Spatial Statistics 23 (2018) 72–90

Contents lists available at ScienceDirect

Spatial Statistics

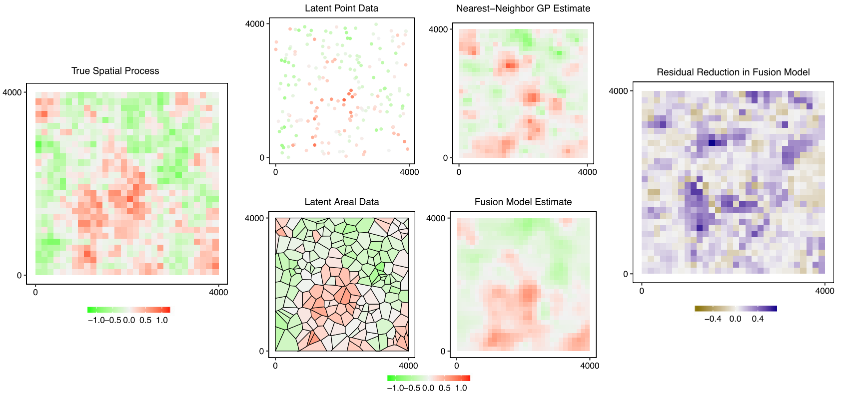
journal homepage: www.elsevier.com/locate/spasta

Generalized spatial fusion model framework for joint analysis of point and areal data

Craig Wang [a], Milo A. Puhan [b], Reinhard Furrer [a,c,*], for the SNC Study Group

[a] Department of Mathematics, University of Zurich, Zurich, Switzerland
[b] Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland
[c] Department of Computational Science, University of Zurich, Zurich, Switzerland

GRAPHICAL ABSTRACT



HIGHLIGHTS

- We propose a new generalized spatial fusion model framework.
- Our framework allows jointly analyzing non-Gaussian point and areal data.
- A simulation study shows fusion models improve prediction performance.

* Correspondence to: Department of Mathematics, University of Zurich, Winterthurerstrasse 190, CH-8057, Zurich Switzerland.

E-mail address: reinhard.furrer@math.uzh.ch (R. Furrer).

https://doi.org/10.1016/j.spasta.2017.11.006

- We identify situations where fusion models can lead to large improvements.
- A Gaussian–Poisson fusion model is applied to an epidemiological dataset.

ARTICLE INFO

Article history:
Received 6 October 2017
Accepted 22 November 2017
Available online 5 December 2017

Keywords:
Data fusion
Bayesian hierarchical model
Nearest neighbor Gaussian process
Change-of-support
Spatial epidemiology

ABSTRACT

The availability of geo-referenced data increased dramatically in recent years, motivating the use of spatial statistics in a variety of research fields, including epidemiology, environmental science, remote sensing, and economics. Combining data measured at both point and areal support can improve parameter estimation and increase prediction accuracy. We propose a new generalized spatial fusion model framework for jointly analyzing point and areal data. Assuming a common latent spatial process, we take a Bayesian hierarchical approach to model both types of data without distributional constraints. The models are implemented with nearest neighbor Gaussian process in Stan modeling language to increase computational efficiency and flexibility. Our simulation study shows that generalized fusion models under this framework model the latent process better than spatial process models. We identify scenarios where fusion models can offer large improvements. We then apply the framework to epidemiological data to identify the spatial risk pattern of respiratory diseases and lung cancer in Canton of Zurich, Switzerland.

1. Introduction

An increase in geo-referenced data spurred the use of spatial statistics in a variety of research areas, including epidemiology, environmental science, remote sensing, and economics (Gelfand et al., 2005; Shi and Cressie, 2007; Lawson et al., 2016; Paci et al., 2017). Often, researchers analyze a single spatial dataset, but a single source of spatial data may not be the best choice for parameter inference due to problems such as missing data and selection bias. For example, in remote sensing, cloud cover can interfere with regional observations. In disease mapping, data collection methods can cause selection bias in certain populations. In addition, there can be modeling difficulties, small sample size or weak spatial correlation may make it hard to estimate parameters (Irvine et al., 2007). In these situations, using multiple data sources can offer an advantage. Data may be collected at different spatial support for a variety of reasons, including budget constraints and privacy considerations. These data can be combined with the assumption of a common underlying spatial process in spatial fusion models. For example, air pollution modeling can be done based on measurements from monitoring stations, or numerical model output from computer simulations. We can assume that the same pollution process influences both measurements taken at the station, and the results of the simulation. Another example can be found in epidemiology, where, for privacy reasons, aggregated case counts at the district level are much more common than individual case locations of a disease. We can assume the same disease risk pattern drives the occurrence of individual cases and aggregated counts. As more database hosts and organizations collaborate, it is becoming easier to link datasets, providing more opportunities to carry out fusion tasks.

The approach of jointly analyzing multiple data sources that have different spatial support is called data fusion or data assimilation (Banerjee et al., 2014), or Bayesian melding (Fuentes and Raftery, 2005; Liu et al., 2011) in different literatures. They take a slightly different approach to modeling, but the basic idea is the same: combining point and areal data in a single statistical model. Fuentes and Raftery (2005) proposed one of the first fusion models, which predicts the spatial distribution of air pollution level. The model specifies both point-referenced measurements from monitoring