Multivariate spatiotemporal modeling with applications to stroke mortality and data privacy

Harrison Quick (Drexel University)
Joint work with Lance Waller (Emory) and Michele Casper (CDC)

The findings and conclusions in this presentation are mine and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Table of Contents

Introduction

Methods

- Multivariate space-time CAR model
- Generation and evaluation of synthetic data

Results

- Analysis of the stroke mortality data
- Generation/Evaluation of synthetic data

Summary and Discussion

Table of Contents

Introduction

Methods

- Multivariate space-time CAR model
- Generation and evaluation of synthetic data

Results

- Analysis of the stroke mortality data
- Generation/Evaluation of synthetic data

Summary and Discussion

Goal for this Talk

The charge of agencies such as the CDC includes the following:

- Conduct surveillance into epidemiologic issues
  - e.g., develop/implement statistical models to better estimate and/or predict trends in the data
- Disseminate information (e.g., data) for public use
  - e.g., publishing articles/reports, release data via CDC WONDER
  - Must be cognizant of potential risks of disclosure when sharing information based on confidential/private data

The goal for this talk will be to develop a statistical framework which is useful for both of these charges.

Today's Example: Stroke Mortality

Background information on stroke mortality:

- Stroke is the fourth leading cause of death in the US
- Mortality rates increase exponentially with age
- Previous work has identified strong spatial patterns in stroke mortality (e.g., "the stroke belt")

Our data consists of the number of stroke deaths, Y_{ikt}, and the population size, n_{ikt}, from:

- $i = 1, \ldots, N_s$=3,099 counties (or county equivalents) from the contiguous United States
- $t = 1, \ldots, N_t$=41 years of data (1973 – 2013)
- US citizens ages 65 and older.
  - $k = 1, \ldots, N_g = 3$ age brackets (65–74, 75–84, 85+)

Because stroke mortality is quite rare, many of our $N_s \times N_g \times N_t = 381{,}177$ counts are quite small.

Data Dissemination Challenges

When releasing these data for public use, CDC WONDER uses NCHS's recommendation of *suppressing* instances where $Y_{ikt} < 10$

- Leads to nearly 70% of the data analyzed here being suppressed.

This has an impact on the types and quality of inference that outside researchers can conduct *using the public-use data*.

- Analyzing all 380,000+ observations would require censored data methods (or otherwise accounting for the missingness) — this is likely an unreasonable expectation.
- Others may restrict their analyses to counties in which complete data are available (i.e., urban centers), or aggregate spatially or across age to obtain larger counts.
- Analyses for more specific demographic groups are left unstudied (e.g., mortality rates by age/race/sex), as the issue will only be compounded.

Our Proposal

To obtain more reliable estimates from the data and to provide unrestricted access to high-quality public-use data, we propose the following:

1. Analyze the data using a Bayesian statistical model which accounts for (a) spatial structure, (b) temporal structure, and (c) between-age-group structure
   - To do so, we will use the multivariate space-time conditional autoregressive (MSTCAR) model of Quick et al. (2017).
2. Using the posterior distribution from the Bayesian model, we will generate multiply-imputed *synthetic* data to replace sensitive counts
   - The resulting synthetic data will preserve the complex spatial, temporal, and between-age dependencies (along with any covariate relationships) that we accounted for in our model.

Table of Contents

Introduction

Methods

- Multivariate space-time CAR model
- Generation and evaluation of synthetic data

Results

- Analysis of the stroke mortality data
- Generation/Evaluation of synthetic data

Summary and Discussion

Disease mapping — the univariate case

Following the convention set forth by Besag et al. (1991), we may assume

$$Y_{ikt} \mid \lambda_{ikt} \sim \text{Pois}\left(n_{ikt}\lambda_{ikt}\right) \text{where} \log \lambda_{ikt} \sim \text{Norm}\left(\mathbf{x}_{ikt}^T \boldsymbol{\beta}_{kt} + Z_{ikt}, \tau_k^2\right),$$

where

- $\mathbf{x}_{ikt}^T \boldsymbol{\beta}$ denotes a regression where $\mathbf{x}_{ikt}$ denotes a vector of county-level covariates
  - For this analysis, our covariates include % non-white and % male within each age group at each time period
- Z_{ikt} denotes a spatiotemporal random effect
- τ_k^2 denotes the variance of the log mortality rates

Conditional autoregressive (CAR) models

To induce spatial correlation in the random effects, Besag et al. (1991) assumed

$$Z_{ikt} \mid \mathbf{Z}_{(i)kt}, \sigma^2_{kt} \sim \text{Norm}\left(\sum_{j\sim i} Z_{jkt}/m_i, \sigma^2_{kt}/m_i\right)$$

$$\pi\left(\mathbf{Z}_{\cdot kt} \mid \sigma^2_{kt}\right) \propto \left(\sigma^2_{kt}\right)^{-(N_s-1)/2} \exp\left[-\frac{\mathbf{Z}^T_{\cdot kt}\left(D-W\right)\mathbf{Z}_{\cdot kt}}{2\sigma^2_{kt}}\right]$$

where

- $\mathbf{Z}_{(i)kt}$ is the vector $\mathbf{Z}_{\cdot kt} = \left(Z_{1kt}, \ldots, Z_{N_s kt}\right)^T$ with the ith element removed.
- $j \sim i$ denotes that counties i and j are neighbors.
- W is an adjacency matrix with $w_{ij} = 1$ if $j \sim i$ and $w_{ij} = 0$ otherwise.
  - $m_i = \sum_j w_{ij}$, the number of neighbors
  - D is a diagonal matrix with elements m_i
- σ^2_{kt} is an age/time-specific variance parameter.

Extension to multiple disease mapping

When modeling data from multiple diseases (or in our case, mortality rates for multiple age groups over time), a multivariate extension of the CAR model can be used (e.g., the multivariate CAR (MCAR) of Gelfand and Vounatsou, 2003).

$$\mathbf{Z}_{i\cdot\cdot} \,|\, \mathbf{Z}_{(i)\cdot\cdot}, \Sigma_Z \sim \text{Norm}\left(\sum_{j\sim i} \mathbf{Z}_{j\cdot\cdot}/m_i, \frac{1}{m_i}\Sigma_Z\right)$$

$$\pi\left(\mathbf{Z} \,|\, \Sigma_Z\right) \propto |\Sigma_Z|^{-(N_s-1)/2} \exp\left[-\frac{1}{2}\mathbf{Z}^T\left\{(D-W)\otimes\Sigma_Z^{-1}\right.\right. \quad \mathbf{Z}\Big],$$

where

- $\mathbf{Z}$ is a $N_s N_g N_t \times 1$ vector of spatiotemporal random effects which allows for correlation between age groups
- Σ_Z is the multivariate analog of σ^2 from the univariate case

Multivariate space-time model for $\mathbf{Z}$

Based on the MCAR of Gelfand and Vounatsou (2003),

$$\mathbf{Z}_{i\cdot\cdot} \,|\, \mathbf{Z}_{(i)\cdot\cdot}, \Sigma_Z \sim \text{Norm}\left(\sum_{j\sim i} \mathbf{Z}_{j\cdot\cdot}/m_i, \frac{1}{m_i}\Sigma_Z\right)$$

- Spatial associations are accounted for via the neighborhood structure in the mean and variance.
- Thus, Σ_Z can be thought of as a (scaled) covariance matrix which accounts for the multivariate and temporal dependencies in $\mathbf{Z}$.
  - We'll allow for differing degrees of temporal correlation within each each age-bracket, denoted by $\boldsymbol{\rho} = \left(\rho_1, \ldots, \rho_{N_g}\right)^T$.
  - Between age-bracket dependencies will be allowed to vary over time, denoted by $\mathcal{G} = \{\mathbf{G}_1, \ldots, \mathbf{G}_{N_t}\}$.

We denote this structure by $\mathbf{Z} \sim \text{MSTCAR}\,(\mathcal{G}, \boldsymbol{\rho})$.

Hierarchical model

Putting these pieces together, our full hierarchical model is as follows:

$$\begin{aligned}\pi\left(\boldsymbol{\beta}, \mathbf{Z}, \mathbf{G}, \mathcal{G}, \boldsymbol{\rho}, \left\{\tau_k^2\right\}, \boldsymbol{\lambda} \mid \mathbf{Y}\right) \propto & \prod_{i,k,t} \text{Pois}\left(Y_{ikt} \mid n_{ikt}\lambda_{ikt}\right) \\ & \times \prod_{i,k,t} \text{Norm}\left(\log \lambda_{ikt} \mid \mathbf{x}_{ikt}^T \boldsymbol{\beta}_{ikt} + Z_{ikt}, \tau_k^2\right) \\ & \times \text{MSTCAR}\left(\mathbf{Z} \mid \mathcal{G}, \boldsymbol{\rho}\right) \times \text{Norm}\left(\boldsymbol{\beta} \mid \mathbf{0}, \boldsymbol{\Sigma}_\beta\right) \\ & \times \prod_{t} \text{InvWish}\left(\mathbf{G}_t \mid \mathbf{G}, \nu\right) \times \text{Wish}\left(\mathbf{G} \mid \mathbf{G}_0, \nu_0\right) \\ & \times \prod_{k} \left[\text{Beta}\left(\rho_k \mid a_\rho, b_\rho\right) \times \text{IG}\left(\tau_k^2 \mid a_\tau, b_\tau\right)\right],\end{aligned}$$

where $\boldsymbol{\Sigma}_\beta = 100\mathbf{I}_{pN_gN_t}$ and $\mathbf{X}$ is the $(N_sN_gN_t \times p)$ matrix of covariates.

We fit this model using Markov chain Monte Carlo (MCMC) and obtain samples from the posterior distribution for each model parameter.

- e.g., $\lambda_{ikt}^{(1)}, \ldots, \lambda_{ikt}^{(L)}$, where L is the number of iterations

Synthetic data

Given our samples for λ_{ikt}, we can generate synthetic counts for our suppressed Y_{ikt} from a truncated Poisson of the form

$$Y_{ikt}^{*(\ell)} \mid \lambda_{ikt}^{(\ell)}, \{Y_{ikt} < 10\} \sim \text{Pois}\left(n_{ikt}\lambda_{ikt}^{(\ell)}\right) \times I\left\{Y_{ikt}^{*(\ell)} < 10\right\}.$$

If desired, this approach could be modified to preserve aggregate totals (e.g., state-level counts) which would be publicly available.

To assess the quality of these synthetic data, we will compare them to synthetic data that could be generated by fitting the MSTCAR model to the publicly available (i.e., *suppressed*) data.

- Counts below 10 will be imputed as part of the model
- We consider this to be the best available alternative for both public users and for ill-intentioned users (or "intruders")

Measuring disclosure risk and utility

- Disclosure risk will be computed as

$$P\left(Y^*_{ikt} = y \,|\, \mathbf{Y}, Y_{ikt} = y\right) \text{ for } y = 0, 1, \ldots, 9.$$

  In particular, we will look at the risk when $y = 1$ (the value we're most concerned about).

- Utility will be compared by fitting a model of the form

$$Y_{ikt} \sim \text{Pois}\left(n_{ikt} \exp\left[\gamma_{0kt} + \text{rural}_{ikt}\gamma_{1kt}\right]\right),$$

  where rural_{ikt} denotes a 0/1 variable taking value 1 if county i has a population (across all age groups) less than 50,000 during year t.

  - Estimates from synthetic data will also be compared to the estimates from the confidential data (i.e., the "truth").

Table of Contents

Introduction

Methods

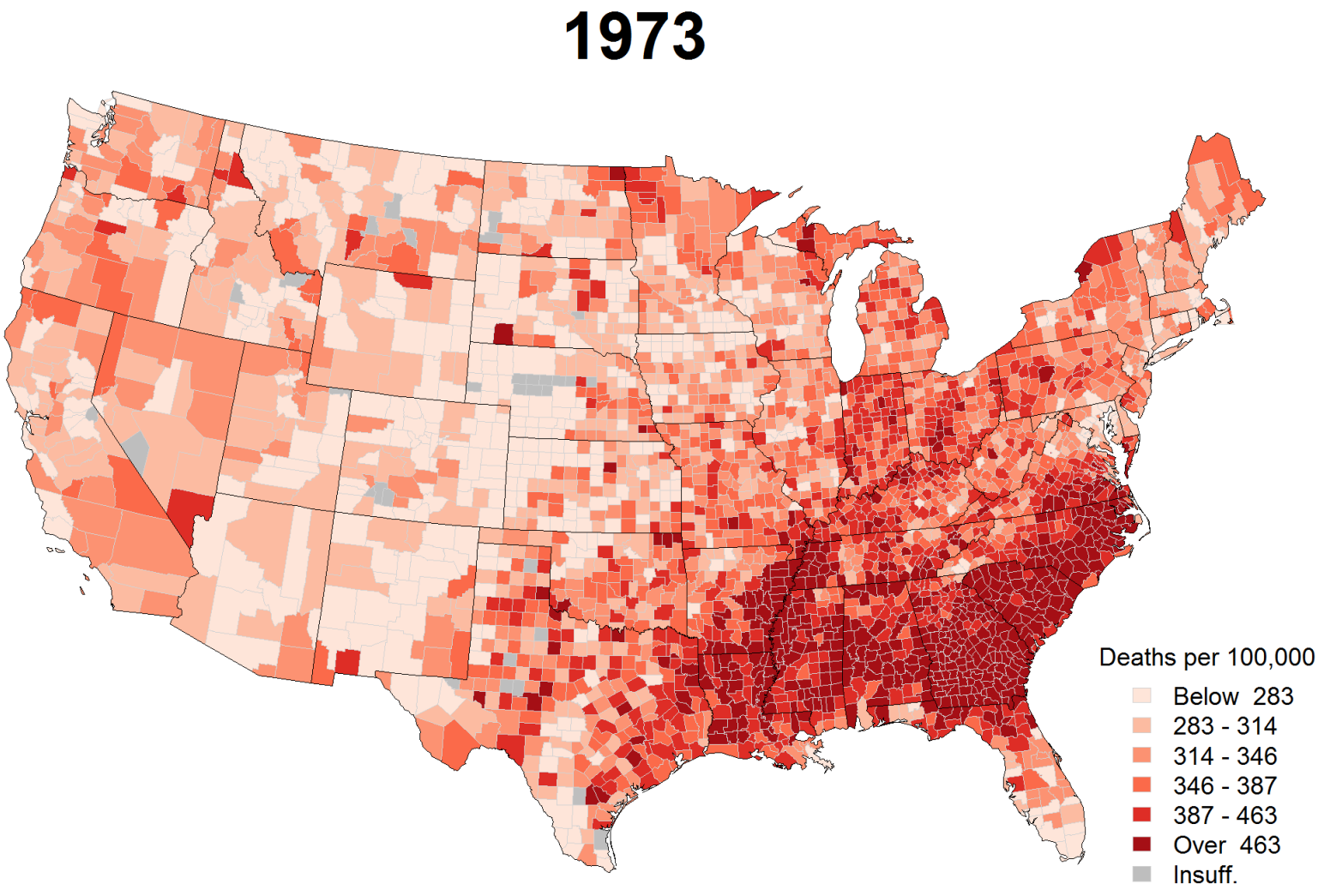
- Multivariate space-time CAR model
- Generation and evaluation of synthetic data

Results

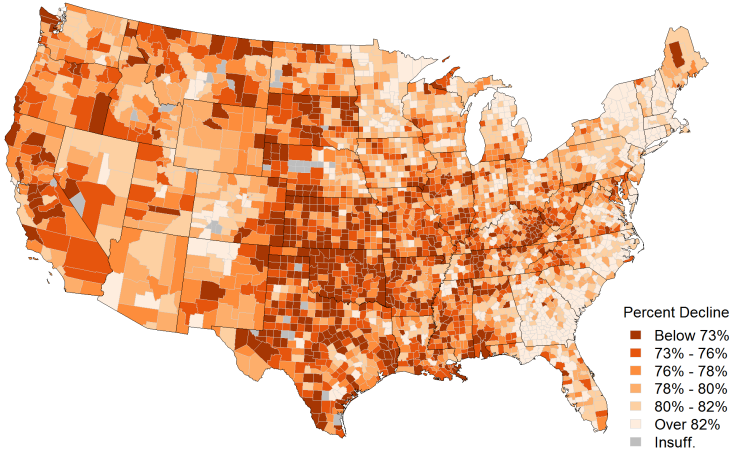
- Analysis of the stroke mortality data
- Generation/Evaluation of synthetic data

Summary and Discussion

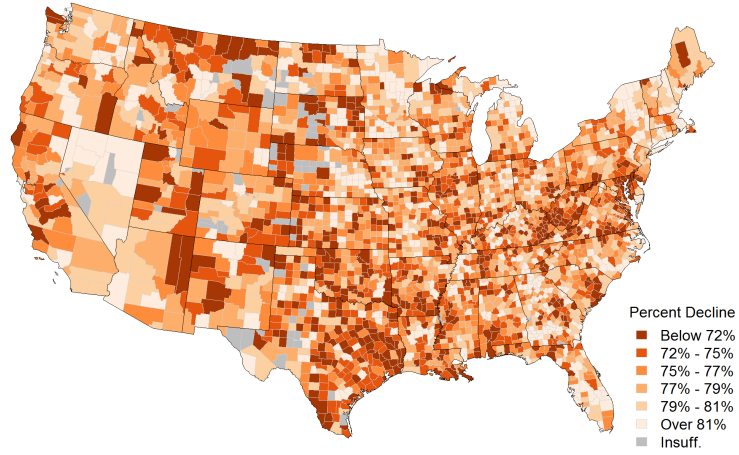
Stroke mortality: ages 65–74



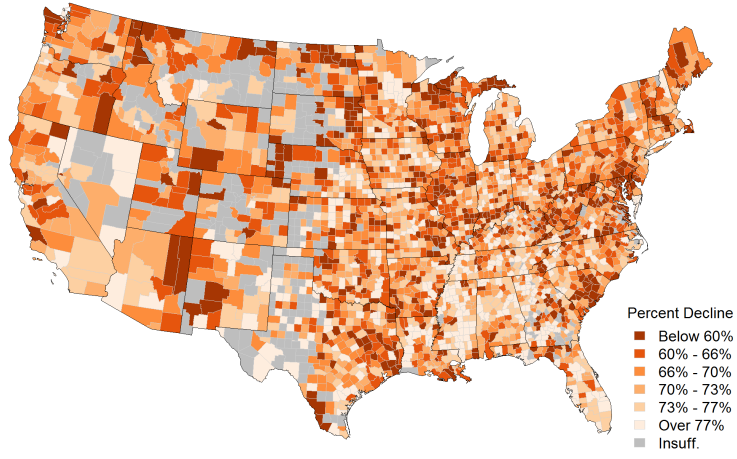
Overall declines in stroke mortality



(a) Ages 65–74

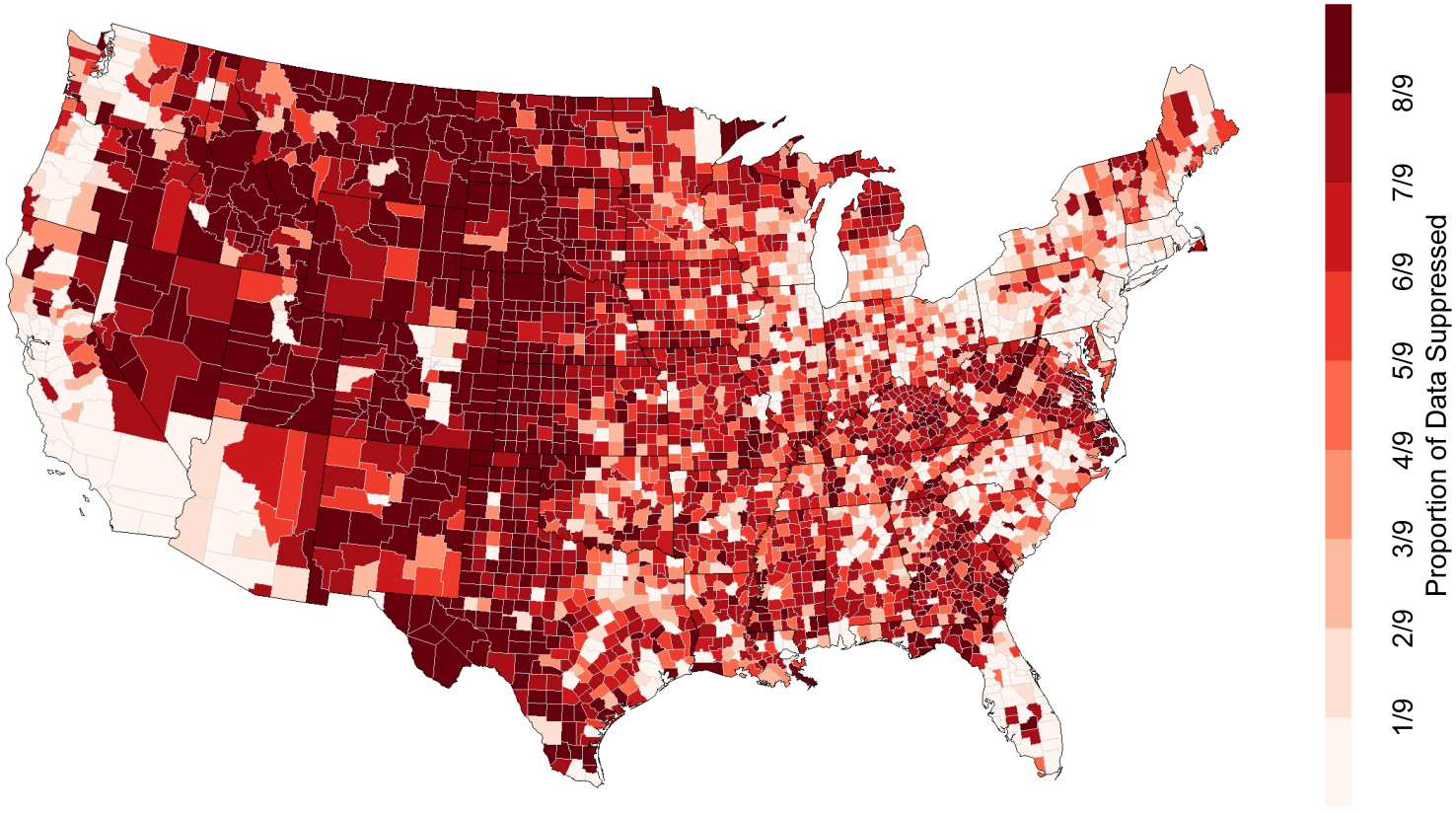


(b) Ages 75–84

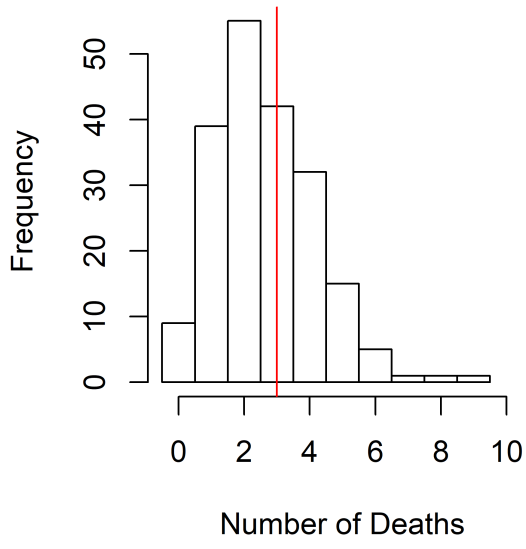


(c) Ages 85+

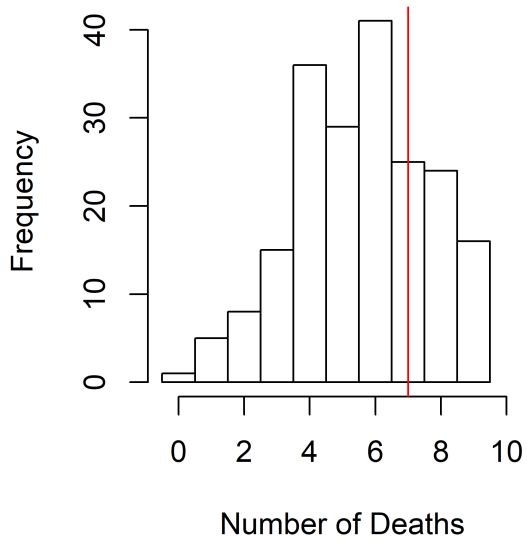
How much of these data are suppressed to the public?



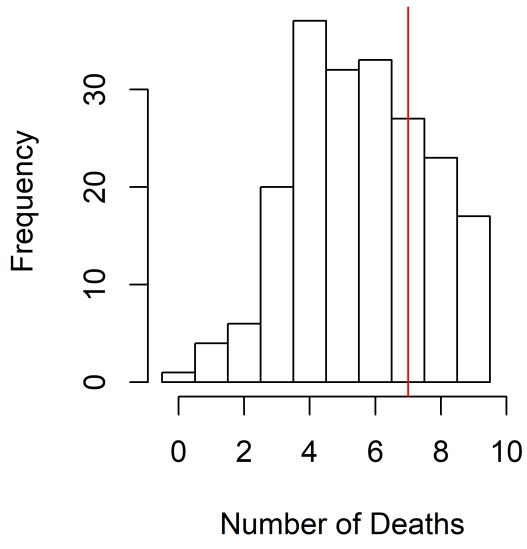
Example: 1986* in Montour County, PA



(a) Ages 65–74



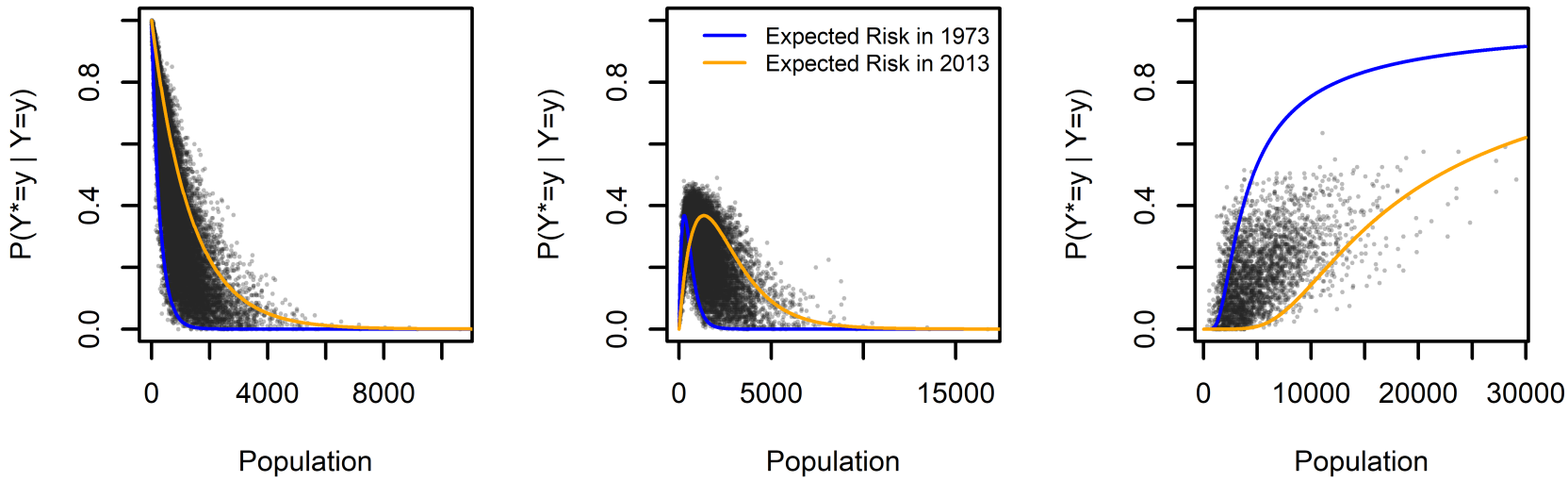
(b) Ages 75–84



(c) Ages 85+

∗ Data *since* 1989 is suppressed on CDC Wonder, but data *prior* to 1989 is unsuppressed and publicly available.

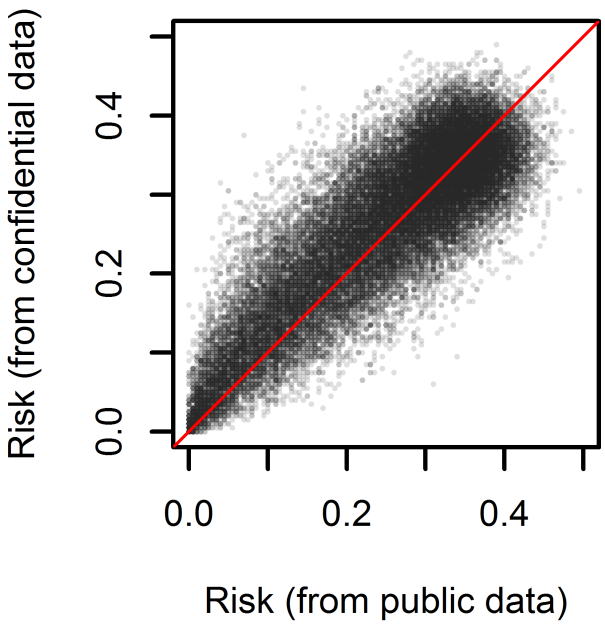
Disclosure risk



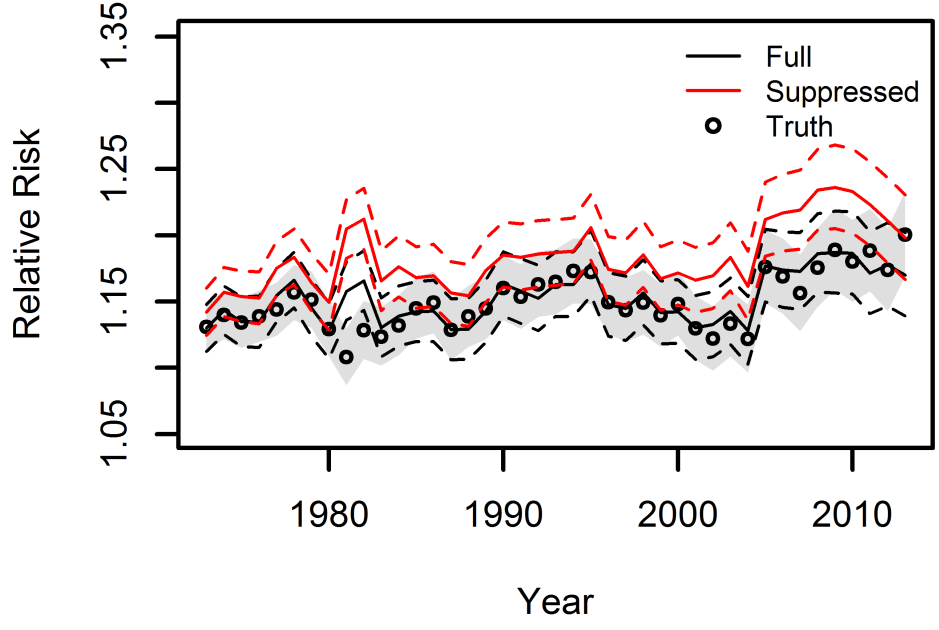
(a) $P(Y^*_{ikt} = 0 \mid Y_{ikt} = 0)$ (b) $P(Y^*_{ikt} = 1 \mid Y_{ikt} = 1)$ (c) $P(Y^*_{ikt} = 9 \mid Y_{ikt} = 9)$

- Red and green lines denote the expected risk probabilities at the beginning and end of the study, respectively.
- These risk probabilities are highest at the boundary values.
  - If $Y_{ikt} = 0$, there is no one's privacy to be concerned about.
  - We set the upper bound to some conservative value.
- Interior values are essentially what we would "expect"

Disclosure Risk and Utility



(a) $P\left(Y^*_{ikt} = 1 \mid Y_{ikt} = 1\right)$



(b) Ages 75–84

Table of Contents

Introduction

Methods

- Multivariate space-time CAR model
- Generation and evaluation of synthetic data

Results

- Analysis of the stroke mortality data
- Generation/Evaluation of synthetic data

Summary and Discussion

Summary

Recall that the goal of this talk was to develop a statistical framework which is useful for both public health surveillance and the dissemination of information, thereby avoiding a redundancy of tasks. Thus, we claim:

- The MSTCAR is well-suited for conducting public health surveillance.
  - The posterior distribution yields inference on rates, aggregates of rates, rate ratios, declines, etc.
- The MSTCAR shows promise for generating synthetic data for public-use
  - Using the MSTCAR should yield synthetic data with very high utility
  - That said, it is not without its weaknesses

Limitations / Future Work

- No clear connection (yet) between this approach and a form of differential privacy
  - We see some similarities between our framework and that used for OnTheMap (Machanavajjhala et al., 2008), but the question is how to express the "informativeness" of our model.
- Not practical for *BIG* examples without *BIG* assumptions
  - A similar analysis with $N_g = 24$ age/race/sex subgroups takes 2+ weeks to run
- Aspects of utility unclear
  - e.g., we assume (but haven't proven) that by accounting for spatial structure, we will preserve relationships for spatially-structured covariates *not included* in the model

Our vision: For this approach to ultimately be used for a series of one-offs rather than to generate a "Synthetic CDC WONDER"

- e.g., CDC researchers study trends in stroke mortality, publish their research, and make the synthetic data available for further analysis by outside researchers

Questions?

hsq23@drexel.edu