Business Systems Research | Vol. 15 No. 2 |2024

Beyond Parametric Bounds: Exploring Regional Unemployment Patterns Using Semiparametric Spatial Autoregression

Andrea Furková, Peter Knížat
University of Economics in Bratislava, Slovak Republic

Abstract

Background: It is a well-known phenomenon that nonlinearities that are inherent in the relationship among economic variables negatively affect the commonly used estimators in the econometric models. The nonlinearities cause an instability of the estimated parameters that, in particular, are unable to capture a local relationship between the response and the covariate. **Objectives:** The main aim of the paper is the simultaneous consideration of spatial effects as well as nonlinearities through an advanced semiparametric spatial autoregressive econometric model. The paper seeks to contribute to empirical studies of regional science focused on the application of semiparametric spatial autoregressive econometric models. **Methods/Approach:** We outline an approach that can be used to correct nonlinearities by incorporating a semiparametric idea within the framework of econometric models. We use an expansion by penalised basis splines that are highly flexible and are able to capture local nonlinearities between variables. **Results:** In the empirical study, we fit different econometric models that attempt to explain the dynamics of the European Union's regional unemployment. **Conclusions:** The results show that regional unemployment exhibits significant spatial dependence, indicating interconnectedness among neighbouring regions and suggesting the adoption of a semiparametric spatial autoregressive model for improved modelling flexibility, surpassing traditional parametric approaches.

Keywords: regional unemployment; linear regression; semiparametric model; generalised additive model; spline regression; spatial autoregressive semiparametric model
JEL classification: C14, C21, C52
Paper type: Research article

Received: Feb 2, 2024
Accepted: Jun 9. 2024

Acknowledgements: This work was supported by the Grant Agency of the Slovak Republic VEGA 1/0047/23.

Citation: Furková, A. Knížat, P. (2024). Beyond Parametric Bounds: Exploring Regional Unemployment Patterns Using Semiparametric Spatial Autoregression. *Business Systems Research, 15*(2), 48-66.
DOI: doi.org/10.2478/bsrj-2024-0017

Introduction

The use of semiparametric models in regional science is not as common as traditional parametric models, but it has gained attention and traction in recent years. Regional science involves the study of spatial patterns, relationships, and dynamics within specific geographic areas, making it a field where spatial econometrics plays a crucial role. The choice between a semiparametric model and a parametric spatial autoregressive (SAR) model (for more details, see, e.g., Anselin and Rey, 2014; Lung-Fei, 2022 or Chi and Zhu, 2019) depends on the specific characteristics of the data and the underlying assumptions of the modelling approach. Both approaches have their advantages and disadvantages, and the decision should be made based on the context and goals of the analysis. In regions with complex spatial patterns, where simple parametric forms do not easily capture the relationships between variables, semiparametric models can offer advantages. These models allow for more flexibility in accommodating spatial heterogeneity (see, e.g., Anselin and Rey, 2014). Semiparametric models (see Basile and Mínguez, 2018; Perperoglou et al., 2019), particularly those incorporating splines or other flexible functions, are useful when dealing with nonlinear relationships. Regional data may exhibit nonlinear patterns that linear parametric models cannot adequately capture. Semiparametric models are well-suited for capturing local variations within regions. Regional science often involves studying spatially heterogeneous phenomena, and semiparametric models can adapt to these variations more effectively than rigid parametric models. Semiparametric models may be more robust to misspecification and outliers, which can be important in the presence of spatial dependencies and complex regional dynamics. On the other hand, parametric models often have the advantage of being more interpretable, with coefficients corresponding to specific parameters. However, as the understanding and acceptance of semiparametric models grow, researchers are finding ways to interpret and communicate results from these models effectively. While semiparametric models offer advantages, researchers should carefully consider the trade-offs, including model complexity, interpretability, and computational demands.

One of the motivational factors of this paper is to contribute to filling the gap of empirical studies in regional science on the application of semiparametric spatial autoregressive econometric models. In this paper, we deal with the problem of unemployment in the regions of the European Union (EU). The novelty of the study can be seen in the simultaneous consideration of spatial effects (spatial autocorrelation and spatial heterogeneity) as well as nonlinearities in the functional form through an advanced semiparametric spatial autoregressive econometric model. This paper aims to investigate the dynamics of regional unemployment through the application of advanced econometric models, with a primary focus on the semiparametric spline spatial autoregressive model. The overarching goal is to enhance our understanding of the spatially-dependent nature of unemployment patterns across different regions. Our specific objectives – hypotheses of this research include:

- *Hypothesis 1 (Spatial Dependence Hypothesis): Regional unemployment rates exhibit significant spatial dependence, indicating that the unemployment rates influence the unemployment status in one region in neighbouring regions.*
- *Hypothesis 2 (Nonlinear Relationships Hypothesis): The relationship between regional unemployment and its determinants is nonlinear, and a semiparametric spline spatial autoregressive model can more effectively capture these nonlinearities than traditional parametric models.*

We commence our study by reviewing a classical linear regression model, where the model's parameters are estimated through the method of ordinary least squares.

The study of linear models is now part of any standard textbook on the introduction to multivariate statistical analysis, e.g., the manuscript in Fahrmeir et al. (2021). To extend the linear model to a more flexible framework, i.e., in the case when the relationship of the response and covariates exhibits some nonlinearities, we introduce basis spline functions and use these to transpose individual covariates into the functional form, which is then regressed onto the response. A basis spline functions are piece-wise polynomials joined in breakpoints, also called knots, which can easily be extrapolated onto a real-valued variable. A thorough theoretical discussion and application of basis spline functions are discussed by Perperoglou et al. (2019).

In the context of the nonlinear spline regression, the estimated parameters correspond to each polynomial, with a number of polynomials (spline curves) defined by the user or through a cross-validation procedure. On the other hand, a generalised additive model (GAM) uses a local smoothing algorithm for the estimation of the regression function and hence belongs to a family of fully nonparametric models. The advantage of GAM is that it further relaxes the assumption of the linearity that parameters in the model would normally restrict. GAM was first proposed by Hastie and Tibshirani (see Hastie and Tibshirani, 1986; 1990), and its most recent theory, including the application in R software, is provided by Wood (2017) and Wood (2023). In Mínguez et al. (2022) authors introduce a new R package for the estimation of flexible semiparametric spatial autoregressive models, which makes it possible to control for spatial dependence simultaneously, nonlinearities in the functional form, and spatiotemporal heterogeneity.

It is a well-known phenomenon that economic data that are observed in specific locations are affected by observations from neighbouring locations, which is called a spill-over effect. The manuscripts Lung-Fei (2022) provide a comprehensive review of spatial regression models that are used for spatial observations in the context of econometrics. We outline a theoretical framework of spatial autoregressive models that incorporate spatial spillover effects. Similarly, we follow by extending the spatial parametric model into the more flexible nonparametric approach, which should capture the nonlinearities between the response and covariates and hence improve the functional form of the estimated model. The paper by Basile et al. (2014) demonstrates the estimation technique for the spatial semiparametric model, which is carried out by using a 2-step "control function" approach since the two-stage least squares method might lead to inconsistent estimates of the regression parameters. In Basile and Mínguez (2018), a critical review of parametric and semiparametric spatial econometric approaches can be found. The author focuses on the capability of each class of models to fit the main features of spatial data (such as strong and weak cross-sectional dependence, spatial heterogeneity, nonlinearities, and time persistence).

As we have already stated, the application of semiparametric models in regional science is not as common as traditional parametric models. We can find the use of this approach in the works of Wahyuni and Fajri (2020) or Mínguez et al. (2022). However, only a few empirical works apply the semiparametric spatial autoregressive econometric approach in connection with the modelling of regional economic problems. From this point of view, we believe that this paper might contribute to supplement empirical analyses of this nature.

The rest of the paper was structured as follows: the methodology section provides the main theoretical background, and the results section presents an overview of a study area, a description of the data, model specification, and main empirical results. The main concluding remarks are presented in Discussion and Conclusion sections. The paper closes with References.

Methodology

Let us assume that we observe a matrix **X** = (x_{ik}), where $i = 1, 2, \ldots, N$ refers to the sample unit that is observed for each covariate $k = 1, 2, \ldots, K$. Let $\mathbf{x}_1, \ldots, \mathbf{x}_N$ be the fixed vectors in $\mathbb{R}^K$ and let $y_1, \ldots, y_N$ be dependent variables. In general, the functional relationship between the response vector y_i and the covariate $\mathbf{x}_i$ can be expressed as:

$$y_i = f(\mathbf{x}_i) + \varepsilon_i, \qquad i = 1, 2, \ldots, N \tag{1}$$

where $\varepsilon_1, \ldots, \varepsilon_N$ are independent and identically distributed (*i.i.d.*) random errors with mean zero and variance σ_ε^2, i.e., $\varepsilon_i \sim i.i.d.\left(0, \sigma_\varepsilon^2\right)$. The function $f(.)$ can be of the parametric or nonparametric form.

In the following subsections, we introduce parametric regression models that can be described by a finite number of estimated parameters. The estimated parameters determine the model's functional form. Subsequently, we outline nonparametric regression models that do not require a predetermined functional form but are constructed according to information derived from data.

Linear regression model and its extension to nonlinear spline regression

A linear Ordinary Least Squares (OLS) regression model can be expressed as (Fahrmeir et al., 2021):

$$y_i = \alpha + \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i, \qquad \varepsilon_i \sim i.i.d.\left(0, \sigma_\varepsilon^2\right) \tag{2}$$

where $\mathbf{x}_i$ represents a $1 \times K$ vector of covariates with associated parameters $\boldsymbol{\beta}$ contained in a $K \times 1$ vector and α is the intercept. The OLS method is used to estimate the parameters, which yields the following (ibid):

$$\boldsymbol{\beta} = \left(\mathbf{X}^{\mathrm{T}} \mathbf{X}\right)^{-1} \mathbf{X}^{\mathrm{T}} \mathbf{y} \tag{3}$$

where $\boldsymbol{\beta} = \left(\alpha, \beta_1, \beta_2, \ldots \beta_p\right)$, noting that the matrix **X** includes ones in its first column for the estimation of α and **y** denotes a vector of dependent variable. The matrix $\left(\mathbf{X}^{\mathrm{T}} \mathbf{X}\right)$ is of full rank in order to be invertible.

In many regression scenarios, the relationship between the response and covariates exhibit local nonlinearities, which implies that the parametric model can be too restrictive. In this case, the functional form of the Eq. (1) is mis-specified and its estimated values y_i lie far off the observed values y_i.

To capture the local nonlinearities between the response and covariates, a nonlinear regression model with basis spline functions, which are extrapolated onto each covariate, and hence replace matrix **X**, could be more suitable. The covariate matrix can be expressed in terms of basis spline expansion as follows (Perperoglou et al., 2019):

$$\mathbf{X} = \mathbf{B}_\delta^m\left(\tau_l\right) \mathbf{C}_\delta \tag{4}$$

where the spline curves $\mathbf{B}_{\delta}^{m}\left(\tau_{l}\right)$ are piecewise polynomials of order *m* that are merged at the break points, also called knots, τ_l, $l=1,2,\ldots,L-1$, where $\delta=1,2,\ldots,\Delta$, refers to the number of spline curves. $\mathbf{C}_{\delta}$ is a $\Delta\times K$ matrix of parameters that needs to be estimated. The reader interested in more theoretical details of basis spline functions can consult Perperoglou et al. (2019).

Within the context of the nonlinear spline regression model, the estimation of parameters in the matrix $\mathbf{C}_{\delta}$ is carried out through an OLS method that minimises the sum of squares errors, which yields the following (omitting τ_l and superscript *m* for simplification):

$$\mathbf{C}_{\delta}=\left(\mathbf{B}_{\delta}^{\mathrm{T}}\mathbf{B}_{\delta}\right)^{-1}\mathbf{B}_{\delta}^{\mathrm{T}}\mathbf{y} \tag{5}$$

The interpretation of estimated parameters $\mathbf{C}_{\delta}$ is more elaborative than in the classical linear regression since each $\mathbf{C}_{\delta}$ is linked to intervals of responses y_i and the covariate $\mathbf{x}_i$. Therefore, the model is able to capture the local nonlinearities between the response and covariates specific to these intervals.

Generalised additive regression model

The generalised additive model (GAM) is considered a nonparametric version of the nonlinear model, where the linear form $\boldsymbol{\beta}\mathbf{X}$ is replaced by a sum of unspecified functions $g\left(\mathbf{x}_i\right)$ that are estimated through a method of the local backfitting (smoothing) algorithm, first proposed by Hastie and Tibshirani (1986). The user can define various forms of smooth functions in *g*(.). We opt to use basis spline functions that have various advantages, refer to Perperoglou et al. (2019) for details.

GAM can be expressed as (Wood, 2017):

$$y_i=\alpha+\sum_{p=1}^{P}g\left(\mathbf{x}_i\right)+\varepsilon_i,\qquad \varepsilon_i\sim i.i.d.\left(0,\sigma_{\varepsilon}^{2}\right) \tag{6}$$

The estimation method of the GAM is based on the minimisation of the cross-validation sum of squares (*CVSS*):

$$CVSS\left(k\right)=\frac{1}{N}\sum_{i=1}^{N}\left(y_i-\alpha-\sum_{p=1}^{P}g^{-i}\left(\mathbf{x}_i\right)\right)^{2} \tag{7}$$

where $g^{-i}\left(\mathbf{x}_i\right)$ is the basis spline function with k number of basis, having removed one observation $\left(x_i,y_i\right)$ from the sample at each iteration. The estimation procedure of minimising *CVSS*(*k*) is a repetitive smoothing of the dependent variable y_i on $\mathbf{X}_i$, which is carried out through *a local backfitting* algorithm. The iterative procedure is described in details in Wood (2017), with an application in R software.

Spatial autoregressive regression model and its extension to nonlinear spline regression

In the socioeconomic problem, we usually observe data from regional economic activities that are known to be regionally correlated, i.e., an observation from the location (region) *i* is affected by observations from other locations *j*, where *i* ≠ *j*, also called spatial spill-over effects. In general, a formal expression of the spatial correlation between different locations can be defined in terms of corresponding non-zero covariances (Anselin and Rey, 2014):

$$\mathrm{cov}\left[y_i y_j\right] = \mathrm{E}\left[y_i y_j\right] - \mathrm{E}\left[y_i\right]\mathrm{E}\left[y_j\right] \neq 0 \qquad i \neq j \tag{8}$$

where E refers to the expected value and y_i and y_j are observed values from regions *i* and *j*, respectively. So, any influences that spread from one location to nearby ones (spatial spill-over effects) should be taken into account when building the regression model. Traditional spatial econometric estimation framework is based on models with spatially autoregressive process, the models that explicitly allow for spatial dependence through spatially lagged variables. The type of spatial model can be determined using *LM* tests (see, e.g., Anselin and Rey, 2014). One of the well–known model from this class is SAR (Spatial Autoregressive) model, which assumes spatial spillover effects within the dependent variable *y*. We present this model in relation to our empirical analysis. The SAR model is formulated as follows (Anselin and Rey, 2014):

$$y_i = \rho \mathbf{W}\mathbf{y} + \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i, \qquad \varepsilon_i \sim i.i.d.\left(0, \sigma_\varepsilon^2\right) \tag{9}$$

where ρ is a spatial autoregressive parameter, **Wy** denotes a spatially lagged dependent variable and **W** is a $N \times N$ spatial weighting matrix. In this paper, queen contiguity spatial weighting matrix was used in all spatial econometric models and spatial statistics calculations. Due to possible problems with isolated units as well as with high variability of neighbouring regions resulting from other approaches, the queen contiguity form seemed to be suitable for determining spatial regional structures. In the case of spatial weights, for instance, based on a distance function, inverse or radial (see, e.g., Pavlovčič-Prešeren et al., 2019), there can be a problem with the bimodal distribution when some regions have very few neighbours. On the other hand, the other regions have very many neighbouring units. In the scientific and empirical literature, there are many other traditional definitions for the spatial structure among spatial locations (see, e.g., Lung-Fei, 2022 or Chi and Zhu, 2019).

Estimation of models with spatial autocorrelation and/or spatial heterogeneity requires special estimation methods and procedures. For instance, the estimation of spatial autoregressive models (e.g., SAR model) is affected by the presence of the spatially lagged variable **Wy** on the right-hand side of the regression equation, which causes problems with endogeneity. Therefore, OLS is not a suitable estimation method. The estimation of such models is based on familiar econometric estimation methods, but they must be modified with respect to spatial aspects: Maximum Likelihood (ML), Two-Stage Least Squares (2SLS) or Generalised Moment Method (GMM). A review of these estimation methods can be found in Anselin and Rey (2014) or Chi and Zhu (2019).

In general, spatial autoregressive models are sometimes unfeasible in the presence of model misspecification. Geniaux and Martinetti (2018) pointed out that it can often be problematic to disentangle between a real spatial autocorrelation and different sources of violation of *i.i.d.*, such as spatial heterogeneity through unobserved covariates and spatially varying relationships. Modelling spatial data requires flexible econometric tools that allow us to control spatial dependence, spatial heterogeneity, non-linearities and other possible model specification biases. To address this demand for flexibility, the adoption of the nonparametric structure or semiparametric structure of the spatial regression model is advisable. Similarly, using basis spline functions for covariates, the spatial autoregressive semiparametric model can be defined as (Basile et al., 2014):

$$y_i = \rho \mathbf{W}\mathbf{y} + \alpha + \mathbf{x}_i \boldsymbol{\beta} + g\left(\mathbf{x}_i\right) + \varepsilon_i, \qquad \varepsilon_i \sim i.i.d.\left(0, \sigma_\varepsilon^2\right) \tag{10}$$

where spline basis expansions of original covariates are defined in Eq. (4). Some covariates could enter Eq. (10) in the parametric form, which can be determined through preliminary statistical analysis.

The estimation of Eq. (10) can be carried out by using either a restricted maximum likelihood (REML) or a 2-step "control function" approach, refer to Basile et al. (2014) for theoretical details. The REML approach combines penalised regression spline (PS) methods (see, e.g., Perperoglou et al., 2019) with standard spatial autoregressive models such as SAR defined in Eq. (9), Spatial Error Model (SEM) or Spatial Durbin Model (SDM). An important advantage of such models is that they make it possible to capture local nonlinearities within the specification of spatial autoregressive terms, i.e., to capture spatial interaction effects and parametric and nonparametric relationships. In addition, a geoadditive term, i.e., a smooth function of the spatial coordinates can be included in Eq. (10) to capture a spatial trend effect (to capture spatially autocorrelated unobserved heterogeneity).

Results

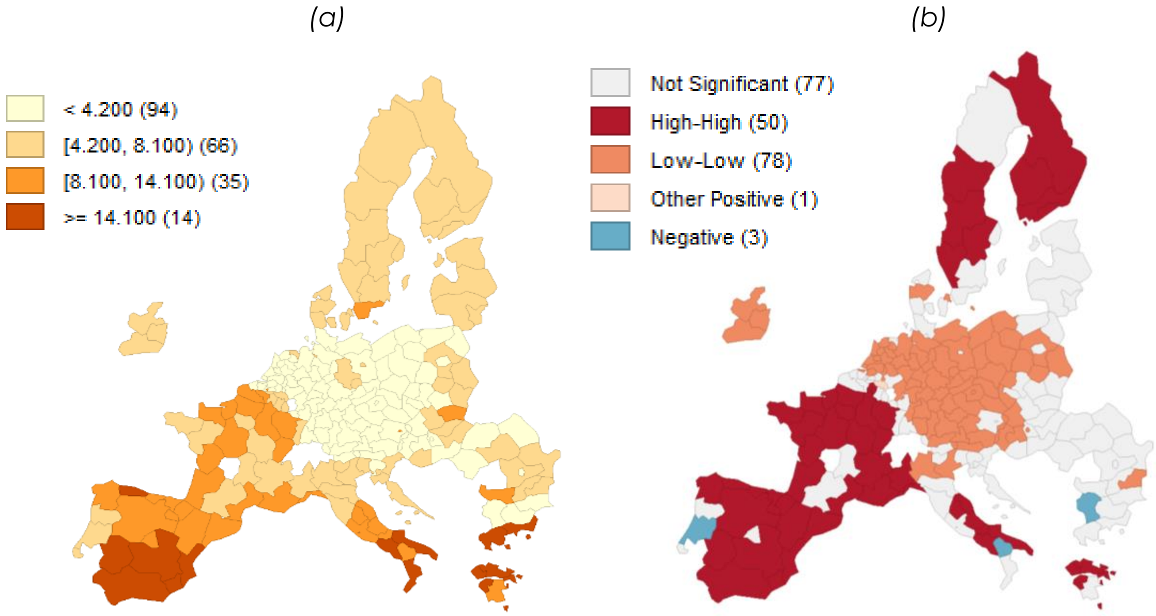
In this section, we apply the theoretical framework outlined in the previous section. We perceive the semiparametric SAR model defined by Eq. 10 to be highly useful for modelling cross-sectional spatial data considering nonlinearities, spatial dependence, and spatial heterogeneity. We empirically illustrate this model's performance in modelling the European unemployment problem.

Regional Unemployment Data

The paper uses data from the Eurostat regional statistical database (Eurostat, 2023). After excluding isolated observations (island regions) and missing data, the corrected database contains 209 European regions at the NUTS 2 level (NUTS—Nomenclature of territorial units for statistics). Figure 1 provides an overview of the study area. This figure shows a real spatial distribution (a) and local Geary cluster map (b) of Unemployment rates in 2019 across the EU regions.

The maps presented in Figure 1 already indicate disparities among the EU regions. In addition to regional disparities, we can also notice that regions are considerably clustered. The existence of strong positive spatial autocorrelation indicates the statistically significant value of global Moran's I statistic (0.683 with pseudo-p-value 0.001). The local Geary cluster map (see e.g., Chi and Zhu, 2019) provides more evidence about indicated unequal distribution and spatial clustering of the EU unemployment. Based on Figure 1 (b), we identify statistically significant locations – regions with positive spatial autocorrelation so-called hot spots and cold spots locations (50 high-high and 78 low-low locations). The high-high locations are mainly the regions of Spain and France. These regions are regions where high values of unemployment rates are clustered. Low-low values are mainly concentrated in the regions of Germany, Austria, the Netherlands, and some regions of Eastern Europe, such as the Czech Republic and Poland. This suggests that the geographic position of the region and the spatial regional spillovers probably affect the level of regional unemployment.

Figure 1
Spatial distribution of Unemployment rates in 2019 – natural breaks map (a) and Local Geary cluster map (b)

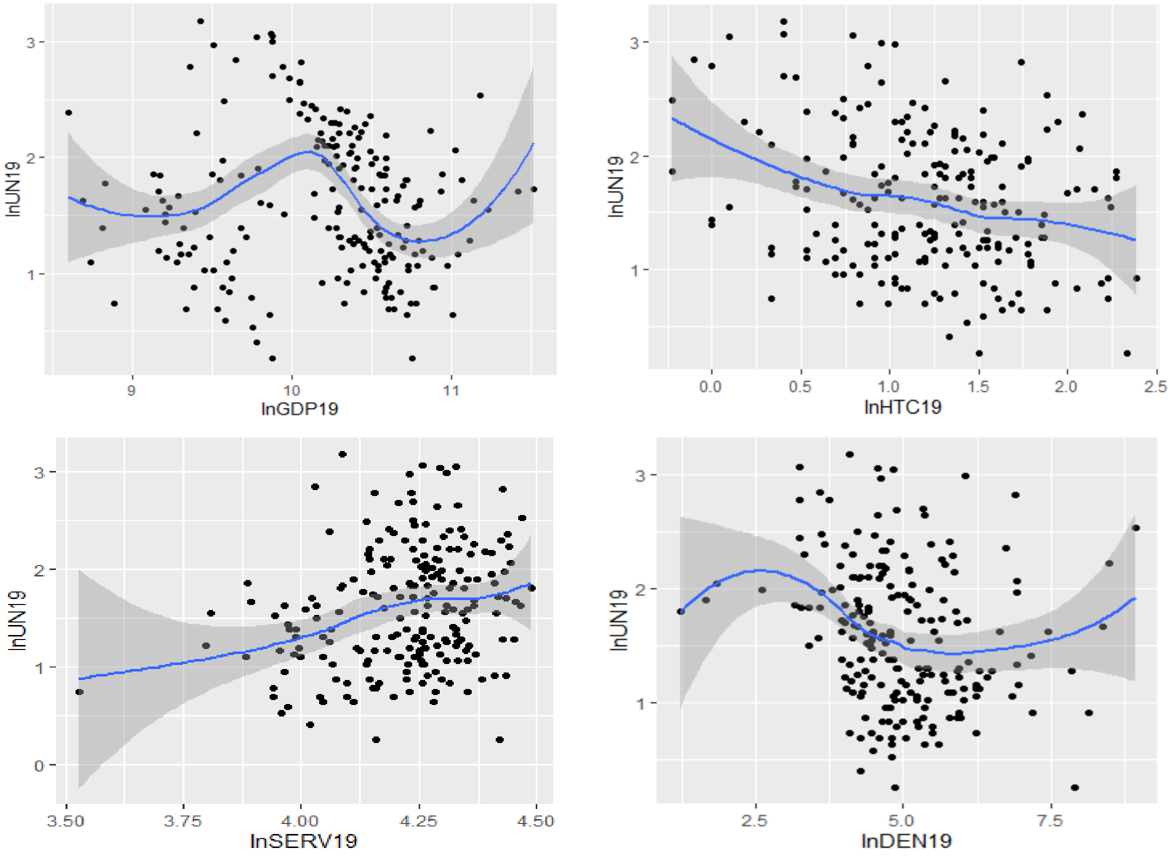


Source: Authors' work.
Note: The numbers in brackets indicate the number of regions in the category. Compared to the quartile map, the natural breaks criterion is better at grouping extreme observations. Interestingly, unlike quantile maps, the number of observations in each category can be highly unequal.

The initial empirical analysis will be based on a linear regression model that draws from the "regional competitiveness" theory (Formánek, 2019) explaining unemployment dynamics in terms of its key determining factors *GDP* – gross domestic product (Euro per inhabitant) and two convenient labour-force structure and competitiveness indicators: *HTC - employment in technology and knowledge-intensive sectors – high-technology sectors* (percentage of total employment), *SERV - employment in technology and knowledge-intensive sectors - services* (percentage of total employment). In addition, we also consider variable *DEN - population density* (persons per square kilometre) as a possible determinant of unemployment. All variables have logarithmically transformed forms and the observed period is 2019. Due to the skewed distribution of the dependent variable – Unemployment rates, we use log-transformation.

In the next step, we briefly examine the relationship between the response and each covariate. Figure 2 displays *one-to-one* relationships, with a fitted line using the *scatterplot* smoother and its corresponding 95% confidence intervals. It was created using the ggplot function in R.

Figure 2
Scatterplots of Unemployment vs four key determinant factors (GDP, HTC, SERV, DEN)



Source: Authors' work.

Figure 2 clearly shows that the *one-to-one* relationship between unemployment and all other key determinant factors is nonlinear. Short preliminary analyses indicate that an OLS regression might be far from sufficient to investigate the determinants of EU regional unemployment. First of all, we have seen, that spatial preliminary analysis reveals the problem of spatial autocorrelation and heterogeneity. On the other hand, the scatterplots of unemployment versus four key determinant factors, i.e. one-to-one relationships analysis, point to the problem of nonlinearity. It follows that the nonparametric regression could be a more flexible modelling of the effects of continuous covariates on the dependent variable since the classical linear model might not sufficiently capture local nonlinearities.

Econometric Models

The empirical part of the paper consists of the estimation of five econometric models to determine the factors affecting regional unemployment and to compare the performance of different specifications of the econometric models:

- **Model1**
  - non-spatial parametric (linear) model - OLS regression:

$$y_i = \alpha + \sum_{k=1}^{K} \beta_k x_{k,i} + \varepsilon_i, \quad i = 1,2,\ldots,N \qquad \varepsilon_i \;\Box\; i.i.d.\left(0,\sigma_\varepsilon^2\right) \tag{11}$$

- **Model2**
  - non-spatial nonlinear (non-parametric) model – spline regression:

$$y_i = \alpha + \sum_{\delta=1}^{\Delta} g_\delta\left(x_{\delta,i}\right) + \varepsilon_i, \quad i = 1,2,\dots,N \qquad \varepsilon_i \,\square\, i.i.d.\left(0,\sigma_\varepsilon^2\right) \tag{12}$$

- **Model3**
  - SAR parametric (linear) model:

$$y_i = \rho \sum_{j=1}^{N} w_{ij} y_j + \alpha + \sum_{k=1}^{K} \beta_k x_{k,i} + \varepsilon_i, \quad i = 1,2,\dots,N \qquad \varepsilon_i \,\square\, i.i.d.\left(0,\sigma_\varepsilon^2\right) \tag{13}$$

- **Model4**
  - semiparametric (nonlinear) SAR model without spatial trend:

$$y_i = \rho \sum_{j=1}^{N} w_{ij} y_j + \sum_{\delta=1}^{\Delta} g_\delta\left(x_{\delta,i}\right) + \varepsilon_i, \quad i = 1,2,\dots,N \qquad \varepsilon_i \,\square\, i.i.d.\left(0,\sigma_\varepsilon^2\right) \tag{14}$$

- **Model5**
  - semiparametric (nonlinear) SAR model with spatial trend:

$$y_i = \rho \sum_{j=1}^{N} w_{ij} y_j + \sum_{\delta=1}^{\Delta} g_\delta\left(x_{\delta,i}\right) + \tilde{f}\left(s_{1i}, s_{2i}\right) + \varepsilon_i, \ i = 1,2,\dots,N \ \varepsilon_i \,\square\, i.i.d.\left(0,\sigma_\varepsilon^2\right) \tag{15}$$

where y_i denotes the response variable and $x_{k,i}$ denotes the individual predictors, all defined in the section Regional Unemployment Data. Eq. 15 represents the spatial trend and denotes the spatial coordinates of ith region. We have already defined the other remaining terms in the section Methodology.

The spatial regression models defined in (13), (14) and (15) contain a spatial lag of the dependent variable. This means that the expected value of unemployment in the *i*th region is no longer influenced only by exogenous regional characteristics but also by the exogenous characteristics of all other regions through a spatial multiplier (for more details, see, e.g., Chi and Zhu, 2019)). The specifications of all spatial econometric models are based on the queen contiguity weights (matrix **W**) – these binary weights indicate whether regions share a boundary or not. As the last model, we introduce a semiparametric spatial model with a spatial trend (see Eq. 15) in order to control for unobserved spatial heterogeneity.

We estimated all models defined by (11) – (15) equations in the R package pspatreg (Mínguez et al., 2022). The non-parametric terms (either trends or covariates) were modeled using P-Splines. The estimation methods were maximum likelihood (ML) and restricted maximum likelihood (REML).

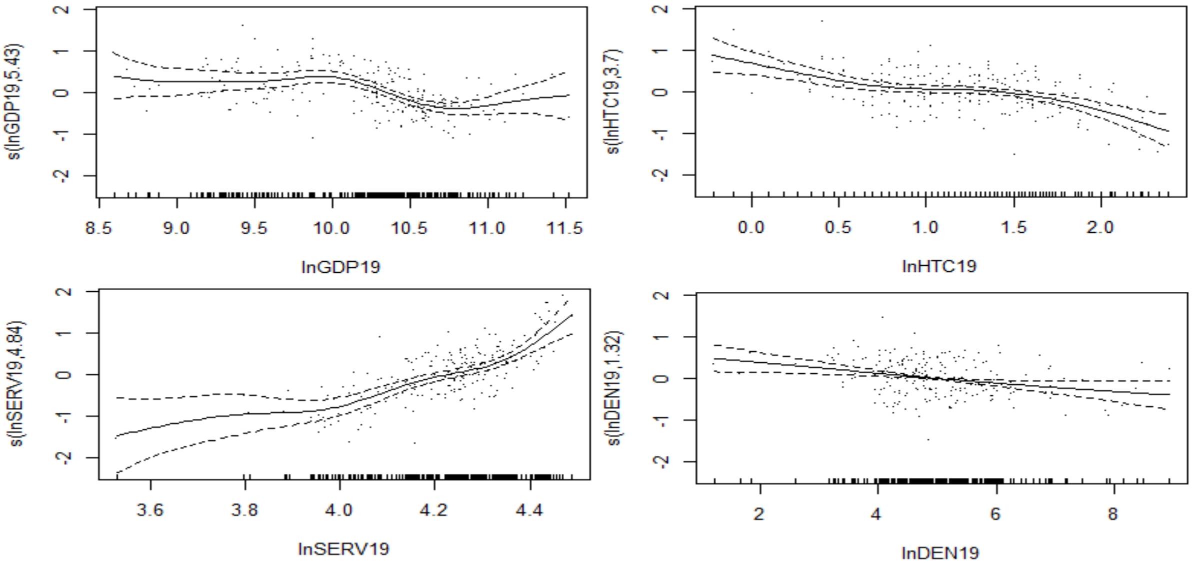
The performed analysis of one-to-one relationships pointed to the problem of nonlinearity. Nonparametric functions *g*(.) in models (12), (14) and (15) are formed by transposing a real-valued covariate by spline basis expansion. The spline functions have high flexibility and can handle data that changes in subintervals, which relates to local nonlinearities. In this context, it was crucial to determine which variables belong to the parametric or non-parametric component, to select the optimal knots and their location. We relied on the starting GAM model defined in Eq. 6, and we utilised adaptive knot selection methods used in the mgcv package in R (Wood, 2023), which automatically select knot locations based on data characteristics. The approach used for knot selection involves automatic smoothness selection using penalized likelihood methods. The resulting number of knots was 9 and was used in all non-parametric and semi-parametric models.

Empirical Results

Since the estimations of models (11) - (15) provide extensive estimation outputs, it is not possible to list them within the scope of this article. In this section, we present the most important outputs that allow us to evaluate the hypothesis stated in the Introduction. Other outputs are available at the request of the authors.Figure 3 shows plots of non-parametric covariates resulting from GAM model estimation.

Figure 3

Plots of terms of non-parametric covariates - GAM model



Note: Pointwise confidence intervals in dashed lines.
Source: Authors' work.

Figure 3 shows that both the left and right tail of confidence intervals, for all variables, are very wide, which could indicate a potential disturbance by extreme values in the estimation of classic econometric models, i.e. the assumption of normality of estimated residuals can be violated at its tails. Hence, it further emphasises that semiparametric models with spline functions could be more appropriate. From a statistical point of view, regarding the issue of identifying statistically significant determinants of EU regional unemployment, we can conclude that the results show a statistically significant influence of all selected factors, and the parameter estimates have the expected signs.

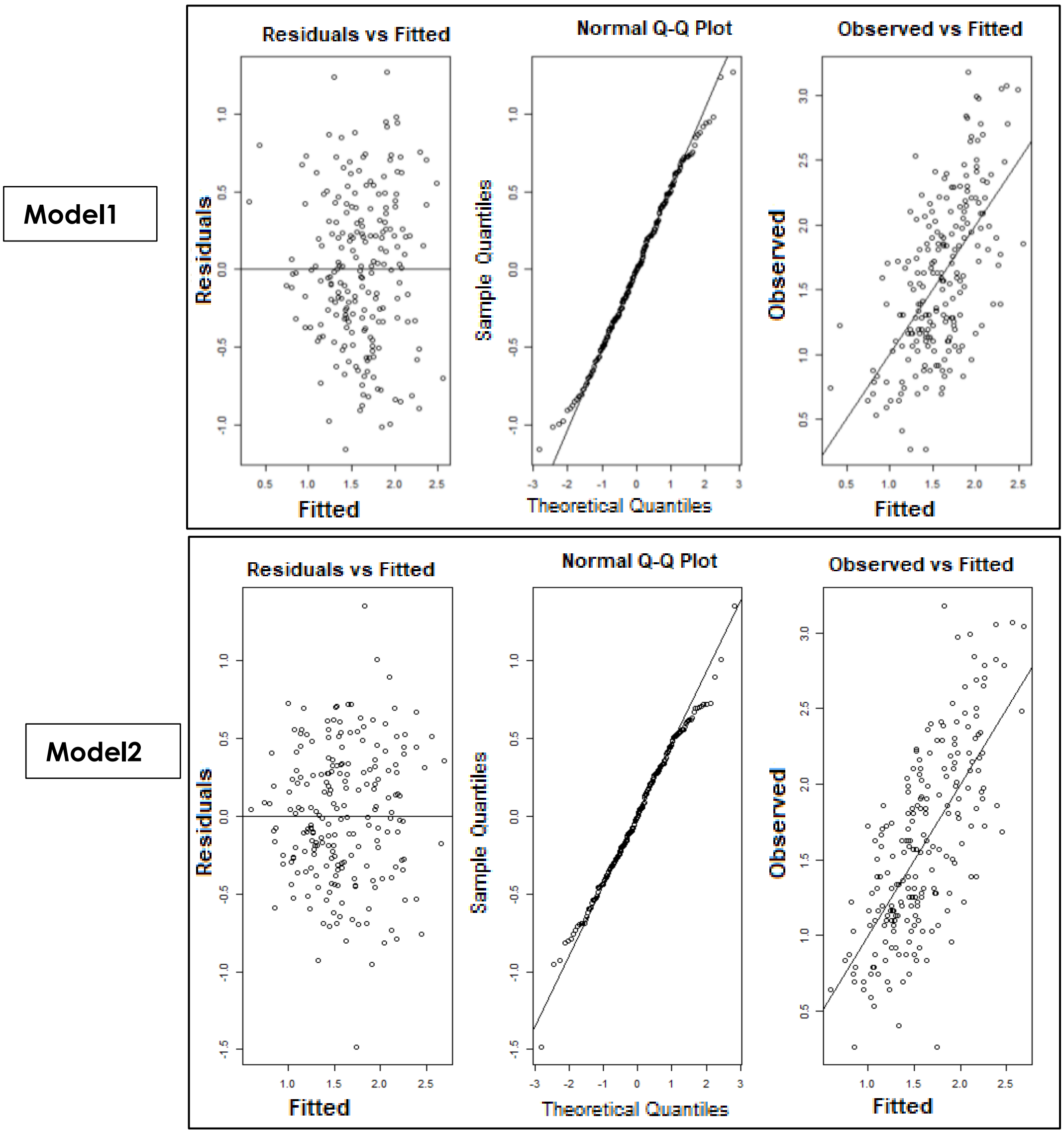
Also, the consideration of the spatial aspect in spatial SAR models suggests strong spatial spillover effects between regions. The statistical significance of the spatial autoregressive parameter and its high positive value (approx. 0.67) in both spatial models contribute to the confirmation of the hypothesis of spatial regional connectivity. In this context, it was necessary to calculate and verify the statistical significance of the average direct, average indirect and average total effect of all explanatory variables due to the correct interpretation of the model parameters. In addition, the assumption of non-linearity in SAR models (14) and (15), i.e., the situation that we consider a non-parametric smooth function for all covariates (except spatial lag variable), caused these effects to have a non-parametric character.

In the following figures, we show the specific results of individual econometric models, Model1 – Model5, which allows us to evaluate the stated hypothesis of the paper. Particularly, we look at whether the assumptions of each model are met, which determines the stability of the estimated parameters and their corresponding statistical tests of significance. Figures 4 – 6 display three plots: i) the residuals versus fitted values, ii) the Q-Q plot of the residuals, and iii) the observed values versus fitted values. If the first plot shows no pattern, it implies that residuals are independent and identically distributed; any other pattern could indicate a correlation in residuals or an unstable variance of residuals. The second plot assesses if the residuals come from a normal distribution, which is met if all values are close to the diagonal line. By the third plot, we are able to examine the predictive power of the model, i.e., the closer the spread of the observed versus fitted values to the diagonal line, the better the model's fit to the observed data (noting a risk of overfitting if the values are too close to the diagonal line).

Figure 4 shows a comparison between the estimate of ordinary least squares and its amended version, where we use spline functions in the estimator. Both models show distorted residuals, i.e., the first plot shows a fan-shaped pattern that indicates an unstable, nonconstant variance of the residuals. The second diagram shows that the assumption of normality is violated by extreme values, which is made clear by the deviation of the values at both ends. The third diagram shows that a model with spline functions (Model 2) performs slightly better in terms of predictive power.

Figure 4

Residuals vs Fitted, Normal Q-Q plot and Observed vs Fitted – Model1 and Model2

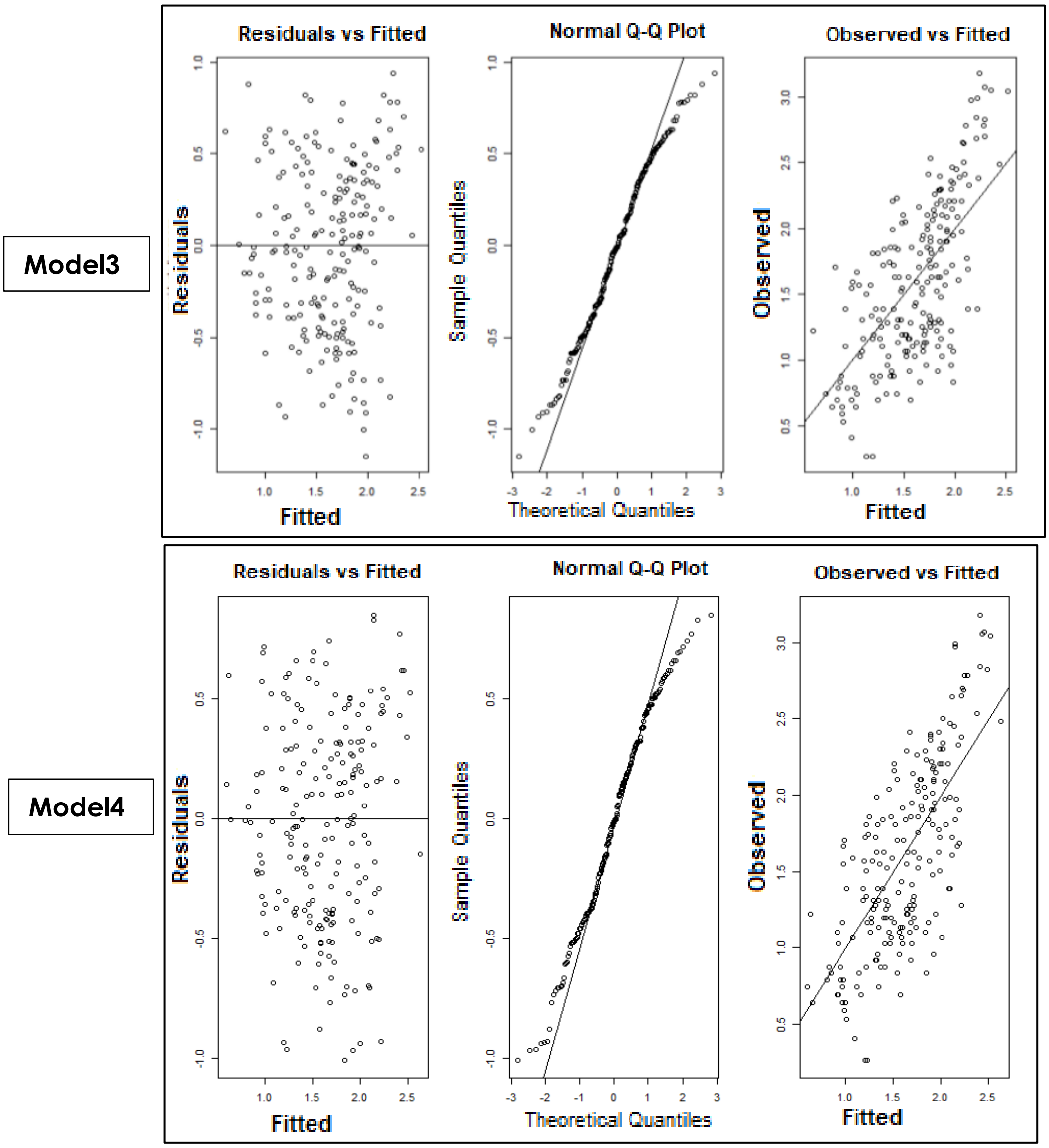


Source: Authors' work.

Figure 5 compares the regression models with spatial effects that are expected to improve the model's stability since they correct the estimator for the spatial correlation across observations. Similarly, the first model is a classic spatial autoregressive (SAR) model, with the second one being its counterpart with spline functions in the estimator. Contrary to expectations, neither of the models with spatial effects (Model 3 and Model 4) significantly improves the estimation results in terms of the normality of the residuals and the predictive power of the model. The normality of residuals is still violated, as can be seen in the second plot, and the predicted power of the models, shown in the third plot (see Figure 5), is similar to regressions with no spatial effects. However, a spatial regression attains an improvement in the stability of the estimated models, i.e., the first plot of residuals versus fitted values shows a somehow random pattern, in both the classic and spline SAR models.

Figure 5

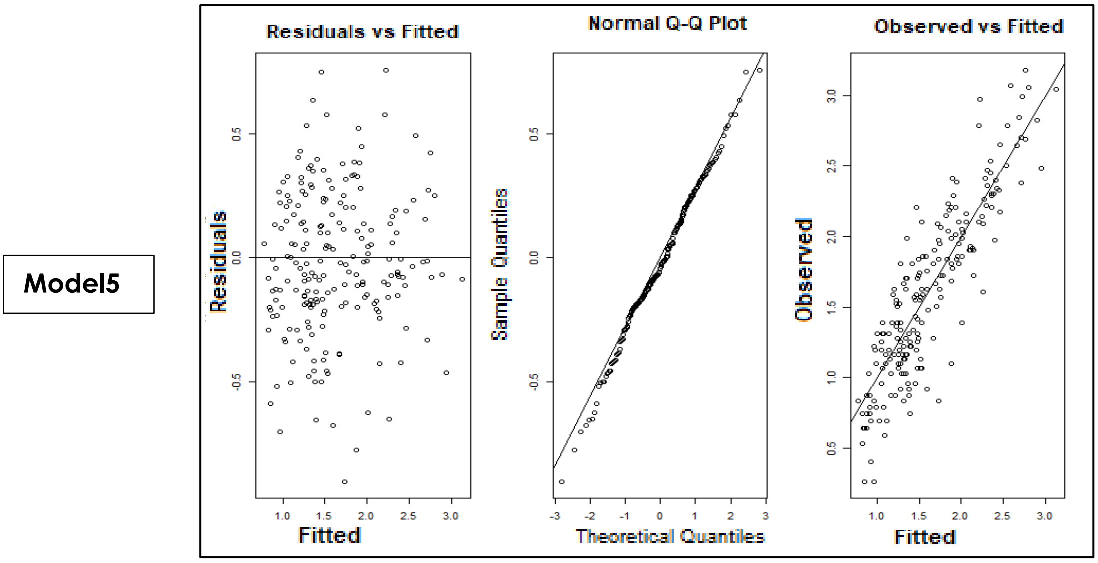
Residuals vs Fitted, Normal Q-Q plot and Observed vs Fitted – Model3 and Model4



Source: Authors' work.

Figure 6 displays the same matrix of the evaluation measures for the spatial regression model with spline functions, which also include a spatial trend in the estimator. We intentionally show the performance of this model on a standalone basis since it attains a significant improvement in all of the performance measures. Including the spatial trend in the model's estimator corrected the deviation from the normality at the tails of the Q-Q plot, i.e., the extreme values do not cause any disturbance, and we can safely keep them in the dataset (some analyses exclude extreme values). The other two plots also show a more accurate behaviour of the model when compared to its previous spatial counterparts.

Figure 6
Residuals vs Fitted, Normal Q-Q plot and Observed vs Fitted – Model5



Source: Authors' work.

Moreover, we further examine the analysis of variance between models by looking at different conventional statistics which are calculated as part of the models' estimation. This allows us to pick the best estimator.

Table1
Linear vs Nonlinear with splines (Model1 vs Model2)

	logLik(1)	rlogLik(2)	EDF(3)	AIC(4)
Linear	45.51	32.32	5.00	-81.02
Nonlinear with splines	49.83	45.90	15.46	68.74

Note: (1) Note: Log-Likelihood; (2) restricted Log-Likelihood; (3) Effective degrees of freedom; (4) Akaike information criterion
Source: Authors' work.

Table 2
SAR vs SAR with splines (Model3 vs Model4)

	logLik(1)	rlogLik(2)	EDF(3)	AIC(4)
SAR	109.31	94.254	6.000	-206.62
SAR with splines	111.14	104.366	13.118	-196.05

Note: (1) Note: Log-Likelihood; (2) restricted Log-Likelihood; (3) Effective degrees of freedom; (4) Akaike information criterion
Source: Authors' work.

Table 3
Nonlinear with spline vs SAR with splines (Model2 vs Model4)

	logLik(1)	rlogLik(2)	EDF(3)	AIC(4)
Nonlinear with splines	49.83	45.90	15.46	68.74
SAR with splines	111.14	104.366	13.118	-196.05

Note: (1) Note: Log-Likelihood; (2) restricted Log-Likelihood; (3) Effective degrees of freedom; (4) Akaike information criterion
Source: Authors' work.

Table 4
SAR with splines vs SAR with splines and spatial trend (Model4 vs Model5)

	logLik(1)	rlogLik(2)	EDF(3)	AIC(4)
SAR with splines	111.14	104.366	13.118	-196.05
SAR with splines and spatial trend	122.48	118.49	27.040	-190.88

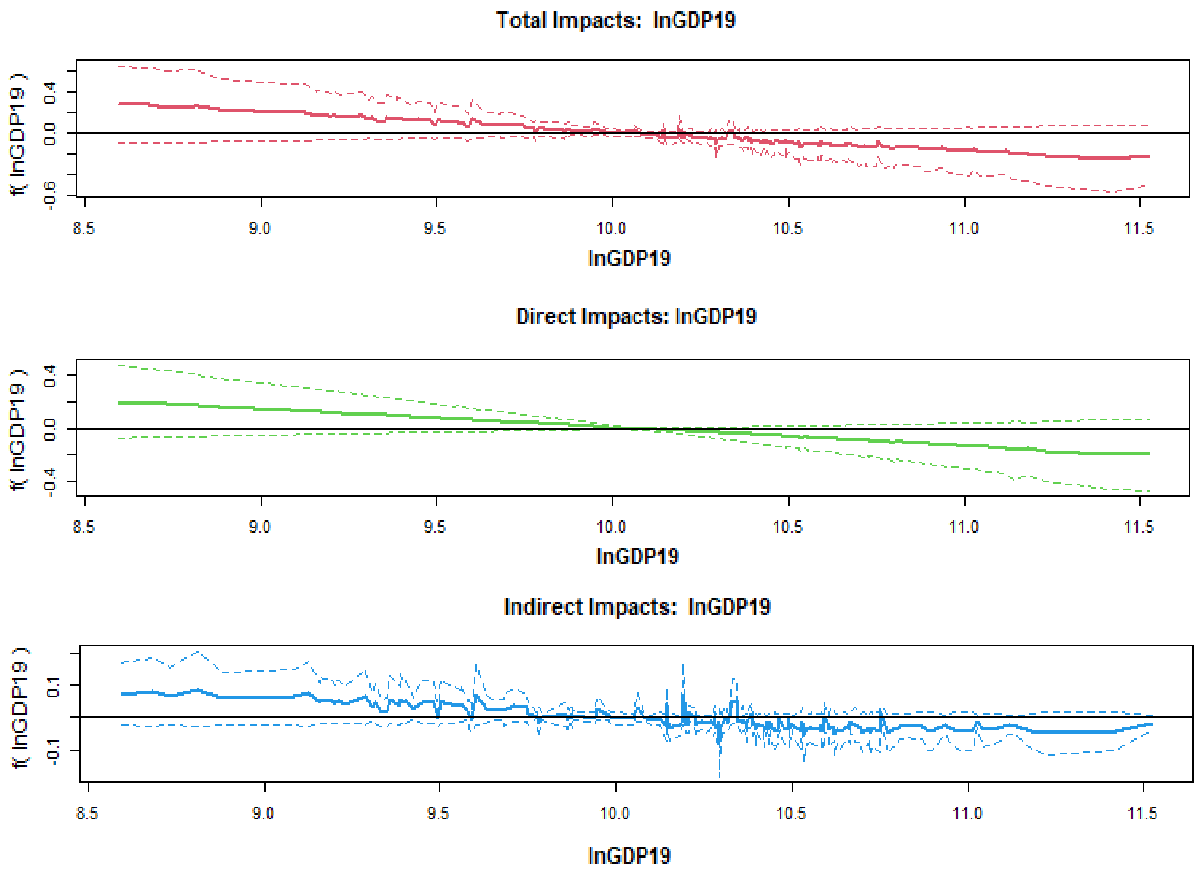
Note: (1) Note: Log-Likelihood; (2) restricted Log-Likelihood; (3) Effective degrees of freedom; (4) Akaike information criterion
Source: Authors' work.

From tables 1-4, we can re-state the conclusion that the SAR model with spline functions, which also includes the spatial trend, has the highest log-likelihood, restricted log-likelihood, and EDF numbers. The EDF measurement shows that all models are highly nonlinear. Regarding the information criteria, the AIC values are similar for the spatial models, and these values speak in their favour.

The semiparametric SAR model with a spatial trend (Model5) appears to be extremely useful for modelling spatial data with respect to nonlinearities, spatial dependence, and spatial unobserved heterogeneity when this heterogeneity is smoothly distributed over space. Figure 7 shows selected estimation results based on Model 5 as a brief preview of the results. Based on the estimation results of Model 5, we were able to calculate total, direct and indirect (or spillover) effects for all smooth (non-parametric) terms. Graphs of non-parametric covariate terms for the SAR model with spline and spatial trend (Model5) are presented in Figure 7, the results are presented only for the GDP variable.

Interpreting non-parametric impacts from a spatial semiparametric autoregressive model involves understanding the direct, indirect, and total effects of predictor variables on the response variable. The interpretation of these results compared to the results provided by the parametric SAR model is more complicated but probably provides very useful insights into the influence of predictors on the response variable. Figure 7 provides interesting information, e.g., regarding indirect effects. Indirect effects capture the impact of a predictor variable on the response variable through spatial dependencies, considering interactions with neighbouring regions. Positive indirect effects suggest that an increase in the predictor variable not only affects the response variable in the same region but also spills over to positively influence neighbouring regions. Conversely, negative indirect effects imply a negative spillover effect. In the case of the GDP variable for its different levels, we see that these impacts are different, and we notice that higher GDP values correspond to negative spillover effects. This means that a higher level of GDP in neighbouring regions contributes to reducing the level of unemployment in a particular region.

Figure 7
Plots of non-parametric direct, indirect and total impacts – GDP variable - Model5


Total Impacts: lnGDP19

Direct Impacts: lnGDP19

Indirect Impacts: lnGDP19

Note: Pointwise confidence intervals in dashed lines.
Source: Authors' work.

In general, comparing direct, indirect, and total effects across different predictor variables helps prioritise their importance in influencing the response variable. Understanding the spatial dynamics can be very helpful for potential policy implications. Above all, regional policies and interventions can be targeted much more precisely.

Discussion

The empirical findings demonstrate that it is of the utmost importance to choose an appropriate theoretical framework for the econometric model, including its corresponding estimator. An incorrect model leads to weak estimated parameters, which are important when interpreted in the context of the economic impact. The model may also suffer from poor predictive power. The complex econometric models have a difficult structural form and might require more elaboration in their interpretation. However, as shown by the empirical analysis, they can lead to stable estimated parameters and improvements in the predictive power, which is crucial when using the economic interpretation of the estimated parameters to draw conclusions that can have implications for the decision of macroeconomic policies.

From the methodological perspective, we observed that the relationship between economic variables is usually subject to local nonlinearities that are not possible to be captured by the classic linear econometric models. The local nonlinear behaviour can be captured through the application of the spline function in the model's estimator. The spline functions are piece-wise polynomials that are fitted to the observed data within the specified periods – a number of spline functions determine a degree of smoothing – which directly models the local nonlinearities.

Conclusion

The main objective of the paper is to outline a theoretical framework of econometric models with different forms in their regressor function. We start with a classic linear regression which is extended to a more flexible nonlinear form by transforming its covariates into spline functions. The spline functions have the advantage that they can capture local nonlinearities that are usually present in the relationship between economic variables. In the follow-up models, we include spatial spillover effects that are common in the observations from different regions. We start with a classic spatial autoregressive (SAR) model, which is further extended to have spline functions as its covariates, with an additional version that includes a spatial trend in the estimator.

In the empirical analysis, we apply these models to the economic dataset, which contains 209 European regions, with the aim of explaining the dynamics of the unemployment rate through four key economic determining factors.

The preliminary analysis shows that all of the determining factors have a strong nonlinear relationship with the unemployment rate on a standalone basis, which indicates that a simple linear model might not be the best estimator. The findings essentially confirm the importance of the identified determinants, and, in addition, the spatial econometric model estimates also highlight the significant spatial interdependence in the context of regional unemployment in the EU. The results show that the models with spline functions are a better fit than their classic counterparts. However, the only model that corrects the instability of the estimated parameters, which is caused by the violation of normality in residuals, is the spatial regression model with spline functions that also contain a spatial trend in the regressor function.

We conclude that a more complex model can correct local nonlinearities that cause the distortion in the models' estimates. Even though these models might be more elaborate in terms of economic interpretation, they eliminate the instability in the estimated parameters that might lead to incorrect conclusions that are used for decision-making in economic policies.

Our research can be further expanded to include more variables, and it can be tested in different economic scenarios.

References

1. Anselin L., & Rey, S. J. (2014). Modern Spatial Econometrics in Practice. GeoDa Press LLC, Chicago.
2. Basile, R., Durbán, M., Mínguez, R., María Montero, J., & Mur, J. (2014). Modeling regional economic dynamics: Spatial dependence, spatial heterogeneity and nonlinearities. *Journal of Economic Dynamics and Control, 48*, 229-245. https://doi.org/10.1016/j.jedc.2014.06.011
3. Basile, R., & Mínguez, R. (2018). Advances in Spatial Econometrics: Parametric vs. Semiparametric Spatial Autoregressive Models, In: Commendatore, P., Kubin, I., Bougheas, S., Kirman, A., Kopel, M., Bischi, G. (eds) The Economy as a Complex Spatial System., pp. 81-106, Springer Proceedings in Complexity. Springer.
4. Chi, G., & Zhu, J. (2020). Spatial Regression Models for the Social Sciences. https://doi.org/10.4135/9781544302096

5. Eurostat. (2023). Regional statistics, available at https://ec.europa.eu/eurostat/web/regions/database (15 March 2023).
6. Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. D. (2021). Regression Models, Methods and Applications, Springer Berlin, Heidelberg.
7. Formánek, T. (2019). Spatial econometric analysis with applications to regional macroeconomic dynamics. Habilitation Thesis, University of Economics, Prague.
8. Geniaux, G., & Martinetti, D. (2018). A new method for dealing simultaneously with spatial autocorrelation and spatial heterogeneity in regression models. *Regional Science and Urban Economics, 72*, 74-85. https://doi.org/10.1016/j.regsciurbeco.2017.04.001
9. Hastie, T., & Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science, 1*(3). https://doi.org/10.1214/ss/1177013604
10. Hastie, T. J., Tibshirani, R. J. (1990). Generalized Additive Models, Chapman & Hall/CRC.
11. Lung-Fei, L. (2022). Spatial Econometrics: Spatial Autoregressive Models , World Scientific Publishing Company, p. 896.
12. Mínguez, R., Basile, R., & Durbán, M. (2022). An introduction to pspatreg: A new R package for semiparametric spatial autoregressive analysis. *REGION,* 9(2), R1-R15. https://doi.org/10.18335/region.v9i2.450
13. Pavlovčič-Prešeren, P., Stopar, B., & Sterle, O. (2019). Application of different radial basis function networks in the illegal waste dump-surface modelling. *Central European Journal of Operations Research, 27*(3), 783-795. https://doi.org/10.1007/s10100-018-0586-z
14. Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., & Schmid, M. (2019). A review of spline function procedures in R. *BMC Medical Research Methodology, 19*(1). https://doi.org/10.1186/s12874-019-0666-3
15. Wahyuni, S. A., Ratnawati, R., Indriyani, I., & Fajri, M. (2020). Spline Regression Analysis to Modelling The Open Unemployment Rate in Sulawesi. *Natural Science: Journal of Science and Technology,* 9(2). https://doi.org/10.22487/25411969.2020.v9.i2.15202
16. Wood, S. N. (2017). Generalized Additive Models: An Introduction with R, Second Edition (2nd ed.). Chapman and Hall/CRC.
17. Wood, S. (2023). Mixed GAM Computation Vehicle with Automatic Smoothness Estimation, available at https://cran.r-project.org/web/packages/mgcv/mgcv.pdf (15 January 2024).

About the authors

Andrea Furková works as an Associate professor at the University of Economics in Bratislava. Her research interests are spatial econometrics and multi-criteria optimization. She participated as co-researcher, deputy head and head of several completed VEGA grant (the Grant Agency of the Slovak Republic) projects and COST (European Cooperation in Science and Technology) projects. The author can be contacted at **andrea.furkova@euba.sk**.

Peter Knížat is an external PhD student at the University of Economics in Bratislava. His research interests are spatial econometrics and functional data analysis. He works full-time at the Statistical Office of the Slovak Republic as a statistician, where he is responsible for proposing a statistical methodology for big data analysis. The author can be contacted at email: **peter.knizat@euba.sk**.