Spatial Convolution Splines for Multivariate Spatial Data

Sackineh Shamil Jasim [*1]

[1] Department of Statistics - College of Administration and Economics -University of Karbala-Iraq
Emails: sackineh.sh@uokerbala.edu.iq

Abstract

The Spatial Convolution Splines Multivariate Regression Model (SCSMRM) were used on the data represented a diabetes disease measurements across different regions in Iraq (Basrah, Baghdad, Babylon, Sulaimanya) while considering multiple risk factors such as age, BMI, weight , income, education level, blood pressure for the same geographic location for (200) patient, and combine the health data with the risk factor data to create a comprehensive dataset. Each record in the dataset should include the geographic location, diabetes status, and values for each risk factor we applied (SCSMRM), the results showed that significant the model and the risk factors studied in the model explain 61% of the changes that occur in the diabetes. It also showed the significance of the factors (age - weight - body mass index (BMI) - educational level - blood pressure) and the non-significance of the variable (income), and these results are consistent with the actual reality of the disease.

Keywords: Spatial; Convolution; Splines; Regression; Multivariate Spatial Data.

1. Introduction

Convolutional neural networks are considered a type of deep machine learning that mimics the human brain, while traditional neural networks rely only on input matrices and do not use convolution on data. Convolutional neural networks have been trained on millions of images, but in the case of very large data and complex data, we need a faster time to complete the training. This network takes its idea from the communication between artificial neurons and natural cells of humans or animals, which achieves greater speed and lower error rates than traditional neural networks, as it is very effective in classification. This type of uncertainty has been applied in drug discovery and in identifying potential treatments by predicting interactions between biological molecules and proteins. This type of network was developed and published for the first time through the discovery of handwritten numbers. Since then, this type of network has been used to read postal codes and security numbers.

Over the last ten years, there has been a growing focus and effort on developing multivariate spatial models. The subject of multivariate spatial modeling is seeing significant growth. However, the majority of existing models are not practical for analyzing huge spatial processes. The use of "Spatial Convolution Splines" models to analyze multivariate spatial data is due to the ability of this model to capture the complex spatial relationships between different variables more accurately compared to traditional models. Spatial heterogeneity and correlations between sites can be modeled more effectively. There is also flexibility in modeling, as the use of Splines in the model allows for flexibility in the modeling process, making it possible to capture non-linear patterns and heterogeneous relationships in the data. The increasing amount of space-time datasets with several factors has prompted these endeavors. In the atmospheric sciences, for example, hundreds of state variables are included in climate models and weather forecasts, and these variables interact with one other in complex and subtle ways. Elevation and sea surface temperature are only two examples of the many variables that may be found in remote sensing datasets at extremely high geographic resolution. Spatial and health care econometric datasets, which are often limited to census tract levels,

Doi: https://doi.org/10.54216/GJMSA.0100207
Received: September 22, 2023 Revised: December 12, 2023 Accepted: January 12, 2024

and includes a wide range of measurements. Some diseases such as (heart, lung, diabetes, etc.) can be treated using multivariate spatial modeling. We need methods that are accurate in their results and represent complex interactions. We need optimal models to forecast dynamic spatial methods [1]. Presentation of a versatile set of non-stationary multivariate stochastic models for analyzing geographic data. William et al. [2] used many variables for geographic data through simulation. William et al., [2] presented a method that has multiple uses and resolutions, is simple, and is spatially scalable for multivariate.

2. Spatial convolution

If the time variable (t) is replaced by a function f that depends on the spatial variable (x), it turns into a spatial convolution process.

Convolution is an important method for every computational technique that performs smoothing, and when used on two-dimensional functions such as images, the analysis is based on image accuracy, movement recognition, image matching, etc. Convolution is basically defined as a method that combines two functions f(x) and g(x) on the continuous variable x [3]:

$$\mathrm{H(x)} \otimes \mathrm{L(x)} = \int_{-\infty}^{\infty} H(t).L(x-t)dt = \qquad (1)$$

Where $\otimes$ means convolution and (.) means ordinary multiplication. The integral may be seen as the outcome of convolving a function H(x) with a point spread function (PSF) L(x) and summing the contributions at each point. If the point spread function (PSF) is very narrow, preferably resembling a delta function, then the convolution will be exactly the same as the original function H(x). One may easily see the function H(x) as being expanded or distributed due to the impact of L(x). This reasoning may create the perception that convolution always results in the smudging of the original function. However, this is not always the case, particularly when the Point Spread Function (PSF) exhibits a distribution of both positive and negative values [4].

When convolution is used to digital pictures, the previous formulation undergoes two modifications: (1) a double integral is required for the two dimensions and (2) integration is replaced by discrete summation. The revised version of the convolution is [3]:

$$\mathbf{H(x,y)} * \mathbf{L(x.y)} = \int_{t_1=-\infty}^{\infty} \int_{t_2=-\infty}^{\infty} \boldsymbol{H(t_1,t_2).L(x-t_1,y-t_2)dt_1\,dt_2} \quad (2)$$

And,

$$\boldsymbol{\gamma(X,Y) = H(x,y) * L(x,y) = \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} H(i,j) * L(x-i,y-j)} \qquad (3)$$

L is now designated as a spatial convolution mask. The need to invert the mask before to its application is problematic for viewing the convolution process, especially when dealing with matching operations such as corner location [5].

3. Splines

Spline" functions are large groups of functions often employed in applications that need data interpolation or data smoothing. One or more dimensions may be present in the data. When determining the spline functions for interpolation, one often applies the interpolation constraints to identify the minimizers of appropriate roughness measures (such as integral squared curvature).

Doi: https://doi.org/10.54216/GJMSA.0100207
Received: September 22, 2023 Revised: December 12, 2023 Accepted: January 12, 2024

Smoothing splines may be seen as extensions of interpolation splines when they are selected to minimize a weighted combination of the average squared approximation error across observed data and the roughness measure. The finite dimensionality of spline functions has been shown to be applicable to several important formulations of the roughness measure. This characteristic is the primary factor contributing to their utility in representation and computation. In the rest of this section, we focus only on one-dimensional polynomial splines and adhere strictly to this specific definition of "spline" [6,7].

A spline is a kind of piecewise polynomial function. The function S maps values from the interval [a,b] to the set of real numbers, R. $\rho : [\boldsymbol{A}, \boldsymbol{B}] \in \boldsymbol{R}$

4. Spatial Convolution Splines (SCS)

Spatial convolution in statistics is a versatile tool used for smoothing, interpolating, detecting patterns, identifying edges, extracting features, detecting anomalies, processing images, and transforming spatial data. These applications are crucial across various fields such as environmental science, geology, urban planning, epidemiology, and remote sensing, providing deeper insights and more accurate analysis of spatial phenomena. This technique combines the concepts of spatial convolution and splines to enhance image representation and feature extraction [8].

5. Spatial Convolution Splines Multivariate Regression Model (SCSMRM)

The link between spatial convolution points and splines smoothing functions is intimate, as this relationship is represented by applying the convolution states to the original data, producing convolved data, and then applying the splines functions, which produces smoothed, smooth, and useful data that gives accurate results in estimation.

Let us have a set of spatial data consisting of a set of explanatory variables represented by a vector X with dimension pxn, where p number of explanatory variable, n sample size, and one dependent variable, Y, has a dimension of nx1, and this data is represented by the matrix Z. If we have a kernel function and let it be a Gaussian function, then the ordered pair (x, y), then the data convolution process for each location is as follows:

$$\boldsymbol{\gamma}(\mathbf{x},\mathbf{y}) = \boldsymbol{H}(\mathbf{x},\mathbf{y}) * \boldsymbol{L}(\mathbf{x},\mathbf{y}) = \sum_{\mathbf{i}=\frac{\mathbf{L}}{\mathbf{2}}}^{\frac{\mathbf{L}}{\mathbf{2}}} \sum_{\mathbf{j}=-\frac{\mathbf{L}}{\mathbf{2}}}^{\frac{\mathbf{L}}{\mathbf{2}}} \boldsymbol{Z}(\boldsymbol{x}+\boldsymbol{i},\boldsymbol{y}+\boldsymbol{j}))\boldsymbol{k}(\boldsymbol{i},\boldsymbol{j}) \qquad (\mathbf{4})$$

After convolution, spline interpolation is used to smooth the convolved features. Let {(xi, yi)} be the set of spatial locations, and Z′(xi, yi) be the convolved values. The spline S(x, y) is fitted to these points:

$$\rho(\boldsymbol{X},\boldsymbol{Y}) = \sum_{\mathbf{i}=\mathbf{1}}^{\mathbf{n}} \beta_{\boldsymbol{i}} \boldsymbol{B_i}(\boldsymbol{x},\boldsymbol{y}) \qquad \ldots (\mathbf{5})$$

B_i (x, y) basis functions for the spline, and$\beta_{\boldsymbol{i}}$ are the coefficients.

The response variable $\mathcal{G}$ at location (x, y) is modeled as a function of the smoothed (spline-interpolated) predictors as following:

$$\boldsymbol{\mathcal{G}}(\boldsymbol{x},\boldsymbol{y}) = \beta_{\mathbf{0}} + \sum_{\mathbf{p}=\mathbf{1}}^{\mathbf{P}} \beta_{\boldsymbol{p}} \rho_{\boldsymbol{p}}(\boldsymbol{X},\boldsymbol{Y}) + \boldsymbol{u}(\boldsymbol{x},\boldsymbol{y}) \qquad \ldots (\mathbf{6})$$

Where β0 intercept. $\beta_{\boldsymbol{p}}$the coefficients for each smoothed predictor $\boldsymbol{S_p}(\boldsymbol{X},\boldsymbol{Y})$ ϵ(x, y) is the error term.

Then fit the regression model using a suitable method Ordinary Least Squares to estimate the coefficients β_p and validation and evaluation the model using Cross-Validation by perform spatial

Doi: https://doi.org/10.54216/GJMSA.0100207
Received: September 22, 2023 Revised: December 12, 2023 Accepted: January 12, 2024

cross-validation to assess model performance, and model diagnostics by check for spatial autocorrelation in the residuals and other diagnostics to ensure model validity.

6. Applied data

The data represented a diabetes disease measurements across different regions in Iraq (Basrah, Baghdad, Babylon, Sulaimanya) while considering multiple risk factors such as age, BMI, weight , income, education level, blood pressure for the same geographic location for (200) patient, and combine the health data with the risk factor data to create a comprehensive dataset. Each record in the dataset should include the geographic location, diabetes status, and values for each risk factor we applied (SCSMRM) to analysis the risk factors in all regions and the results as following:

Table 1: Estimated regression coefficients and model parameters

Variable	Estimate	SE	tStat	P-value
Intercept	120.031	0.909	132.047	0.000
x_1	7.5567	0.198	38.165	0.000
x_2	22.5567	0.192	117.483	0.000
x_3	18.897	0.072	262.458	0.000
x_4	-0.1567	0.962	-0.163	0.789
x_5	-4.545	0.163	-27.883	0.001
x_6	12.1967	0.514	23.729	0.000
R^2= 0.605		R^2= 0.593		
F Sig.: 0.000	F Calculated.: 60.700		F Tabulated(5,193,1%): 3.114	

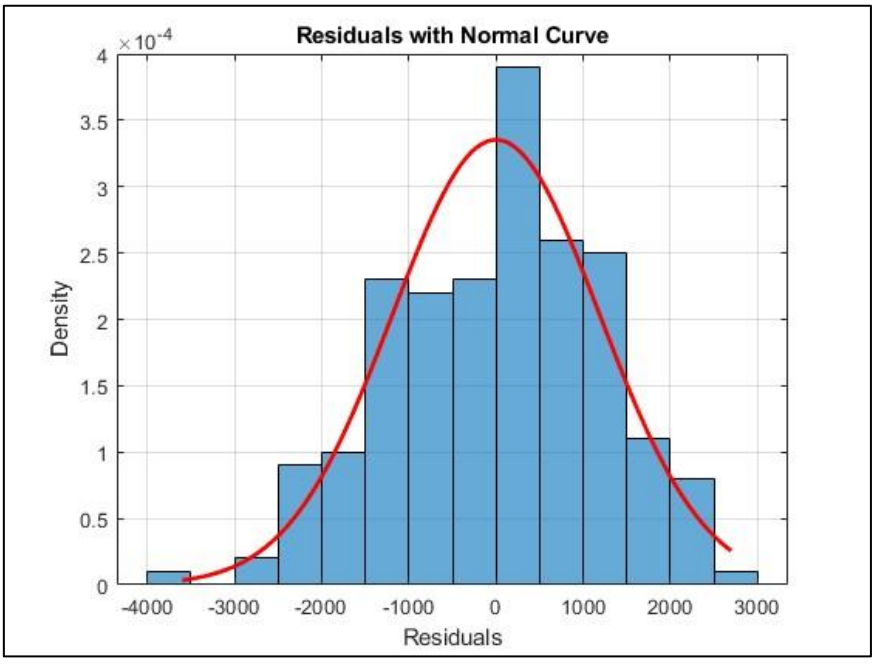

Figure 1: Residuals for the estimated model

Doi: https://doi.org/10.54216/GJMSA.0100207
Received: September 22, 2023 Revised: December 12, 2023 Accepted: January 12, 2024

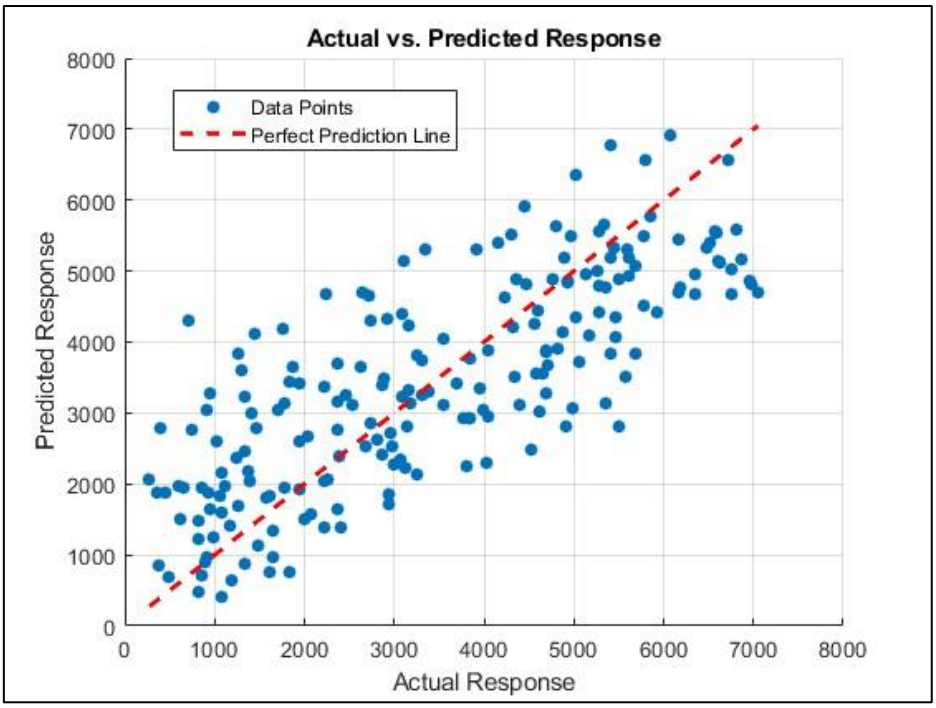

Figure 2: Actual values versus estimated values.

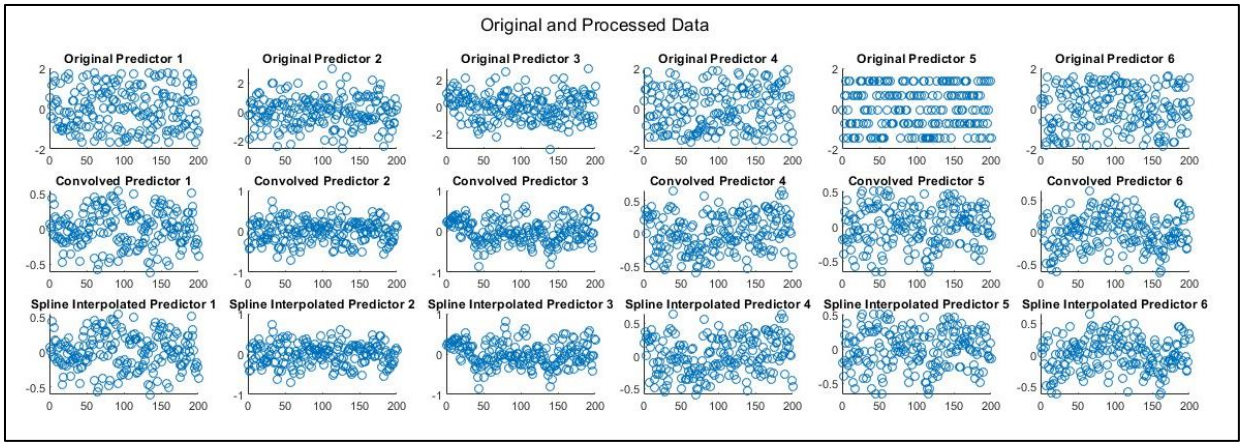

Figure 3: The original values, the convolved values, and the values interpreted by the Spline function for Predictors variables

7. Discussion of results

We note from Table (1) that the significance of the model estimated using the (SCSMRM) model is because the value of F Sig.: 0.000 is less than the significance level of 1%. The estimated model also showed that the coefficient of determination had a value of R-squared: 0.605, meaning that the risk factors studied in the model explain 61% of the changes that occur in the incidence of diabetes. It also showed the significance of the factors (age - weight - body mass index (BMI) - educational level - blood pressure) and the non-significance of the variable (income), and these results are consistent with the actual reality of the disease. What confirms the accuracy of the model is that the residuals have a normal distribution, as in Figure (1), and that the estimated values are highly consistent with the real values, as in Figure (2), and that the values interpreted by the Spline smoothing function are highly accurate, as shown in Figure (3).

References

[1] Majumdar A, Paul D, Bautista D. A generalized convolution model for multivariate nonstationary spatial

Doi: https://doi.org/10.54216/GJMSA.0100207
Received: September 22, 2023 Revised: December 12, 2023 Accepted: January 12, 2024

processes. Stat Sin 2010:675–95.
[2] Kleiber W, Nychka D, Bandyopadhyay S. A model for large multivariate spatial data sets. Stat Sin 2019;29:1085–104.
[3] Davies ER. Computer vision: principles, algorithms, applications, learning. Academic Press; 2017.
[4] Chen G, Guo Y, Zeng Q, Zhang Y. A Novel Cellular Network Traffic Prediction Algorithm Based on Graph Convolution Neural Networks and Long Short-Term Memory through Extraction of Spatial-Temporal Characteristics. Processes 2023;11:2257.
[5] Sun Y, Pang S, Zhang J, Zhang Y. Porosity prediction through well logging data: A combined approach of convolutional neural network and transformer model (CNN-transformer). Phys Fluids 2024;36.
[6] Micula G, Micula S. Handbook of splines. vol. 462. Springer Science & Business Media; 2012.
[7] Schumaker LL. Spline functions: computational methods. SIAM; 2015.
[8] Chui CK, Ron A. On the convolution of a box spline with a compactly supported distribution: linear independence for the integer translates. Can J Math 1991;43:19–33.

Doi: https://doi.org/10.54216/GJMSA.0100207
Received: September 22, 2023 Revised: December 12, 2023 Accepted: January 12, 2024