An empirical evaluation of space-time models for surveillance of disease maps

Carmen L. Vidal Rodeiro, Andrew B. Lawson

Department of Epidemiology and Biostatistics . Norman J. Arnold School of Public Health. University of South Carolina. Columbia, SC 29208.

Introduction

Many of the methods applied in disease mapping can be applied as a surveillance tool. Disease mapping is frequently constrained to a single time period but data in public health are available for time windows of several years. New cases are added to disease registries regularly and for surveillance purposes it is important to reanalyse the data in periodic intervals. If a geographically health hazard suddenly occurs, we would like to have a surveillance system that can pick up an increase in disease risk or incidence as quickly as possible. Then, a natural extension of disease mapping is to investigate the dynamics of a disease.

Objective: Compare space-time models for counts of disease in a surveillance context via an empirical evaluation based on simulated data that represent different risk scenarios over space and time. The purpose of these analyses is to assess the ability of the models to detect changes in risk patterns over space and time.

Notation: Counts of disease in fixed spatial sites indexed as i=1,…,m and temporal periods j=1,…,T are denoted as y_{ij}; e_{ij} and θ_{ij} denote, respectively, the expected number of cases and the relative risk in the ith region in time period j.

Data and Methods

In surveillance, it is important that the models are *capable of describing the overall behaviour of the disease in space and time* and also that they will be *sensitive to changes in the spatio-temporal structure*. To this end, the models should be able to capture spatial, temporal and spatio-temporal effects and their parameterization should allow for changes in time. The use of McMC algorithms to provide samples of the posterior distributions is required. They are straightforward to implement but the sequential re-fitting of the models in a surveillance context presents a problem, as each year new data is added and new parameters included. An alternative approach to this problem, adopted also in this work, is the use of a sliding windows (3, 5 and 8 years) within which the models are fitted.

We use as our study region the U.S. state of South Carolina by county and a time period of 20 years.

Although in our work we fitted nine different space-time models in a wide range of scenarios, we show here some results in three different scenarios, which represent some possible changes in the risk patterns over space and time:

Scenario 1: a global jump in risk occurs in a year (j=10) and then disappear.

Scenario 2: local jumps in risk of different intensities are generated for different regions in South Carolina in years 10 and 12.

Scenario 3: an increase in the risk occurs in year 7 in a small area in the centre of the region (two counties) and it spreads to neighbouring counties in the following years. This effect stops in year 11.

The models used to fit the data assume the following expressions for the relative risk:

Model 1

$$\log\theta_{ij} \sim normal\left(\log\theta_{ij-1}, \sigma_\theta^2\right)$$

where θ_{ij-1} is the relative risk in the ith region in time period j-1.

Model 2

$$\log\theta_{ij} = u_i + v_i + t_j$$

where u_i is a spatially correlated random effect, v_i is an uncorrelated spatial random effect and t_j is a temporal random effect.

Monitoring the changes

In order to monitor the process and detect the changes in risk patterns, several measures can be used, in particular:

• surveillance residuals: difference between the observed data for a new year and the data expected under a model when it is fitted for previous years. Draws of the posterior predictive distribution are used to represent what we can expect under the model. In the case of count data they can be approximated by

$$r_{ij} = y_{ij} - \frac{1}{G}\sum_{g=1}^{G} e_{ij}\hat{\theta}_{ij-1}^{(g)} \qquad i=1,\ldots,m \qquad j=2,\ldots,T$$

where $\left\{\hat{\theta}_{ij-1}^{(g)}\right\}$ is a set of parameter values sampled from the posterior distribution.

• p-value surface for the ranking of the ordinary Bayesian residuals computed via a parametric bootstrap.

• changes in parameter values: they can indicate, for example, changes in the variability across the map, changes in the global spatial correlation structure, jumps in risk that occur in time or changes in risk at particular space-time locations. Here we examine changes in the variance of the random effects by computing the ratios between their values for two successive time periods.

• probability of θ_{ij} being bigger than 1: This probability tells us if there are extreme values in the relative risk (in those cases the value of p_{ij} would be very close to 1; 0.95 could be use as a cut-off point).

$$p_{ij} = prob\left(\theta_{ij} > 1\right) = \frac{\#\left\{\hat{\theta}_{ij}^{(g)} > 1;\ g=1,\ldots,G\right\}}{G}$$

• Moran's I statistic: measures the degree of residual autocorrelation. It usually ranges between -1 and 1, with large positive values indicating neighbourhood similarity of the residuals and values close to zero indicating absence of spatial autocorrelation.

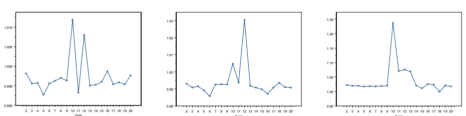
Results

Due to lack of space we are showing only an example of the ample range of results that were obtained in this study.

DIC results (Deviance Information Criterion) show that there are significant increases (that suggests worse fit) in the years where changes in risk were generated. Moving windows models give a worse fit than the year-by-year model in the left side of the time period (boundary effects). The residual sum of squares (RSS) displays high values also for those years. This measure of goodness of fit also shows that model 2 fits better the data in all scenarios; for this reason, we are showing here results for model 2.

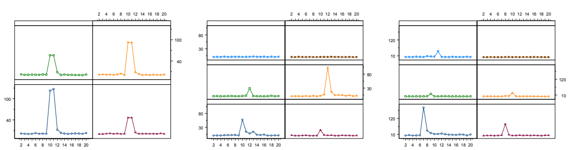
The following pictures show the empirical ratios of the variance of the random effects for two successive years (σ_u^2, σ_v^2 and σ_t^2) for model 2 in scenario 2.

For the random effect u, there are changes in years 10 and 12. For any other year the value for their ratio is close to 1. When we are using sliding windows we lose this information as the values for the ratios highly vary over the years. The variance of the random effect v has the same behaviour. Finally there are increases in the variance of the random effect t in years 10 and 12.



σ_u^2 σ_v^2 σ_t^2

In order to study the pointwise surveillance residuals, we randomly selected counties which are located in the regions where the jumps in risk are located. Also, residuals for counties that are located in regions where the risk does not suffer variations are shown (for scenario 2, the two pictures on the top and for scenario 3 the picture on the right top corner).



Scenario 1 Scenario 2 Scenario 3

From these pictures it is clear that something unusual happened in some counties in certain years of the study period because the surveillance residuals are bigger than for other years and therefore, for those years, the data is not representative of what is expected under the model.

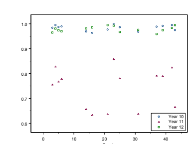
Moran's I statistic for the residuals lies between 0.1 and 0.5 for all years but for the years where changes in risk take place. For them, the value of the Moran's I is positive and around 1, which indicates the presence of spatial autocorrelation.

The following maps display the p-value surface for year 10 and year 12 in the scenario 2 fitted using model 2. Unusual values appear for some counties in years 10 and 12. These counties correspond with those where we generated the change in risk. So, for them, the data are not representative of what we can expect under the model. In the counties where no changes in risk where generated, the p-values, as a measure of discrepancy between the assumed model and the current data, show that the model produces a good fit.

p-value surface for year 10. Model 2 (fitted year by year), scenario 2.

p-value surface for year 12. Model 2 (fitted year by year), scenario 2.

In the second scenario, this picture shows, for randomly selected counties where jumps in risk where generated, the probability of θ_{ij} being bigger than 1 for years 10, 11, and 12. It is clear than in years 10 and 12 the probability is bigger than 0.95, showing that there are extreme values of the relative risk.



Conclusions

Statistical approaches to the continuous monitoring of the incidence of diseases in a population represent an important challenge to public health surveillance.

As is clear from some of the results, while is possible to employ many methods already developed outwith the surveillance context, there is the need to develop special methods which are more sensitive to the sequential nature of the surveillance task.