Recinto Universitario de Mayagüez

Modeling Spatial Dependence: A Simulation-Based Comparison of Parametric and Semi-Parametric Approaches

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, Mayagüez

October 29, 2025

Abstract

This research aims to compare the performance of spatial regression models that rely on predefined weight matrices with that of semi-parametric regression models using spatial smoothers.

1 Proposal Keywords

Spatial simulation, Spatial Regressions, GAMs, Tensor Products

2 Introduction

Regression models are commonly used to simplify and represent complex real-world relationships. Since all models are approximations, additional components are often introduced to better capture underlying patterns. Spatial models are no exception and are not a novel concept. Through simulation, this research seeks to understand the limitations of various spatial models and identify contexts in which specific models are preferable.

We also aim to address key issues using Bayesian inference on one of the models. As a practical application, we will implement the models using data from the Quarterly Census of Employment and Wages (QCEW).

3 Background and Motivation

Spatial regression models are relatively simple yet expressive tools. The Spatial Durbin Model (SDM) can be expressed as:

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

In this formulation, $X\beta$ represents the standard linear component, and $\rho W X$ incorporates spatial influence through a known weight matrix W. However, in real-world applications, W is not truly known and must be inferred or chosen by the researcher. For example, W can be represented in different forms, such as:

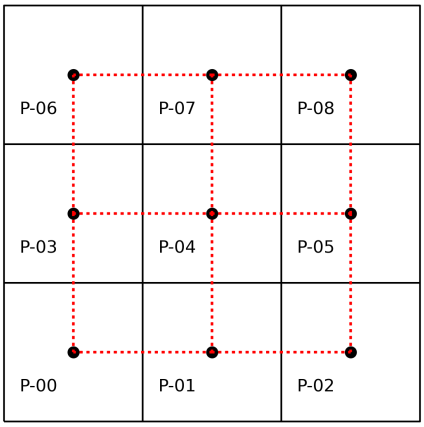

Figure 1: Rook Contiguity Matrix

Mathematically, the rook model can be represented as:

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

Another popular variant is the Queen's contiguity model, which includes all bordering neighbors:

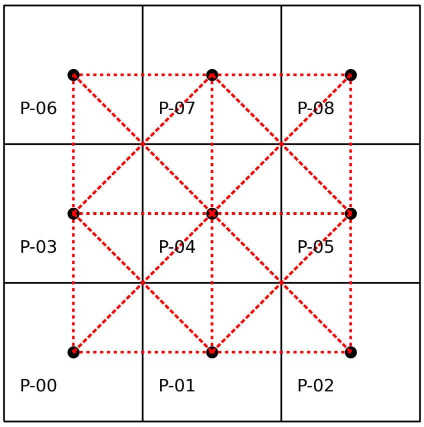

Figure 2: Queen Contiguity Matrix

Its corresponding matrix:

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

Another approach is the k-nearest neighbors (KNN) model, where influence decreases with distance:

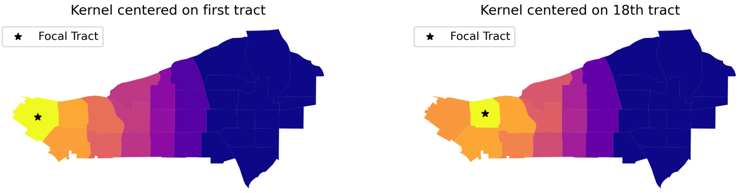

Figure 3: K-Nearest Neighbors (KNN) Model

Clearly, there are numerous ways to define the W matrix, and no universally accepted method exists for choosing the most appropriate one. In practice, W is unknown and selecting it remains a major modeling challenge.

4 Systematic Literature Review

5 Aims and Objectives

The primary objective of this research is to understand and compare the Spatial Durbin Model (SDM) and a semi-parametric model using tensor product smoothers. We will investigate under what circumstances each model performs better in terms of prediction accuracy and reliability. Furthermore, we aim to evaluate trade-offs such as ease of implementation, assumptions, computational cost, and interpretability.

We restate the SDM as:

$$y = X\beta + \rho W X + \epsilon \tag{2}$$

Where W is predefined by the researcher, though in practice the actual spatial process remains unknown. Choosing an appropriate W is often left to domain experts.

6 Research Plan and Methodology

6.1 Model Comparison

We begin with the classical Ordinary Least Squares (OLS) regression:

$$y_i = \alpha + \sum_{i=1}^{p} x_i \beta_i + \epsilon \tag{3}$$

To account for spatial correlation, we introduce a spatial weight matrix W into the regression:

$$y = X\beta + \rho W X + \epsilon \tag{4}$$

Which expands to:

$$y_{it} = \alpha + \sum_{i=1}^{p} x_{it} \beta_i + \rho \sum_{j=1}^{N} w_{ij} x_{jt} + \epsilon \tag{5}$$

Alternatively, spatial terms may be incorporated into the dependent variable:

$$y = \alpha + X\beta + \rho W Y + \epsilon \tag{6}$$

Or into the error structure:

$$\begin{aligned} y &= \alpha + X\beta + u \\ u &= \gamma W u + \epsilon \end{aligned} \tag{7}$$

The semi-parametric model with a spatial smoother can be expressed as:

$$y = \alpha + \sum_{i=1}^{p} x_{it} \beta_i + f(C_i) + \epsilon \tag{8}$$

Where C_i is the centroid of each observation and $f(C_i)$ is a spatial smoothing function.

We hypothesize that semi-parametric models may outperform spatial regression models, especially when the true W matrix is unknown or misspecified.

6.2 Spatial Smoother

Previously it was motioned that $f(C_i)$ would be the penalized spatial smoother used as an alternative to using the weight to remove the spatial correlation. Stratign from the basis function we can express any X as a sum of cubic function. This is can be showen in the following equation:

$$f(x) = \sum_{j=1}^{k} b_j(x) \beta_j \tag{9}$$

This would generate the following graph.

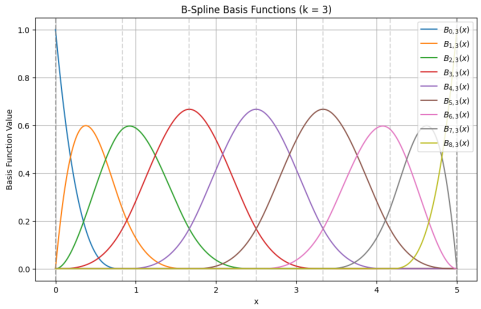

Figure 4: Basis spline functions

Addapting to include other variables can be done by first selecting the basis function for each factor.

$$f_x(x) = \sum_{j=1}^{k} b_j(x)\beta_j \quad f_z(z) = \sum_{j=1}^{k} d_j(z)\delta_j$$

we can then let $b_j(x)$ vary smoothly accoriding to z this give us the following:

$$\beta_i(z) = \sum_{l=1}^{L} \delta_{jl} d_l(z)$$

which gives:

$$f_{xz}(x,z) = \sum_{j=1}^{k} \sum_{l=1}^{L} \delta_{jl} d_l(z) \beta_i(x) \tag{10}$$

implementing this would give the following grpah

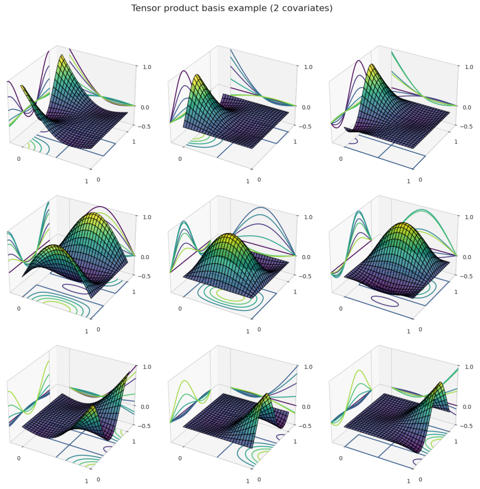

Figure 5: Basis spline functions

using a penalized of the tensor folloss the following format:

$$\min ||y + f_{xz}||_2^2 + \lambda ||f_{xz}||_2^2$$

6.3 Cobb-Douglas Implementation

An implementation of this method were it would be of interest is in the production function Cobb-Douglas where there is not much literature in the effects of spatial component. The Cobb-Douglas is given by the following equation:

$$Y = AK^{\alpha}L^{\beta}$$

Where:

(a) Y is the total output (real GDP) produced.

(b) A is total factor productivity (TFP), capturing the efficiency with which inputs are used.

(c) K is the quantity of physical capital used in production.

(d) L is the quantity of labor employed.

(e) α is the output elasticity of capital, representing the percentage change in output resulting from a 1% change in capital, holding labor constant.

(f) β is the output elasticity of labor, representing the percentage change in output resulting from a 1% change in labor, holding capital constant.

Then we add an error term ϵ where it comes from a normal distribution with mean 0 and constant variance σ^2

$$\log(Y) = \log(A) + \alpha\log(K) + \beta\log(L) + \epsilon; \quad \epsilon \sim N(0, \sigma^2) \tag{11}$$

This functional form assumes constant returns to scale if $\alpha+\beta = 1$, increasing returns to scale if $\alpha+\beta > 1$, and decreasing returns to scale if $\alpha + \beta < 1$.

Turnig this to a stochastic model we convert log both side to get the following:

$$\begin{aligned} Y &= AK^{\alpha}L^{\beta} \\ \log(Y) &= \log(AK^{\alpha}L^{\beta}) \\ \log(Y) &= \log(A) + \alpha\log(K) + \beta\log(L) + \epsilon \end{aligned}$$

Returning to the original form would give:

$$Y = AK^{\alpha}L^{\beta}e^{\epsilon} \tag{12}$$

Now Turning the Cobb-Douglas into a panel form is given by the follwoing model:

$$y_{it} = \mu_i + \gamma_t + \alpha k_{it} + \beta l_{it} + \epsilon_{it} \tag{13}$$

where:

(a) $\log(A)_{it} \equiv \mu_i + \gamma_t + \epsilon_{it}$

(b) $\log(L)_{it} \equiv l_{it}$

(c) $\log(K)_{it} \equiv k_{it}$

Using the notation form the previos section we can add the spatial componant to the Cobb-Douglas function and returning as the orinal function is as follwos:

$$\begin{aligned} y_{it} &= \mu_i + \gamma_t + \alpha k_{it} + \beta l_{it} + \rho\sum_{j=1}^{N} w_{ij}y_{jt} + \epsilon_{it} \\ \log(Y_{it}) &= \log(A_{it}) + \alpha\log(K_{it}) + \beta\log(L_{it}) + \rho\sum_{j=1}^{N} w_{ij}\log(Y_{jt}) \\ Y_{it} &= A_{it}K_{it}^{\alpha}L_{it}^{\beta}\exp\left\{\rho\sum_{j=1}^{N} w_{ij}\log(Y_{jt})\right\} \\ Y_{it} &= A_{it}K_{it}^{\alpha}L_{it}^{\beta}e^{\rho\sum_{j=1}^{N} w_{ij}\log(Y_{jt})} \end{aligned}$$

$$Y_{it} = A_{it}K_{it}^{\alpha}L_{it}^{\beta}\prod_{j=1}^{N} Y_{it}^{\rho w_{ij}} \tag{14}$$

7 Prototype Design and Implementation

We will simulate data using the following structure:

$$y \sim \alpha + \sum_{i=1}^{p} x_{it}\beta_i + \rho \sum_{j=1}^{N} w_{ij} y_{jt} + \epsilon \tag{15}$$

With $\epsilon \sim N(0, \sigma^2)$ and multiple forms of W (rook, queen, KNN, etc.) used for robustness testing.

8 Success and Impact

SDM models are highly sensitive to the choice of W, for which no standardized selection method exists. Often, domain expertise is required. The semi-parametric model, however, does not depend on W, making it potentially more robust.

We will evaluate model performance using mean squared error (MSE) of predicted outcomes:

$$\frac{\sum_{n=1}^{N} (y_n - \hat{y}_{nSDM})^2}{N}; \quad \frac{\sum_{n=1}^{N} (y - \hat{y}_{GAM})^2}{N} \tag{16}$$

And by comparing estimated vs. true coefficients:

$$\frac{\sum_{n=1}^{N} (\beta - \hat{\beta}_{SDM})^2}{N}; \quad \frac{\sum_{n=1}^{N} (\beta - \hat{\beta}_{GAM})^2}{N} \tag{17}$$