Spatial modelling of Lupus incidence over 40 years with changes in census areas

Ye Li, Patrick Brown, Håvard Rue, Mustafa al-Maini and Paul Fortin

February 12, 2011

Abstract

Clinical data on the location of residence at the time of diagnosis of new Lupus cases in Toronto, Canada, for the 40 years to 2007 are modelled with the aim of finding areas of abnormally high risk. Inference is complicated by numerous irregular changes in the census regions on which population is reported. A model is introduced consisting of a continuous random spatial surface and fixed effects for time and ages of individuals. The process is modelled on a fine grid and Bayesian inference performed using Integrated Nested Laplace Approximations. Predicted risk surfaces and posterior exceedance probabilities are produced for Lupus and, for comparison, Psoratic Arthritis data from the same clinic. Simulations studies are also carried out to better understand the performance of the proposed model as well as to compare with existing methods.

Keywords: integrated nested Laplace approximation; changing boundaries; Bayesian inference; disease mapping

1 Introduction

1.1 Lupus and Psoratic Arthritis

Lupus is an autoimmune disease characterized by acute and chronic inflammation of various tissues of the body including the skin. It can be limited to the skin or can involve several organ systems. When internal organs are involved, the condition is referred to as SLE. This is an uncommon but incurable disease with a prevalence

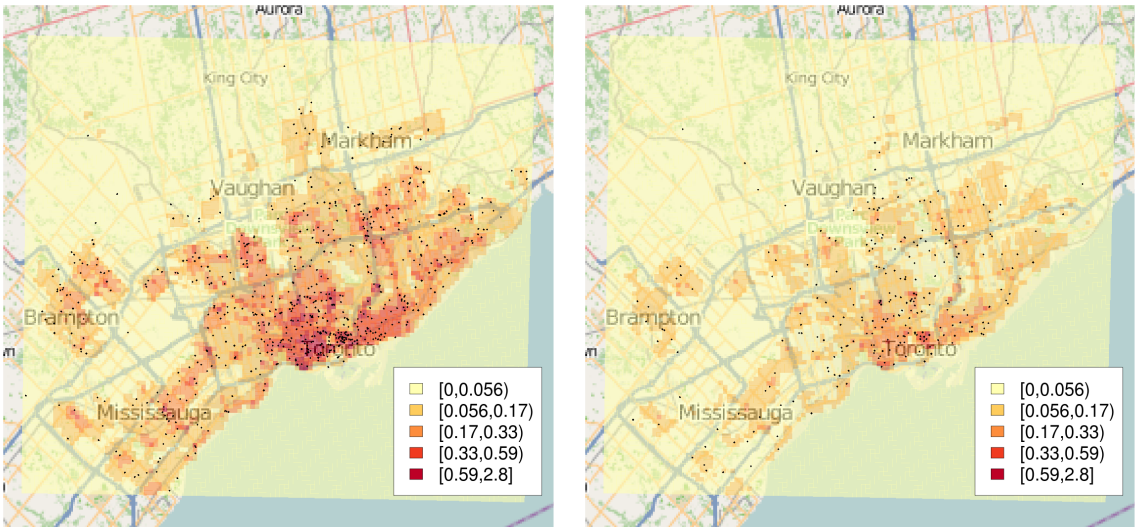
estimated at 1:1000, and more than 90% of the cases are female (Simard & Costenbader, 2007; Bernatsky, 2007). The etiology remains unclear but there is mounting evidence that both genetic predispositions and environmental exposures are important in the etiopathogenesis (Cooper et al., 2002). Psoriasis is a skin disease that presents with an erythematosquamous rash that occurs most frequently on the elbows, knees and scalp, but can cover much of the body. When arthritis occurs concomitantly with psoriasis, it is called Psoriatic arthritis (PsA). PsA affects women and men equally with an incidence of approximately 6 per 100,000 per year and a prevalence of about one to two per 1,000 (Brockbank & Gladman, 2002). Genetic predispositions are thought to contribute most to psoriasis and PsA and, contrary to SLE, there are fewer studies addressing the environmental risk factors that may contribute to PsA (Alamanos et al., 2008; Pattison et al., 2008).

The aim of this paper is to make inferences about the spatial distribution of SLE in Toronto, Canada, with the objective of identifying areas of elevated incidence rates which might be indicative of an underlying environmental risk factor. PsA is used as a comparison group, which might be expected to share reporting biases and structural patterns with SLE though should not exhibit any spatial dependence caused by environmental risk factors. To accumulate an adequate number of cases, data amassed over a period of 40 years is considered. During this time the population of Toronto has increased, with growth more pronounced in suburban areas, and the boundaries of regions on which population figures are reported have changed with every census. The main statistical challenge, therefore, is to address the problem spatial modelling with changing census boundaries.

SLE and PsA cases in Toronto are referred to a single specialized clinic, the Centre for Prognostic Studies in Rheumatic Diseases, with date of and residence at diagnosis being recorded. There were 875 lupus cases referred from 1970 to 2006, with 88% being female, and 527 PsA cases between 1978 and 2007, with 41% female. Details regarding data collection and processing are described by al Maini (2008). The Census of Canada provides population data and digitized boundaries of reporting regions for the years 1971, 1981, and thereafter 5-yearly until 2006. The population (rounded to the nearest 5) by five year age and sex group is provided for each census region. The smallest census region for which data are available is the Dissemination Area (or Enumeration Area before 1996), which contain on the order of 400 individuals. The 1986 and earlier censuses have digitized boundaries only for the larger Census Tracts

containing roughly 20000 individuals.

Figure 1 shows expected incidence rate (as described in section 2.3) over the whole study period, in cases per km^2, with case locations superimposed as dots. A small amount of random noise has been added to the locations of the cases . The figures shows the two most common sex and disease combinations, Female Lupus and Male PsA and as expected, cases are more concentrated in the higly populated areas. Additional graphs for female lupus and male PsA are shown in Appendix B.



(a) Female lupus

(b) Male psoriatic arthritis

Figure 1: Expected intensity of cases per km^2 (calculated from population data and incidence rates) and observed case locations, for the period 1965-2007 in Toronto, Canada

1.2 Spatial methods for disease incidence data

The most common approach for disease mapping is to model the case counts for a set of non-overlapping subregions as Poisson distributed conditional on a normally-distributed sub-region-level random effect. Spatial dependence is induced by having each sub-region's random effect depending on its neighbours using a Markov Random Field model (see Lawson, 2008). The neighbourhood structure makes statistical inference computationally feasible even when the number of sub-regions is large. The Besag-York-Mollié model is the most commonly used, having the risk of a sub-region depending on its immediate neighbours with equal weight (Besag et al., 1991).

When location data, as opposed to spatially aggregated case counts, are available, spatial point process methodology is often applied (see Diggle, 2003). The Log-Gaussian Cox Processes (LGCP) is a useful and popular model for location data which are distributed inhomogeneously due to both deterministic effects (such as variations in population) and stochastic, possibly environmental, heterogeneity. LGCP's are equivalent to the Poisson-Markov Random Fields described above if population and risk are assumed to be constant within sub-regions.

Second order properties of disease risk can be explored with the use K-functions (see Diggle, 2003, ch. 4) for point process data, and for aggregated data with spatial variograms and tests such as Moran's I (Cliff & Ord, 1981). Such methods are useful for evaluating a null hypothesis of spatial independence. For predicting and making inference about the spatial distribution of risk (first order properties), the non-Gaussian nature of case counts suggest Bayesian inference based on Markov chain Monte Carlo (MCMC) algorithms. Rue et al. (2009) introduce Integrated Nested Laplace Approximations (INLA), which is a less computationally intensive and potentially more robust alternative to MCMC.

2 Models and Methods

Separate models are fit for Lupus and PsA, and for each disease separate models are fit for males and females. An alternative would be to assume the risk surfaces were identical for each sex and to perform a combined analysis. Separate analyses were undertaken as this assumption was deemed to be unreasonable due to possible differences in risk factors for males and females for the diseases in question.

2.1 Model

Let (S_{jk}, T_{jk}) be the location in space and time for case k in the jth age group, and $P_j(s,t)$ and $\Lambda_j(s,t)$ be the population and risk surface respectively for location s and time t. The locations and times are realisations from a spatio-temporal inhomogeneous

Poisson process with

$$\{T_{jk}, S_{jk}|R(s); k = 1...K_j\} \sim \text{Poisson Process}[\Lambda_j(s,t)P_j(s,t)]$$
$$\Lambda_j(s,t) = \Gamma(t)\Theta_j R(s)$$

Here $\Gamma(t)$ represents variations in risk over time, Θ_j is the age group effect and $R(s)$ is the random spatial risk surface. $R(s)$ is modelled using the Log Gaussian Cox Process as:

$$\log(R(s)) = \mu + U(s)$$
$$\text{Cov}[U(s), U(s+h)] = \sigma^2 \text{Matérn}(|h|/\phi, \nu).$$

where ϕ, ν are the range and roughness parameters respectively and h is the distance between two points on the plain. An alternative would be to use a shot-noise Cox process type model where R(s) would be the sum of basis functions with random centres. The principle reason for the choice of the log-Gaussian process is because of its being the natural extension of the Besag-York-Mollié model commonly used for disease mapping problems with aggregated data. Further, the latent Gaussian assumption allows for the use of the Integrated Nested Laplace approximation technique, greatly reducing computational requirements in comparison with MCMC algorithms.

2.2 Inference

2.2.1 Inference on age and time effects

We make a number of simplifying assumptions in order to carry out inference on the parameters and spatial random effect. Population is taken to be constant between the midpoints of census years (5 year intervals after 1981), and uniformly distributed within census regions (CT's or DA's). Let C_i denote the ith census interval and write $P_j(s,t) = P_{ij}(s); t \in C_i$, the population at location s during the ith census. As shown in Appendix A.1, inference on the spatial surface $R(s)$ depends only on the integrated

time effects. Hence we define Γ_i as

$$\Gamma_i = \int_{C_i} \Gamma(t)dt$$

As a result, we estimate only Γ_i rather than the full time trend $\Gamma(t)$. The case counts by census period and age group $M_{ij} = ||k; C_{i-1} <= T_{jk} < C_i||$ are sufficient for making inference on the Γ_i and age effects Θ_j as shown in Appendix A.2. More specifically, $[\Theta_j, \Gamma_i|S_{jk}, T_{jk}, R(\cdot)] \propto [\Theta_j, \Gamma_i|M_{ij}, R(\cdot)]$ and estimation of these parameters is accomplished using

$$\begin{aligned} M_{ij}|R(\cdot) &\sim \text{Poisson}\left(\int_{C_i}\int_{\Omega} \Lambda_j(s,t)P_j(s,t)dsdt\right) \\ &\sim \text{Poisson}\left(\Gamma_i\Theta_j \int_{\Omega} P_{ij}(s)R(s)ds\right) \end{aligned}$$

where Ω is the extent of the study region.

2.2.2 Inference on the Spatial effects

The next assumption made is to assume that $R(s)$ is constant within regular lattice of squared grid cells $G_\ell, \ell = 1 \dots N$ covering the study region Ω, with $R(s) = R_\ell$ when $s \in G_\ell$. This serves two purposes. First, it allows the Matérn correlation of the $U(s)$ to be approximated by a Markov random field where each grid cell depends only an a small number of nearby cells (Lindgren et al., 2011; Lindgren & Rue, 2007). This provides an analytical formula for the inverse of the variance matrix, which is sparse, and results in the computations being feasible for even fine grids with many thousands of cells. Second, evaluating the risk surface on grid cells rather than the census regions provides geographic boundaries which are constant over time, and having the risk surface constant within cells allows the problem to be reduced to a Generalized Linear Mixed Model with Poisson distributed cell counts.

Conditioning on Θ_j and Γ_i, the distribution of R_ℓ depends only the total case count Y_ℓ for that cell, with Appendix A.3 showing that $[R_\ell|\Theta_j, \Gamma(t), S_{jk}, T_{jk}] = [R_\ell|\Theta_j, \Gamma_i, Y_\ell]$. Integrating under the intensity surface and summing over age groups gives the cell counts Y_ℓ being Poisson distributed:

$$
\begin{aligned}
Y_\ell &\sim \text{Poisson}(O_\ell R_\ell) \\
O_\ell &= \sum_{ij} \Gamma_i \Theta_j P_{ij\ell} \\
\log(R_\ell) &= \mu + U_\ell \\
\text{Cov}(U_\ell, U_m) &= \sigma^2 \text{Matérn}(|G_\ell|/\phi; \nu)
\end{aligned} \tag{1}
$$

The O_ℓ is an offset parameter, interpretable as the expected number of cases when $R_\ell = 1$. $|G_\ell|$ is size of the grid cell, which is the same as the distance between centroids of two grid cells. The $P_{ij\ell}$ are the total populations in each grid cell ℓ by census period i and age group j, formally defined as $P_{ij\ell} = \int_{G_\ell} P_{ij}(s)ds$. With the assumption that populations are uniformly distributed within census regions, the population of each region (as given by the census) is allocated to grid cells with which it intersects according to the proportion of the area of the region contained within each cell.

This model is a fairly standard Generalized Linear Mixed Model, with U_ℓ approximated by a Gaussian Markov Random Field. The number of neighbours which need conditioning on depends on the roughness parameter ν and the weights for each neighbour are a deterministic function of the range parameter ϕ.

2.3 Implementation

Although it would be possible to use the preceding results in a fully Bayesian Gibbs sampling algorithm, we instead use a two stage process which ignores uncertainty in the Θ_j and Γ_i but greatly reduces the dimensionality of the problem. First, the Maximum Likelihood Estimates for Θ_j and Γ_i are estimated with the spatial random effects U_ℓ set to zero i.e. using $M_{ij} \sim \text{Poisson}(\exp(\mu)\Theta_j\Gamma_i\tilde{P}_{ij})$,where $\tilde{P}_{ij}$ is the total population of group j in the ith census. Then treating the $\hat{\Theta}_j$ and $\hat{\Gamma}_i$ as fixed known values, the offset parameters for each grid cell $O_\ell = \sum_{ij} \hat{\Gamma}_i\hat{\Theta}_j P_{ij\ell}$ are computed (and are mapped in Figure 1). Inference on the spatial surface U_ℓ, its variance σ^2 and range ϕ, and the intercept(mean) parameter μ are made using the purely spatial Generalized Linear Mixed Model with spatial random effects as given in (1).

Ignoring uncertainty in the age effects Θ_j, and ignoring spatial dependence in estimating them, is fairly common in purely spatial models of aggregated data (see Waller & Gotway, 2004). The data are sufficiently informative about the age and time effects

that the parameters are estimated with sufficient precision that their standard errors are negligible, though the standard errors are admittedly underestimated since they ignore spatial dependence. Also note that the intercept parameter μ is not treated as fixed, and in effect the age and time parameters are only assumed known up to a constant of proportionality.

Even with the Markov random field approximation, computational limitations require the grid cells to number at most 20,000 to 25,000. As a result, a roughly 58km by 53km central section of the Greater Toronto Area was covered with a grid of 115 by 104 cells spaced 500m apart ($|G_\ell| = 0.5$km). This was a compromise between creating a region as large as possible to capture the greatest number of cases yet still using a fairly fine grid in order minimize the effects of the assumption that risk is constant within grid cells. Note that the age and time parameters Γ and Θ were estimated from the full dataset, not only the data in the rectangular region.

The integrals of the populations over grid cells to compute the $P_{ij\ell}$ are the most computationally demanding part of the algorithm, though efficient routines have been developed by the Geographic Information Systems research community and are available in R through the "raster" package. Rasterizing produces a pixel image from a set of polygons by taking the average of the values of the surface in each polygon intersecting a given grid cell, weighted by the area of overlap. The offsets O_ℓ are produced by calculating the offsets at the census region level for each census year, rasterizing and adding the rasterized images over census periods pixel by pixel.

2.4 Prior Distributions

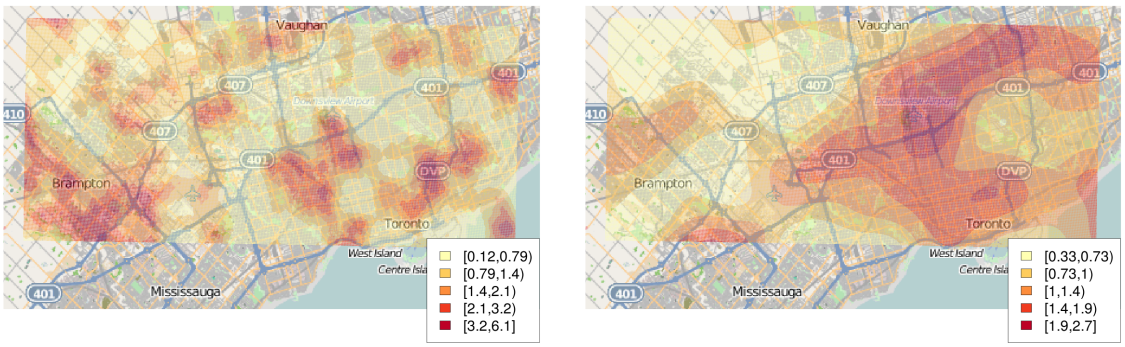
The roughness parameter ν and range parameter ϕ are typically not well identified in spatial models (see Clark & Gelfand, 2006). Stein (1999) shows that the predicted spatial surface depends on the product $\phi\sigma^2$ but is insensitive to changes in the individual parameters. As a result, we fix the roughness parameter at 2 and put a informative prior on the range.

In consultation with collaborating clinicians, we determined that the range of spatial dependence is most likely to be a small number of kilometres. A 95% prior interval between 400m and 4km was chosen, resulting in a Gamma prior with mean 2km and shape parameter 6.5. The precision $1/\sigma^2$ has a weaker prior of Gamma with shape 1 and scale 0.01, giving a prior mean for σ of 0.17 and 95% interval (0.052, 0.63). The upper 97.5% quantile for σ would give 95% of the U_ℓ between -1.2 and 1.2 on the log

scale, or relative risks between 0.3 and 3.3 on the natural scale. Identical priors were used for all disease and sex group combinations. In order to test the sensitivity of the result to the priors, $Gamma(1, 5 \cdot 10^{-5})$ priors for ϕ and σ are also tried and results are compared.

3 Simulation Study

A simulation study was used to evaluate the proposed methodology and compare it to the more established Besag, York and Mollié model. To this end, spatial random fields were simulated for a number of different spatial range parameters and the resulting estimated risk surfaces were compared to the true values. As the BYM model does not allow for changes in spatial boundaries, only the 2006 census boundaries and populations were used. Published incidence rates for lung cancer and stomach cancer were used to simulate case locations, since lupus rates are too small to simulate cases in a single census period and choosing a constant to scale rates by is arguably more arbitrary than selecting more prevalent disease rates. Stomach cancer rates yielded on average 320 cases, the same order of magnitude as the 716 and 277 cases observed for female lupus and male PsA respectively. Lung cancer is much more prevalent, yielding roughly 2,200 cases in each simulation, and was included to contrast the methods' performance in high and low sample sizes. Forty Gaussian random fields with mean 0, variance 0.5 and Matérn correlation with roughness 2 were simulated over central Toronto with 4 ranges between $1km$ and $4km$. The target area of the simulation is smaller than the actual study area in order to save computational time. Figure 2 gives two simulated risk surfaces with ranges 1km and 4km, showing the simulations include both rough and smooth surfaces.



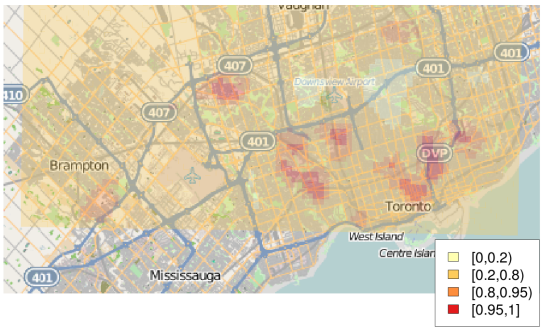
(a) Range 1km

(b) Range 4km

Figure 2: Simulated log-normal risk surfaces with mean 0 and variance 0.5 for two different range parameters.

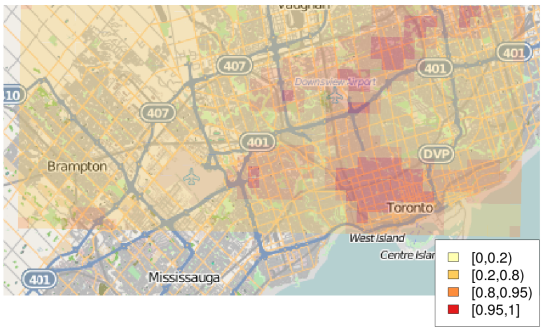
Risk surfaces are estimated using the proposed model with grid cell of different sizes($|G_\ell| = 200m,\ 500m,\ 1km$ and $1.5km$) as well as with the BYM model. Priors on ϕ and τ are Gamma$(1, 5 \cdot 10^{-5})$ and priors on variances of BYM model are LogGamma$(0.1, 0.1)$.

In disease mapping problems, the question of interest is often to identify regions of abnormally high risk. We use C to denote the criteria for 'abnormal' with, for example, $C = 1$ and $C = 2$ meaning that a region would be considered high risk if its risk were above or twice above the average risk. Figure 3 shows the probability that the risk is above the average or $pr(R(s) > 1|Y)$ for the BYM model and the grid approximation model with $|G_\ell| = 0.5km$ & $1km$ for the two simulated surface shown in Figure 2 respectively. The grid approximation assigns higher posterior probabilities to many of the regions with relative risk above 1 in comparison with the BYM model. This difference is equally apparent with the rough and smooth surfaces. The results for larger C values and the more prevalent lung cancer display a more noticeable improvement of the grid approximation over BYM (not shown).



(a) Range 1km, Grid approximation (0.5km)

(b) Range 1km, BYM Model



(c) Range 4km, Grid approximation (1km)

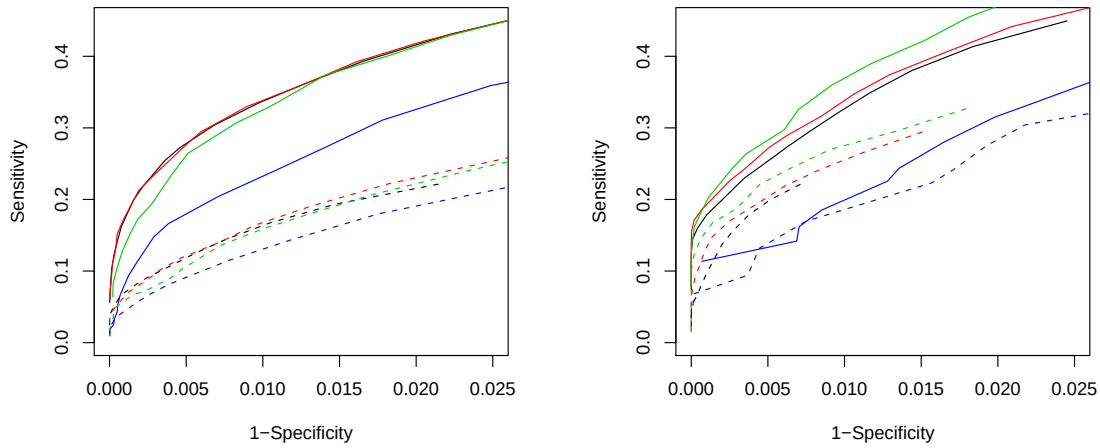
(d) Range 4km, BYM Model

Figure 3: Posterior probabilities of risk being above average $pr(R(s) > 1|\text{data})$ for simulated stomach cancer data. Simulated spatial surfaces are shown in Figure 2. Inference was performed using the grid approximation model and the Besag, York and Mollié (BYM) model.

To quantify the performance of BYM and the proposed model, sensitivity and specificity are calculated, let Ω be extent of the study region:

$$Sens = \frac{\int_\Omega I(pr(\hat{R}(s) > C) > \alpha) I(R(s) > C) ds}{\int_\Omega I(R(s) > C) ds} \qquad Spec = \frac{\int_\Omega I(pr(\hat{R}(s) > C) < \alpha) I(R(s) < C) ds}{\int_\Omega I(R(s) < C) ds}$$

Here $pr(R(s) > C)$ is a $100C\% - 1$ exceedance probability and α is a probability threshold. For example $C = 2$ and $\alpha = 0.95$ seeks to identify regions with 100% higher risk (or twice of the risk) than average and would identify a region as such if it's posterior probability of meeting this criteria were above 95%. ROC curves are produced for different C values through a range of α.



(a) Range 1km

(b) Range 4km

Figure 4: ROC curves computed from simulated data using grid approximations with cell sizes ($|G_\ell|$) of 200m (—),500m (—) and 1km (—) as well as the BYM model (—) using census tract geographies. Lung cancer (—) and stomach cancer (- - -) are shown for 100% exceedance probabilities $pr(R(s) > 2)$.

Figure 4 shows the Receiver Operating Characteristic (ROC) curves for detecting regions above exceedance thresholds of $R(s) = 2$, with the curves obtained by varying the posterior probability required for defining a region as being above a threshold. For instance, comparing a 0.8 probability criteria to a 0.5 probability criteria, or $pr(R(s) > 2|Y) > 0.8$ vs $pr(R(s) > 2|Y) > 0.5$, would result in the 0.5 having higher sensitivity for detecting parts of the surface where the simulated $R(s)$ was above two. However, the 0.8 criteria would have greater specificity, having a smaller proportion of the surface where $R(s) < 2$ incorrectly labelled as being above two. Varying the probability criteria between 0.1 and 0.99 for each fitted model and each threshold, and comparing the proportions of the regions correctly and incorrectly identified as being above the threshold, produces a range of sensitivity and specificity values which are connected as ROC curves.

The grid approximation consistently outperforms the BYM model, with the ROC curves closer to the desirable values in the top left, and with the difference being more pronounced with the rough surface (range 1km). The grid approximation appears to be robust to the choice of grid size with larger grid size performing better on a smoother surface (range 4km). Using grid cells much larger than the spatial range produced less

favourable results (not shown), which would be expected as the true surface would likely vary within grid cells under such circumstances.

To quantify the method's ability of making predictions of the risk surface, the Integrated Mean Square Error is computed for both the BYM and grid approximation models. Figure 5 shows MSE of BYM model and the grid approximation model ($|G_\ell| = 0.5km$) as a function of the spatial range parameter, with the grid model consistently having lower MSE than the BYM model and the difference getting smaller as the true surface gets smoother. The approximations on different grid sizes have similar MSE and they all outperform the BYM model (not shown). Figures in Appendix C show predicted risk surfaces corresponding to Figure 3.

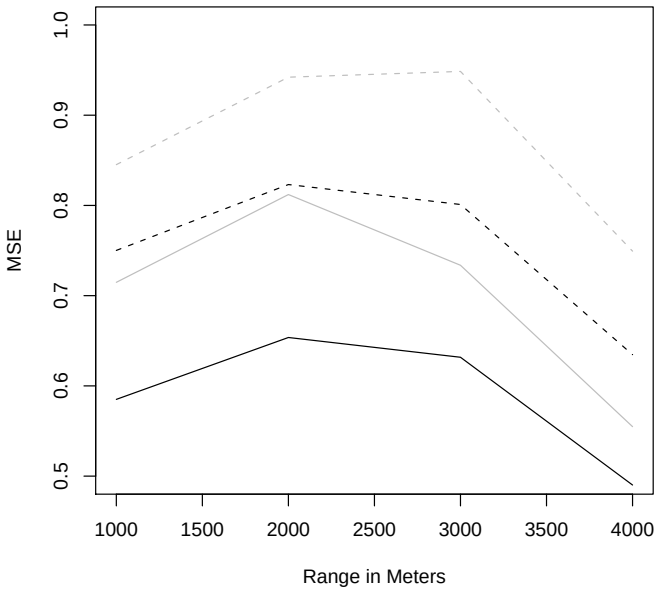


Figure 5: Mean Square Error for the grid approximation model with cell size 0.5km (—) and BYM model (—) for lung (—) cancer and stomach (- - -) cancer, as a function of the spatial range parameter.

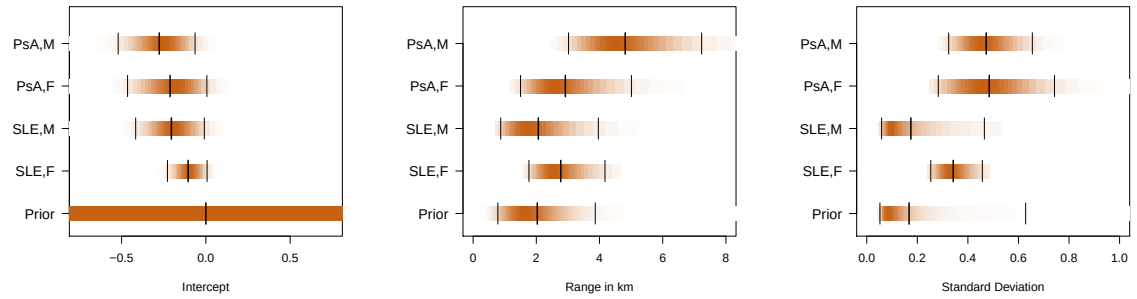
4 Results

Table 1 shows the total number of cases and the number of cases in the rectangular region used for the spatial model, for each of the four disease and sex combinations.

Diseases	Total Cases	Cases used in model
Female Lupus	767	716
Male Lupus	108	106
Female PsA	216	156
Male PsA	311	277

Table 1: Number of cases for each sex and disease combination, both total numbers and numbers in the rectangular region on which the spatial model was fit

Figure 6 shows the prior and posterior densities of the intercept (μ), range (ϕ) and standard deviation (σ) for each outcome as specified in Model 1, where vertical bars locate the mean and 95% credible intervals. Female Lupus is the most prevalent outcomes and correspondingly has posterior distributions which are tighter than male lupus and PsA. As expected, the data are for the most part uninformative regarding the range parameter with the posteriors in Figure 6c being similar to the prior. Female lupus and male PsA, being the two most common outcomes, show posteriors with some significant departure from the prior with male lupus being indistinguishable from the prior.



(a) Intercept μ (b) Range ϕ (c) SD σ

Figure 6: Prior and Posterior distributions of model parameters for each disease outcome: Male psoriatic arthritis, Female psoriatic arthritis, Male lupus, and Female lupus. Shown are means (|), 95% intervals (|), and density strips ( ).

Figure 7 shows the predicted values of the risk $R(s)$, or $\mathrm{E}(R(s)|Y)$, for each disease and gender on a common scale. Male lupus and female PsA show flat risk surfaces,

with the more prevalent female lupus and male PsA show more spatial variation in risk. As the precision with which risk is estimated varies throughout the study region, regions with high predicted risk are not necessarily evidence of a "hot spot" or cluster. Rather, this could merely reflect a region where risk is poorly estimated due to low population. Figure 8 reflects the uncertainty in estimation by plotting 20% exceedance probabilities, or $\text{pr}(R(s) > 1.2|S,T)$. Using 20% as a rough figure for the excess risk that would be expected in the presence of an environmental risk factor, regions with an exceedance probability above, say, 80% or 95% would be indicative or warranting further investigation.

(a) Female Lupus

(b) Male Lupus

(c) Female PsA

(d) Male PsA

Figure 7: Posterior means of the relative risk surface $E(R(s)|\text{data})$, using separate analyses for each of the disease and sex combinations

A sizable area in the bottom right corner of Figure 8A shows strong evidence of

elevated female Lupus risk. In the centre of this area is the former Toronto Wellesley Hospital, where the Lupus clinic was located until 1997. Other areas of high predicted risk, including Markham, do not prove to be significant. A significant male PsA cluster exists to the north of the clinic, with further modest evidence of clustering to the north. Male Lupus appears to have relative risk below 1.2 throughout the region, whereas for female PsA the results are inconclusive. This is consistent with the upper 97.5% posterior quantile of the intercept and standard deviation being lower for male Lupus than female PsA.

(a) Female Lupus

(b) Male Lupus

(c) Female PsA

(d) Male PsA

Figure 8: Posterior probability of relative risks being more than 20% above average, $pr[R(s) > 1.2|\text{data}]$, for each of the disease and sex combinations

The priors used in this analysis are reasonably informative, and alternative priors (Gamma with shape 1 and scale 0.00005) were subsequently tried to assess the sensitiv-

ity of these results to the prior specification. As female PsA and male Lupus are much sparser datasets, these priors with smaller shape parameters understandably yielded very wide posterior distributions and flat risk surfaces. Male PsA and female lupus gave similar parameter posterior distributions and exceedance probability plots with the alternate priors, see Appendix D.

5 Model Diagnostics and Reporting Bias

Further analyses were undertaken to explore both possible reporting bias and the assumption that the spatial risk surface is constant over time. The area of significantly elevated Lupus risk in Figure 8a is centred around the former Toronto Wellesley Hospital, which until 1997 was the site of the Lupus clinic where the data were collected. As discussed in Section 6, proximity to the clinic is not the only unique feature of this neighbourhood, though were reporting bias responsible for the apparent elevated risk it would be expected that risk would fall following the clinic's move in 1997.

Define lupus hotspot as $H = \{s; pr(R(s) > 1.2|S,T) > 0.95\}$, or the red region in Figure 8a. The following model was fit to the female Lupus data, where the spatial variation in risk depends only on whether or not a location s is inside H:

$$\begin{aligned} \{T_{jk}, S_{jk}|R(s); k = 1...K_j\} &\sim \text{Poisson Process}[\Lambda_j(s,t)P_j(s,t)] \\ log(\Lambda_j(s,t)) &= \psi(t) + \vartheta_j + [\alpha + \beta max(0, t - 1997)]I(s \in H) \end{aligned} \tag{2}$$

where α is the elevated risk in H prior to 1997 and β is the change in risk in H every year after 1997. Assuming the risk is constant within each year and age group, and population is uniformly distributed for census regions crossed by the boundaries of H, the total case count for each year-age-hotspot group is the sufficient statistics for the model parameters.

The estimated model parameters are shown in Table 2, and the results are displayed graphically in Figure 9. The observed count within the hotspot is plotted for each year, along with the predicted or expected counts under four scenarios. Using the model in 2, the fitted model parameters and the population within the hotspot are used to calculate expected counts for each year. In addition, a null model where the reporting bias parameter β is set to zero is fit and the expected counts shown. Finally, the

predicted risk from the full spatial model is multiplied by the model offsets to produce expected counts.

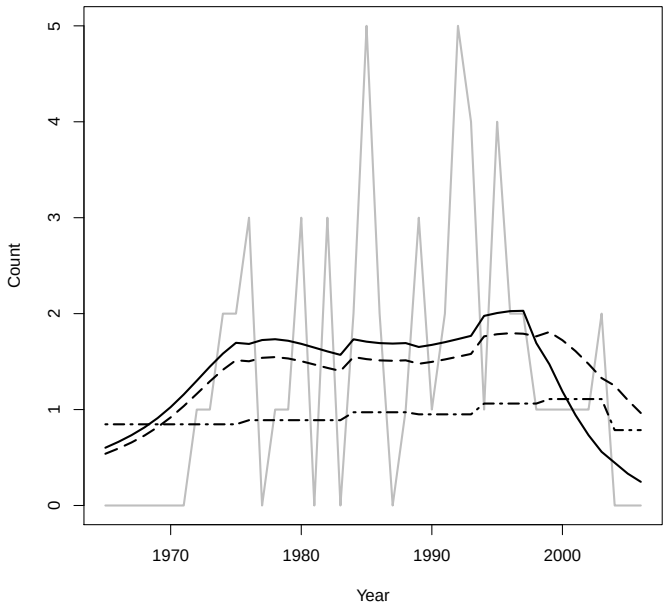


Figure 9: Observed yearly case count in H (—), predicted count from the full spatial model (– · –), predicted count from (2) (—) and predicted count from (2) when β=0, (– – –).

Notice that the expected and observed number of cases fall dramatically near the end of the followup period, due to the time lag between lupus cases being identified and patients being present at the specialised lupus clinic. The ability to estimate the relative risk at the end of the study period is likely to be adversely affected by this low expected and observed case count. For the full spatial model, the time trend $\phi(t)$ is only estimated for five year intervals, and the plot is slightly misleading as a 5 year average rate is shown for this model and a yearly rate is shown for the others. Also, the spatial smoothness of this model has shrunk this risk in comparison with the other models which allow for sharp changes in risk at the borders of H. The flatness of the expected count can also be explained by the population within this hotspot being fairly stable over time, with Toronto's population growth since 1965 occurring primarily outside the city centre.

Variable	Estimate	SE	P-value
α	1.52	0.15	< 0.001
β	-0.17	0.09	0.062

Table 2: Estimate from the model in (2), showing that the elevated risk prior to 1997 (α) is highly significant and the decreasing risk after the clinic moved in 1997 (β) is not significant

Comparing the model allowing for reporting bias ($\hat{\beta} = -0.17$) with the null model $\beta = 0$, the p-value of 0.062 for $\hat{\beta}$ is marginally significant and the visual fit for the model incorporating reporting bias is arguably better. In regards to the aim of identifying reporting bias, the results should be interpreted as inconclusive and additional years worth of data should better establish whether risk has fallen subsequent to the clinic's move. In terms of model validation, the assumption of the spatial risk surface being constant over time would be refuted by either a significant value of $\hat{\beta}$, or some evidence of dependence in the counts with respect to the expected counts of the null model. There is no obvious dependence in the counts, and a generalized additive model allowing for the relative risk in H to vary smoothly over time did not produce a statistically significant result (not shown). Model diagnostics are inhibited by low counts, for example fitting the spatial model separately for different time periods would yield few cases and low power, though the results presented in this section do not produce strong evidence against the model assumptions.

6 Discussion

The primary goal of this work was to identify areas of Toronto where spatially varying social or environmental factors could be causing higher incidence of Lupus than would be expected given the population. The conclusion that must be drawn from this analysis is that the area in the vicinity of the Lupus clinic is the region which fits this description. Reporting bias has not been conclusively demonstrated to be the source of this excess risk, a model testing for reduction of risk following the clinic's move out of the area yielded a suggestive yet not definitive p-value of 0.06. The clinic's male PsA data do not suggest reporting bias in PsA case ascertainment, yet the data are sufficiently informative to detect excess risk of PsA in a different region of the

city. The presence of the Wellesley hospital is not the only unique feature of this neighbourhood, for example Cusimano et al. (2010) have found this area to have the highest violent crime incidence in the city. It is possible that the characteristics of this neighbourhood have affected not only lupus incidence but also the decision to locate the hospital and clinic there.

Spatially referenced covariates could have been included to attempt to explain the spatial variation. Average income, proportion of residents from various ethnic groups, and distance from the clinic would be straightforward to gather or calculate. However, a spatial analysis of this sort is not necessarily the optimal method for investigating the hypotheses generated here. Returning to the paper medical records would yield information on the education and ethnicity of cases and severity at time of diagnosis, which would yield more information on social, genetic, and reporting effects respectively than using spatial covariates as a proxy for individual effects.

Allowing changes in the risk surface over time will be an important future extension to this work. A spatio-temporal Gaussian random field $U(s,t)$ in place of the current purely spatial $U(s)$ would accomplish this. However, the relatively small number of cases may not be sufficient to identify these changes over time. Exploratory analyses were attempted using parametric basis function centred on the Wellesley hospital with the coefficients on these basis functions changing over time. Lags, sometimes years, between diagnosis and referral to the clinic should result in a gradual rather than abrupt change in reporting bias following the clinic's move in 1997. These results were inconsistent, with negative reporting bias at some ranges and positive at others, likely because radially symmetric basis functions are not sufficiently flexible to capture the intricacies of the data. Fully non-parametric estimation or a spatio-temporal random field would be an improvement.

The use of Integrated Nested Laplace Approximations for inference has resulted in fast and efficient estimation with the chain convergence problems often encountered with Markov Chain Monte Carlo algorithms. Approximating the continuous surface on a regular grid has also been essential for reducing the computational burden of the algorithm, and the simulation study has shown that this grid approximation outperforms the more common BYM model. Computing the offset parameters O_ℓ proved the most time consuming step as it involves computing the areas of overlap between grid cells and the irregular census regions. However, the price paid for the speed and robustness of INLA has been the need to ignore uncertainty in the age and time effects.

Acknowledgements

The authors would like to thank the Natural Sciences and Engineering Research Council of Canada for providing a Postgraduate Scholarship for the first author and a Discovery grant for the second author. Funding was also provided by Cancer Care Ontario's Population Studies Network. Todd Norwood provided much of the population data and spatial boundary files and assisted in interpreting them.

References

al Maini, M. (2008). Mapping *Systemic Lupus Erythematosus* and *Psoratic Arthritis* in Greater Toronto area using geographic information systems. Master's thesis, University of Toronto Institute of Medical Science. MSc Thesis. U. of T. IMS.

Alamanos, Y., Voulgari, P., & Drosos, A. (2008). Incidence and prevalence of psoriatic arthritis: a systematic review. *Journal of Rheumatology*, *35*, 1354.

Bernatsky, S. (2007). A population-based assessment of systemic lupus erythematosus incidence and prevalence results and implications of using administrative data for epidemiological studies. *Rheumatology*, *46*, 1814.

Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, *43*(1), 1–20.

Brockbank, J. & Gladman, D. (2002). Diagnosis and management of psoriatic arthritis. *Journal of Clinical Epidemiology*, *62*, 2447–2457.

Clark, J. S. & Gelfand, A. E. (2006). *Hierarchical modelling for the environmental sciences: statistical methods and applications*. Oxford University Press.

Cliff, A. D. & Ord, J. K. (1981). *Spatial processes: models & applications*. Taylor & Francis.

Cooper, G., Dooley, M., Treadwell, E., Clair, E., & Gilkeson, G. (2002). Risk factors for development of systemic lupus erythematosus Allergies, infections, and family history. *Journal of clinical epidemiology*, *55*(10), 982–989.

Cusimano, M., Marshall, S., Rinner, C., Jiang, D., & Chipman, M. (2010). Patterns of Urban Violent Injury: A Spatio-Temporal Analysis. *PLoS ONE*, *5*.

Diggle, P. J. (2003). *Statistical analysis of spatial point patterns*. Oxford University Press.

Lawson, A. B. (2008). *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology*. CRC Press.

Lindgren, F. & Rue, H. (2007). Explicit construction of gmrf approximations to generalised Matern fields on irregular grids. Technical Report 12, Centre for Mathematical Sciences, Mathematical Statistics, Lund Institute of Technology, Lund University.

Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The SPDE approach. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *73*(4). To appear.

Pattison, E., Harrison, B. J., Griffiths, C. E. M., Silman, A. J., & Bruce, I. N. (2008). Environmental risk factors for the development of psoriatic arthritis: results from a case control study. *Annals of the Rheumatic Diseases*, *67*, 672–676.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B(Statistical Methodology)*, *71*(2), 319–392.

Simard, J. F. & Costenbader, K. H. (2007). What can epidemiology tell us about systemic lupus erythematosus? *International Journal of Clinical Practice*, *61*, 1170.

Stein, M. L. (1999). *Interpolation of spatial data: some theory for kriging*. Springer.

Waller, L. A. & Gotway, C. A. (2004). *Applied spatial statistics for public health data*. Wiley-IEEE.

A Proofs

A.1 Intergrated temporal effect

As we assumed popluation is constant between two census years, as a consequence, let N(s) be number of obseved cases at location s

$$
\begin{aligned}
N(s)|R(s) &\sim \text{Poisson}\left(\int_S \int_T \Lambda_j(s,t) P_j(s,t) dsdt\right) \\
E(N(s)|R(s)) &= \int_S \sum_{ij} P_{ij}(s) \int_T \Lambda_j(s,t) dsdt \\
&= \int_S \sum_{ij} P_{ij}(s) R(s) \int_T \Gamma(t)\Theta_j dsdt \\
&= \int_S \sum_{ij} P_{ij}(s) R(s) \Theta_j \int_T \Gamma(t) dsdt \\
&= \int_S \sum_{ij} P_{ij}(s) R(s) \Theta_j \Gamma_i dsdt
\end{aligned}
$$

which only depends on the intergrated temperoal effect Γ_i instead of $\Gamma(t)$

A.2 Sufficiency of M_{ij}

Write $\bar{\mathbf{S}} = \{S_{jk}; j = 1 \dots J, k = 1 \dots K_j\}$, $\bar{\mathbf{T}} = \{T_{jk}; j = 1 \dots J, k = 1 \dots K_j\}$, $\bar{\mathbf{M}} = \{M_{ij}; i = 1 \dots I, j = 1 \dots J\}$ $\bar{\mathbf{\Theta}} = \{\Theta_j; j = 1 \dots J\}$ and $\bar{\mathbf{\Gamma}} = \{\Gamma_i; i = 1 \dots I\}$. To show $[\bar{\mathbf{\Theta}}, \bar{\mathbf{\Gamma}} | \bar{\mathbf{S}}, \bar{\mathbf{T}}, R(\cdot)] \propto [\bar{\mathbf{\Theta}}, \bar{\mathbf{\Gamma}} | \bar{\mathbf{M}}, R(\cdot)]$,

$$
Pr(\bar{s}, \bar{t} | \bar{R}, \bar{\mathbf{\Theta}}, \bar{\mathbf{\Gamma}}) = \prod_j \left(\prod_k \Lambda_j(s_{jk}, t_{jk}) P_j(s_{jk}, t_{jk}) exp\left(- \int_S \int_T \Lambda_j(s,t) P_j(s,t) dsdt \right) \right)
$$

$$
\begin{aligned}
Pr(\bar{\mathbf{\Theta}}, \bar{\mathbf{\Gamma}} | \bar{\mathbf{S}}, \bar{\mathbf{T}}, \bar{R}) &= \prod_{ijk} R(s_{ijk}) \Theta_j \Gamma_i P_j(s_{ijk}) \prod_{ij} exp\left(-\Gamma_i \Theta_j \int_{B_r} R(s) P_{ij}(s) ds \right) \\
&= \prod_{ijk} R(s_{ijk}) \Theta_j \Gamma_i P_j(s_{ijk}) \prod_{ijr} exp\left(-\Gamma_i \Theta_j P_{ijB_r} \int_{B_r} R(s) ds \right) \\
&= \prod_{ij} \left(\Gamma_i \Theta_j \prod_{k} R(s_{ijk}) P_j(s_{ijk}) \right) \prod_{ij} exp(-\Gamma_i \Theta_j P_{ijB_r}) exp\left(-\sum_r \int_{B_r} R(s) ds \right) \\
&= \prod_{ij} \left(\Gamma_i \Theta_j \tilde{M}_{ij} \right) \prod_{ij} exp(-\Gamma_i \Theta_j P_{ijB_r}) exp\left(-\sum_r \int_{B_r} R(s) ds \right)
\end{aligned}
$$

where $\tilde{M}_{ij} = \prod_k R(s_{ijk}) P_j(s_{ijk})$ is sufficient, under the condition that $R(\cdot)$ is 1, $\tilde{M}_{ij} = P_j(s_{ijk})^{M_{ij}}$, therefore M_{ij} is sufficient.

Note that $\tilde{M}_{ij} = exp\left(\sum_k log(P_j(s_{ijk})) + \sum_k log(R(s_{ijk})) \right)$

A.3 Sufficiency of Y_ℓ

Write $\mathbf{S}_j = \{S_{jk}; k = 1 \dots K_j\}$, $\mathbf{T}_j = \{T_{jk}; k = 1 \dots K_j\}$, and $\mathbf{N}_j = \{N_{j\ell}; \ell = 1 \dots L\}$, where$N_{j\ell} = ||k, S_{jk} \in G_\ell||$. To show $[\Lambda_j(\cdot,\cdot) | \mathbf{S}_j, \mathbf{T}_j] = [\Lambda_j(\cdot,\cdot) | Y_\ell, \mathbf{T}_j]$ it suffices, from Bayes theorem, to prove that $[\mathbf{S}_j, \mathbf{T}_j | \Lambda_j(\cdot,\cdot)] \propto [Y_\ell, \mathbf{T}_j | \Lambda_j(\cdot,\cdot)]$.

$$
\begin{aligned}
Pr[\mathbf{S}_j, \mathbf{T}_j | \Lambda_j(\cdot,\cdot)] &= \prod_j \left(\prod_k \Lambda_j(s_{jk}, t_{jk}) P_j(s_{jk}, t_{jk}) exp\left(-\int_S \int_T \Lambda_j(s,t) P_j(s,t) ds dt \right) \right) \\
&= \prod_{jk} R(s_{jk}) \Theta_j \Gamma(t_{jk}) P_j(s_{jk}, t_{jk}) \prod_j exp\left(-\Gamma(t_j) \Theta_j \int_S \int_T R(s) P_{ij}(s,t) ds dt \right)
\end{aligned}
$$

With the grid-cell process where, $R(s) = R_\ell, s \in G_\ell$ then

$$Pr(\mathbf{S}_j, \mathbf{T}_j | \Lambda_j(\cdot,\cdot)) = R_\ell^{Y_\ell} * \tag{3}$$

$$\prod_j \Theta^{N_{j\ell}} \prod_{jk} \Gamma(t_{jk}) * \tag{4}$$

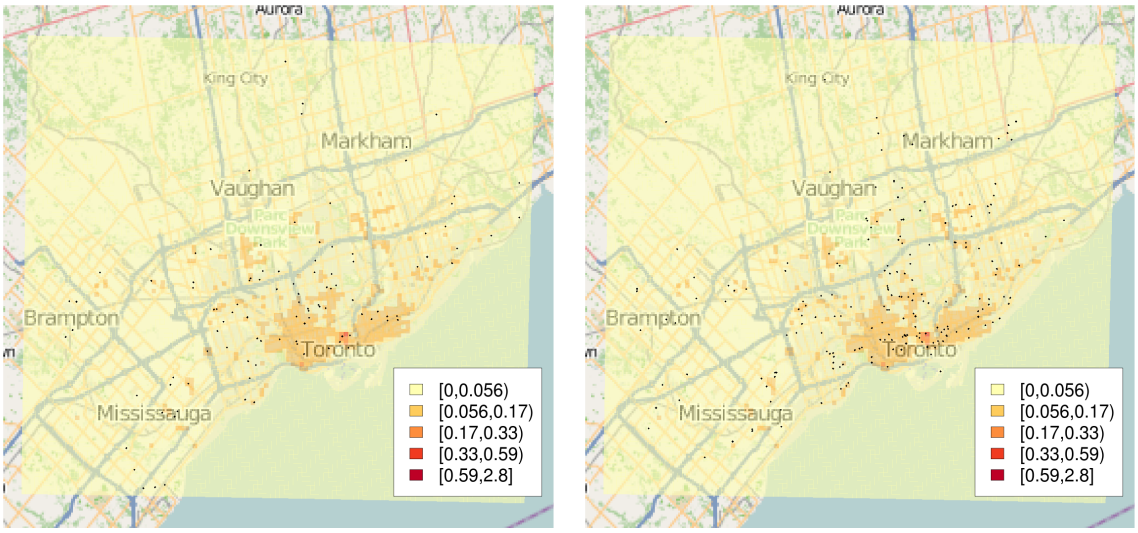
$$\prod_j exp\left(-\Gamma(t_j)\Theta_j \int_S \int_T R(s) P_{ij}(s,t) dsdt\right) \tag{5}$$

where (3) does not depends on $R(\cdot)$, (4) is constant and in (5):

$$\int_S \int_T R(s) P_j(s,t) dsdt = \sum_\ell \int_C R_\ell P_{j\ell}(t) dt$$

is independent of $\mathbf{S}_j$ and $R(\cdot)$, where $P_{j\ell}(t) = \int_{G_\ell} P_j(s,t)$. Therefore this conditional distribution is independent of $\mathbf{S}_j$ given $(Y_\ell, \mathbf{T}_j), P_{j\ell}(t))$, which is proportion to $[Y_\ell, \mathbf{T}_j | \Lambda_j(\cdot,\cdot)]$, therefore $(Y_\ell, \mathbf{T}_j)$ is sufficient for R_ℓ.

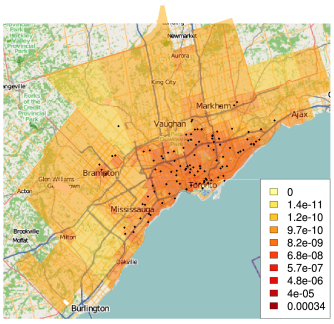
B Additional figures



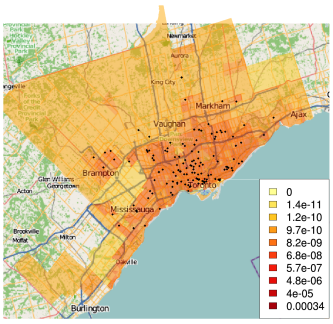
(a) Male Lupus

(b) Female PSA

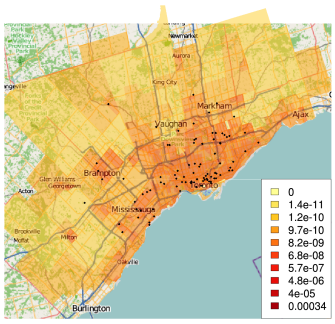
Figure 10: Expected incidence rate per km^2 and observed case locations for period 1965-2007 of Male Lupus and Female Psoriatic Arthritis in Toronto, Canada



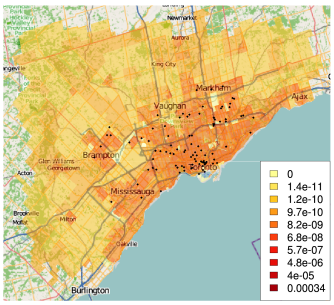
(a) 1971



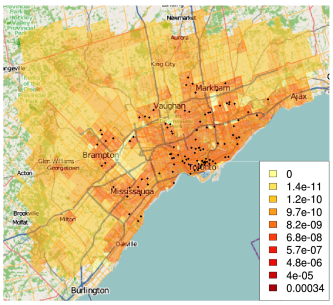
(b) 1981



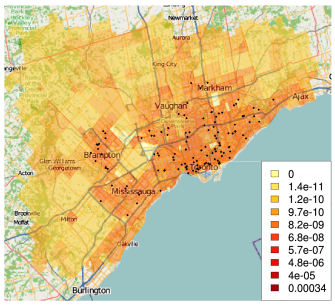
(c) 1986



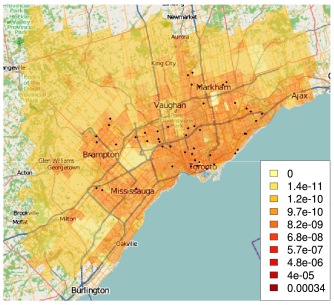
(d) 1991



(e) 1996

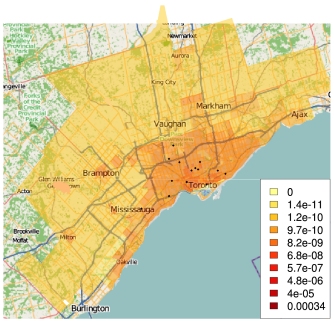


(f) 2001

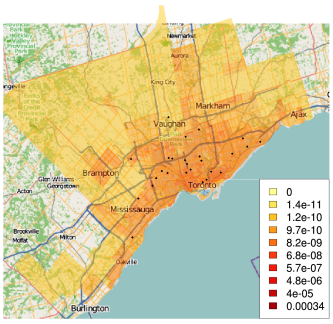


(g) 2006

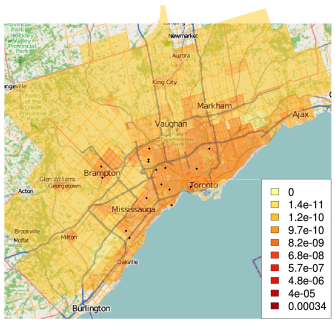
Figure 11: Maps of Female expected and observed Lupus cases for each census year



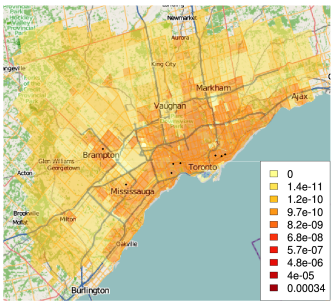
(a) 1971



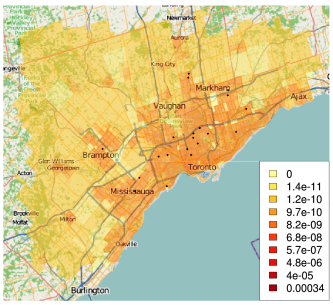
(b) 1981



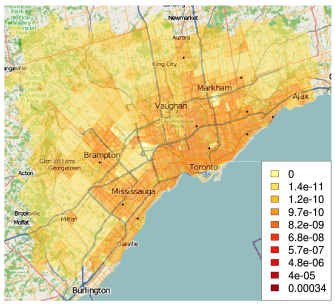
(c) 1986



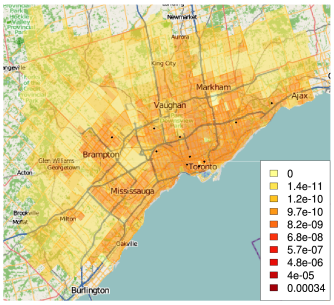
(d) 1991



(e) 1996

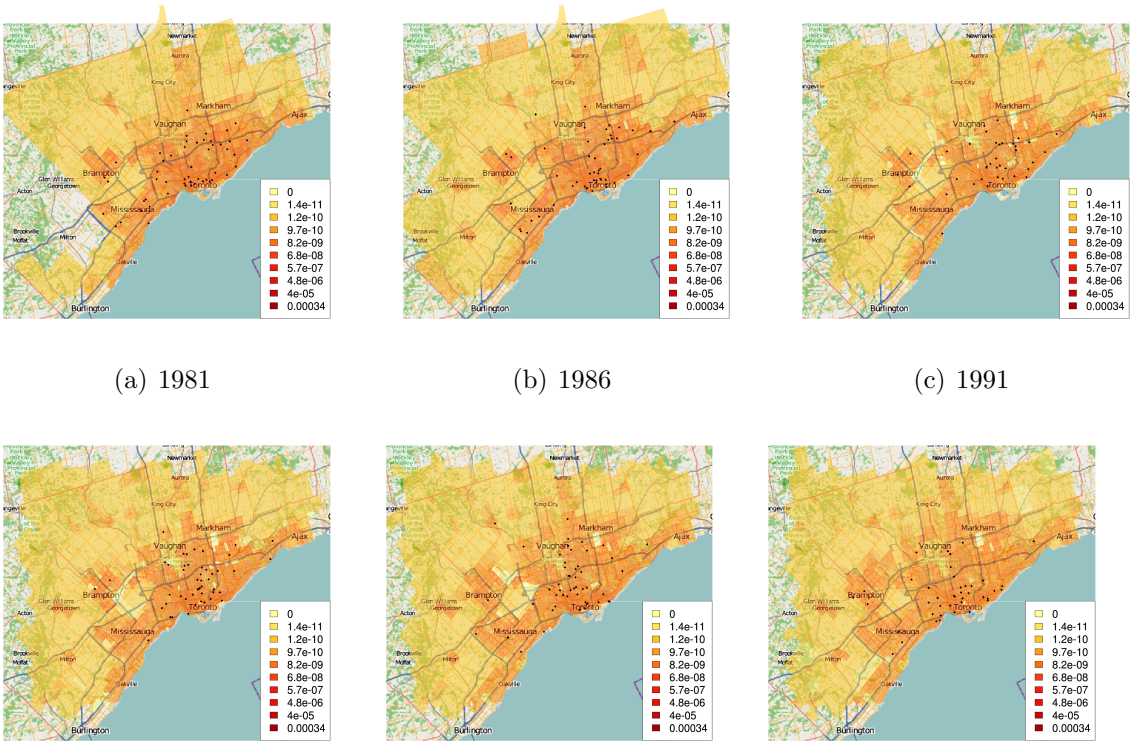


(f) 2001



(g) 2006

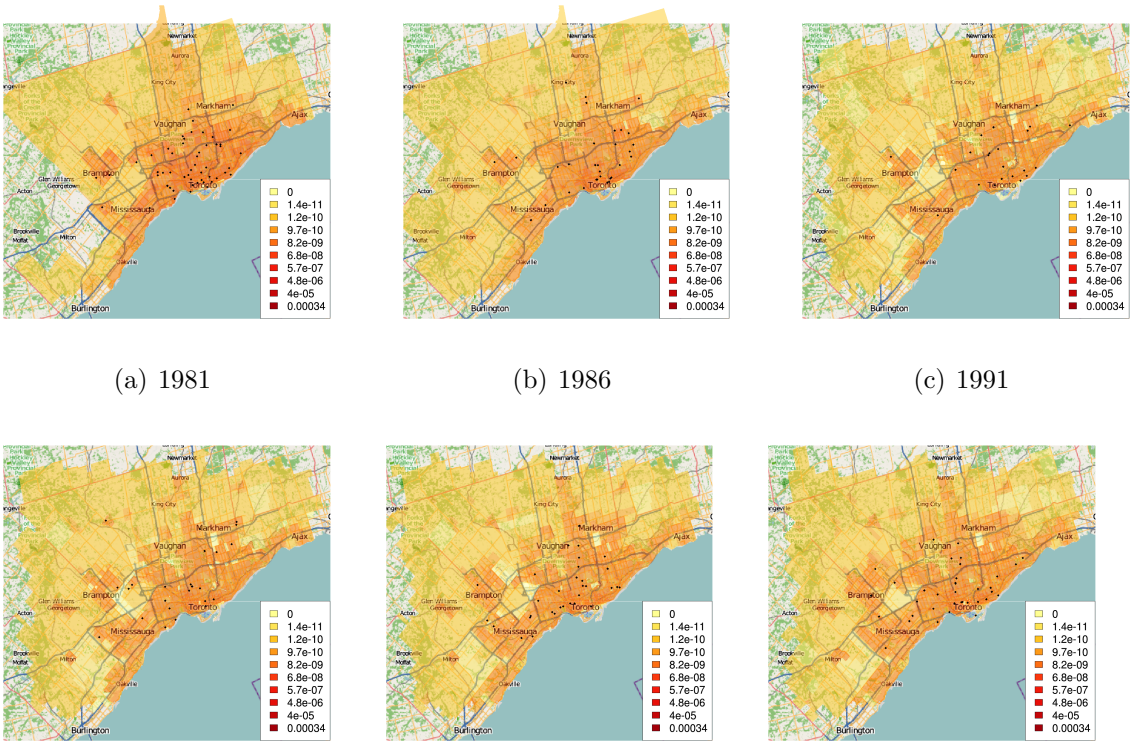
Figure 12: Maps of Male expected and observed Lupus cases for each census year



(a) 1981 (b) 1986 (c) 1991

(d) 1996 (e) 2001 (f) 2006

Figure 13: Maps of Male expected and observed PsA cases for each census year

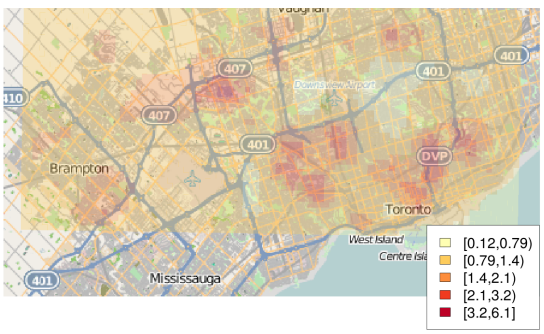


(a) 1981 (b) 1986 (c) 1991

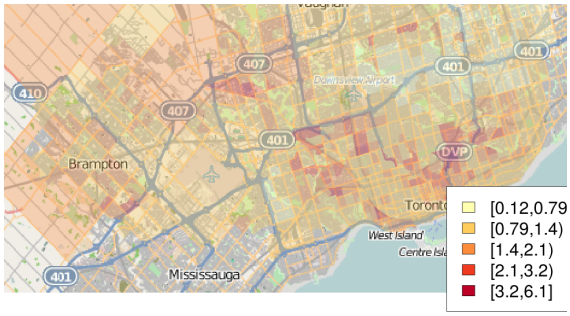
(d) 1996 (e) 2001 (f) 2006

Figure 14: Maps of Female expected and observed PsA cases for each census year

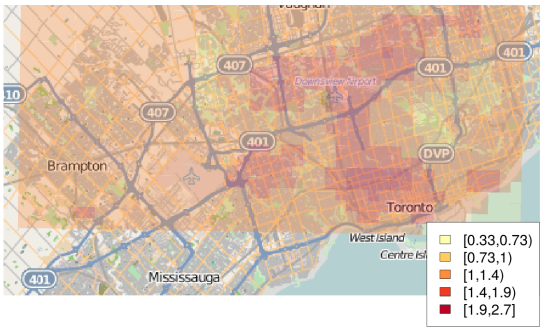
C Further simulation results



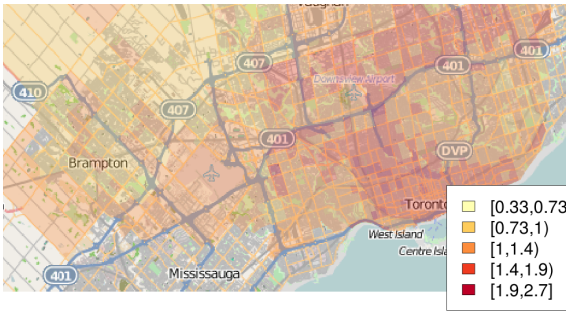
(a) Range 1km, Propsed Model (0.5km)



(b) Range 1km,BYM Model



(c) Range 4km, Propsed Model (1km)



(d) Range 4km,BYM Model

Figure 15: Predicted risk $\mathrm{E}(R(s)|Y)$ from Stomach Cancer

D Additional Results

(a) Female Lupus

(b) Male Lupus

(c) Female PsA

(d) Male PsA

Figure 16: $pr(R(s) > 1.2|Y)$ for alternative prior: shape =1, scale = 0.00005