Biostatistics (2010), **0**, 0, *pp.* 1–34

doi:10.1093/biostatistics/bio000

Bayesian model choice for detecting unusual temporal patterns in small area data

Guangquan Li [*,1], Nicky Best[1,2], Anna Hansell[1,2], Ismail Ahmed[1], Sylvia Richardson[1,2]

[1] *Department of Epidemiology & Biostatistics, Imperial College, London, UK*

[2] *MRC-HPA Centre for Environment and Health, Imperial College, London, UK*

SUMMARY

Space-time modelling of small area data is often used in epidemiology for mapping chronic disease rates and by government statistical agencies for producing local estimates of, for example, unemployment or crime rates. Although, temporal changes in most local areas tend to resemble each other closely, abrupt changes over time exhibited in some areas may suggest, e.g., the emergence of localized predictors/risk factor(s) or impact of a new policy. Detection of areas with "unusual" temporal patterns is therefore important in warranting further investigations.

In this paper, we propose a novel detection method for short time series of small area data using Bayesian model choice between estimates that resemble the overall temporal pattern and those retaining the local temporal structures. For each area, evidence of belonging to the area-specific versus the common trend is synthesised by the model weight. By comparing the weight to its distribution under the null hypothesis, a test of signif-

*To whom correspondence should be addressed: guang.li@imperial.ac.uk

icance is performed to classify the local time trend as "unusual" or not. As no closed form is available, we have developed a Monte Carlo procedure to approximate the null distributions. Placed in the multiple testing context, classification rules are derived from the method of Benjamini and Hochberg (1995) to control for the false discovery rate. A comprehensive simulation study has demonstrated the consistent good performance of the proposed method in detecting various realistic departure patterns, in addition to ensuring that the FDR is well-controlled at the desired level. The proposed method is applied to mortality data on chronic obstructive pulmonary disease (COPD) in England and Wales between 1990 and 1997 a) to test a hypothesis that a government policy increased the diagnosis of COPD and b) to perform surveillance. While results showed no evidence supporting the hypothesis regarding the policy, an identified unusual district (Tower Hamlets in inner London) was later recognised to have higher than national rates of hospital readmission and mortality due to COPD by the National Health Service, which initiated various local enhanced services to tackle the problem.

Keywords: Bayesian spatio-temporal analysis; disease surveillance; detection; FDR; COPD.

1. INTRODUCTION

For many areas of application such as small area estimates of income, unemployment, crime rates and rates of chronic diseases, smooth time changes are expected. However, due to changes to social structure, policy implementation or emergence of localized risk factor(s), some areas may exhibit unexpected changes over time. Therefore, detection of areas with unusual temporal patterns is an important issue in spatio-temporal analysis of small area data.

In the small area context, observed data for each spatial unit are often too sparse

to provide reliable estimates. Bayesian hierarchical models offer a flexible framework which, through the use of spatially and/or temporally structured random effects, allows information to be shared between areas and across time points. Uncertainty of estimates can hence be reduced. As a natural extension to the purely spatial models such as those discussed in Best *and others* (2005), time trends are often modelled independently of the spatial pattern. For example, in disease mapping, the effects of space and time are typically modelled additively on the log or logit scale as $u_i + \gamma_t$ where u_i and γ_t are smoothed random effects capturing the spatial and temporal patterns respectively (Waller *and others*, 1997; Knorr-Held and Besag, 1998). The separation of space and time encapsulated in the additive formulation assumes that all areas in the study region behave identically over time and therefore display the same temporal structure, namely, γ_t, an assumption that ignores any localized behaviours. To relax this assumption, Knorr-Held (2000) extended the separable framework by including a space-time interaction term, which captures the additional variations that are not modelled by the space+time main effects. In a series of papers by MacNab and colleagues (2001; 2007) time series data for each spatial unit is modelled by a combination of a so-called "global" trend and a "regional" trend, both estimated using splines. Gaussian Markov random field structure is further imposed on the spline coefficients such that areas nearby tend to have similar trend patterns. While these models can accommodate flexibly a variety of time trend structures, the focus is on providing estimates but not detecting areas with unusual behaviours. To detect excess space-time variability, a recent paper by Abellan *and others* (2008) specified a mixture of two normal distributions, one with a larger variance than the other, for the space-time interaction term. Under this framework, allocation of an interaction term to the normal with a larger variance indicates excess variability present in the observed

data. Classification of areas into "stable" and "unstable" risk clusters is then based on summary statistics of the selection probability (Abellan *and others*, 2008). However, by construction, this model may not be particularly sensitive when the departures exhibit certain structures, for example, higher risks occurring at some consecutive time points.

Besides the model-based methods, detection of areas with unexpected changes based on test statistics has a far-longer history, e.g., the Knox test (1964) and Mantel's test (1967). A test-based method that is close in spirit to the method proposed in this paper is the space-time permutation scan statistic by Kulldorff *and others* (2005), a refinement of the space-time scan statistic of Kulldorff (2001). Readers are referred to a recent paper by Robertson *and others* (2010) for a thorough discussion of other test-based methods. This space-time permutation scan statistic is designed to test for the presence of excessive space-time interactions. Given a cylindrical volume containing a number of geographical areas over a specific period of time, the observed number of cases in that volume is compared to what would have been expected if the cases were independently distributed over space and time. Designed to detect space-time interactions, this test automatically adjusts for pure spatial and pure temporal effects so, for example, if all areas showed a doubling risk at time t compared to $t-1$, then no areas would be highlighted. However, if this only occurs in one area, the scan statistic is designed to detect such an area. Implemented in SaTScan^{TM}, this method and its space-only version have been applied to many problems in disease surveillance. However, the construction of the cylindrical scanning volume makes it inefficient to detect isolated clusters (i.e., elevated risk in a single or very small number of areas). Furthermore, inherited from the purely spatial scan statistic, this space-time extension is conservative in detecting secondary and subsequent clusters (Haining, 2003, page 257). In the simulation study here, we will

compare the performance of our proposed detection framework to that of this popular permutation test approach.

Multiple comparison is one crucial issue to address under any detection model. Due to the large number of tests performed, some proportion of the declared areas are bound to arise by pure chance. To tackle this problem, we employ the procedure proposed by Benjamini and Hochberg (1995), hereafter referred to as the BH algorithm, that provides a decision rule with a control of the false discovery rate (FDR), defined as the expected proportion of the declared areas induced by a decision rule that are false positives. In applying the BH algorithm, a Monte Carlo step is required to approximate the distributions of the model selection criterion, namely the model weights that will be introduced in Section 2,under the null hypothesis. Various approximation procedures will be considered in order to reduce the computational burden.

With two substantive questions in mind, we analyse a set of mortality data on Chronic Obstructive Pulmonary Disease (COPD) in England and Wales (1990-1997) using the proposed method. COPD is a common chronic condition characterized by slowly progressive and irreversible decline in lung function. It is responsible for approximately 5% of deaths in the UK (Hansell *and others*, 2003). While smoking is the main risk factor, exposure to high levels of dusts and fumes in industries such as mining are associated with higher risks of COPD (Coggon and Taylor, 1998; Miller and MacCalman, 2010). In a spatial analysis of COPD mortality covering 1981-1999, higher rates of COPD mortality were noted in districts in England and Wales containing mining areas (Best and Hansell, 2009). Industrial Injuries Disablement Benefit was made available for miners developing COPD from 1992 onwards in the UK (Rudd, 1998; Seaton, 1998). As miners with other respiratory problems with similar symptoms (e.g., asthma) could potentially

have benefited from this scheme, our first question was to test whether this policy may have differentially increased the likelihood of a COPD diagnosis in mining areas. Spatial variability in COPD mortality has been shown to correlate well with spatial variability of COPD in hospital admissions and GP contacts (Hansell *and others*, 2003), so mortality is likely to be a good proxy for COPD morbidity and prevalence. Therefore, one might expect to see a relative increase in rates of COPD mortality in men living in mining districts (very few miners are women), occurring against the known national trend of decreasing COPD mortality rates in men of all ages since the late 1980s (Lopez *and others*, 2006) related to changes in UK smoking trends over time. In addition to this, our second task is to explore the use of this detection method as a tool for disease surveillance to highlight areas with a potential need for further investigation and/or intervention.

The structure of the paper goes as follows. In Section 2, we will first describe the detection framework. The Monte Carlo procedure for approximating the null distributions and the implementation of the BH algorithm will also be outlined in Section 2. The COPD mortality data used in our case study will be described in Section 3. In Section 4,we will investigate the performance of the proposed model by a simulation study. Application of the method to the COPD data will be detailed in Section 5.

2. DETECTION BASED ON BAYESIAN MODEL SELECTION

2.1 A general detection framework

Let $y_{i,t}$ and $E_{i,t}$ be the observed and expected numbers of disease cases, respectively, in area i at time t. When the disease of interest is rare, a Poisson distribution is often assumed to model the count data. Specifically, at the first level of the model hierarchy, we have, $y_{i,t} \sim Poisson(\theta_{i,t} \cdot E_{i,t})$ with $i = 1, ..., N$ and $t = 1, ..., T$.

With the aim of detecting areas with temporal trends that differ from the common trend, we propose to describe the distribution of relative risk $\theta_{i,t}$, by two alternative models, one that assumes a space-time separability for all areas and one that provides time trend estimates for each spatial unit individually. To be precise, at the second level of the hierarchy, $\theta_{i,t}$ is modelled as

$$log(\theta_{i,t}) = \begin{cases} \alpha_0 + \eta_i + \gamma_t & \text{Model 1 for all } i,\, t \\ u_i + \xi_{i,t} & \text{Model 2 for all } i,\, t. \end{cases} \tag{2.1}$$

Model 1 (or the common trend model) combines the effects of space, η_i, and time, γ_t, additively (on the log scale), and consequently, the temporal trend pattern is the same for all areas, an assumption that can over-smooth local trends that display true departures. Representing the null hypothesis, Model 1 will also be referred to as the null model. In order to accommodate substantial departures from the common trend pattern, the alternative Model 2 (or the area-specific trend model) is formulated such that the temporal trends are estimated independently for each area. Here, u_i is the area-specific intercept and $\xi_{i,t}$ depicts the local trend patterns. Using a model choice formulation, a model indicator z_i indicates for each area whether Model 1 ($z_i = 1$) or Model 2 ($z_i = 0$) is supported by the data. The posterior model weight, $w_i = P(z_i = 1|\text{data})$, is then calculated to quantify the evidence for retaining the null hypothesis that the given area shows no departures from the common trend pattern. A small value of w_i indicates that the trend pattern of area i is unlikely to follow that of the common trend, γ_t.

To fully specify the above modelling framework, priors are to be assigned to the model components. For Model 1, we assign a convolution prior for the spatial random effect term, η_i, and a Gaussian random walk model of order 1 (RW(1)) to the temporal random effect term γ_t. Introduced by Besag *and others* (1991), the spatial convolution

prior (or the BYM prior) combines a spatially structured random effect term, to which we assign the conventional conditional autoregressive model (CAR), and a spatially unstructured random effect term, which follows $N(0, \sigma_\eta^2)$. More specifically, for the spatial CAR prior, we impose the neighborhood structure by defining an adjacency matrix $\mathbf{W}$ of size $N \times N$ such that the diagonal entries $w_{i,i} = 0$ and the off-diagonal entries $w_{i,j} = 1$ if areas i and j share a common boundary. Otherwise, $w_{i,j} = 0$. To implement the temporal RW(1) prior, we use its equivalent form of a one-dimensional CAR model (see e.g. Fahrmeir and Lang (2001)). Similar to the spatial CAR prior, the temporal neighborhood structure is defined though a matrix $\mathbf{Q}$ where $q_{h,t} = 1$ if $|h - t| < 2$ and $q_{h,t} = 0$ otherwise with h and t indexing units of time. A global intercept, α_0, is also included since both the CAR prior on η_i and the RW(1) prior on γ_t are constrained to sum-to-zero. Although a BYM+RW(1) setting is assigned here, specification of Model 1 is application-specific, details of which will be provided in Section 5.

For Model 2, the same RW(1) prior structure is used on $\xi_{i,t}$. Because of the sum-to-zero constraint on the RW(1) prior, the estimated trend patterns are additively adjusted according to the observed data by an area-specific intercept u_i. A vague prior is assigned to each u_i so that no information is borrowed from other areas in estimating terms in the area-specific trend model, ensuring that each area is treated independently.

Putting everything together, the full specification of the proposed framework is as follows.

$$\begin{array}{cc} \text{Model 1} & \text{Model 2} \\ \alpha_0 \sim Uniform(-\infty, +\infty) & u_i \sim N(0, 1000) \\ \eta_i \sim N(v_i, \sigma_\eta^2) \text{ and } v_{1:N} \sim CAR(\mathbf{W}, \sigma_v^2) & \xi_{i,1:T} \sim CAR(\mathbf{Q}, \sigma_\xi^2) \\ \gamma_{1:T} \sim CAR(\mathbf{Q}, \sigma_\gamma^2) & \sigma_\xi^2 = (\sigma_\gamma \cdot s)^2 \end{array} \tag{2.2}$$

A weakly informative half Normal prior $N(0,1)$ bounded strictly below by 0 is assigned to σ_η, σ_v and σ_γ, as suggested by Gelman (2006). Expressing no prior information on the superiority of the two models, we have $z_i \sim Bernoulli(0.5)$. Definition of s will be given in Section 2.2.

The proposed detection framework was implemented in WinBUGS (Lunn *and others*, 2000). The two competing models are fitted separately to the same set of data, inspired by the idea of pseudoprior (Carlin and Chib, 1995). At each iteration, the model indicator z_i then selects a trend estimate from one of the two models for each area. Model fitting and model selection are embedded within one WinBUGS program facilitated by the *cut* function, which ensures that the estimation of the two models is not affected by the selection. The model is represented as a directed acyclic graph (DAG) in Figure 1 and annotated WinBUGS code is given in the Supplementary Material.

2.2 Specification of σ_ξ^2, the variance of the area-specific trends

Through a common prior, data from all areas will contribute to the estimation of σ_ξ^2. However, in a situation where there are only a small number of areas with truly unusual trends and hence larger variability, this specification can lead to an oversmoothed setting that does not necessarily accommodate well the detection purpose because the variance estimate reflects only the small variability of the common trend pattern, which the majority of areas follow. In Equation 2.2, we set $\sigma_\xi^2 = (\sigma_\gamma \cdot s)^2$ where s is a scaling parameter that we fix *a priori* at values greater than 1 in line with our prior intention of capturing by Model 2 abrupt changes in time trend patterns that have a larger variance than σ_γ^2. Selection of s should reflect adequately the uncertainty of the estimated local trends in addition to providing good fits. For example, when the expected counts $E_{i,t}$ are small, a larger value of s is required. The role of s in the model selection procedure is

further discussed in Section 6.

In the simulation study, we will compare the detection performance with various settings of s together with a setting where σ_ξ^2 is estimated (the corresponding s value can be calculated by $\sqrt{\sigma_\xi^2/\sigma_\gamma^2}$). Coupled with the detection rule outlined below, the detection performance is shown to be robust to different values of s, given it is sufficiently large. We have also provided a tool in the Supplementary Material to help select s in practice.

2.3 Detection rules with control of FDR

Detection rules are derived from the Benjamini and Hochberg (BH) algorithm (1995) that controls the false discovery rate. This algorithm operates on the p-values. Since no closed form is available, the distributions of w_i under the null hypothesis are approximated by Monte Carlo simulations. It should be noted that these null distributions differ from area to area because of the differences in the Poisson mean, $\mu_{i,t} = E_{i,t} \cdot \exp\{\alpha_0 + \eta_i + \gamma_t\}$, that characterizes the null model. Hence approximating the null distributions has to be done in principle on area-specific replications (see further comments at the end of this section).

The Monte Carlo procedure comprises the following 5 steps.

Step 1 Fit the null model (Model 1 in Equation 2.1) to the observed data;

Step 2 Generate N_{null} data replicates from the null model using the resulting estimates (e.g., posterior means of the model parameters) from Step 1;

Step 3 For each replicate dataset D_j, $j = 1, ..., N_{null}$, fit the full detection model (specified by Eq. 2.1 and 2.2) and extract the estimates of the model weights, $\hat{w}_{ij}$, for each area;

Step 4 The distribution of w_i under the null is then formed by $\hat{w}_{i,1:N_{null}}$;

Step 5 The p-value for area i, p_i, is then calculated as the proportion of values $\hat{w}_{ij}$ less than the estimated posterior model weight w_i obtained from applying our model selection procedure to the real data.

At Step 2, we implicitly assume that only a small proportion of the areas display trends with substantive departures so that parameters in the null model can be well estimated using all observed data.

At a given FDR level, say α, a maximum integer-valued k is sought such that $p_{(k)} \leqslant \frac{k \cdot \alpha}{N}$ with $p_{(1:N)} \equiv 0 \leqslant p_{(1)} \leqslant p_{(2)} \leqslant \cdots \leqslant p_{(N)}$ denoting the vector of p-values in an ascending order and N denoting the number of tests performed (here $N =$ number of areas). The corresponding areas with $p_{(1:k)}$ are then classified as unusual and, on average, no more than $\alpha \cdot k$ of these would have been truly usual, i.e., false positives, as ensured by the BH algorithm.

As demonstrated in Supplementary Material, the number of null simulations required depends on the number of areas in the study region and the FDR level that one wishes to control at. For example, with a region of 354 areas, at the 5% FDR level we need at least 7080 samples in the Monte Carlo procedure to achieve the required precision. If we estimate one null distribution for each area, this means fitting the full detection model to 7080 replicate datasets, which is computationally extremely burdensome. However, since areas with similar Poisson means in the null model tend to have similar null distributions, we can greatly reduce the computational costs by pooling null samples from areas with similar values of $\mu_{i,t}$ to form stratum-specific nulls. For both the simulation and the COPD analysis, 10 μ-stratum are used. The area-specific nulls can further be approximated by a global distribution using sample points from all areas. Estimating from the same number of null simulations, performance of these two approximations will be

compared to that of the area-specific nulls.

3. Data description

National mortality data on COPD (ICD9 490-496) from 1990-1997 were provided by the UK Small Area Health Statistics Unit (SAHSU) at Local Authority District (LAD) level. Analyses were conducted in men aged 45 years and over as our hypothesis related to rates in men and there are very few COPD cases in younger adults or children. For the sake of illustration, we only used data from England (354 LADs) in the simulation study. For the real data analysis in Section 5, we used data from both England and Wales (374 LADs), excluding the City of London and Isles of Scilly, which have very small populations and virtually no cases. Expected counts for each LAD were standardized by 5-year age group with the age-specific reference rates calculated over the 8-year period in England and Wales. Tables 1 summarises the expected counts. Table 1 in Supplementary Material summarises the standardized mortality ratios (SMR).

4. Simulation study

4.1 Generating the data

Simulated data were generated for 8 time points for LADs in England. Model 1, with BYM for the spatial component and RW(1) for the temporal component, was first fitted to the real COPD data and the posterior mean of the fitted model was then used for generating the simulated data. Data under various departure scenarios were simulated using either the original set of expected counts from the real COPD data or a reduced set. The latter is formed by multiplying the original set of expected counts by 1/5 to examine sensitivity of detection performance to the size of expected counts. Aggregated at the annual-district level, these reduced expected cases represent a situation where the

disease of interest is extremely rare, a challenging situation for any detection methods.

Reflecting the amount of information one area possesses, the expected counts and the overall spatial risk partially influence how difficult it is to detect an area if its temporal pattern differs from the common pattern. Fifteen areas, approximately 4% of a total 354, were chosen to provide a good contrast. Selection of these areas is detailed in Supplementary Material.

Denoting $\gamma_{1:T}$ as the estimated common trend (on the log scale) from Model 1, $\gamma^*_{1:T}$ as the time trend with departures for the 15 selected areas and θ (> 0) as the magnitude of the departure, the following 3 departure patterns were considered.

1. A sudden increase of risk at time points 3 and 4, i.e., $\gamma^*_t = \gamma_t + log(\theta)$ for $t = \{3, 4\}$ and $\gamma^*_t = \gamma_t$ otherwise;

2. Increased risks appear at the first two time points, i.e., $\gamma^*_t = \gamma_t$ for $t \geqslant 3$ and $\gamma^*_t = \gamma_t + log(\theta)$ for $t = \{1, 2\}$;

3. The unusual time trend fluctuates around the common trend: $\gamma^*_t = \gamma_t + log(\theta)$ for $t = \{1, 7\}$ and $\gamma^*_t = \gamma_t - log(\theta)$ for $t = 4$.

Illustrated in Figure 2, these three departure patterns are representations of those seen in real analyses (e.g., in Glass (1998)). For each departure pattern, two different departure magnitudes are used, $\theta = 1.5$ and 2. These two chosen levels of departure are realistic for epidemiological studies. For the remaining 339 areas, the common trend $\gamma_{1:T}$ is assigned. Fifty sets of data, referred to as departure simulations hereafter, were generated under each of the 12 simulation scenarios (3 departure patterns $\times$ 2 magnitudes $\times$ 2 sets of expected counts).

4.2 Fitted models

In addition to the setting with a variable σ^2_ξ (to be denoted as s-vary), 8 model settings ($s \in \{1.5, 2, 2.5, 3, 3.5, 4, 5, 10\}$) and 6 model settings ($s \in \{2, 3, 4, 5, 6, 7\}$) were considered for data generated from the original set and the reduced set of expected counts,

respectively. See Supplementary Material for how to select the range for s in practice. For comparison, the space-time permutation test in SaTScan was also fitted to the simulated data. The threshold p-value, under which excess is declared, is set at 0.05.

For each combination of model setting and set of expected counts, the distributions of w_i under the null hypothesis are approximated by 200 null simulations, using the procedures outlined in Section 2.3. It should be noted that these simulations need only be carried out once for each setting-expected number combination. The p-values are calculated based on (a) the area-specific null distributions, (b) the stratum-specific null distributions and (c) the global null distribution, common to all areas.

Model performance is summarized by sensitivity and empirical FDR. To calculate the empirical FDR, we take the mean of the false proportion rates, FP$= \frac{V}{R}$, where V is the number of declared areas that are truly usual and R is the total number of declared areas. When $R = 0$, FP is set to zero. For the sensitivity, we record the percentage of times (out of 50 departure simulations) that each of the 15 truly unusual areas was correctly identified. The Receiver Operation Characteristic (ROC) curve, a conventional tool for comparing different binary classification methods, is not used here because it suppresses the differences in sensitivity for areas with different levels of expected counts.

4.3 Results for the original expected counts

Under the three departure patterns, Figure 3 demonstrates how good the BH algorithm is in controlling the empirical FDR at the pre-defined levels. Four model settings are compared, s-vary, $s = 2$, $s = 5$ and $s = 10$. Based on the stratum-specific null distributions, the latter three settings yielded empirical FDRs that are well controlled below the desired levels. Consistent positive biases can be seen with s-vary (first columns). Comparing the solid and the dashed lines, the false proportion rates are less variable for

the larger departure magnitude. While the global null distribution provided similar good FDR controls, all empirical estimates based on the area-specific null distributions were substantially inflated over the controlled levels, suggesting that the individual null distributions based only on 200 simulations are, as expected, not sufficiently precise (results not shown). In fact, with the current set of data, both the stratum-specific and global null approximations can dramatically reduce the required number of null simulations while achieving similar control of FDR as the "gold standard" area-specific nulls with a sufficiently large number of simulations (see Supplementary Material Figure 1).

With p-value set to 0.05, the empirical estimates of FDR from SaTScan were in general greater than 0.2 with highly variable false proportion rates (for example, $\hat{FDR} = 0.19$ with a 95% sampling interval (0.00, 0.78) with $\theta = 2$ under pattern 2) under the departure scenarios considered (results not shown).

For the sake of illustration, sensitivity is evaluated using an FDR nominally controlled at the 0.1 level and stratum-specific null distributions. With departure magnitudes $\theta = 1.5$ and 2, respectively, Figures 4 and 5 summarize the ability to detect the 15 truly unusual areas using three model settings (s-vary, $s = 2$ and $s = 5$) and SaTScan (column-wise) under three departure patterns (row-wise). In each plot, the probabilities of correctly detecting the 5 truly unusual areas, each having a median expected count at one of the 5 percentiles, are joined by the solid line (low spatial risks), the dashed line (median spatial risks) and the dotted line (high spatial risks). All lines consistently show an overall increasing pattern, indicating that the three settings all tend to be more powerful in detecting changes in areas with larger expected counts, an expected result. The detection power depends also on the level of spatial risk. Not surprisingly, it is easier to detect departures when the associated area has a higher risk averaged over time,

although the impact on power due to the difference in spatial risks is relatively minor compared to that due to the difference in expected counts.

Results from $s = 2$ (the second columns in Figures 4 and 5) and $s = 5$ (the third columns) are in general comparable yet considerably better than those obtained from s-vary (the first columns) and SaTScan (the last columns). Compared to SaTScan, our method is particularly more powerful in detecting areas with sparse data, i.e., those at the lower percentiles of the expected count distribution and/or with lower averaged spatial risks. It is interesting to point out that SaTScan achieved slightly higher or similar probabilities of detecting the two low spatial risk areas with relatively large expected counts (the right hand tails of the solid lines in the last column of Figures 4 and 5). This is because these two areas happen to be close together, a situation where the scanning windows used in SaTScan has the advantage.

Besides the expected number and the level of spatial risks, the detection power also somewhat depends on the pattern of departure. Departure patterns 2 and 3 appear to be easier to detect than departures with pattern 1. This is probably because departures under pattern 1 occurred at the middle of the observation period, whereas patterns 2 and 3 have departures at the beginning and/or end of the period. Thus there will be more smoothing of pattern 1 by the information borrowing from the previous and next time points imposed by the RW(1) prior than for either pattern 2 or 3. Such a difference is less marked as expected counts, spatial risks and/or departure magnitudes become higher.

4.4 Results for the reduced expected counts

When data are generated from the reduced expected numbers, neither SaTScan nor our model could pick up areas with departures of $\theta = 1.5$ (results not shown) at such a high level of sparsity. With $\theta = 2$, departures of patterns 2 and 3 can be detected by

our method (Figure 6). The performance is particularly satisfactory, with sensitivity at 0.6 or more, for detecting areas with relatively large expected counts. Departures with pattern 1, on the other hand, can rarely be identified. SaTScan also failed to capture any areas with any types of departure with the detection probability barely going above 0.2. In terms of controlling the FDR, despite some positive biases under departure pattern 1 (first row of Figure 7), when s is fixed at 3 or above, the detection rules from the BH algorithm with either stratum-specific null distributions or the global null distribution controlled the empirical FDR reasonably well at the predefined levels (Figure 7). Compared to the scenarios using the original expected counts, a slightly larger s is required to achieve good performance because of greater uncertainty inherent in the sparse data (see further discussion in Section 6). In addition, the 95% sampling intervals are generally wider than those from data generated using the original set of expected counts.

5. Application: Chronic obstructive pulmonary disease (COPD)

Motivated by the interrogation discussed at the beginning of this paper, we analyze the COPD data using our detection model to formally examine the evidence of the policy impact and to explore the ability of our method to perform disease surveillance.

Various space-time separable models with different specifications for the spatial and temporal components, for example CAR + RW(1) and BYM + RW(1), were fitted to the COPD data. The model with BYM for the spatial component and RW(1) for the temporal component produced the smallest Deviance Information Criterion (DIC, Spiegelhalter *and others* (2002)) and hence was used as the specification for the common trend model in Equation 2.1. The scaling parameter s was fixed at a number of values, namely, $s \in \{1.5, 2, 2.5, 3, 3.5, 4, 5, 10\}$ and also s-vary. Setting of $s = 1.5$ was chosen by the tool provided in the Supplementary Material. Shown to be sufficient in the Simulation Study,

200 null simulations were carried out to generate stratum-specific null distributions (with 10 strata).

The analyses did not find evidence to support our hypothesis that the introduction of Industrial Injuries Disablement Benefit for miners developing COPD in 1992 increased the likelihood of COPD diagnosis and therefore COPD mortality in mining areas. At the FDR level of 0.05, settings with $s = 1.5$ and 2 both identified four local authority districts (Figure 8 (a)), amongst which only two (Rotherham and Carmarthenshire) were in mining areas (out of a total 40 mining districts). The reason for this may be that a very large number of mines closed from the mid 1980s in the UK, which would have dramatically reduced the impact of mining dust exposures on COPD development (and subsequently mortality) in an area. Working conditions improved and dust control measures were noted to have reduced relationships between dust exposures and lung function in some areas by the 1990s (Seaton, 1998). Additionally, in some mining areas, doctors writing death certificates may have continued to put pneumoconiosis (another compensable illness) on the death certificate for miners dying of a respiratory disease, instead of COPD (Seaton, 1998). Further, while the local trend increased in Carmarthenshire, it decreased in Rotherham (Figure 2 in Supplementary Material). These departures of time trend patterns may be the result of several other changes occurred over this time period, in addition to any potential impact of the policy.

The other two unusual districts with a statistically significant local increasing trend (against a national decreasing trend) were in inner London (Lewisham and Tower Hamlets in Figure 2 in Supplementary Material). They are very deprived areas with high levels of in-migration and large ethnic minority populations especially from Africa and the Indian subcontinent and therefore might not show similar trends to the rest of the UK.

In fact, Tower Hamlets has been commissioning Local Enhanced Services since 2008 in order to optimise patient management and to reduce COPD (TowerHamlets-Council, 2009; NHS-TowerHamlets, 2009), but the rising trend in COPD could potentially have been recognised earlier in the 1990s through using this type of surveillance statistic.

Rotherham and Carmarthenshire are consistently identified by settings with $s \leqslant 5$ whereas Lewisham and Tower Hamlets were only detected by $s = 1.5$ and $s = 2$. With $s = 10$, only Rotherham was identified. In practice, results using a variety of settings should be explored. In addition, we note that stratum-specific and global nulls yielded similar results while the number of detected areas hardly changed with FDR=0.1 or 0.15 (not shown).

Two circular clusters of large numbers of areas were detected by SaTScan, as shown in Figure 8 (b), both of which were in mining areas. The one in the north of England, containing 46 LADs, expressed an excess risk of 1.05 during 1990-1992 while the one in Wales and the south-west with 19 LADs showed an increased risk of 1.12 between 1995 and 1996. Although the second smaller cluster may appear to be consistent with our hypothesis of the impact of government policy on mortality data, these results should be interpreted with great care since, as shown by our simulation study, a considerable number of these detected areas would be false discoveries. In addition, SaTScan missed out completely on identifying the two LADs in inner London.

6. Conclusion and discussion

The proposed detection framework has demonstrated its superior performance in detecting various realistic departure scenarios in the simulation study and its usefulness in terms of both assessing policy impact and performing surveillance in the COPD application, while tightly controlling the false discovery rate.

We formulated the detection problem in the Bayesian model selection framework instead of the conventional mixture modelling approach, which, due to the complexity of the two competing models, has problems achieving convergence. We utilized the posterior model weights $w_i|data$ to choose between two models. The Bayes factor (BF) would be an alternative for the selection criterion. The crucial, and often challenging, task is to calibrate a criterion in order to transform the resultant measure into a classifier of, in our case, being "usual"or "unusual". In the case of BF, there are methods, such as Jeffrey's interpretation (Jeffreys, 1961) and a simulation-based method by Vlachos and Gelfand (2003) to tackle this task. Here, we set out to control for the FDR, an important quantity in the detection context. Coupled with the BH algorithm, calibration of the model weights is achieved through a Monte Carlo procedure, which is equally applicable to the case where BF is used.

Under our detection approach, departures are easier to detect when the target area has large expected counts and/or high overall spatial risks. In the simulation, the reduced set of expected counts presents a minimal level of information beyond which this method is not likely to perform well. Below this level, one may have to aggregate the data over either a longer period of time or at a higher geographical level or both. Note that the power of our detection method is not affected by the geographical distribution of the unusual areas since all areas are treated independently under the area-specific model. This feature also helps to target individual areas, making the test more specific, rather than clusters of areas, which SaTScan usually identifies.

In order to meaningfully define the common trend, the proposed method assumes that the proportion of unusual areas is small, perhaps no more than 10%. This is also implicitly assumed in approximating the null distributions (Step 2 of the MC procedure).

Suitable applications of this method would be in monitoring early disease outbreaks, detecting elevation of crime rates and assessing impact of a small-scaled implementation of a policy.

The choice of FDR is important when considering the use of our model to assist in surveillance of chronic disease. This was explored in the COPD application which showed a consistent detection pattern of 4 LADs at the FDR levels of 5%, 10% and 15%. Although 5% seems to be commonly used in epidemiological studies (e.g., Harris *and others* (1998); Charlesworth *and others* (2010)), choice of the FDR threshold should reflect practical and application-specific considerations. For example, if subsequent investigation of the identified areas is costly, one may apply a stringent rule such as FDR=0.05, making the detection more specific. On the other hand, if the disease under monitoring has a high incidence rate, one might want a more sensitive test (by using a higher FDR level) to have better detection of true positives. Perhaps a more advisable approach is to present results at various FDR levels (for example, Ventrucci *and others* (2010)). Subject-specific experts can then be sought to interpret the findings.

The FDR control depends on the precision of the MC procedure. Since the variability in expected counts is relatively small, both stratum-specific nulls and the global null approximate well the area-specific null distributions (supported by Figure 3 in the Supplementary Material) and hence worked well with the current dataset. However, if the expected counts in some parts of the region are some orders of magnitude higher than others, for example in a study region comprising rural- and urban-only small areas, the global null distribution may not be appropriate, though the stratum-specific approach can still be applied. Finer stratifications of areas (e.g., >10 strata) may be required and a sensitivity of results on different stratifications is recommended.

In the simulation study, we only considered two departure magnitudes, namely $\theta = 1.5$ and 2. Departure magnitudes above 2, though easily detected using the proposed framework, rarely occur in practice. As an additional check, we also considered a departure pattern where the temporal patterns of the common trend and the area-specific trend are the same but they only differ by the overall level (i.e., one is above the other). As expected, areas with this unusual trend were not captured by the proposed framework. This is because, under our model construction, such a shift in overall level is captured by the spatial term, η_i in the common trend model, so all areas are therefore considered to be usual in trend pattern.

Under our detection framework, the role of s in the specification of σ_ξ^2 is twofold. It not only controls the smoothness of the estimated local trends but also reflects the uncertainty of the trend estimates, two aspects that influence the model selection procedure. As s increases, the goodness of fit (GOF) improves as the fitted local trends becomes more flexible in depicting the observed data and hence in terms of model selection, the area-specific trend model is preferred. However, beyond a point where the GOF improvement ceases, the uncertainty dominates, leading to favour the less variable estimates from the common trend model.

Since it is difficult to learn from data, we chose to fix $s > 1$ to express *a priori* our goal of detection.The extensive simulation study has shown that with s between 2 and 10, our detection method achieved robust performance. Fixing s to values beyond 10 may not be justifiable since departures with such a large variability can easily be identified by just plotting the raw data. In practice, if one wishes to report findings from only one setting, the posterior predictive criteria as discussed in Section 1 of the Supplementary Material can be used. However, it is recommended to carry out sensitivity analyses of

the detection outcomes on various model settings.

Public health surveillance systems are commonly used to monitor infectious diseases e.g. notifications of intestinal infections, where the aim is to identify statistically significant departures from background levels so that public health measures can be initiated. However, similar routine monitoring systems are uncommon for chronic disease. As demonstrated, detection methods such as that proposed in this paper have potentially high policy relevance for national or regional chronic disease surveillance to help identify departures from common trends that may require explanation and investigation and targeted interventions. For example, our analyses showed that COPD mortality trends rose in three districts when the national trend was a decrease. Such findings could be used to improve local health care facilities for COPD prevention and management. This is indeed the case in Tower Hamlets but various schemes were only initiated in 2008 (TowerHamlets-Council, 2009; NHS-TowerHamlets, 2009).

The proposed detection framework can be readily adapted to monitor infectious diseases, where areas with departures are likely to form local clusters. Spatial dependence of the model choice can be induced through a Gaussian random field (e.g., in Fernandez and Green (2002)) such that choice of model depends not only on the data but also on the hypothesised spatial structure of the alternative, potentially achieving higher power.

The time window over which changes are detected also needs to be considered. The detection method has been applied to data with 8 time points. For a longer time span (e.g., > 10 time points), the model indicator z_i currently used is perhaps too restrictive. The detection framework may need to be extended so that the model indicator is specific to both area and time point, namely, $z_{i,t}$. Furthermore, more than one departure may occur during a long time period. To help pinpoint the periods with departure, we are

currently developing a sequential fitting of the detection framework where data are fed one time point at a time. This sequential framework can also be useful to initiate public health measures promptly.

Acknowledgments

This research was funded by the ESRC National Centre for Research Methods (BIAS II node, grant RES-576-25-0015). We would like to thank the Small Area Health Statistics Unit (SAHSU) for provision of the COPD data. The work of SAHSU was funded by a grant from the Department of Health for England and the U.K. Department for Environment, Food and Rural Affairs.

REFERENCES

ABELLAN, J., RICHARDSON, S. AND BEST, N. (2008). Use of space-time models to investigate the stability of patterns of disease. *Environmental Health Perspectives* **116**(8), 1111–1119.

BENJAMINI, Y AND 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**(1), 289–300.

BESAG, J., YORK, J. AND MOLLIÉ, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* **43**(1), 1–20.

BEST, N. AND HANSELL, A. (2009). Geographic variations in risk: adjusting for unmeasured confounders through joint modeling of multiple diseases. *Epidemiology* **20**(3), 400–10.

BEST, N., RICHARDSON, S. AND THOMSON, A. (2005, Feb). A comparison of bayesian spatial models for disease mapping. *Stat Methods Med Res* **14**(1), 35–59.

CARLIN, B.P. AND CHIB, S. (1995). Bayesian model choice via markov chain monte carlo methods. *JRSS(B)* **57**(3), 473–484.

CHARLESWORTH, J., CURRAN, J., JOHNSON, M.P., GÖRING, H., DYER, T.D., DIEGO, V.P., KENT, J.W JR, MAHANEY, M.C., ALMASY, L., MACCLUER, JEAN W., MOSES, E.K. *and others*. (2010). Transcriptomic epidemiology of smoking: the effect of smoking on gene expression in lymphocytes. *BMC Med Genomics* **3**, 29.

COGGON, D. AND TAYLOR, N. (1998). Coal mining and chronic obstructive pulmonary disease: a review of the evidence. *Thorax* **53**(5), 398–407.

FAHRMEIR, L. AND LANG, S. (2001). Bayesian inference for generalized additive mixed models based on markov random field priors. *Applied Statistics*, 201–220.

FERNANDEZ, C AND GREEN, PJ. (2002). Modelling spatially correlated data via mixtures: a bayesian approach. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **64**, 805–826.

GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayes. Anal.* **1**(3), 515–533.

GLASS, GENE V. (1998). Interrupted time-series quasi-experiments. *Technical Report*, Arizona State Unversity.

HAINING, R. (2003). *Spatial Data Analysis: Theory and Practice*. Cambridge University Press.

HANSELL, A., HOLLOWELL, J., MCNIECE, R., NICHOLS, T. AND STRACHAN, D. (2003*a*). Chronic obstructive pulmonary disease (copd) and asthma-interpreting the routine data. *Europ. Respiratory Joun.* **21**(2), 279–286.

HANSELL, A., WALK, J. AND SORIANO, J. (2003*b*). What do chronic obstructive pulmonary disease patients die from? a multiple cause coding analysis. *Eur Respir J* **22**(5), 809–14.

HARRIS, L., LUFT, F., RUDY, D., KESTERSON, J. AND TIERNEY, W. (1998). Effects of multidisciplinary case management in patients with chronic renal insufficiency. *American Journal of Medicine* **105**(6), 464–471.

JEFFREYS, H. (1961). *Theory of Probability*. Oxford University Press London.

KNORR-HELD, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine* **19**(1718), 2555–2567.

KNORR-HELD, L. AND BESAG, J. (1998). Modelling risk from a disease in time and space. *Statistics in Medicine* **17**(18), 2045–2060.

KNOX, E. AND BARTLETT, M. (1964). The detection of space-time interactions. *JRSS(C)* **13**(1), 25–30.

KULLDORFF, M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic. *JRSS(A)* **164**, 61–72.

KULLDORFF, M., HEFFERNAN, R., HARTMAN, J., ASSUNÇÃO, R. AND MOSTASHARI, F. (2005). A space-time permutation scan statistic for disease outbreak detection. *PLoS Medicine* **2**(3), e59.

LOPEZ, A., SHIBUYA, K., RAO, C., MATHERS, C., HANSELL, A., HELD, L., SCHMID, V. AND BUIST, S. (2006). Chronic obstructive pulmonary disease: current burden and future projections. *Eur Respir J* **27**(2), 397–412.

LUNN, D., THOMAS, A., BEST, N. AND SPIEGELHALTER, D. (2000). Winbugs-a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* **10**(4), 325–337.

MACNAB, Y. (2007, (a)). Spline smoothing in bayesian disease mapping. *Environmetrics* **18**(7), 727–744.

MACNAB, Y. AND DEAN, C. (2001). Autoregressive spatial smoothing and temporal spline smoothing for mapping rates. *Biometrics* **57**(3), 949–956.

MANTEL, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Res.* **27**, 209–220.

MILLER, B. AND MACCALMAN, L. (2010). Cause-specific mortality in british coal workers and exposure to respirable dust and quartz. *Occup Environ Med* **67**(4), 270–6.

NHS-TOWERHAMLETS. (2009). Nhs tower hamlets strategic plan 2009-2010 to 2012-2013. *Technical Report*, National Health Services (NHS, UK).

ROBERTSON, COLIN, NELSON, TRISALYN A., MACNAB, YING C. AND LAWSON, ANDREW B. (2010). Review of methods for space–time disease surveillance. *Spatial and Spatio-temporal Epidemiology* **1**, 105–116.

RUDD, R. (1998). Coal miner's respiratory disease litigation. *Thorax* **53**, 337–340.

SEATON, A. (1998). The new prescription: industrial injuries benefits for smokers? *Thorax* **53**(5), 335–6.

SPIEGELHALTER, D., BEST, N., CARLIN, B. AND VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B* **64**, 583–616.

TOWERHAMLETS-COUNCIL. (2009). Health and wellbeing in tower hamlets - joint strategic needs assessment. *Technical Report*, Tower Hamlets Council.

VENTRUCCI, M., SCOTT, M. AND COCCHI, D. (2010). Multiple testing on standardized mortality ratios: a bayesian hierarchical model for fdr estimation. *Biostatistics*.

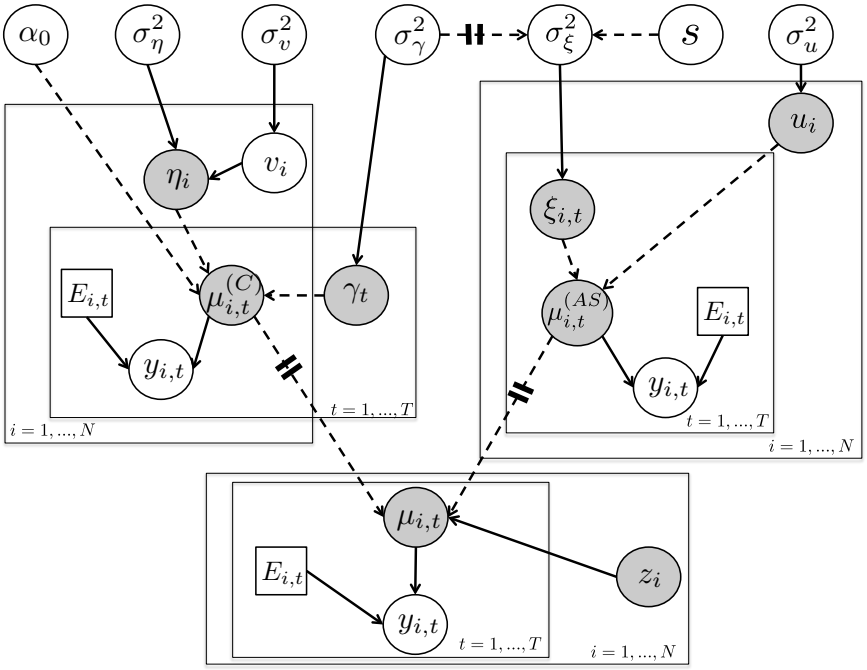
VLACHOS, P. AND GELFAND, A. (2003). On the calibration of bayesian model choice criteria. *Journal of Statistical Planning and Inference* **111**(1-2), 223–234.

WALLER, L., CARLIN, B., XIA, H. AND GELFAND, A. (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association* **92**, 607–617.

Table 1. Summary of the age-adjusted expected counts in England and Wales (without City of London and Isle of Scilly).

	Median	1990	1991	1992	1993	1994	1995	1996	1997
min	9.78	9.65	9.27	9.36	9.91	9.43	10.23	11.62	11.75
10%	24.15	22.97	22.91	23.22	23.69	24.46	25.14	25.39	25.90
25%	30.67	29.06	30.04	30.29	30.50	30.64	31.37	32.46	33.50
50%	42.17	39.47	40.40	41.04	41.77	42.34	43.50	44.54	45.53
75%	56.95	55.98	56.59	56.57	57.06	57.53	59.18	60.35	62.08
90%	79.80	77.83	78.21	78.71	79.07	80.27	82.04	83.40	85.25
max	318.36	315.85	317.19	316.60	317.98	318.74	326.34	329.04	331.97

Fig. 1. A graphical representation of the detection framework using a directed acyclic graph (DAG). Nodes in gray appear in Equation 2.1 and the bold equal sign denotes the application of the *cut* function in WinBUGS.



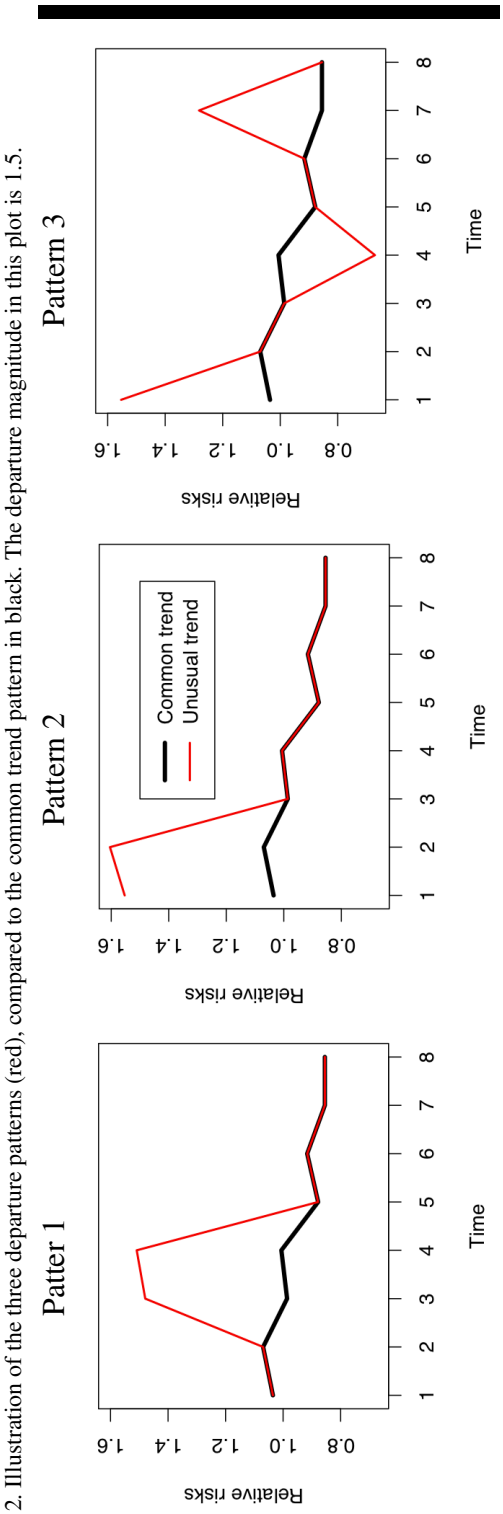
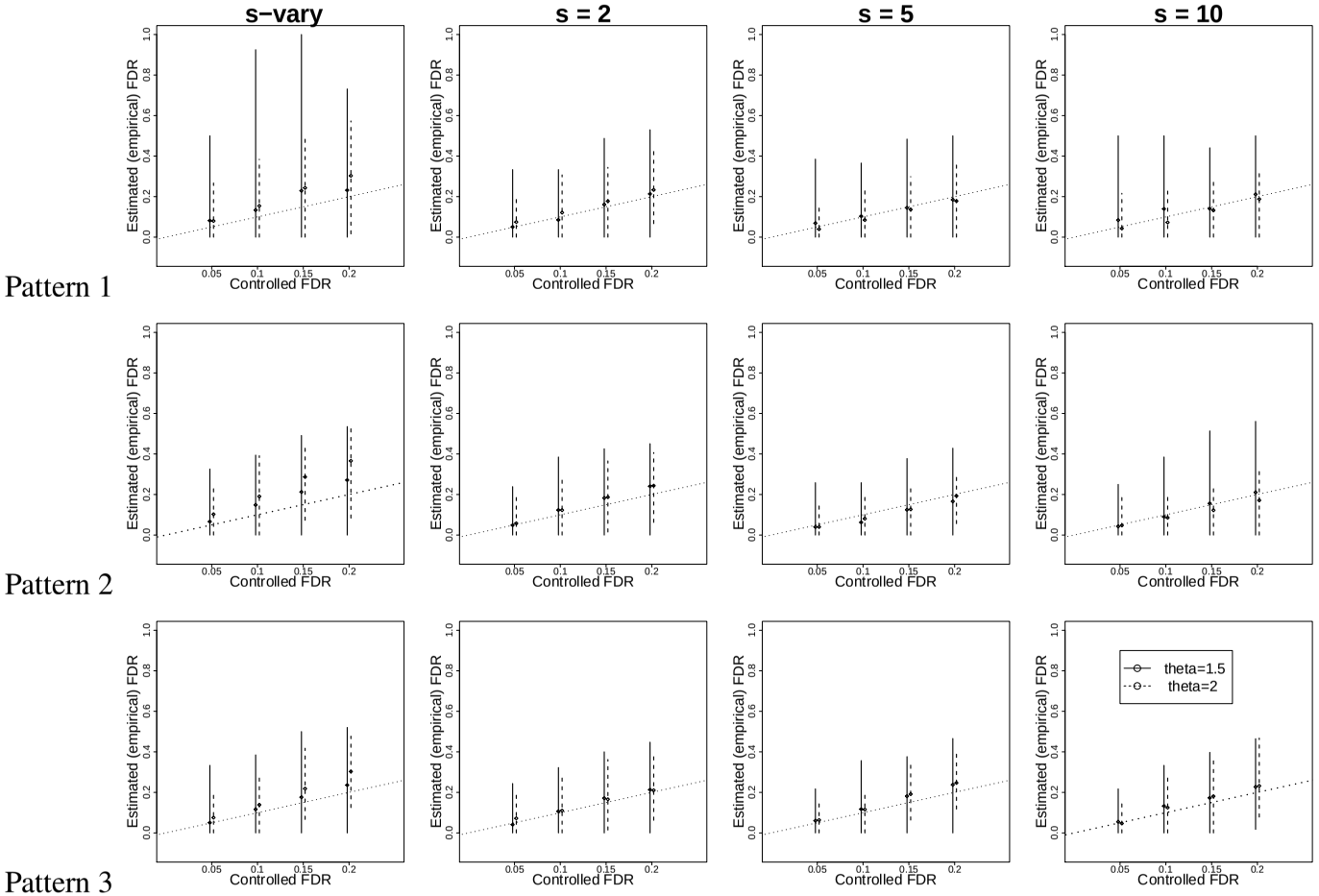


Fig. 2. Illustration of the three departure patterns (red), compared to the common trend pattern in black. The departure magnitude in this plot is 1.5.

Fig. 3. Comparisons of the empirical FDR to the corresponding controlled level using the stratum-specific nulls. Based on the original set of expected counts, four model settings (s-vary, s =2, 5 and 10) are compared under 3 patterns with 2 departure magnitudes $\theta = 1.5$ and 2. The empirical FDR (the mean false proportion rate across the 50 replicate datasets) and the 95% sampling interval are represented by the point and the vertical bar, respectively.



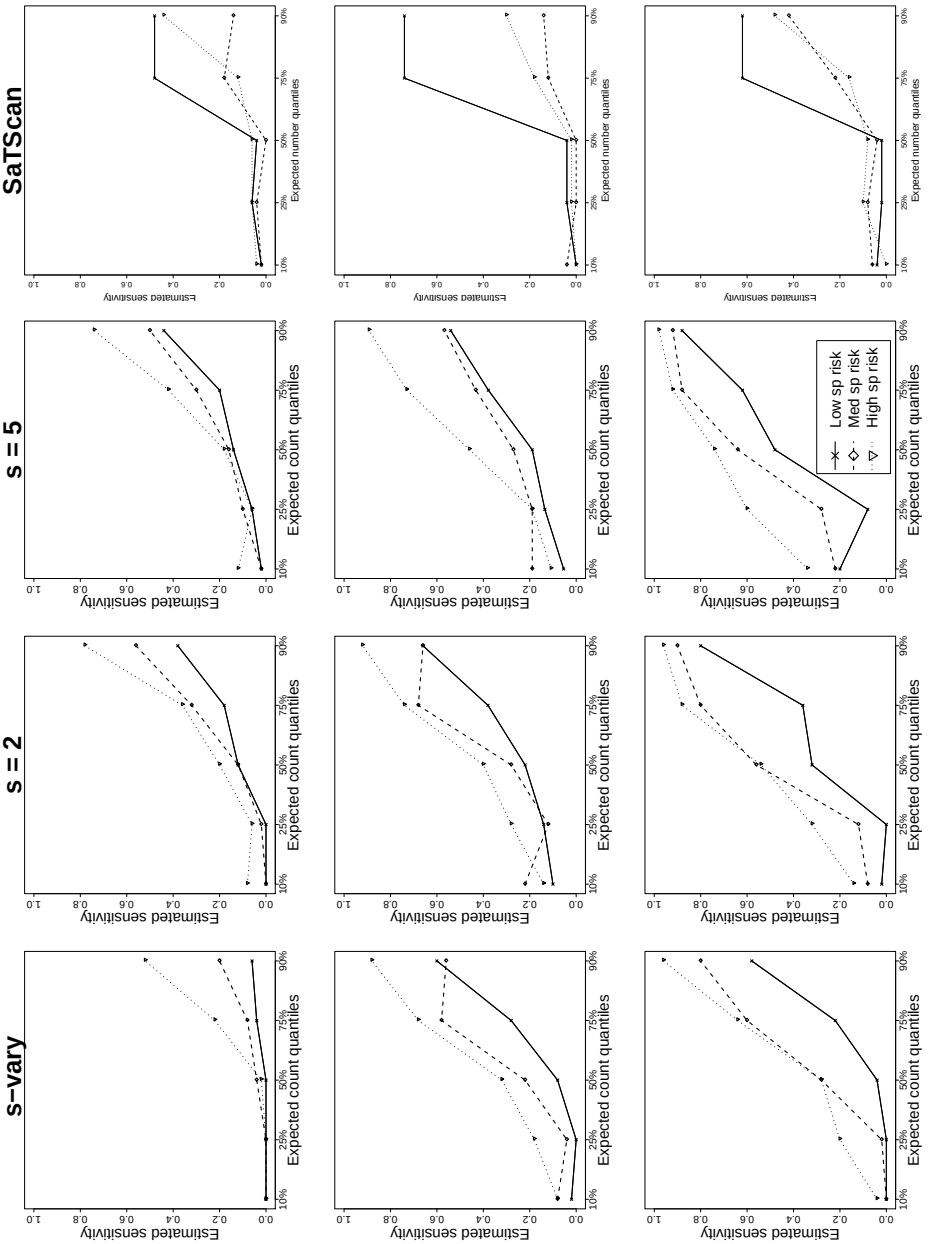


Fig. 4. Sensitivity of detecting the 15 truly unusual areas with departures magnitude $\theta = 1.5$ using various model settings and SaTScan. Data were generated from the original set of expected counts. Stratum-specific nulls are used.

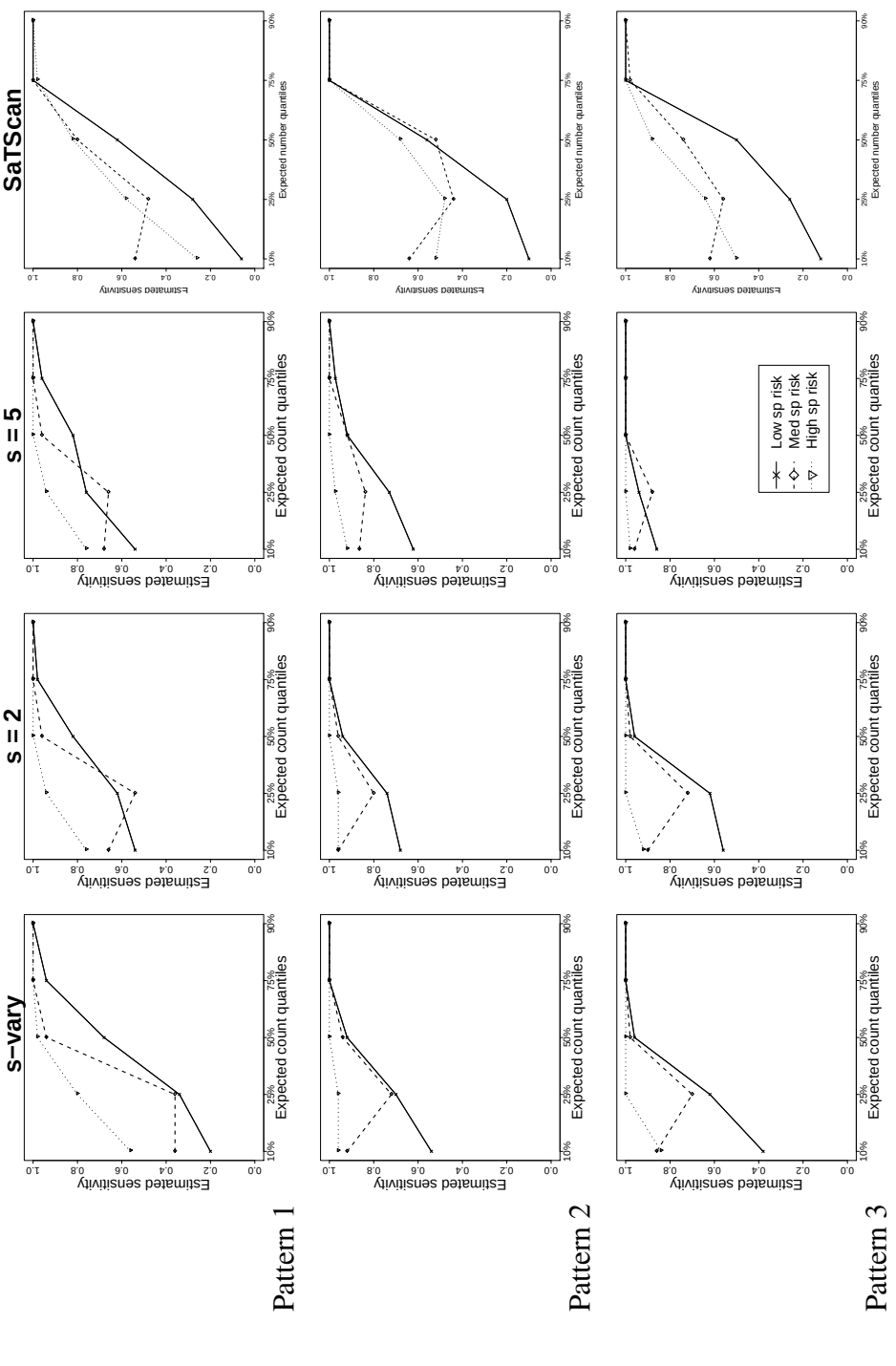


Fig. 5. Sensitivity of detecting the 15 truly unusual areas with departures magnitude $\theta = 2$ using various model settings and SaTScan. Data were generated using the original set of expected counts. Stratum-specific nulls are used.

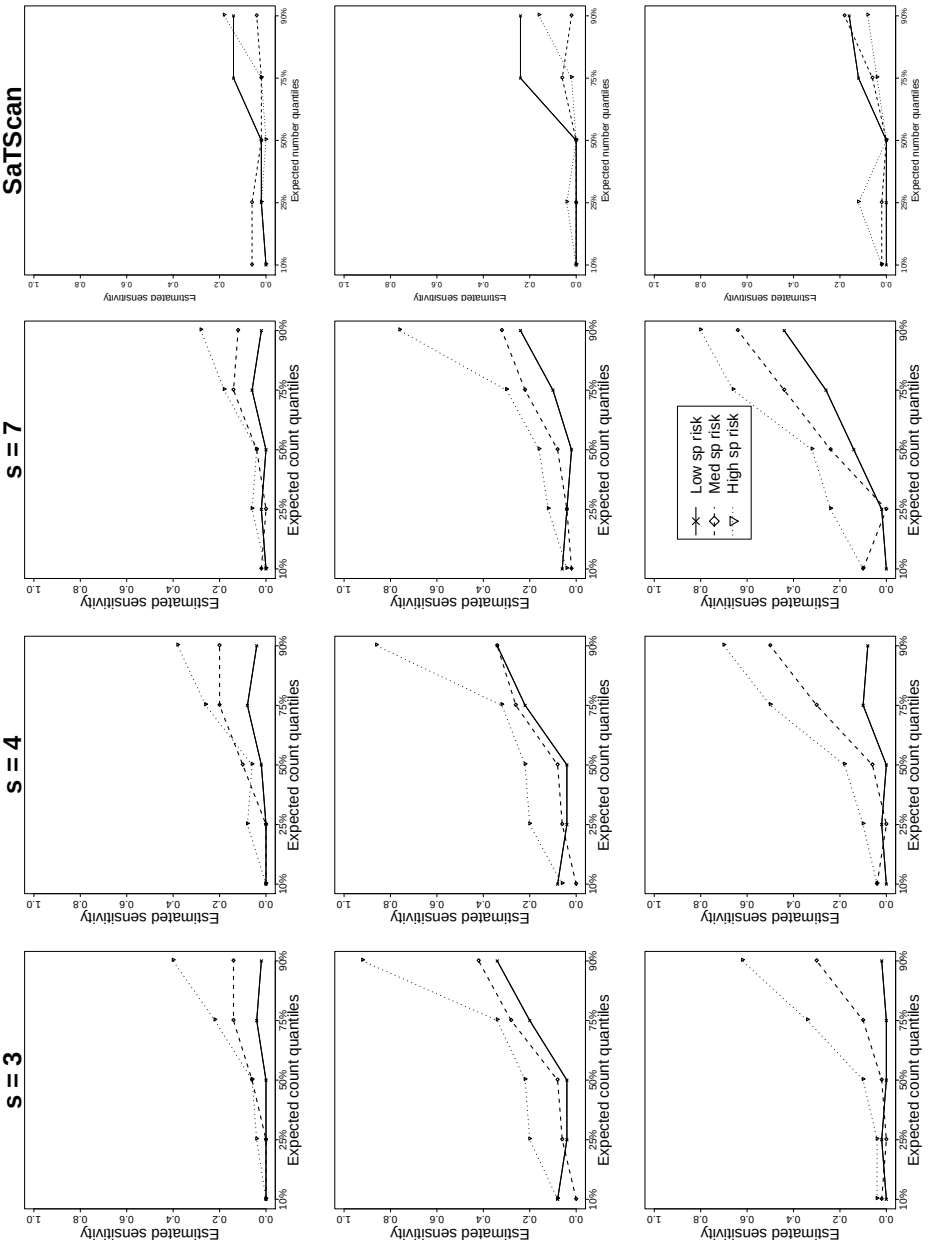


Fig. 6. Sensitivity of detecting the 15 truly unusual areas with departures magnitude $\theta = 2$ using various model settings and SaTScan. Data were simulated from the reduced set of expected counts. Stratum-specific nulls are used.

Fig. 7. Comparisons of the empirical FDR to the corresponding controlled level using the stratum-specific nulls. Based on the reduced set of expected counts, three model settings (s=3, 4 and 7) are compared under 3 patterns with departure magnitude $\theta = 2$. The empirical FDR (the mean false proportion rate) and the 95% sampling interval are represented by the point and the vertical bar, respectively.

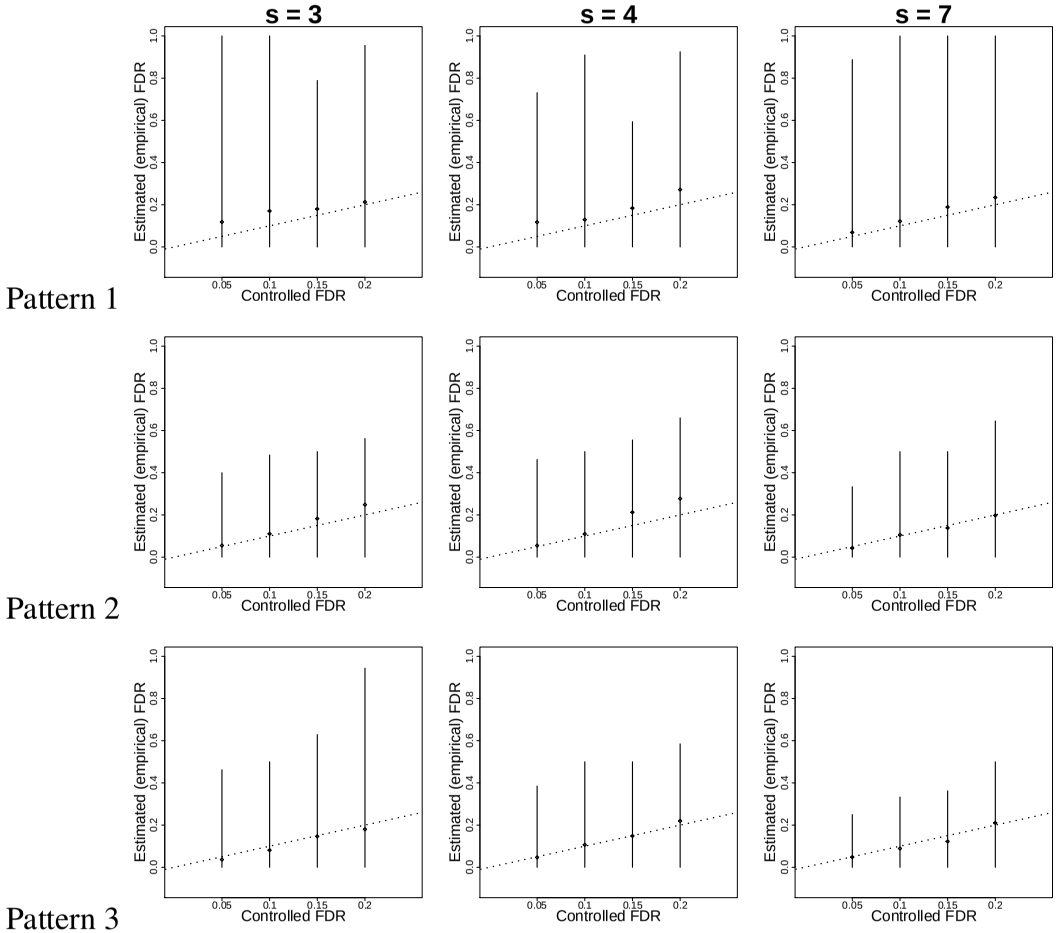


Fig. 8. Locations of the identified Local Authority Districts by (a) the proposed method using s=1.5 or s=2 with FDR controlled at different levels using stratified null distributions and (b) the space-time permutation test in SaTScan with p-value=0.05.