Connected Health Cities – End of Project Report

Workforce Development:

A spatially discrete approximation to log-Gaussian Cox processes for modelling aggregated disease count data

Contents:

30th March 2020

Olatunji Johnson, Peter Diggle, Emanuele Giorgi

Abstract:

In this project, we develop a computationally efficient discrete approximation to log-Gaussian Cox process (LGCP) models for the analysis of spatially aggregated disease count data.

Our approach overcomes an inherent limitation of spatial models based on Markov structures, namely, that each such model is tied to a specific partition of the study area, and allows for spatially continuous prediction.

We compare the predictive performance of our modelling approach with LGCP through a simulation study and an application to primary biliary cirrhosis incidence data in Newcastle upon Tyne, UK.

Our results suggest that, when disease risk is assumed to be a spatially continuous process, the proposed approximation to LGCP provides reliable estimates of disease risk both on spatially continuous and aggregated scales.

The proposed methodology is implemented in the open-source R package SDALGCP.

Keywords: disease mapping, geostatistics, log-Gaussian Cox process, Monte Carlo maximum likelihood

30th March 2020
Olatunji Johnson, Peter Diggle, Emanuele Giorgi

Introduction:

In this paper, our concern is to make inference on a spatially continuous disease risk surface using aggregated counts of reported disease cases, say, y_i, over regions i forming a partition of a geographical area of interest A.

In this context, information on risk factors and the population at risk may also be available, possibly at different spatial scales. We shall denote these by $d(x)$ and $m(x)$, respectively, when available on a spatially continuous scale, and by d_i and m_i when they are spatially aggregated.

Existing methods from small area estimation (SAE) only allow spatial prediction at the aggregated level of the regions i and are usually based on a Gaussian Markov random field (GMRF) structure.

Typically, nonzero elements of the precision matrix of a GMRF are restricted to contiguous pairs of the i.

Hence, the formulation and interpretation of a GMRF are tied to the specific partition of A, which will usually have been drawn up for administrative, historical, or other reasons unrelated to the disease aetiology.

The use of such models also becomes impractical when the spatial units i change over time. Wall[1] points out that the use of GMRFs is especially problematic when dealing with irregular geometries, which can induce counter-intuitive forms for the correlation structure between variables associated with the i.

The geostatistical paradigm, unlike SAE, treats disease risk as a spatially continuous phenomenon irrespective of the data format.

Diggle et al[2] argue that the analysis of spatially aggregated counts can be regarded as a special case of the class of geostatistical problems and propose to model the y_i as an aggregated realisation of a log-Gaussian Cox process (LGCP).

Unlike GMRFs, LGCPs allow for prediction of disease risk at any spatial scale, while avoiding the ecological fallacy. However, fitting of LGCP models using the aggregated counts y_i is

computationally demanding due to the iterative imputation of the unobserved locations for each reported case within a region i.

In this paper, our objective is to develop a computationally efficient approximation to LGCPs in order to predict disease risk at any desired spatial scale.

We argue that this provides a more realistic alternative to GMRF models when LGCPs are not computationally feasible and can also be used as an exploratory tool in order to inform more complex modelling approaches based on LGCPs.

The method has been implemented in the open-source R package SDALGCP[3], available from the Comprehensive R Network Archive.

Methods:

We developed a spatially discrete approximation to LGCP models in order to carry out spatial prediction of disease risk at any desired spatial scale using spatially aggregated disease count data.

The details of the methodology can be found in our paper published in Statistics in Medicine, titled "A spatially discrete approximation to log-Gaussian Cox processes for modelling aggregated disease count data". The link to the paper can be found in (https://onlinelibrary.wiley.com/doi/full/10.1002/sim.8339) (Johnson et al[4]).

3.0 APPLICATION: MAPPING OF PRIMARY BILIARY CIRRHOSIS RISK

In order to test our proposed method, we analyse incidence data on PBC in Newcastle upon Tyne, UK, the data set is freely available from the lgcp R package.

The data consist of geo-referenced cases of definite or probable PBC between 1987 and 1994. The objective of this analysis is to quantify the difference in the predictive inferences between the gold-standard LGCP model and the proposed SDA, on PBC incidence at LSOA level and the spatially continuous relative risk surface.

In the case of SDA, we fit the population weighted (SDA I) and simple average (SDA II) versions described in the previous section. We also consider the exponential variogram (EV) model proposed by Wall 3 consisting of a geostatistical Poisson model for the counts whose spatial structure is defined using the centroids of each LSOA.

Finally, we fit the Besag, York, and Mollié (BYM) model. In all five models, we use the index of multiple deprivation (IMD) as a covariate of the linear predictor.

The regression coefficients for the IMD are denoted by β_i in the LGCP model and by β_i in the BYM, EV, and SDA models, with i = 0 corresponding to the intercept and i = 1 the effect of IMD.Table 1 shows the estimates of the parameters of the model while Figure 1 shows the map of the estimated continuous relative risk surface exp{S(x)} over a 300 × 300 m regular grid covering the whole of the study area.

Conclusion/Discussion:

In this project, we have developed an SDA to LGCP models in order to carry out spatial prediction of disease risk at any desired spatial scale using spatially aggregated disease count data.

As variation in disease risk occurs in a spatial continuum irrespective of the format in which the data are available, we consider the LGCP framework to be a natural statistical paradigm for modelling aggregated disease count data.

However, when computational constraints make the fitting of an LGCP infeasible, we argue that SDA provides a computationally efficient solution while respecting the spatially continuous nature of disease risk.

SDA also overcomes some of the limitations inherent to other spatially discrete models, such as CAR models. In addition to providing spatially continuous predictions, SDAs can also deal with the issue of changing administrative boundaries over time and allow incorporation of covariates available at any spatial scale.

We conclude that SDA is a reliable approximation to LGCP for carrying out predictions at area level, both in terms of point predictions and in the quantification of uncertainty. It also provides spatially continuous predictions in disease risk that are comparable to those from LGCP, but with larger standard errors and more conservative predictions intervals.

References:

1. Wall MM. A close look at the spatial structure implied by the CAR and SAR models. J Stat Plan Inference. 2004;121(2):311-324.

2. Diggle PJ, Moraga P, Rowlingson B, Taylor BM. Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm. Statistical Science. 2013;28:542-563.

3. Johnson O, Giorgi E, Diggle P. SDALGCP: Spatially Discrete Approximation to Log-Gaussian Cox Processes for Aggregated Disease Count

Data. https://CRAN.R-project.org/package=SDALGCP. R package version 0.1.0; 2018

4. Johnson, O., Diggle, P., & Giorgi, E. (2019). A spatially discrete approximation to log-Gaussian Cox processes for modelling aggregated disease count data. *Statistics in medicine*, *38*(24), 4871-4887.

Author/Main Contact:

Olatunji Johnson, Peter Diggle, Emanuele Giorgi

30th March 2020

Olatunji Johnson, Peter Diggle, Emanuele Giorgi

Model	Parameter	Estimate	95% CI
SDA I	σ^2	1.043	(0.907, 1.180)
	ϕ	742.857	(453.153, 1005.405)
	β_0^*	−8.080	(−8.248, −7.912)
	β_1^*	0.008	(0.004, 0.011)
SDA II	σ^2	1.020	(0.898, 1.142)
	ϕ	857.143	(489.590 1037.638)
	β_0^*	−7.876	(−8.029, −7.722)
	β_1^*	0.006	(0.002, 0.010)
EV	σ^2	0.316	(0.246, 0.369)
	ϕ	525.570	(367.719, 949.950)
	β_0^*	−8.069	(−8.177, −7.957)
	β_1^*	0.009	(0.006, 0.011)
BYM	τ^2	0.108	(0.012, 0.470)
	ν^2	0.023	(0.003, 0.173)
	β_0^*	−7.917	(−8.167, −7.694)
	β_1^*	0.007	(0.001, 0.014)
LGCP	σ^2	0.479	(0.237, 0.914)
	ϕ	1163.877	(528.618, 1967.756)
	β_0	−19.333	(−19.738, −19.013)
	β_1	0.008	(0.001, 0.015)

TABLE 1 Point estimates and 95% confidence/credible intervals (CIs) for the model parameters of the spatially to log-Gaussian Cox process (LGCP) using a population-weighted log-intensity average (SDA

30th March 2020

Olatunji Johnson, Peter Diggle, Emanuele Giorgi

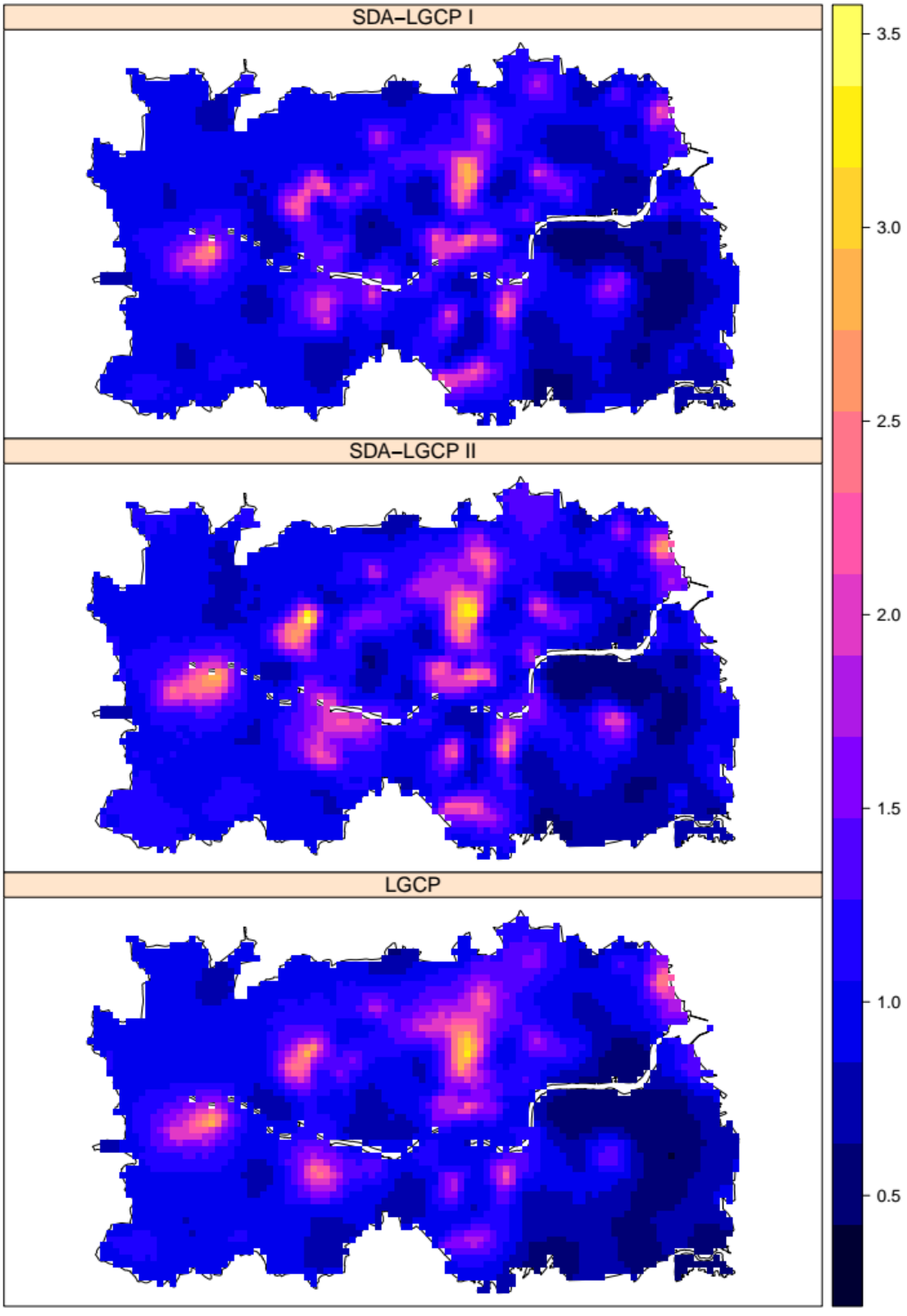
SDA–LGCP I
SDA–LGCP II
LGCP
3.5
3.0
2.5
2.0
1.5
1.0
0.5

FIGURE 1 Maps of the predicted relative risk surface exp{S(x)} from the fitted spatially discrete approximation (SDA) to log-Gaussian Cox process (LGCP) using a population-weighted log-intensity average (SDA I, upper panel) and a simple average (SDA II, middle panel), and the exact LGCP model (lower panel)