Spatial Statistics 18 (2016) 300–317

Contents lists available at ScienceDirect

Spatial Statistics

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

Emerging patterns in multi-sourced data modeling uncertainty

Alexander Kolovos [a,*], Lynette M. Smith [b], Aimee Schwab-McCoy [c], Sarah Gengler [d], Hwa-Lung Yu [e]

[a] SpaceTimeWorks, San Diego, CA, USA

[b] Medical Center, University of Nebraska, Omaha, NE, USA

[c] Department of Mathematics, Xavier University, Cincinnati, OH, USA

[d] Earth and Life Institute, Environmental Sciences. Université catholique de Louvain, Louvain-la-Neuve, Belgium

[e] Department of Bioenvironmental Systems Engineering, National Taiwan University, Taipei, Taiwan

ARTICLE INFO

Article history:
Received 20 December 2015
Accepted 29 May 2016
Available online 11 June 2016

Keywords:
Uncertainty
Poisson
Binomial kriging
Minimum norm approximations
Bayesian maximum entropy

ABSTRACT

The abundance of spatial and space–time data in many research fields has led to an increasing interest in the analytics of spatial data information. This development has renewed the attention to predictive spatial methodologies and advancing geostatistical tools. In this context, the present work reviews a series of cross-discipline studies that utilize multiple monitoring sources, and promote applied approaches in spatial and spatiotemporal modeling to improve our understanding of uncertainty. As multi-sourced information gives birth to new aspects of uncertainty, we explore emerging patterns in dealing with uncertainty in sources across structured, unstructured, and incomplete spatial data. We also illustrate how additional forms of information, such as secondary data and physical models, can further support and benefit research in the characterization and modeling of natural attributes.

* Corresponding author.
E-mail addresses: alexander.kolovos@spacetimeworks.com (A. Kolovos), lmsmith@unmc.edu (L.M. Smith), schwaba1@xavier.edu (A. Schwab-McCoy), sarah.gengler@uclouvain.be (S. Gengler), hlyu@ntu.edu.tw (H.-L. Yu).

http://dx.doi.org/10.1016/j.spasta.2016.05.005
2211-6753/

1. Introduction

Analysis of spatial and space–time data plays a prominent role in many research fields. The present work is a collection of cross-discipline spatial studies that help us illustrate approaches to address uncertainty found in multi-sourced information. Uncertainty typically stems from ontological (or aleatoric) causes and epistemic limitations. Our presentation takes a closer look at dealing with the latter category, where limitations are specifically related to the available informational content. In particular, epistemic uncertainty is explored from the following classification viewpoints:

a. Uncertainty in structured data due to a known data probabilistic distributional form.
b. Uncertainty in unstructured data.
c. Uncertainty due to lack of adequate data information.

In the following, each of the above categories is represented in context within application studies that adopt suitable methodologies to tackle a spatial problem effectively.

With respect to viewpoint (a) of structured data that follow a given probabilistic distribution, first we review a theoretical approach for prediction with variables that follow Poisson distribution in the presence of covariates. In public health and epidemiological settings such variables are very common and may represent disease incidence or mortality rates. Prediction is often assisted by auxiliary variables that represent relevant environmental factors such as temperature and precipitation, and we present a bivariate Generalized Linear Mixed Model approach for this type of analysis. Within the same viewpoint (a) we visit a different example in the health sciences where data are used for derivation of spatial proportions, and are considered as variables that follow a binomial distribution (Schwab and Marx, 2015). In this scenario, we present a kriging-based approach to model these proportions in the cases when an auxiliary variable is either present or not.

Viewpoint (b) represents a frequently occurring phenomenon of the digital age, where lack of structure is commonly observed in data collection through the internet (e.g., via sentiment analysis text, large-scale data collection from unspecified sources, etc.) We examine a case study of unstructured data collected through internet crowdsourcing (Bogaert and Gengler, 2014). Although inexpensive, crowdsourcing-based analysis can entail significant uncertainty due to the lack of data quality assurance. The analysis suggests crowdsourcing information coding can be enabled even in the absence of elementary quality assurance.

The last viewpoint (c) covers a broad range of spatial studies where analysis can be hampered by lack of observed or accurate data, thus contributing to prediction uncertainty. In one case from the field of atmospheric pollution studies, preferential sampling of particulate matter (PM) leaves a significant part of the domain of interest without input information about the pollutant (Yu et al., 2015). A solution comes in the form of available secondary background information that is associated to PM, and contributes with PM soft values to yield results throughout the prediction domain. In a second case study, we illustrate a framework to select optimal locations for the installation of solar energy stations (Zagouras et al., 2015). In a framework application in California, we see how modeling secondary satellite-derived soft data mitigates the lack of actual observations, and enables location selection with optimality based on physical coherence, geometrical, and geostatistical criteria. In both of the above examples, Bayesian maximum entropy provides the theoretical background to process uncertain information in a geostatistical context.

2. Uncertainty in structured data

2.1. Bivariate generalized linear mixed model prediction for Poisson data

Spatially correlated counts that follow a Poisson distribution are encountered often in public health settings to describe phenomena such as disease incidence and mortality rates. A particular case of interest is the West Nile virus (WNV), which is an arthropod-borne virus most commonly spread by infected mosquitoes, with most infections occurring during the summer months (Centers for Disease Control and Prevention, 2015). Being able to predict areas where WNV infection is likely to