Spatial Statistics 21 (2017) 27–41

Contents lists available at ScienceDirect

Spatial Statistics

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

A geostatistical model for combined analysis of point-level and area-level data using INLA and SPDE

Paula Moraga [a,*], Susanna M. Cramb [a,b], Kerrie L. Mengersen [a,c], Marcello Pagano [d]

[a] ARC Centre of Excellence for Mathematical & Statistical Frontiers, Queensland University of Technology (QUT), Brisbane, Australia
[b] Cancer Council Queensland, Brisbane, Australia
[c] Cooperative Research Centre for Spatial Information, Australia
[d] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, United States

ARTICLE INFO

Article history:
Received 26 July 2016
Accepted 18 April 2017
Available online 1 June 2017

Keywords:
Misalignment
Data fusion
Model-based geostatistics
INLA
SPDE

ABSTRACT

In this paper a Bayesian geostatistical model is presented for fusion of data obtained at point and areal resolutions. The model is fitted using the INLA and SPDE approaches. In the SPDE approach, a continuously indexed Gaussian random field is represented as a discretely indexed Gaussian Markov random field (GMRF) by means of a finite basis function defined on a triangulation of the region of study. In order to allow the combination of point and areal data, a new projection matrix for mapping the GMRF from the observation locations to the triangulation nodes is proposed which takes into account the types of data to be combined. The performance of the model is examined and compared with the performance of the method RAMPS via simulation when it is fitted to (i) point, (ii) areal, and (iii) point and areal data to predict several simulated surfaces that can appear in real settings. The model is applied to predict the concentration of fine particulate matter ($PM_{2.5}$), in Los Angeles and Ventura counties, United States, during 2011.

* Correspondence to: School of Mathematical Sciences, Queensland University of Technology, GPO Box 2434, Brisbane, QLD 4001, Australia.
E-mail address: paula.moragaserrano@qut.edu.au (P. Moraga).
URL: https://Paula-Moraga.github.io/ (P. Moraga).

http://dx.doi.org/10.1016/j.spasta.2017.04.006
2211-6753/

1. Introduction

Spatial and spatio-temporal data arise in a wide range of scientific disciplines, including the environmental, epidemiological, geographical and ecological fields (Cressie, 1993). Data are typically observed either at points in space (point data), or over areal units such as counties or postal codes (areal data). Examples include air pollution measurements taken at a set of ambient stations, temperature and precipitation measurements from weather stations, and population sizes from census tracts. In epidemiology, point data arise when the locations at which cases of disease occur are available, and areal data are often reported when point data are aggregated over geographical subregions of the region of study due to ethical concerns over data use and patient confidentiality (Lawson, 2012).

Spatially misaligned data are becoming increasingly common due to advances in both data collection and management, as well as due to the ability to merge data from large databases such as disease registries. When information is available from multiple sources on different scales, data may be fused to examine just one variable, such as disease counts recorded in different administrative units. Here the aim is interpolation (Banerjee et al., 2014). Alternatively, we might wish to relate one variable to other variables that are available at different spatial resolutions and alignments. An example is determining whether the risk of an adverse outcome provided at zip level is related to exposure to an environmental pollutant measured at a network of stations, after adjusting for population at risk and other county level demographic information. Here the aim is regression (Banerjee et al., 2014).

In this paper we will focus on the data fusion problem which seeks to learn about a particular variable by combining data that are available at different spatial scales. Others have previously developed Bayesian models enabling fusion of data obtained at areal and point-referenced resolutions via the use of latent point-level processes (Fuentes and Raftery, 2005), hierarchical downscaling (Berrocal et al., 2010), modeling data conditional on the resolution (Wikle and Berliner, 2005), and the use of algorithms such as the reparameterized and marginalized posterior sampling (RAMPS) (Cowles et al., 2007).

The previous approaches use Bayesian predictive inference implemented via Markov chain Monte Carlo (MCMC) based methods. These methods have made a great impact on statistical practice by making Bayesian inference tractable for complex models but they also present a wide range of problems in terms of convergence and computational time (Taylor and Diggle, 2014). In this paper we propose general and flexible hierarchical Bayesian models to analyze spatially misaligned data. In order to fit the models, we resort to the Integrated Nested Laplace approximation (INLA) (Rue et al., 2009) and the Stochastic Partial Differential Equation (SPDE) (Lindgren et al., 2011) approaches which are a computationally effective alternative to MCMC for Bayesian inference. In order to allow the combination of data at different spatial resolutions, we propose a new projection matrix for mapping the GMRF in the SPDE method which takes into account how the different types of data are collected. This new approach is fast and flexible.

The outline of the paper is as follows. First, we present flexible models for handling spatial misaligned data in fusion problems. Then, we briefly introduce the INLA and SPDE approaches for Bayesian inference, and present the projection matrix that allows the combination of point and areal data. In Section 3, a simulation study is carried out to compare the performance of the model when estimating several simulated surfaces using point, areal, and point and areal data combined. Then, in Section 4 we evaluate the model in comparison to the RAMPS alternative method for data fusion by applying the methods to several simulated data scenarios. In Section 5, we present an application of the model to real data showing spatial misalignment. In this application, we obtain the spatial distribution of fine particulate matter ($PM_{2.5}$), in Los Angeles and Ventura counties, United States, during 2011. Finally, the conclusions are presented.

2. Models and inference

2.1. Models

The models proposed assume that there is a spatially continuous variable underlying all observations that can be modeled using a Gaussian random field process. This process is denoted by