Generating Spatially Referenced, Differentially Private Synthetic Data Using a Poisson-lognormal Approach

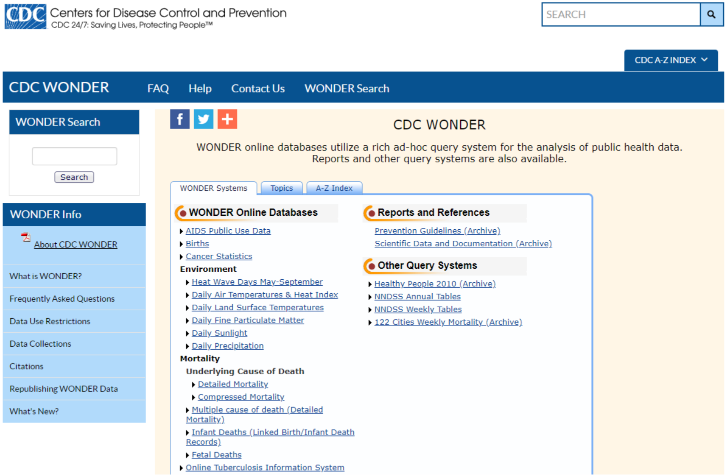
(a work in progress)

Harrison Quick (University of Minnesota)

Table of Contents

Table of Contents

CDC WONDER


CDC
Centers for Disease Control and Prevention
CDC 24/7: Saving Lives, Protecting People™
SEARCH
CDC A-Z INDEX
CDC WONDER
FAQ
Help
Contact Us
WONDER Search
WONDER Search
Search
WONDER Info
About CDC WONDER
What is WONDER?
Frequently Asked Questions
Data Use Restrictions
Data Collections
Citations
Republishing WONDER Data
What's New?
CDC WONDER
WONDER online databases utilize a rich ad-hoc query system for the analysis of public health data.
Reports and other query systems are also available.
WONDER Systems
Topics
A-Z Index
WONDER Online Databases
AIDS Public Use Data
Births
Cancer Statistics
Environment
Heat Wave Days May-September
Daily Air Temperatures & Heat Index
Daily Land Surface Temperatures
Daily Fine Particulate Matter
Daily Sunlight
Daily Precipitation
Mortality
Underlying Cause of Death
Detailed Mortality
Compressed Mortality
Multiple cause of death (Detailed Mortality)
Infant Deaths (Linked Birth/Infant Death Records)
Fetal Deaths
Online Tuberculosis Information System
Reports and References
Prevention Guidelines (Archive)
Scientific Data and Documentation (Archive)
Other Query Systems
Healthy People 2010 (Archive)
NNDSS Annual Tables
NNDSS Weekly Tables
122 Cities Weekly Mortality (Archive)

CDC WONDER

County-level heart disease-related death counts for ages 35–44 in 2016 from all races and all genders

Compressed Mortality, 1999-2016 Results

Request Form | Results | Map | Chart | About

Compressed Mortality Data | Dataset Documentation | Other Data Access | Help for Results | Printing Tips | Help with Exports | Save | Export | Reset

Quick Options | More Options | Top | Notes | Citation | Query Criteria

County	Deaths	Population	Crude Rate Per 100,000
Autauga County, AL (01001)	Suppressed	7,190	Suppressed
Baldwin County, AL (01003)	14	24,545	57.0 (Unreliable)
Barbour County, AL (01005)	Suppressed	3,171	Suppressed
Bibb County, AL (01007)	Suppressed	3,043	Suppressed
Blount County, AL (01009)	Suppressed	7,090	Suppressed
Bullock County, AL (01011)	Suppressed	1,301	Suppressed
Butler County, AL (01013)	Suppressed	2,262	Suppressed
Calhoun County, AL (01015)	19	13,460	141.2 (Unreliable)

All counts less than 10 are suppressed in public-use datasets

CDC WONDER

While CDC WONDER offers a wealth of data and *does* implement privacy protections, there is still room for improvement:

- Utility: Suppression of small counts affects users' ability to assess...
- Privacy

Is there a way that CDC can address these issues?

CDC WONDER

While CDC WONDER offers a wealth of data and *does* implement privacy protections, there is still room for improvement:

- Utility: Suppression of small counts affects users' ability to assess...
  - Urban/Rural disparities
- Privacy

Is there a way that CDC can address these issues?

CDC WONDER

While CDC WONDER offers a wealth of data and *does* implement privacy protections, there is still room for improvement:

- Utility: Suppression of small counts affects users' ability to assess...
  - Urban/Rural disparities
  - Racial disparities
- Privacy

Is there a way that CDC can address these issues?

CDC WONDER

While CDC WONDER offers a wealth of data and *does* implement privacy protections, there is still room for improvement:

- Utility: Suppression of small counts affects users' ability to assess...
  - Urban/Rural disparities
  - Racial disparities
  - Differences by sex
- Privacy

Is there a way that CDC can address these issues?

CDC WONDER

While CDC WONDER offers a wealth of data and *does* implement privacy protections, there is still room for improvement:

- Utility: Suppression of small counts affects users' ability to assess...
  - Urban/Rural disparities
  - Racial disparities
  - Differences by sex
  - Differences by age
- Privacy

Is there a way that CDC can address these issues?

CDC WONDER

While CDC WONDER offers a wealth of data and *does* implement privacy protections, there is still room for improvement:

- Utility: Suppression of small counts affects users' ability to assess...
  - Urban/Rural disparities
  - Racial disparities
  - Differences by sex
  - Differences by age
  - Differences by cause-of-death
- Privacy

Is there a way that CDC can address these issues?

CDC WONDER

While CDC WONDER offers a wealth of data and *does* implement privacy protections, there is still room for improvement:

- Utility: Suppression of small counts affects users' ability to assess...
  - Urban/Rural disparities
  - Racial disparities
  - Differences by sex
  - Differences by age
  - Differences by cause-of-death
- Privacy
  - Targeted attacks by clever intruders can overcome data suppression to uncover the true counts

Is there a way that CDC can address these issues?

Synthetic Data

One option to address the issue of data suppression would be to release *synthetic data*: e.g., if

- $\mathbf{y} = (y_1, \ldots, y_I)^T$ denotes a restricted-use dataset of I observations,
- $p(\mathbf{y} \mid \phi)$ is an appropriate statistical model for $\mathbf{y}$ with parameters ϕ, and
- $p(\phi \mid \psi)$ is a prior distribution for ϕ given hyperparameters, ψ,

then we can generate a synthetic dataset, $\mathbf{z} = (z_1, \ldots, z_I)^T$, from the posterior predictive distribution,

$$p(\mathbf{z} \mid \mathbf{y}, \psi) = \int p(\mathbf{z} \mid \phi)\, p(\phi \mid \mathbf{y}, \psi)\, d\phi.$$

That is, we can sample ϕ^* from $p(\phi \mid \mathbf{y}, \psi)$ and then sample $\mathbf{z}$ from $p(\mathbf{z} \mid \phi^*)$.

- Natural next question: How do we know if synthetic data generated from $p(\mathbf{z} \mid \mathbf{y}, \psi)$ are sufficiently protective?

Differential Privacy (Dwork, 2006)

The standard typically used for demonstrating formal privacy guarantees is the concept of *differential privacy* (Dwork, 2006).
In this context, $p\left(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\psi}\right)$ is ϵ-differentially private if for any similar[1] dataset, $\mathbf{x}$,

$$\left| \log \frac{p\left(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\psi}\right)}{p\left(\mathbf{z} \mid \mathbf{x}, \boldsymbol{\psi}\right)} \right| \leq \epsilon. \tag{1}$$

While $\boldsymbol{\psi}$ *can* be viewed as a vector of model parameters, *in practice* the elements of $\boldsymbol{\psi}$ are merely specified to satisfy ϵ-differential privacy.

[1] $\|\mathbf{x} - \mathbf{y}\| = 2$ and $\sum_i x_i = \sum_i y_i$ — i.e., there exists i and i' such that $x_i = y_i - 1$ and $x_{i'} = y_{i'} + 1$ with all other values equal

Table of Contents

Poisson-Gamma model (Quick, 2021)

Motivated by the field of disease mapping — where death data are typically modeled as being Poisson distributed — Quick (2021) proposed assuming

$$y_i \mid \lambda_i \sim \text{Pois}\left(n_i \lambda_i\right) \text{ and } \lambda_i \sim \text{Gamma}\left(a_i, b_i\right)$$

which implies $\lambda_i \mid y_i \sim \text{Gamma}\left(y_i + a_i, n_i + b_i\right)$. Now recall that if the y_i are (conditionally) independent Poisson random variables and if $y_\cdot = \sum_i y_i$, then

$$\mathbf{y} \mid \boldsymbol{\lambda}, \sum_i y_i = y_\cdot \sim \text{Mult}\left(y_\cdot, \left\{\frac{n_i \lambda_i}{\sum_j n_j \lambda_j}\right\}\right)$$

Thus, we can generate synthetic data by:

1. Sampling λ_i^* from $\text{Gamma}\left(y_i + a_i, n_i + b_i\right)$ for $i = 1, \ldots, I$
2. Sampling $\mathbf{z} \sim \text{Mult}\left(z_\cdot, \left\{n_i \lambda_i^* / \sum_j n_j \lambda_j^*\right\}\right)$

But under what conditions will this satisfy ϵ-differential privacy?

Poisson-Gamma model — ϵ-differential privacy

It *can* (but won't) be shown that the Poisson-gamma synthesizer, denoted $p\left(\mathbf{z} \mid \mathbf{y}, \mathbf{a}, \mathbf{b}\right)$, will satisfy ϵ-differential privacy if

$$a_i \geq \frac{z.}{e^{\epsilon}/\nu_i - 1} \tag{2}$$

where $\nu_i \in [1, 2]$ is a function of $\mathbf{n}$, $\mathbf{a}$, and $\mathbf{b}$ is *generally* ≈ 1 when the number of observations is large. Later, Quick (2022) proposed using the *prior predictive distribution* to truncate the synthetic data to a "reasonable" range of values, $z_i \in [L_i, U_i]$, which yields the requirement that

$$a_i \geq \frac{U_i - L_i}{e^{\epsilon}/\nu_i - 1} - 2 * L_i, \text{ where } \nu_i = \frac{y. - L_i + a_{(i)} + y. - L_i - 1}{y. - U_i + a_{(i)} + y. - L_i - 1}. \tag{3}$$

e.g., if $n_i = 100$ and I *expect* $\lambda_i \approx 0.01$ — and thus I *expect* $y_i \approx 100 \times 0.01 = 1$ — then it's probably fair to assume that $y_i \in [0, 20]$ even if $\sum_i y_i = 10{,}000$.

Table of Contents

BYM CAR model framework

Rather than smooth all observations toward a common rate, we'd like to take a page from the spatial statistics and disease mapping literature and consider the conditional autoregressive (CAR) model framework of Besag, York, and Mollié (BYM; 1991), which assumes:

$$
\begin{aligned}
y_i \mid \lambda_i &\sim \text{Pois}\left(n_i \lambda_i\right) \\
\log \lambda_i \mid \beta_0, \mathbf{z}, \tau^2 &\sim \text{Norm}\left(\beta_0 + z_i, \tau^2\right) \\
\mathbf{z} \mid \sigma^2 &\sim \text{CAR}\left(\sigma^2\right),
\end{aligned}
$$

where $\mathbf{z} \sim \text{CAR}\left(\sigma^2\right)$ implies

$$
z_i \mid \mathbf{z}_{(i)}, \sigma^2 \sim \text{Norm}\left(\sum_{j \sim i} z_j / m_i, \sigma^2 / m_i\right),
$$

where $j \sim i$ indicates that counties i and j are neighbors and m_i denotes the number of counties that neighbor county i.

Quantifying the informativeness of the BYM CAR model

While the CAR model framework is nice, it's not straightforward to quantify how informative the model is (relative to the gamma prior in the previous framework). To that end, recent work by Quick et al. (2021; *SSTE*) started by establishing a relationship between $\lambda_i \sim \text{Gamma}\left(a_i, b_i\right)$ and $\lambda_i \sim \text{LogNorm}\left(\mu_i, \sigma_i^2\right)$, which yielded an approximation of the form:

$$\widehat{a}_i = \frac{1}{\exp \sigma_i^2 - 1}. \tag{4}$$

Quick et al. (2021) then extended this concept to the BYM CAR model for a region with m_0 neighbors by integrating the CAR random effects out of the model, yielding:

$$\widehat{a}_0 = \frac{1}{\exp\left[\tau^2 + \left(\sigma^2 + \tau^2\right) / m_0\right] - 1}. \tag{5}$$

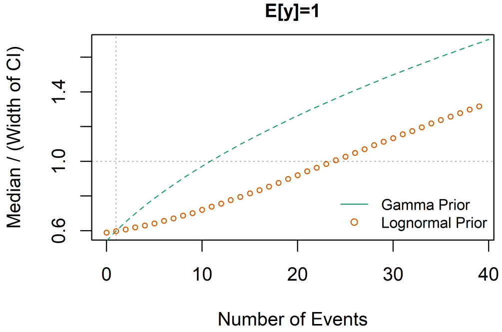
Because each region can have its own number of neighbors — and to facilitate comparisons between different maps — we write $\widehat{a}_0$ using $m_0 = 3$ as a rule-of-thumb.

Comparing the Poisson-Gamma and Poisson-Lognormal

To help establish the model informativeness calculation for the Poisson-lognormal framework (and, by extension, the BYM CAR model), Quick et al. (2021) proposed the use of the *relative precision*, which is defined as:

$$\mathrm{RP}\left(\lambda_i \,|\, \mathbf{y}\right) = \frac{\text{Posterior Median of } \lambda_i}{\text{Width of the 95\% CI for } \lambda_i}$$

- Under the Poisson-gamma model, the relative precision is simply a function of $y + a$
- Under the Poisson-lognormal, the relative precision is a function of both $y + \widehat{a}$ *and* the discrepancy between the observed y and $E\left[y \,|\, \mu, \sigma^2\right]$



Hand-wavy Differential Privacy

Based on this, I'm claiming that the Poisson-lognormal framework *approximately* satisfies ϵ-differential privacy if a Poisson-gamma framework with $a_i \leq \widehat{a}_0$ for all i would satisfy ϵ-differential privacy.

- Note: This is in no way related to any other "approximate differential privacy" definition that I'm aware of. I'm essentially claiming there's a "transitive property" of differential privacy.

Why do I expect this to be an attractive strategy?

- The aforementioned Quick et al. (2021) demonstrates that the BYM CAR model framework tends to produce *very* informative models.
  - Quick et al. (2021) criticized this as oversmoothing, but this is actually ideal for privacy because it should provide yield improved utility (smoothing toward regional averages rather than state-level or national averages) *and* improved privacy protections (typical levels of smoothing can correspond to $\epsilon \approx 1$).

Table of Contents

CVD-related Deaths in Minnesota Census Tracts in 2011

Attribute	**Levels**
Census Tract	$i = 1, \ldots, 1{,}336$ Census tracts in Minnesota
Age	$a = 1, \ldots, 12$ Levels Ages 30–34; Ages 35–39; Ages 40–44; Ages 45–49; Ages 50–54; Ages 55–59; Ages 60–64; Ages 65–69; Ages 70–74; Ages 75–79; Ages 80–84; Ages 85 and older

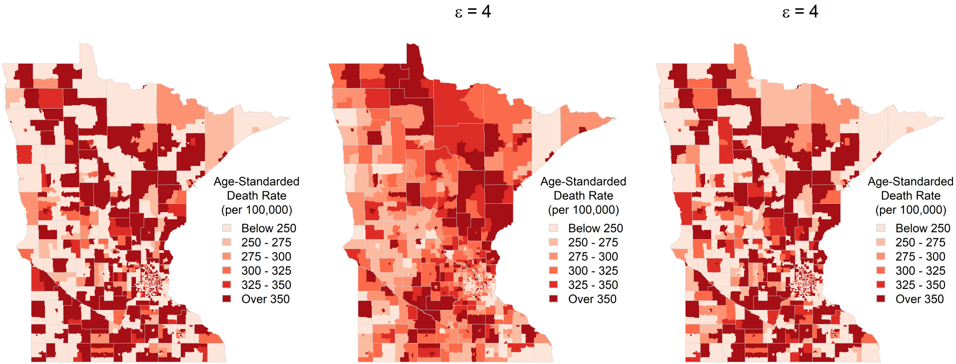
In total, there were $y_{\cdot} = \sum_{ia} y_{ia} = 4{,}187$ CVD-related deaths for white men in MN in 2011 belonging to these $1{,}336 \times 12 = 16{,}032$ strata.

- We have data for other demographic groups, other causes-of-death, and other years, but I tried to keep it simple as a "proof-of-concept".
- Over 80% of the death counts are zero and the largest value is $y_{ia} = 9$

Prior information

- Both models take advantage of tract-level population estimates commissioned by the National Cancer Institute's (NCI's) Surveillance, Epidemiology, and End Results (SEER) program.
- The Poisson-gamma model's prior distributions are designed to smooth toward overall age-specific death rates
  - I cheat in this example by using MN's 2011 state-level death rates, but I typically use national-level rates published annually by the CDC — in practice, this shouldn't make much of a difference.
- The variance parameters in the BYM-CAR model are fixed such that $\widehat{a}_0$ matches the requirement for ϵ-differential privacy under the P-G framework, but no external information is used to inform any model parameters.
- Most importantly, my goal will be to estimate urban/rural disparities in age-standardized death rates, and neither model accounts for anything about urban/rural disparities.

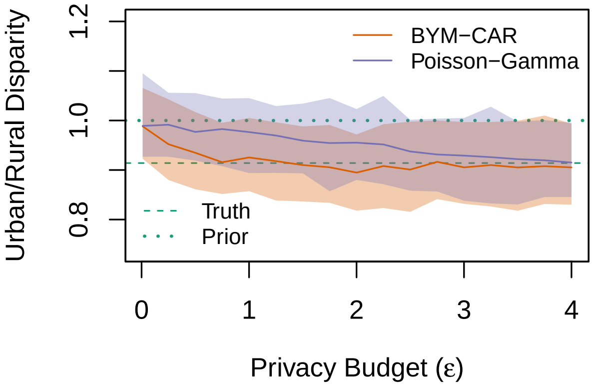
Tract-level Age-Standardized CVD Death Rates



(a) True Rates (b) BYM CAR Model (c) Poisson-Gamma

Figure 1: Degradation in utility for the age-standardized rates as ϵ decreases.

Urban/Rural Disparities in Age-Standardized CVD Death Rates



- For large ϵ, both models preserve the urban/rural disparity in age-standardized CVD death rates by virtue of noninformative priors
- As ϵ decreases, the BYM CAR model still preserves geographic (and urban/rural) disparities, whereas the gamma prior does not.

Table of Contents

Summary

Key background:

- The Poisson-gamma model can produce differentially private synthetic data (Quick, 2021; 2022)
- Past work has established a relationship between the informativeness of the prior specification in the Poisson-gamma framework and that of the Poisson-lognormal (and, by extension, the BYM-CAR model; Quick et al., 2021)

Key claim:

- A BYM-CAR model whose informativeness matches a Poisson-gamma model that satisfies ϵ-differential privacy will *approximately* satisfy ϵ-differential privacy

Key results:

- The BYM-CAR model preserves geographic and urban/rural disparities even for small ϵ, whereas the Poisson-gamma framework gradually shifts from the true disparity toward no disparity.