geostan: An R package for Bayesian spatial analysis

17 September, 2022

Summary

Analyses of data collected across areal units, such as census tracts and states, are now ubiquitous in the social and health sciences. Data sources include surveys (especially large government-back surveys like the US Census Bureau's American Community Survey (ACS)), vital statistics systems, and disease registries (particularly cancer registries). These data sources can provide crucial information about population health and socio-economic outcomes, but many standard (non-spatial) statistical methods and workflows are either not applicable to spatial data or require adjustment (Cressie 2015; Haining and Li 2020).

This paper introduces **geostan**, an `R` package for analyzing spatial data using Bayesian inference. The primary focus of the package is areal data for socio-economic and health research. The package provides tools for a complete workflow for spatial regression and disease mapping, and has unique spatial measurement error (ME) models suitable for researchers using ACS estimates as covariates (Donegan, Chun, and Griffith 2021).

Statement of need

The distinguishing characteristic of spatial data is that maps of the data typically contain moderate to strong spatial patterns, or spatial autocorrelation, which reduces effective sample size (ESS) and renders many standard statistical methods inappropriate ("Student" [W.S. Gausset] 1914; Clifford, Richardson, and Hémon 1989). In addition, spatial patterns are often of direct interest—for example, disease mapping studies are concerned primarily with understanding how disease or mortality risk vary over space.

A major challenge for spatial analysis is data quality, particularly for researchers using survey-based covariates. A single spatial analysis may use dozens, or even thousands, of error-laden survey estimates. Sampling error in ACS estimates is often substantial in magnitude and socially patterned (Folch et al. 2016; Donegan, Chun, and Griffith 2021), which can have real consequences on communities and service providers (Bazuin and Fraser 2013). Spatial ME models are required to avoid ME biases and unwarranted levels of confidence in results.

Existing R packages with spatial modeling functions include **spatialreg** (R. Bivand and Piras 2015), **INLA** (Rue, Martino, and Chopin 2009), **ngspatial** (Hughes and Cui 2020), **BayesX** (Belitz et al. 2022; Umlauf et al. 2015), **CARBayes** (Lee 2013), **nimble** (de Valpine et al. 2017). Custom spatial models can be built using **rstan** (Stan Development Team 2022a), **INLA*, and **nimble**, including spatial ME models, but this requires specialized programming and statistical skills. **geostan** fills two gaps in this software landscape. First, **geostan** offers spatial ME models that are appropriate for survey-based covariates. Second, **geostan** provides spatial model diagnostic functions that make it easy for users to evaluate model results even if they are unfamiliar with Markov chain Monte Carlo (MCMC).

Functionality

geostan provides tools for spatial data visualization, construction of spatial weights matrices, spatial ME models, models for censored count data, and multiple types of spatial statistical models for continuous and discrete data types. Spatial weights matrices are created by calling the `spdep` package (R. S. Bivand, Pebesma, and Gomez-Rubio 2013) to identify neighboring areas, and the results are returned in sparse matrix format using the `Matrix` package (Bates, Maechler, and Jagan 2022).

geostan uses Markov chain Monte Carlo (MCMC) for inference, which allows users to conduct formal inference on generated quantities of interest. The models are built using the `Stan` modeling language, a state-of-the-art platform for MCMC sampling (Gabry, Goodrich, and Lysy 2020; Stan Development Team 2022a, 2022b), but users only need to be familiar with the standard R formula interface. Because **geostan** returns `stanfit` objects from **rstan**, it is compatible with the **rstan** ecosystem of packages including **shinystan** for visual summaries of model parameters and MCMC diagnostics (Gabry 2018), **tidybayes** for working with MCMC samples (Kay 2022), and **bridgesampling** for model comparison using Bayes factors (Gronau, Singmann, and Wagenmakers 2020).

Exploratory spatial data analysis (ESDA)

The package provides convenience functions for visualizing spatial patterns and conducting ESDA, including

- Moran scatter plot for visualizing spatial autocorrelation (Chun and Griffith 2013)
- Moran coefficient and Geary Ratio for measuring global spatial autocorrelation (Chun and Griffith 2013)
- Local Moran's I and local Geary's C for measuring and visualizing local spatial autocorrelation (Anselin 1995)
- The Approximate Profile Likelihood (APLE) estimator for measuring spatial autocorrelation (Li, Calder, and Cressie 2007)
- Effective sample size (ESS) calculation (D. A. Griffith 2005)

These tools are provided for exploratory analysis, not 'cluster detection'; p-values are not provided.

geostan also provides a convenience function for obtaining a quick visual summary of a variable (see Figure 1). When a fitted model is provided, the `sp_diag` function returns graphical diagnostics for model residuals.

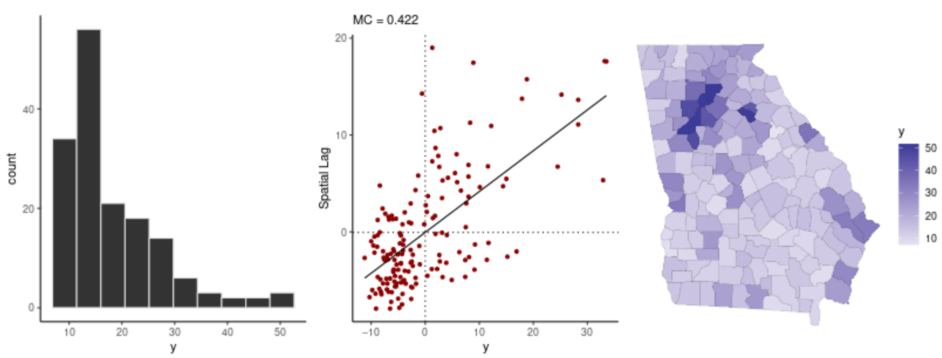

Figure 1: Spatial diagnostic summary for percent college educated, Georgia counties.

Spatial models

Table 1 lists the types of spatial models that are implemented in **geostan**. In addition to (non-spatial) generalized linear models (GLMs), options include spatial conditional autoregressive (CAR) models (Donegan 2021), intrinsic conditional autoregressive (ICAR) models including the BYM (Besag, York, and Mollié 1991) and BYM2 specifications (Riebler et al. 2016; Morris et al. 2019; Donegan and Morris 2021), simultaneously-specified spatial autoregressive (SAR) models (Cliff and Ord 1981) (which are referred to as the spatial error model (SEM) in the econometrics literature (LeSage 2014)), and eigenvector spatial filtering (ESF) (D. Griffith, Chun, and Li 2019; Donegan, Chun, and Hughes 2020).

All of the models allow for a set of exchangeable 'random effects' to be added, and spatially lagged covariates (SLX) can also be added to any of the models. While proper CAR models have been avoided in the past due to their computational burden, the CAR model is the most efficient spatial model in **geostan**. It is fast enough to work interactively on a laptop with $3,000+$ observations, such as U.S. county data.

Table 1: Spatial models currently implemented in **geostan**.

	Gaussian	Student's t	Poisson	Binomial
CAR	x		x	x
ESF	x	x	x	x
GLM	x	x	x	x
ICAR			x	x
SAR	x		x	x

A set of functions for working with model results conveniently extract fitted values, marginal effects, residuals, spatial trends, and posterior (or prior) predictive distributions. Users are encouraged to always undertake a thoughtful spatial analysis of model residuals and other quantities to critique and improve their models through successive rounds of ESDA (cf. Gabry et al. 2019).

Spatial ME models

ME models can be added to any **geostan** model. These are models for covariates measured with error, particularly small-area survey estimates with standard errors. The ME models treat the true covariate values as unknown parameters or latent variables, which are assigned a spatial CAR prior model. Users provide the scale of observational uncertainty or ME (e.g., survey standard errors) as data (Donegan, Chun, and Griffith 2021; cf. Bernardinelli et al. 1997; Xia and Carlin 1998; Kang, Liu, and Cressie 2009; Logan et al. 2019). All uncertain inferences from the ME models are automatically propagated throughout the regression or disease mapping model, and graphical diagnostics are provided for evaluating results of spatial ME models.

Acknowledgements

I am grateful for support this project received from Esri Inc. and the Geospatial Information Sciences program at The University of Texas at Dallas.

References

Anselin, Luc. 1995. "Local Indicators of Spatial Association—LISA." *Geographical Analysis* 27 (2): 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x.

Bates, Douglas, Martin Maechler, and Mikael Jagan. 2022. *Matrix: Sparse and Dense Matrix Classes and Methods*. https://R-Forge.R-project.org/projects/matrix/.

Bazuin, Joshua Theodore, and James Curtis Fraser. 2013. "How the ACS Gets It Wrong: The Story of the American Community Survey and a Small, Inner City Neighborhood." *Applied Geography* 45 (12): 292–302. https://doi.org/10.1016/j.apgeog.2013.08.013.

Belitz, Christiane, Andreas Brezger, Thomas Kneib, Stefan Lang, and Nikolaus Umlauf. 2022. *BayesX: Software for Bayesian Inference in Structured Additive Regression Models*. https://www.uni-goettingen.de/de/bayesx/550513.html.

Bernardinelli, Luisa, Cristian Pascutto, Nicola G. Best, and Walter R. Gilks. 1997. "Disease Mapping with Errors in Covariates." *Statistics in Medicine* 16 (7): 741–52. https://doi.org/10.1002/(sici)1097-0258(19970415)16:7%3C741::aid-sim501%3E3.0.co;2-1.

Besag, Julian, Jeremy York, and Annie Mollié. 1991. "Bayesian Image Restoration, with Two Applications in Spatial Statistics." *Annals of the Institute of Statistical Volume* 43: 1–20. https://doi.org/10.1007/BF00116466.

Bivand, Roger S., Edzer Pebesma, and Virgilio Gomez-Rubio. 2013. *Applied Spatial Data Analysis with R, Second Edition*. Springer, NY. https://asdar-book.org/.

Bivand, Roger, and Gianfranco Piras. 2015. "Comparing Implementations of Estimation Methods for Spatial Econometrics." *Journal of Statistical Software* 63 (18): 1–36. https://doi.org/10.18637/jss.v063.i18.

Chun, Yongwan, and Daniel A Griffith. 2013. *Spatial Statistics and Geostatistics: Theory and Applications for Geographic Information Science and Technology*. Los Angeles: Sage.

Cliff, AD, and JK Ord. 1981. *Spatial Processes: Models and Applications*. Pion.

Clifford, Peter, Sylvia Richardson, and Denis Hémon. 1989. "Assessing the Significance of the Correlation Between Two Spatial Processes." *Biometrics* 45: 123–34. https://doi.org/10.2307/2532039.

Cressie, Noel. 2015. *Statistics for Spatial Data*. John Wiley & Sons.

de Valpine, Perry, Daniel Turek, Christopher Paciorek, Cliff Anderson-Bergman, Duncan Temple Lang, and Ras Bodik. 2017. "Programming with Models: Writing Statistical Algorithms for General Model Structures with NIMBLE." *Journal of Computational and Graphical Statistics* 26: 403–13. https://doi.org/10.1080/10618600.2016.1172487.

Donegan, Connor. 2021. "Building Spatial Conditional Autoregressive (CAR) Models in the Stan Programming Language." *OSF Preprints*. https://doi.org/10.31219/osf.io/3ey65.

Donegan, Connor, Yongwan Chun, and Daniel A. Griffith. 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, Connor, Yongwan Chun, and Amy E Hughes. 2020. "Bayesian Estimation of Spatial Filters with Moran's Eigenvectors and Hierarchical Shrinkage Priors." *Spatial Statistics* 38: 100450. https://doi.org/10.1016/j.spasta.2020.100450.

Donegan, Connor, and Mitzi Morris. 2021. "Flexible Functions for ICAR, BYM, and Bym2 Models in Stan." https://github.com/ConnorDonegan/Stan-IAR (accessed on July 13, 2022).

Folch, David C., Daniel Arribas-Bel, Julia Koschinsky, and Seth E. Spielman. 2016. "Spatial Variation in the Quality of American Community Survey Estimates." *Demography* 53: 1535–54. https://doi.org/10.1007/s13524-016-0499-1.

Gabry, Jonah. 2018. *Shinystan: Interactive Visual and Numerical Diagnostics and Posterior Analysis for Bayesian Models*. https://CRAN.R-project.org/package=shinystan.

Gabry, Jonah, Ben Goodrich, and Martin Lysy. 2020. *Rstantools: Tools for Developing r Packages Interfacing with 'Stan'*.

Gabry, Jonah, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. 2019. "Visualization in Bayesian Workflow." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 182 (2): 389–402. https://doi.org/10.1111/rssa.12378.

Griffith, Daniel A. 2005. "Effective Geographic Sample Size in the Presence of Spatial Autocorrelation." *Annals of the Association of American Geographers* 95 (4): 740–60. https://doi.org/10.1111/j.1467-8306.2005.00484.x.

Griffith, Daniel, Yongwan Chun, and Bin Li. 2019. *Spatial Regression Analysis Using Eigenvector Spatial Filtering*. London: Academic Press. https://doi.org/10.1016/C2017-0-01015-7.

Gronau, Quentin F., Henrik Singmann, and Eric-Jan Wagenmakers. 2020. "bridgesampling: An R Package for Estimating Normalizing Constants." *Journal of Statistical Software* 92 (10): 1–29. https://doi.org/10.18637/jss.v092.i10.

Haining, Robert P., and Guangquan Li. 2020. *Modelling Spatial and Spatio-Temporal Data: A Bayesian Approach*. Boca Raton, FL, USA: CRC Press.

Hughes, John, and Xiaohui Cui. 2020. *ngspatial: Fitting the Centered Autologistic and Sparse Spatial Generalized Linear Mixed Models for Areal Data*. Frederick, MD.

Kang, Emily L., Desheng Liu, and Noel Cressie. 2009. "Statistical Analysis of Small-Area Data Based on Independence, Spatial, Non-Hierarchical, and Hierarchical Models." *Computational Statistics & Data Analysis* 53: 3016–32. https://doi.org/10.1016/j.csda.2008.07.033.

Kay, Matthew. 2022. *tidybayes: Tidy Data and Geoms for Bayesian Models*. https://doi.org/10.5281/zenodo.1308151.

Lee, Duncan. 2013. "CARBayes: An R Package for Bayesian Spatial Modeling with Conditional Autoregressive Priors." *Journal of Statistical Software* 55 (13): 1–24. https://www.jstatsoft.org/htaccess.php?volume=55&type=i&issue=13.

LeSage, James P. 2014. "What Regional Scientists Need to Know about Spatial Econometrics." *The Review of Regional Studies* 44: 13–32.

Li, Honfei, Catherine A. Calder, and Noel Cressie. 2007. "Beyond Moran's I: Testing for Spatial Dependence Based on the Spatial Autoregressive Model." *Geographical Analysis* 39 (4): 357–75. https://doi.org/10.1111/j.1538-4632.2007.00708.x.

Logan, John R, Cici Bauer, Jun Ke, Hongwei Xu, and Fan Li. 2019. "Models for Small Area Estimation for

Census Tracts." *Geographical Analysis* 52 (3): 325–50. https://doi.org/10.1111/gean.12215.

Morris, Mitzi, Katherine Wheeler-Martin, Dan Simpson, Stephen J Mooney, Andrew Gelman, and Charles DiMaggio. 2019. "Bayesian Hierarchical Spatial Models: Implementing the Besag York Mollié Model in Stan." *Spatial and Spatio-Temporal Epidemiology* 31: 100301. https://doi.org/10.1016/j.sste.2019.100301.

Riebler, Andrea, Sigrunn H Sørbye, Daniel Simpson, and Håvard Rue. 2016. "An Intuitive Bayesian Spatial Model for Disease Mapping That Accounts for Scaling." *Statistical Methods in Medical Research* 25 (4): 1145–65. https://doi.org/10.1177/0962280216660421.

Rue, Håvard, Sara Martino, and Nicholas Chopin. 2009. "Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations (with Discussion)." *Journal of the Royal Statistical Society B* 71: 319–92. https://doi.org/10.1111/j.1467-9868.2008.00700.x.

Stan Development Team. 2022a. "RStan: The R Interface to Stan." https://mc-stan.org/.

———. 2022b. "Stan Modeling Language Users Guide and Reference Manual, 2.30." https://mc-stan.org/.

"Student" [W.S. Gausset]. 1914. "The Elimination of Spurious Correlation Due to Position in Time and Space." *Biometrika* 10: 179–80. https://doi.org/10.1093/biomet/10.1.179.

Umlauf, Nikolaus, Daniel Adler, Thomas Kneib, Stefan Lang, and Achim Zeileis. 2015. "Structured Additive Regression Models: An R Interface to BayesX." *Journal of Statistical Software* 63 (21): 1–46. https://www.jstatsoft.org/v63/i21/.

Xia, Hong, and Bradley P Carlin. 1998. "Spatio-Temporal Models with Errors in Covariates: Mapping Ohio Lung Cancer Mortality." *Statistics in Medicine* 17 (18): 2025–43. https://doi.org/10.1002/(sici)1097-0258(19980930)17:18%3C2025::aid-sim865%3E3.0.co;2-m.