Biostatistics (2012), **13**, 4, *pp.* 695–710
doi:10.1093/biostatistics/kxs005
Advance Access publication on March 26, 2012

BaySTDetect: detecting unusual temporal patterns in small area data via Bayesian model choice

GUANGQUAN LI*, NICKY BEST, ANNA L. HANSELL
Department of Epidemiology & Biostatistics, Imperial College, London W2 1PG, UK and MRC-HPA Centre for Environment and Health, Imperial College, London W2 1PG, UK
guang.li@imperial.ac.uk

ISMAÏL AHMED,
Department of Epidemiology & Biostatistics, Imperial College, London, UK and INSERM, CESP Centre for Research in Epidemiology and Population Health, F-94807 Villejuif, France

SYLVIA RICHARDSON
Department of Epidemiology & Biostatistics, Imperial College, London, UK and MRC-HPA Centre for Environment and Health, Imperial College, London, W2 1PG, UK

SUMMARY

Space–time modeling 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 there is typically a general temporal trend, which affects all areas similarly, abrupt changes may occur in a particular area, e.g. due to emergence of localized predictors/risk factor(s) or impact of a new policy. Detection of areas with "unusual" temporal patterns is therefore important as a screening tool for further investigations. In this paper, we propose BaySTDetect, a novel detection method for short-time series of small area data using Bayesian model choice between two competing space–time models. The first model is a multiplicative decomposition of the area effect and the temporal effect, assuming one common temporal pattern across the whole study region. The second model estimates the time trends independently for each area. For each area, the posterior probability of belonging to the common trend model is calculated, which is then used to classify the local time trend as unusual or not. Crucial to any detection method, we provide a Bayesian estimate of the false discovery rate (FDR). A comprehensive simulation study has demonstrated the consistent good performance of BaySTDetect in detecting various realistic departure patterns in addition to estimating well the FDR. The proposed method is applied retrospectively 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 recognized 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. Our method would have led to an early detection of this local health issue.

Keywords: Bayesian spatiotemporal analysis; Disease surveillance; Detection; FDR; COPD.

*To whom correspondence should be addressed.

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 in 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 spatiotemporal 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. Reliable estimates can hence be obtained. As a natural extension to the purely spatial models such as those discussed in Best *and others* (2005), time trends are often modeled independently of the spatial pattern. For example, in disease mapping, the effects of space and time are typically modeled 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 behaviors. 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 modeled by the space + time main effects. In a series of papers by MacNab and Dean (2001) and MacNab (2007) time-series data for each spatial unit are modeled 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 behaviors. 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 particular time patterns, 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 (Knox and Bartlett, 1964) and Mantel's test (Mantel, 1967). A recent paper by Robertson *and others* (2010) provides a thorough discussion on various test-based detection methods. Among those, the space–time permutation scan statistic by Kulldorff *and others* (2005), a refinement of the space–time scan statistic of Kulldorff (2001) is close in spirit to the method proposed in this paper. Implemented in SaTScan, this method and its space-only version have been widely applied 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, p. 257). In the simulation study here, we will compare the performance of our proposed detection framework (referred to as BaySTDetect hereafter) 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 induced by a decision rule is bound to arise by pure chance. Under the frequentist framework, the seminal paper by Benjamini and Hochberg (1995) offered a procedure to estimate the false discovery rate (FDR), which is defined as the expected proportion

of the declared areas induced by a decision rule that are false positives. More recently, Newton *and others* (2004) and Müller *and others* (2004) provided a fully Bayesian treatment of this problem. Here, we employ the procedure introduced in Newton *and others* (2004) to classify areas into usual and unusual based on the posterior probabilities that each area follows the common trend. This procedure aims to ensure that the expected false positive proportion is close to the predefined FDR level given the data and the model.

With two substantive questions in mind, we analyze a set of mortality data on chronic obstructive pulmonary disease (COPD) in England and Wales (1990–1997) using BaySTDetect. COPD is a common chronic condition characterized by slowly progressive and irreversible decline in lung function. It is responsible for ∼5% of deaths in the United Kingdom (Hansell *and others*, 2003b). 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 onward in the United Kingdom (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 primary care contacts (Hansell *and others*, 2003a), 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 U.K. smoking trends over time. In addition to this, our second task is to explore the use of BaySTDetect 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 BaySTDetect, which involves a modeling framework and an FDR-based classification procedure. The COPD mortality data used in our case study will be described in Section 3. In Section 4, we will investigate the performance of BaySTDetect by a simulation study. Application of the method to the COPD data will be detailed in Section 5.

2. BaySTDetect: detection based on Bayesian model selection

2.1 *A general modeling framework*

We formulate the modeling framework within a disease-mapping context, although this framework is completely general for any type of small area data that can be modeled by latent Gaussian Markov random field (Rue and Held, 2005).

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 random variable, $Y_{i,t}$, is often assumed to model the count data. Specifically, at the first level of the model hierarchy, we have, $Y_{i,t} \sim \text{Poisson}(\mu_{i,t} \cdot E_{i,t})$ with $i = 1, \ldots, N$ and $t = 1, \ldots, T$.

With the aim of detecting areas with temporal trends that differ from the common trend, we propose to describe the distribution of relative risk $\mu_{i,t}$, by two alternative models, one that assumes 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, $\mu_{i,t}$ is modeled as

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

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 oversmooth local trends that display true departures. 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 Bayesian model selection formulation, for each area, a model indicator z_i is introduced to select estimates from either Model 1 ($z_i = 1$) or Model 2 ($z_i = 0$). The posterior frequency of selecting the common trend model, $f_i = P(z_i = 1|\text{data})$, can then be obtained, where a small value of f_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 modeling 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 (henceforth denoted BYM) 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 a Gaussian $N(0, \sigma_\eta^2)$ distribution. 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, and, otherwise $w_{i,j} = 0$. To implement the temporal RW(1) prior, we use its equivalent form of a 1D CAR model (see, e.g. Fahrmeir and Lang, 2001). Similar to the spatial CAR prior, the temporal neighborhood structure is defined through a matrix $\mathbf{Q}$, where $q_{h,t} = 1$ if $|h - t| = 1$ 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, i.e. $\sum_i v_i = 0$ (as in (2.2) below) and $\sum_t \gamma_t = 0$. 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. However, estimates of the area-specific temporal variances, $\sigma_{i,\xi}^2$, can be unreliable if modeled independently. Therefore, these (log) variances are modeled hierarchically.

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

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

A weakly informative half Normal prior $N(0, 1)$ bounded strictly below by 0 is assigned to each of σ_η, σ_v and σ_γ, as suggested by Gelman (2006). For the model indicator z_i, we have $z_i \sim$ Bernoulli(0.95). This prior reflects the surveillance nature of the analysis where we expect to find only a small number of unusual areas a priori. We return to this prior specification in Section 6. An $N(0, 1000)$ prior is assigned to the hyperparameter a, the mean of the log area-specific trend variances ($\sigma_{i,\xi}^2$), whereas for the variance parameter b^2, a moderately informative prior is assigned, namely, $b \sim N(0,2.5^2)$ bounded strictly below by 0. This prior on the standard deviation corresponds to our prior expectation that roughly 10% of areas would have local temporal variability 10 times greater than the population median, i.e. $P(\sigma_{i,\xi}^2 > 10 \cdot \exp(a)) \approx 0.10$.

The above modeling framework was implemented in WinBUGS (Lunn *and others*, 2000). The two competing models are fitted separately to the same set of data. At each Monte Carlo Markov chain (MCMC) iteration, an overarching mixture model, with a model indicator z_i taking values 0 or 1, compares probabilistically the fit of Models 1 and 2 in each area and selects one or the other model. The relative frequency (i.e., proportion of the total MCMC iterations) that each model is selected can be interpreted as the posterior probability of choosing that model. 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 but that variability of fit is properly taken into account. A schematic diagram of the modeling framework is given in Fig. 1 and annotated WinBUGS code is given in the supplementary material available at *Biostatistics* online.

For each model fitting, two MCMC chains with different starting values were run for 20,000 iterations with the first half discarded as burn-in and storing every fifth iteration of the second half for inference. Convergence of parameters (apart from z_i) was examined using the Gelman–Rubin statistics (Gelman and Rubin, 1992; Brooks and Gelman, 1998), which were below 1.05 for both the simulation study and the COPD application.

2.2 *Detection rules based on the Bayesian FDR*

The posterior model probability f_i, which is calculated at the end of the MCMC run as the posterior mean of the model indicator z_i, indicates how likely area i is to follow the common trend. A small f_i suggests a high probability that the corresponding area is unusual. But, put another way, if we declare this area to

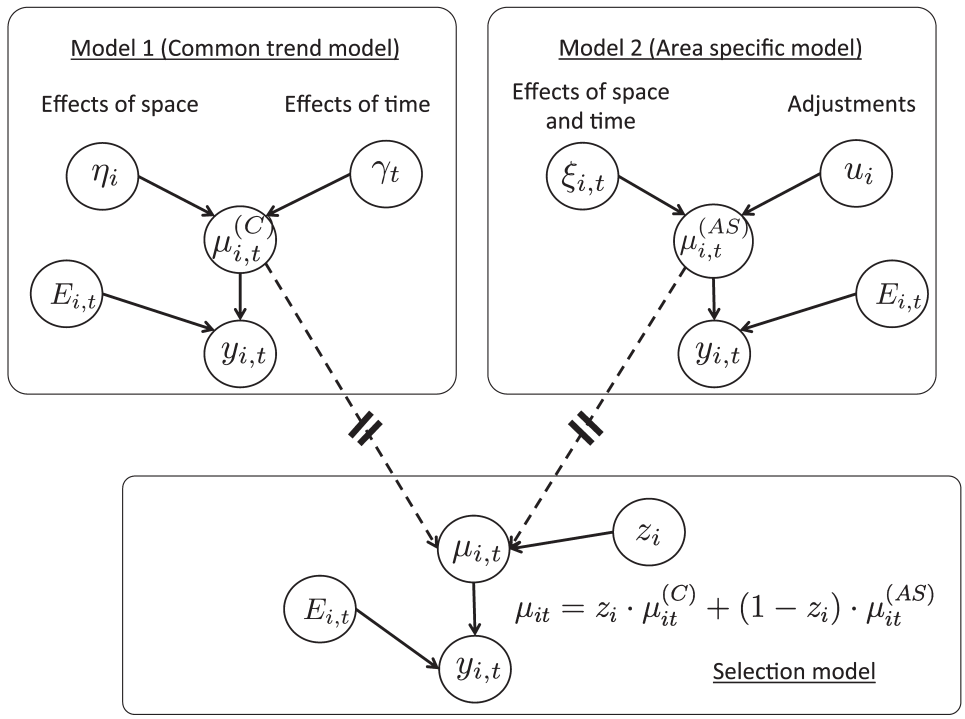

Fig. 1. A schematic diagram of the BaySTDetect modeling framework.

be unusual, f_i can be thought of as the probability of false detection for this area. Wakefield (2007) refers to this probability as the "Bayesian false discovery probability." We can utilize the dual interpretations of f_i to classify areas as well as to provide an estimate of the FDR.

To classify areas, we choose a suitable threshold, C, for f_i so that areas with $f_i \leqslant C$ are declared to be unusual. Furthermore, following Newton *and others* (2004) and Ventrucci *and others* (2011), the threshold C is chosen in such a way that the average probability of false detection (i.e. the average value of f_i) among areas declared to be unusual is less than or equal to some preset level α. That is, $C = f_{(k)}$, where k is the maximum integer such that $1/k \cdot \sum_{j=1}^{k} f_{(j)} \leqslant \alpha$, with $f_{(j)}$ denoting the jth ordered posterior model probability. This procedure aims to ensure that, on average, no more than $(k \cdot \alpha)$ of these detected areas truly follow the common trend, i.e. are false positives.

3. DATA DESCRIPTION

National mortality data on COPD (ICD9 490–496) from 1990 to 1997 were provided by the U.K. 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 (mainly occupied by businesses, particularly financial services, with very few residents) 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. Table 1 in supplementary material available at *Biostatistics* online summarizes the standardized mortality ratios.

4. SIMULATION STUDY

In this section, we present a small simulation study designed to evaluate the (frequestist) operating characteristics of BaySTDetect under various realistic scenarios.

4.1 *Generating the data*

Data were generated for 8 time points for 354 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 (summarized in Table 2 in supplementary material available at *Biostatistics* online) from the real COPD data or a reduced set. The latter is formed by multiplying the original set by one-fifth in order 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 method.

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, $\sim$4% of a total 354, were chosen to be unusual areas; selection of these areas was based on ensuring a good contrast of expected counts and levels of risk (see supplementary material available at *Biostatistics* online).

Figure 2 illustrates the three departure patterns considered, which are representations of those seen in real analyses (e.g. Glass, 1998). Construction of these patterns is provided in supplementary material available at *Biostatistics* online. Two realistic departure magnitudes are used, $\theta = 1.5$ and 2, for each

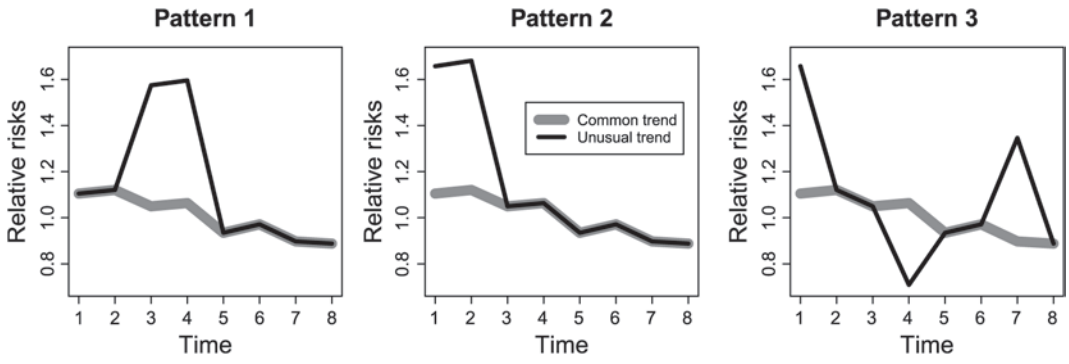

Fig. 2. Illustration of the three departure patterns (black), compared to the common trend pattern in gray. The departure magnitude θ in these plots is 1.5.

departure pattern. For the remaining 339 areas, the common trend is assigned. Fifty sets of data were generated under each of the 12 simulation scenarios (3 departure patterns × 2 magnitudes × 2 sets of expected counts).

4.2 *Results*

Model performance is summarized by empirical FDR and sensitivity. To calculate the empirical FDR, we take the mean (over the 50 replicates) of the false positive proportion, FP $= \frac{V}{R}$, where V is the number of false positives and R is the total number of declared areas. When $R = 0$, FP is set to 0. For the sensitivity, we record the percentage of times (out of 50 replicates) that each of the 15 truly unusual areas was correctly identified. The receiver operation characteristic (ROC) curve is not used here for comparison because it suppresses the differences in sensitivity for areas with different levels of expected counts and overall risks.

For comparison, the space–time permutation test in SaTScan (Kulldorff, 2010) was also fitted to the 50 replicates. The threshold p value, under which excess is declared, is set at 0.05.

Figure 3 shows that the empirical FDR estimated from BaySTDetect is close to the corresponding preset levels (column 1) or lower (column 3), where the latter suggests the FDR-based detection procedure appears to be somewhat "conservative" when data are very sparse. When data were simulated from the original expected counts with $\theta = 2$ (middle column), a moderate positive bias appears, although the empirical FDR is still reasonably close to the preset 0.05 level. The 95% sampling intervals of the false positive proportion over 50 replicates are generally wider under the reduced expected counts than those under the original expected counts and also wider when $\theta = 1.5$ than when $\theta = 2$.

With the p value threshold set to 0.05, the empirical FDRs from SaTScan were quite high (in general $>$0.15) with highly variable false positive proportions under the departure scenarios considered (Table 3 in supplementary material available at *Biostatistics* online).

For the sake of illustration, sensitivity of BaySTDetect is evaluated at the nominal $\alpha = 0.1$ FDR level. Using the original expected counts, Fig. 4 summarizes the ability to detect the 15 truly unusual areas using BaySTDetect (Columns 1 and 3) and SaTScan (Columns 2 and 4) under three departure patterns (rowwise). In each plot, the probabilities of correctly detecting the 5 truly unusual areas, each having an overall expected count at one of the 5 percentiles, are joined by solid line (low spatial risks), dashed line (median spatial risks), and dotted line (high spatial risks). Overall, BaySTDetect is more powerful in identifying the truly unusual areas than SaTScan, in particular when the data are sparse, e.g. detecting

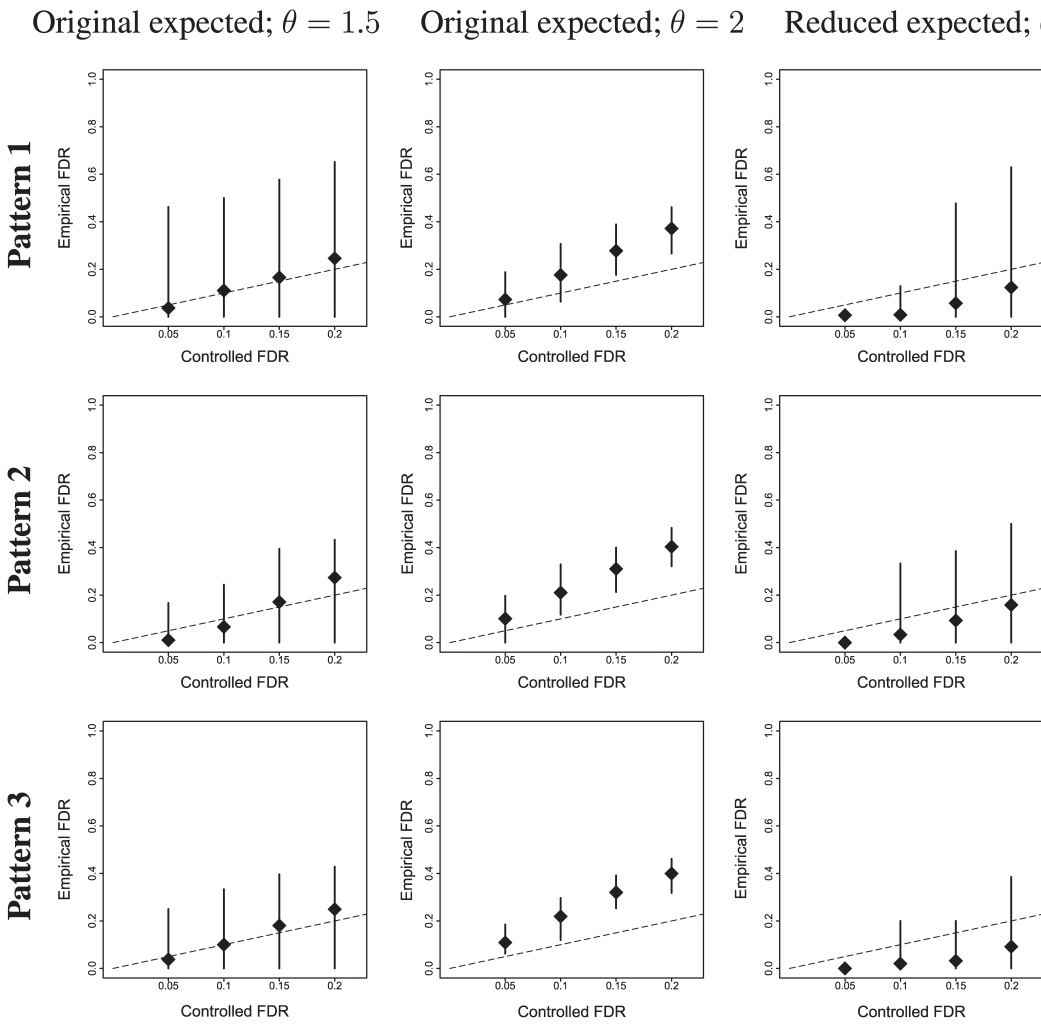

Fig. 3. Comparisons of the empirical FDR from BaySTDetect to the corresponding preset level based on the original set of expected counts (first two columns; $\theta = 1.5$ and 2, respectively) and the reduced expected counts (third column with $\theta = 2$) under three departure patterns (rowwise). The mean and 95% sampling interval of the false positive proportion (over 50 replicates) are represented by the point and the vertical bar, respectively. The dash line represents the 45° line.

areas at the lower percentiles of the expected count distribution (when $\theta = 2$) or with lower spatial risks (when $\theta = 1.5$).

Within each plot in Fig. 4, all lines consistently show an overall increasing pattern, indicating that both BaySTDetect and SaTScan tend to be more powerful in detecting changes in areas with larger expected counts, an unsurprising result. The detection power also generally increases with the increasing level of spatial risk (from the solid line, to the dashed line, to the dotted line). 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 second and fourth columns of Fig. 4). This is because these two areas happen to be close together, a situation where the scanning windows used in SaTScan has the advantage.

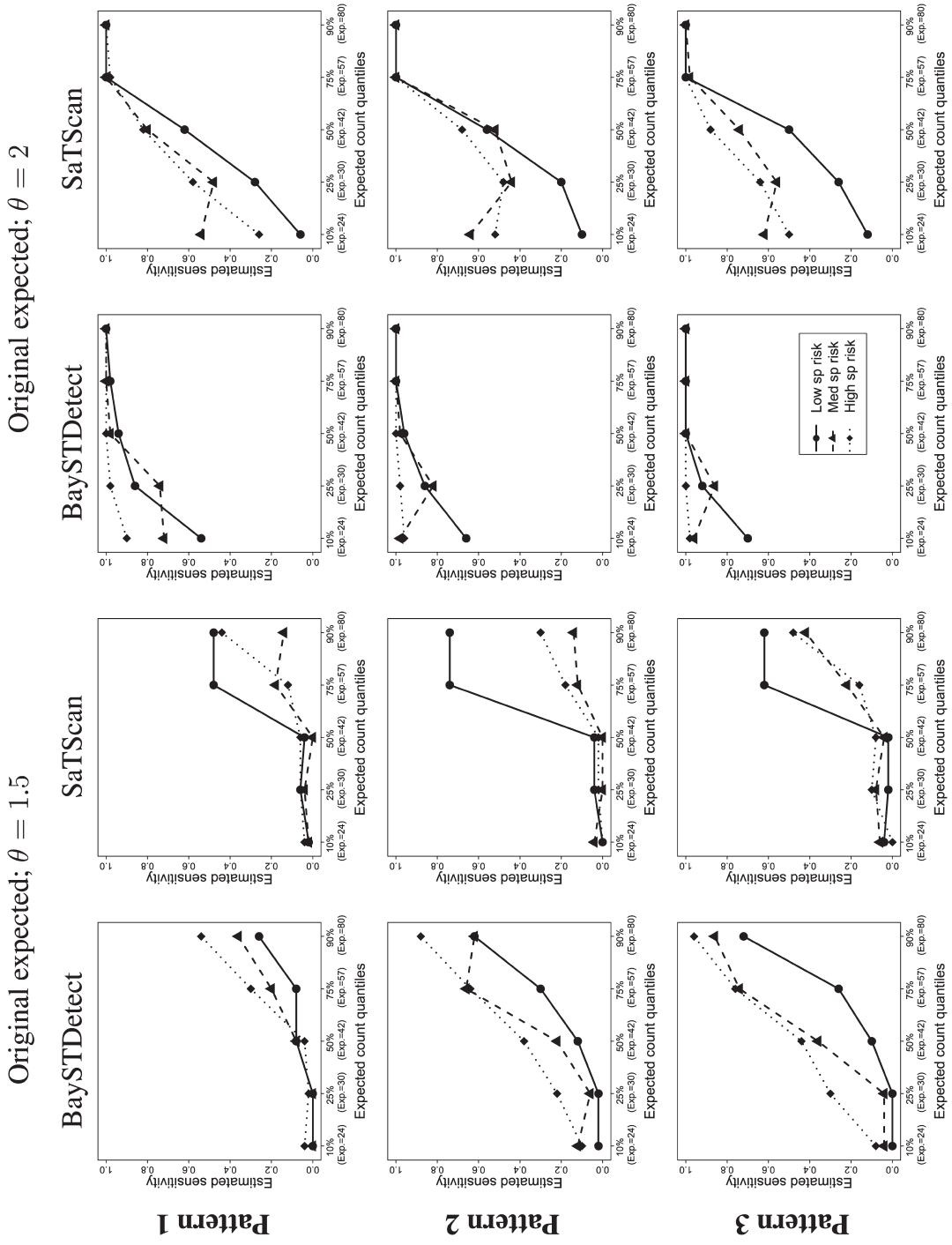

Fig. 4. Sensitivity of detecting the 15 truly unusual areas with departures magnitudes of $\theta = 1.5$ and $\theta = 2$ using BaySTDetect (with FDR preset at 0.1) and SaTScan (threshold for p value $= 0.05$). Data were generated from the original set of expected counts. The expected number of cases at each percentile is given in brackets.

Besides the expected number and the level of spatial risks, the detection power also somewhat depends on the departure pattern. 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 in 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 patterns 2 or 3. Such a difference is less marked as expected counts, spatial risks, and/or departure magnitudes become higher.

When data were generated from the reduced expected numbers, neither SaTScan nor BaySTDetect could pick up areas with departures of $\theta = 1.5$ (results not shown) at such a high level of sparsity. Although shown to be somewhat "conservative," BaySTDetect demonstrates an overall better performance than SaTScan with $\theta = 2$ (Fig. 5), particularly in detecting areas with medium (dashed lines) or high (dotted lines) spatial risks under departure patterns 2 and 3.

The detection results from BaySTDetect were robust against different priors on b, the standard deviation of the log temporal variances (results not shown). Changing the prior on z_i to a "less informative" one, e.g. Bern(0.5), appeared to increase the estimated FDR slightly, pushing the points in Fig. 3 slightly upward but was robust in terms of power (see Section 6 for further discussion).

5. APPLICATION: COPD

Motivated by the policy and public health issues discussed at the beginning of this paper, we analyze the COPD data using BaySTDetect 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 (Spiegelhalter *and others*, 2002) and hence was used as the specification for the common trend model in (2.1).

At the FDR level of 0.05, five LADs (out of 374) were identified (Fig. 6(a)) among which three (Rotherham, Carmarthenshire, and Barnsley) were in mining areas (out of a total 40 mining districts). Although a high proportion of the detected areas are mining areas, Carmarthenshire and Barnsley displayed an overall increasing trend during 1990–1997 while Rotherham had a faster decreasing trend compared to the common trend (Fig. 1 in supplementary material available at *Biostatistics* online). In addition to any potential impact of the policy, several other changes that occurred over this time period could have contributed to these departures of time trend patterns. For example, a very large number of mines closed from the mid 1980s in the United Kingdom, which would have dramatically reduced the impact of mining dust exposures on COPD development (and subsequently mortality) in mining areas. 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). The inconsistency of the unusual trend patterns associated with these three detected mining areas did not provide evidence to support the policy-related hypothesis set out in Section 1.

The other two unusual districts with a statistically significant local increasing trend between 1990 and 1997 (against a national decreasing trend) were detected in inner London (Lewisham and Tower Hamlets in Fig. 1 in supplementary material available at *Biostatistics* online). 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 United Kingdom. In

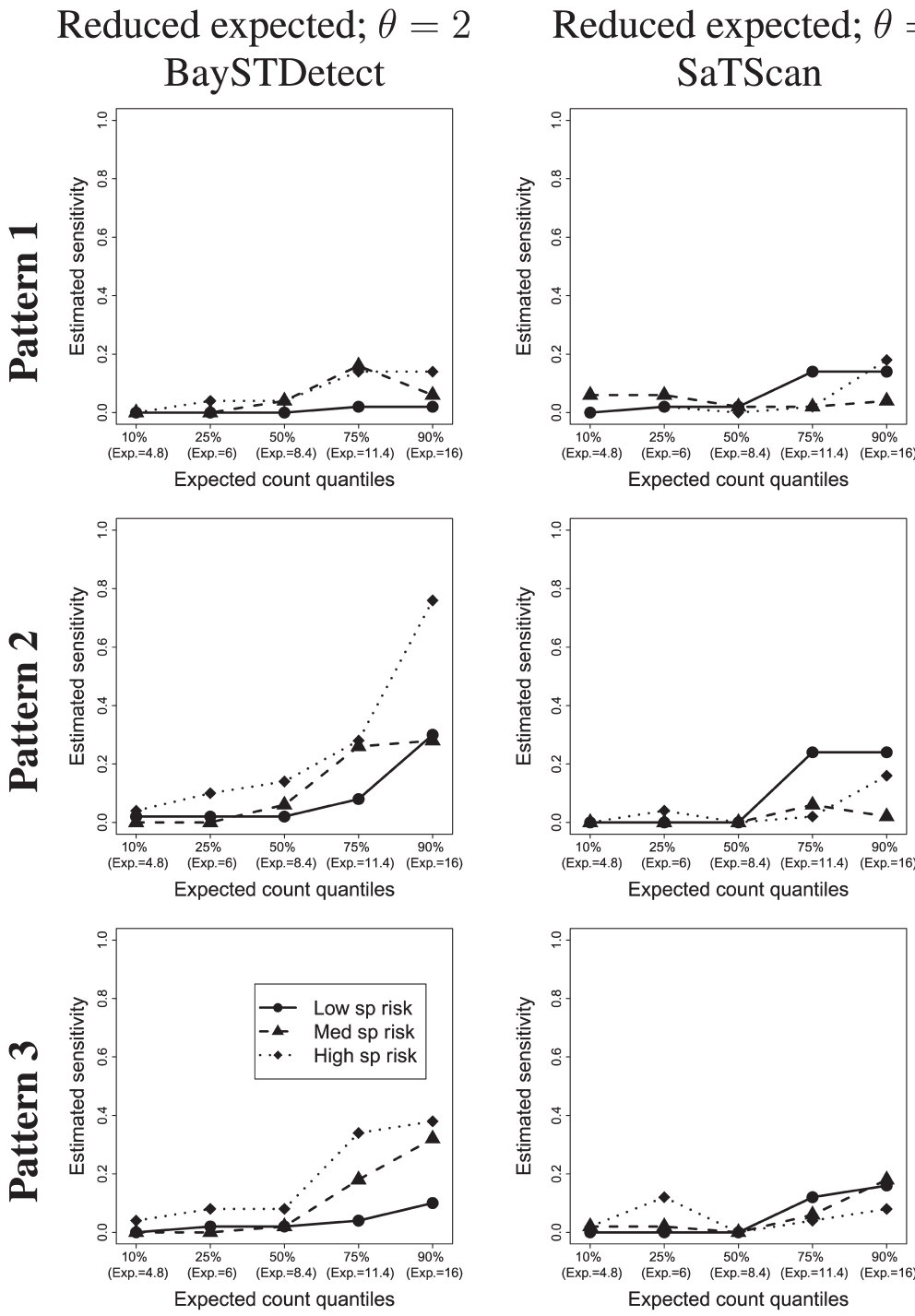

Fig. 5. Sensitivity of detecting the 15 truly unusual areas with departures magnitude $\theta = 2$ using BaySTDetect (with FDR preset at 0.1) and SaTScan (threshold for p value $= 0.05$). Data were simulated from the reduced set of expected counts. The expected number of cases at each percentile is given in brackets.

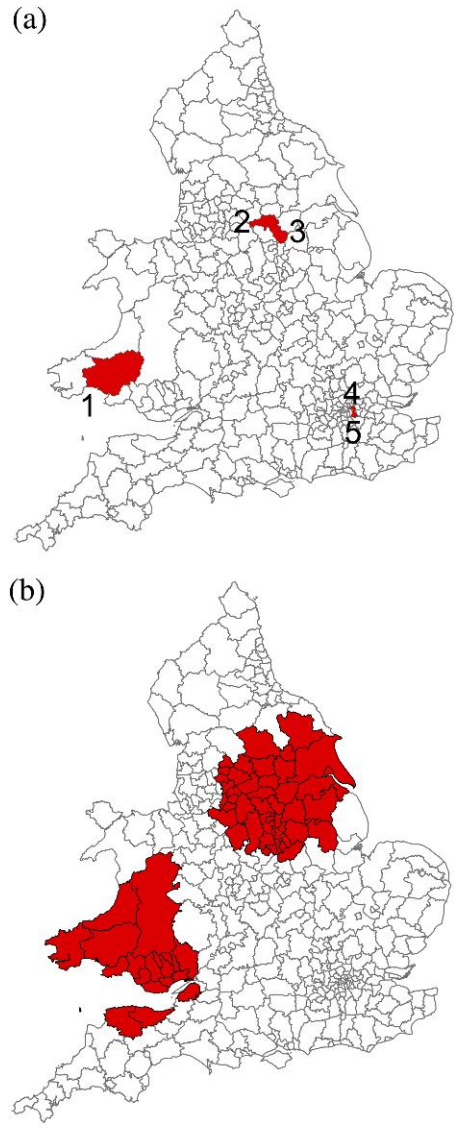

Fig. 6. Locations of the identified LADs by (a) BaySTDetect with FDR predefined at 0.05 (1. Carmarthenshire, 2. Barnsley, 3. Rotherham, 4. Tower Hamlets, and 5. Lewisham) and (b) the space–time permutation test in SaTScan with threshold for p value set at 0.05.

fact, Tower Hamlets has been commissioning Local Enhanced Services since 2008 in order to optimize patient management and to reduce COPD (Tower Hamlets Council, 2009; NHS-Tower Hamlets, 2009), but the rising trend in COPD could potentially have been recognized earlier in the 1990s through using BaySTDetect.

As we increase the FDR threshold to 0.1 and 0.15, there were 7 and 9 LADs, respectively, identified as unusual (Fig. 2 in supplementary material available at *Biostatistics* online).

Two circular clusters of large numbers of areas were detected by SaTScan (Fig. 6 (b)), both of which were associated with mining areas. The one in the north of England, containing 46 LADs (including 14

mining areas), expressed an excess risk of 1.05 during 1990–1992 while the one in Wales and the southwest with 20 LADs (including 7 mining areas) 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

BaySTDetect 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 well estimating the FDR.

In BaySTDetect, we use the posterior model probability of selecting the common trend model, f_i, for detection. Several procedures have been proposed to calibrate this statistic, such as the decision theoretic approach in Müller *and others* (2006) and Wakefield (2007) in genetic studies, and through predefining the FDR level (e.g. Newton *and others*, 2004; Ventrucci *and others*, 2011). Here, we follow the latter and obtain the classification threshold based on a predefined FDR. If the costs associated with false positives and false negatives can be specified, other thresholding criteria could be used.

How well the FDR is estimated depends somewhat on the prior on the model indicator z_i. This is partly because the derivation of the threshold C in Section 2.2 does not take into account the prior on z_i. An unrealistic "uniform" prior (i.e. Bern(0.5)) on z_i would result in underestimating the true FDR if, for example, only 10% of areas are expected to be unusual (see supplementary material available at *Biostatistics* online for explanation). In fact, an implicit assumption of our model formulation is that the proportion of unusual areas is small a priori, perhaps no more than 10%, so that a common trend can be meaningfully defined and estimated. This suggests a lower bound of about 0.9 for the prior on z_i and means that our method is suitable for detecting relatively rare events, such as early disease outbreaks or assessing impact of a policy which only targets a small proportion of areas. An alternative way to calibrate the f_i is to approximate its sampling distribution under the null via Monte Carlo simulation and adopt a frequentist procedure for controlling the FDR. Although computationally intensive, the performance in bounding the preset FDR is independent of the prior specification as this calibration is model specific. Comparison of these different procedures is a subject of ongoing research.

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 study, the reduced set of expected counts presents a minimal level of information beyond which BaySTDetect 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. In addition, the power of BaySTDetect 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 detection more specific, rather than clusters of areas, which SaTScan usually identifies.

Public health surveillance systems are commonly used to monitor infectious diseases but rarely performed for chronic diseases. As demonstrated, the proposed method has high policy relevance for national or regional chronic disease surveillance to help identify departures from common trends that may require investigation and, perhaps, interventions. For example, the detection results from BaySTDetect using the COPD mortality data 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 over 10 years after the increasing trend started (Tower Hamlets Council, 2009; NHS-Tower Hamlets, 2009).

The BaySTDetect modeling framework can be readily adapted to monitor infectious diseases, where areas with departures are likely to form local clusters. Spatial dependence of model choice can be induced

through a Gaussian random field prior on z_i (e.g. Fernandez and Green, 2002) such that choice of model depends not only on the data but also on the hypothesized spatial structure of the alternative, potentially achieving higher power.

The time window over which changes are detected also needs to be considered. BaySTDetect 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 since it assumes that the unusual trend applies to the entire time period. The modeling framework may need to be extended so that the model indicator is specific to both area and time point, namely, $z_{i,t}$. Furthermore, we are currently developing a sequential fitting of BaySTDetect, where data are fed one time point at a time in order to pinpoint the time of departure and to estimate the magnitude of departure from the common trend. This sequential framework could also be useful to initiate public health measures promptly.

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

We would like to thank the Small Area Health Statistics Unit (SAHSU) for provision of the COPD data. The authors thank both reviewers and the Associate Editor for their constructive comments. *Conflict of Interest:* None declared.

FUNDING

This research was funded by the ESRC National Centre for Research Methods (BIAS II node, RES-576-25-0015). 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**, 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**, 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**, 1–20.

BEST, N. AND HANSELL, A. (2009). Geographic variations in risk: adjusting for unmeasured confounders through joint modeling of multiple diseases. *Epidemiology* **20**, 400–410.

BEST, N., RICHARDSON, S. AND THOMSON, A. (2005). A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research* **14**, 35–59.

BROOKS, S. P. AND GELMAN, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* **7**, 434–455.

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

FAHRMEIR, L. AND LANG, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors. *Applied Statistics* **50**, 201–220.

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

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

GELMAN, A. AND RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457–472.

GLASS, G. V. (1998). Interrupted time-series quasi-experiments. *Technical Report.* Phoenix, Arizona: Arizona State University.

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

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

HANSELL, A., WALK, J. AND SORIANO, J. (2003b). What do chronic obstructive pulmonary disease patients die from? A multiple cause coding analysis. *European Respiratory Journal* **22**, 809–814.

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

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

KNOX, E. AND BARTLETT, M. (1964). The detection of space-time interactions. *Journal of the Royal Statistical Society Series C* **13**, 25–30.

KULLDORFF, M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society Series A* **164**, 61–72.

KULLDORFF, M. (2010). *Software for the Spatial and Space-Time Scan Statistics* (SaTScan v9.0.1). Available from http://www.satscan.org/.

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**, 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. *European Respiratory Journal* **27**, 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**, 325–337.

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

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

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

MILLER, B. AND MACCALMAN, L. (2010). Cause-specific mortality in british coal workers and exposure to respirable dust and quartz. *Occupational and Environmental Medicine* **67**, 270–276.

MÜLLER, P., PARMIGIANI, G. AND RICE, K. (2006). FDR and Bayesian multiple comparisons rules. Johns Hopkins University, Department of Biostatistics Working Papers, 115.

MÜLLER, P., PARMIGIANI, G., ROBERT, C. AND ROUSSEAUR, J. (2004). Optimal sample size for multiple testing: the case of gene expression microarrays. *Journal of the American Statistical Association* **99**, 990–1001.

NEWTON, M. A., NOUEIRY, A., SARKAR, D. AND AHLQUIST, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5**, 155.

NHS-TOWER HAMLETS. (2009). NHS Tower Hamlets strategic plan 2009-2010 to 2012-2013. *Technical Report*, National Health Services (NHS, UK).

ROBERTSON, C., NELSON, T. A., MACNAB, Y. C. AND LAWSON, A. 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.

RUE, H. AND HELD, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, Vol. 104. London: Chapman & Hall.

SEATON, A. (1998). The new prescription: industrial injuries benefits for smokers? *Thorax* **53**, 335–336.

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.

TOWER HAMLETS COUNCIL. (2009). Health and wellbeing in Tower Hamlets—joint strategic needs assessment. *Technical Report*, Tower Hamlets Council.

VENTRUCCI, M., SCOTT, M. AND COCCHI, D. (2011). Multiple testing on standardized mortality ratios: a Bayesian hierarchical model for FDR estimation. *Biostatistics* **12**, 51–67.

WAKEFIELD, J. (2007). A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *The American Journal of Human Genetics* **81**, 208–227.

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.

[*Received November 12, 2010; revised February 8, 2012; accepted for publication February 11, 2012*]