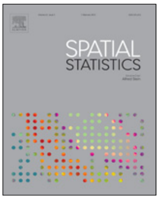
Spatial Statistics 14 (2015) 439–451

Contents lists available at ScienceDirect

Spatial Statistics

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



Bayesian marked point process modeling for generating fully synthetic public use data with point-referenced geography

Harrison Quick[a,*], Scott H. Holan[b], Christopher K. Wikle[b], Jerome P. Reiter[c]

[a] Division of Heart Disease and Stroke Prevention, Centers for Disease Control and Prevention, Atlanta, GA 30329, United States

[b] Department of Statistics, University of Missouri-Columbia, 146 Middlebush Hall, Columbia, MO 65211-6100, United States

[c] Department of Statistical Science, Box 90251, Duke University, Durham, NC 27708-0251, United States

ARTICLE INFO

Article history:
Received 7 March 2015
Accepted 20 July 2015
Available online 13 August 2015

Keywords:
Dimension reduction
Disclosure
Marked point processes
Multiple imputation
Privacy
Predictive process

ABSTRACT

Many data stewards collect confidential data that include fine geography. When sharing these data with others, data stewards strive to disseminate data that are informative for a wide range of spatial and non-spatial analyses while simultaneously protecting the confidentiality of data subjects' identities and attributes. Typically, data stewards meet this challenge by coarsening the resolution of the released geography and, as needed, perturbing the confidential attributes. When done with high intensity, these redaction strategies can result in released data with poor analytic quality. We propose an alternative dissemination approach based on fully synthetic data. We generate data using marked point process models that can maintain both the statistical properties and the spatial dependence structure of the confidential data. We illustrate the approach using data consisting of mortality records from Durham, North Carolina.

Published by Elsevier B.V.

* Corresponding author.
E-mail address: HQuick@cdc.gov (H. Quick).

http://dx.doi.org/10.1016/j.spasta.2015.07.008
2211-6753/Published by Elsevier B.V.

1. Introduction

Many statistical agencies, research centers, and individual researchers – henceforth all called agencies – collect confidential data that they intend to share with others as public use files. Many agencies are also obligated ethically, and often legally, to protect the confidentiality of data subjects' identities and sensitive attributes. This can be particularly challenging for agencies seeking to include fine levels of geography, e.g., street addresses or tax parcel identifiers, in the public use files. While detailed spatial information offers enormous benefits for analysis, it also can enable ill-intentioned users – henceforth called intruders – to easily identify individuals in the file.

Because of these confidentiality risks, agencies typically alter geographies and sensitive attributes before disseminating public use files. Perhaps the most common redaction method is to aggregate geographies to high levels like states (or not to release geography at all). Unfortunately, aggregation sacrifices analyses that require finer geographic detail and potentially creates ecological fallacies (Wang and Reiter, 2012). Furthermore, when the file includes other variables known to intruders like demographic information, aggregation alone may not suffice to protect confidentiality. Another strategy is to move each record's observed location to another randomly drawn location, e.g., within some circle of radius r centered at the original location. When large movements are needed to protect confidentiality – as can be the case when released data include demographic and other variables possibly known by intruders – inferences involving spatial relationships can be seriously degraded (Armstrong et al., 1999; VanWey et al., 2005). Suppression and aggregation also are commonly used to redact non-geographic attributes (Willenborg and de Waal, 2001; Hundepool et al., 2012), as are perturbative methods like data swapping (Dalenius and Reiss, 1982) and adding noise to values (Fuller, 1993). When applied with high intensity, these methods can result in files having poor analytic quality without adequate confidentiality protection (Winkler, 2007; Holan et al., 2010; Drechsler and Reiter, 2010).

An alternative to aggregation and perturbation is to release multiply-imputed synthetic data, in which confidential values are replaced with draws from statistical models designed to capture important distributional features in the collected data. Synthetic data come in two flavors. Partially synthetic data comprise the original units surveyed with some collected values replaced with multiple imputations (Little, 1993; Kennickell, 1997; Abowd and Woodcock, 2004; Reiter, 2003, 2004; An and Little, 2007; Toth, 2014), and fully synthetic data comprise entirely simulated records (Rubin, 1993; Reiter, 2002, 2005a; Raghunathan et al., 2003). In this article, we focus on fully synthetic data; see Reiter and Raghunathan (2007) for a review of the differences in the two flavors. Fully synthetic data can offer low disclosure risks as the released data cannot be meaningfully matched to external databases, while allowing secondary analysts to make valid inferences for wide classes of estimands via standard likelihood-based methods (Raghunathan et al., 2003; Reiter, 2005b).

Many of the existing approaches for generating synthetic data have been used primarily for data with no (or highly aggregated) geographical information (e.g., Hawala, 2008; Kinney et al., 2011) or with moderately aggregated geography like block groups or areal regions (e.g., Machanavajjhala et al., 2008; Burgette and Reiter, 2013; Paiva et al., 2014), where the goal is to estimate a model that predicts areal units from individuals' attributes, and then to assign each synthetic individual to an areal unit based on their attributes. These areal modeling approaches, however, do not apply when the goal is to release non-aggregated, point-referenced geography, although Paiva et al. (2014) make an *ad hoc* suggestion to randomly assign each individual to a point within its synthetic areal region. One exception is the work of Wang and Reiter (2012), who proposed that agencies treat latitude and longitude just like other continuous variables, approximating their conditional distributions given non-synthesized variables and releasing simulated locations by sampling from the models. They use regression trees (Reiter, 2005c) to approximate the conditional distributions. They also use trees to synthesize attributes conditional on latitude and longitude, treating the geographies as predictors in the tree models.

While the approach of Wang and Reiter (2012) is computationally efficient, it may not preserve local spatial dependence in the confidential data. To do so more effectively, we propose to take advantage of models developed specifically for point patterns – in particular models for marked point processes (e.g., Liang et al., 2009; Taddy and Kottas, 2012) – to generate fully synthetic data with locations and attributes. In marked point process modeling, there are two general approaches to