The geostan R package for Bayesian spatial analysis

Connor Donegan

Doctoral Candidate
Geospatial Information Sciences
The University of Texas at Dallas

Research Assistant
Peter O'Donnell Jr. School of Public Health
University of Texas Southwestern Medical Center

StanConnect Through Space and Time
October 31, 2022
Virtual Conference

Connor.Donegan@UTDallas.edu
https://connordonegan.github.io

What geostan offers

- Exploratory spatial data analysis & model evaluation tools
- Disease mapping and spatial regression models for areal data
- Compatible with the RStan ecosystem of R packages
  - All models were built using Stan and the `rstantools` package
- Interface to Stan models using the familiar R formula inferface
- Tools for building custom spatial models in Stan[1]

[1] See Donegan 2021, 'Building spatial conditional autoregressive models in the Stan programming language' *OSF Pre-prints* https://osf.io/3ey65/

```
> library(geostan)
> data(georgia)
> C <- shape2mat(georgia)
> cars <- prep_car_data(C)
> fit <- stan_car(deaths.male ~ offset(log(pop.at.risk.male)),
                  data = georgia,
                  car = cars,
                  family = poisson()
                  )

> sp_diag(fit, georgia)
```

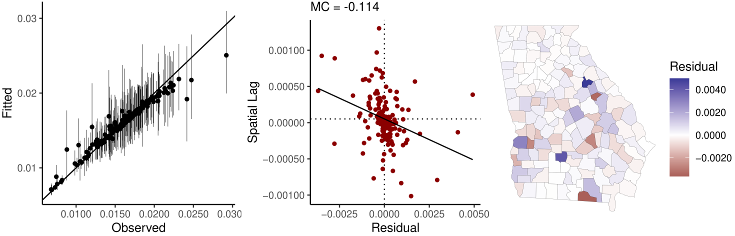


Figure 1: Spatial model diagnostics from geostan::sp_diag.

All models can also include:

- Spatial measurement error (ME) models for covariates
- Non-spatial information pooling (unstructured ‘random effects’)
- Spatially-lagged covariates

Table 1: Models implemented in **geostan**.

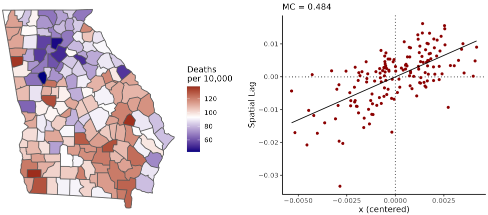
	Gaussian	Student	Poisson	Binomial
CAR	x		x	x
ESF	x	x	x	x
GLM	x	x	x	x
ICAR			x	x
SAR	x		x	x

What is spatial autocorrelation (SA)?

Visually conspicuous map pattern:[2]

- Positive SA: Similar places are near to each other
- Negative SA: Dissimilar places are near to each other
- SA mixtures: Overlapping PSA-NSA patterns

Defining property of PSA: information-redundancy → *error-inflation*



(a) GA counties

(b) Moran scatter plot

Figure 2: Georgia county-level mortality risk, women ages 55-64, years 2014–2018.

[2] Chun and Griffith, 2013; D. Griffith et al., 2019.

‘Small scale’ does not mean ‘precise’

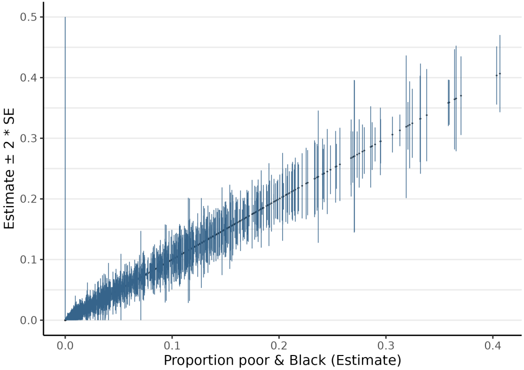


Figure 3: ACS estimates and their margins of error, U.S. counties (2006–2010).

Survey estimate reliability varies:

- idiosyncratically
- geographically
- by survey question
- by social positionality
- by spatial and conceptual resolution

Inference with observational uncertainty

Implications of ignoring ME

Exaggerated confidence in results that are prone to bias.

Jointly model covariates and disease risk

Probability model for (unknown) variable x_i given survey estimates z_i, standard errors s_i, and geographic data:

$$z_i \sim \text{Normal}(x_i, s_i^2) \tag{1}$$

$$g(\boldsymbol{x}) \sim \text{Normal}(\mu, \boldsymbol{\Sigma}) \tag{2}$$

$$\boldsymbol{\Sigma} = (\boldsymbol{I} - \rho \boldsymbol{C})^{-1} \boldsymbol{M} \tag{3}$$

Modeled values $\boldsymbol{x}$ enter the regression/disease model:

$$\boldsymbol{y} \sim \text{Poisson}(e^{\boldsymbol{O} + \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{\phi}}) \tag{4}$$

Logical implications of covariate uncertainty pass into inferences on disease risk.[3]

[3] Donegan et al., 2021; cf. Bernardinelli et al., 1997; Kang et al., 2009; Logan et al., 2019

Ethico-political aspects of public health monitoring

Colorectal cancer (CRC) prevention

- Screening for CRC enables early detection *and prevention* through removal of polyps
- Trends in CRC risk reflect past prevention work, and guide future work
- Choice of variables to stratify on is agenda-setting

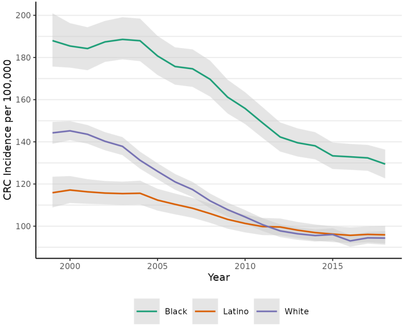


Figure 4: Age-standardized CRC incidence rates, screening age adults (50-79), Texas metro areas, by race-ethnicity.

Modeling census-tract CRC incidence, Dallas-Fort Worth, Texas

Guiding questions

1. Do racial-ethnic patterns in CRC risk map onto the space of the city-region?
2. What is learned by stratifying by nativity and social class?

Model specification

$y \sim \text{Poisson}(E \cdot e^{\alpha + X\beta + \phi})$

y: CRC cases, ages 50–79*
E: Expected cases given age composition
ϕ: Spatial trend (CAR model)
X: Foreign born (%), income, race-ethnicity

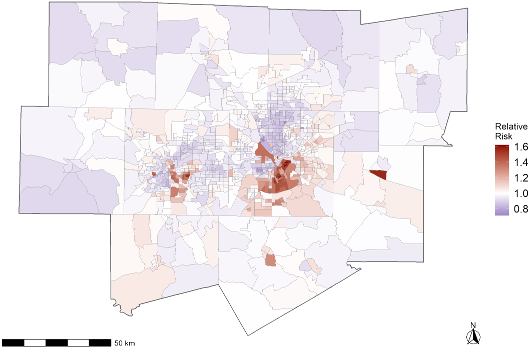
Apply first to 1999–2003 ($n = 6,959$), then 2015–2019 ($n = 9,386$).

geostan code

```
> fitted_model <- stan_car(
    Cases ~ offset(log_E) +
    nativity + log(income) +
    api + black + white,
    car_parts = cars,
    data = cancer,
    centerx = TRUE
    )
```

* Texas Cancer Registry (TCR)

Modeling census tract CRC risk, 1999–2003



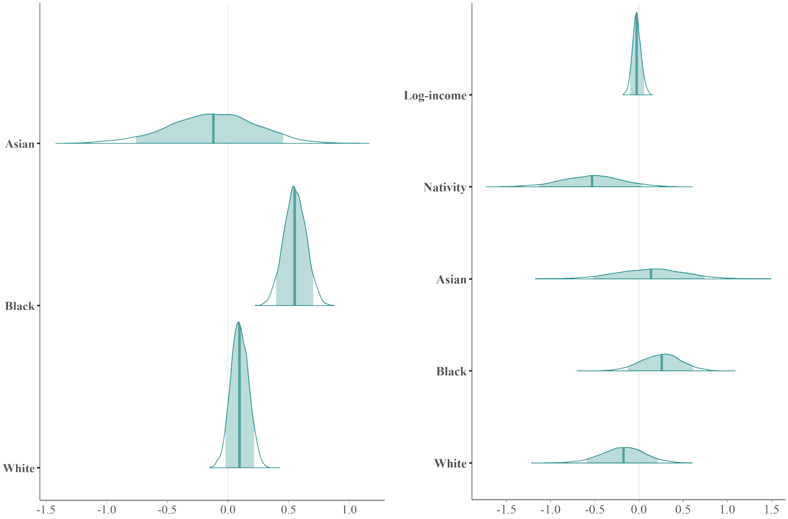
```
> f <- fitted(
   fitted_model
    )

> geo_data$f <- f$mean

> ggplot(geo_data) +
   geom_sf(aes(fill = f)) +
   scale_fill_gradient2(
    midpoint = 1
   )
```

Figure 5: CRC risk by census tract, DFW, 1999–2003.

```
> samples <- as.matrix(fitted_model, pars = ‘beta’)
> bayesplot::mcmc_areas(samples, prob = 0.9)
```

or:

```
> stanfit <- fitted_model$stanfit
> samples <- as.matrix(stanfit, pars = ‘beta’)
> bayesplot::mcmc_areas(samples, prob = 0.9)
```



(a) $\mathcal{M}_1$

(b) $\mathcal{M}_2$

Figure 6: Posterior distributions of β, 1999–2003. *Note:* proportion Hispanic/Latino is serving as a reference category.

Modeling census tract CRC risk, 2015–2019

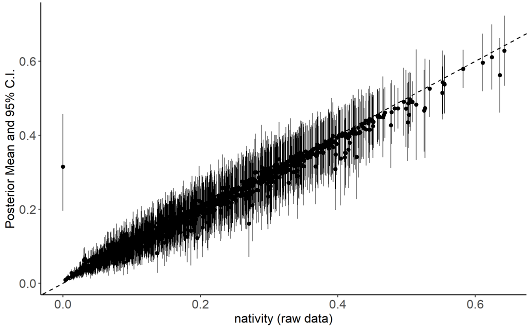


Figure 7: ME model diagnostic for % foreign born: 2019 ACS estimates v. fitted values with 95% CIs (from `geostan::me_diag`).

```
> C <- shape2mat(geo_data, 'B')

> cars <- prep_car_data(C)

> se <- data.frame(
   log_inc = data$log_inc_se,
   nativity = data$nativity_se,
   api = data$api_se,
   black = data$black_se,
   white = data$white_se
   )

> ME <- prep_me_data(
   se = se,
   car_parts = cars,
   logit = c(F, rep(T, 4))
   )

> fitted_model <- stan_icar(
   formula,
   type = 'bym',
   data = data,
   C = C,
   ME = ME,
   ...
   )

> me_diag(
   fitted_model,
   'navitiy',
   geo_data
   )
```

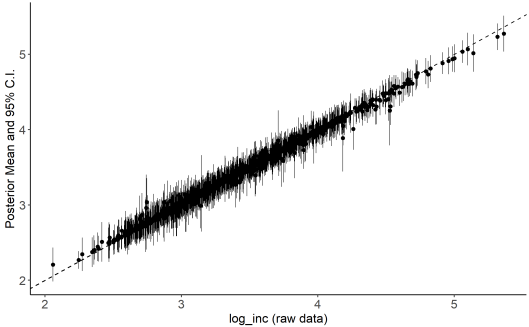


Figure 8: ME model diagnostic for log-income: 2019 ACS estimates v. fitted values with 95% CIs (from `geostan::me_diag`).

```
> me_diag(fitted_model, 'log_inc', geo_data)
```

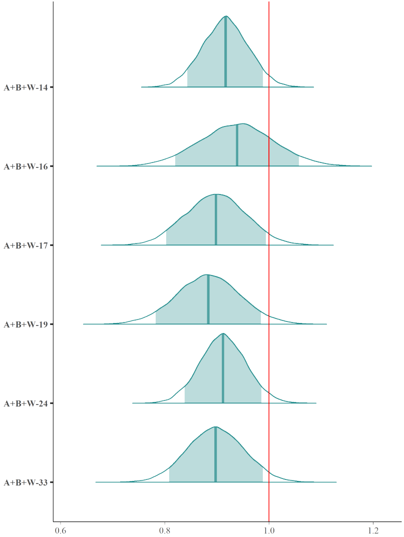
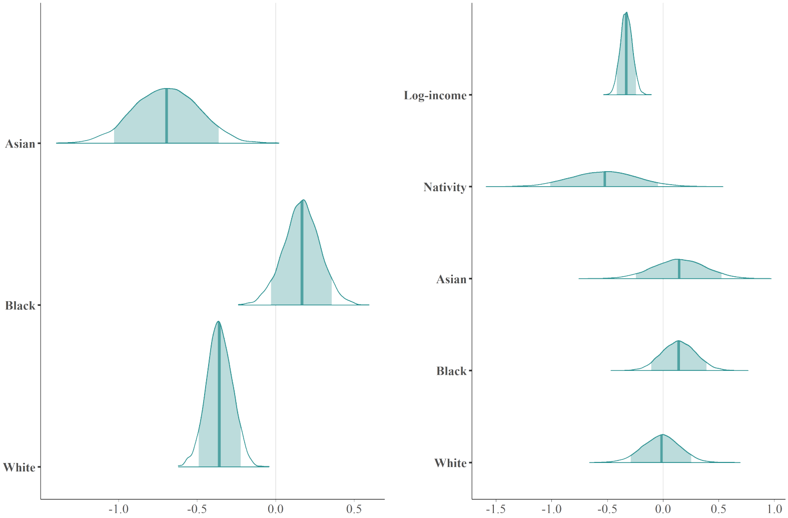


Figure 9: Posterior distribution for Asian + Black + White (population proportions), select census tracts.

```
> X <- as.matrix(fitted_model, pars = "x_true")

> W <- X[, grep("white", colnames(X))]
> B <- X[, grep("black", colnames(X))]
> A <-  X[, grep("api", colnames(X))]

> Z <- A + B + W
```



(a) $\mathcal{M}_1$

(b) $\mathcal{M}_2$

Figure 10: Posterior distributions of β, 2015–2019.

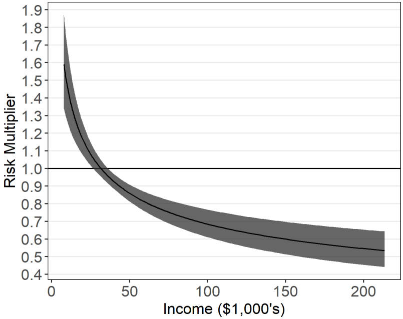
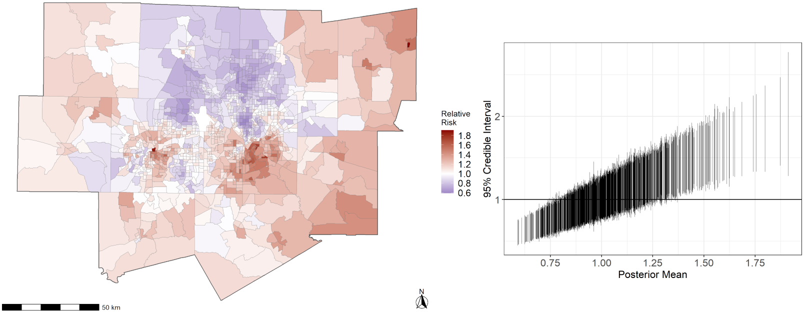


Figure 11: Areal income-CRC risk gradient, 2015–2019.

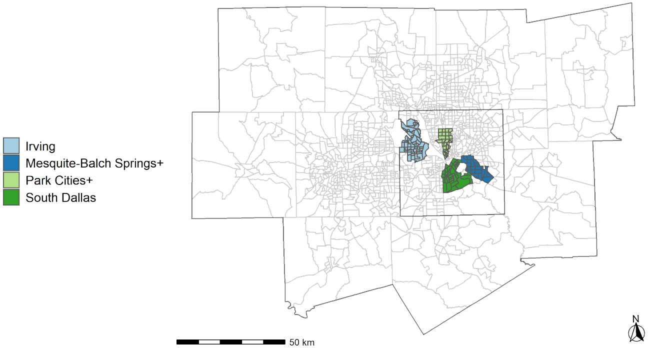
```
> new_data <- data.frame(
   log_inc = seq(min(data$log_inc), max(data$log_inc),
                 length.out = 200),
   nativity = mean(data$nativity),
   ...
   )

> preds <- predict(
   fitted_model,
   data = new_data,
   type = ‘response’
   )

> ggplot(preds) +
    geom_ribbon(aes(exp(log_inc),
                    ymin = ‘2.5%‘, ymax = ‘97.5%‘),
      alpha = .75) +
    geom_line(aes(exp(log_inc), mean))
```



(a) Posterior means

(b) 95% credible intervals

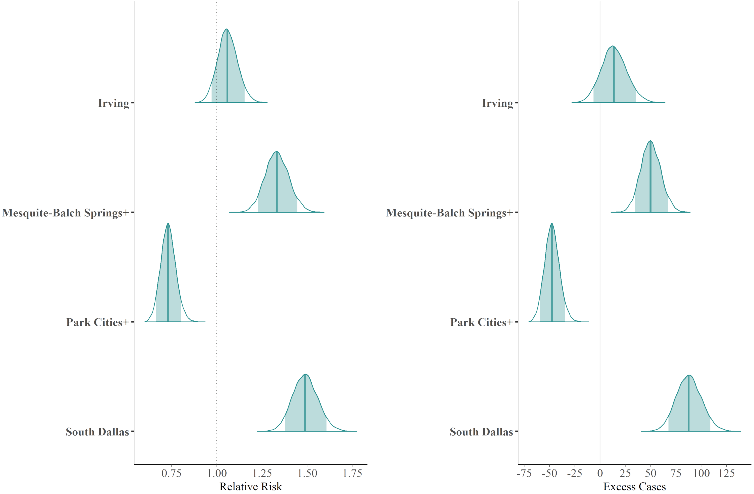
Figure 12: CRC risk by DFW census tract, 2015–2019.



		$1,000s	%			
	Population	Income⋆	Asian	Black	Latino	White
Irving	228,784	29	17	13	41	27
Mesquite-Balch Springs+	146,194	14	0	24	65	10
Park Cities+	99,207	105	4	2	9	82
South Dallas	120,020	14	0	68	29	2

⋆ Income per-capita.

Figure 13: Select cities and neighborhoods in Dallas County.



(a) Fitted/Expected

(b) Fitted - Expected

Figure 14: Probability distributions for area CRC risk, 2015–2019.

Issues and help

connor.donegan@gmail.com
https://github.com/connordonegan/geostan/issues

Documentation

https://connordonegan.github.io/geostan

Installation

```
> install.packages('geostan')
```

References I

Anselin, L. (1995). Local indicators of spatial association—LISA. *Geographical Analysis*, *27*(2), 93–115.

Bernardinelli, L., Pascutto, C., Best, N. G., & Gilks, W. R. (1997). Disease mapping with errors in covariates. *Statistics in Medicine*, *16*(7), 741–752. https://doi.org/10.1002/(sici)1097-0258(19970415)16:7⟨741::aid-sim501⟩3.0.co;2-1

Chun, Y., & Griffith, D. A. (2013). *Spatial statistics and geostatistics: Theory and applications for geographic information science and technology*. Los Angeles, Sage.

Donegan, C. (2021). Building spatial conditional autoregressive models in the Stan programming language. *OSF Pre-print*. https://doi.org/10.31219/osf.io/3ey65

Donegan, C., Chun, Y., & Griffith, D. A. (2021). Modeling community health with areal data: Bayesian inference with survey standard errors and spatial structure. *Int J Env Res Public Health*, *18*(13), 6856. https://doi.org/10.3390/ijerph18136856

Donegan, C., Chun, Y., & Hughes, A. E. (2020). Bayesian estimation of spatial filters with Moran's eigenvectors and hierarchical shrinkage priors. *Spatial Statistics*, *38*, 100450.

References II

Donegan, C., Hughes, A. E., & Lee, S. J. C. (2022). Colorectal cancer incidence, inequalities, and prevention priorities in urban texas: Surveillance study with the “surveil” software package. *JMIR Public Health and Surveillance*, *8*(8), e34589.

Gabry, J., Goodrich, B., & Lysy, M. (2020). *Rstantools: Tools for developing r packages interfacing with 'stan'*. https://mc-stan.org/rstantools/

Griffith, D., Chun, Y., & Li, B. (2019). *Spatial regression analysis using eigenvector spatial filtering*. London, Academic Press.

Griffith, D. A. (2011). Positive spatial autocorrelation, mixture distributions, and geospatial data histograms. In *IEEE international conference on spatial data mining and geographical knowledge services*.

Kang, E. L., Liu, D., & Cressie, N. (2009). Statistical analysis of small-area data based on independence, spatial, non-hierarchical, and hierarchical models. *Computational Statistics & Data Analysis*, *53*, 3016–3032. https://doi.org/10.1016/j.csda.2008.07.033

Logan, J. R., Bauer, C., Ke, J., Xu, H., & Li, F. (2019). Models for small area estimation for census tracts. *Geographical Analysis*, *52*(3), 325–350. https://doi.org/10.1111/gean.12215

References III

- Morris, M., Wheeler-Martin, K., Simpson, D., Mooney, S. J., Gelman, A., & DiMaggio, C. (2019). Bayesian hierarchical spatial models: Implementing the Besag York Mollié model in Stan. *Spatial and Spatio-temporal Epidemiology*, *31*, 100301.
- Stan Development Team. (2022). Stan modeling language users guide and reference manual, version 2.30. https://mc-stan.org/