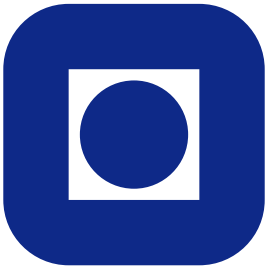
NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

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

by

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

PREPRINT
STATISTICS NO. 7/2010



NORWEGIAN UNIVERSITY OF SCIENCE AND
TECHNOLOGY
TRONDHEIM, NORWAY

This preprint has URL http://www.math.ntnu.no/preprint/statistics/2010/S7-2010.pdf
Håvard Rue has homepage: http://www.math.ntnu.no/∼hrue
E-mail: hrue@math.ntnu.no
Address: Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway.

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

April 15, 2010

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.

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

1 Introduction

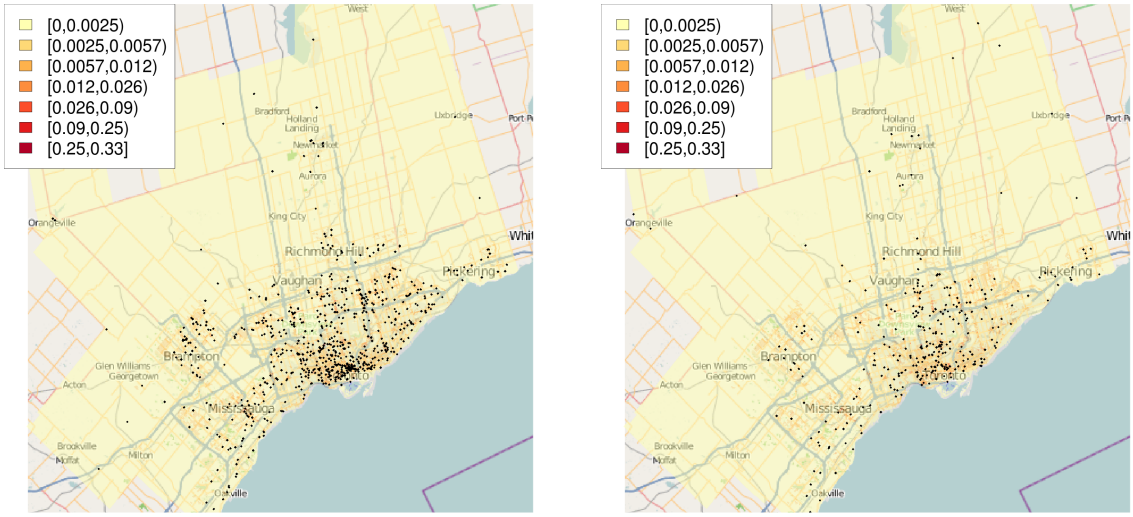
1.1 Lupus and Psoratic Arthritis

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 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 the population density for census year 2006 (per square meter) and the case locations of Female Lupus and Male PsA.



(a) Female Lupus

(b) Male PSA

Figure 1: 2006 population density by Dissemination Area and observed case locations for all years

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 it's 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.

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 (Diggle et al., 1998) 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 are 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} and T_{jk} be the locations in space and time for case k in the jth age group, $P_j(s,t)$ be the population and $\lambda_j(s,t)$ be the risk surface for location s and time t. The locations and times are realisations from a spatio-temporal inhomogeneous Poisson point process with

$$\begin{aligned}
\{T_{jk}, S_{jk}|U(s); k = 1...K_j\} &\sim \text{Poisson Point Process}[\lambda_j(s,t)P_j(s,t)] \\
\lambda_j(s,t) &= \lambda(s)\exp(\gamma(t) + \theta_j) \\
\log(\lambda(s)) &= \mu + U(s) \\
\text{Cov}[U(s), U(s+h)] &= \sigma^2 \text{Matèrn}(|h|/\varphi, \nu).
\end{aligned}$$

Here $\gamma(t)$ is the fixed time effect and θ_j is the age group effect, $\lambda(s)$ is the pure spatial risk surface and $U(s)$ is the spatial relative risk surface on the log scale, which has a Matèrn correlation structure with roughness ν and range φ.

2.2 Inference

2.2.1 Approximation on Grid

The first assumption made is to assume that $U(s)$ is constant within cells of a grid covering the study region, with $U(s) = U_\ell$ when s is in grid cell G_ℓ. This serves two purposes. First, it allows the Matèrn 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., 2010; 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. The number of cases in cell G_ℓ from age group j, written $N_{j\ell} = ||k, S_{jk} \in G_\ell||$, is along with the times T_{jk} sufficient for performing inference on the spatial surface, i.e. $[\lambda_j(s_{jk}, t_{jk})|S_{jk}, T_{jk}] = [\lambda_j(s_{jk}, t_{jk})|N_{j\ell}, T_{jk}]$ (see Appendix A.1). By integrating the spatio-temporal risk $\lambda_k(s,t)$ over grid cell G_ℓ and through time, the distribution of of the cell counts is derived as

$$N_{j\ell}|U_\ell \sim \text{Poisson}\left[\exp(U_\ell\theta_j)\int P_{j\ell}(t)e^{\gamma(t)}dt\right]$$
$$P_{j\ell}(t) = \int_{G_\ell} P_{j\ell}(s,t)ds.$$

2.2.2 Inference on age and time effects

An assumption regarding the distribution of population within census regions and between census periods is necessary in order to obtain populations within grid cells. Population density is taken to be constant between the midpoints of census years (5 year intervals after 1981), and within census regions (CT's or DA's). Let C_i denote the ith census interval, and write $P_j(s,t) = P_{ij}(s); t \in C_i$ and $P_{ij\ell} = \int_{G_\ell} P_{ij}(s)ds$. Note that grid cells cross census region boundaries and populations are not assumed to be constant within grid cells.

As shown in the following section, inference on the spatial surface $U(s)$ depends

only on the integrated time effects. Hence we define γ_i as

$$\gamma_i = \log \int_{C_{i-1}}^{C_i} exp(\gamma(t))dt.$$

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, $[\theta_j, \gamma_i | S_{jk}, T_{jk}, U_\ell, \mu] \propto [\theta_j, \gamma_i | M_{ij}, U_\ell, \mu]$ and estimation of these parameters is accomplished using

$$\begin{aligned} M_{ij}|U_\ell &\sim \text{Poisson}\left(\int_{C_i}\int_{R^2} \lambda_j(s,t)P_j(s,t)dsdt\right) \\ &\sim \text{Poisson}\left(\sum_{i\ell} \exp(\mu + \theta_j + \gamma_i + U_\ell)P_{ij\ell}\right) \end{aligned}$$

2.2.3 Inference on the Spatial effects

Conditioning on θ_j and γ_i, the distribution of U_ℓ depends only the total case count Y_ℓ for that cell, with Appendix A.3 showing that $[U_\ell|\theta_j, \gamma(t), \mu, S_{jk}, T_{jk}] = [U_\ell|\theta_j, \gamma_i, \mu, Y_\ell]$. Integrating under the intensity surface and summing over age groups gives the cell counts Y_ℓ being Poisson distributed

$$\begin{aligned} Y_\ell &\sim \text{Poisson}(O_\ell \bar{\lambda}_\ell) \\ O_\ell &= \sum_i \int_{G_\ell} \gamma_i \theta_j P_{ij\ell} |C_i| ds \\ \log(\bar{\lambda}_\ell) &= \mu + U_\ell \\ \text{Cov}(U_\ell, U_m) &= \sigma^2 \text{Matèrn}(|G_\ell - G_m|/\phi; \nu) \end{aligned} \tag{1}$$

The O_ℓ are an offset parameter, interpretable as the expected number of cases when $\lambda(s) = 1$. The $\bar{\lambda}_\ell$ is the risk in cell G_ℓ and $\bar{\lambda}_\ell = \int_{G_\ell} \lambda(s)ds$.

The model above 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 and weights for each neighbour are determined by the scale 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 from M_{ij} with the spatial random effects U_ℓ set to zero i.e. using $M_{ij} \sim \text{Poission}(\exp(\mu + \theta_j + \gamma_i)P_{ij})$. Then treating the $\hat{\theta}_j$ and $\hat{\gamma}_i$ as fixed known values, the offset parameters for each grid cell O_ℓ are computed. Inference on the spatial surface U_ℓ, it's variance σ and range ϕ, and the mean parameter μ are made using the purely spatial Generalized Linear Mixed Model with spatial random effects 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 in this instance time) effects that the parameters are estimated with sufficient precision that the effect of uncertainty in these parameter estimates is believed to be negligible. 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 45km by 20km central section of the Greater Toronto Area was covered with a grid of 104 by 230 cells spaced 200m apart. 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 add further computational burden to model fitting, 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 multiplying by constants related to the grid cell size and area of each census region, 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 (see Clark & Gelfand, 2006) in spatial models, 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 strong prior on the range.

The range parameter is given a Gamma prior with 95% support between 400m and 4km, giving a mean of 2km and shape parameter 6.5. This reflects the expert opinion solicited on lupus and its risk factors. The precesion $1/\sigma^2$ has a weaker prior of Gamma with range 1 and shape 0.01, giving a prior mean for σ of 0.16 and 95% interval (0.052, 0.63). Idential priors were used for all disease and sex group combinations.

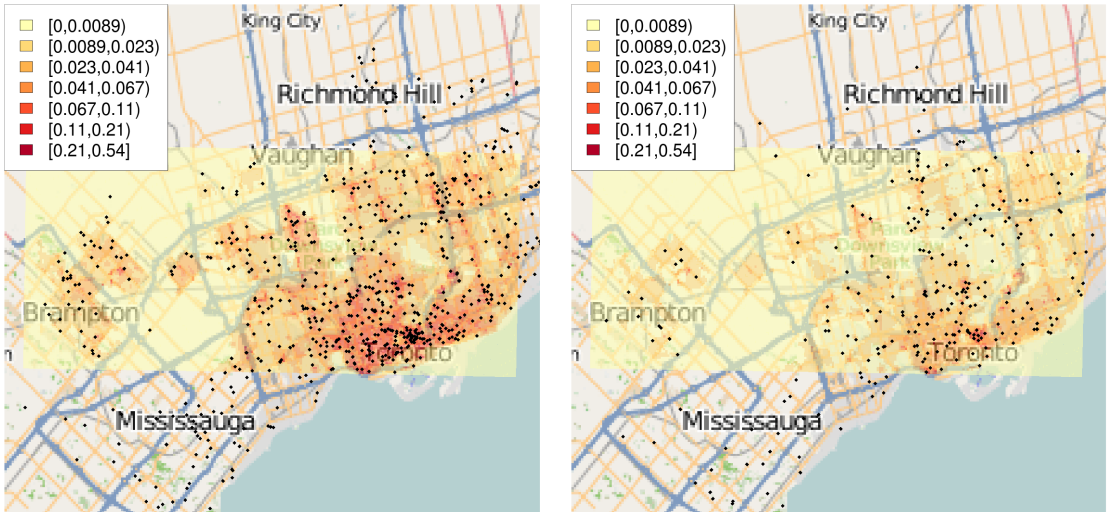
3 Results

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	555
Male Lupus	108	75
Female PsA	216	156
Male PsA	311	222

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.

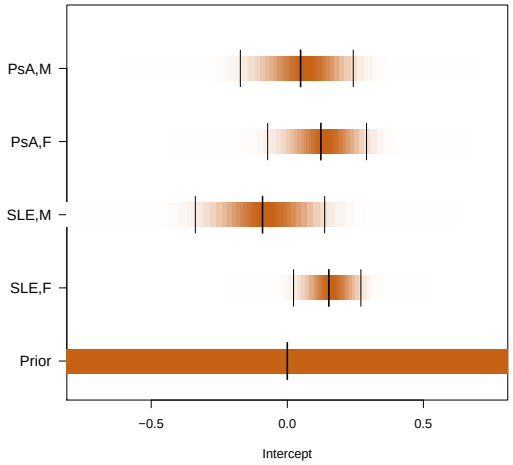
Figure 2 shows the expected incidence rate, in cases per km^2 over the study period, calculated as $O_\ell/0.4^2$ with case locations superimposed as dots. Additional graphs for female lupus and male PsA, as well as maps for individual census periods, are shown in Appendix B.



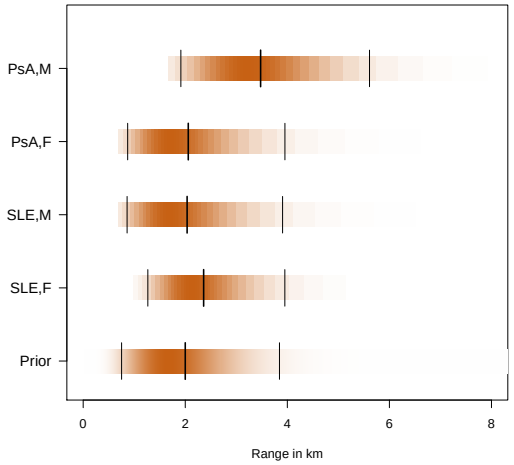
(a) Female Lupus (b) Male PSA

Figure 2: Expected count and observed case locations

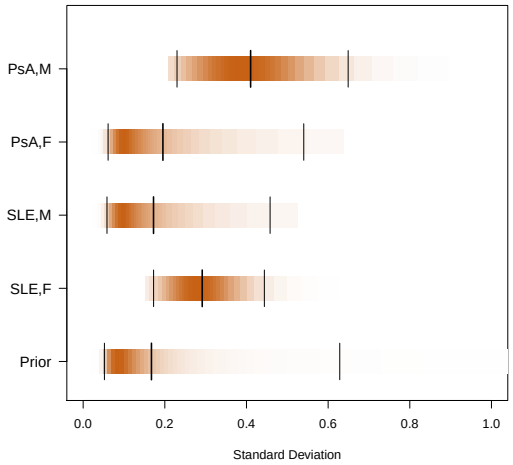
Figure 3 shows the prior and posterior densities of the intercept, range and standard deviation parameters for each outcome. Female Lupus is the most prevalent outcomes and correspondingly has posterior distributions which are tighter than male lupus and and PsA. As expected, the data are for the most part uninformative regarding the range parameter with the posteriors in Figure 3c being similar to the prior. Female lupus and male PsA, being the two most common outcomes, show posteriors with some departure from the prior with female PsA and male lupus being indistinguishable from the prior.



(a) Intercept μ



(b) Range φ

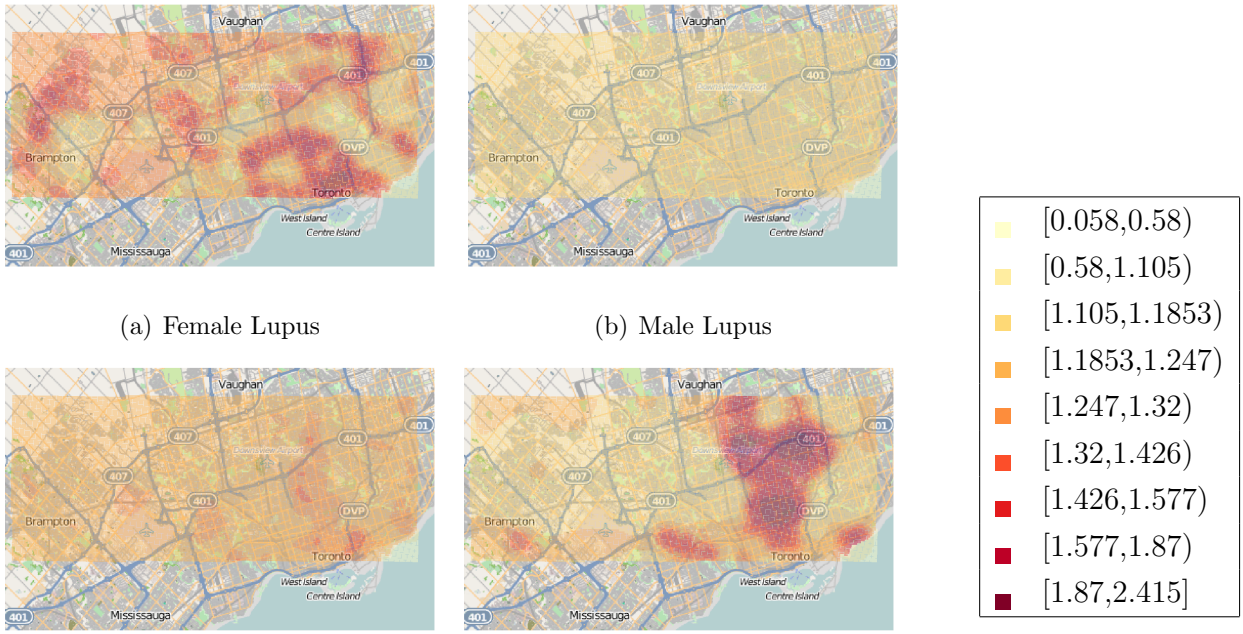


(c) Standard Deviation $sd(U(s))$

Figure 3: Prior and Posterior of Hyperparameters for Female Lupus and Male PsA

Figure 4 shows the predicted values of the risk $\lambda(s)$, or $\mathrm{E}(\lambda(s)|Y)$, 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 5 reflects uncertainty in estimation by plotting 20% exceedance probabilities, or $\text{pr}(\lambda(s) > 1.2|S,T)$. 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 probablity above, say, 80% or 95% would be indicative or warranting further investigation.



(a) Female Lupus (b) Male Lupus

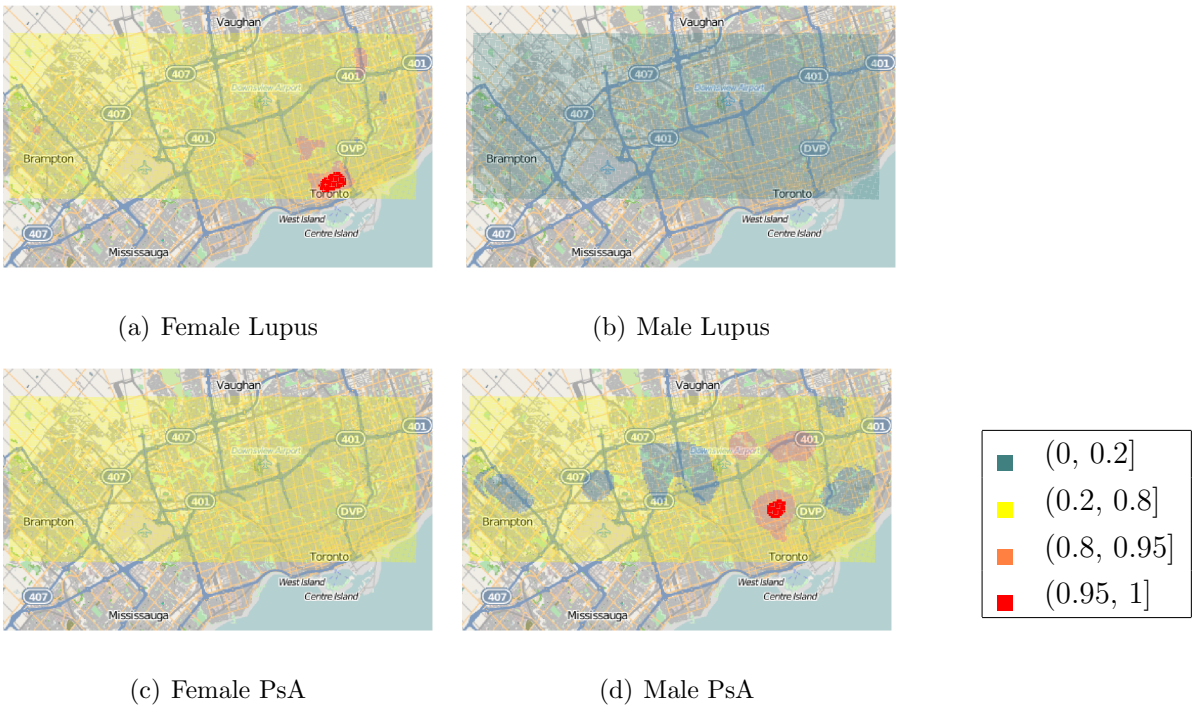
(c) Female PsA (d) Male PsA

Figure 4: Posterior mean of relative risks, $E(\lambda(s)|Y)$ for each of the disease and sex combinations

A sizable area in the bottom right corner of Figure 5A 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 Brampton, do not prove to be significant. A significant male PsA cluster exists to the north and west of the Lupus cluster, 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 5: Posterior probability of relative risks being more than 20% above the average rate, $pr(\lambda(s) > 1.2|Y)$ for each of the disease and sex combinations

4 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. While reporting bias should certainly be suspected and investigated, the presence of the Wellesley hospital is not the only unique feature of this neighbourhood. Cusimano et al. (2010) have found this area to have the highest violent crime incidence in the city, for instance. The absence of a PsA cluster around the hospital also suggests reporting bias might not be the overriding factor. 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 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, though given the small number of cases in the area concerned an in-depth analysis of the medical charts by a clinician would be a more sensible first step.

The use of INLA for inference has resulted in fast and efficient estimation. 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.

Although the clinical data used here provided full street addresses which were geocoded as points, many spatial surveillance problems involve the use of spatially censored data aggregated to the postal code or census region. Although a large proportion of cases will be in dense areas where these regions are quite small, maps produced are often dominated by larger less populated regions where the spatially censored data is less precise. Methods for spatially aggregated data with changing boundaries would be

more complex but would address a wide range of problems in epidemiology and public health.

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.

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.

Diggle, P. J., Tawn, J. A., & Moyeed, R. A. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *47*(3), 299–350.

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

Lindgren, F., Lindström, J., & Rue, H. (2010). An explicit link between Gaussian fields and Gaussian Markov random fields: The SPDE approach. Technical report 5, Department of mathematical sciences, Norwegian University of Science and Technology.

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.

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 Sufficiency of $N_{j\ell}$

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

$$\begin{aligned}
\log Pr[\mathbf{S}_j, \mathbf{T}_j|\lambda_j(\cdot,\cdot)] &= \sum_{k=1}^{K_j} \log(\lambda_j(s_{jk}, t_{jk})P_j(s_{jk}, t_{jk})) - \int_S \int_T \lambda_j(s,t)P_j(s,t)dsdt \\
&= \sum_k (\mu + U(s_{jk}) + \theta_j + \gamma(t_{jk}) + \log(P_j(s_{jk}, t_{jk}))) \\
&\quad - \int_S \int_T \lambda_j(s,t)P_j(s,t)dsdt
\end{aligned}$$

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

$$\begin{aligned}
\log Pr(\mathbf{S}_j, \mathbf{T}_j|\lambda_j(\cdot,\cdot)) =& N_{j\ell}U_\ell + (\mu + \theta_j)\sum_\ell N_{j\ell} + \sum_k \gamma(t_{jk}) && (2)\\
&+ \sum_k \log(P_j(s_{jk}, t_{jk})) && (3)\\
&- \int_S \int_T \lambda_j(s,t)P_j(s,t)dsdt && (4)
\end{aligned}$$

where (2) does not depends on S_j, (3) is constant with respect to S_j and (4),

$$\int_S \int_T \lambda_j(s,t)P_j(s,t)dsdt = \sum_\ell \int_C \exp[\mu + U_\ell + \theta_j + \gamma(t)]P_{j\ell}(t)dt$$

is independent of $\mathbf{S}_j$. Therefore this conditional distribution is independent of $\mathbf{S}_j$ given $(\mathbf{N}_j, \mathbf{T}_j), P_{j\ell}(t))$, which is proportion to $[\mathbf{N}_j, \mathbf{T}_j|\lambda_j(\cdot,\cdot)]$, therefore $(\mathbf{N}_j, \mathbf{T}_j)$ is sufficient for U_ℓ.

A.2 Sufficiency of M_{ij}

Write $\bar{\mathbf{S}} = \{S_{jk}; j = 1 \dots J, k = 1 \dots K_j\}$, $\bar{\mathbf{T}} = \{T_{jk}; j = 1 \dots J, k = 1 \dots K_j\}$, $\bar{\mathbf{M}} = \{M_{ij}; i = 1 \dots I, j = 1 \dots J\}$ $\bar{\boldsymbol{\theta}} = \{\theta_j; j = 1 \dots J\}$ and $\bar{\boldsymbol{\gamma}} = \{\gamma_i; i = 1 \dots I\}$. To show $[\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\gamma}} | \bar{\mathbf{S}}, \bar{\mathbf{T}}, U_\ell, \mu] \propto [\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\gamma}} | \bar{\mathbf{M}}, U_\ell, \mu]$, similar to the proof in Appendix A.1,

$$\log Pr(\bar{s}, \bar{t} | \bar{U}, \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\gamma}}, \mu) \propto \sum_j \left(\sum_k \log(\lambda_j(s_{jk}, t_{jk}) P_j(s_{jk}, t_{jk})) - \int_S \int_T \lambda_j(s,t) P_j(s,t) ds dt \right)$$

$$\begin{aligned}
\log Pr(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\gamma}} | \bar{\mathbf{S}}, \bar{\mathbf{T}}, \bar{U}, \mu) &= \sum_{ij} \left(\sum_k \mu + U(s_{ijk}) + \theta_j + \gamma_i + \log P_{ij}(s_{ijk}) \right) - \sum_{ij} \int_S \lambda_{ij}(s) P_{ij}(s) ds \\
&= \mu \sum_{ij} M_{ij} + \sum_{ijk} U(s_{ijk} + \sum_{ij} \theta_j + \sum_{ij} \gamma_i + \sum_{ijk} \log P_{ij}(s_{ijk}) \\
&- \sum_{ij\ell} \exp((\mu + \theta_j + \gamma_i + U_\ell) P_{ij\ell}) \\
&= \mu \sum_{ij} M_{ij} + \sum_\ell U_\ell Y_\ell + I \sum_j \theta_j + J \sum_i \gamma_i + \sum_{ijk} \log P_{ij}(s_{ijk}) \\
&- e^\mu \sum_{ij} e^{\theta_j} e^{\gamma_i} \sum_\ell e^{U_\ell} P_{ij\ell}
\end{aligned}$$

which is independent of $\bar{\mathbf{S}}$ and $\bar{\mathbf{T}}$. Note that if $U_\ell = 0$, it becomes the poission likelihood.

A.3 Sufficiency of Y_ℓ

$\bar{\mathbf{N}} = \{N_{ij}; i = 1 \dots I, j = 1 \dots J\}$ and $\bar{\mathbf{Y}} = \{Y_\ell; \ell = 1 \dots L\}$, $\bar{\mathbf{U}} = \{U_\ell; \ell = 1 \dots L\}$. To show $\bar{\mathbf{Y}}$ is sufficient for $\bar{\mathbf{U}}$, it suffices to show that $[\bar{\mathbf{U}} | \bar{\boldsymbol{\theta}}, \gamma(\cdot), \bar{\mathbf{N}}, \bar{\mathbf{T}}, \mu] \propto [\bar{\mathbf{U}} | \bar{\mathbf{Y}}, \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\gamma}}, \mu]$

$$\begin{aligned}
[U_\ell | \bar{\boldsymbol{\theta}}, \gamma(\cdot), N_{j\ell}, T_{jk}, \mu] &\propto [N_{j\ell}, T_{jk} | U_\ell, \bar{\boldsymbol{\theta}}, \gamma(\cdot), \mu] [U_\ell | \bar{\boldsymbol{\theta}}, \gamma(\cdot), \mu] \\
&\propto [T_{jk} | N_{j\ell}, U_\ell, \bar{\boldsymbol{\theta}}, \gamma(\cdot), \mu] [N_{j\ell} | U_\ell, \bar{\boldsymbol{\theta}}, \gamma(\cdot), \mu] [U_\ell | \bar{\boldsymbol{\theta}}, \gamma(\cdot), \mu]
\end{aligned}$$

We have shown that $(\bar{\mathbf{N}}, \bar{\mathbf{T}})$ is sufficient for $\bar{\mathbf{U}}$, so $[\bar{\mathbf{T}} | \bar{\mathbf{N}}, \bar{\mathbf{U}}, \bar{\boldsymbol{\theta}}, \gamma(\cdot), \mu] = [\bar{\mathbf{T}} | \bar{\mathbf{N}}, \bar{\boldsymbol{\theta}}, \gamma(\cdot), \mu]$, which is constant with respect to $\bar{\mathbf{U}}$, therefore,

$$
\begin{aligned}
[\bar{\mathbf{U}}|\bar{\boldsymbol{\theta}}, \gamma(\cdot), \bar{\mathbf{N}}, \bar{\mathbf{T}}, \mu] &\propto [\bar{\mathbf{N}}|\bar{\mathbf{U}}, \bar{\boldsymbol{\theta}}, \gamma(\cdot), \mu][\bar{\mathbf{U}}|\bar{\boldsymbol{\theta}}, \gamma(\cdot), \mu] \\
&\propto [\bar{\mathbf{N}}|\bar{\mathbf{U}}, \bar{\boldsymbol{\theta}}, \gamma(\cdot), \mu, \bar{\mathbf{Y}}][\bar{\mathbf{Y}}|\bar{\mathbf{U}}, \bar{\boldsymbol{\theta}}, \gamma(\cdot), \mu][\bar{\mathbf{U}}|\bar{\boldsymbol{\theta}}, \gamma(\cdot), \mu]
\end{aligned}
$$

Again $[\bar{\mathbf{N}}|\bar{\mathbf{U}}, \bar{\boldsymbol{\theta}}, \gamma(\cdot), \mu, \bar{\mathbf{Y}}]$ is constant as $Y_\ell = \sum_j N_{j\ell}$, therefore,

$$
\begin{aligned}
[\bar{\mathbf{U}}|\bar{\boldsymbol{\theta}}, \gamma(\cdot), \bar{\mathbf{N}}, \bar{\mathbf{T}}, \mu] &\propto [\bar{\mathbf{Y}}|\bar{\mathbf{U}}, \bar{\boldsymbol{\theta}}, \gamma(\cdot), \mu][\bar{\mathbf{U}}|\bar{\boldsymbol{\theta}}, \gamma(\cdot), \mu] \\
&\propto [\bar{\mathbf{Y}}|\bar{\mathbf{U}}, \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\gamma}}, \mu][\bar{\mathbf{U}}|\bar{\boldsymbol{\theta}}, \gamma(\cdot), \mu]* \\
&\propto [\bar{\mathbf{U}}|\bar{\mathbf{Y}}, \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\gamma}}, \mu][\bar{\mathbf{U}}|\bar{\boldsymbol{\theta}}, \gamma(\cdot), \mu][\bar{\mathbf{U}}] \\
&\propto [\bar{\mathbf{U}}|\bar{\mathbf{Y}}, \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\gamma}}, \mu]
\end{aligned}
$$

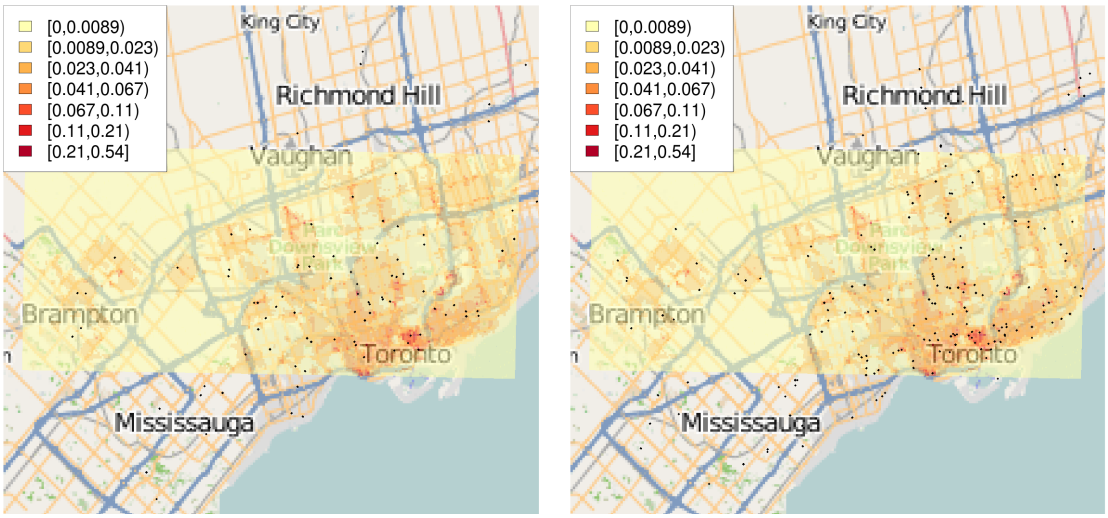
So $\bar{\mathbf{Y}}$ is sufficient for $\bar{\mathbf{U}}$.

*Note that as we assumed popluation is constant between two census years, as a consequence,

$$
\begin{aligned}
Y_\ell|U_\ell &\sim \text{Poisson}\left(\int_{G_\ell}\int_T \lambda_j(s,t)P_j(s,t)dsdt\right) \\
E(Y_\ell|U_\ell) &= \int_{G_\ell} \sum_i P_{ij}(s) \int_T \lambda_j(s,t)dsdt \\
&= \int_{G_\ell} \sum_i P_{ij}(s) \int_T exp(\gamma(t) + \theta_j + U_\ell + \mu)dsdt \\
&= \int_{G_\ell} \sum_i P_{ij}(s)exp(\theta_j + U_\ell + \mu) \int_T exp(\gamma(t))dsdt \\
&= \int_{G_\ell} \sum_i P_{ij}(s)exp(\theta_j + U_\ell + \mu)exp(\gamma_i)ds \\
&= \int_{G_\ell} \sum_i P_{ij}(s)exp(\theta_j + U_\ell + \mu + \gamma_i)ds
\end{aligned}
$$

which only depends on the intergrated temperoal effect γ_i instead of $\gamma(t)$

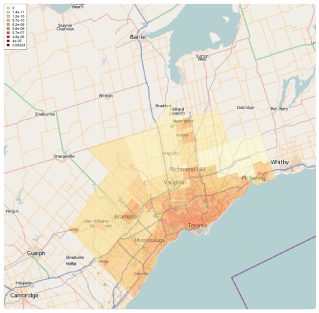
B Additional figures



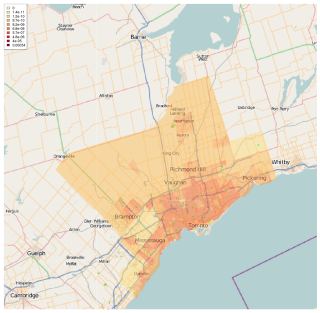
(a) Male Lupus

(b) Female PSA

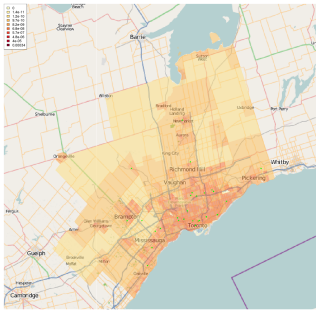
Figure 6: Expected Count and observed case locations



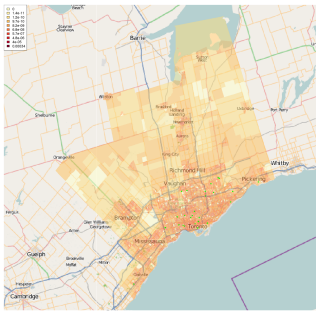
(a) 1971



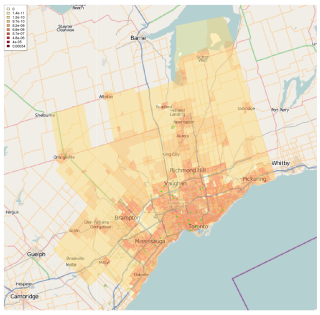
(b) 1981



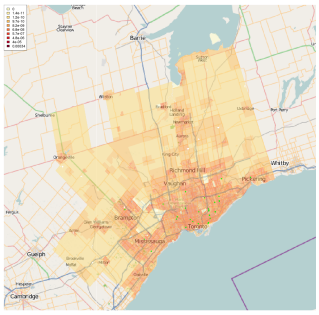
(c) 1986



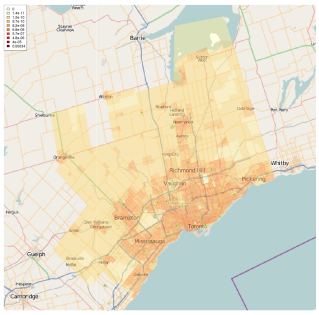
(d) 1991



(e) 1996

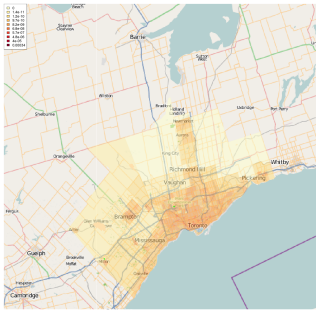


(f) 2001

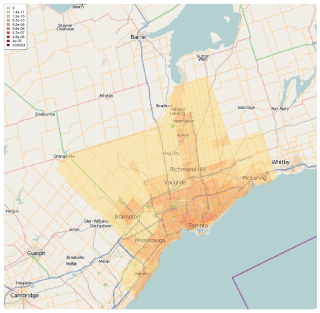


(g) 2006

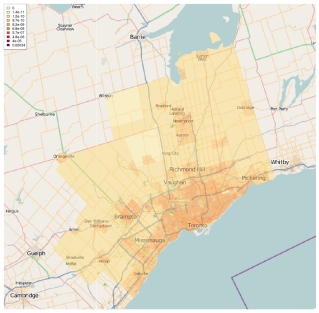
Figure 7: Maps of Female expected and observed Lupus cases for each census year



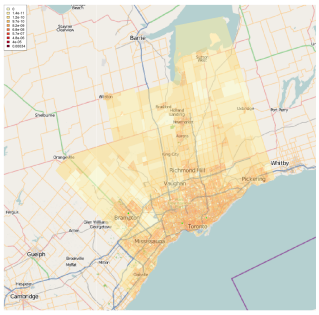
(a) 1971



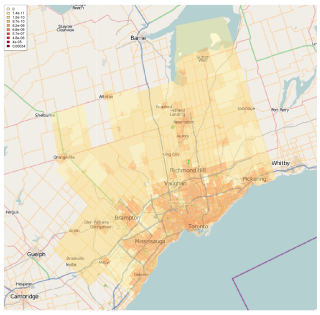
(b) 1981



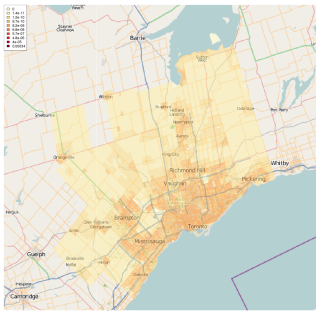
(c) 1986



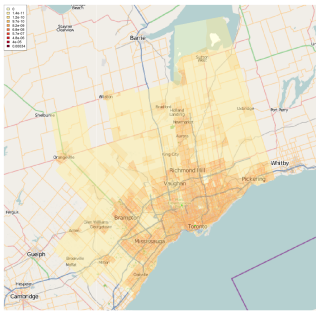
(d) 1991



(e) 1996

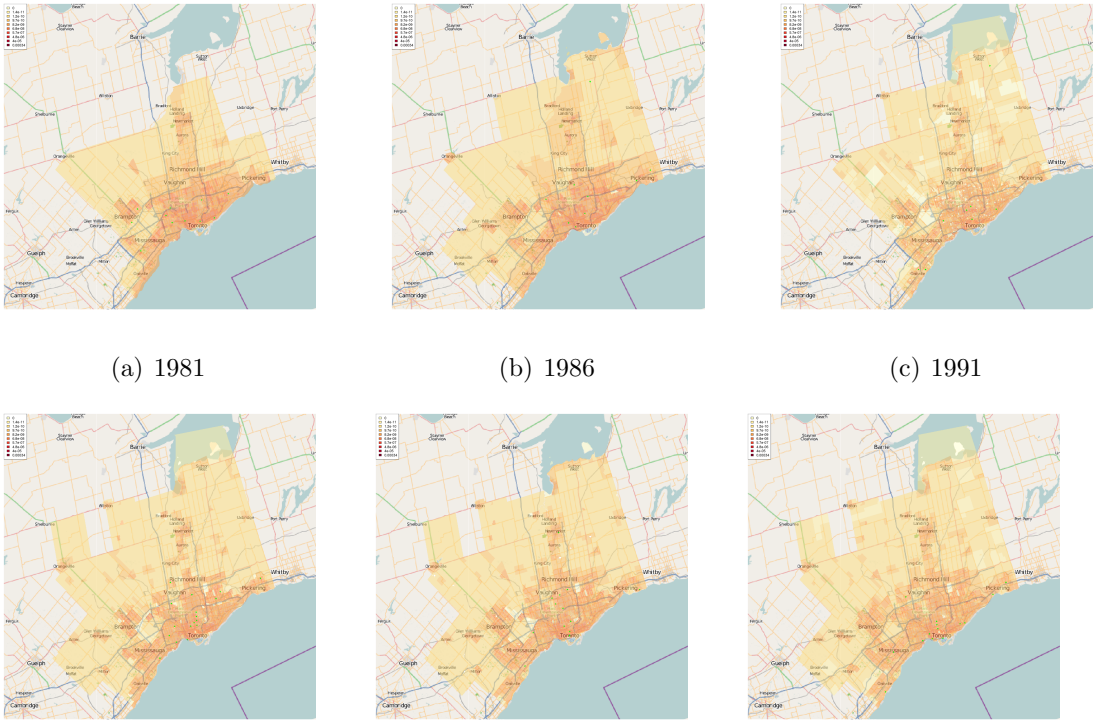


(f) 2001



(g) 2006

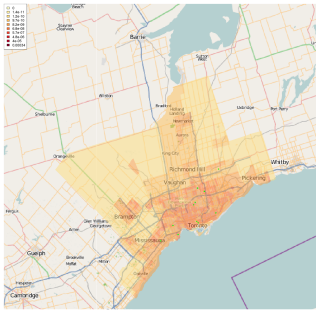
Figure 8: Maps of Male expected and observed Lupus cases for each census year



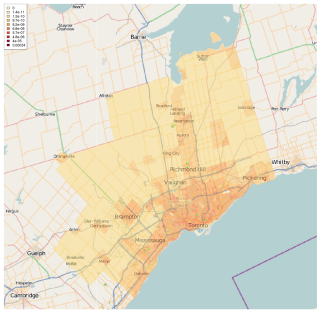
(a) 1981 (b) 1986 (c) 1991

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

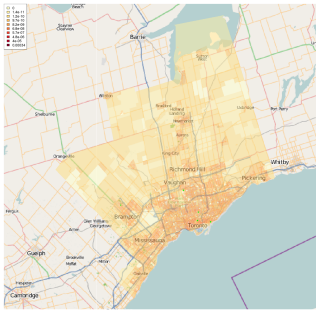
Figure 9: Maps of Male expected and observed PsA cases for each census year



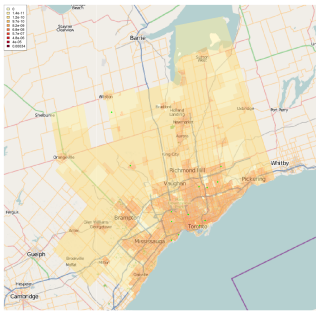
(a) 1981



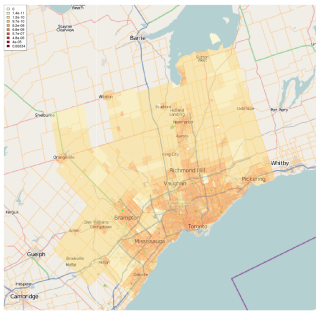
(b) 1986



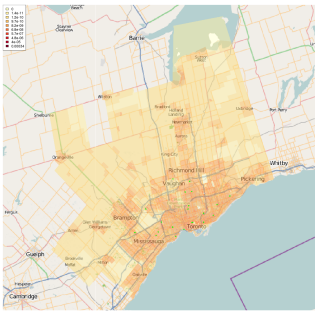
(c) 1991



(d) 1996



(e) 2001



(f) 2006

Figure 10: Maps of Female expected and observed PsA cases for each census year