Spatial Statistics 21 (2017) 75–95

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta

A local-EM algorithm for spatio-temporal disease mapping with aggregated data

Jonathan S.W. Lee [a], Paul Nguyen [b], Patrick E. Brown [c,*], Jamie Stafford [d], Nathalie Saint-Jacques [e]

[a] *Capital One, Canada*
[b] *Institute for Clinical Evaluative Sciences, Canada*
[c] *University of Toronto and St. Michael's Hospital, Canada*
[d] *University of Toronto, Canada*
[e] *Dalhousie University and Cancer Care Nova Scotia, Canada*

ARTICLE INFO

Article history:
Received 9 January 2017
Accepted 3 May 2017
Available online 3 June 2017

Keywords:
Local-EM
Kernel smoothing
Disease mapping
Nonparametric inference
Spatio-temporal statistics

ABSTRACT

Spatial data on disease incidence locations are often aggregated to regional counts to preserve privacy, and spatio-temporal modelling of such can be problematic when there are boundary changes over the study period. Here an inhomogeneous Poisson process with intensity depending on variations in population (known a priori) and a smoothly varying relative risk is estimated with a local-Expectation–Maximization (or local-EM) algorithm. Using incidence data for male bladder cancer in Nova Scotia, Canada, the question of whether the data are consistent with spatially varying but temporally constant relative risk is examined. Areas where there is evidence that relative risk is substantially greater than 1 are identified with the intention of assessing the possible presence of environmental risk factors.

This paper extends existing work by incorporating a temporally varying risk surface and an explicit data structure which contains a mixture of point locations and locations aggregated to non-nested areas. This added flexibility allows the modelling of data amalgamated from different sources and collected over many years. While local-EM leads naturally to an Expectation–Maximization–Smoothing algorithm, the extension to mixtures of aggregations

* Corresponding author.
E-mail addresses: jonnnny@gmail.com (J.S.W. Lee), nguy82@gmail.com (P. Nguyen), patrick.brown@utoronto.ca (P.E. Brown), stafford@utstat.toronto.edu (J. Stafford), nathalie.st-jacques@ccns.nshealth.ca (N. Saint-Jacques).

http://dx.doi.org/10.1016/j.spasta.2017.05.001

leads to a modified algorithm that includes an additive term at every iteration to account for observed point locations.

1. Introduction

One difficulty that often arises when modelling small-scale spatial variation in disease risk is that the relevant data are often spatially aggregated. Residential locations of individual cases are typically reported as geographical regions either to preserve privacy or because full street addresses are not collected or retained. When locations are aggregated to non-overlapping and static regions, such as census areas or administrative units, disease risk is often treated as being spatially discrete (or equivalently piecewise-constant) at the region level. The Besag–York–Mollié model (or BYM, see Besag et al., 1991) continues to be the dominant methodology for modelling random spatial variation in disease risk for data of this sort, using a spatially autoregressive random effect to allow for possible dependence between risk in neighbouring regions. While this approach is often adequate for data collected over a short time interval, it can be problematic when data span a long time period containing more than one census of population. Spatial boundaries may change during the study period, and the degree of spatial aggregation of locations tends to lessen over time. Further, temporal effects are harder to ignore as the time period of interest is lengthened and spatio-temporal variation becomes more important to account for. Modelling data collected over many years is of particular importance when studying a rare disease in an area with low population density, as the number of cases in a single year or census period will be small. Many environmental exposures of interest to epidemiologists (i.e. involving industrial facilities or human consumption of ground water) occur in areas of relatively low population, and the health outcomes with which they are postulated to have an association are often very specific (i.e. childhood cancers). Combining these issues with the fact that administrative regions tend to be large outside of urban areas results in a large class of problems where the BYM model is of limited usefulness.

The approach taken here is to treat aggregated locations as a "missing data" or censoring problem, constructing a spatial point process model for case locations and making inferences conditional on the unknown locations being within their respective aggregation areas. The first attempt at explicitly modelling area-level data as an aggregated continuous process appears to be Brillinger (1990), who used a kernel smoothing algorithm of which the methods used here are a direct extension. More recently, Li et al. (2012) specify a fully parametric log-Gaussian Cox process (or LGCP) model for the case locations and use a Markov Chain Monte Carlo (MCMC) data-augmentation algorithm for inference where the locations are an unobserved latent variable, an approach improved and extended by Taylor et al. (2015). The methodology used by Banerjee et al. (2014) specifies a latent BYM model in place of a latent LGCP, where the intensity is piecewise constant on a set of regions nested within the boundaries of the regions on which data are observed.

A different approach involving spatially aggregated data began with Prentice and Sheppard (1995), and modelled the intensity of events (or probability with which events are cases or controls) as log-linear functions of spatial covariates. Location uncertainty in this context is manifested as uncertainty as to which values of the covariates are ascribe to each case, and Best et al. (2000) and Huang et al. (2014) are amongst the papers advancing methodology in this area. The difference between the two sets of approaches is that the work in the vein of Prentice and Sheppard (1995) is primarily concerned with estimating the contribution to disease risk from specific covariates of interest. In contrast, Brillinger (1990) and the Bayesian MCMC papers are concerned with explicit estimation and inference for spatial variation not explained by modelled covariates.

Here the local Expectation Maximization (local-EM) methods of Fan et al. (2011) and Nguyen et al. (2012) are extended from purely spatial models to allow for spatio-temporal variation in the latent risk surface. A simulation-based test for comparing the spatio-temporal model to a spatially-varying temporally-constant model is developed. Furthermore, an explicit structure for the presence of multiple types of aggregation in a dataset (mixtures of locations, disjoint regions, and irregular,