**Remembering the Importance of Spatial Relationships in Regressions**

Keelin Haynes

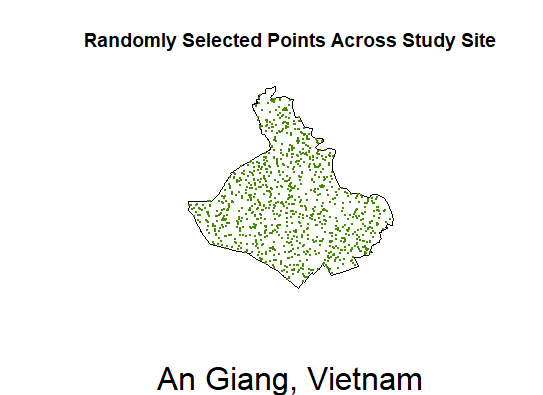
This paper explores the importance of incorporating spatial relationships in regression models. It discusses the importance of accounting both for collinearity and spatial autocorrelation within datasets. It uses a stepwise variance inflation factor analysis, ordinary least squares, spatial autoregressive and spatial error models on a custom made dataset to explore how soil flood drainage time is affected by difference environmental explanatory variables.

**Introduction**

Most statisticians understand the importance of accounting for collinearity and covariate significance when conducting analysis. What sometimes goes unaccounted for however is the role that spatial relationships can play in statistical analysis. Large amounts of research has gone into demonstrating the importance of accounting for spatial relationships in statistical analysis (Chun and Griffith, 2013; Darmofal, 2015). This tutorial demonstrates how to account for spatial relationships in analysis by using a custom dataset to see how the time that a environment is flooded is explained by various environmental explanatory variables. This paper lays out the results of the tests conducted in the tutorial, which is attached at the end of the document as a github commit containing a r markdown and html files, as well as data needed for the analysis.

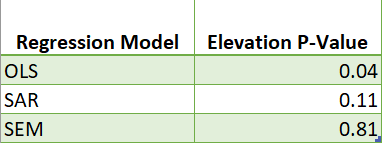
**Data and Methods**

The dataset used in the analysis was created by combining elevation data from the USGS and soil data from the Mekong River Commission (NASA JPL, 2013; and Mekong, 2002). The dataset used is a data frame containing 1,000 observations, with the x and y coordinates for each observation, as well as elevation, slope, anion fixation, mineral reserve, inundation depth, inundation duration, drainage, soil stability, topsoil texture, and soil depth information.



Our analysis explores how inundation duration is affected by the other explanatory factors. As with other statistical analysis, we first run through how to determine collinearity of covariates and eliminate those with high levels of collinearity. This is done in the tutorial using a stepwise variance inflation factor (VIF) analysis (Marquardt, 1970). The VIF showed that soil depth and topsoil texture should be eliminated from the analysis as they showed high collinearity with the other variables. Eliminating those two, we were left with seven covariates.

With the best covariates selected for analysis, the tutorial then demonstrates how to run an ordinary least squares (OLS) regression (Wilkinson et al, 1973); spatial autoregressive (SAR) regression (Wall, 2002); and spatial error model (SEM) regression (Bivand, 2015). The regression analysis showed varying levels of significance for each of the covariates, however of particular note is how elevation is interpreted by the models. In the OLS and SAR, elevation had a p value indicating significance, whereas the SEM showed that the elevation is not significant.



Moving forward from this analysis, it would be interesting to explore if these results are reproducible in both nearby and distant locales, and if these covariate significances vary by location. Further it is important to note that the soil data used in this analysis was derived from vector data that was converted to a raster format, before having sample points taken from the raster object. This conversion of the data could have caused issues of projection and the location of points being slightly off.

**Conclusion**

Statisticians working with spatial data must not only remember to account for covariate collinearity, but the impact the spatial relationships can have on the results of analysis. Additionally, it is important to remember that any analysis that is being performed is only as good as the data that is being analyzed.

A r markdown file, an html containing the tutorial and results, as well as the data needed for the tutorial can be found here:

<https://github.com/keelindhaynes/SpatialRegressioninR>

*For further information on this project contact:*

PI: Keelin Haynes

Affiliation: Dept. of Geography, Miami University

Email: hayneskd@miamioh.edu

**References**

Bivand, Roger, Piras, Gianfranco (2015). Comparing

Implementations of Estimation Methods for Spatial Econometrics. Journal of Statistical Software, 63(18), 1-36. <https://www.jstatsoft.org/v63/i18/>.

Chun, Y. and Griffith, D.A. 2013. Spatial Statistics &

Geostatistics . SAGE Publications Inc., Thousand Oaks, CA, 181 pages. ISBN: 9781446201749.

Darmofal, D. (2015) “Spatial Lag and Spatial Error

Models,” in *Spatial Analysis for the Social*

*Sciences*. Cambridge: Cambridge University Press (Analytical Methods for Social Research), pp. 96–118. doi: 10.1017/CBO9781139051293.007.

Marquardt, D. (1970). Generalized Inverses, Ridge

Regression, Biased Linear Estimation, and

Nonlinear Estimation. *Technometrics,* *12*(3), 591-612. doi:10.2307/1267205

Mekong River Commission. Soil Map of the Lower

Mekong Basin. 2002.

NASA JPL (2013). NASA Shuttle Radar Topography

Mission Global 1 arc second [Data set]. NASA EOSDIS Land Processes DAAC. Accessed 2019-12-13 from <https://doi.org/10.5067/MEaSUREs/SRTM/SRTMGL1.003>

Wall, Melanie M. “A Close Look at the Spatial Structure Implied by the CAR and SAR Models.” *Journal of Statistical Planning and Inference*, vol. 121, no. 2, 2004, pp. 311–324., doi:10.1016/s0378-3758(03)00111-3.

Wilkinson, G. N. and Rogers, C. E. (1973). Symbolic

descriptions of factorial models for analysis of variance. Applied Statistics, **22**, 392--399. 10.2307/2346786.