

## **Exploratory Spatial Data Analysis: Semivariogram Clouds and Plots**

Michael G. McManus (he/him), U.S. EPA

[mcmamus.michael@epa.gov](mailto:mcmamus.michael@epa.gov)

*The views expressed in this presentation are those of the author and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency.*

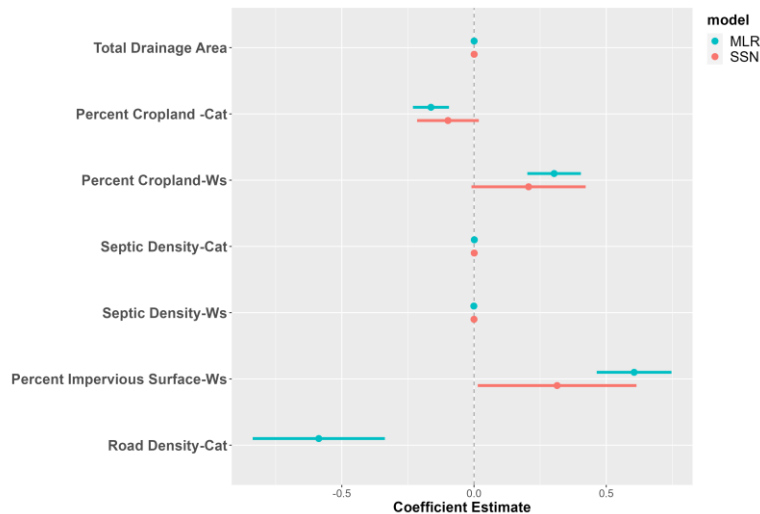
Exploratory spatial data analysis is meant answer the question: do things near each resemble each other? Think real estate, do homes that are nearer to each have more similar sale values than homes that are further apart? Think air or water quality monitoring, do stations that are near each other have similar pollutant concentrations than sites that are further apart? When spatial features that are near each resemble each other in some characteristic that is call spatial autocorrelation.

**“Spatial data are often analyzed as if each observation were independent of one another, which is generally untrue. Incorporating spatial dependence into statistical models can uncover more realistic ecological and environmental patterns and notably improve predictions at unobserved locations”.**

**Mike Dumelle  
Spatial Statistician  
U.S. EPA/ORD  
To ORD R-User Group  
June 26, 2024**

Motivating quote

## Slope Estimates from a Multiple Linear Regression and Spatial Models with 95% Confidence Limits

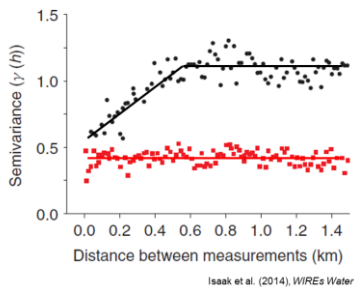


Office of Research and Development

Motivating example. There were 600 observations from 163 monitoring stations in this dataset from an Indiana watershed where we are studying total phosphorous concentrations as a function of landcover variables from EPA's StreamCat dataset. The MLR assumed those observations were independent. We see small confidence limits, the blue lines, around the slope estimates, the blue points. The spatial stream network model, a type of spatial linear model, did not assume independence and, instead, incorporated spatial autocorrelation. We get larger confidence limits, the red lines. Why is that?

## Objectives

1. Work with spatial data in R as a simple feature objects and shapefiles and use a geopackage geodatabase to organize our spatial features.
2. How to make and interpret a semivariogram plot.



This is what spatial modeling, random forest regression kriging, spatial stream network models, and other spatial data analysis methods handle.

This is what multiple linear regression, random forests, and many data analysis methods assume.

Incorporating spatial dependence can produce better fitting models that make better predictions.