Transportation Research Part E 114 (2018) 121–143

Contents lists available at ScienceDirect

Transportation Research Part E

journal homepage: www.elsevier.com/locate/tre

Semi-parametric spatial autoregressive models in freight generation modeling

Tamás Krisztin

International Institute for Applied Systems Analysis, IIASA, Austria

ARTICLE INFO

Keywords:
Non-linear spatial autoregressive models
Genetic algorithms
Semi-parametric modeling
Freight generation

ABSTRACT

This paper proposes for the purposes of freight generation a spatial autoregressive model framework, combined with non-linear semi-parametric techniques. We demonstrate the capabilities of the model in a series of Monte Carlo studies. Moreover, evidence is provided for non-linearities in freight generation, through an applied analysis of European NUTS-2 regions. We provide evidence for significant spatial dependence and for significant non-linearities related to employment rates in manufacturing and infrastructure capabilities in regions. The non-linear impacts are the most significant in the agricultural freight generation sector.

1. Introduction

Regional freight generation models are widely used to model the volume of freight originating from regions. This modeling approach is a popular way of predicting future freight volumes, especially in the context of the so-called four-stage model of freight analysis. Moreover, such models are an important cornerstone in transportation planning (Ortúzar and Willumsen, 2011; Sánchez-Díaz, 2017; Sánchez-Díaz et al., 2015).

The classic freight generation model does not take spatial dependencies between the modeled regions into account. This limitation has been pointed out in recent literature, among others in Novak et al. (2011) and Sánchez-Díaz et al. (2016). Both studies emphasize the importance of spatial lags of the dependent variable in freight generation models. The theoretical motivation for such dependencies are economic spillovers, as well as the shared transportation infrastructure. Moreover, spatial dependence is more likely than spatial independence. Ignoring such dependencies can lead to severely biased estimates, as noted by Anselin and Bera (1998), Anselin et al. (2004), LeSage and Pace (2009) and Fischer and Wang (2011) among others.

A second shortcoming of the classic freight generation model, is that it does not control for non-linear impacts of the independent variables. Recent literature provides strong evidence for the presence of such non-linearities (Ranaiefar et al., 2013; Chow et al., 2010; De Grange et al., 2010; Hesse and Rodrigue, 2004). However, there is a lack of consensus in the literature over which functional form to use for the explanatory variables. Sánchez-Díaz et al. (2016) suggest multiple non-linear transformations (for example logarithmic, quadratic or exponential transformations, as in Novak et al. (2011)) of the explanatory variables, until a sufficient value of the chosen measure of fit is achieved. As noted by Tavasszy et al. (2012) and Rodrigue (2006) such an exploratory approach shows some drawbacks: first, it is difficult to specify ex ante which functional form would be the most appropriate for each variable. Second, testing for a wide number of transformations can be computationally burdensome, and third, including polynomials of higher order can lead to numerical instability.

Such non-linearities in the parameters (besides the non-linearities in the variables introduced by spatial dependencies), however, seem to play a central role in freight generation. Suggestions to deal with this issue include regression trees (Rodrigue, 2006; Holguín-

E-mail address: krisztin@iiasa.ac.at.

https://doi.org/10.1016/j.tre.2018.03.003
Received 16 September 2017; Received in revised form 7 February 2018; Accepted 2 March 2018

Veras and Patil, 2008; Ranaiefar et al., 2013; Al-Deek and El-Maghraby, 2000). These approaches, however, neglect to simultaneously control for spatial dependencies in the model. While some non-linear approaches, such as those by Ranaiefar et al. (2013) and Al-Deek and El-Maghraby (2000) might implicitly model spatial spillovers, they do not explicitly measure the intensity of the spatial dependencies between freight generating regions. Such information, however, can be of value for policy makers. Novak et al. (2011) account for spatial dependence in the error term, while opting for an explorative approach in testing out non-linear specifications for selected covariates. They conclude that there is strong support for non-linearities in the parameters even while taking into account spatial dependencies in the error terms.

While Chow et al. (2010) and De Jong et al. (2013) neglect taking the spatial aspect of freight generation into account, they capture non-linearities in the parameters in the freight generation model through a semi-parametric approach. Semi-parametric modeling in the context of spatial autoregressive (SAR) models was recently addressed by Basile (2008), Del Bo and Florio (2012), Fotopoulos (2012), Basile et al. (2014). These papers use a form of parameter expansion called basic splines,[1] based on locally defined piecewise polynomials, to model explanatory variables in a flexible way (Ruppert et al., 2003). Such basic splines have been a popular way for modeling non-linearities in a semi-parametric fashion, ever since their introduction in the seminal work by DeBoor (1978). The main advantage of this approach lies in the fact that each piecewise polynomial only forms a local basis, with unit integrals, and overlaps only with a limited number of other polynomials. Moreover, the upper range of basis function is limited, and the differentials of basic splines are readily available, as they are composed of piecewise polynomials themselves (Eilers and Marx, 1996). All of these properties ensure that the spline functions are easily tractable, both numerically and analytically (Fahrmeier et al., 2004). Moreover, in the case of spatially dependent explanatory variables, basic spline models can be estimated in the same fashion as classic SAR models (Basile, 2008).

The main disadvantage of using basic splines lies in the fact that the modeler has to choose a set of support points for the spline. On the one hand, if this set of support points is too small, the splines may not adequately reflect the non-linearity of the modeled function. On the other hand, if the number of support points is too large, the model may be severely overparameterized. This issue is quite relevant in the context of regional freight modeling, where usually cross-sectional data or small-scale panels are used for inference. Multiple approaches have been proposed to deal with this problem. As suggested by e.g. Koop and Poirier (2004), one could vary the number of spline support points in order to minimize certain criteria, for example, some form of information criterion (such as the one proposed by Akaike or the Bayesian information criterion). Another approach relies on selecting a priori a relatively large number of uniformly spline support points and using a form of Bayesian shrinkage through adequate choice of hyperpriors, such as in Eilers and Marx (1996).

This paper addresses the issue of spatial dependence and non-linearities in the parameters in freight generation models. Spatial dependence is addressed by using a SAR model, which features a spatial lag of the dependent variable, for freight generation modeling (Anselin, 1988; LeSage and Pace, 2009; Fischer and Wang, 2011). The theoretical motivation for such dependencies are economic spillovers amongst regions, as well as their shared transportation infrastructure. Moreover, spatial dependence is more likely than spatial independence. Ignoring such dependencies can lead to severely biased estimates, as noted by Anselin and Bera (1998), Anselin et al. (2004), LeSage and Pace (2009), Fischer and Wang (2011) among others.

The novelty of our approach lies in combining a spatial econometric model with a semi-parametric framework in an adaptive manner. Current spatial econometric approaches, such as Basile et al. (2014), rely on setting a fixed number of equidistant support points over the range of the non-linearly modeled variables. We argue, that such an approach does not adequately capture the non-linearities, and instead propose an adaptive method using a variant of genetic algorithms to find the – in terms of AIC – optimal number and position of support points for modeling each co-variate. We aim to demonstrate in this paper through multiple Monte Carlos studies that this approach leads to lower bias in parameter estimates, especially in the presence of moderate non-linearities. The basic principles of the adaptive spline knot selection algorithm are based on the ideas presented in Koch and Krisztin (2011). We differ from the approach presented in Krisztin (2017) by using an adaptive strategy for direct selection of spline knots, instead of a Bayesian approach with a penalization term (which relies on a large number of pre-selected spline knots). While the penalization strategy is a valid approach to this problem, the genetic algorithm approach presented in this paper allows for a more flexible estimation and does not have to resort to Bayesian methods.

Section 2 introduces the classic SAR model in the context of freight generation, and discusses common issues in parameter interpretation and estimation. Section 3 puts forth a semi-parametric variant of the classic SAR model, as a possible way of modeling non-linearities in the parameters coupled with a spatial lag of the dependent variable. This section develops a semi-parametric SAR model, which uses basic splines, coupled with a numerical optimization procedure, to adequately limit the problem of over-parametrization. Section 4 discusses in detail the estimation algorithm and further econometric issues related to the estimation procedure. Section 5 provides evidence in the context of a Monte Carlo simulation study that the proposed approach can adequately model non-linearities in the parameters, without over-fitting. Moreover, the proposed estimation method is compared to a classic SAR model (with no semi-parametric modeling). Section 6 includes an application in freight generation modeling. In the context of this application, the proposed semi-parametric approach is applied to European freight generation data, covering 258 NUTS-2 regions in 2011. The dataset includes both aggregate freight generation over all sectors and sector specific data for the agricultural, mining and food sectors. Based on this data, the applicability of the proposed semi-parametric estimation approach is demonstrated. Moreover,

[1] Basic splines are a class of semi-parametric basis function, which can be used to approximate non-linearities in the parameters. This is achieved by using a large set of overlapping piecewise polynomials (the so-called bases) in order to model each explanatory variable (DeBoor, 1978). Note, however, that basic splines are linear in the parameters and only approximate non-linearities in the parameters by transforming the explanatory variable in a non-linear fashion.

the proposed modeling approach is contrasted with a classic SAR model, which is linear in the parameters. The results provide evidence for significant spatial dependence. The semi-parametric variant of the model performs better, both in terms of in-sample fit and in predicting freight generation in 2012. The results suggest significant non-linear impacts in regions' road infrastructure and in the share of manufacturing employment. Finally, Section 7 concludes.

2. The spatial autoregressive model of freight generation

Let us consider a set of N freight generating regions and let i denote a specific region ($i = 1,...,N$). Further, let us denote the volume of freight originating from these regions by the $N \times 1$ vector $\boldsymbol{y}$. The core assumption of the SAR model is that the volume of freight originating from region i does not solely depend on K explanatory variables and a normally distributed error term, but also on the freight generation of neighboring regions. More formally, we assume that $\boldsymbol{y}$ can be modeled in the following fashion:

$$\boldsymbol{y} = \rho \mathbf{W}\boldsymbol{y} + \boldsymbol{\iota}_N \beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (1)$$

$$\boldsymbol{\varepsilon} \sim \mathscr{N}(0,\mathbf{I}_N \sigma^2)$$

where $\mathbf{W}$ is an exogenously given $N \times N$ spatial weight matrix of known constants. $w_{i,j}$ is a typical element of $\mathbf{W}$ in the i-th row and j-th column ($i,j = 1,...,N$). If regions i and j are considered to be neighbors, $w_{i,j} > 0$, otherwise $w_{i,j} = 0$. Moreover, no region can be considered a neighbor to itself, therefore $w_{i,j} = 0 \ \forall \ i = j$. We assume that $\mathbf{W}$ is doubly stochastic, that is $\sum_i^N w_{i,j} = \sum_j^N w_{i,j} = 1$. ρ denotes the spatial autoregressive parameter. We assume that ρ is restricted to the parameter space $-1 \leqslant \rho \leqslant 1$. The SAR model in Eq. (1), subsumes the classic linear freight generation model as a special case, in the case of $\rho = 0$. $\boldsymbol{\iota}_N$ an $N \times 1$ vector of ones, β_0 the corresponding intercept, $\mathbf{X}$ is an $N \times K$ matrix of explanatory variables, and $\boldsymbol{\beta}$ the corresponding $K \times 1$ coefficient vector. $\boldsymbol{\varepsilon}$ is an $N \times 1$ vector of independently and identically distributed error terms, with zero mean and $\mathbf{I}_N \sigma^2$ variance, where $\mathbf{I}_N$ denotes an $N \times N$ identity matrix.

A special feature of the spatial autoregressive model in Eq. (1) is that the volume of freight generated by region i does not only depend on the explanatory variables associated with i, but on its neighbors as well (and these in turn depend on their neighbors). Due to the spatial dependence of $\boldsymbol{y}$, it would not be correct to estimate the model in Eq. (1) via ordinary least squares (OLS), using $[\mathbf{W}\boldsymbol{y},\boldsymbol{\iota}_N,\mathbf{X}]$ as a matrix of explanatory variables. Besides leading to biased estimates (Anselin and Bera, 1998; Anselin et al., 2004; LeSage and Pace, 2009; Qu and Lee, 2015), such a model would suffer from endogeneity and from serial correlation in the errors.

Instead the model should be estimated as a system of equations. This is apparent if we re-write the model its reduced form:

$$\boldsymbol{y} = (\mathbf{I}-\rho\mathbf{W})^{-1}(\boldsymbol{\iota}_N \beta_0 + \mathbf{X}\boldsymbol{\beta}) + (\mathbf{I}-\rho\mathbf{W})^{-1}(\boldsymbol{\varepsilon}). \qquad (2)$$

This is a non-linear model, which cannot be estimated via OLS. However, estimates for $\rho,\beta_0,\boldsymbol{\beta}$, and σ^2 in Eq. (2) can be found by maximizing the log-likelihood function (see LeSage and Pace, 2009, p. 47):

$$\mathscr{L} = -N\log(\pi\sigma^2)/2 + \log[\det(\mathbf{A})] - \boldsymbol{e}'\boldsymbol{e}/2\sigma^2 \qquad (3)$$

$$\boldsymbol{e} = \mathbf{A}\boldsymbol{y} - \boldsymbol{\iota}_N \beta_0 - \mathbf{X}\boldsymbol{\beta}$$

where $\mathscr{L}$ denotes the log-likelihood, $\mathbf{A} = (\mathbf{I}_N - \rho\mathbf{W})$ and $\det(\mathbf{A})$ denotes the determinant of $\mathbf{A}$. Conditional on $\hat{\rho}$ (let estimates be denoted with $\wedge$), estimates for $\hat{\beta}_0,\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are available in closed form. LeSage and Pace (2009) provide efficient computational algorithms to estimate $\hat{\rho}$, conditional on the closed form estimators for $\hat{\beta}_0,\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$. These can be used of the parameter $\rho \in [1/wmin,1/wmax]$ (where $wmin$ and $wmax$ denote the smallest and largest eigenvalue of $\mathbf{W}$, respectively) and if the matrix $\mathbf{W}$ is row-stochastic.

As opposed to models containing no spatial lag of the dependent variable, in SAR models interpreting the impact of the k-th ($k = 1,...,K$) explanatory variable $\boldsymbol{x}_k$ (where $\boldsymbol{x}_k$ denotes the k-th column of $\mathbf{X}$) on the dependent variable is richer, but more complicated. This is due to the spatial connectivity relationships incorporated in the model. Consider, that a change in a single explanatory variable in region i has not only a "direct" impact on the volume of freight generated by region i (denoted as y_i), but also on the volume of freight generated by region j as well (where $j \neq i$). More formally, we can re-formulate the SAR model as:

$$\boldsymbol{y} = \mathbf{A}^{-1}\boldsymbol{\iota}_N \beta_0 + \sum_{k=1}^{K} \mathbf{S}_k \boldsymbol{x}_k + \mathbf{A}^{-1}\varepsilon \qquad (4)$$

$$\mathbf{S}_k = \mathbf{A}^{-1}(\mathbf{I}_N \beta_k)$$

where the $N \times N$ matrix $\mathbf{S}_k$ contains the partial derivative impacts of a change in $\boldsymbol{x}_k$. Its (i,j)-th element – denoted by $S_k(i,j)$ – contains the partial derivative impact of the k-th co-variate on y_i, that is:

$$\frac{\partial y_i}{\partial x_{j,k}} = S_k(i,j). \qquad (5)$$

This implies that the standard interpretation of the estimated parameters as partial derivatives does not apply in the case of SAR models. However, interpreting the full $N \times N$ partial derivative matrix $\mathbf{S}_k$ is not practicable. To alleviate this issue, LeSage and Pace (2009) introduce scalar summary impact measures. They label the average of the diagonal elements of $\mathbf{S}_k$ – that is $S_k(i,i)$ for all $i = 1,...,N$ – as the average *direct effect* of the k-th variable on $\boldsymbol{y}$. These effects include all impacts of the k-th variable on the freight generation of region i, as well as the feedbacks to the observation itself, stemming from neighboring regions. Average *indirect effects* arise as the average of the changes in all typical elements of the k-th explanatory variable $x_{j,k}$, where $j \neq i$. They can be obtained by taking the average of the off-diagonal elements of the i-th row of the matrix $\mathbf{S}_k$, for each observation i. The sum of average direct and indirect effects of the k-th covariate equal the average *total effects*.

Note, that the model in Eq. (1) is non-linear in the case of $\rho \neq 0$. The SAR model term, $\rho\mathbf{W}\boldsymbol{y}$, gives rise to this non-linearity. In essence, each element y_i of the dependent variable $\boldsymbol{y}$ depends upon its neighbor's freight generation (as they depend on their neighbor's in turn, etc.). The strength of this non-linear influence is determined by the coefficient ρ, and the structure is given by the exogenous matrix $\mathbf{W}$.

The model, however, is linear in the influence of the explanatory variables on the term $\mathbf{A}\boldsymbol{y}$, that is, the model in Eq. (1) is linear in the parameters $\boldsymbol{\beta}$. This can be an issue if the SAR model is to be used in the context of freight generation, where multiple studies (Hesse and Rodrigue, 2004; Chow et al., 2010; Tavasszy et al., 2012; Rodrigue, 2006; Novak et al., 2011; De Jong et al., 2013) provide evidence on non-linearities in the parameters. The classical approach to address this issue (see Novak et al., 2011; Ortúzar and Willumsen, 2011) is to explore multiple candidate models using non-linear transformations of the explanatory variable (such as logarithmic or quadratic functions). The most appropriate of these candidate models can then be determined using a fit statistic, such as the coefficient of determination. Such an approach of course only explores a very limited number of non-linear transformations and explanatory variables. Moreover, using polynomial transformations can lead to orthogonality and numerical issues in the matrix of explanatory variables.

3. A semi-parametric extension using splines

In order to incorporate non-linearities in the parameters, we extend the basic SAR freight generation model from Eq. (1). In this context, let us consider two sets of explanatory variables. The $N \times Q$ matrix $\mathbf{X}_1$ contains the first set of explanatory variables. These are modeled in a fashion that takes into account non-linearities in the parameters. The $N \times D$ matrix $\mathbf{X}_2$ contains the second set of explanatory variables, which are modeled linearly in the parameters. We model $\mathbf{X}_1$ through an unknown function $\mathscr{F}(\cdot)$:

$$\boldsymbol{y} = \rho\mathbf{W}\boldsymbol{y} + \iota_N\beta_0 + \mathscr{F}(\mathbf{X}_1) + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \tag{6}$$

where $\boldsymbol{\beta}_2$ is the $D \times 1$ coefficient vector corresponding to $\mathbf{X}_2$.

We further specify the function $\mathscr{F}(\cdot)$ as being the sum of Q unknown functions $f(\cdot)$. Each function $f(\cdot)$ models a column of $\mathbf{X}_1$, which we denote by the $N \times 1$ vector $\boldsymbol{x}_q$ (where $q = 1,...,Q$):

$$\mathscr{F}(\mathbf{X}_1) = \sum_{q=1}^{Q} f(\boldsymbol{x}_q,\boldsymbol{\theta}_q). \tag{7}$$

Note, that each $f(\cdot)$ also depends on an $L_q \times 1$ parameter vector $\boldsymbol{\theta}_q$, where the length L_q of the parameter vector can differ from covariate to covariate and $L_q < \infty$, as well as $L_q \in \mathbb{Z}^+$. In the general formulation presented in Eq. (7), the parametrization of $f(\cdot)$ may be non-linear in the explanatory variable $\boldsymbol{x}_q$, non-linear in the parameters $\boldsymbol{\theta}_q$, or it can be non-linear in both.

The advantage of using the non-linear function $f(\cdot)$ to model $\boldsymbol{x}_q$ over a simple linear representation is that this allows a greater flexibility, and – in principle – a greater accuracy in prediction and impact estimation (see White, 2000). Offsetting these two advantages are some potentially severe disadvantages: first, non-linear models can overfit the data, which would result in inferior performance in comparison to linear models. Second, the estimation might pose a serious computational challenge. Finally, the resulting parameter estimates can be difficult to interpret.

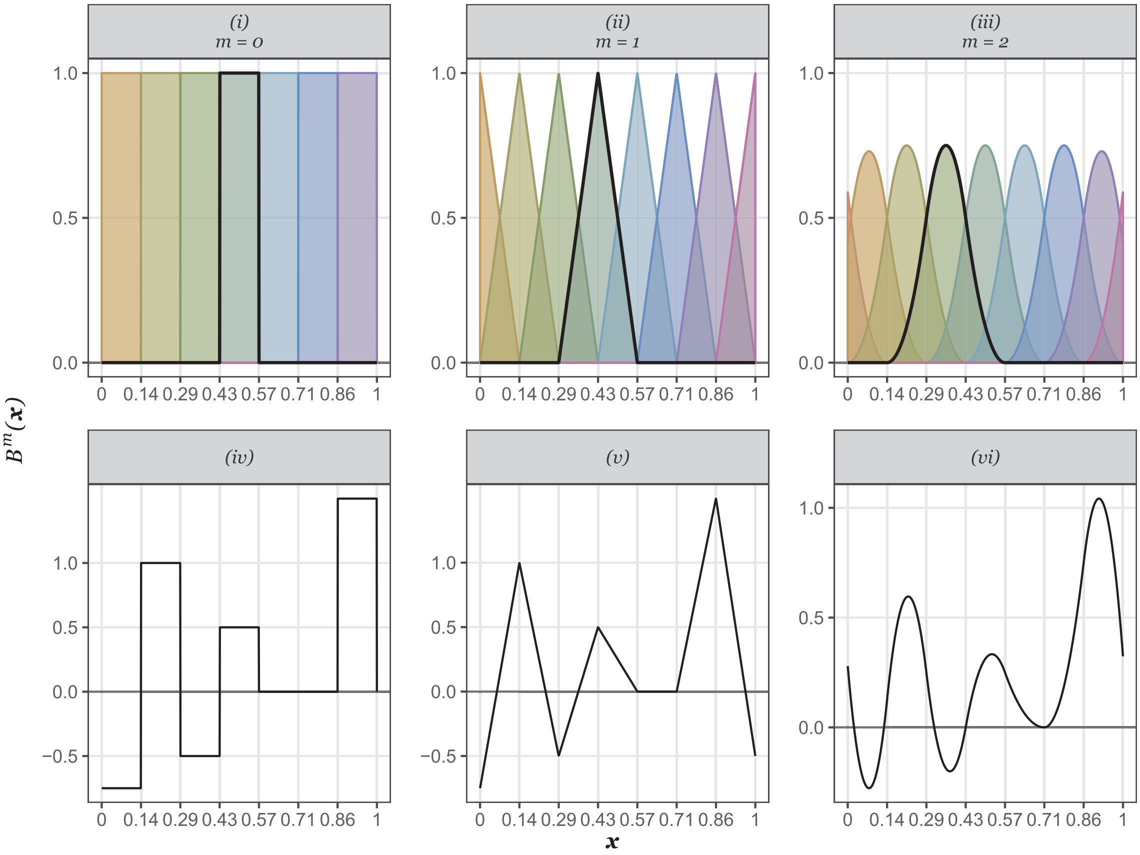
A direct approach might assume that the function $f(\cdot)$ is truly non-linear in the parameters. In such a case, there generally is no closed form solution for an optimal coefficient vector $\boldsymbol{\theta}_q$ (White, 2000). A potentially useful estimation algorithm might be constructed through an iterative approach, which tries out successive candidate values until convergence – as defined by a suitably chosen metric – is achieved. Such an optimization procedure, however, can be challenging to apply in practice. It involves fine-tuning the optimization algorithm, which might not be well behaved. Moreover, convergence is not guaranteed, and when confronted with a complex solution space featuring a large number of local optimum values, the algorithm might fail to converge altogether. Even if convergence was achieved, it is impossible to determine whether the algorithm converged merely to a local or to the desired global optimum.

Since these computational challenges arise from $f(\cdot)$ being truly non-linear in the parameters, one could find a class of functions, which – although linear in the parameters – possess sufficient flexibility to approximate $f(\cdot)$. This motivates semi-parametric approaches, which use a form of parameter expansion for $f(\cdot)$, divided into L_q (potentially overlapping) segments of basis functions, denoted as $B_l(\cdot)$ (where $l = 1,...,L_q$). Subsequently, $f(\cdot)$ is modeled as a linear combination of the L_q basis functions:

$$f(\boldsymbol{x}_q,\boldsymbol{\theta}_q) \approx \sum_{l=1}^{L_q} \theta_{l,q}B_l(\boldsymbol{x}_q) \tag{8}$$

where we denote the l-th element of the parameter vector $\boldsymbol{\theta}_q$ as $\theta_{l,q}$.

This spline parametrization is non-linear in $\boldsymbol{x}_q$, but linear in the parameters $\boldsymbol{\theta}_q$, thus it allows for flexible modeling of $f(\cdot)$. This eliminates the computational challenges arising from non-linearity in the parameters: the model is not prone to be stuck at local minima and it can be estimated in the same fashion as the classic SAR model. For the purpose of interpreting the summary impact measures of the SAR model, it is essential that the basis functions be smooth, continuous polynomials, with continuous first-order



Notes: The splines model a $10,000 \times 1$ vector $\boldsymbol{x}$, containing uniformly distributed observations between zero and one. The position of the equally spaced support points is marked on the x-axis of each panel. For constructing the functional forms in panels (iv) to (vi) the coefficient vector $\boldsymbol{\theta} = [1.00, -0.75, 1.00, -0.50, 0.50, 0.00, 0.00, 1.50, -0.50]'$ was used.

Fig. 1. Illustration of spline basis functions of the 0-th (i), first (ii), and second degree (iii), as well as their respective functional forms (iv) to (vi). **Notes**: The splines model a 10,000 × 1 vector $\boldsymbol{x}$, containing uniformly distributed observations between zero and one. The position of the equally spaced support points is marked on the x-axis of each panel. For constructing the functional forms in panels (iv) to (vi) the coefficient vector $\boldsymbol{\theta} = [1.00,-0.75,1.00,-0.50,0.50,0.00,0.00,1.50,-0.50]'$ was used.

derivates. A further condition is that the basis functions be orthogonal. We follow evidence form Basile (2008), Basile et al. (2014), and Basile et al. (2013), and use basic splines to model $B_l(\cdot)$.

3.1. Basic splines

Basic splines are essentially locally defined polynomials of the m-th (with $m = 0,...,M$) degree (DeBoor, 1978). They consist of recursively defined basis functions $B_l^m(\cdot)$. Each basis function $B_l^m(\cdot)$ is a polynomial and is defined over only a partial range of the modeled explanatory variable vector $\boldsymbol{x}_q$. Let the $(L_q + m + 1) \times 1$ vector $\boldsymbol{\kappa}_q$, denote the total set of support points for $f(\boldsymbol{x}_q)$, with the l-th element of $\boldsymbol{\kappa}_q$ being denoted as $\kappa_{l,q}$. Then, the l-th basis function $B_l^m(\cdot)$ is defined only between the knots $\kappa_{l,q}$ and $\kappa_{l+m+1,q}$ and is zero otherwise.

The total vector of support points $\boldsymbol{\kappa}_q$ must satisfy the requirement that:

$$\kappa_{1,q} \leqslant ... \leqslant \kappa_{l,q} \leqslant ... \leqslant \kappa_{L_q+m+1,q} \tag{9}$$

where the first and last support points are defined so that $\kappa_{1,q} = min(\boldsymbol{x}_q)$ and $\kappa_{L_q+m+1,q} = max(\boldsymbol{x}_q)$. Here, $min(\boldsymbol{x}_q)$ denotes the smallest and $max(\boldsymbol{x}_q)$ the largest element of $\boldsymbol{x}_q$, respectively.

The nature of basis functions of varying degree is illustrated in Fig. 1. The top row [panels (i) to (iii)] depicts spline basis functions. The bottom row [panels (iv) to (vi)] shows the resulting functional fit, when the basis functions multiplied by the parameter vector $\boldsymbol{\theta} = [1.00,-0.75,1.00,-0.50,0.50,0.00,0.00,1.50,-0.50]'$, respectively.

Panels (i) to (iii) illustrate basis functions of the 0-th, first and second degree, respectively. In all three panels the range of the semi-parametrically modeled variable is subdivided into nine equal segments by eight equally distributed spline knots. The position of spline knots is marked on the x-axis. The varying colors denote different basis functions, which are positive only in a specific range and zero otherwise. For illustration purposes, the fourth basis function is outlined in black in panels (i) to (iii).

The full functional forms of the spline functions using 0-th, first and second degree basic functions and multiplied by a specific parameter vector θ are illustrated in panels (iv) to (vi) in Fig. 1. Splines of 0-th degree ($m = 0$) [in panel (i)] are only defined in the interval between two spline knots, where they are equal to one, and zero otherwise. They do not overlap and their resulting spline function in panel (iv) is not continuous. Basis functions of the first degree ($m = 1$) – depicted in panel (ii) – span the interval of exactly two spline knots and overlap with exactly one of their respective left and right neighbors. The left- and rightmost basis functions extend beyond the range of the modeled variable. First degree basis functions, however, also do not produce a continuous spline function [see panel (v)]. The second order basis functions ($m = 2$) in panel (iii) are quadratic polynomials, defined exactly over the range of three spline knots and they overlap with exactly four neighboring basis functions. In an analogous fashion to $m = 1$ in panel (ii), the two left- and rightmost basis functions extend beyond the range of the modeled variable. As illustrated in panel (vi), quadratic basis functions produce a continuous non-linear spline curve in conjunction with an additive parameter vector. The functional form of the resulting quadratic spline polynomial is depicted in panel (vi).

Basic splines have a recursive definition. Following DeBoor (1978) we can write the definition of a 0-th degree basic spline (where $m = 0$) as:

$$B_l^0(\boldsymbol{x}_q) = \mathscr{I}_{[\varkappa_{l,q},\varkappa_{l+1,q}]}(\boldsymbol{x}_q) = \begin{cases} 1 & \varkappa_{l,q} \leqslant x_{i,q} < \varkappa_{l+1,q} \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where $x_{i,q}$ is the i-th element of $\boldsymbol{x}_q$, and $\mathscr{I}_{[\varkappa_{l,q},\varkappa_{l+1,q}]}(\cdot)$ is a function, which takes on the value of one if $x_{i,q}$ lies in the interval $\varkappa_{l,q}$ to $\varkappa_{l+1,q}$, but is zero otherwise. It is evident that the basis function of order $m = 0$ in Eq. (10) does not overlap with any neighboring splines. It should also be obvious that a basic spline representation of the 0-th degree of $\boldsymbol{x}_q$ is not a continuous function.

A first order basic spline, with $m = 1$ is defined in a recursive fashion:

$$B_l^1(\boldsymbol{x}_q) = \frac{\boldsymbol{x}_q - \varkappa_{l,q}}{\varkappa_{l+1,q} - \varkappa_{l,q}} B_l^0(\boldsymbol{x}_q) + \frac{\varkappa_{l+2,q} - \boldsymbol{x}_q}{\varkappa_{l+2,q} - \varkappa_{l+1,q}} B_{l+1}^0(\boldsymbol{x}_q) \tag{11}$$

where $B_l^1(\cdot)$ denotes a spline basis of the first degree. It can be readily observed that the basic spline in Eq. (11) is simply the product of two overlapping basic splines of the 0-th degree from Eq. (10). This can be extended for basic splines of arbitrary degree $m > 1$, with:

$$B_l^m(\boldsymbol{x}_q) = \frac{\boldsymbol{x}_q - \varkappa_{l,q}}{\varkappa_{l,q+m} - \varkappa_{l,q}} B_l^{m-1}(\boldsymbol{x}_q) + \frac{\varkappa_{l+m+1,q} - \boldsymbol{x}_q}{\varkappa_{l+m+1,q} - \varkappa_{l+1,q}} B_{l+1}^{m-1}(\boldsymbol{x}_q). \tag{12}$$

Such basis functions exhibit a number of desirable properties, which makes them numerically easy to handle. Thus they pose an attractive choice for semi-parametric modeling purposes (Eilers and Marx, 1996; Ruppert et al., 2003; Fahrmeir et al., 2009). First of all, they form a local basis. At any given point in the range of $\boldsymbol{x}_q$, only $m + 1$ piecewise functions are simultaneously not equal to zero. Moreover, all local basis functions have the same functional form, they are merely shifted along the horizontal axis. Second, basic splines bases have unit sums, that is $\sum_{l=1}^{L_q} B_l^m(\boldsymbol{x}_q) = 1$. Furthermore, their upper range is limited to be lower or equal to one. This implies desirable numerical properties for regression analysis. Third, their derivatives are easy to calculate, due to the recursive definition of basic spline piecewise polynomials. The first-order derivative of each basis function can be expressed as:

$$\frac{\partial}{\partial \boldsymbol{x}_q} B_l^m(\boldsymbol{x}_q) = m\left(\frac{B_l^{m-1}(\boldsymbol{x}_q)}{\varkappa_{l+m,q} - \varkappa_{l,q}} - \frac{B_{l+1}^{m-1}(\boldsymbol{x}_q)}{\varkappa_{l+m+1,q} - \varkappa_{l+1,q}}\right) \tag{13}$$

where $\frac{\partial}{\partial x_q} B_l^m(\boldsymbol{x}_q)$ denotes the first-order derivative of the basic spline function of order m. Thus, the total derivative of the q-th semi-parametric function $f(\cdot)$ from Eq. (8) can be expressed as:

$$\frac{\partial}{\partial \boldsymbol{x}_q} f(\boldsymbol{x}_q) \approx \frac{\partial}{\partial \boldsymbol{x}_q} \sum_{l=1}^{L_q} \theta_{l,q} B_l^m(\boldsymbol{x}_q) = m \sum_{l=1}^{L_q} \frac{\theta_{l,q} - \theta_{l-1,q}}{\varkappa_{l+m} - \varkappa_l} B_l^{m-1}(\boldsymbol{x}_q). \tag{14}$$

This implies that the derivative of a full polynomial basic spline can be expressed using the differences in the knot points of a basic spline of degree $m-1$. Therefore, estimating the parameter vector θ_q not only provides an estimate for the polynomial form of $\boldsymbol{x}_q$, but also for its derivatives.

We can write the full set of basis functions of $f(\boldsymbol{x}_q)$ in terms of a $N \times L_q$ matrix $\mathbf{Z}_q = [B_1^m, \ldots, B_{L_q}^m]$, so that $f(\boldsymbol{x}_q) = \mathbf{Z}_q\theta_q$. In a similar fashion, the total set of spline functions in Eq. (8) can be written in terms of an $N \times \sum_{q=1}^{Q} L_q$ design matrix $\mathbf{Z} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_Q]$. Thus, the semi-parametric SAR model from Eq. (7) can be expressed as:

$$\boldsymbol{y} = \rho \mathbf{W}\boldsymbol{y} + \iota_N \beta_0 + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \tag{15}$$

where $\boldsymbol{\gamma} = [\theta_1', \ldots, \theta_Q']'$ denotes the $\left[\sum_{q=1}^{Q} L_q\right] \times 1$ vector of basic spline coefficients.

Given the set of design knots, a maximum likelihood estimation approach could provide an estimate for $\boldsymbol{\gamma}$. This estimate is, of course, dependent on the exact position and number of design knots, which we did not yet specify, except for the general requirements in Eq. (9). Note, that the exact number and position of the design knots has a considerable influence on the abilities of $\mathscr{F}(\cdot)$ to approximate non-linear functions. This would indicate that choosing a large number of equally spaced knot points is advantageous. Consider, however, that such an approach drastically increases the size of the design matrix $\mathbf{Z}$ and the number of parameters to estimate. To avoid such problems, the next section proposes an estimation method incorporating a novel approach for determining the optimal number and position of spline knots.

4. Adaptive spline knot selection

Due to the fact that basic splines are a parametric expansions of the modeled variables, we can use well documented procedures for model estimation, conditional on an a priori given number and position of spline support knots. Given a set of spline knots $\boldsymbol{\kappa}_q$ for ($q = 1,...,Q$), estimators for the coefficients $\rho,\boldsymbol{\gamma},\beta_0,\boldsymbol{\beta}_2,\sigma^2$ in Eq. (15) can be found by maximizing the log-likelihood conditional on ρ, in an analogous fashion to Eq. (3):

$$\mathscr{L} = -N\log(\pi\sigma^2)/2 + \log[\det(\mathbf{A})] - \boldsymbol{e}_o{}'\boldsymbol{e}_o/2\sigma^2 \tag{16}$$

$$\boldsymbol{e}_o = \mathbf{A}\boldsymbol{y} - \boldsymbol{\iota}_N\beta_0 - \mathbf{Z}\boldsymbol{\gamma} - \mathbf{X}_2\boldsymbol{\beta}_2.$$

While there generally is no closed form solution for an optimal set of spline knots, however, given a total set of candidate spline knots $\boldsymbol{\kappa}$ (where $\boldsymbol{\kappa} = [\boldsymbol{\kappa}_1',...,\boldsymbol{\kappa}_Q']$), we can evaluate the likelihood of the resulting model. Thus, a possible way of selecting the optimal number and position of spline knots would be using measures to compare models with differing number of spline knots. Such a comparison can be performed, for instance using the information criterion proposed by Akaike (1974)[2]:

$$AIC_c(\boldsymbol{\kappa}) = 2\mathscr{K} - 2\mathscr{L} + \frac{(\mathscr{K}+1)2\mathscr{K}}{N-\mathscr{K}-1} \tag{17}$$

$$\mathscr{K} = \sum_{q=1}^{Q} L_q + D + 3$$

where $\mathscr{K}$ denotes the total number of parameters in Eq. (15) and D is the number of covariates modeled in a linear in the parameters fashion. The Akaike information criterion is particularly well suited for the task of optimization, as it provides a convex optimization space, with no local minima (Awad, 1996; Akaike, 1974).

The classic approach to spline knot selection is to subdivide the range of the explanatory variables into equal segments and select a fixed number of spline knots. We refer to this spline knot selection strategy as *fixed spline knot selection*. The main drawback of this approach is that the optimal number of spline knots is difficult to evaluate ex ante, and that placing equidistant knots could put a large number of knots in the range of values where we have relatively few observations on explanatory variables.

Instead we elect to use an adaptive strategy for spline knot selection, which we denote as *adaptive spline knot selection*. Our main assumption is that the position of the spline knots for each spline $f(\boldsymbol{x}_q)$ is not on a continuous scale, but restricted to the observations in $\boldsymbol{x}_q$, that is $\kappa_{l,q} \in \boldsymbol{x}_q \ \forall \ l,q$. The main drawback of this assumption is that we can't model covariates with less than m unique observations (such as dummy variables).[3] Moreover, $\boldsymbol{x}_q$ should not contain a large number of outlying observations, as this might impact the ranges of the basis functions. The advantage of this assumption is that it limits the number of possible spline knot locations and transforms the problem of selecting a suitable $\boldsymbol{\kappa}$ to a (potentially large-scale) combinatorial optimization problem.

To facilitate combinatorial optimization, we can – based on the assumptions above – encode the candidate spline knots $\boldsymbol{\kappa}$ as a binary candidate vector $\boldsymbol{s}$. For this purpose, consider the matrix $\mathbf{H}$, which corresponds to the matrix of covariates $\mathbf{X}_1$, with each column having its elements sorted in an ascending fashion (that is $h_{1,q} \leqslant h_{2,q} \leqslant ... \leqslant h_{N,q} \ \forall \ q$, where $h_{i,q}$ is a typical element of the matrix $\mathbf{H}$). Let the $Z \times 1$ (where $Z = NQ$) vector be $\boldsymbol{h} = vec(\mathbf{H})$, where the $vec(\cdot)$ operator stacks all columns of a matrix. Let the $Z \times 1$ candidate solution vector $\boldsymbol{s}$ be a binary vector, with a value of one at position z (with $z = 1,...,Z$) if a spline knot is placed on the z-th element of the vector $\boldsymbol{h}$, and zero otherwise. Thus, the total set of spline knots $\boldsymbol{\kappa}$ for a candidate solution $\boldsymbol{s}$ can be obtained by $\boldsymbol{\kappa} = [\boldsymbol{h} \odot \boldsymbol{s} | \boldsymbol{s} = 1]$, where $\odot$ denotes the Hadamard product. We will denote this encoding process via the function $b(\cdot)$ (where $b(\boldsymbol{s}) \rightarrow \boldsymbol{\kappa}$).

4.1. A genetic algorithm approach

We utilize a genetic algorithm approach, in order to solve the combinatorial optimization problem of adaptively selecting the optimal number and positions of spline knots. Genetic algorithms are stochastic search algorithms inspired by evolutionary processes, and have proved to be successful in tackling large scale combinatorial problems (Fischer and Leung, 1998), such as the well-known traveling salesman problem (Kazemi et al., 2009; Aydin and Fogarty, 2004). The underlying mechanism involves an iterative procedure, where in each iteration a population of candidate solutions to a given problem is evaluated, based on some form of scoring measure, termed as the *fitness function*. Such an approach favors a stochastic exploration of a large part of the solution space. While a genetic algorithm in itself requires only simple computational steps (such as generating uniformly distributed numbers), their main drawback is that a large number of fitness evaluations have to be performed.

The classical genetic algorithm – as outlined in the seminal work by Holland (1992) – is suited to tackle discrete optimization problems formulated in the following fashion:

$$\min\{g(\boldsymbol{s}) | \boldsymbol{s} \in \Psi\} \tag{18}$$

where the non-constant function $g(\cdot)$ is the so-called *fitness function* mapping the total candidate solution space Ψ, so that $g\colon \Psi \rightarrow \mathbb{R}$.

[2] In practice, it is recommendable to use the adjusted AIC, termed as AIC_c, proposed by Hurvich et al. (1998), which corrects for finite sample size. The formula provided here corresponds to AIC_c. The AIC_c has generally the same convexity properties as the original Akaike information criterion.

[3] This is due to the fact that the basic spline representation of a given vector needs a minimal number of observations.

We assume a single solution vector $\boldsymbol{s} \in \{0,1\}^Z$, where Z is defined as $Z = NQ$, that is the total number of elements in $\mathbf{X}_2$. Note, that the search space Ψ is a hypercube with dimensions 2^Z.

The genetic algorithm is initialized with a population set P, containing N_P candidate solutions ($\boldsymbol{s}_1,...,\boldsymbol{s}_{N_P}$), with $P \subset \Psi$. N_P-1 of these solutions are randomly generated, while the first of the initial solutions equals a model with a fixed set of spline knots. The reasoning for the randomly generated initial population is that it is that is not known a priori where the globally optimal spline knots in Ψ are located. The fixed spline knot initial solution, however, ensures that the algorithm performs at least as well as the fixed spline knot model. Each individual candidate solution $\boldsymbol{s}_r$ (with $r = 1,...,N_P$) represents a feasible solution to the problem in Eq. (18), and its function value $g(\boldsymbol{s}_k)$ is termed as its fitness score, which has to be minimized. We use AIC_c as a *fitness function*, therefore:

$$g(\boldsymbol{s}_r) = AIC_c[b(\boldsymbol{s}_r)]. \tag{19}$$

Running the genetic algorithm involves an iterative execution of processing steps, which modify and delete members of P. Let us denote the maximum number of iterations as T and a specific iteration as t (with $t = 0,...,T$), with the population at time t being denoted as $P(t)$. Based on the population $P(0)$ subsequent populations $P(t)$ are generated by using three genetic operators: *selection*, *mutation* and *crossover*.

The *selection* into the sets P_{mate} and P_{death} is carried out via roulette wheel selection (Goldberg et al., 1989). The set P_{mate} contains all the population members selected for crossover and mutation and the P_{death} set contains the population members to be removed from the population. This is a proportional random selection technique, where a fixed number ($N_{P_{mate}}$ and $N_{P_{death}}$, respectively) of random experiments are carried out in succession. The probability of candidate solution $\boldsymbol{s}_k$ being selected to be in the sets P_{mate} and P_{death}[4] is:

$$p(\boldsymbol{s}_r \in P_{mate}) = \frac{\max[P(t)]-g(\boldsymbol{s}_r)}{\sum_{r=1}^{N_P}\{\max[P(t)]-g(\boldsymbol{s}_r)\}} \tag{20}$$

$$p(\boldsymbol{s}_r \in P_{death}) = \frac{g(\boldsymbol{s}_r)-\min[P(t)]}{\sum_{r=1}^{N_P}\{g(\boldsymbol{s}_r)-\min[P(t)]\}} \tag{21}$$

where $\max[P(t)]$ and $\min[P(t)]$ denote the largest and smallest fitness scores from the population at t, respectively. In essence, candidate solutions are selected (and copied into their respective intermediate populations) with probabilities according to Eqs. (20) and (21), based on their fitness relative to the fitness of all other candidate solutions in the population. Note, that a single candidate solution can be copied multiple times into the intermediate populations with this procedure.

After selection, the genetic operators of single point *crossover* and bit string *mutation* are applied subsequently to the intermediate population P_{mate}, in order to create the population set in the next iteration $P(t + 1)$. These two operators serve the purpose of generating new sample points from the total solution space Ψ, while partially preserving the distribution of spline knots across the hyperplanes present in the population at t. Single point *crossover* involves recombining existing spline knot sets to explore new regions of Ψ. The $N_{P_{mate}}$ candidate solutions are randomly matched in a pairwise fashion. Each pair (parents) is combined by choosing a random point between one and Z with a uniform probability distribution. Then both binary vectors are cut into two parts and juxtaposed, by using the first part from one parent up to the randomly selected point and then using the binary information from the second parent. Through this procedure two new candidate solutions are generated. These are termed the offspring of the selected pair of parents.

Finally, after the crossover operation has been completed, the bit string *mutation* operator is applied with a uniform probability to randomly selected members of P_{mate}. This operation involves a bit-wise recombination of the binary vector $\boldsymbol{s}_r$. The role of this operation is to ensure that the entire solution space of Ψ remains accessible and that the algorithm does not get stuck at local optima. During this operation a random position between one and Z is selected in $\boldsymbol{s}_r$. If the value of $\boldsymbol{s}_r$ at the z-th place was zero, then it is flipped to one. If the value was one, it is flipped to zero. After this mutation step is complete, the selected candidate solutions, together with the offspring from the crossover step, are copied to replace the ones selected in the step P_{death}. Then the whole process is repeated for $t + 1$: with first evaluating the fitness of each population member, then selecting them into two intermediate sets and finally applying the crossover and mutation operators.

The steps of the genetic algorithm can be expressed as:

- Generate N_P candidate solutions form a uniform distribution. Let us denote the population at $t = 0$ as $P(0) = \{\boldsymbol{s}_1,...,\boldsymbol{s}_{N_P}\}$, with $P(0) \subset \Psi$.
- Evaluate the fitness score (AIC_c) of each $\boldsymbol{s}_r$, that is calculate $g(\boldsymbol{s}_r)\ \forall\ r$ in the current population $P(t)$.
- Apply the roulette wheel *selection* operator in order to generate two intermediate subpopulations of the same length ($N_{P_{mate}}$, with $N_{P_{mate}} < N_P$): the so-called mating pool P_{mate} [$P_{mate} \subset P(t)$] and the death pool P_{death} [$P_{death} \subset P(t)$].
- Generate new candidate solution from the mating pool P_{mate} via the single point *crossover* and the bit string *mutation* operators and replace all members of the death pool P_{death} with the newly generated candidate solutions.
- Set $t = t + 1$ and proceed with *Step 2*, until one or both of the stopping criteria have been reached: either $t = T$ or the relative gain in the best fitness score is smaller than 10^{-4}. Once the algorithm has stopped, designate the candidate solution with the highest

[4] Note, that effectively two separate roulette wheel selections are carried out: one for P_{mate} and one for P_{death}. For this reason $p(\boldsymbol{s}_r \in P_{mate})$ in Eq. (20) assigns high probability of being selected to population members with high ranking fitness scores, while $p(\boldsymbol{s}_r \in P_{death})$ in Eq. (21) assigns a high probability of being selected to population members with low ranking fitness scores. This implies that it is entirely possible for a population member $\boldsymbol{s}_r$ to be part of P_{mate} and P_{death} at t.

fitness as the result of the genetic algorithm.

Note, that the above algorithm's computational time to convergence increases sharply with the number of spline knots and modeled explanatory variables. In order to alleviate this problem, we use an extension of the genetic algorithm approach, proposed by Sycara (1998). Herein, we employ multiple genetic algorithms, which are executed in a parallel processing environment. Each algorithm has an own independent population set. However, the population size, the size of the set P_{mate} and P_{death} can differ among the genetic algorithms. After a fixed number of iterations, or when a set amount of processing time has passed, one randomly selected member is deleted from each of the independent population sets. Finally, one randomly selected population member from each independent population is transferred to another population set. This step ensures, that no genetic algorithm is stuck in a local optimum. Talukdar et al. (1998) used the well-known shortest path problem as a benchmark for this approach. By applying this methodology to the problem they demonstrated that the solutions provided by decentralized genetic algorithms converge quicker towards an optimal solution.

5. Performance on artificial data

In order to asses the performance of the semi-parametric SAR model, we run a series of Monte Carlo studies, based on artificially generated datasets. Our benchmark data generating process is a semi-parametric SAR model from Eq. (6), without a constant and containing two semi-parametrically modeled explanatory variables ($Q = 2$):

$$(\mathbf{I}_N - \rho \mathbf{W})\boldsymbol{y} = \mathscr{F}(\mathbf{X}_1) + \boldsymbol{\varepsilon} \tag{22}$$

$$\mathscr{F}(\mathbf{X}_1) = f_1(\boldsymbol{x}_1, \beta_1) + f_2(\boldsymbol{x}_2, \beta_2) \tag{23}$$

where $\boldsymbol{\varepsilon} = \mathscr{N}(0,\sigma^2), \mathbf{X}_1 = [\boldsymbol{x}_1, \boldsymbol{x}_2], \boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are $N \times 1$ vectors with uniformly distributed elements between zero and one, and β_1 and β_2 are the corresponding scalar coefficients. $\mathbf{W}$ represents a row and column standardized spatial weight matrix (see Pace and LeSage, 2002). That is, each row and column of $\mathbf{W}$ sums up exactly to unity. The spatial weight matrix is constructed based on a randomly generated spatial pattern, where the locations of the observations in two-dimensional space were randomly generated from a uniform distribution. The concept of neighborhood is based on k-nearest neighbors[5], where the nearest neighbors are determined via the Euclidean distance between the randomly generated coordinates.

We assess the performance of the proposed semi-parametric estimation method with regard to sample size, signal to noise ratio and strength of spatial dependencies. More specifically, we let $N = \{100,350,700\}$. and $\rho = \{0.1,0.5,0.8\}$. We do not set σ^2 directly to different levels, as the total model variance would also be dependent on the spatial structure and the autoregressive parameter. Instead, we directly set the signal to noise ratio SNR, which is measured by:

$$SNR = \frac{Var(\hat{\boldsymbol{y}})}{Var(\boldsymbol{y})} = \frac{Var[(\mathbf{I}_N - \rho \mathbf{W})^{-1}(\mathscr{F}(\mathbf{X}_1))]}{Var(\boldsymbol{y})} \tag{24}$$

where $Var(\cdot)$ denotes the variance of a vector. We set σ^2 in such a fashion that the desired number of $SNR = \{0.1,0.5,0.8\}$ is obtained via Eq. (24).

In both of the presented Monte Carlo studies, we explore different configurations of the non-linear functions $f_1(\cdot)$ and $f_2(\cdot)$. Table 1 presents an overview of the Monte Carlo studies and the non-linear functions used as benchmark data generating processes. In both studies two combinations of models are compared: (i) the classic SAR model [see Eq. (1)], and its counterpart with adaptive spline knot selection via genetic algorithms (ii). The classic SAR model in this Monte Carlo study does not take the non-linearity in the parameters of $\mathscr{F}(\mathbf{X}_1)$ into account and instead models $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ as linear in the parameters. The semi-parametric SAR model with adaptive spline knot selection corresponds to the model presented in Eq. (15) and utilises the algorithm described in Section 4 to determine the optimal number and position of spline knots[6]. In context of this Monte Carlo study we set the number of genetic algorithms executed in a parallel fashion to twelve, each with a population size of 100 and $N_{P_{mate}} = 10$, and a mutation rate of 2%[7]. We execute the algorithm for up to a maximum of $T = 100{,}000$ cycles, where every 50 cycles the genetic algorithms exchange one randomly selected member of their population.

The first Monte Carlo study [denoted as (a) in Table 1] aims to quantify the bias associated with using a linear in the parameter model – such as the classic SAR model for modeling a non-linear in the parameters problem – and to assess what amount of this bias is corrected by using a semi-parametric estimation method. In this spirit the semi-parametric part of the data generating process contains both a quadratic and a square root term. These are often suggested non-linear transformation to explore in the freight generation literature (see, e.g. Novak et al., 2011), thus it is important to establish that the proposed semi-parametric estimation algorithm accurately assesses the impacts associated with such non-linearities in the parameters.

The second Monte Carlo study [denoted as (b) in Table 1] investigates the performance of the semi-parametric SAR model in the absence of non-linearities in the parameters. For this purpose, we compare a classic SAR model [see Eq. (1)] to its semi-parametric counterpart [see Eq. (13)] using the adaptive spline knot selection algorithm in Section 4. The data generating process used for

[5] We use $k = 7$ neighbors for the Monte Carlo studies presented in this paper.

[6] For the fixed spline knot solution included in the initial population, we use eleven equidistant spline knots (with $\boldsymbol{\kappa}_1 = \boldsymbol{\kappa}_2 = [0.0,0.1,0.2,...,0.9,1.0]$) to model $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, respectively.

[7] We explore the impact of varying the number of generic algorithms in Table 7 in the Appendix.

Table 1
Specification of $f_1(\cdot)$ and $f_2(\cdot)$ in both Monte Carlo studies.

Monte Carlo	$f_1(\boldsymbol{x}_1)$	$f_2(\boldsymbol{x}_2)$
(a)	$2\boldsymbol{x}_1^2$	$1.2\sqrt{\boldsymbol{x}_2 + 1}$
(b)	$2\boldsymbol{x}_1$	$1.2\boldsymbol{x}_2$

benchmarking both models contains no non-linearities in the parameters.

The results of the first Monte Carlo [denoted as (a) in Table 1] study are presented in Table 2. Each row of the table corresponds to 300 Monte Carlo runs. The values in the first four columns characterize each set of Monte Carlo runs, in terms of signal to noise ratio, spatial autoregressive parameter ρ, the sample size and the implied variance parameter σ^2. Columns (i) to (iv) present the mean bias for σ^2,ρ and the total effects obtained when estimating the data generating process with the linear in the parameters SAR model. Columns (v) to (viii) contain the same information for the semi-parametric SAR model with adaptive spline knots.

Turning our attention to the results, we can observe that in the group of Monte Carlo studies where the signal to noise ratio is low ($SNR = 0.1$), no clear best model emerges. In terms of σ^2, in all but two cases (where the adaptive spline knot model performs the best) with $SNR = 0.1$ the classic SAR model outperforms its semi-parametric counterparts. In terms of absolute bias in ρ, the semi-parametric models show less absolute bias than the classic SAR in cases of $\rho = 0$ and $SNR = 0.1$, while in cases of $\rho = 0.8$ and

Table 2
Monte Carlo study comparing the performance of the classic SAR model to its semi-parametric counterparts, with and without adaptive spline knots, under the presence of moderate non-linearity in the parameters.

SNR	ρ	Sample size	σ^2	SAR				Semi-parametric SAR			
				Mean bias σ^2	Mean bias ρ	Mean bias $\partial\boldsymbol{y}/\partial\boldsymbol{x}_1$	Mean bias $\partial\boldsymbol{y}/\partial\boldsymbol{x}_2$	Mean bias σ^2	Mean bias ρ	Mean bias $\partial\boldsymbol{y}/\partial\boldsymbol{x}_1$	Mean bias $\partial\boldsymbol{y}/\partial\boldsymbol{x}_2$
				(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)
0.1	0.0	100	1.54	0.0253	−0.1110	−0.5856	−0.4309	0.0443	0.1028	−0.0695	−0.0680
		350	1.83	−0.0134	−0.1151	−0.5788	−0.5245	0.0050	0.0274	0.0278	−0.0242
		700	1.83	0.0110	−0.1265	−0.5117	−0.5064	0.0278	0.0026	0.0048	−0.0045
	0.5	100	1.97	0.0876	−0.1894	−0.0785	−0.1123	0.0902	−0.1723	−0.3837	0.7221
		350	1.86	0.0630	−0.2327	−0.0377	0.1098	0.0646	−0.2355	−0.1191	0.3378
		700	1.75	0.0413	−0.2331	−0.0363	0.0742	0.0437	−0.2329	−0.1371	0.0033
	0.8	100	1.73	0.1028	−0.1191	−1.5414	1.2652	0.1014	−0.1440	−0.1552	0.1603
		350	2.23	0.1259	−0.1599	−4.9549	4.3297	0.1322	−0.1822	−0.4223	0.5296
		700	2.35	0.1250	−0.1584	−0.5791	1.7182	0.1330	−0.1793	−1.2797	2.0644
0.5	0.0	100	0.63	−0.0333	−0.3341	−0.1467	−0.0493	0.0084	0.0245	−0.0132	−0.0162
		350	0.61	−0.0412	−0.3504	−0.0742	−0.1031	0.0031	0.0001	0.0015	−0.0005
		700	0.59	−0.0431	−0.3853	−0.1117	−0.0923	0.0022	−0.0040	−0.0017	0.0035
	0.5	100	0.61	−0.0274	−0.1855	−0.0207	−0.0279	0.0085	−0.0237	−0.0083	0.0378
		350	0.60	−0.0300	−0.2139	−0.1632	−0.0034	0.0032	−0.0342	−0.0075	0.0017
		700	0.57	−0.0315	−0.1994	−0.1694	−0.0819	0.0021	−0.0308	−0.0087	0.0422
	0.8	100	1.00	0.0272	−0.0878	−0.2291	0.0589	0.0386	−0.0856	−0.1200	0.1746
		350	0.62	−0.0235	−0.0876	−0.2512	−0.0042	0.0024	−0.0180	−0.0327	0.0726
		700	0.62	−0.0236	−0.0884	−0.2466	−0.0188	0.0024	−0.0243	−0.0304	0.0905
0.8	0.0	100	0.31	−0.1005	−0.3615	−0.1903	−0.1163	0.0063	0.0040	0.0039	0.0022
		350	0.32	−0.1011	−0.3733	−0.1400	−0.0700	0.0022	−0.0004	0.0021	0.0033
		700	0.31	−0.1100	−0.3731	−0.0725	−0.0475	0.0011	0.0008	0.0006	−0.0001
	0.5	100	0.30	−0.0888	−0.1774	−0.2410	−0.1900	0.0091	−0.0028	0.0026	0.0076
		350	0.30	−0.0915	−0.1995	−0.1423	−0.0791	0.0019	−0.0021	−0.0001	0.0036
		700	0.31	−0.0914	−0.2001	−0.1161	−0.0348	0.0010	−0.0014	−0.0006	0.0040
	0.8	100	0.30	−0.0971	−0.0890	−0.4791	−0.0145	0.0047	0.0033	0.0006	0.0111
		350	0.31	−0.0849	−0.0831	−0.2491	−0.0799	0.0029	0.0009	0.0003	0.0062
		700	0.30	−0.0883	−0.0841	−0.2718	−0.0427	0.0004	−0.0009	−0.0003	0.0061

Notes: $\mathscr{F}(\mathbf{X}_1) = 2\boldsymbol{x}_1^2 + 1.2\sqrt{\boldsymbol{x}_2 + 1}$. *SNR* stands for signal to noise ratio; the reported partial derivatives correspond to the *total effects* summary impact measures from LeSage and Pace (2009). Each row reports average values over 300 Monte Carlo runs. We used twelve parallel genetic algorithms, which exchange one member of their population every 50 cycles. The adaptive spline knot algorithm ran for a maximum of $T = 100{,}000$ cycles, with a population size of $N_P = 100$ and $N_{P_{mate}} = 10$, and mutation probability of 2%.

Table 3
Monte Carlo study comparing the performance of the classic SAR model to its semi-parametric counterpart with adaptive spline knots for a DGP without non-linearity in the parameters.

SNR	ρ	Sample size	σ^2	SAR				Semi-parametric SAR			
				Mean bias σ^2 (i)	Mean bias ρ (ii)	Mean bias $\partial \boldsymbol{y}/\partial \boldsymbol{x}_1$ (iii)	Mean bias $\partial \boldsymbol{y}/\partial \boldsymbol{x}_2$ (iv)	Mean bias σ^2 (v)	Mean bias ρ (vi)	Mean bias $\partial \boldsymbol{y}/\partial \boldsymbol{x}_1$ (vii)	Mean bias $\partial \boldsymbol{y}/\partial \boldsymbol{x}_2$ (viii)
0.1	0.0	100	1.98	0.0924	0.0707	−0.1296	−0.0635	0.4879	0.1188	0.2186	0.1075
		350	1.89	0.0271	0.0281	−0.0496	−0.0577	0.1264	0.0458	−0.1262	−0.0424
		700	2.13	0.0158	0.0141	−0.0144	−0.0384	0.0772	0.0147	−0.0083	0.0025
	0.5	100	1.90	0.0932	−0.1540	0.3198	0.7685	0.4614	−0.1863	1.7962	1.4238
		350	2.16	0.0562	−0.2149	0.5432	0.5538	0.1931	−0.2471	0.0235	0.0963
		700	1.89	0.0334	−0.1939	0.4517	0.5145	0.0886	−0.2478	−0.2656	−0.0074
	0.8	100	2.41	0.1828	−0.1334	−0.5227	1.6868	0.7959	−0.1452	0.9145	−1.5881
		350	1.85	0.0637	−0.1223	0.6437	1.0073	0.1788	−0.1816	−0.3686	−0.1889
		700	2.21	0.0971	−0.1484	0.7928	1.1767	0.1893	−0.1861	−0.7829	−1.3713
0.5	0.0	100	0.59	0.0136	0.0045	−0.0022	−0.0035	0.0449	0.0532	0.0117	−0.0144
		350	0.66	0.0027	0.0043	0.0008	−0.0076	0.0145	0.0201	0.0026	0.0173
		700	0.68	0.0026	0.0026	−0.0022	−0.0019	0.0087	0.0042	0.0112	0.0149
	0.5	100	0.69	0.0142	−0.0004	0.0087	0.0212	0.0643	−0.0653	−0.1241	−0.0397
		350	0.64	0.0020	−0.0115	0.0333	0.0305	0.0147	−0.1013	−0.0370	0.0015
		700	0.68	0.0010	−0.0097	0.0257	0.0312	0.0080	−0.0760	−0.0528	0.0086
	0.8	100	0.69	0.0118	−0.0012	0.0233	0.0340	0.0632	−0.0672	−3.3999	−0.7793
		350	0.79	0.0032	−0.0084	0.0439	0.0698	0.0263	−0.0800	−0.2824	−0.0277
		700	0.60	0.0021	−0.0037	0.0258	0.0259	0.0087	−0.0502	−0.1096	−0.0871
0.8	0.0	100	0.32	0.0078	0.0043	0.0044	0.0077	0.0179	−0.0002	0.0068	−0.0045
		350	0.33	0.0011	0.0023	0.0010	0.0016	0.0045	0.0058	0.0031	0.0050
		700	0.34	0.0012	0.0002	0.0004	−0.0006	0.0028	0.0000	−0.0027	0.0072
	0.5	100	0.36	0.0028	0.0026	−0.0018	0.0086	0.0198	0.0084	0.0155	−0.0025
		350	0.35	0.0022	0.0008	0.0028	0.0025	0.0061	−0.0054	−0.0016	−0.0100
		700	0.33	0.0004	−0.0003	0.0025	0.0014	0.0023	−0.0061	−0.0028	−0.0032
	0.8	100	0.30	0.0028	0.0062	0.0060	0.0052	0.0137	−0.0002	0.0050	0.0097
		350	0.33	0.0015	0.0007	0.0038	0.0035	0.0042	−0.0042	−0.0026	−0.0034
		700	0.34	0.0004	0.0006	0.0025	0.0047	0.0027	−0.0042	−0.0070	−0.0099

Notes: $\mathscr{F}(\mathbf{X}_1) = 2\boldsymbol{x}_1 + 1.2\boldsymbol{x}_2$. *SNR* stands for signal to noise ratio; the reported partial derivatives correspond to the *total effects* summary impact measures from LeSage and Pace (2009). Each row reports average values over 300 Monte Carlo runs. We used twelve parallel genetic algorithms, which exchange one member of their population every 50 cycles. The adaptive spline knot algorithm ran for a maximum of $T = 100{,}000$ cycles, with a population size of $N_P = 100$ and $N_{P_{mate}} = 10$, and mutation probability of 2%.

$SNR = 0.1$, the classic SAR exhibits the least absolute bias in the spatial autoregressive parameter. However, in the test cases with $SNR = 0.5$ and $SNR = 0.8$, the semi-parametric model with adaptive spline knots exhibits the least absolute bias in σ^2 in all but one case ($SNR = 0.5, \rho = 0.8$, sample size 100). Moreover, the semi-parametric model with adaptive spline knots outperforms the classic SAR in terms of absolute bias in ρ for all test cases with $SNR = 0.5$ and $SNR = 0.8$. The semi-parametric models outperform the classic SAR in terms of absolute bias in total effects of $\boldsymbol{x}_1$ in all Monte Carlo cases with $SNR = 0.5$ and $SNR = 0.8$. In the case of absolute bias in total effects of $\boldsymbol{x}_2$, the semi-parametric models only consistently outperform the classic SAR model in cases of $SNR = 0.8$. In the case if $SNR = 0.5$, the semi-parametric model performs better with lower ρ values, while with $\rho = 0.8$ and $SNR = 0.5$, the classic SAR exhibits less overall bias in the total effects of $\boldsymbol{x}_2$.

Table 3 presents the results of the second Monte Carlo study [denoted as (b) in Table 1]. Each row of the table represents the average over 300 Monte Carlo runs. The first four columns contain the signal to noise ratio, the spatial autoregressive parameter, the sample size and the implied variance, that characterizes each set of Monte Carlo runs. Columns (i) to (iv) and columns (v) to (viii) show the bias in terms of σ^2, ρ and total impact effects of both variables, for the SAR and the semi-parametric SAR model, respectively.

These results indicate that overall in 20% of cases the classic SAR model outperforms the semi-parametric SAR in terms of absolute bias. In the case of absolute bias in σ^2, the classic SAR outperforms its semi-parametric counterpart in all cases. The coefficient estimates for ρ exhibit in all but two (with $SNR = 0.8, \rho = 0.0$, sample size 100, and $SNR = 0.8, \rho = 0.5$ and sample size 100) of the Monte Carlo cases less absolute bias in the classic SAR model.

In terms of absolute bias in the total impact estimates of the coefficients the semi-parametric SAR model outperforms its linear in the parameters counterpart in ~ 63% of Monte Carlo test cases with a low signal to noise ratio ($SNR = 0.1$). In the test cases with

$SNR = 0.5$ and $SNR = 0.8$ the semi-parametric model performs worse in terms of absolute total impact estimate bias, only outperforming the classic SAR model in ∼ 16% of Monte Carlo test cases. It should be noted that while the classic SAR model clearly outperforms its semi-parametric counterpart, the magnitude of the differences in absolute biases strongly decreases in the Monte Carlo cases with signal to noise ratios $SNR = 0.5$ and $SNR = 0.8$ and sample size $n = 700$. This implies that if no non-linearities are present in the data generating process, the linear in the parameters SAR model exhibits less absolute bias overall. However, while it generally remains larger, the relative difference in absolute bias of the classic SAR and the semi-parametric model decreases as the *SNR* and sample size increases.

The results of both Monte Carlo studies suggest that the proposed semi-parametric adaptive spline knot approach can accurately estimate, not only the functional form of $f(\cdot)$, but also its derivative. With increasing sample size, the accuracy of the estimates increases as well, but note that the mean bias of σ^2 does not exhibit drastic changes. This indicates that despite the increased sample size, the algorithm does not tend to overfit the model. The estimation procedure seems to be robust regarding different levels of ρ, though it should be noted, that with higher ρ and lower signal-to-noise ratio, the classic SAR performs marginally better. This seems logical, since in very noisy dependent variables a simpler model is expected to perform better. Finally, we have demonstrated that the estimation approach performs on a comparable level to the classic SAR, when confronted with a data generating process without any non-linearities in the parameters.

6. European freight generation

We assess the performance of the proposed semi-parametric SAR estimation method on a real world example. For this purpose, our dependent variable is the volume of *road freight generated* by European NUTS-2 regions in 2011. We consider the total volume of road freight generated (across all sectors). The freight generation dataset was obtained from *Eurostat*. We measure the volume of road freight in million tons. The dataset covers 258 European NUTS-2 regions (2006 version of regions), which include all EU-27 countries, with the exception of Cyprus and Malta. Furthermore, all island NUTS-2 regions were excluded from the study, since they usually do not report road freight generation. For a complete list of the NUTS-2 regions included in this study, refer to Table 6 in the Appendix. Fig. 4 in the Appendix displays a map of NUTS-2 level freight generation [panel (i)].

Our explanatory variables were selected based on Novak et al. (2011). We explain the volume of regional freight generated in 2011, by a set of explanatory variables measured in 2010. Following Novak et al. (2011) we use the regional share of employment and the regional share of employment in agriculture (NACE rev 2 A to B) and manufacturing (NACE rev. 2 C to I) as an indicator of a region's sectoral configuration. All of these explanatory variables stem from *Eurostat*. Further explanatory variables are the length of the road network (measured in 10,000 km) and the distance to the closest seaport (measured in travel time in minutes). The latter two variables are available from *ESPON*.

For the construction of the spatial weight matrix, the geodesic distance between regions' centroids was used. The results are based on a k-nearest spatial weight matrix configuration, with $k = 7$. [8]

6.1. Aggregate freight generation

Table 4 presents the results of the estimation using the classic SAR [panel (i)] and the semi-parametric SAR with adaptive spline knots [panel (ii)]. For the initial population of the semi-parametric SAR model, a SAR model with ten equidistant spline knots per variable was selected. The semi-parametric SAR adaptive spline knot algorithm was evaluated with $T = 100{,}000$ cycles. We used twelve parallel genetic algorithms, which exchange one member of their population every 50 cycles. The population and mating pool lengths were set to $N_P = 100$ and $N_{P_{mate}} = 10$, while mutation probability was 2%.

For the purposes of reporting the results, we make use of the summary impact measures suggested by LeSage and Pace (2009). Direct effects are the average impact of an explanatory variable on a region's generated freight, without affecting its neighbors. Indirect effects summarize the average impact of an explanatory variable on only a region's neighbor. Finally, total effects are the average impacts of an explanatory variable, which both affect the region itself and its neighbors. For the semi-parametric SAR model, the summary impact measures were calculated using the first order derivatives of the splines. The convergence of the semi-parametric SAR algorithm – in terms of its best, median and average fitness score – is illustrated in Fig. 5 in the Appendix.

The first point to note is that the semi-parametric SAR model performs better when the two models are compared using their respective AIC_c. A likelihood ratio test confirms these results, with a value of −635.684 and $p < 0.001$ in favor of the semi-parametric SAR model. The same is true if we turn our attention to the coefficient of determination corrected for sample size: the $\overline{R}^2$. The classic SAR exhibits an $\overline{R}^2 = 0.71$, while the semi-parametric SAR has an $\overline{R}^2 = 0.87$. [9]

Second, the comparatively better performance – in terms of AIC_c and $\overline{R}^2$ – of the semi-parametric SAR model, might raise concerns that the model is in fact over-fitting. To rule out this problem, we compare the predictive performance of both models by projecting them one year ahead. For this purpose, we collected a corresponding matrix of explanatory variables from 2011 (the sources of our data are the same as in the original data set). Then, using the estimated co-efficients and the 2011 explanatory

[8] Configurations with $k = [3,5,9]$ were tested as well, however the estimation results did not change in a significant manner.

[9] In fact the high $\overline{R}^2 = 0.87$ might raise concerns of multi-collinearity bing present in the model. In order to rules this out, we have performed tests for multicollinearity for the full explanatory variable matrices of both models in Table 4. Both the eigenvalues based condition index, as well as the χ^2 test proposed by Farrar-Glauber indicates no multicollinearity for the models, with values of $CI = 1.839$ ($CI = 2.701$ for the semi-parametric SAR case), and $\chi^2 = 1.353$ ($\chi^2 = 1.957$ for the semi-parametric SAR case), respectively.

Table 4
Summary impact measures of the SAR (i) and semi-parametric SAR with adaptive spline knots (ii) freight generation models.

Variable	Direct	Indirect	Total
	(i) – Spatial autoregressive model		
Total employment (in %)	0.577^{***}	0.933^{***}	1.510^{***}
Agricultural employment (in %)	-1.076^{***}	-1.755^{**}	-2.830^{**}
Manufacturing employment (in %)	-1.821	-3.049	-4.870
Length of road network (10,000 km)	0.742^{***}	1.221^{***}	1.963^{***}
Travel time to seaport (min.)	0.082^{**}	0.135^{*}	0.216^{**}
ρ	0.654^{***}		
σ^2	0.155		
Log-likelihood	-35.181		
AIC_c	80.600		
RMSE	0.389		
$\overline{R}^2$	0.71		
Number of observations	258		
Number of parameters	8		
	(ii) Semi-parametric spatial autoregressive model		
Total employment (in %)	-0.096	-0.039	-0.135
Agricultural employment (in %)	-1.642^{**}	-0.535^{*}	-2.177^{**}
Manufacturing employment (in %)	-1.408	-0.461	-1.869
Length of road network (10,000 km)	1.101^{***}	0.364^{*}	1.465^{***}
Travel time to seaport (min.)	0.087	0.029	0.116
ρ	0.277^{***}		
σ^2	0.140		
Log-likelihood	-153.023		
AIC_c	8.695		
RMSE	0.328		
$\overline{R}^2$	0.87		
Number of observations	258		
Number of parameters	24		

Notes: *** and ** denote statistical significance at the one percent level and five percent level, respectively (significance levels calculated using the algorithm proposed by LeSage and Pace (2009)). RMSE denotes the root mean squared error, when projecting the model to 2012 using explanatory variables from 2011. The semi-parametric SAR reached convergence after 23,291 cycles (no AIC_C improvement). A spatial neighborhood matrix configuration with seven nearest neighbors was used.

variables, we project the model and compare it with observed freight generation from 2012. The accuracy of this projection is measured by the root mean squared error (RMSE), in Table 4. If the semi-parametric SAR model indeed over-fits the data, we would expect to observe a lower predictive performance. This seems to be not the case, with the classic SAR having an $RMSE = 0.389$ and the semi-parametric SAR having an $RMSE = 0.328$.

Third, note the differing spatial autocorrelation parameter between the SAR ($\rho = 0.654$) and the semi-parametric SAR ($\rho = 0.277$) estimates. This – in conjunction with the log-likelihood test and the $\overline{R}^2$ – implies that the SAR model might contain a bias in the spatial estimates. Nonetheless, both models indicate a highly significant degree of spatial autocorrelation, thus confirming our hypothesis that the freight generation of one region depends on the freight generation of its neighbors. This is in line with the findings of Novak et al. (2011) and Krisztin (2017).

Turning our attention to the linear in the parameter SAR model impact estimates [panel (i)]: *total employment* is significant and positive both for the own endowment of regions, as well as for neighboring regions. This is inline with the results from Novak et al. (2011). Similarly, the negative impact of *agricultural employment*, both directly and for neighbors, is similar to the results of Novak et al. (2011). Also in line with the study by Novak et al. (2011) is the fact that the impact of manufacturing employment is negative. However, as opposed to the results presented in Novak et al. (2011), the impacts are not significant. The significantly positive influence of infrastructural variables, such as *road network* is well documented in the literature (Novak et al., 2011; Chun et al., 2012; Lawson et al., 2012). The *travel time to seaport* is not significant for neighboring regions' freight generation, only for a region's own freight generation, where it plays a comparatively minor role (0.082).

The semi-parametric SAR results in panel (ii) of Table 4 show that only *agricultural employment* seems to play a significant ($p < 0.001$) and negative role in directly affecting a region's freight generation. A ceteris paribus increase of 1% of agricultural employment in a region would decrease the region's own freight generation by 1.6 million tons. Note, that the indirect effects from *agricultural employment* are only weakly significant, thus an increase or decrease in a region's agricultural employment does not necessarily lead to a significant change in neighboring regions' freight generation. Moreover, *Manufacturing employment* does not exhibit a significant influence. This is in contrast to the classic SAR model, where manufacturing employment was weakly significant. The *length of the road network*, which proxies the transportation infrastructure in the model, exhibits a positive effect, both directly and indirectly on a region's neighbors. This provides evidence for the fact that increasing a region's transportation infrastructure does not only benefit the region itself, but also its neighbors. Increasing the length of the road network by 10,000 km, would indicate an

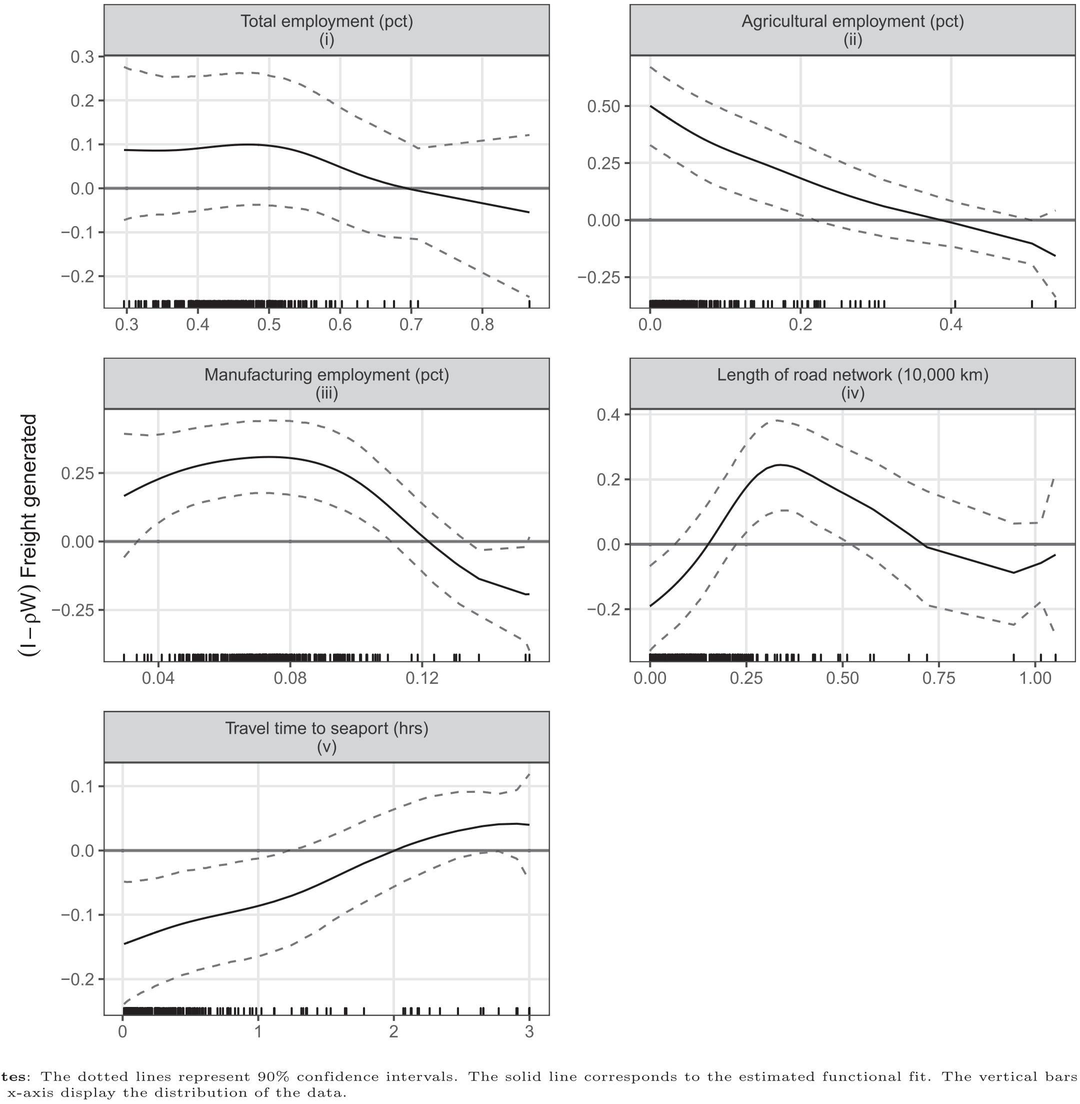


Fig. 2. Functional fit of the semi-parametric SAR model terms. **Notes**: The dotted lines represent 90% confidence intervals. The solid line corresponds to the estimated functional fit. The vertical bars along the x-axis display the distribution of the data.

increase of 1.231 million tons in a region's freight generation, while its neighbors would generate 0.436 million tons more freight. Note, that in contrast to the classic SAR model, the travel time to seaports in minutes is not significant in the semi-parametric model.

While Table 4 displays aggregate impact measures, the semi-parametric spline based representation also allows us to plot the functional form of the variable over its range. This visualizes the underlying non-linearities and allows us to interpret possible non-linear influences. Fig. 2 shows the estimated functional form for the explanatory variables. The vertical bars along the *x*-axis display the distribution of the data. Note, that the significance measures in Table 4, panel (ii) assess whether a summary impact measure is significantly different from zero over its full functional form. In order to assess whether non-linear impacts of a variable are significant over specific intervals only, we need to estimate error bounds for the estimated functional forms. These error bounds (confidence intervals) were obtained using bootstrapping. For this purpose, we iteratively estimate the model using the adaptive spline knot procedure and resample the full set of observations from the model residuals. We used 1000 iterations for the functional fits.

Turning our attention to the functional forms in Fig. 2, we can observe that in the case of *total employment* (i), *agricultural employment* (ii), and *length of road network* (iv) the support for spline knots in higher ranges of the covariates is sparse. The explanatory variables *manufacturing employment* (iii) and *length of road network* (iv) exhibit clearly significant non-linear functional forms. For manufacturing employment, the impact is initially increases until around 7% of employment, after which it starts to decrease and the non-linear impact becomes insignificant at around 11%. In the case of road network, the functional fit would indicate that up to about 31 million kilometres the impact of road network is positive, but after this it seems to decrease and becomes insignificant at around 56 million kilometres, after which the frequency of observations on the covariate decreases very strongly. In the case of agricultural employment and travel times to seaport, the slope of the coefficient is relatively linear and not significant.

Table 5
Summary impact measures of the semi-parametric SAR freight generation model for the mining (i), food (ii), and agricultural (iii) sectors.

Variable	Direct	Indirect	Total
	(i) – Mining		
Total employment (in %)	−0.075	−0.040	−0.115
Agricultural employment (in %)	−0.411	−0.204	−0.615
Manufacturing employment (in %)	−0.277	−0.140	−0.417
Length of road network (10,000 km)	0.481^{***}	0.238^{**}	0.718^{***}
Travel time to seaport (min.)	0.062	0.032	0.094
ρ	0.336^{***}		
σ^2	0.033		
Log-likelihood	42.498		
AIC_c	−82.940		
$\bar{R}^2$	0.73		
Number of parameters	16		
	(ii) – Food		
Total employment (in %)	0.027	0.007	0.034
Agricultural employment (in %)	-0.168^{**}	−0.050	-0.218^{**}
Manufacturing employment (in %)	−0.015	−0.002	−0.017
Length of road network (10,000 km)	0.098^{***}	0.030^{*}	0.128^{***}
Travel time to seaport (min.)	0.005	0.002	0.007
ρ	0.231^{***}		
σ^2	0.080		
Log-likelihood	385.433		
AIC_c	−768.466		
$\bar{R}^2$	0.84		
Number of parameters	18		
	(ii) - Agriculture		
Total employment (in %)	−0.045	−0.026	−0.072
Agricultural employment (in %)	−0.068	−0.038	−0.106
Manufacturing employment (in %)	0.029	0.019	0.048
Length of road network (10,000 km)	0.079^{***}	0.045^{**}	0.124^{***}
Travel time to seaport (min.)	0.018	0.010	0.028
ρ	0.371^{***}		
σ^2	0.115		
Log-likelihood	420.232		
AIC_c	−810.628		
$\bar{R}^2$	0.78		
Number of parameters	14		

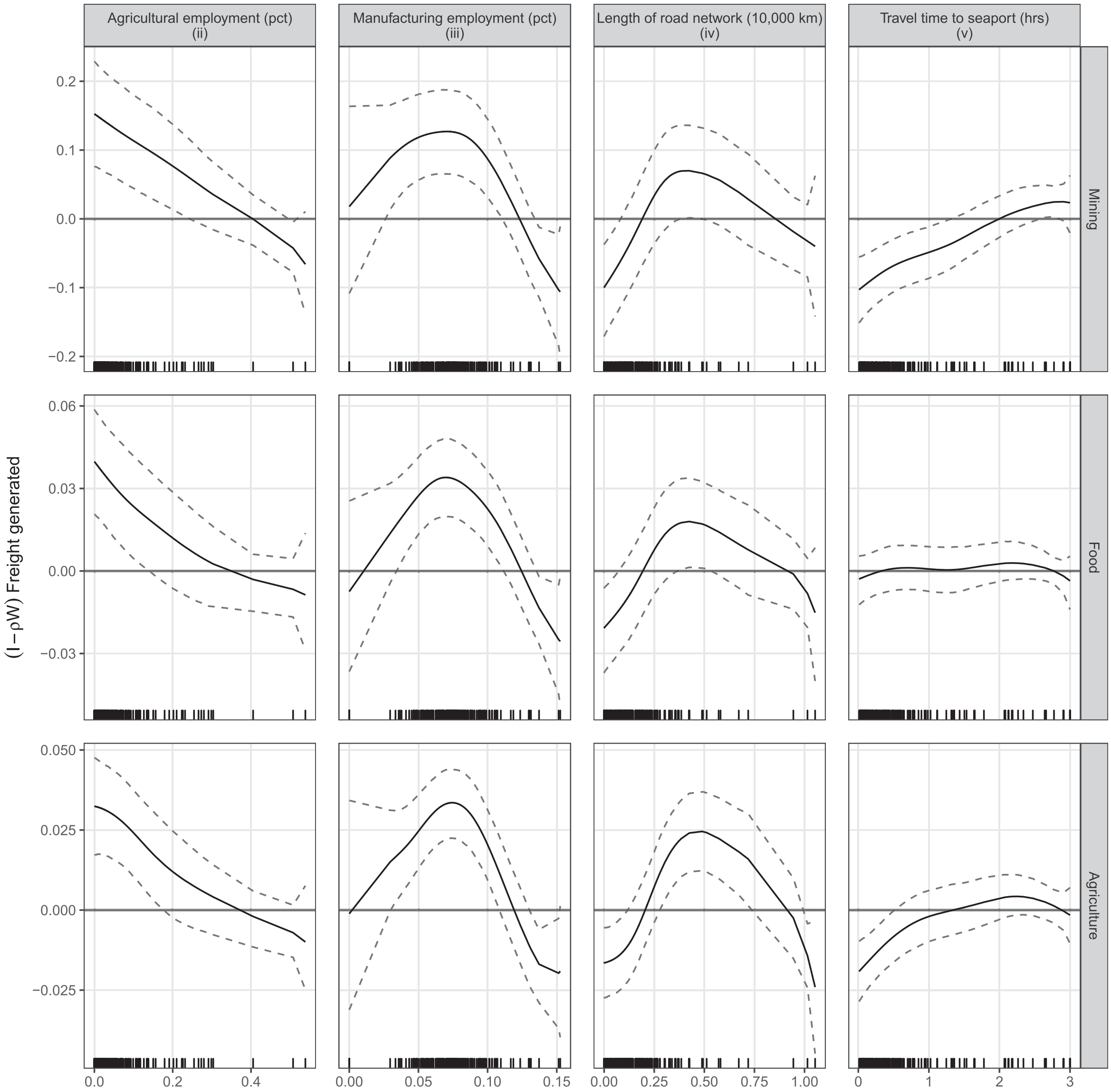
Notes: *** and ** denote statistical significance at the one percent level and five percent level, respectively (significance levels calculated using the algorithm proposed by (LeSage and Pace, 2009)). The semi-parametric SAR reached convergence (no significant AIC_C improvement) after 24,932,47,102, and 62,654 cycles. A spatial neighborhood matrix configuration with seven nearest neighbors was used.

6.2. *Sectoral freight generation*

While the previous results highlight potential non-linearities in aggregate freight generation, the question remains open, whether this pattern is also prevalent across different sectors of freight generation. Therefore, in a second step of our analysis, we consider the case of sectoral freight generation, in contrast to aggregate freight generation. More specifically, we consider three sectors of freight generation, according to the European Commission's NST/R (NST/R - *Nomenclature uniforme des marchandises pour les Statistiques de Transport, Révisée*) classification: mining and minerals (NST/R codes *GT03* and *GT09*), food products (NST/R code *GT04*), and agricultural products (NST/R code *GT04*). The choice of these sectors is motivated by the fact that they generated the total largest volume of freight in Europe in 2011. Fig. 4 in the Appendix display maps of the volume of sectoral freight generated on a NUTS-2 level [panels (ii) - (iv)]. The same set of explanatory variables is used, as in the aggregate freight generation case. For the purpose of inference, a separate semi-parametric SAR model was run for each sectoral dataset. The algorithm settings for each sectoral model exactly correspond to the aggregate freight generation model from the previous subsection.

Table 5 presents summary direct, indirect, and total impacts, as well as summary statistics for the mining (i), food (ii), and agricultural sector (iii). First, note that the signal to noise ratio ($\bar{R}^2$) of the sectoral models is comparable to the aggregate semi-parametric freight generation model, with the model of the food sector having the highest $\bar{R}^2$ of 0.84. [10] Second, in all three sectoral models the spatial autoregressive parameter ρ is positive and highly significant. This confirms the evidence from the aggregate model

[10] Note, that the log-likelihood and AIC_c scores are not directly comparable between models, as the independent variable vector $\boldsymbol{y}$ differs per sector.



Notes: The dotted lines represent 90% confidence intervals. The solid line corresponds to the estimated functional fit. The vertical bars along the x-axis display the distribution of the data.

Fig. 3. Functional fit of selected semi-parametric SAR model terms for the mining, food, and agricultural sectors. **Notes**: The dotted lines represent 90% confidence intervals. The solid line corresponds to the estimated functional fit. The vertical bars along the x-axis display the distribution of the data.

that spatial dependence plays an important role in freight generation.

The signs of the summary impact estimates provide for all sectors a similar pattern, as in the aggregate freight generation case. Note, however, that the agricultural employment is only significant for the direct impacts in the food sector [panel (ii) in Table 5]. That is, a ceteris paribus one percent change in the agricultural employment of a region would lead to a decrease of 0.168 million tons of freight originating from that region, but would not significantly impact the region's neighbors. For the freight generation in the mining and agricultural sectors, agricultural employment is insignificant. The length of the road networks – our proxy for regional road infrastructure – is positive and significant for all three models. It plays, however, a smaller role in the food (0.128 total impact) and agricultural sectors (0.124 total impact), as in the case of the aggregate freight generation over all sectors (1.565 total impact). In the case of the mining sector, however, an increase of 10,000 kilometres in the road network would lead to a corresponding increase of

0.481 million tons of mining and mineral goods in the region itself and to an increase of 0.238 million tons in its neighbors. This reflects the fact that infrastructure is of particular interest to the mining industry.

The number of parameters per sectoral model provides an indication of the underlying non-linearities. Note, that the sectoral models converge on a lower number of parameters (14 to 18 parameters, respectively) as the aggregate freight generation (24 parameters). This indicates that the underlying curves might be flatter, and is also reflected in the lower number of significant summary impact estimates.

Fig. 3 displays the functional fit of the B-spline curves per sector (rows) and explanatory variable (columns). The variable *total employment* was not depicted, as its functional form is flat in all case and never significantly different from zero. The overall functional forms of the sectoral B-splines is similar to the aggregate freight generation case in Fig. 2, where the explanatory variables *manufacturing employment* (iii) and *length of road network* (iv) exhibit clearly significant non-linear functional forms. In case of the manufacturing employment, both the mining sector and the agricultural sector exhibit clearly non-linear patterns.

7. Closing remarks

This paper addresses two key weaknesses in freight generation modeling. First, the classic regional freight generation model assumes that observations are independent. This is not correct, since spatial dependencies have been shown to play a role in freight generation (Novak et al., 2011). However, the approach by Novak et al. (2011) only takes spatial dependence in the error terms into account. This paper introduces spatial lags of the dependent variable in the classic freight generation model.

The second weakness of freight generation modeling that is addressed in the paper is the presence of non-linearities in the parameters in freight generation models. Such non-linearities can arise in freight generation modeling, where the derivative impact of regional GDP or employment in manufacturing can first be positive, but this effect is hypothesized to decrease with higher values of the covariate. In this paper we put forth a novel estimation method for a SAR model with non-linearities in the parameters. Non-linearities in the parameters can be adequately captured through semi-parametric techniques, such as basic splines. Spline models rely on subdividing the impact space of the modeled variable based on a set of so-called support knots. Each potentially overlapping subdivision of the impact space is modeled in a linear in the parameters fashion. Obviously the ability of spline-based semi-parametric approaches to adequately capture non-linearities in the parameters is based on the number and relative position of support knots over the impact space of the explanatory variables.

Current approaches, such as that by Basile et al. (2014) rely on setting a fixed number and position of spline knots. Such an approach is inefficient, since it might well overparametrize the problem and might lead to overfitting. Instead we suggest an adaptive approach, based on a variant of genetic algorithms, which selects the optimal – in terms of AIC – number and position of spline knots per explanatory variable. Moreover, we show that this approach is compatible with the classic summary spatial impact estimates, proposed by LeSage and Pace (2009).

We investigate the performance of the proposed semi-parametric SAR estimation method with adaptive spline knots in a series of Monte Carlo studies. First, we show that the proposed method exhibits on average lower bias in the parameters than both a linear in parameters SAR model, and a semi-parametric SAR with fixed spline knots. Second, we demonstrate that in the absence of non-linearities in the parameters, the bias of the proposed semi-parametric estimation method is comparable to that of the classic linear in the parameters SAR model. This indicates that the estimation method is robust to overfitting in cases where there are no non-linearities in the parameters. Additionally, in the presence of high non-linearities, the performance of our proposed estimation method exhibits lower bias in the signal to noise ratio and spatial autoregressive parameter, as the semi-parametric SAR model with fixed spline knots.

Finally, we demonstrate the applicability of the approach to freight generation modeling, in an applied case study for European NUTS-2 level regions in 2010, based on the study by Novak et al. (2011). In this case study we demonstrate that (i) spatial dependence plays a key role in European freight generation modeling, (ii) not taking prevalent non-linearities in the parameters into account leads to biased estimates, even when controlling for spatial lags of the dependent variable. Moreover, (iii) we present evidence for significant non-linearities in freight generation stemming from the length of the road network and manufacturing employment.

Appendix A

(See Tables 6–8).
(See Figs. 4 and 5).

Table 6
List of countries and NUTS-2 regions.

Code	Region name	Code	Region name	Code	Region name	Code	Region name
	Austria		**Denmark**		**Italy**		**Sweden**
AT11	Burgenland (AT)	DK01	Hovedstaden	ITC1	Piemonte	SE11	Stockholm
AT12	Niedersterreich	DK02	Sjlland	ITC2	Valle d'Aosta	SE12	stra Mellansverige
AT13	Wien	DK03	Syddanmark	ITC3	Liguria	SE21	Smland med arna
AT21	Krnten	DK04	Midtjylland	ITC4	Lombardia	SE22	Sydsverige
AT22	Steiermark	DK05	Nordjylland	ITF1	Abruzzo	SE23	Vstsverige
AT31	Obersterreich		**Estonia**	ITF2	Molise	SE31	Norra Mellansverige
AT32	Salzburg	EE00	Eesti	ITF3	Campania	SE32	Mellersta Norrland
AT33	Tirol		**Greece**	ITF4	Puglia	SE33	vre Norrland
AT34	Vorarlberg	EL11	Anatoliki Makedonia, Thraki	ITF5	Basilicata		**Slovenia**
	Belgium	EL12	Kentriki Makedonia	ITF6	Calabria	SI01	Vzhodna Slovenija
BE10	Rgion de Bruxelles-Capitale	EL13	Dytiki Makedonia	ITG1	Sicilia	SI02	Zahodna Slovenija
BE21	Prov. Antwerpen	EL14	Thessalia	ITG2	Sardegna		**Slovakia**
BE22	Prov. Limburg (BE)	EL21	Ipeiros	ITH1	Provincia Autonoma di Bolzano	SK01	Bratislavsk kraj
BE23	Prov. Oost-Vlaanderen	EL22	Ionia Nisia	ITH2	Provincia Autonoma di Trento	SK02	Zpadn Slovensko
BE24	Prov. Vlaams-Brabant	EL23	Dytiki Ellada	ITH3	Veneto	SK03	Stredn Slovensko
BE25	Prov. West-Vlaanderen	EL24	Sterea Ellada	ITH4	Friuli-Venezia Giulia	SK04	Vchodn Slovensko
BE31	Prov. Brabant Wallon	EL25	Peloponnisos	ITH5	Emilia-Romagna		**United Kingdom**
BE32	Prov. Hainaut	EL30	Attiki	ITI1	Toscana	UKC1	Tees Valley, Durham
BE33	Prov. Lige	EL41	Voreio Aigaio	ITI2	Umbria	UKC2	Northumberland, Tyne and Wear
BE34	Prov. Luxembourg (BE)	EL42	Notio Aigaio	ITI3	Marche	UKD1	Cumbria
BE35	Prov. Namur	EL43	Kriti	ITI4	Lazio	UKD3	Greater Manchester
	Bulgaria		**Spain**		**Latvia**	UKD4	Lancashire
BG31	Severozapaden	ES11	Galicia	LT00	Lietuva	UKD6	Cheshire
BG32	Severen tsentralen	ES12	Principado de Asturias		**Luxembourg**	UKD7	Merseyside
BG33	Severoiztochen	ES13	Cantabria	LU00	Luxembourg	UKE1	East Yorkshire and
BG34	Yugoiztochen	ES21	Pas Vasco		**Lithuania**	Northern Lincolnshire	
BG41	Yugozapaden	ES22	Comunidad Foral de Navarra	LV00	Latvija	UKE2	North Yorkshire
BG42	Yuzhen tsentralen	ES23	La Rioja		**Netherlands**	UKE3	South Yorkshire
	Czech Republic	ES24	Aragn	NL11	Groningen	UKE4	West Yorkshire
CZ01	Praha	ES30	Comunidad de Madrid	NL12	Friesland (NL)	UKF1	Derbyshire, Nottinghamshire
CZ02	Stredn Cechy	ES41	Castilla y Len	NL13	Drenthe	UKF2	Leicestershire, Rutland and
CZ03	Jihozpad	ES42	Castilla-la Mancha	NL21	Overijssel	Northamptonshire	
CZ04	Severozpad	ES43	Extremadura	NL22	Gelderland	UKF3	Lincolnshire
CZ05	Severovchod	ES51	Catalua	NL23	Flevoland	UKG1	Herefordshire, Worcestershire and
CZ06	Jihovchod	ES52	Comunidad Valenciana	NL31	Utrecht	Warwickshire	
CZ07	Stredn Morava	ES53	Illes Balears	NL32	Noord-Holland	UKG2	Shropshire, Staffordshire
CZ08	Moravskoslezsko	ES61	Andaluca	NL33	Zuid-Holland	UKG3	West Midlands
	Germany		**Finland**	NL34	Zeeland	UKH1	East Anglia
DE11	Stuttgart	FI19	Lnsi-Suomi	NL41	Noord-Brabant	UKH2	Bedfordshire, Hertfordshire
DE12	Karlsruhe	FI1B	Helsinki-Uusimaa	NL42	Limburg (NL)	UKH3	Essex
DE13	Freiburg	FI1C	Etel-Suomi		**Poland**	UKI1	Inner London
DE14	Tbingen	FI1D	Pohjois- ja It-Suomi	PL11	Ldzkie	UKI2	Outer London
DE21	Oberbayern		**France**	PL12	Mazowieckie	UKJ1	Berkshire, Buckinghamshire and
DE22	Niederbayern	FR10	le de France	PL21	Malopolskie	Oxfordshire	
DE23	Oberpfalz	FR21	Champagne-Ardenne	PL22	Slaskie	UKJ2	Surrey, East, West Sussex
DE24	Oberfranken	FR22	Picardie	PL31	Lubelskie	UKJ3	Hampshire, Isle of Wight
DE25	Mittelfranken	FR23	Haute-Normandie	PL32	Podkarpackie	UKJ4	Kent
DE26	Unterfranken	FR24	Centre (FR)	PL33	Swietokrzyskie	UKK1	Gloucestershire, Wiltshire, Bristol

(*continued on next page*)

Table 6 (*continued*)

Code	Region name	Code	Region name	Code	Region name	Code	Region name
DE27	Schwaben	FR25	Basse-Normandie	PL34	Podlaskie	UKK2	Dorset, Somerset
DE30	Berlin	FR26	Bourgogne	PL41	Wielkopolskie	UKK3	Cornwall, Isles of Scilly
DE40	Brandenburg	FR30	Nord - Pas-de-Calais	PL42	Zachodniopomorskie	UKK4	Devon
DE50	Bremen	FR41	Lorraine	PL43	Lubuskie	UKL1	West Wales, The Valleys
DE60	Hamburg	FR42	Alsace	PL51	Dolnoslaskie	UKL2	East Wales
DE71	Darmstadt	FR43	Franche-Comt	PL52	Opolskie	UKM2	Eastern Scotland
DE72	Gieen	FR51	Pays de la Loire	PL61	Kujawsko-Pomorskie	UKM3	South Western Scotland
DE73	Kassel	FR52	Bretagne	PL62	Warminsko-Mazurskie	UKM5	North Eastern Scotland
DE80	Mecklenburg-Vorpommern	FR53	Poitou-Charentes	PL63	Pomorskie	UKM6	Highlands, Islands
DE91	Braunschweig	FR61	Aquitaine		**Portugal**	UKN0	Northern Ireland (UK)
DE92	Hannover	FR62	Midi-Pyrnes	PT11	Norte		
DE93	Lneburg	FR63	Limousin	PT15	Algarve		
DE94	Weser-Ems	FR71	Rhne-Alpes	PT16	Centro (PT)		
DEA1	Dsseldorf	FR72	Auvergne	PT17	rea Metropolitana de Lisboa		
DEA2	Kln	FR81	Languedoc-Roussillon	PT18	Alentejo		
DEA3	Mnster	FR82	Provence-Alpes-Cte d'Azur		**Romania**		
DEA4	Detmold	FR83	Corse	RO11	Nord-Vest		
DEA5	Arnsberg		**Hungary**	RO12	Centru		
DEB1	Koblenz	HU10	Kzp-Magyarorszg	RO21	Nord-Est		
DEB2	Trier	HU21	Kzp-Dunntl	RO22	Sud-Est		
DEB3	Rheinhessen-Pfalz	HU22	Nyugat-Dunntl	RO31	Sud - Muntenia		
DEC0	Saarland	HU23	Dl-Dunntl	RO32	Bucuresti - Ilfov		
DED2	Dresden	HU31	szak-Magyarorszg	RO41	Sud-Vest Oltenia		
DED4	Chemnitz	HU32	szak-Alfld	RO42	Vest		
DED5	Leipzig	HU33	Dl-Alfld				
DEE0	Sachsen-Anhalt		**Ireland**				
DEF0	Schleswig–Holstein	IE01	Border, Midland, Western				
DEG0	Thringen	IE02	Southern, Eastern				

Table 7
Monte Carlo simulation results with varying number of genetic algorithms.

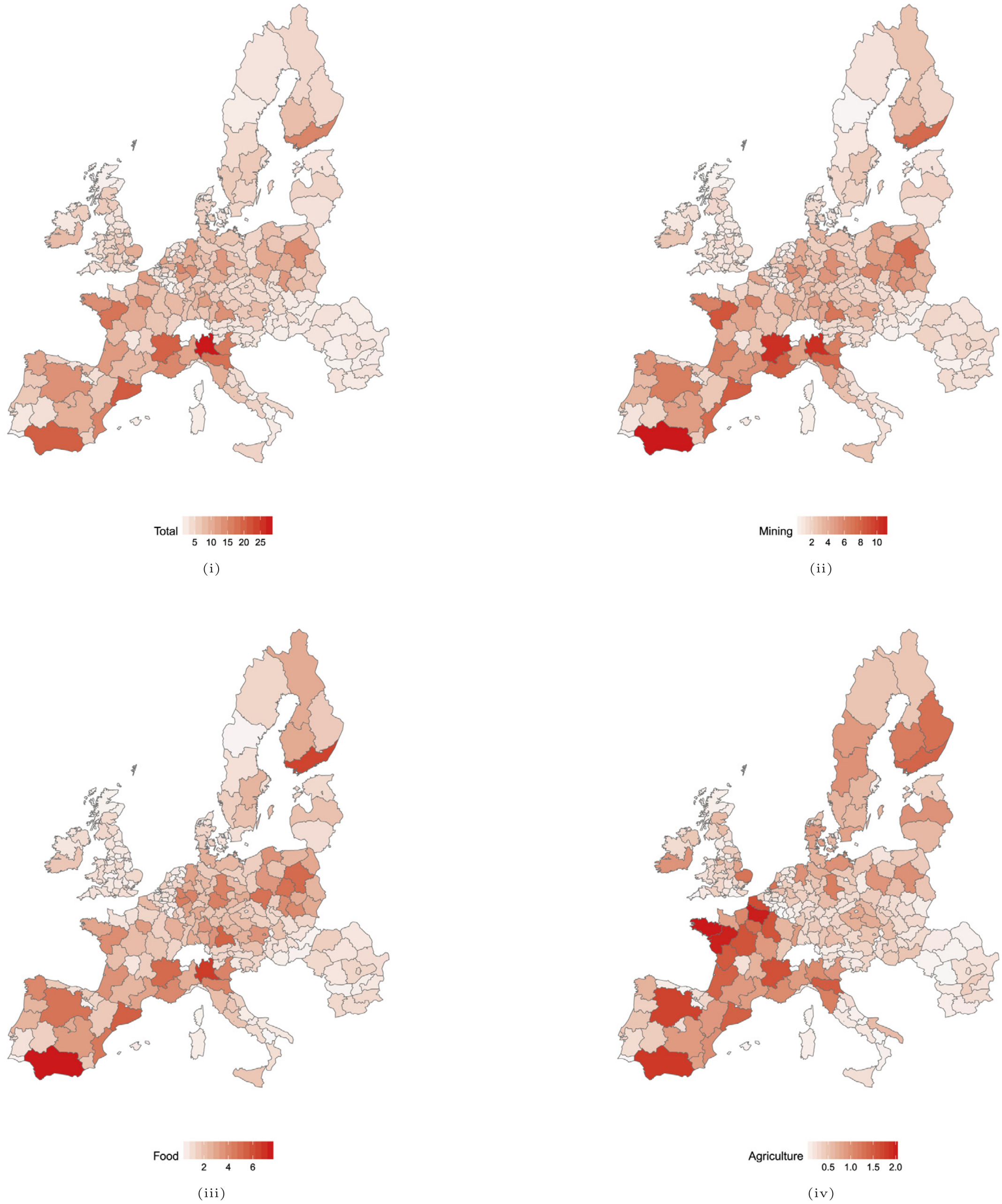
			7 GAs				10 GAs				15 GAs			
SNR	ρ	Sample size	Mean bias σ^2	Mean bias ρ	Mean bias $\partial y/\partial \boldsymbol{x}_1$	Mean bias $\partial y/\partial \boldsymbol{x}_2$	Mean bias σ^2	Mean bias ρ	Mean bias $\partial y/\partial \boldsymbol{x}_1$	Mean bias $\partial y/\partial \boldsymbol{x}_2$	Mean bias σ^2	Mean bias ρ	Mean bias $\partial y/\partial \boldsymbol{x}_1$	Mean bias $\partial y/\partial \boldsymbol{x}_2$
0.1	0.0	100	0.57	0.02	−0.09	0.14	0.54	0.06	0.00	0.14	0.62	0.02	−0.09	0.30
		350	−0.53	0.00	−0.07	0.06	−0.56	−0.02	−0.21	0.06	−0.49	0.00	−0.07	0.14
		700	−0.61	0.00	−0.04	−0.02	−0.63	0.02	0.12	−0.02	−0.65	0.00	−0.07	0.06
	0.5	100	0.94	0.03	−0.06	0.15	0.96	−0.01	−0.16	0.15	0.94	0.02	0.03	0.05
		350	−0.53	0.00	−0.07	0.03	−0.49	0.02	−0.15	0.03	−0.53	0.00	−0.06	0.08
		700	−0.78	0.00	0.00	−0.03	−0.74	−0.04	0.18	−0.03	−0.76	0.00	−0.06	−0.09
	0.8	100	1.30	0.03	−0.03	0.16	1.25	0.04	0.12	0.16	1.25	0.03	−0.02	−0.01
		350	−0.52	0.00	−0.06	−0.01	−0.48	−0.04	−0.17	−0.01	−0.48	0.00	−0.13	−0.04
		700	−0.94	0.00	0.04	−0.03	−0.90	−0.04	−0.13	−0.03	−0.97	0.00	0.01	0.15
0.5	0.0	100	0.32	−0.03	−0.10	−0.07	0.28	−0.06	0.01	−0.07	0.29	−0.03	−0.19	0.09
		350	−0.29	0.01	−0.11	0.12	−0.26	−0.03	0.03	0.12	−0.33	0.01	−0.19	0.22
		700	0.13	0.00	−0.98	−0.09	0.14	0.03	−0.82	−0.09	0.17	0.00	−0.89	0.08
	0.5	100	0.51	−0.02	−0.78	0.12	0.50	−0.01	−0.67	0.12	0.56	−0.02	−0.69	−0.01
		350	−0.47	0.01	0.09	0.08	−0.48	0.05	0.18	0.08	−0.43	0.01	0.09	0.04
		700	−0.35	−0.01	0.12	−0.06	−0.32	0.02	0.17	−0.06	−0.39	−0.01	0.10	−0.02
	0.8	100	−0.08	0.04	−0.22	−0.25	−0.12	0.05	−0.10	−0.25	−0.11	0.04	−0.32	−0.39
		350	0.12	−0.02	0.09	0.07	0.07	−0.03	−0.13	0.07	0.10	−0.02	0.13	−0.08
		700	0.10	0.01	0.12	−0.03	0.05	0.05	0.01	−0.03	0.13	0.01	0.20	−0.21
0.8	0.0	100	−0.19	−0.03	−0.12	−0.28	−0.14	−0.01	0.13	−0.28	−0.23	−0.03	−0.12	−0.27
		350	−0.15	0.01	−0.01	0.18	−0.16	−0.01	−0.15	0.18	−0.11	0.01	−0.04	0.11
		700	−0.08	0.01	−0.14	−0.15	−0.05	0.02	−0.03	−0.15	−0.11	0.01	−0.21	−0.21
	0.5	100	−0.27	0.00	−0.11	−0.17	−0.24	0.02	−0.33	−0.17	−0.22	0.01	−0.13	−0.27
		350	0.10	0.02	0.03	0.13	0.08	0.04	−0.05	0.13	0.12	0.02	−0.07	0.17
		700	−0.02	0.03	−0.14	−0.09	−0.04	0.03	−0.08	−0.09	−0.01	0.03	−0.12	0.02
	0.8	100	−0.17	0.00	−0.09	−0.65	−0.19	−0.02	−0.15	−0.65	−0.18	0.00	−0.11	−0.80
		350	−0.09	0.00	0.07	0.15	−0.12	0.02	0.11	0.15	−0.07	0.00	0.04	0.04
		700	−0.07	0.01	−0.14	−0.03	−0.09	0.04	−0.21	−0.03	−0.11	0.01	−0.07	0.16

Notes: The numbers in the first row denote the number of genetic algorithms (GAs) executed in a parallel fashion, based on the algorithm by Sycara (1998). $\mathscr{F}(\mathbf{X}_1) = 5\boldsymbol{x}_1^3 + 0.8\sqrt{\boldsymbol{x}_2 + 1.5}$. *SNR* stands for signal to noise ratio; the reported partial derivatives correspond to the *total effects* summary impact measures from LeSage and Pace, 2009. Each row reports average values over 300 Monte Carlo runs. The algorithms exchange one member of their population every 50 cycles. The adaptive spline knot algorithm ran for a maximum of $T = 100{,}000$ cycles, with a population size of $N_P = 100$ and $N_{P_{mate}} = 10$, and mutation probability of 2%.

Table 8
Summary impact measures of the semi-parametric SAR with fixed spline knots freight generation model.

Variable	Direct	Indirect	Total
Total employment (in %)	2.646	1.586	4.232
Agricultural employment (in %)	0.810	1.919	2.729
Manufacturing employment (in %)	−2.520	−1.245	−3.297
Length of road network (10,000 km)	1.405^{*}	0.764^{*}	2.170^{*}
Travel time to seaport (min.)	−0.582	−0.386	−0.968
ρ	0.396^{***}		
σ^2	0.151		
Log-likelihood	−96.641		
AIC_c	320.123		
$\overline{R}^2$	0.23		
Number of observations	258		
Number of parameters	50		

Notes: *** and ** denote statistical significance at the one percent level and five percent level, respectively (significance levels calculated using the algorithm proposed by LeSage and Pace (2009)). A spatial neighborhood matrix configuration with seven nearest neighbors was used.



(i)

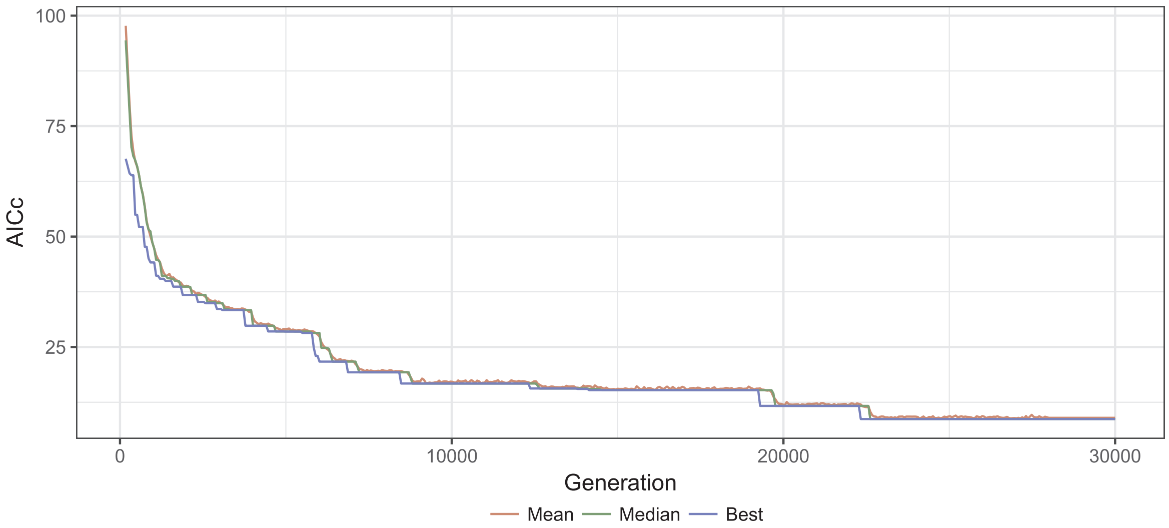
(ii)

(iii)

(iv)

Notes: *Mining* – NST/R GT03 Metal ores and other mining and quarrying products, peat, uranium and thorium ores; NST/R GT09 Other non-metallic mineral products; *Food* – NST/R GT04 Food products, beverages and tobacco; *Agriculture* – NST/R GT01 Products of agriculture, hunting, and forestry, fish and other fishing products. Source: *Eurostat*.

Fig. 4. Yearly million tons of freight generated by NUTS-2 regions in 2011; (i) aggregate over all sectors, (ii) mining, (iii) food, and (iv) agricultural sector. **Notes**: *Mining* – NST/R GT03 Metal ores and other mining and quarrying products, peat, uranium and thorium ores; NST/R GT09 Other non-metallic mineral products; *Food* – NST/R GT04 Food products, beverages and tobacco; *Agriculture* – NST/R GT01 Products of agriculture, hunting, and forestry, fish and other fishing products. Source: *Eurostat*.



Notes: The semi-parametric SAR adaptive spline knot algorithm was evaluated with $T = 100,000$ cycles, but convergence was reached after $23,291$ cycles. We used twelve parallel genetic algorithms, which exchange one member of their population every 50 cycles. The population and mating pool lengths were set to $N_P = 100$ and $N_{P_{mate}} = 10$, while mutation probability was 2%.

Fig. 5. Convergence of the genetic algorithm. **Notes**: The semi-parametric SAR adaptive spline knot algorithm was evaluated with $T = 100{,}000$ cycles, but convergence was reached after 23,291 cycles. We used twelve parallel genetic algorithms, which exchange one member of their population every 50 cycles. The population and mating pool lengths were set to $N_P = 100$ and $N_{P_{mate}} = 10$, while mutation probability was 2%.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control.

Al-Deek, J., El-Maghraby, M., 2000. Truck trip generation models for seaports with container and trailer operation. Transport. Res. Rec.: J. Transport. Res. Board 1719, 1–9.

Anselin, L., 1988. Spatial Econometrics: Methods and Models. Kluwer, Dordrecht.

Anselin, L., Bera, A., 1998. Spatial dependence in linear regression models with an introduction to spatial econometrics. In: Ullah, A., Giles, D.E.A. (Eds.), Handbook of Applied Economic Statistics. Marcel Dekker, New York, pp. 237–289.

Anselin, L., Florax, R., Rey, S.J., 2004. Advances in Spatial Econometrics: Methodology, Tools and Applications. Springer, Berlin Heidelberg New York.

Awad, A.M., 1996. Properties of the akaike information criterion. Microelectron. Reliab. 36, 457–464.

Aydin, M.E., Fogarty, T.C., 2004. Teams of autonomous agents for job-shop scheduling problems: an experimental study. J. Intell. Manuf. 15, 455–462.

Basile, R., 2008. Regional economic growth in Europe: a semiparametric spatial dependence approach. Papers Region. Sci. 87, 527–544.

Basile, R., Benfratello, L., Castellani, D., 2013. Geoadditive models for regional count data: an application to industrial location. Geograph. Anal. 45, 28–48.

Basile, R., Mínguez, R., Montero, J.M., Mur, J., 2014. Modelling regional economic dynamics: spatial dependence, spatial heterogeneity and nonlinearities. J. Econ. Dynam. Control 48, 229–245.

Chow, J.Y.J., Yang, C.H., Regan, A.C., 2010. State-of-the art of freight forecast modeling: lessons learned and the road ahead. Transportation 37, 1011–1030.

Chun, Y., Kim, H., Kim, C., 2012. Modeling interregional commodity flows with incorporating network autocorrelation in spatial interaction models: an application of the US interstate commodity flows. Comput., Environ. Urban Syst. 36, 583–591.

De Grange, L., Fernández, E., De Cea, J., 2010. A consolidated model of trip distribution. Transport. Res. Part E: Logist. Transport. Rev. 46, 61–75.

De Jong, G., Vierth, I., Tavasszy, L., Ben-Akiva, M., 2013. Recent developments in national and international freight transport models within Europe. Transportation 40, 347–371.

DeBoor, C., 1978. A Practical Guide to Splines. Springer, Berlin Heidelberg New York.

Del Bo, C.F., Florio, M., 2012. Infrastructure and growth in a spatial framework: evidence from the EU regions. Eur. Plan. Stud. 20, 1393–1414.

Eilers, P.H.C., Marx, B.D., 1996. Flexible smoothing with B-splines and penalties. Stat. Sci. 11, 89–121.

Fahrmeier, L., Kneib, T., Lang, S., 2004. Penalized structured additive regression for space-time data: a bayesian perspective. Stat. Sin. 14, 715–745.

Fahrmeir, L., Kneib, T., Lang, S., 2009. Regression: Modelle, Methoden und Anwendungen, 2nd ed. Springer, Berlin Heidelberg New York.

Fischer, M.M., Leung, Y., 1998. A genetic-algorithms based evolutionary computational neural network for modelling spatial interaction data. Annals Region. Sci. 32, 437–458.

Fischer, M.M., Wang, J., 2011. Spatial Data Analysis: Models, Methods and Techniques. Springer Heidelberg Dordrecht London, New York.

Fotopoulos, G., 2012. Nonlinearities in regional economic growth and convergence: the role of entrepreneurship in the European union regions. Annals Region. Sci. 48, 719–741.

Goldberg, D., Kalyanmoy, D., Korb, B., 1989. Messy genetic algorithms: motivation, analysis, and first results. Complex Syst. 3, 493–530.

Hesse, M., Rodrigue, J.P., 2004. The transport geography of logistics and freight distribution. J. Transp. Geogr. 12, 171–184.

Holguín-Veras, J., Patil, G., 2008. A multicommodity integrated freight origin-destination synthesis model 8, 309–326.

Holland, J.H., 1992. Genetic algorithms. Scient. Am. 267, 66–72.

Hurvich, C.M., Simonoff, J.S., Tsai, C.-L., 1998. Smoothing parameter selection in nonparametric regression using an improved Akaike Information Criterion. J. Roy. Stat. Soc. Ser. B, Stat. Methodol. 60, 271–293.

Kazemi, A., Zarandi, M.F., Husseini, S.M., 2009. A multi-agent system to solve the production–distribution planning problem for a supply chain: a genetic algorithm approach. Int. J. Adv. Manuf. Technol. 44, 180–193.

Koch, M., Krisztin, T., 2011. Applications for asynchronous multi-agent teams in nonlinear applied spatial econometrics. J. Internet Technol. 12, 1007–1014.

Koop, G., Poirier, D., 2004. Bayesian variants of some classical semiparametric regression techniques. J. Econom. 123, 259–282.

Krisztin, T., 2017. The determinants of regional freight transport: a spatial, semiparametric approach. Geogr. Anal. 49, 268–308.

Lawson, C., Holguín-Veras, J., Sánchez-Díaz, I., Jaller, M., Campbell, S., Powers, E., 2012. Estimated generation of freight trips based on land use. Transport. Res. Rec.: J. Transport. Res. Board 2269, 65–72.

LeSage, J.P., Pace, R.K., 2009. Introduction to Spatial Econometrics. CRC Press, Boca Raton London New York.

Novak, D.C., Hodgdon, C., Guo, F., Aultman-Hall, L., 2011. Nationwide freight generation models: a spatial regression approach. Networks Spatial Econ. 11, 23–41.

Ortúzar, J.d.D., Willumsen, L.G., 2011. Modelling Transport, 4th ed. John Wiley & Sons, Chichester.

Pace, R.K., LeSage, J.P., 2002. Semiparametric maximum likelihood estimates of spatial dependence. Geogr. Anal. 34, 76–90.

Qu, X., Lee, L.F., 2015. Estimating a spatial autoregressive model with an endogenous spatial weight matrix. J. Econom. 184, 209–232.

Ranaiefar, F., Chow, J.Y.J., Rodriguez-Roman, D., Veiga de Camargo, P., Ritchie, S.G., 2013. Structural commodity generation model that uses public data: geographic scalability and supply chain elasticity analysis. Transport. Res. Rec., pp. 73–83.

Rodrigue, J.P., 2006. Challenging the derived transport-demand thesis: geographical issues in freight distribution. Environ. Plan. A 38, 1449–1462.

Ruppert, D., Wand, M.P., Carroll, R.J., 2003. Semiparametric Regression. Cambridge University Press, Cambridge.

Sánchez-Díaz, I., 2017. Modeling urban freight generation: a study of commercial establishments freight needs. Transport. Res. Part A: Policy Pract. 102, 3–17.

Sánchez-Díaz, I., Holguín-Veras, J., Ban, X.J., 2015. A time-dependent freight tour synthesis model. Transport. Res. Part B: Methodol. 78, 144–168.

Sánchez-Díaz, I., Holguín-Veras, J., Wang, X., 2016. An exploratory analysis of spatial effects on freight trip attraction. Transportation 43, 177–196.

Sycara, K.P., 1998. Multiagent systems. AI Magaz. 19, 79.

Talukdar, S., Baerentzen, L., Gove, A., De Souza, P., 1998. Asynchronous teams: cooperation schemes for autonomous agents. J. Heurist. 4, 295–321.

Tavasszy, L.A., Ruijgrok, K., Davydenko, I., 2012. Incorporating logistics in freight transport demand models: state-of-the-art and research opportunities. Transp. Rev. 32, 203–219.

White, H., 2000. Asymptotic Theory for Econometricians, 3rd ed. Academic Press, Orlando San Diego San Francisco New York London.