UPR
Recinto Universitario de Mayagüez

Proposal Title

Master in Statistics Mathematics

Alejandro M. Ouslan

Supervisors:
Dr. Raul E. Macchiavelli
Dra. Damaris Santana
Dr. Julio C. Hernandez
Dr. Roberto Rivera Santiago

University of Puerto Rico, Mayaguez

September 19, 2025

Abstract

This research looks to compare the preformance of Spatial regressions using a predefined weights matrix and semi parametric regessions with a spatila smoother

1 Proposal Keywords

Spatial simulation, Spatial Regressions, GAMs, Tensor Products

2 Introduction

Regressions are use to simplfy and model complex relationship of the real world. Every model is wrong and we add additional compones with the objective of better representing the real world relationship. The Spatial models shown in this paper are no different and are not a nobel consept. Through the use of simulations we are trying to understand the limitations of this model and understand what are their limitations and show on what context one should pick thouse models. Later next we would use some Bayesian inferecne to address one of the underlining problems using one of the models. Finaly we will implement the model using data from the QCEW as an applied example of the methods.

3 Background and Motivation

The spatial regersion are a very simple model with very express. The SDM can be express in the following model:

4 Systematic Literature Review

5 Aims and Objectives

The overall objectives of this research are to understand and compare the SDM model and Semi parametric model using tensor products. We intend to look at on which surcumstances does each model preform better than the other and generate better predictors closer to the truth. Additionaly we look to the cost benefits of each model like the dificulty of immplementing teh model, asumptions, computational cost and other other observations.

$$y = X\beta + \rho WX + \epsilon \tag{1}$$

Where we have our classic liner componet of $X\beta$ plus our spatial componet ρWX. This is the data generating process,the W is a matrix that the researcher knows and chooses. However in practice the researcher does not know how this process is conducted and needs to infer what type of realationships does W represent. For example this W can be represened in the following ways:

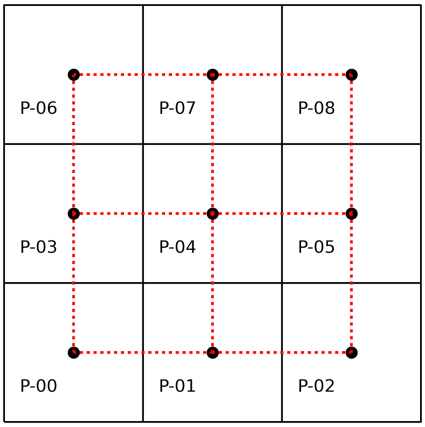

Figure 1: Rook Model

Represented mathematicaly this would be the following matrix

$$\begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

Another popular model is the Queens model which model all neighbors that are touching bordars

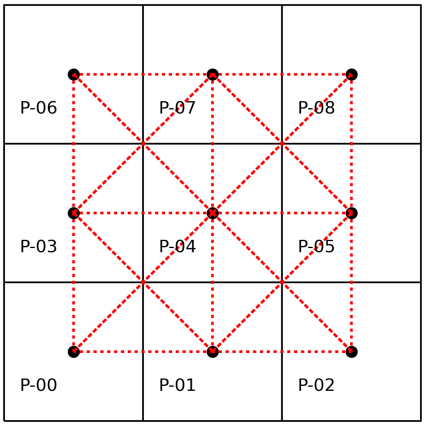

Figure 2: Queens Model

Which has the following matrix attach to ti:

$$\begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \end{bmatrix}$$

There are also much more compress modell like the KNN model which followis the proces decreasing weights as you become further way by K units away. You can see it in the folowing graph.

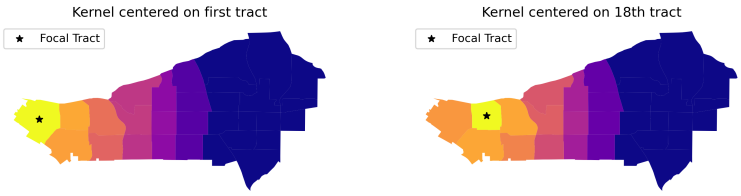

Figure 3: KNN Model

It is clear to see that there numeruse ways to represent this W matrix and no clear methode to deside which is apropiete given that in practice W is actualy an unknow given that when do not know the data generating proces.

6 Research Plan and Methodology

Starting for a simple regression of ordinary least squares (OLS) can be defined as follows:

$$y_i = \alpha + \sum_{i=1}^{p} x_i \beta_i + \epsilon \tag{2}$$

where y is our expicatory variable and x is the independent variables. If we wanted to study the corrolation that a location i has on its neighbors, the classic approach would be to define a relationship matrix W which details how each location relates to all other locations. This classic Spatial regressions are an extension of the normal regresion with with the addition of a spatialy term, which can be defined as follows:

$$y = X\beta + \rho WX + \epsilon \tag{3}$$

expanding the matrixes we get the following:

$$y_{it} = \alpha + \sum_{i=1}^{p} x_{it} \beta_i + \rho \sum_{j=1}^{N} w_{ij} x_{jt} + \epsilon \tag{4}$$

This spatial can be addapted spatialy control for either your dependent term

$$y = \alpha + X\beta + \rho WY + \epsilon \tag{5}$$

Or even the error term

$$\begin{aligned} y &= \alpha + X\beta + u \\ u &= \gamma W u + \epsilon \end{aligned} \tag{6}$$

Continuing from the SDM we can express the model a non linear equation as follows:

$$y = \alpha + \sum_{i=1}^{p} x_{it} \beta_i + f(C_i) + \epsilon \tag{7}$$

where C is the centroid of the observations and $f(C)$ is a function given the centroid of the individuals.

The overall hypothesis is whether the semiparametric methode preforms better on average given that we do not have reason to believe what is W. We expect to find what are to mesure the effects of picking a wrong W and wheather we can miticate them with using a spetial smoother.

7 Prototype Design and Implementation

The Strategy for this resarch is to simulate data in the folowing format:

$$y \sim \alpha + \sum_{i=1}^{p} x_{it} \beta_i + \rho \sum_{j=1}^{N} w_{ij} x_{jt} + \epsilon \tag{8}$$

Where $\epsilon \sim N(0, \sigma^2$ and W is defined in multiple ways this could be show in the following examples:

8 Success and Impact

The SDM generaly carries the problem that is very influenced on how W is defined and there is now systematic methedology to pick an apropiet W. This task is mostly delegated to domain experties. Given that the semiparametric model proposed does not depend on W this research intends to look if this model has a better preformance than the SDM model. Preformance of this model is primarly defined as:

$$\frac{\sum_{n=1}^{N} (y_n - \hat{y}_{nSDM})^2}{N}; \quad \frac{(y - \hat{y}_{GAM})^2}{N} \tag{9}$$

where N is the number of simulations, y is the predetermine outcome variable that we picked before the simulations are run. In addition we would also use look at the performance of β which can be shown in the simal maner:

$$\frac{(\beta - \hat{\beta}_{SDM})^2}{N}; \quad \frac{(\beta - \hat{\beta}_{GAM})^2}{N} \tag{10}$$