Spatial Statistics 22 (2017) 89–106

Contents lists available at ScienceDirect

Spatial Statistics

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

Prediction and model comparison for areal unit data

Philip White [a,*], Alan Gelfand [a], Theresa Utlaut [b]

[a] Department of Statistical Science, Duke University, United States
[b] Intel, Hillsboro, OR, United States

ARTICLE INFO

Article history:
Received 2 May 2017
Accepted 4 September 2017
Available online 19 September 2017

Keywords:
Block average
CAR models
Disease mapping
Gaussian processes
Semiconductor chips
Ranked probability score

ABSTRACT

Areal unit or discrete spatial data is customarily modeled with the goal of spatial smoothing, typically using Markov random field models. Examples include image restoration and disease mapping. Here, we focus on a different issue for such data; we consider the set of areal units as only partially observed. One application is to learn about the smoothing behavior of various Markov random field models. That is, if two different smoothing priors are used, how can we quantify the relative smoothing that each imposes? We propose to fit models of interest to a portion of the data and hold out the rest for model comparison. A second application concerns the setting where, in fact, only a portion of the areal units have been observed, and we seek prediction of the remainder. Our motivating context investigates the performance of semiconductor chips, created as dies (the areal units) within wafers within lots, yielding nested modeling structure. Multiple tests are administered to each die involving both binary and continuous measurements. In practice, only a small subset of the dies are sampled, resulting in prediction of performance for the remaining unsampled dies. Furthermore, dies in the same locations are tested on each wafer, and the manufacturing process encourages within wafer, between wafer and between lot dependence. Other missing data applications include damaged images and small area estimation with missing observations for some units. We demonstrate prediction first with an image that is observed at several rates of missingness. Then, a well-studied Ohio lung cancer dataset is used for model comparison

* Corresponding author.
E-mail addresses: paw27@stat.duke.edu (P. White), alan@stat.duke.edu (A. Gelfand), theresa.l.utlaut@intel.com (T. Utlaut).

http://dx.doi.org/10.1016/j.spasta.2017.09.002

with regard to smoothing. Finally, examination of the nested modeling for semiconductor chip data is offered.

1. Introduction

We consider the situation of areal unit data, so-called discrete multivariate spatial data, where we have areal units over a region which are only partially observed, and we seek to infer about the unobserved units. Applications we envision include disease mapping, small area estimation, image analysis, and our motivating, more challenging application, performance of semiconductor chips. Customarily, with areal unit data, the objective is spatial smoothing (Banerjee et al., 2014), employing a complete dataset over the units; missingness is not a concern. With a goal of smoothing it is difficult to assess model performance (see Stern and Cressie, 1999, for early thoughts in this regard). Since visual assessment is qualitative, we might ask how one can quantify one smoothing relative to another? Moreover, under fitting to the full data, with no hold out data for validation, if model performance is assessed by comparison of predicted with observed, it will be impossible to outperform independent local (unit-level) estimation. Smoothing does not seek to minimize a goodness of fit criterion.

Here, we are interested in either of the following two scenarios. The first supposes that a substantial portion of our data is missing. For example, in the case of semiconductor chip data, we have the following setting. We have a *run* consisting of *lots* which are portions of a silicon ingot to which impurities are added in order to affect electrical properties. Each lot is sliced into thin *wafers*, and each wafer is partitioned into 195 areal units called *dies*, illustrated in Fig. 1. After production, a die is tested with respect to meeting measures of performance, e.g., speed, reliability, stress, power usage, in order to determine whether it is acceptable for use as a semiconductor chip. It is infeasible to test all of the dies in all of the wafers within a lot. In practice, performance is measured typically for only a subset of the dies but predictive inference is sought regarding performance for all of the dies on all of the wafers. In our examples, 20% of the dies are observed, see Fig. 1; however, the sampling rate can vary significantly with the application.

The second scenario focuses on model comparison for areal unit data. For instance, in the disease mapping context, we observe counts of disease cases across areal units (see, e.g., Clayton and Kaldor, 1987; Mollié et al., 1996; Green and Richardson, 2002; Lawson, 2013). Typically, spatial random effects are introduced using Markov random field (MRF) models in the form of a conditionally autoregressive (CAR) specification (see below) in the log mean for the counts. Model comparison would seek to compare the various CAR models that have been proposed in the literature (e.g. Besag, 1974; Besag et al., 1991; Leroux et al., 2000; Dean et al., 2001). As above, these models provide smoothing, here, of relative risks. With models fitted to the full set of counts, how can we decide which smoothing of the relative risks is preferred? Minimizing a predictive mean square error criterion will not be appropriate since we are not trying to *fit* the observed counts. Instead, using metrics such as predictive mean square error or rank probability scores (Gneiting and Raftery, 2007) with hold out data provides potentially useful out-of-sample *measures of smoothness*; their use in this context does not seem to be suggested in the literature. Illustratively, we might fit a given model to a portion of the units, selected at random, and predict for the remainder; we might do this several times to *average* over the randomness in the selection of fitting and validation units.

Modeling for areal unit data arises according to the nature of the data. For example, with count data for the units, as in disease mapping, we think only in terms of measurements at areal scales. There exists a conceptual count for any subregion/areal unit of the study area, but we do not imagine a count at a point, i.e., there is no point-referenced surface of counts. Similarly, if we collect proportions as the data for the units, there is no proportion at a point. Such settings result in finite dimensional model specifications with MRFs providing the customary modeling (e.g. Rue and Held, 2005; Banerjee et al., 2014).

Alternatively, we can imagine areal unit measurements arising as averages of a surface over a region. Such a surface is customarily viewed as a realization of a stochastic process, typically a