Bayesian Marked Point Process Modeling for Generating Fully Synthetic Public Use Data with Point-Referenced Geography

Harrison Quick
Joint work with Scott Holan, Chris Wikle, and Jerry Reiter

Postdoctoral Researcher, University of Missouri

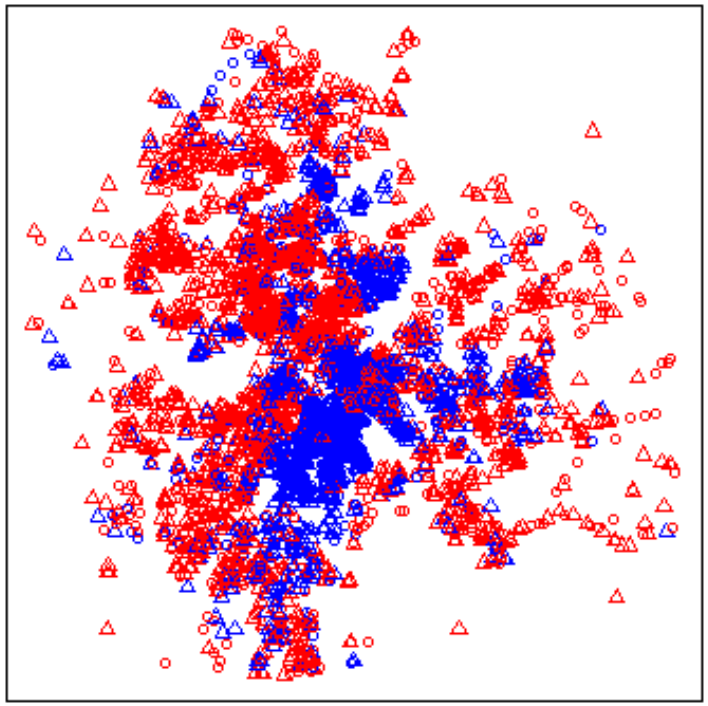
Sept. 12, 2014

Table of Contents

The Problem

Suppose you want to release data to the public with the following information:

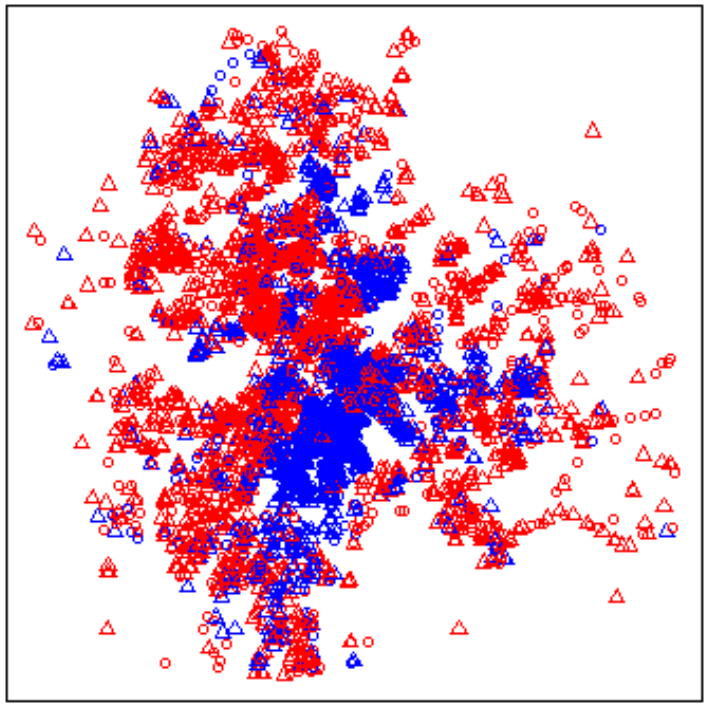
- Race
- Gender
- Income
- Spatial location



The Problem

Suppose you want to release data to the public with the following information:

- Race
- Gender
- Income
- Spatial location
- Sensitive information like disease status



- Releasing these data may present a disclosure risk

Some Current Options

- Release data at the aggregate level (e.g., county, state, etc.)
  - Useful for summarizing, but not individual-level, no fine spatial resolution
- Release synthetic (i.e., fake) data
  - Non-spatial: Plenty of options, including CART and model-based methods
  - Spatial: Spatial components are typically modeled simply as additional covariates (e.g., areal data modeled as a categorical covariate, etc.)...

Some Current Options

- Release data at the aggregate level (e.g., county, state, etc.)
  - Useful for summarizing, but not individual-level, no fine spatial resolution
- Release synthetic (i.e., fake) data
  - Non-spatial: Plenty of options, including CART and model-based methods
  - Spatial: Spatial components are typically modeled simply as additional covariates (e.g., areal data modeled as a categorical covariate, etc.)...
    - ... but do these methods preserve the spatial dependencies in the data?

Our solutions

- Want individual-level microdata with exact spatial locations and without the inherent disclosure risks

- Want to preserve spatial dependencies in the real data

Our solutions

- Want individual-level microdata with exact spatial locations and without the inherent disclosure risks
  - Generate L fully synthetic datasets
  - Say, $L = 10$ or $L = 100$
- Want to preserve spatial dependencies in the real data

Our solutions

- Want individual-level microdata with exact spatial locations and without the inherent disclosure risks
  - Generate L fully synthetic datasets
  - Say, $L = 10$ or $L = 100$
- Want to preserve spatial dependencies in the real data
  - Use a marked point process model
  - This will allow us to have categorical and non-categorical marks, as well as spatial locations, for each observation in our synthetic data

General strategy

1. Fit a hierarchical model to our data
   - We present a general framework here, but the exact model used will depend on what is most appropriate for a given dataset
2. Use posterior predictive distribution to generate fully synthetic datasets

Table of Contents

Components of marked point process model

We will build our marked point process sequentially using the following components

- Categorical marks
  - e.g., race, gender, etc.
- Spatial locations (given categories)
- Non-categorical marks (given spatial locations and categories)
  - e.g., income, age, etc.

Categorical mark model

Suppose our data consist of a total of K combinations of categorical marks, e.g., race, gender, etc., and let n_k denote the number of individuals belonging to the k-th mark group, $k = 1, \ldots, K$. A natural choice for modeling the vector $\mathbf{n} = (n_1, \ldots, n_K)'$ is a multinomial random variable

$$\mathbf{n} \,|\, N, \boldsymbol{\theta} \sim Mult(N, \boldsymbol{\theta}),$$

where $N = \sum_k n_k$ denotes the total sample size and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)'$ denotes the vector of group probabilities.

Model for spatial locations

To model locations, we take an approach similar to that of Liang et al. (2009)

- Locations modeled as arising from a log-Gaussian Cox process
- Integral approximated using numerical integration

$$LGCP\left\{\mathcal{S}_k \mid \lambda_k(\cdot), n_k\right\} = \exp\left\{-\int_D \lambda_k(\mathbf{s})d\mathbf{s}\right\}\prod_{i=1}^{n_k}\lambda_k(\mathbf{s}_{i,k})$$

$$\approx \exp\left\{-\frac{|D|}{T}\sum_{i=1}^{n_{ni}}\lambda_k(\mathbf{s}_{i,ni})\right\}\prod_{i=1}^{n_k}\lambda_k(\mathbf{s}_{i,k}),$$

where

- $\lambda_k(\mathbf{s}) = \exp\left[\mathbf{x}_\lambda(\mathbf{s})'\boldsymbol{\beta}_{\lambda\,|\,k} + w_{\lambda\,|\,k}(\mathbf{s})\right]$ is the intensity surface for category k
- $\mathbf{x}_\lambda(\mathbf{s})$ is a vector of spatial predictors with a corresponding vector of regression coefficients, $\boldsymbol{\beta}_{\lambda\,|\,k}$
- $w_{\lambda\,|\,k}(\mathbf{s})$ is a spatial random effect

Model for spatial locations, cont.d

For large values of n_{ni} (the number of integration points), this can be computationally burdensome. As such, we've used the *predictive process* model of Banerjee et al. (2008) to reduce the dimension to a more manageable level

$$LGCP\left\{\mathcal{S}_k \mid \lambda_k(\cdot), n_k\right\} = \exp\left\{-\int_D \lambda_k(\mathbf{s})d\mathbf{s}\right\}\prod_{i=1}^{n_k}\lambda_k(\mathbf{s}_{i,k})$$
$$\approx \exp\left\{-\frac{|D|}{T}\sum_{i=1}^{n_{ni}}\widetilde{\lambda}_k(\mathbf{s}_{i,ni})\right\}\prod_{i=1}^{n_k}\widetilde{\lambda}_k(\mathbf{s}_{i,k}),$$

where $\widetilde{\lambda}_k(\mathbf{s}_{i,k})$ denotes the low-rank approximation of the intensity surface corresponding to a predictive process model on $w_{\lambda\,|\,k}(\mathbf{s})$

Non-categorical marks model

Given our categories and our locations, we can model non-categorical marks using standard approaches from the geostatistical literature. For instance,

$$Y_k(\mathbf{s}) \sim N(\mathbf{x}_Y(\mathbf{s})'\boldsymbol{\beta}_{Y\,|\,k} + w_{Y\,|\,k}(\mathbf{s}), \sigma_k^2),$$

for a continuously varying response or

$$Y_k(\mathbf{s}) \sim Pois\left(\exp\left\{\mathbf{x}_Y(\mathbf{s})'\boldsymbol{\beta}_{Y\,|\,k} + w_{Y\,|\,k}(\mathbf{s})\right\}\right),$$

for a positive integer-valued variable such as age.

Again, we may choose to use a predictive process approach to achieve dimension reduction.

Table of Contents

Generating fully synthetic microdata

Having fit our hierarchical model and obtained posterior distributions for all of our model parameters, we can now generate our L synthetic datasets.
First, we assign our $N^{\dagger(\ell)}$ synthetic individuals to groups

$$\mathbf{n}^{\dagger(\ell)} \mid N^{\dagger(\ell)}, \boldsymbol{\theta}^{(\ell)} \sim Mult\left(N^{\dagger(\ell)}, \boldsymbol{\theta}^{(\ell)}\right),$$

where $\ell = 1, \ldots, L$. For the ease of notation, $\boldsymbol{\theta}^{(\ell)}$ denotes the ℓth post-burn-in sample from our posterior.

Generating fully synthetic microdata, cont.d

Next, we sample geographies for these individuals from

$$\mathcal{S}_k^{\dagger(\ell)} \,|\, \widetilde{\lambda}_k^{(\ell)}(\cdot), n_k^{\dagger(\ell)} \sim LGCP\left\{\widetilde{\lambda}_k^{(\ell)}(\cdot), n_k^{\dagger(\ell)}\right\}, \quad k = 1, \ldots, K.$$

For computational reasons, these points will be sampled *with replacement* — i.e., it is possible for multiple individuals to be at the same location.

Finally, we sample our non-categorical marks, say

$$Y_k^{\dagger(\ell)}(\mathbf{s}_i^{\dagger(\ell)}) \,|\, \mu_k^{(\ell)}(\mathbf{s}_i^{\dagger(\ell)}), \sigma_k^{2(\ell)} \sim N\left(\mu_k^{(\ell)}(\mathbf{s}_i^{\dagger(\ell)}), \sigma_k^{2(\ell)}\right), \quad k = 1, \ldots, K.$$

Evaluating utility of the synthetic data

- To investigate spatial dependence, we estimate the K and L functions

$$\widehat{K}(h) = \frac{|D_s|}{N} \sum_{\substack{i=1 \\ i \neq j}}^{N} \sum_{j=1}^{N} I(||\mathbf{s}_i - \mathbf{s}_j|| \leq h)/N$$

$$\widehat{L}(h) = \sqrt{\widehat{K}(h)/\pi} - h,$$

  where $|D_s|$ is the area of spatial domain.

- We also make comparisons of parameter estimates from various statistical models of interest

Assessing disclosure risk in synthetic data

First we define two terms/expressions:

- *spatially close*: Two locations are said to be spatially close if the distance between them is less than some ϵ_s; denoted $\sim \mathbf{s}_0$
- *similar attributes*: Two individuals are said to have similar attributes if they belong to the same mark categories and their non-categorical marks are within ϵ_a of each other; denoted $\sim Y_k(\mathbf{s}_0)$

Assessing disclosure risk in synthetic data, cont.d

- "Type S Risk": If you know the spatial location, how well can you estimate the attributes?

$$P(\sim Y_k(\mathbf{s}_0) \mid \sim \mathbf{s}_0) \approx p_s^{\dagger(\ell)}(Y_k(\mathbf{s}_0), \mathbf{s}_0)$$
$$= \frac{\sum_i I\{Y_k^{\dagger(\ell)}(\mathbf{s}_i^{\dagger(\ell)}) \sim Y_k(\mathbf{s}_0) \mid \mathbf{s}_i^{\dagger(\ell)} \sim \mathbf{s}_0\}}{\sum_i I\{\mathbf{s}_i^{\dagger(\ell)} \sim \mathbf{s}_0\}}$$

for $\ell = 1, \ldots, L$

- "Type A Risk": If you know the attributes, how well can you estimate the spatial location?

$$P(\sim \mathbf{s}_0 \mid \sim Y_k(\mathbf{s}_0)) \approx p_a^{\dagger(\ell)}(Y_k(\mathbf{s}_0), \mathbf{s}_0)$$
$$= \frac{\sum_i I\{\mathbf{s}_i^{\dagger(\ell)} \sim \mathbf{s}_0 \mid Y_k^{\dagger(\ell)}(\mathbf{s}_i^{\dagger(\ell)}) \sim Y_k(\mathbf{s}_0)\}}{\sum_i I\{Y_k^{\dagger(\ell)}(\mathbf{s}_i^{\dagger(\ell)}) \sim Y_k(\mathbf{s}_0)\}}$$

Table of Contents

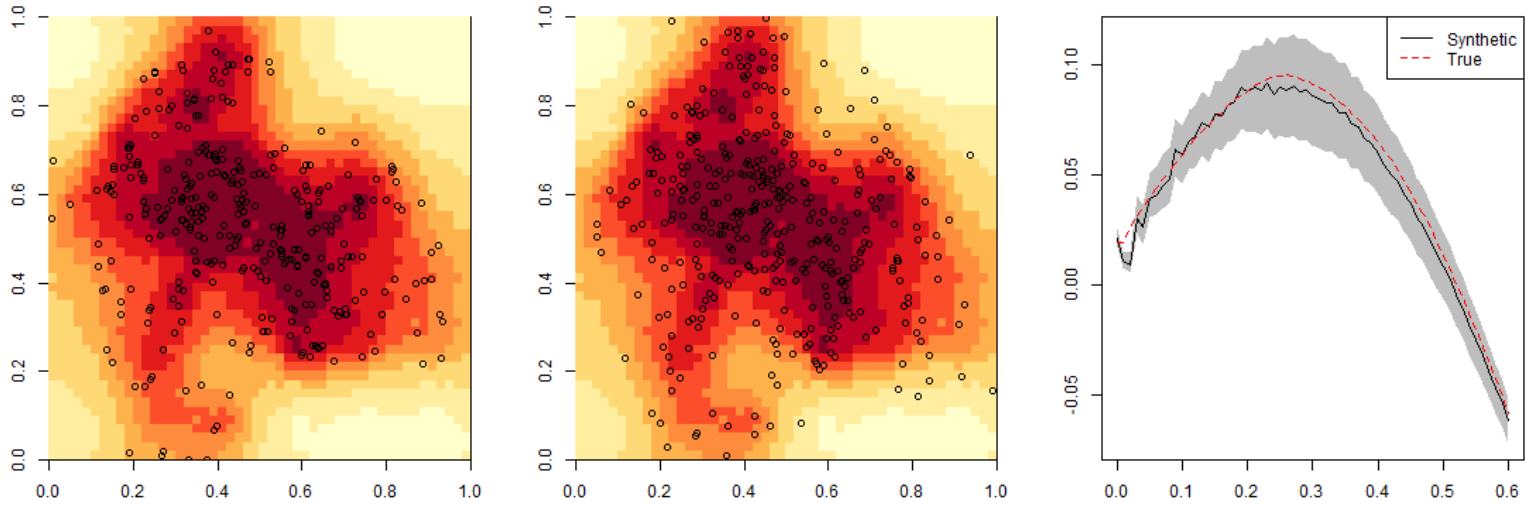
Durham, NC mortality data

These data consist of addresses for $N = 6294$ individuals with the following attributes

- Race (black/white)
- Gender
- Cause of death (cancer-related vs. other)
- Education ($<$ HS, HS, some college)
- Age

We model our data as having 24 categorical groups and assume a (truncated) Poisson distribution for age. Based on these data, we will generate $L = 100$ synthetic datasets, each comprised of N synthetic individuals

Comparison of spatial dependencies



(a) True Locations (b) Synthetic Locations (c) L Function

Figure 1: Panels (a) and (b) display locations for one group of observations from the true and a synthetic dataset, respectively, both overlayed on our estimated intensity surface. Panel (c) displays $\widehat{L}(h)$ over a range of values of h for our true (dashed red line) and synthetic data (black line), along with the pointwise empirical 95% CI from the synthetic data (gray band).

Comparison of parameter estimates

We may also wish to compare parameter estimates from the following models:

- Poisson Regression:

$$\text{Age}_i \sim \textit{Pois}(\lambda_i) \text{ where } \log(\lambda_i) = \sum_{j=1}^{24} \beta_{1j} I(\text{Group}_i = j),$$

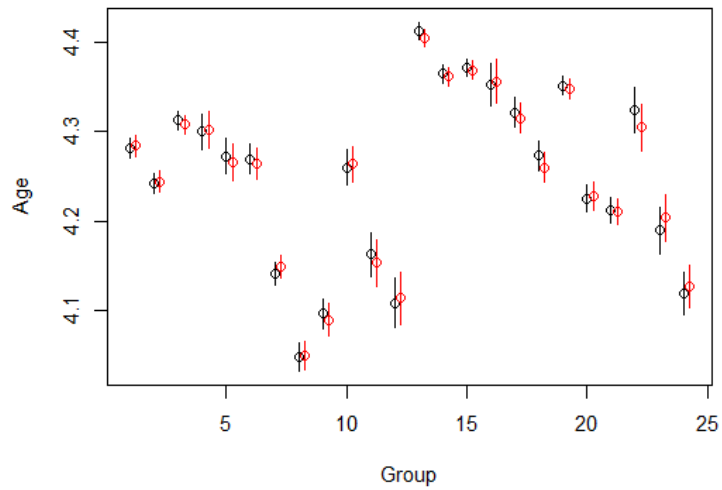
and Group_i is an integer between 1 and 24 denoting which combination of race, gender, education level, and cause of death the individual belongs to.

- Logistic Regression:

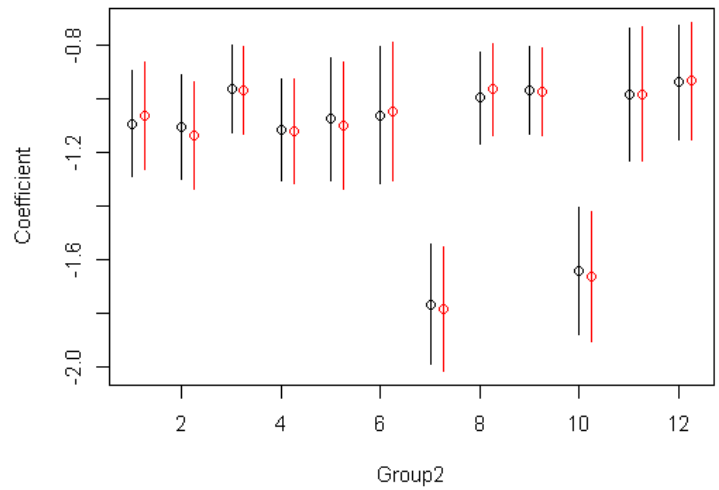
$$\text{CoD}_i \sim \textit{Ber}(\pi_i) \text{ where } \text{logit}(\pi_i) = \sum_{j=1}^{12} \beta_{2j} I(\text{Group2}_i = j)$$

and Group2_i be an integer between 1 and 12 denoting which combination of race, gender, and education level each individual belongs to.

Comparison of parameter estimates, cont.d



(a) Poisson Regression: Age



(b) Logistic Regression: Cause of Death

Figure 2: Comparison of parameter estimates from a Poisson regression on age and a logistic regression on the cause of death. Circles denote mean estimates while bars denote the 95% CI. Estimates from the real data are displayed in black, while estimates obtained from the synthetic data are displayed in red.

Risk Assessment

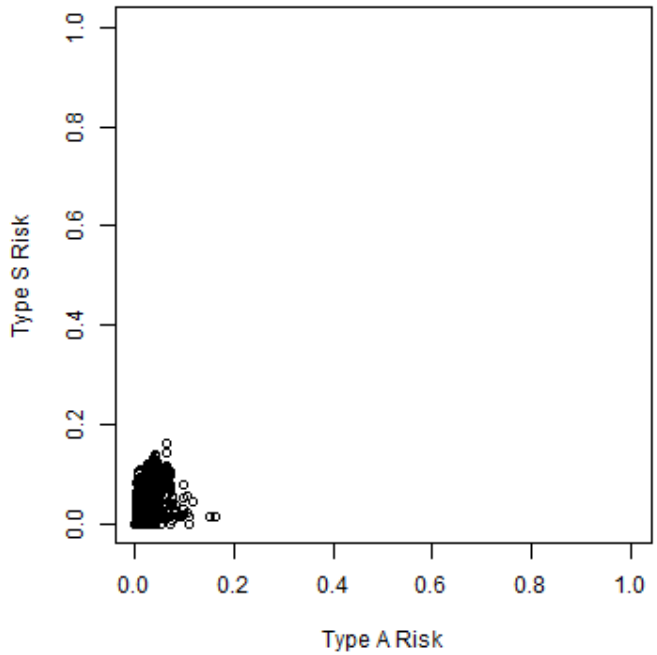


Figure 3: Comparisons of the estimated Type S Risk versus the estimate Type A Risk in our mortality example.

Table of Contents

Summary

- We have used a fully Bayesian framework to generate fully-synthetic microdata for the purpose of disclosure limitation
- While not explicitly discussed here, we believe our framework can be quite computationally efficient by taking advantage of dimension reduction techniques as well as parallel computing
- Based on the results from our mortality dataset, our synthetic data offer a high degree of data utility without a substantial disclosure risk

Thanks!

My collaborators:

- Scott Holan (Missouri)
- Chris Wikle (Missouri)
- Jerry Reiter (Duke)

This research was partially supported by the U.S. National Science Foundation (NSF) and the U.S. Census Bureau under NSF grants SES-1132031 and SES-1131897, funded through the NSF-Census Research Network (NCRN) program.