The Journal of Open Source Software

geostan: An R package for Bayesian spatial analysis

Connor Donegan [1,2]

1 Geography and Geospatial Information Sciences, The University of Texas at Dallas **2** Peter O'Donnell Jr. School of Public Health, University of Texas Southwestern Medical Center

DOI: 10.21105/joss.04716

Software

- Review
- Repository
- Archive

Editor: Mehmet Hakan Satman

Reviewers:

- @wcjochem
- @becarioprecario

Submitted: 14 July 2022
Published: 11 November 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-backed 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 they require adjustment (Cressie, 2015; Haining & Li, 2020).

This paper introduces **geostan**, an R (R Core Team, 2021) package for analyzing spatial data using Bayesian inference. **geostan**'s spatial models were built using Stan, a platform for Markov chain Monte Carlo (MCMC) sampling (Gabry et al., 2020; Stan Development Team, 2022a, 2022b). 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 et al., 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 typically reduces effective sample size (ESS) and renders many standard statistical methods inappropriate (Clifford et al., 1989; "Student" [W.S. Gausset], 1914). 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 (Donegan et al., 2021; Folch et al., 2016), which can have real consequences on communities and service providers (Bazuin & 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 & Piras, 2015), **INLA** (Rue et al., 2009), **ngspatial** (Hughes & Cui, 2020), **BayesX** (Belitz et al., 2022; Umlauf et al., 2015), **CARBayes** (Lee, 2013), and **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 MCMC analysis.

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. The `shape2mat` function creates spatial weights matrices by first calling the **spdep** package (R. S. Bivand et al., 2013) to identify the adjacency structure of the spatial data, and results are returned to the user in sparse matrix format using the **Matrix** package (Bates et al., 2022).

geostan uses 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 et al., 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 et al., 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 & Griffith, 2013)
- Moran coefficient and Geary Ratio for measuring global spatial autocorrelation (Chun & 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 et al., 2007)
- Effective sample size (ESS) calculation (D. A. Griffith, 2005)

These tools are provided for exploratory analysis, but not for detection of clusters. For this and other reasons, p-values are not provided. Graphics are created with **ggplot2** (Wickham, 2016).

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.

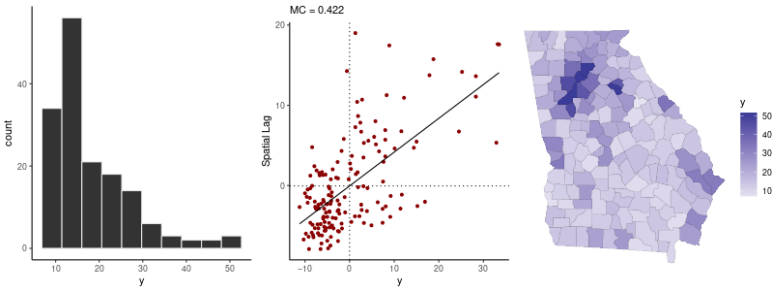

Figure 1: Spatial diagnostic summary for percent college educated, Georgia counties.

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 et al., 1991) and BYM2 specifications (Donegan & Morris, 2021; Morris et

al., 2019; Riebler et al., 2016), simultaneously-specified spatial autoregressive (SAR) models (Cliff & Ord, 1981) (which are referred to as the spatial error model (SEM) in the econometrics literature (LeSage, 2014)), and eigenvector spatial filtering (ESF) (Donegan et al., 2020; D. Griffith et al., 2019).

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

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 more than 3000 observations, such as U.S. county data (Donegan, 2021).

A set of functions for working with model results conveniently extracts 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 et al., 2021; cf. Bernardinelli et al., 1997; Kang et al., 2009; Logan et al., 2019; Xia & Carlin, 1998). All uncertain inferences from the ME models are automatically propagated throughout the regression or a 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 from the Geography and Geospatial Information Sciences program at The University of Texas at Dallas.

References

Anselin, L. (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, D., Maechler, M., & Jagan, M. (2022). *Matrix: Sparse and dense matrix classes and methods*. https://R-Forge.R-project.org/projects/matrix/

Bazuin, J. T., & Fraser, J. C. (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, C., Brezger, A., Kneib, T., Lang, S., & Umlauf, N. (2022). *BayesX: Software for Bayesian inference in structured additive regression models*. https://www.uni-goettingen.

de/de/bayesx/550513.html

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%3C741::aid-sim501%3E3.0.co;2-1

Besag, J., York, J., & Mollié, A. (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, R. S., Pebesma, E., & Gomez-Rubio, V. (2013). *Applied spatial data analysis with R, second edition*. Springer, NY. https://asdar-book.org/

Bivand, R., & Piras, G. (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, Y., & Griffith, D. A. (2013). *Spatial statistics and geostatistics: Theory and applications for geographic information science and technology*. Sage.

Cliff, A., & Ord, J. (1981). *Spatial processes: Models and applications*. Pion.

Clifford, P., Richardson, S., & Hémon, D. (1989). Assessing the significance of the correlation between two spatial processes. *Biometrics*, *45*, 123–134. https://doi.org/10.2307/2532039

Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.

de Valpine, P., Turek, D., Paciorek, C., Anderson-Bergman, C., Temple Lang, D., & Bodik, R. (2017). Programming with models: Writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, *26*, 403–413. https://doi.org/10.1080/10618600.2016.1172487

Donegan, C. (2021). Building spatial conditional autoregressive (CAR) models in the Stan programming language. *OSF Preprints*. 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. https://doi.org/10.1016/j.spasta.2020.100450

Donegan, C., & Morris, M. (2021). *Flexible functions for ICAR, BYM, and BYM2 models in Stan*. https://github.com/ConnorDonegan/Stan-IAR (accessed on July 13, 2022).

Folch, D. C., Arribas-Bel, D., Koschinsky, J., & Spielman, S. E. (2016). Spatial variation in the quality of American Community Survey estimates. *Demography*, *53*, 1535–1554. https://doi.org/10.1007/s13524-016-0499-1

Gabry, J. (2018). *Shinystan: Interactive visual and numerical diagnostics and posterior analysis for Bayesian models*. https://CRAN.R-project.org/package=shinystan

Gabry, J., Goodrich, B., & Lysy, M. (2020). *Rstantools: Tools for developing R packages interfacing with 'Stan'*.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (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, D. A. (2005). Effective geographic sample size in the presence of spatial autocorrelation. *Annals of the Association of American Geographers*, *95*(4), 740–760. https://doi.org/10.1111/j.1467-8306.2005.00484.x

Griffith, D., Chun, Y., & Li, B. (2019). *Spatial regression analysis using eigenvector spatial filtering*. Academic Press. https://doi.org/10.1016/C2017-0-01015-7

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (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, R. P., & Li, G. (2020). *Modelling spatial and spatio-temporal data: A Bayesian approach*. CRC Press.

Hughes, J., & Cui, X. (2020). *ngspatial: Fitting the centered autologistic and sparse spatial generalized linear mixed models for areal data*.

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

Kay, M. (2022). *tidybayes: Tidy data and geoms for Bayesian models*. https://doi.org/10.5281/zenodo.1308151

Lee, D. (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, J. P. (2014). What regional scientists need to know about spatial econometrics. *The Review of Regional Studies*, *44*, 13–32. https://doi.org/10.2139/ssrn.2420725

Li, H., Calder, C. A., & Cressie, N. (2007). Beyond Moran's I: Testing for spatial dependence based on the spatial autoregressive model. *Geographical Analysis*, *39*(4), 357–375. https://doi.org/10.1111/j.1538-4632.2007.00708.x

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

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. https://doi.org/10.1016/j.sste.2019.100301

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Riebler, A., Sørbye, S. H., Simpson, D., & Rue, H. (2016). An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Statistical Methods in Medical Research*, *25*(4), 1145–1165. https://doi.org/10.1177/0962280216660421

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society B*, *71*, 319–392. 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/

Stan Development Team. (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*, 179–180. https://doi.org/10.1093/biomet/10.1.179

Umlauf, N., Adler, D., Kneib, T., Lang, S., & Zeileis, A. (2015). Structured additive regression models: An R interface to BayesX. *Journal of Statistical Software*, *63*(21), 1–46. https://www.jstatsoft.org/v63/i21/

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4

Xia, H., & Carlin, B. P. (1998). Spatio-temporal models with errors in covariates: Mapping Ohio lung cancer mortality. *Statistics in Medicine*, *17*(18), 2025–2043. https://doi.org/10.1002/(sici)1097-0258(19980930)17:18%3C2025::aid-sim865%3E3.0.co;2-m