UPR
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

February 10, 2026

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

Economic 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. Especially in micro economices where the assumption of independent and identically distributed (i.i.d) can not be reasonably assumed due to its time factor or close proximity making them influence each outer. Spatial models 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.

More specifically in this research we seek to find reasonable strategies to best specify spatial models, where are it can be reasonably be presumed that there is some spatial effect but there is no clear answer on how spatial agents relate to each other.

3 Background and Motivation

Starting with out, economic theory tells that you can model the productivity of a given economy as a function of labor (L) and capital (K) this can be model in the following equation.

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

Where:

- Y is the total output (real GDP) produced.
- A is total factor productivity (TFP), capturing the efficiency with which inputs are used.
- K is the quantity of physical capital used in production.
- L is the quantity of labor employed.
- α is the output elasticity of capital, representing the percentage change in output resulting from a 1% change in capital, holding labor constant.
- β is the output elasticity of labor, representing the percentage change in output resulting from a 1% change in labor, holding capital constant.

In practice, though varios data sources L, K and Y are known and α and β are unknown. The unknown parameters are of special interest to economist and for public policy as they highlight areas where local government could invest to increase productivity. However as this models are representations of more complex models is reasonable to asume there is some unseen components in our models. An example of this could be the pretense of underground economies, this referring to transactions that could be reflected in the productivity but can not be reasonably measured through the official means and this could introduce uncertainty into our model.

Turning the Cobb-Douglas equation into a stochastic model can be done in the following format:

$$\begin{aligned} Y &= AK^{\alpha}L^{\beta} \\ \log(Y) &= \log\left(AK^{\alpha}L^{\beta}\right) \\ \log(Y) &= \log(A) + \alpha\log(K) + \beta\log(L) + \epsilon; \epsilon \sim N(0, \sigma) \end{aligned}$$

Returning to the original form would give:

$$Y = AK^{\alpha}L^{\beta}e^{\epsilon} \tag{1}$$

3.1 Spatial Modeling

Turning the Cobb-Douglas into a stochastic model that follows the conventual linear regression gives access to the ability to introduce space into the model. However it is important to define space in this context. Spatial regression models are relatively simple yet expressive tools. The Spatial Durbin Model (SDM) can be expressed as:

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

In this formulation, $X\beta$ represents the standard linear component, and ρWX 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:

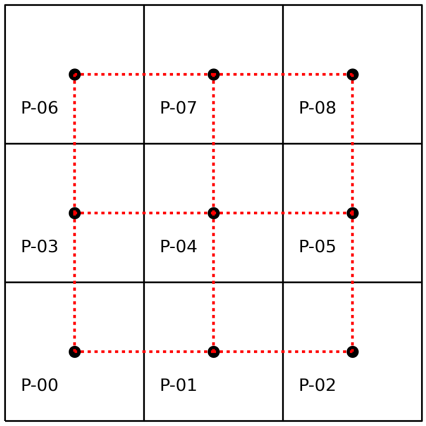

Figure 1: Rook Contiguity Matrix

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:

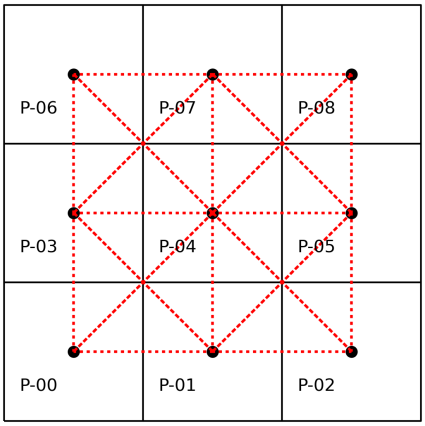

Figure 2: Queen Contiguity Matrix

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:

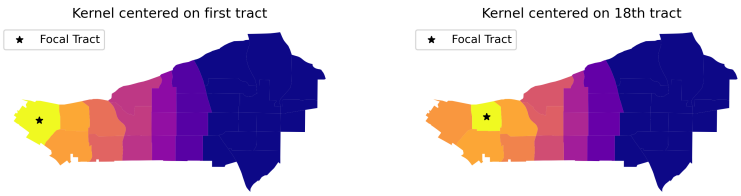

Figure 3: K-Nearest Neighbors (KNN) Model

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.

3.2 spatial Cobb-Douglas

To introduce space into the Cobb-Douglas we tu turn the model into panel. This is shown in the following model:

$$y_{it} = \mu_i + \gamma_t + \alpha k_{it} + \beta l_{it} + \epsilon_{it} \tag{3}$$

Where:

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

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

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

Contextualizing the pannel form of the Cobb-Douglas tell us that you can deconstroc a given economy into smaller sub economies that share the same elasticities for their labor and capital but differ in their quantities of capital and labor. 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:

$$y_{it} = \mu_i + \gamma_t + \alpha k_{it} + \beta l_{it} + \rho \sum_{j=1}^{N} w_{ij} y_{jt} + \epsilon_{it}$$

$$\log(Y_{it}) = \log(A_{it}) + \alpha \log(K_{it}) + \beta \log(L_{it}) + \rho \sum_{j=1}^{N} w_{ij} \log(Y_{jt})$$

$$Y_{it} = A_{it} K_{it}^{\alpha} L_{it}^{\beta} \exp\left\{ \rho \sum_{j=1}^{N} w_{ij} \log(Y_{jt}) \right\}$$

$$Y_{it} = A_{it} K_{it}^{\alpha} L_{it}^{\beta} e^{\rho \sum_{j=1}^{N} w_{ij} \log(Y_{jt})}$$

$$Y_{it} = A_{it} K_{it}^{\alpha} L_{it}^{\beta} \prod_{j=1}^{N} Y_{it}^{\rho w_{ij}} \tag{4}$$

Introducing space in this way it tell how this smaller economies are affected by the overall output of neighbors with a given relationship W. However as expalain previously there is no universally accepted relationship W and it is possible that depending the economy looked at this W might be compleatly different and more complex that would be virtually imposible to pick an appropriate W for the model.

4 Aims and Objectives

The primary objective of this research is to examine and compare the performance of the Spatial Autoregressive Model (SAR) of the Cobb-Douglas with that of a semi-parametric regression model that incorporates tensor-product smoothers. The analysis focuses on identifying the conditions under which each modeling approach yields superior predictive accuracy and more reliable inference. In addition, this study evaluates key trade-offs between the two approaches, including ease of implementation, underlying assumptions, computational cost, and interpretability.

5 Research Plan and Methodology

5.1 Model Comparison

We begin with the classical Ordinary Least Squares (OLS) regression model, specified as

$$y_i = \alpha + \sum_{j=1}^{p} x_{ij} \beta_j + \epsilon_i, \tag{5}$$

where the error terms are assumed to be independently and identically distributed.

To account for spatial dependence, a spatial weights matrix W is incorporated into the regression framework. One such specification is the Spatial Durbin Model (SDM), given by

$$y = X\beta + \rho WX + \epsilon, \tag{6}$$

which expands to

$$y_{it} = \alpha + \sum_{j=1}^{p} x_{it,j} \beta_j + \rho \sum_{k=1}^{N} w_{ik} x_{kt} + \epsilon_{it}. \tag{7}$$

Alternative spatial specifications incorporate spatial dependence directly into the response variable, as in the Spatial Autoregressive (SAR) model,

$$y = \alpha + X\beta + \rho WY + \epsilon, \tag{8}$$

or into the error structure, as in the Spatial Error Model (SEM),

$$\begin{aligned} y &= \alpha + X\beta + u, \\ u &= \gamma W u + \epsilon. \end{aligned} \tag{9}$$

As an alternative to parametric spatial regression, we consider a semi-parametric specification that incorporates spatial dependence through a smooth function of location:

$$y = \alpha + \sum_{j=1}^{p} x_{it,j}\beta_j + f(C_i) + \epsilon, \tag{10}$$

where C_i denotes the centroid coordinates of observation i, and $f(C_i)$ is a spatial smoothing function constructed using tensor-product splines.

We hypothesize that the semi-parametric model may outperform parametric spatial regression models when the true spatial weights matrix W is unknown or misspecified, particularly in terms of predictive accuracy and robustness.

5.2 Spatial Smoother

As previously noted, the term $f(C_i)$ represents a penalized spatial smoother used as an alternative to specifying a spatial weights matrix in order to account for spatial correlation. Beginning with a univariate basis expansion, any smooth function of a single covariate can be expressed as a linear combination of spline basis functions. In particular, a cubic spline representation may be written as

$$f(x) = \sum_{j=1}^{k} b_j(x)\,\beta_j. \tag{11}$$

This representation generates a smooth curve constructed from a set of local basis functions, as illustrated in Figure 4.

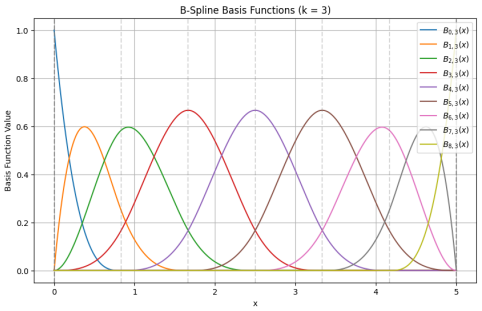

Figure 4: Basis spline functions

To extend this framework to multiple variables, a separate basis expansion is first specified for each covariate. For two variables x and z, these expansions take the form

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

The coefficients of the basis functions in one dimension may then be allowed to vary smoothly as a function of the other variable. Specifically, letting the coefficients β_j vary smoothly with respect to z yields

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

Substituting this expression into the original basis expansion produces the tensor–product smoother

$$f_{xz}(x,z) = \sum_{j=1}^{k}\sum_{l=1}^{L} \delta_{jl}\, b_j(x)\, d_l(z). \tag{12}$$

Implementing this construction results in a smooth surface over the joint domain of (x, z), as illustrated in Figure 5.

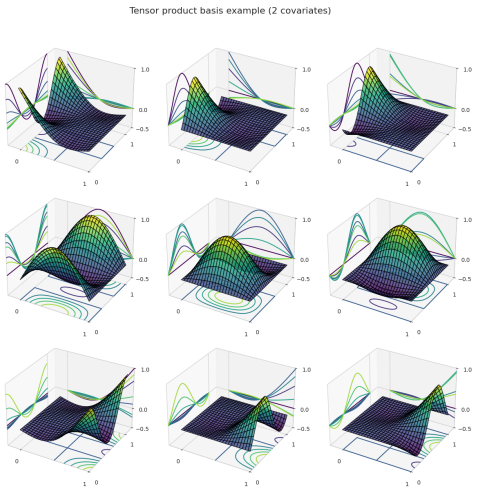

Figure 5: Tensor-product spline surface

The tensor–product smoother can be expressed compactly using the tensor (Kronecker) product design matrix

$$X_{\text{tensor}} = B \otimes D,$$

where B denotes the spline basis matrix for x and D denotes the spline basis matrix for z.

Estimation of the smoother coefficients is carried out by solving the ridge–regularized least-squares problem

$$\min_{\delta} \ \|y - X_{\text{tensor}}\delta\|_2^2 \ + \ \alpha \, \|\delta\|_2^2,$$

where the penalty parameter α controls the smoothness of the estimated surface.

5.3 Cobb–Douglas Interpretation

Incorporating a tensor-product spatial smoother into the Cobb–Douglas production function yields the following specification:

$$Y_{it} = K_{it}^{\alpha} L_{it}^{\beta} \exp\{f(\text{lat}_i, \text{lon}_i) + \gamma_t + \epsilon_{it}\}, \tag{13}$$

This equation serves as the primary model employed in this research. The parametric components preserve the standard Cobb–Douglas interpretation of output elasticities with respect to capital and labor, while the tensor-product function $f(\text{lat}_i, \text{lon}_i)$ flexibly captures spatial variation in productivity. By modeling spatial effects nonparametrically, this approach avoids the need to specify a spatial weights matrix and aims to produce more reliable and robust parameter estimates in the presence of unknown or complex spatial dependence.

5.4 Data Sources and Variable Specification

To apply the proposed models to the Puerto Rican economy, we utilize high-frequency sub-national data. Given the lack of official municipal-level GDP reports in real-time, the following proxies and data sources are employed:

- **Output (Y):** Following established econometric literature that uses electricity consumption as a high-correlation proxy for economic activity, we utilize municipal energy consumption data provided by **LUMA Energy**. This allows for a more granular temporal analysis than traditional annual proxies.
- **Capital (K):** Physical capital stocks are derived from the **Centro de Recaudación de Ingresos Municipales (CRIM)**. Specifically, data is extracted from the *Personal Property Tax Return (Planilla de Contribución sobre la Propiedad Mueble AS-29.1)*. Key indicators include:

- **Current Assets:** Inventory, materials, and supplies (Encasillado C).
- **Fixed Assets:** Machinery, equipment, and leasehold improvements.

- **Labor (L):** Labor input is measured using employment and wage data from the **Quarterly Census of Employment and Wages (QCEW)**. This dataset provides a comprehensive count of establishments, employment, and wages for workers covered by State unemployment insurance laws, categorized by North American Industry Classification System (NAICS) codes.

The integration of these datasets provides a panel structure desegregated by municipality (i) and time (t), which is essential for the spatial and semi-parametric models described in Equations 8 and 10.

6 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:

$$\frac{\sum_{n=1}^{N}(y_n - \hat{y}_{nSDM})^2}{N}; \quad \frac{\sum_{n=1}^{N}(y - \hat{y}_{GAM})^2}{N} \tag{14}$$

And by comparing estimated vs. true coefficients:

$$\frac{\sum_{n=1}^{N}(\beta - \hat{\beta}_{SDM})^2}{N}; \quad \frac{\sum_{n=1}^{N}(\beta - \hat{\beta}_{GAM})^2}{N} \tag{15}$$

7 Preliminary results

In the following results are for a regar regression model specified in the following way:

We consider a set of N observations generated from a spatial regression framework. Let y denote the response vector, $X = (X_1, X_2, X_3)$ the explanatory variables, and W a spatial weights matrix (Rook, Queen, or k-nearest neighbors with $k = 6$). The spatial Durbin model (SDM) used in the simulation study is given by

$$y = \rho W y + X\beta + \varepsilon, \tag{16}$$

where ρ is the spatial autoregressive parameter, β are the direct effects

The error term follows

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I_N).$$

The true parameter values in the data-generating process are

$$\beta = (4, 5, 6, 7)^\top, \qquad \rho = 0.7.$$

Looking over the results we see that the best preforming model is the Queens model as this is the model used to generate the data. The second best preforming model is the penalized tensor products. This shows that if there is no reasonable argument for picking the spatial weights matrix the penalized tensor products are a reasonable starting point for controlling for spatial variability.

Table 1: Simulation results for 1000 runs comparing spatial specifications

	Rook	**Queen**	**KNN (6)**	**Tensor**	**Base**
Intercept	3973.28	0.40	92.49	8252.62	7177.34
X_1	0.74	0.0032	0.30	0.21	1.22
X_2	0.90	0.0032	0.30	0.26	1.57
X_3	1.04	0.0031	0.39	0.27	1.92
ρ	0.43	0.000024	0.0060	–	–