|y)$, high risk cluster member probabilities $\widehat{\pi}_{11i}$ and high risk outlier probabilities $\widehat{\pi}_{10i}$. Also shown are the average of these two probabilities in the 98 areas not in the high risk clusters. Table 2 further shows the number of areas adjacent to each of the 15 high risk areas, and the number of such neighbours which are high risk under the particular clustering scenario shown in Figure 1. For example, area 61 is a "cluster edge" area, adjacent to two high risk areas, but also to three areas with background risk.

It can be seen from Table 2 that for a relatively low frequency event (average $e_i = 20$),

the four cluster centre areas (21, 23, 25 and 66) have posterior mean π_{11i} of 0.96, 0.98, 0.98 and 0.96 respectively, only slightly below the estimated E_i. The 11 remaining areas which are in the high risk clusters, but are also adjacent to some background risk areas, have lower $\widehat{\pi}_{11i}$ values (between 0.69 and 0.82), and $\widehat{\pi}_{10i}$ values which are a higher proportion of the total exceedance probability estimate. The lowest estimated $\widehat{\pi}_{11i} = 0.69$ is for the "cluster edge" area 61. The remaining 98 non-cluster areas have $\widehat{\pi}_{11i}$ averaging 0.18, and with a maximum of 0.47. The pair of clustering probabilities $\{\widehat{\pi}_{11i},, \widehat{\pi}_{10i}\}$ thus together distinguish the high risk clusters from remaining background risk areas, and cluster centres from cluster edges.

For more frequent events (with average e_i of 60 or 100), the classification error for the cluster centres is eliminated, with estimated E_i and π_{11i} both equal to 1 for areas 21, 23, 25 and 66. The cluster edge status of the remaining 11 high risk areas also becomes more apparent: for the highest frequency (average e_i=100), five such areas (17,18, 28, 61 and 72) have $\widehat{\pi}_{11i}$ under 0.6, while $\widehat{E}_i = 1$.

Another scenario (scenario B) is developed to compare the cluster status probability estimates between two situations: (i) when high risk characterises all neighbours surrounding area i (so area i is a true cluster centre), and risk is evenly distributed among such neighbors, and (ii) when high risk is not common to all neighbours, but unevenly concentrated among a few neighbors, so area i is a potential cluster edge. Consider the most westerly cluster (cluster 3) in Figure 1. Situation (i) is already considered under scenario A, which has a relative risk of 1.75 across areas (61,63,66,68,72,74), and 66 represents a true cluster centre. To represent the uneven risk situation (ii), instead of assuming area 68 is high risk, we assume $r_{68} = 0.34$, with area 66 remaining with risk 1.75, while areas 61, 63, 72 and 74 have risks set at 2.15 (see Figure 2). The average risk across the six areas remains similar to scenario A, namely 1.78, but the risk pattern is more uneven, with area 66 now no longer a cluster centre (and even a potential cluster edge), and area 68 now potentially a low risk outlier. The risk configuration across the remaining 107 areas remains as in scenario (A), and the average e_i is set at 60.

Primary goals here are to assess whether the π_{11i} distinguish between the two scenarios regarding the west cluster, especially the status of area 66, and whether the π_{01i} reflect the low risk in area 68. Table 3 shows that under situation (ii), area 68 now has $\widehat{E}_i = 0$, and also has a low risk outlier probability of $\widehat{\pi}_{01i} = 0.65$, reflecting its own low risk but its contiguity to high risk areas. As to area 66, the even risk scenario (upper panel) has $\widehat{E}_i = \widehat{\pi}_{11i} = 1$, while the uneven risk scenario (lower panel) has $\widehat{E}_i = 1$, but $\widehat{\pi}_{11i}$ reduced to 0.75. Despite the fact that three of its four neighbours have elevated risk (RR=2.15), this area is no longer recognized as a cluster centre. Given that the relevant estimated $\widehat{E}_i$ are either 1 or 0, the value $\widehat{\pi}_{11i} = 0.75$ reflects the fact that one of the four neighbours is now low risk.

4 Case Study 1: Impacts of Deprivation on COPD Emergency Admissions.

We now consider real data for the same set of areas as in the preceding simulations. Thus

chronic obstructive pulmonary disease (COPD) is a major cause of emergency (unplanned) hospital admission, and in the UK there is concern about the rise in such admissions (Jones, 2009). Emergency admissions for respiratory conditions such as COPD may be classed as ambulatory sensitive, that is potentially avoidable with effective preventive care in primary and community settings (Tian et al, 2012). The geographical distribution of COPD emergency admissions is therefore policy relevant.

The analysis uses admission totals y_i over the period 2006/07 to 2010/11 in the 113 MSOAs. Expected admissions e_i are based on an England wide schedule of age specific rates, with scaling to ensure $\sum_i y_i = \sum_i e_i$. The average admission count is 77.1. We consider models with and without ecological predictors, and show how a cluster configuration scheme (see section 3.2) may show further light on heterogeneity in event risk and predictor effects.

A baseline model (model 1) is provided by the Besag et al (1991) convolution or BYM model with intercept only:

$log(r_{i0}) = \beta_{00} + u_{i0} + s_{i0}$.

A flat prior on β_{00} is assumed, and a gamma prior on the precision $1/\sigma_s^2$ with index 1 and shape 0.001 (e.g. Besag et al, 1995; Higdon, 2007). Improved convergence is obtained by linking random effect precisions; thus $1/\sigma_u^2 = \rho/\sigma_s^2$, where ρ is assigned an exponential prior with rate 1. Inferences from this and subsequent models are based on the second halves of two chain runs of 50,000, with convergence assessed according to BGR statistics (Brooks & Gelman, 1998). Fit is assessed using the Deviance Information Criterion or DIC (Spiegelhalter et al, 2002).

Estimates of localized (i.e. hotspot) status $\widehat{E}_i$ are obtained by monitoring over iterations t the binary indicators $b_i^{(t)}$ (when $r_i^{(t)} > \tau_r = 1$), from which one obtains local join counts $J_{11i}^{(t)}, J_{10i}^{(t)}, J_{01i}^{(t)}$ and $J_{00i}^{(t)}$. Various thresholds on $\widehat{E}_i$ for assigning elevated risk make little difference to totals of areas classed as having hotspot status. For example, taking $\widehat{E}_i > 0.95$ gives $n_1 = 38$ high risk areas, while taking $\widehat{E}_i > 0.75$ gives $n_1 = 44$ high risk areas.

Cluster configuration categories (section 3.2) may be defined on the basis of this baseline model. Thus areas are defined as high risk cluster centres if both $\widehat{E}_i > 0.95$ and $\widehat{\pi}_{11i} > 0.9$, and other high risk if $\widehat{E}_i > 0.95$ but $\widehat{\pi}_{11i} < 0.9$, giving $n_{11} = 10$ and $n_{10} = 28$. Low risk areas are defined for $\widehat{D}_i > 0.95$, as low risk cluster centres if both $\widehat{D}_i > 0.95$ and $\widehat{\pi}_{00i} > 0.9$, and as other low risk if $\widehat{D}_i > 0.95$ but $\widehat{\pi}_{00i} < 0.9$, giving $n_{00} = 21$ and $n_{01} = 23$. There are $n_I = n - (n_1 + n_0) = 113 - (38 + 44) = 31$ unclassified (intermediate risk) areas.

Differences in COPD risk levels according to these categories is apparent first if admissions and expected events are cumulated over areas within the categories, providing what are sometimes called standard hospitalisation ratios (SHR). The SHR for the high risk cluster centres is 1.78, for other high risk areas is 1.57. By contrast, the SHR for low risk cluster

centres is 0.56, and for other low risk areas is 0.65. The averages of the modelled relative risks (r_i) over areas within each of the cluster categories are very similar: 1.74 for high risk cluster centres, 1.58 for other high risk areas, 0.57 for low risk cluster centres and 0.66 for other low risk areas.

Figure 3 maps out these categories and provides a single graphical summary of varying types of local clustering. It can be seen that there is a belt of nine contiguous areas in the centre south of the region which are all classified as high risk cluster members. So definitive and spatially continuous high risk, spreading across administrative area boundaries, is most evident here, and would provide a basis for a public health intervention to establish why there are so many emergencies, or promote higher uptake of preventive strategies such as COPD self-management education (Purdy, 2010). High risk also appears in the south west of the region but is more fragmented, with intermediate risk areas also present. There are two belts of areas containing low risk cluster centres, demonstrating definitive low risk that spreads across administrative area boundaries.

Emergency hospital admissions are often positively related to area social deprivation (Simpson and Hippisley-Cox, 2010), and so a predictor x_i is provided by the logarithm of the percent of households in poverty (2007/2008) in each MSOA. In the model

$log(r_{i1}) = \beta_{01} + u_{i1} + s_{i1} + x_i\beta_{11}$,

a N(0,1000) prior is assigned to β_{11}. This model provides a better fit, although the reduction in DIC is not large (898.2 as against 901.1). The posterior mean (95% CrI) on β_{11} is 0.76 (0.48, 1.05). In order to assess spatial heterogeneity in predictor effects (e.g. Sridharan et al, 2011) a varying coefficient model is applied with

$log(r_{i2}) = \beta_{02} + u_{i2} + s_{i2} + x_i(\beta_{12} + v_{i2})$,

where v_{i2} is a random effect following a conditional autoregressive prior (see section 3.1) with zero mean over all n areas. This model produces a further slight reduction in DIC to 897.1.

However, despite the modest gain in global fit, other forms of heterogeneity may be present. Of possible policy or epidemiological significance is variability in predictor effects across cluster configuration categories; for example, if the effect of deprivation on COPD emergency admissions is more marked in high risk cluster centres, this could reflect less effective community care in deprived areas. Accordingly the averages of the coefficients $\beta_{12i} = \beta_{12} + v_{i2}$ over areas within each of the five configuration categories are obtained at each iteration. Let $\{B_k^{(t)}, k = 1, .., 5\}$ denote these averages. Heterogeneity in predictor effects between configuration categories can be assessed via the indicators $C_{jk}^{(t)} = I(B_j^{(t)} > B_k^{(t)})$, with posterior means providing estimates that $\Pr(B_j > B_k|y)$.

Table 4 shows posterior intervals for the B_k, and estimates for the probabilities $\Pr(B_j > B_k|y)$. It is apparent that the deprivation effect is stronger for high risk areas, whether such areas are cluster centres ($k = 5$) or other high risk areas ($k = 4$). All but one of the ten comparisons between-cluster category coefficients B_k show significant differences.

5 Case Study 2: Bivariate Clustering in Health Risk, Child Obesity Rates.
The clustering criteria developed above are designed for the case where area risk status is uncertain a priori and is based on a statistical model. Since a Bayesian approach is adopted one facet of such a model is the prior specification. Thus assessment of bivariate (or more generally multivariate) clustering patterns may be affected by alternative assumptions regarding random effects interdependence, for example, whether random effects are correlated between outcomes as well as over areas. Such an assessment can be made using indicators of joint outcome high risk clustering.

This case study accordingly considers interdependence in clustering patterns between two outcomes (denoted A, B) under different assumptions. The outcomes are again for the 113 MSOAs in Outer NE London, and are: y_{Ai}, totals of children assessed as obese (accumulated over annual readings during a three year period 2008/9 to 2010/11) at pre-primary ages (4 to 5); y_{Bi}, totals of children assessed as obese at end-primary ages (10 to 11). The average obese count in each MSOA at ages 4-5 is 34.1, and the average at ages 10-11 is 55.8. Expected totals e_{Ai} and e_{Bi} are obtained by multiplying total child populations in each MSOA by the regional obesity rates, 0.117 and 0.214. This ensures $\sum_i y_{Ai} = \sum_i e_{Ai}$, and $\sum_i y_{Bi} = \sum_i e_{Bi}$.

Rising child obesity is a major focus of public health concern in the UK and elsewhere (National Audit Office, 2006), and elevated bivariate risk (high obesity at both pre-primary and end-primary ages) in spatially contiguous areas provides a rationale for targeted intervention (National Obesity Observatory, 2011). It would also be possible to focus on pre-primary obesity and increased obesity between pre-primary and end-primary ages, since some areas may show greater increases in child obesity during primary ages.

Consider the joint high risk indicators

$b_{ABi} = I(r_{Ai} > 1, r_{Bi} > 1)$,

which can be obtained from monitoring $r_{iA}^{(t)}$ and $r_{iB}^{(t)}$ in the models

$log(r_{iA}) = \beta_{0A} + u_{iA} + s_{iA}$,

$log(r_{iB}) = \beta_{0B} + u_{iB} + s_{iB}$.

One may obtain exceedance probabilities $E_{ABi} = E(b_{ABi})$, as for the univariate case. To assess high risk joint clustering, the corresponding local join-counts under binary adjacency, and assuming $b_{Ai} = I(r_{Ai} > 1)$, and $b_{Ai} = I(r_{Bi} > 1)$, are

$J_{AB11i} = b_{ABi} \sum_{j \in N_i} b_{ABj} = b_{Ai} b_{Bi} \sum_{j \in N_i} b_{Aj} b_{Bj}$.

The proportion of joins π_{AB11i} focussed on area i that are joint high risk over both outcomes, defined by

$E(J_{AB11i}) = L_i \pi_{AB11i}$,

provides a summary probability index of that area i is a member of a high risk cluster on both outcomes. From sampled indicators $b_{ABi}^{(t)}$ at MCMC iterations $t = 1, .., T$, bivariate local join counts, namely $J_{AB11i}^{(t)}$ can be obtained, with estimates of π_{AB11i} obtained as

$\widehat{\pi}_{AB11i} = \sum_{t=1}^{T} J^{(t)}_{AB11i}/(TL_i)$. The remaining three local join counts and cluster status probabilities (adapted to the bivariate case) can be obtained analogously.

The number and spatial pattern of areas with elevated bivariate clustering may be compared according to different prior assumptions regarding random effects. Two alternative prior specifications regarding the random effects are compared. Under the first (model 1), the effects u_{iA} and u_{iB} are a priori uncorrelated with each other. However, the spatial effects s_{iA} and s_{iB} are taken to follow a bivariate conditional autoregressive (CAR) prior (e.g. Neelon et al, 2012). Then the conditional density of $s_i = (s_{iA}, s_{iB})$ is bivariate normal, with means

$E(s_{iA}|s_{[i]A}) = \sum_{j \in N_i} s_{jA}/L_i, \qquad E(s_{iB}|s_{[i]B}) = \sum_{j \in N_i} s_{jB}/L_i,$

and within area 2×2 precision matrix

$Prec(s_i|s_{[i]}) = L_i\zeta,$

where ζ is assigned a Wishart prior with 2 degrees of freedom and an identity scale matrix. This model has a DIC of 1534.6. Considering high risk exceedance probabilities, there are respectively 14, 16 and 4 areas with $\widehat{E}_{Ai} > 0.9, \widehat{E}_{Bi} > 0.9$, and $\widehat{E}_{ABi} > 0.9$. Numbers of areas with univariate cluster member probabilities $\widehat{\pi}_{A11i} > 0.85$ and $\widehat{\pi}_{B11i} > 0.85$ (representing high risk clustering in each outcome separately) are respectively 5 and 5. There are, however, no areas with probabilities of joint cluster member status ($\widehat{\pi}_{AB11i}$) exceeding 0.85. By contrast, considering low risk clustering, there are respectively 18 and 35 areas with univariate cluster member probabilities $\widehat{\pi}_{A00i} > 0.85$ and $\widehat{\pi}_{B00i} > 0.85$, and 14 areas with probabilities of joint cluster member status ($\widehat{\pi}_{AB00i}$) exceeding 0.85.

A second option (model 2) is a shared factor or shared component model, specified as

$log(r_{iA}) = \beta_{0A} + u_{iA} + \lambda_A s_i,$

$log(r_{iB}) = \beta_{0B} + u_{iB} + \lambda_B s_i,$

where s_i is a univariate conditional autoregressive prior. To ensure identification, it is assumed that $\lambda_A = 1$, with the variance of the s_i an unknown. Specifically, a gamma Ga(1,0.001) prior is assumed on the precision $1/\sigma_s^2$. The unknown loading λ_B is assigned an exponential E(1) prior. This model reduces the DIC to 1522.6, while the number of areas with probabilities $\widehat{\pi}_{A11i}$ and $\widehat{\pi}_{B11i}$ exceeding 0.85 are now respectively 14 and 17. There are also now 6 areas with estimated probabilities $\widehat{\pi}_{AB11i}$ of joint high risk cluster member status exceeding 0.85. As to low risk clustering, there are 25 areas with probabilities of joint cluster member status ($\widehat{\pi}_{AB00i}$) exceeding 0.85.

Figure 4 maps out the location of areas with high cluster member probabilities under this model. The 'high risk cluster' category shows areas with $\widehat{\pi}_{AB11i}$>0.85, and the 'low risk cluster' category shows areas with $\widehat{\pi}_{AB00i}$>0.85.

The proposed clustering indicators thus clearly demonstrate how spatial clustering in multiple indicators of a health outcome can be detected. It is also apparent that inferences about

the spatial patterning of bivariate risk may be influenced by alternative priors for borrowing strength between outcomes.

7 Conclusions

Spatial analyses of health outcomes at a small-area scale are important for assessing geographic heterogeneity in disease risk and detecting areas with elevated risk. Exceedance probabilities for each area separately (hotspot probabilities) are often used to summarise variations in relative risk, but may not necessarily coincide with common elevated risk both in an area and its surrounding locality of nearby areas.

The present study has suggested how local join count statistics can be used to detect risk clustering in area lattice data, and distinguish between cluster centres with elevated risk, cluster centres with depressed risk, cluster edge areas, and also outlier areas with dissimilar risk from their neighbours. A Bayesian perspective is proposed, and has utility when binary indicators of relative risk status are latent and determined by a statistical model. The methodology extends straightforwardly to multiple health outcomes, where routine spatial scan statistics are not available.

Applications of the methodology include development of a typology of areas taking account both of hotspot status and local clustering in high and low risk. This cluster typology detected pronounced risk variation in the analysis of COPD emergency admissions, and also showed heterogeneity in effects of area predictors according to cluster. Another potential application is in gauging sensitivity of joint clustering patterns implied by statistical modelling to alternative borrowing strength assumptions.

Possible adaptations of the methods proposed here include bivariate exploratory spatial analysis considering a health outcome (outcome A) and area risk factor (outcome B). The focus would be on detecting clusters of areas where elevated health risk and elevated levels of the risk factor coincide. Another possible variation would be trinary join counts (Zhang and Zhang, 2008), allowing for an intermediate risk category between low and high risk. A space-time perspective can also be developed using period-specific local join counts.

References

Anselin, L, Syabri L, Kho Y (2006a) GeoDa: An introduction to spatial data analysis. Geographical Analysis 38: 5-22

Anselin L, Lozano N, Koschinsky J (2006b) Rate Transformations and Smoothing, GeoDa Center Research Report (http://geodacenter.asu.edu/learning/tutorials).

Bell N,Schuurman N, Hameed S (2008) Are injuries spatially related? Join-count spatial autocorrelation for small-area injury analysis. Injury Prevention, 14(6), 346-353.

Besag J, York J, Mollie A (1991) Bayesian image restoration, with two applications in spatial statistics. Ann. Inst. Statist. Math., 43, 1-21.

Besag, J., Green, P., Higdon, D., Mengersen, K (1995) Bayesian computation and stochastic

systems, Statistical Science, 10, 3-66.
Boots B (2003) Developing local measures of spatial association for categorical data. Journal of Geographical Systems 5(2): 139-160
Boots B (2006) Local configuration measures for categorical spatial data: binary regular lattices. Journal of Geographical Systems, 8(1): 1-24
Brooks S, Gelman A (1998) General methods for monitoring convergence of iterative simulations. J Computational and Graphical Statistics, 7,434–455
Christensen R, Johnson W, Branscum A (2010) Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians. CRC Press.
Dietz R (2002) The estimation of neighborhood effects in the social sciences: an interdisciplinary approach, Social Science Research 31, 539–575
Han D, Carrow S, Rogerson P, Munschauer F (2005) Geographical variation of cerebrovascular disease in New York State: the correlation with income. Int J Health Geogr. 2005 Oct 21;4:25.
Higdon D (2007) A primer on space-time modeling from a Bayesian perspective. In Finkenstadt B, Held L, Isham, V (eds), Statistical Methods for Spatio-Temporal Systems, CRC, pp 217–279.
Jones R (2009) Trends in emergency admissions. British Journal of Healthcare Management, 15(4): 188-196.
Lawson A (2013) Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology, 2nd edition. CRC Press, New York.
Lawson A, Biggeri A, Boehning D, Lesaffre E, Viel J, Clark A (2000) Disease mapping models: an empirical evaluation. Stat Med 19:2217–2241
McKenzie, J, Egozcue J, Heilbronner R, Hielscher R, Müller A, Schaeben H (2008) Quantifying rock fabrics - a test of independence of the spatial distribution of crystals. CoDaWork 2008, Girona, Spain, May 27-30.
National Audit Office (2006) Tackling Child Obesity – First Steps. NAO London
National Obesity Observatory (2011) Child Obesity Statistics for PCT Clusters. National Obesity Observatory, London.
Neelon B, Anthopolos R, Miranda M (2012) A spatial bivariate probit model for correlated binary data with application to adverse birth outcomes. Stat Methods Med Res. 2012 May 16. [Epub ahead of print]
Purdy S (2010) Avoiding hospital admissions. What does the research evidence say? The Kings Fund, London
Rae A (2009) Isolated Entities or Integrated Neighbourhoods? An Alternative View of the Measurement of Deprivation, Urban Studies, 46(9): 1859–1878
Recchia, A (2010) Contiguity-Constrained Hierarchical Agglomerative Clustering Using SAS. Journal of Statistical Software, http://www.jstatsoft.org/v33/c02
Richardson S; Thomson A; Best N; Elliott P (2004) Interpreting posterior relative risk estimates in disease-mapping studies. Environ Health Perspect. 112:1016-1025.
Scheipl, F, Fahrmeir, L, Kneib, T (2012) Spike-and-slab priors for function selection in structured additive regression models. Journal of the American Statistical Association, 107:

1518-1532.
Simpson C, Hippisley-Cox J (2010) Trends in the epidemiology of chronic obstructive pulmonary disease in England: a national study of 51,804 patients. British Journal of General Practice, 60(576): 483–8.
Spiegelhalter, D, Best, N, Carlin, B, Van Der Linde, A. (2002) Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society, Ser B, 64: 583-639.
Sridharan S, Koschinsky J, Walker J (2011) Does context matter for the relationship between deprivation and all-cause mortality? The West vs. the rest of Scotland. Int J Health Geogr. 2011 May 12;10:33.
Stevens P, Jenkins D (2000) Analyzing species distributions among temporary ponds with a permutation test approach to the join-count statistic. Aquatic Ecology, 34, 91-99
Tian Y, Dixon A, Gao H (2012) Emergency hospital admissions for ambulatory care-sensitive conditions: identifying the potential for reductions. Kings Fund, London (http://www.kingsfund.org.uk/).
Wakefield J, Kim A (2013) A Bayesian model for cluster detection. Biostatistics. 2013 Mar 7. [Epub ahead of print]
Zhang S, Zhang K (2008) The Study on Trinary Join-Counts for Spatial Autocorrelation. Proceedings of the 8th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Shanghai, June 25-27, 2008, pp. 111-118

Appendix 1 Winbugs Code

The following is the code for the model $log(r_i) = \beta + u_i + s_i$ in the simulated data analysis. Convergence is improved by linking the precision parameters; thus $1/\sigma_u^2 = \rho/\sigma_s^2$, where ρ is assigned an exponential prior with rate 1. The vectors "map" in the code contains the binary adjacency matrix. With N areas and NN joins, the vector C is of length NN+1 and contains cumulated L_i, with initial element $C_1 = 0$, next element $C_2 = L_1$, next element the total $C_3 = L_1 + L_2$, following element $C_4 = L_1 + L_2 + L_3$, and so on.

```
model {for (i in 1:N) {# likelihood
y[i] ~dpois(mu[i]); mu[i] <- e[i]*r[i]
log(r[i]) <- beta0+u[i]+s[i]
u[i] ~dnorm(0,tau.u);
# risk status
b[i] <- step(r[i]-1);
# deriving join counts
for (j in C[i]+1:C[i+1]) { join11[i,j] <- b[i]*b.map[j]
join10[i,j] <- b[i]*(1-b.map[j])
join01[i,j] <- (1-b[i])*b.map[j]
join00[i,j] <- (1-b[i])*(1-b.map[j])}
J11[i] <- sum(join11[i,C[i]+1 : C[i+1]]); J10[i] <- sum(join10[i,C[i]+1 : C[i+1]])
J01[i] <- sum(join01[i,C[i]+1 : C[i+1]]); J00[i] <- sum(join00[i,C[i]+1 : C[i+1]])
# pi11, pi10, pi01, pi00
pi.L[1,i] <- J11[i]/L[i]; pi.L[2,i] <- J10[i]/L[i]; pi.L[3,i] <- J01[i]/L[i]; pi.L[4,i] <- J00[i]/L[i]}
```

```
# neighbourhood vector of risk status indicators
for (i in 1:NN) {b.map[i] <- b[map[i]]}
# priors
beta0 ~dflat(); tau.s ~dgamma(1,0.001); rho ~dexp(1); tau.u <- rho*tau.s
s[1:N] ~car.normal(map[], wt[], L[], tau.s)
for (i in 1:NN) { wt[i] <- 1}}
```

Table 1 Glossary of Symbols

r_i	Relative risk in area i
b_i	Binary risk status (=1 for elevated risk, =0 for depressed risk)
τ_r	Relative Risk threshold for defining risk status (e.g. τ_r =1)
w_{ij}	Elements of spatial interaction matrix W, continuous or binary (using adjacency)
S_0	$\sum_i \sum_j w_{ij}$, namely total interactions across entire map
S_{0i}	$\sum_j w_{ij}$, namely total interactions with area i as focus
L_i	Total neighbours of area i ($L_i = S_{0i}$ when W has binary elements)
N_i	Neighbourhood of area i (containing L_i areas adjacent to area i) when w_{ij} are binary
J_{11}	$\sum_i \sum_j w_{ij}\, b_i b_j$. Global black-black join count, namely total joins across entire map where both b_i and b_j are 1 (elevated risk in areas i and j)
J_{00}	$\sum_i \sum_j w_{ij}\, (1-b_i)(1-b_j)$. Global white-white join count, namely total joins across entire map where both b_i and b_j are 0 (depressed risk in areas i and j)
J_{11i}	$b_i \sum_j w_{ij} b_j$. Local black-black join count, with focus on area i, namely total joins where b_i=1 and neighbours also have b_j=1. When w_{ij} binary, $J_{11i}=b_i \sum_{j\in N_i} b_j$
J_{10i}	$b_i \sum_j w_{ij}(1-b_j)$. Local black-white join count, with focus on area i, namely total joins where b_i=1 but neighbours have b_j=0. When w_{ij} binary, $J_{10i}=b_i \sum_{j\in N_i} (1-b_j)$.
J_{01i}	$(1-b_i)\sum_j w_{ij} b_j$. Local white-black join count, with focus on area i, namely total joins where b_i=0 but neighbours have b_j=1. When w_{ij} binary, $J_{01i}=(1-b_i) \sum_{j\in N_i} b_j$.
J_{00i}	$(1-b_i)\sum_j w_{ij}(1-b_j)$. Local white-white join count, with focus on area i, namely total joins where b_i=0 and neighbours also have b_j=0. When w_{ij} binary, $J_{00i}=(1-b_i) \sum_{j\in N_i} (1-b_j)$.
E_i	$\Pr(b_i=1)$. High risk exceedance probability, probability that b_i=1.
D_i	$\Pr(b_i=0)=1-E_i$. Low risk exceedance probability.
π_{11i}	Probability of high risk cluster membership. For binary adjacency, $E(J_{11i})=L_i\, \pi_{11i}$
π_{10i}	Probability of high risk outlier. For binary adjacency, $E(J_{10i})=L_i\, \pi_{10i}$
π_{01i}	Probability of low risk outlier. For binary adjacency, $E(J_{01i})=L_i\, \pi_{01i}$
π_{00i}	Probability of low risk cluster membership. For binary adjacency, $E(J_{00i})=L_i\, \pi_{00i}$

Table 2 Detecting Elevated Risk (Hotspots and Clusters), Simulated Data
Estimated Exceedance and Cluster Status Probabilities For Differing Event Frequencies

			Average e_i=100			Average e_i=60			Average e_i=20		
Area identifier (see Figure 1)	**Total neigh-bours**	**Total high risk neigh-bours**	$\hat{E}_i$	$\hat{\pi}_{11i}$	$\hat{\pi}_{10i}$	$\hat{E}_i$	$\hat{\pi}_{11i}$	$\hat{\pi}_{10i}$	$\hat{E}_i$	$\hat{\pi}_{11i}$	$\hat{\pi}_{10i}$
16	**5**	**3**	**1.00**	**0.68**	**0.32**	**1.00**	**0.71**	**0.29**	**0.98**	**0.79**	**0.19**
17	**6**	**3**	**1.00**	**0.58**	**0.42**	**1.00**	**0.61**	**0.39**	**0.98**	**0.71**	**0.27**
18	**7**	**3**	**1.00**	**0.54**	**0.46**	**1.00**	**0.58**	**0.42**	**0.98**	**0.72**	**0.26**
21	**3**	**3**	**1.00**	**1.00**	**0.00**	**1.00**	**1.00**	**0.00**	**0.98**	**0.96**	**0.02**
22	**4**	**3**	**1.00**	**0.82**	**0.18**	**1.00**	**0.84**	**0.16**	**0.99**	**0.89**	**0.10**
23	**3**	**3**	**1.00**	**1.00**	**0.00**	**1.00**	**1.00**	**0.00**	**0.99**	**0.98**	**0.01**
25	**4**	**4**	**1.00**	**1.00**	**0.00**	**1.00**	**1.00**	**0.00**	**1.00**	**0.98**	**0.01**
27	**6**	**3**	**1.00**	**0.61**	**0.39**	**1.00**	**0.65**	**0.35**	**0.99**	**0.76**	**0.23**
28	**7**	**3**	**1.00**	**0.54**	**0.46**	**1.00**	**0.58**	**0.42**	**0.98**	**0.71**	**0.27**
61	**5**	**2**	**1.00**	**0.54**	**0.46**	**1.00**	**0.61**	**0.39**	**0.97**	**0.69**	**0.28**
63	**6**	**4**	**1.00**	**0.76**	**0.24**	**1.00**	**0.80**	**0.20**	**0.97**	**0.82**	**0.15**
66	**4**	**4**	**1.00**	**1.00**	**0.00**	**1.00**	**1.00**	**0.00**	**0.98**	**0.96**	**0.02**
68	**7**	**4**	**1.00**	**0.65**	**0.35**	**1.00**	**0.69**	**0.31**	**0.97**	**0.75**	**0.22**
72	**7**	**3**	**1.00**	**0.56**	**0.44**	**1.00**	**0.61**	**0.39**	**0.98**	**0.71**	**0.27**
74	**6**	**3**	**1.00**	**0.62**	**0.38**	**1.00**	**0.65**	**0.35**	**0.98**	**0.76**	**0.22**
Average, 15 High Risk Areas			**1.00**	**0.73**	**0.27**	**1.00**	**0.75**	**0.25**	**0.98**	**0.81**	**0.17**
Average, Other 98 Areas			**0.13**	**0.03**	**0.10**	**0.18**	**0.05**	**0.13**	**0.37**	**0.18**	**0.19**

Table 3 Cluster Centres and Cluster Edges under Different Risk Patterns

Known Cluster Pattern	Area identifier (see Figures 1 and 2)	$\hat{E}_i$	$\hat{\pi}_{11i}$ (High risk cluster member)	$\hat{\pi}_{10i}$ (High risk outlier)	$\hat{\pi}_{01i}$ (Low risk outlier)	$\hat{\pi}_{00i}$ (Low risk cluster member)
Scenario i (even risk, area 66 as cluster entre)	61	1	0.61	0.39	0	0
	63	1	0.80	0.20	0	0
	66	1	1	0	0	0
	68	1	0.69	0.31	0	0
	72	1	0.61	0.39	0	0
	74	1	0.65	0.35	0	0
Scenario ii (uneven risk, area 66 no longer cluster centre, area 68 low risk outlier)	61	1	0.37	0.63	0	0
	63	1	0.61	0.39	0	0
	66	1	0.75	0.25	0	0
	68	0	0.00	0.00	0.65	0.35
	72	1	0.59	0.41	0	0
	74	1	0.46	0.54	0	0

Table 4 Deprivation Effects on COPD Emergency Admissions by Cluster Configuration Category*

Deprivation Parameter	Mean	2.5%	97.5%
B_1	0.779	0.574	1.010
B_2	0.706	0.508	0.947
B_3	0.677	0.445	0.935
B_4	0.884	0.681	1.108
B_5	0.903	0.708	1.127

Probabilities of differing deprivation effects

$Pr(B_2>B_1\|y)$	0	$Pr(B_4>B_3\|y)$	1
$Pr(B_3>B_1\|y)$	0	$Pr(B_5>B_1\|y)$	1
$Pr(B_3>B_2y)$	0.034	$Pr(B_5>B_2\|y)$	1
$Pr(B_4>B_1\|y)$	1	$Pr(B_5>B_3\|y)$	1
$Pr(B_4>B_2\|y)$	1	$Pr(B_5>B_4\|y)$	0.878

*Categories: 5 High risk cluster centre, 4 Other high risk, 3 Low risk cluster centre, 2 Other low risk, 1 Intermediate risk

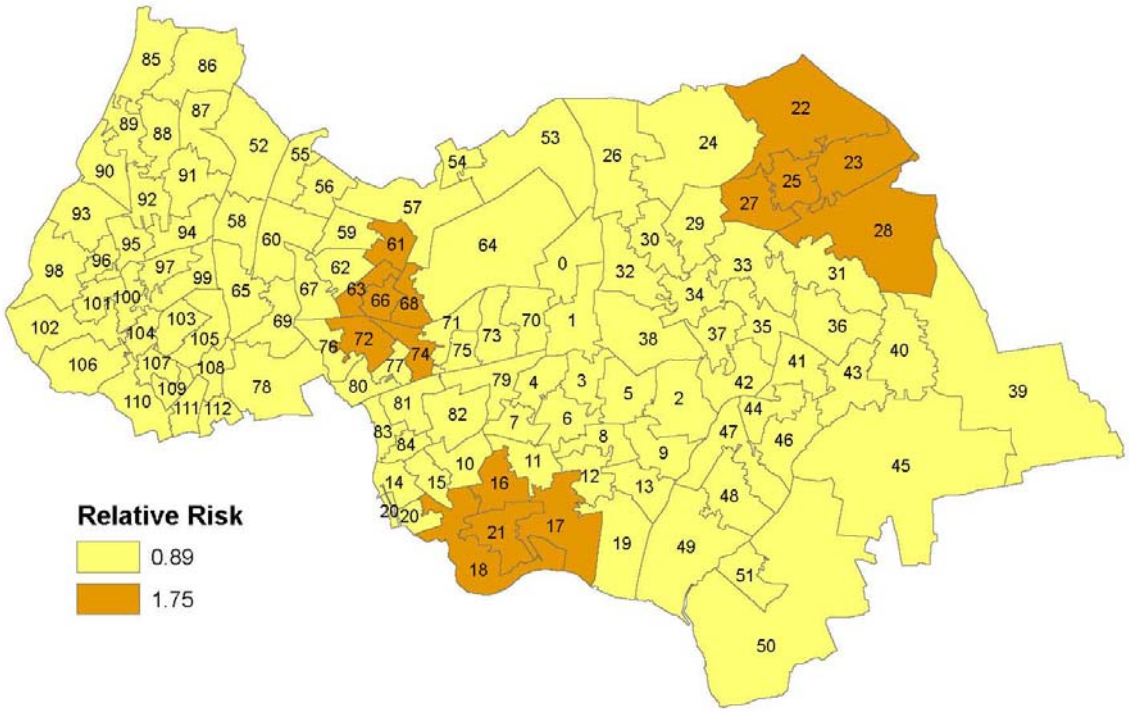

Figure 1 High Risk Clusters, Outer NE London MSOAs. Scenario for Simulations

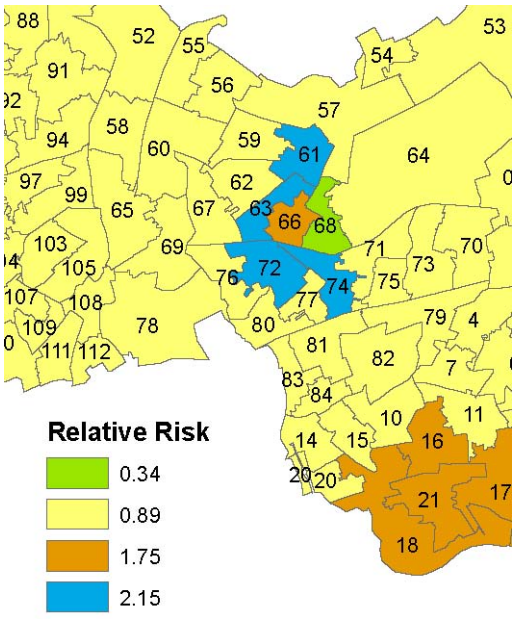

Figure 2 Uneven Risk Pattern, West Cluster. Scenario for Simulations

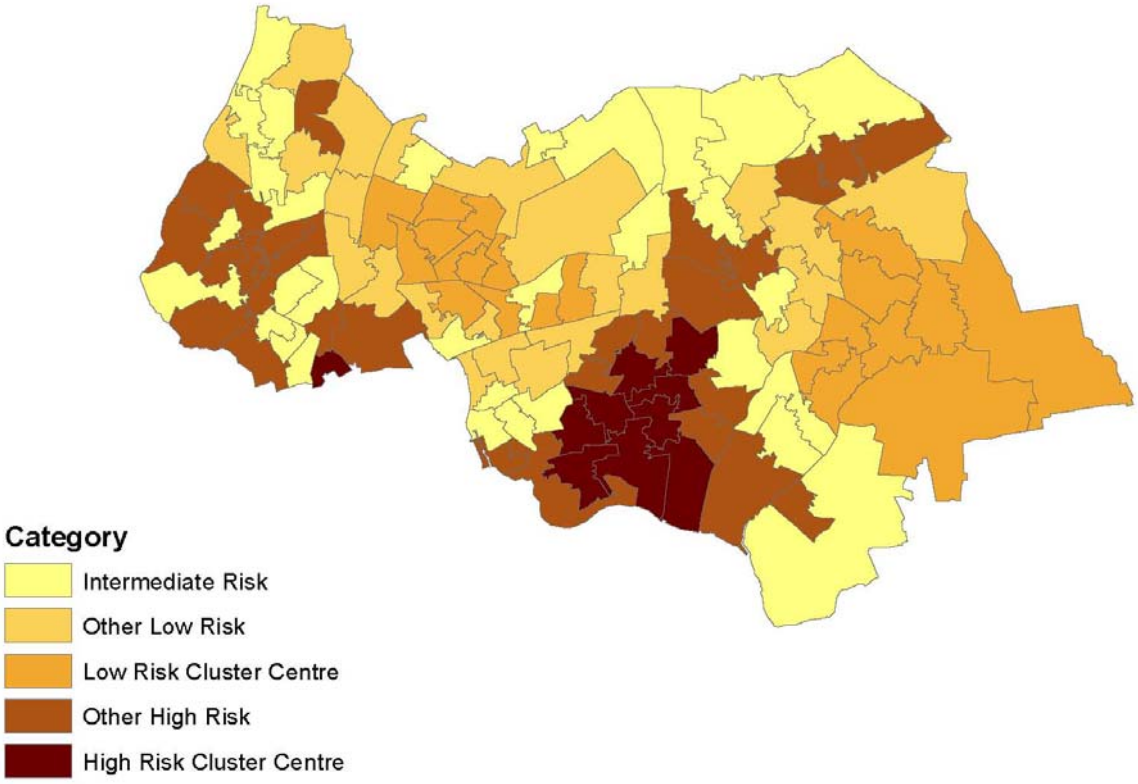

Figure 3 Cluster Configurations, Emergency COPD Admissions

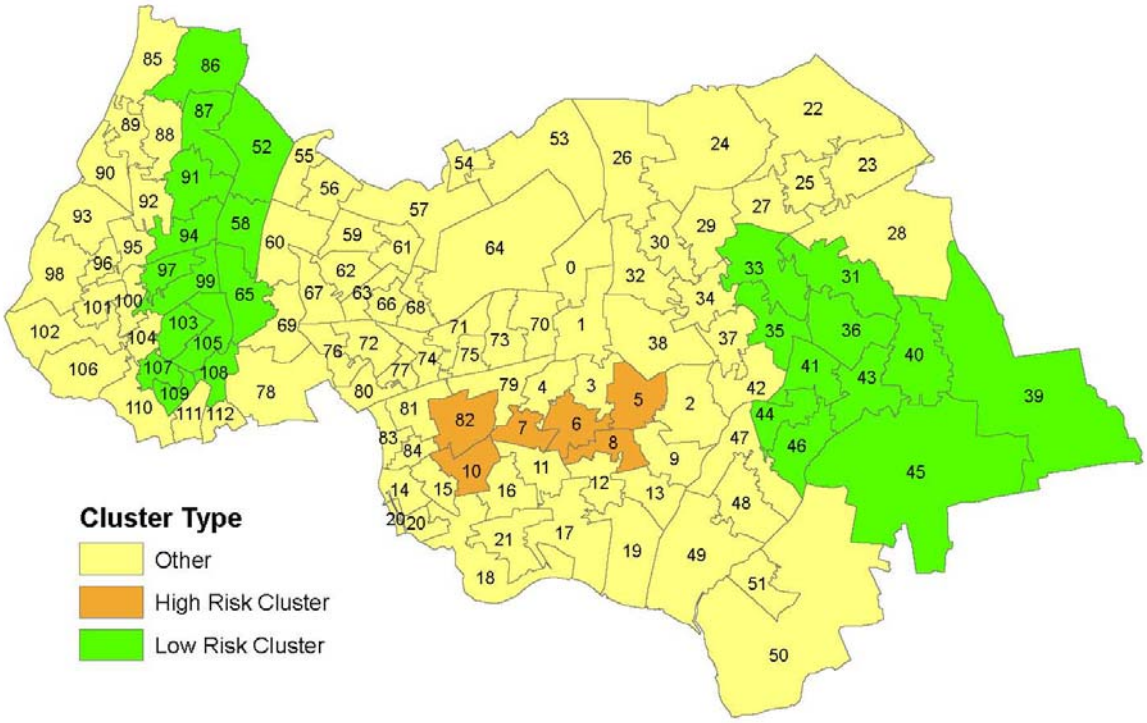

Figure 4 High Risk and Low Risk Bivariate Clusters, Pre-Primary Obesity and End-Primary Obesity