

Data mining cancer registries: retrospective surveillance of small area time trends in cancer incidence using BaySTDetect

Guangquan Li, Sylvia Richardson, Lea Fortunato, Ismaïl Ahmed, Anna Hansell, Mireille Toledano and Nicky Best

Department of Epidemiology and Biostatistics

School of Public Health, Faculty of Medicine

Imperial College London

St Mary's Campus, Norfolk Place, London, W2 1PG, UK

Email: guang.li@imperial.ac.uk (To whom correspondence should be addressed)

Abstract—Space-time modelling of small area data is often used in epidemiology for mapping temporal trends in chronic disease rates. For rare diseases such as cancers, data are sparse, and a Bayesian hierarchical modelling approach is typically adopted in order to smooth the raw disease rates. Although there may be a general temporal trend which affect all areas similarly, abrupt changes may also occur in particular areas due to, for example, emergence of localized risk factor(s) or impact of a new health or screening policy. Detection of areas with “unusual” temporal patterns is therefore important to flag-up areas warranting further investigations.

In this paper, we present a novel area of application of a recently proposed detection method, BaySTDetect, for short time series of small area data. Placed within the Bayesian model choice framework, BaySTDetect detects unusual time trends based on comparison of two competing space-time models. The first model is a straightforward multiplicative decomposition of the area effect and the temporal effect, assuming one single temporal pattern across the whole study region. The second model estimates a local time trend, independently for each area. An area-specific model indicator is introduced to select which model offers a better description of the local data. Classification of an area local time trend as “unusual” or not is based on the posterior mean of this model indicator, which represents the probability that the common trend model is appropriate for that area. An important feature of the method is that the classification rule can be fine-tuned to control the false detection rate (FDR). Based on previous simulation results, we present some further insights of the model specification in relation to the detection performance in practice. BaySTDetect is then applied to data on several different cancers collected by the Thames Cancer Registry in South East England to illustrate its potential in retrospective surveillance.

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

I. INTRODUCTION

Bayesian spatial and spatio-temporal mapping of small area disease rates is now common practice in spatial epidemiology, as a means of visualising spatial and temporal variation in disease risk [1]. The spatial pattern of risk may vary greatly across a study region, but smooth time changes are typically expected in most areas. However, due to emergence of localized risk factors (such as a new source of pollution), local policy implementation (for example, health

awareness or screening campaigns), or changes to health care provision or social structure of the local population, 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 disease data. Similar issues also arise in the analysis of other types of small-area data, such as crime or unemployment rates.

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 (e.g., [2], [3], [4], [5], [6]). Although uncertainty of estimates can be reduced, the smoothing nature of these models is not suitable for the detection purpose. Constructed within the Bayesian hierarchical framework, we recently proposed BaySTDetect for detecting unusual time patterns [7].

A test-based method that is close in spirit to BaySTDetect is the space-time permutation scan statistic introduced in [8], a refinement of the space-time scan statistic of [9]. Implemented in SaTScan, this later method and its space-only version have been popular in disease surveillance. However, our previous simulation study [7] showed that SaTScan was inferior to BaySTDetect in detecting isolated clusters (i.e., elevated risk in a single or very small number of areas), partly because of the construction of the cylindrical scanning volume in SaTScan [7]. Furthermore, in the simulation setting that we have considered, SaTScan generally resulted in large number of false positives.

Multiple comparison is another crucial issue to address under any detection model. Some proportion of the classified areas are bound to arise by pure chance due to the large number of tests performed. To tackle this problem, BaySTDetect utilises a Bayesian procedure in deriving decision rules that enable to control 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.

II. BAYSTDetect: DETECTION BASED ON BAYESIAN MODEL SELECTION

In this section, we recall the construction of BaySTDetect [7] in the context of disease surveillance, although the detection framework can be adapted to other types of small area data such as crime-mapping.

A. A modelling framework for disease surveillance

When the disease of interest is rare and non infectious, a Poisson distribution is often assumed to model the count data. Let $y_{i,t}$ and $E_{i,t}$ be the observed and expected numbers of disease cases, respectively, in area i at time t . We have, at the first level of the model hierarchy,

$$y_{i,t} \sim \text{Poisson}(\theta_{i,t} \cdot E_{i,t})$$

with $i = 1, \dots, N$ and $t = 1, \dots, T$.

The aim here is to detect areas with temporal trends that differ from the common trend so we propose to describe $\theta_{i,t}$, the relative risk by two alternative models. The first model is separable in time and space, in other words it estimates the disease risk in each area and time point as the sum (on the log scale) of a spatial term and a temporal term that is common to all areas. The second model fits a different time series model separately for each area. Specifically, at the second level of the model hierarchy, $\theta_{i,t}$ is modelled as

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

In Model 1, the effects of space, η_i , and time, γ_t , are (on the log scale) additively combined, and hence the temporal trend pattern is the same for all areas, an assumption that can over-smooth local trends that display true departures. The alternative Model 2 (or referred to as the area-specific model) is introduced so that substantial departure from the common trend pattern can be accommodated. In Model 2, the temporal trends are estimated independently for each area, where u_i is the area-specific intercept and $\xi_{i,t}$ depicts the local trend patterns. To choose between the two model formulations, a model indicator z_i is used for each area to indicate whether Model 1 ($z_i = 1$) or Model 2 ($z_i = 0$) is supported by the data. The posterior mean, $p_i = P(z_i = 1 | \text{data})$, then represents the posterior probability that the area shows no departures from the common trend pattern. A small value of p_i indicates that the trend pattern of area i is unlikely to follow that of the common trend, γ_t .

Under the Bayesian modelling framework, priors are to be assigned to all unknown quantities. In 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 [10], 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 $N(0, \sigma_\eta^2)$. 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. 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, an area-specific intercept u_i is used to additively adjust the estimated trend patterns to the local level. 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.

The full prior specification of the two models is as follows.

Model 1

$$\begin{aligned} \alpha_0 &\sim \text{Uniform}(-\infty, +\infty) \\ \eta_i &\sim N(v_i, \sigma_\eta^2) \\ v_{1:N} &\sim \text{CAR}(\mathbf{W}, \sigma_v^2) \\ \gamma_{1:T} &\sim \text{CAR}(\mathbf{Q}, \sigma_\gamma^2) \end{aligned}$$

Model 2

$$\begin{aligned} u_i &\sim N(0, 1000) \\ \xi_{i,1:T} &\sim \text{CAR}(\mathbf{Q}, \sigma_{i,\xi}^2) \\ \log(\sigma_{i,\xi}^2) &\sim N(a, b^2) \end{aligned}$$

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 [11]. For the model indicator z_i , we have $z_i \sim \text{Bernoulli}(\pi)$ where $\pi = 0.95$ in [7]. This prior on z_i reflects the surveillance nature of the analysis where we expect to find only a small number of unusual areas *a priori*. In Section III, we investigate how this prior affects the model performance. A Normal(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 allows for nontrivial probability that the local temporal variance can be 10 times larger than the population median, i.e., $p(\sigma_{i,\xi}^2 > 10 \cdot \exp(a)) > 0$.

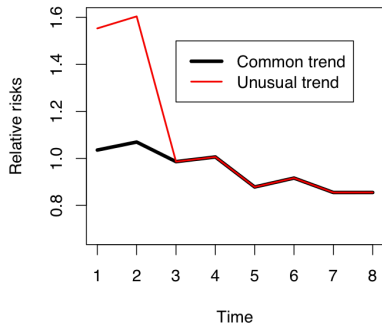
B. Detection rule based on Bayesian FDR

The following procedure is employed to derive a classification threshold. At a given FDR level, say α , a maximum integer-valued k is sought such that

$$\sum_{j=1}^k p_{(j)} \leq k \cdot \alpha$$

with $p_{(1:N)} \equiv 0 \leq p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(N)}$ denoting the vector of the posterior probabilities of the common trend model in an ascending order and N denoting the number

Figure 1. Illustration of the departure pattern (red), compared to the common trend pattern in black. The departure magnitude in this plot is 1.5.



of areas (or the number of tests we perform). The corresponding areas with $p_{(1:k)}$ are then classified as unusual. The simulation study in [7] shows that this procedure can provide reasonably good control for the predefined FDR level. In other words, on average, no more than $k \cdot \alpha$ of k declared unusual areas would have been truly usual, i.e., false positives.

III. SIMULATION STUDY

In [7], a comprehensive simulation study has demonstrated a consistent good performance of BaySTDetect in detecting various realistic departure patterns, in addition to estimating well the FDR, in comparison to SaTScan. In this paper, we first briefly recall some key results from [7]. We then present some further simulation results based on sparse data, a situation often encountered in disease surveillance. In such a case, one may choose to increase the number of false discoveries induced in order to gain sensitivity. We will demonstrate two ways of achieving this under BaySTDetect. Furthermore, we will give recommendation for selecting the “optimal” way to increase sensitivity in practice. Comparison to SaTScan is also carried out.

As in [7], data were generated based on a set of mortality data from England covering 354 areas and 8 years. Fifteen areas are selected to have the “unusual” trend pattern shown in Figure 1. The area selection is based on their spatial risks (low, medium and high) and expected counts (10^{th} , 25^{th} , 50^{th} , 75^{th} and 90^{th} percentiles of the expected count distribution) in order to provide a good contrast of the detection power – data from an area with high overall spatial risk/large expected counts are less noisy than those from an area with low overall spatial risk/small expected counts and hence the former is easier to detect if departure is present. Using the expected counts summarised in Table I, 50 sets of data were generated where the unusual trend pattern was assigned to the 15 selected areas while the common trend pattern was used for the remaining 339 areas. Two departure magnitudes, $\theta = 1.5$ and 2 were considered so the unusual trend pattern was formed by multiplying the common time

trend by θ at time points 1 and 2. Further details on the simulation setup see [7].

Model performance is summarised in terms of sensitivity and empirical FDR. To calculate the empirical FDR, we take the mean of the false proportion rates (FPR), i.e., $\frac{V}{R}$, over the 50 samples, where V and R are the number of false detected areas and the total number of detected areas, respectively. When no areas are detected, i.e., $R = 0$, the ratio is set to zero. For the sensitivity, we record the percentage of times (out of 50 samples) that each of the 15 truly unusual areas was correctly identified.

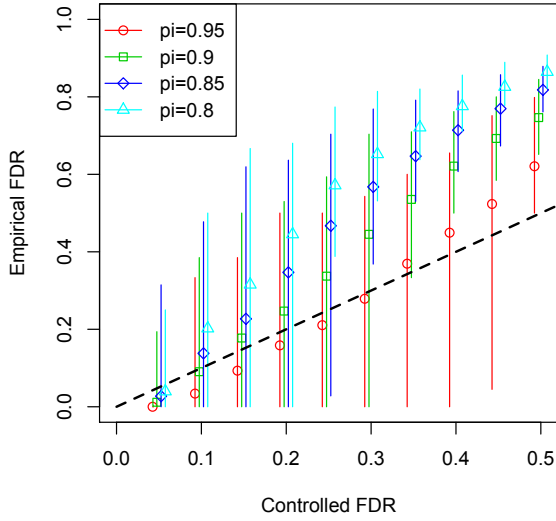
Detection performance based on the set of expected counts in Table I with $\theta = 2$ was previously discussed in [7]. In general, BaySTDetect was more powerful in identifying the truly unusual areas than SaTScan, in particular in detecting areas with medium to high overall spatial risks. The FDR estimate from BaySTDetect was moderately below the predefined level when data are sparse, in other words, being slightly conservative [7].

Because of the excessive noise inherent in sparse data, detecting unusual trends becomes more difficult when the departure magnitude is small $\theta = 1.5$ (see, for example, first column of Figure 4 below). In order to gain power, one can think of modifying parameter(s) of a detection method to allow for more false positives. An obvious choice is to increase the classification threshold, namely, the FDR control level in BaySTDetect or the p-value threshold in SaTScan. Alternatively, specific to BaySTDetect, one can retain the same FDR control level but reduce the prior probability on each area following the common trend model, i.e., specify a smaller value of π in the prior, $z_i \sim \text{Bern}(\pi)$, for the model indicator. Li et al. [7] has shown that π and the empirical FDR are negatively correlated. Therefore, reducing π leads to higher power but also more false positives. We now examine the detection performance of these two modifications in BaySTDetect. Figure 2 compares the empirical FDR to the corresponding controlled levels with various values for π . This clearly shows that, for a given pre-set FDR control level, the empirical FDR (i.e., the actual false positive rate observed in the simulated data) increases as π decreases. Figure 2 also reveals that the empirical FDR is closer to the pre-set FDR control level (i.e., less biased) for higher values of π . Close inspection reveals that certain combinations of π and the FDR control level yield similar empirical FDR estimates. For instance, (FDR=0.15 + $\pi=0.80$), (FDR=0.20 + $\pi=0.85$) and (FDR=0.25 + $\pi=0.90$) all give FDR estimates close to 0.33; likewise (FDR=0.20 + $\pi=0.80$), (FDR=0.25 + $\pi=0.85$), (FDR=0.30 + $\pi=0.90$) and (FDR=0.40 + $\pi=0.95$) all yield empirical FDR estimates of around 0.45 (Figure 2). In each case, the settings with lower π values result in a higher FDR bias. Figure 3 compares the sensitivity of the two pairs of settings highlighted above, suggesting that similar levels of sensitivity can be achieved with different combinations of π and FDR control level.

Table I
SUMMARY OF THE SET OF EXPECTED COUNTS USED IN THE SIMULATION

	Median	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8
Min	0.98	0.97	0.93	0.94	0.99	0.94	1.02	1.16	1.18
10%	2.42	2.30	2.29	2.32	2.37	2.45	2.51	2.54	2.59
30%	3.31	3.13	3.17	3.23	3.30	3.35	3.48	3.55	3.64
50%	4.22	3.95	4.04	4.10	4.18	4.23	4.35	4.45	4.55
70%	5.16	4.95	5.02	5.08	5.12	5.17	5.31	5.42	5.58
90%	7.98	7.78	7.82	7.87	7.91	8.03	8.20	8.34	8.53
Max	31.84	31.59	31.72	31.66	31.80	31.87	32.63	32.904	33.20

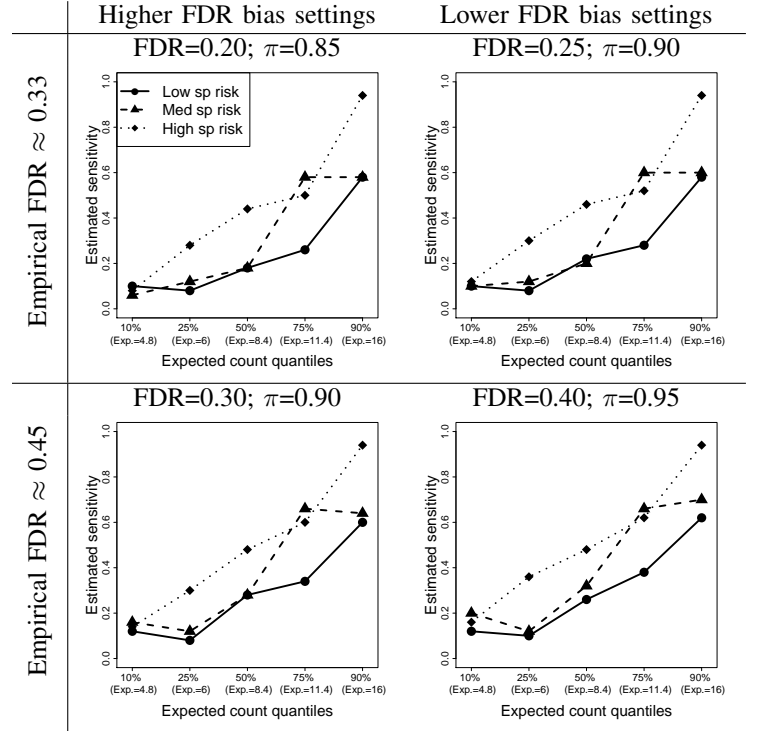
Figure 2. Comparison of the estimated FDR to the corresponding controlled level based on the expected counts in Table I with departure magnitude of $\theta = 2$. Four settings on π are used. The estimated FDR (the mean false discovery proportion across the 50 replicates) and the 95% sampling interval are represented by the point and the vertical bar, respectively. The dash line represents the 45° line.



Since higher values of π lead to better empirical control of the FDR (i.e., the empirical FDR is on average close to the pre-set control level), it is recommended in practice to use a meaningful prior on z_i (typically $\pi=0.95$ or $\pi=0.90$), and then to relax the FDR control level to gain detection power.

Figure 4 demonstrates the (slight) increased power in detecting areas with small departure ($\theta = 1.5$) with an increasing FDR control level from BaySTDetect (first row) and with an increasing p-value threshold from SaTScan (second row). Both methods have low power in detecting these departures, showing, as expected, a limit beyond which neither of the detection methods considered here provides meaningful results. In addition, with similar FDR estimates, BaySTDetect is still marginally better than SaTScan in detecting areas with relatively large expected counts/overall disease risks.

Figure 3. Sensitivity of the two pairs of $(\pi, \text{control FDR})$ settings that yield similar empirical FDR estimates.



IV. APPLICATION: RETROSPECTIVE SURVEILLANCE OF CANCER INCIDENCE DATA

To demonstrate the “data mining” potential of the method, we apply the detection method to carry out a retrospective surveillance of time trends for several cancer types recorded by the Thames Cancer Registry (TCR). TCR collects data on newly diagnosed cases of cancer in the population of London and South East England. It is one of the largest cancer registries in Europe, covering a population of over 12 million, and holds nearly 3 million cancer registration records. In this paper, we will use data for some major cancer types, aggregated by electoral ward of residence (of which there are 1899 in the TCR region) and year, for the period 1981–2008. Age-specific population estimates for each ward and year were based on the 1991 and 2001 Census, and

were used to calculate age-adjusted expected counts for each cancer type. This application is “work in progress”, and results will be presented at the conference.

V. CONCLUSION AND DISCUSSION

Data sparsity is one important issue to be dealt with in analysing small area data. The simulation study here has provided some idea on the limits beyond which detection becomes infeasible, for example, detecting a departure with a small increase in risk of 1.5 relative to the general disease rate in an area with expected cases less than 8 per year. Data aggregation over time/space or both may help deal with this issue. We have explored the tradeoffs between sensitivity and specificity through modifying the FDR control and the prior specification on π . Our recommendation is to increase the predefined FDR control level, rather than to reduce π (equivalent to increasing the prior probability of detection) in order to gain sensitivity, since this leads to empirical false detection rates that are close to the specified control level.

In addition to the cancer surveillance application, other suitable applications of this method include monitoring early disease outbreaks, detecting elevation of crime rates and assessing impact of local implementations of a policy. As demonstrated in the case study, our proposed method also has potentially high policy relevance for national or regional chronic disease surveillance to help identify departures from common trends that may require explanation, investigation and targeted interventions.

REFERENCES

- [1] A. Lawson (2006) *Statistical Methods in Spatial Epidemiology*, John Wiley & Sons Ltd, Chichester, England.
- [2] N. Best, S. Richardson and A. Thomson (2005) *A comparison of bayesian spatial models for disease mapping*, Stat Methods Med Res 14(1), 3559.
- [3] L. Waller, B. Carlin, H. Xia and A. Gelfand (1997) *Hierarchical spatio-temporal mapping of disease rates*, Journal of the American Statistical Association 92, 607617.
- [4] L. Knorr-Held and J. Besag (1998) *Modelling risk from a disease in time and space*, Statistics in Medicine 17(18), 20452060.
- [5] L. Knorr-Held (2000) *Bayesian modelling of inseparable space-time variation in disease risk*, Statistics in Medicine 19(1718), 25552567.
- [6] Y. MacNab (2007). *Spline smoothing in bayesian disease mapping*, Environmetrics 18(7), 727744.
- [7] G. Li, N. Best, A. Hansell, I. Ahmed, and S. Richardson. *Bayesian model choice for detecting unusual temporal patterns in small area data* submitted to Biostatistics and is available from www.bias-project.org.uk/research.htm.
- [8] M. Kulldorff, R. Heffernan, J. Hartman, R. Assuncao and F. Mostashari (2005) *A space-time permutation scan statistic for disease outbreak detection*, PLoS Medicine 2(3), e59.
- [9] M. Kulldorff (2001) *Prospective time periodic geographical disease surveillance using a scan statistic*, J. of Royal Stat. Soc. (A) 164, 6172.
- [10] J. Besag, J. York and A. Mollie (1991) *Bayesian image restoration, with two applications in spatial statistics*, Annals of the Institute of Statistical Mathematics 43(1), 120.
- [11] A. Gelman (2006) *Prior distributions for variance parameters in hierarchical models*, Bayes. Anal. 1(3), 515533.

Figure 4. Sensitivity of detecting the 15 truly unusual areas with small departure ($\theta = 1.5$) when data are sparse. Sensitivity is obtained from BaySTDetect with various FDR control levels (first row) and from SaTScan with various threshold p-values (second row). For BaySTDetect, $\pi = 0.95$.

