REGION
The Journal of ERSA
Powered by WU

Volume 9, Number 2, 2022, R1–R15
DOI: 10.18335/region.v9i2.450

journal homepage: region.ersa.org
ISSN: 2409-5370

An introduction to pspatreg. A new R package for semiparametric spatial autoregressive analysis

Román Mínguez[1], Roberto Basile[2], María Durbán[3]

[1] University of Castilla-La Mancha, Cuenca, Spain
[2] University of L'Aquila, Localita Acquasanta, Italy
[3] University Carlos III of Madrid, Madrid, Spain

Received: 20 July 2022/Accepted: 25 October 2022

Abstract. This article introduces a new R package (*pspatreg*) for the estimation of semiparametric spatial autoregressive models. *pspatreg* fits penalized spline semiparametric spatial autoregressive models via Restricted Maximum Likelihood or Maximum Likelihood. These models are very flexible as they make it possible to simultaneously control for spatial dependence, nonlinearities in the functional form, and spatio-temporal heterogeneity. The package also allows to estimate parametric spatial autoregressive models for both cross-sectional and panel data (with fixed effects), thus avoiding the use of different libraries. The official demos, vignettes, and tutorials of the package are distributed either in CRAN or GitHub. This article illustrates the potentials of the package by applying it to cross-sectional data.

Key words: R package, Spatial dependence, Semiparametric models, Splines

1 Introduction

Modeling spatial and spatio-temporal data requires flexible econometric tools that allow us to control spatial and temporal dependence, spatial heterogeneity, non-linearities, and other possible model specification biases. When combined with standard parametric spatial econometric approaches, semiparametric regression models can provide an answer to this demand for flexibility. New computational methods developed within most modern statistical software (such as R) allow us to overcome all technical problems that arise in this process.

Several packages have recently been proposed to perform spatial econometrics in R (see Bivand et al. 2021, for a recent survey). Focusing on packages and methods dealing with polygonal (or areal) spatial data, the first package was *spdep* (Bivand et al. 2013, Bivand 2022). It was primarily designed for cross-sectional spatial data and to model spatial dependence through the Maximum Likelihood (ML) or the Generalized Method of Moments (GMM) estimation of the spatial lag model (SAR), the spatial error model (SEM), the spatial Durbin model (SDM), and the SARAR model. The estimation functions from *spdep* have recently been moved to the package *spatialreg* (Bivand et al. 2021). Other spatial econometric models for cross-sectional data have been implemented in other packages: *sphet* (Piras 2010) for estimating and testing spatial models with heteroskedastic innovations, *spfilteR* (Juhl 2021) for filtering out spatial dependence in linear models, *spgwr* (Bivand, Yu 2022) for estimating geographycally weighted regression

models, and *spsur* (Lopez et al. 2020) for estimating seemingly unrelated regression equations. Moreover, following several theoretical contributions to the literature on the estimation of static and dynamic spatial panel data models (see Elhorst 2014), other `R` packages for spatial econometric analysis have recently been developed. In particular, *splm* (Millo, Piras 2012) and *SDPDm* implement estimation methods for static and dynamic spatial panel data. All of these `R` packages focus on parametric methods (except *spgwr*, of course), leaving aside issues related to non-linearities in functional form and the estimation of spatio-temporal trends.

The main focus of this article is *pspatreg*, a new `R` package for spatial econometric analysis. *pspatreg* fits penalized spline (PS) semiparametric spatial autoregressive models via Restricted Maximum Likelihood (REML) and ML. This approach combines penalized regression spline methods[1] (Eilers et al. 2015) with standard spatial autoregressive models (such as SAR, SEM and SDM). These types of models (PS-SAR, PS-SEM and PS-SDM) are thoroughly discussed in Mínguez et al. (2020) (see also Montero et al. 2012, Basile et al. 2014, Hoshino 2018).

These models are very flexible as they make it possible to include within the same specification: *i*) spatial autoregressive terms (i.e. spatial lags of dependent and independent variables as well as spatial error terms) to capture spatial interaction or network effects; *ii*) time lags of the dependent variable to capture persistence effects; *iii*) parametric and nonparametric (smooth) terms to identify nonlinear relationships between the response variable and the covariates; *iv*) spatial and spatio-temporal trends, i.e. a smooth interaction between the spatial coordinates and the time trend, to capture site-specific nonlinear time trends.

The proposed method also allows the user to apply an ANOVA decomposition of the spatial or spatio-temporal trend into several components (spatial and temporal main effects, and second- and third-order interactions between them). This gives further insights into the dynamics of the data. Thus, we use the acronym PS-ANOVA-SAR (SEM, SDM, SLX) for the newly proposed data generating process (DGP). The use of nested B-spline bases for the interaction components of the spatio-temporal trend (Lee et al. 2013) contributes to the efficiency of the fitting procedure without compromising the goodness of fit of the model. Finally, we also consider an extension of the PS-ANOVA-SAR (SEM, SDM, SLX), including a first-order time series autoregressive term process (AR1) in the noise to accommodate residual serial correlation. Further extensions to include the time lag of the dependent variable (dynamic spatial model) will be considered in the future.

The next section (Section 2) describes the availability of *pspatreg* with documentation and examples. Section 3 presents a general specification of the semiparametric spatial autoregressive model. Section 4 shows an example of using *pspatreg* with cross-sectional spatial data. The last section presents a conclusion.

2 Documentation of *pspatreg*

The *pspatreg* package is available on both CRAN (https://cran.r-project.org/web/packages/pspatreg/index.html) and GitHub (https://github.com/rominsal/pspatreg) and can be installed in the usual way[2].

Once the package has been installed and loaded, an overview of the functionality of the package, including main functions, methods and databases, can be obtained executing the command `?pspatreg`.

The package includes three vignettes. The first one provides a brief description of the methodology used in the package. The second vignette gives a detailed example of modeling pure spatial data with semiparametric models and spatial lags using the well-

[1] P-splines are a flexible tool for smoothing. They are based on regressions with a large number of local basis functions (called B-splines). A penalty function based on differences between adjacent coefficients is also included in the maximum likelihood function to tune the smoothness of the estimated curve.

[2] You could install *pspatreg* from CRAN executing `install.packages("pspatreg")`. Usually the default options allow to install the package without any problems. Alternatively, to install from GitHub you could use *devtools* package. Once installed, execute the command `devtools::install_github("rominsal/pspatreg")` to install *pspatreg* package.

known *Ames* database included in package *AmesHousing* (Kuhn 2020). It also compares the results of *pspatreg* with the *spatialreg* package for parametric spatial regression models. Lastly, the third vignette provides some insights into spatio-temporal modeling using a panel database of unemployment in Italian provinces. First, this vignette compares the results of spatio-temporal parametric panels with the *splm* package, and then it shows the results of semiparametric spatio-temporal models. Plots of spatio-temporal trends are also included in these examples.

Of course, every function in the package includes reproducible examples. Those included in `pspatfit()`, `impactspar()`, `impactsnopar()`, `plot_sp2d()`, `plot_sp3d()`, and `plot_sptime()` functions are especially interesting. Furthermore, these examples can be also checked using the demos of the package, see `?demo(package = "pspatreg")` for details of the included demos.

3 The Semiparamentric Spatial Autoregressive Model

Let y_{it} be a sample of spatial panel data, where i is an index for the cross-sectional dimension (spatial units), with $i = 1, \ldots, N$, and t is an index for the time dimension (time periods), with $t = 1, \ldots, T$. The general model proposed is written as:

$$y_{it} = \rho \sum_{j=1}^{N} w_{ij,N} y_{jt} + \tilde{f}(s_{1i}, s_{2i}, \tau_t) + \sum_{\delta=1}^{k} g_\delta(x_{\delta_{it}}) + \epsilon_{it},$$

where (s_{1i}, s_{2i}) are the spatial coordinates (latitude and longitude) of individual i (when i refers to areal units: municipality, provinces, etc., the standard convention here is to identify representative points for areal units, the most typical being areal centroids), τ_t is the time period, and $x_{\delta_{it}}$ are independent variables; w_{ij} are the spatial weights, and ρ the spatial autoregressive parameter. The functions $g_\delta(.)$ are parametric or non-parametric smooth functions of the covariates $x_{\delta_{it}}$ (they can be linear, or can accommodate varying coefficient terms, smooth interaction between covariates, smooth by-factor curves, and so on), and $\tilde{f}(s_{1i}, s_{2i}, \tau_t)$ is an unknown non-parametric spatio-temporal trend. The idiosyncratic error term is assumed to follow an AR(1) process, i.e., $\epsilon_{it} = \phi \epsilon_{it-1} + u_{it}$ with $u_{it} \sim N(0, \sigma^2)$.

This semiparametric SAR model turns out to be extremely useful to capture interactive spatial and temporal unobserved heterogeneity when this heterogeneity is smoothly distributed over space and time (Mínguez et al. 2020). The dynamic extension (including y_{it-1} and $\sum_{j=1}^{N} w_{ij,N} y_{it-1}$) is also very promising and merits further theoretical investigation. Finally, the following semiparametric SAR model is very useful for modeling cross-setional spatial data taking into account non-linearities, spatial dependence, and spatial heterogeneity:

$$y_i = \rho \sum_{j=1}^{N} w_{ij,N} y_j + \sum_{\delta=1}^{\Delta} g_\delta(x_{\delta,i}) + \tilde{f}(s_{1i}, s_{2i}) + \epsilon_i$$

$$\epsilon_i \sim i.i.d.(0, \sigma_\epsilon^2).$$

3.1 The Anova Decomposition of the Spatio-temporal Trend

In many situations, the spatial or the spatio-temporal trend to be estimated can be complex, and the use of a single multidimensional smooth function may not be flexible enough to capture the structure in the data. To solve this problem, an ANOVA-type decomposition of $\tilde{f}(s_{1i}, s_{2i}, \tau_t)$ can be used, where spatial and temporal main effects, and second- and third-order interactions between them can be identified:

$$\begin{aligned} \tilde{f}(s_{1i}, s_{2i}, \tau_t) \quad = \quad & f_1(s_{1i}) + f_2(s_{2i}) + f_\tau(\tau_t) + f_{1,2}(s_{1i}, s_{2i}) + \\ & f_{1,\tau}(s_{1i}, \tau_t) + f_{2,\tau}(s_{2i}, \tau_t) + f_{1,2,\tau}(s_{1i}, s_{2i}, \tau_t) \end{aligned}$$

First, the geoadditive terms given by $f_1(s_{1i}), f_2(s_{2i}), f_{1,2}(s_{1i}, s_{2i})$ work as control functions to filter the spatial trend out of the residuals, and transfer it to the mean response in a model specification. Thus, they make it possible to capture the shape of the spatial distribution of y_{it}, conditional on the determinants included in the model. These control functions also isolate stochastic spatial dependence in the residuals, that is, spatially autocorrelated unobserved heterogeneity. Thus, the geoadditive terms can be regarded as an alternative to the use of individual regional dummies to capture unobserved heterogeneity, as long as such heterogeneity is smoothly distributed over space. Regional dummies peak at significantly higher and lower levels of the mean response variable. If these peaks are smoothly distributed over a two-dimensional surface (i.e., if unobserved heterogeneity is spatially autocorrelated), the smooth spatial trend is able to capture them. It is also worth noticing that, in a cross-sectional setting, the inclusion of a smooth spatial trend in the model specification is often the best way to control for unobserved spatial heterogeneity in the absence of degrees of freedom for the introduction of spatial fixed effects.

Second, the smooth time trend, $f_\tau(\tau_t)$, and the smooth interactions between space and time – $f_{1,\tau}(s_{1i}, \tau_t)$, $f_{2,\tau}, (s_{2i}, \tau_t)$, $f_{1,2,\tau}(s_{1i}, s_{2i}, \tau_t)$ – work as control functions to capture the heterogeneous effect of common shocks. Thus, conditional on a smooth distribution of the spatio-temporal heterogeneity, the PS-ANOVA-SAR (SDM, SEM, SLX) model works as an alternative to the models proposed by Bai, Li (2013), Shi, Lee (2018), Pesaran, Tosetti (2011), Bailey et al. (2016) and Vega, Elhorst (2016) which are extensions of common factor models to accommodate both strong cross-sectional dependence (through the estimation of the spatio-temporal trend) and weak cross-sectional dependence (through the estimation of spatial autoregressive parameters).

Furthermore, this framework is also flexible enough to control for the linear and non-linear functional relationships between the dependent variable and the covariates, as well as the heterogeneous effects of these regressors across space. The model inherits all the positive properties of penalized regression splines, such as coping with missing observations by appropriately weighting them and straightforward interpolation of the smooth functions.

3.2 Direct and Indirect (Spillover) Effects of Smooth Terms in the PS-SAR Model

In the case of a semiparametric model without the spatial lag of the dependent variable (PS model), if all regressors are independent of the errors, $\hat{g}_\delta(x_{\delta,it})$ can be interpreted as the conditional expectation of y given x_δ (net of the effect of the other regressors). Blundell, Powell (2003) use the term Average Structural Function (ASF) with reference to these functions. In contrast, in PS-SAR, PS-SDM or in PS-SARAR model, when ρ is different from zero, the estimated smooth functions cannot be interpreted as ASF. Taking advantage of the results obtained for parametric SAR, we can compute the total smooth effect (total–ASF) of x_δ as:

$$\hat{g}_\delta^T(x_\delta) = \Sigma_q \left[\mathbf{I}_n - \hat{\rho}\mathbf{W}_n\right]_{ij}^{-1} b_{\delta q}(x_\delta)\hat{\beta}_{\delta q},$$

where $b_{\delta q}(x_\delta)$ are the B-spline basis functions used to represent the smooth function, and $\hat{\beta}_{\delta q}$ the corresponding estimated parameters.

We can also compute direct and indirect (or spillover) effects of smooth terms in the PS-SAR case as:

$$\hat{g}_\delta^D(x_\delta) = \Sigma_q \left[\mathbf{I}_n - \hat{\rho}\mathbf{W}_n\right]_{ii}^{-1} b_{\delta q}(x_k)\hat{\beta}_{\delta q}$$

$$\hat{g}_\delta^I(x_\delta) = \hat{g}_\delta^T(x_\delta) - \hat{g}_\delta^D(x_\delta).$$

Similar expressions can be provided for the direct, indirect, and total effects of the PS-SDM.

4 Basic Information on *pspatreg*

We are now going to introduce some basic general information about the package. The main function in the *pspatreg* package is `pspatfit()`, which estimates spatio-temporal pernalized spline spatial regression models using either the REML method or the ML method. In its generic form, `pspatfit()` appears as:

```
pspatfit(formula, data, na.action, listw = NULL, type = "sim", method = "eigen",
    Durbin = NULL, zero.policy = NULL, interval = NULL,trs = NULL, cor = "none",
    dynamic = FALSE, control = list())
```

The function `pspatfit()` returns a list of objects of class `pspatreg`, including coefficients of the parametric terms and their standard errors, estimated coefficients corresponding to random effects in mixed model and their standard errors, equivalent degrees of freedom, residuals, fitted values, etc. A wide range of standard methods is also available for the `pspatreg` objects, including `print()`, `summary()`, `coef()`, `vcov()`, `anova()`, `fitted()`, `residuals()`, and `plot()`.

The argument `formula` within the function `pspatfit()` is formula similar to the GAM specification including parametric and non-parametric terms. Parametric covariates are included in the usual way. Non-parametric p-spline smooth terms are specified using `pspl(.)` and `pspt(.)` for the non-parametric covariates and spatial or spatio-temporal trends, respectively. For example:

[1]:
```
formula <- y ~ x1 + x2 + pspl(x3, nknots = 15) + pspl(x4, nknots = 20) +
  pspt(long, lat, year, nknots = c(18,18,8), psanova = TRUE,
       nest_sp1 = c(1, 2, 3),
       nest_sp2 = c(1, 2, 3),
       nest_time = c(1, 2, 2))
```

In the example above, the model includes two parametric terms, two non-parametric terms, and a spatio-temporal trend (with long and lat as spatial coordinates and year as temporal coordinate). The dimension of the basis function, both in `pspl(.)` and `pspt(.)`, is defined by `nknots`. This term should not be less than the dimension of the null space of the penalty for the term (see `null.space.dimension` and `choose.k` from package *mgcv* (Wood 2017) to know how to choose `nknots`). The default number of `nknots` in `pspl(.)` is 10 but, in this example, we have chosen 15 `nknots` for `g_1(x_3)` and 20 `nknots` for `g_2(x_4)`. The default number of `nknots` in `pspt(.)` is `c(10,10,5)`, but we have chosen `c(18,18,8)`.

In this example we also adopt an ANOVA decomposition of the spatio-temporal trend (choosing `psanova = TRUE`). Each effect has its own degree of smoothing which allows a greater flexibility for the spatio-temporal trend. Calculating up to third-order interactions can be computationally expensive. We can select subgroups of interaction effects for the second- and third-order effects to address this problem. We use three parameters available in `pspt()`: `nest_sp1`, `nest_sp2`, and `nest_time` to define these subgroups. These parameters indicate the divisors of the `nknots` parameters. For example, if we set `nest_sp1 = c(1,2,3)`, we will have all knots for the `s_1` effect, 18/2 for each second-order effects with `s_1`, and 18/3 nots for the third order effect with `s_1`[3].

We must set the parameters `f1_main`, `f2_main` or `ft_main` to `FALSE` (the default is `TRUE`) if we want to exclude any main effect. We can also exclude second- or third-order effects setting `f12_int`, `f1t_int`, `f2t_int`, `f12t_int` to `FALSE`.

Using the argument `Type`, we can choose different spatial model specifications: `"sar"`, `"sem"`, `"sdm"`, `"sdem"`, `"sarar"`, or `"slx"`. When creating a `"slx"`, `"sdem"`, or `"sdm"` model, we need to include the formula of the durbin part in the Durbin parameter.

The argument `data` must contain all the variables included in parametric and non-parametric terms of the model. If a `pspt(.)` term is included in `formula`, the data must contain the spatial and temporal coordinates specified in `pspt(.)`. In this case, the

[3] In most empirical cases, the main effects are more flexible than interaction effects and therefore the number of knots in B-Spline bases for interaction effects do not need to be as large as the number of knots for the main effects (Lee et al. 2013).

coordinates must be ordered choosing time as fast index and spatial coordinates as slow indexes.

Both `data.frame` and `sf` class objects can be used as `data` inputs[4]. `sf` objects are recommended since they allow the user to map spatial trends. We use two datasets in `sf` version for our demos.

Plotting the estimated non-parametric smooth terms represents an important step in semiparametric regression analyses. First, the function `fit_terms()` computes estimated non-parametric smooth terms. Then, the functions `plot_sp2d()` and `plot_sp3d()` are used to plot and map spatial and spatio-temporal trends, respectively, while `plot_sptime()` is used to plot the time trend for PS-ANOVA models in 3d. Finally, `plot_terms()` is used to plot smooth non-parametric terms.

The function `impactspar()` computes direct, indirect, and total impacts for continuous parametric covariates using the standard procedure for their computation (LeSage, Pace 2009).

The function `impactsnopar()` computes direct, indirect, and total impacts functions for continuous non-parametric covariates, while the function `plot_impactsnopar()` is used to plot these impacts' functions. It is worth noticing that total, direct, and indirect effects are never smooth over the domain of the variable x_δ due to the presence of the spatial multiplier matrix used in the algorithm for their computation. Indeed, a wiggly profile of direct, indirect, and total effects would appear even if the model was linear. Therefore, in the spirit of the semiparametric approach, we included the possibility of applying a spline smoother to obtain smooth curves (using the argument `smooth=TRUE` in the function `plot_impactsnopar()`).

5 Using *pspatreg* with Cross-sectional Spatial Data

Here, we present the use of *pspatreg* for spatial cross-sectional data (no time dimension involved). In particular, we use Italian province-level data for the estimation of the relationship between labor productivity growth and net internal migration. The standard neoclassical growth model can be specified, in its linear form, as follows:

$$\gamma_i = \alpha + \beta \ln y_{i,0} + \delta m_i + \tau \ln(n_i) + X_i^{'}\psi + \epsilon_i,$$

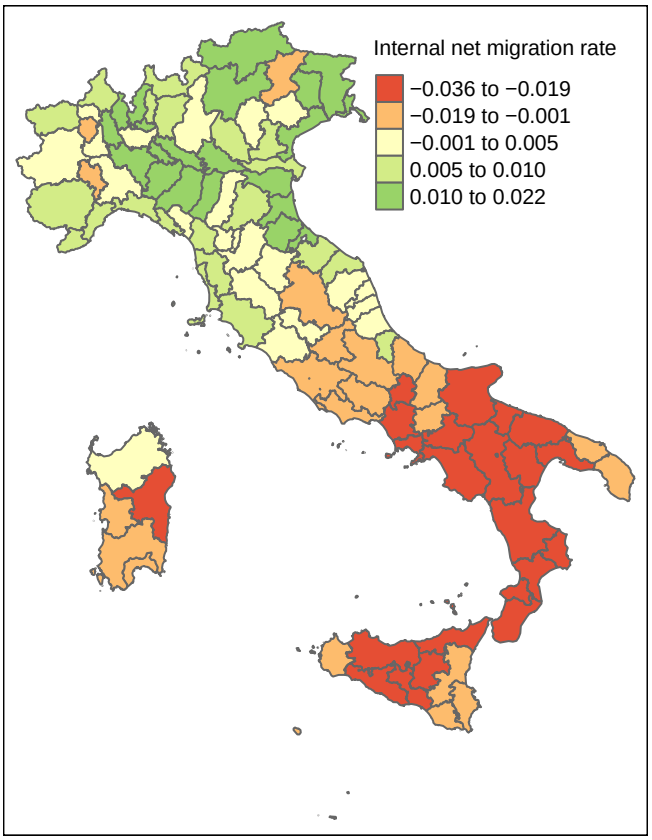
where $\gamma_i = (\ln y_{i,T} - \ln y_{i,0})/T$ is the average annual growth rate of labor productivity (measured as gross value added per worker) computed over T periods (our sample period goes from 2002 to 2018) for each province i (107 Italian provinces), $\ln y_{i,0}$ captures the initial conditions of each province (a negative value of β indicates conditional convergence), m_i is the average annual provincial internal net migration rate (computed as the difference between internal immigration and emigration flows of the working-age population, i.e. people aged 15-65, divided by the total working-age population), $\ln(n_i)$ is the average employment growth rate (the neoclassical growth model suggests a negative value of τ), X_i is a vector of variables controlling for other growth determinants such as physical and human investment rates, and ϵ_i is an identically and independently distributed error term.

Net population movements generally tend to be oriented towards prosperous areas which offer higher real income prospects. This is also true for the Italian case (see Figure 1), where all Southern provinces have negative net migration rates and all Northern provinces have positive rates.

According to the standard neoclassical framework, this pattern of migration should represent a mechanism for reducing spatial economic differentials. Labor migration from poor to rich areas lowers capital intensity (increases the return to capital) in the

[4] `sf` means simple features of spatial vector objects. The geographic vector data model is based on points located within a coordinate reference system (CRS). Points can represent self-standing features (e.g., the location of a house) or they can be linked together to form more complex geometries such as lines and polygons. Most point geometries contain only two dimensions `x` and `y` (3-dimensional CRSs contain an additional `z` value, typically representing height above sea level). `sf` objects provide both a *geometry* information, describing where on Earth the feature is located, and *attributes* information, describing other properties (like the population of the region, the unemployment rate, etc.). `data.frame` objects store only attributes information.

Figure 1: Internal net migration rate from 2002 to 2018 in Italian provinces



destination region and increases capital intensity (lowers the return to capital) in the region of origin. When the same technologies are used everywhere, migration speeds up per worker inter-regional convergence in capital intensity and labor productivity. Therefore, the neoclassical framework predicts a negative value of δ (i.e. net inward migration reduces labor productivity growth). However, alternative theories point to the importance of migrants' characteristics such as youthfulness, entrepreneurship, and skills that, together with their impact on aggregate demand, may have growth-enhancing effects. In terms of aggregate demand, regions losing population through migration may face economic contraction, whereas regions gaining population through migration may benefit from an expansionary effect on output, employment, and income. The transfer of human capital from one place to another is another critical aspect. In particular, skill-selective mobility may have deep effects on origin and destination places. All these alternative contributions predict a positive effect of net migration on growth (i.e. a positive value of δ). Moreover, the presence of a significantly positive effect of net migration is expected to decrease the estimate of β, the parameter associated to the initial conditions (i.e. it is expected to remove the positive omitted variable bias in estimates of β in regressions without the migration variable). Our empirical analysis confirms this intuition. Using our dataset and estimating the model with simple OLS, we actually find a positive effect of net migration on labor productivity growth, in line with several empirical studies:

[2]:
```
formlin_0 <- growth_PROD ~ lnPROD_0+lnoccgr
linear_0 <- lm(formlin_0, data = prod_it)
summary(linear_0, vcov = function(x) vcovHC(x, type = "HC1"))
```

[2]:
```
##
## Call:
## lm(formula = formlin_0, data = prod_it)
##
## Residuals:
```

```
##        Min         1Q     Median         3Q        Max
## -0.0085578 -0.0022181  0.0001756  0.0020764  0.0082702
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.036413   0.033147   1.099   0.2745
## lnPROD_0    -0.001892   0.003171  -0.596   0.5522
## lnoccgr     -0.153926   0.077535  -1.985   0.0498 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.003261 on 104 degrees of freedom
## Multiple R-squared:  0.0847, Adjusted R-squared:  0.0671
## F-statistic: 4.812 on 2 and 104 DF,  p-value: 0.01003
```

[3]:
```
beta_conv_0 <- as.numeric(-log(linear_0$coefficients[2]*16+1)/16)
beta_conv_0
```

[3]:
```
## [1] 0.00192075
```

[4]:
```
formlin <- growth_PROD ~ lnPROD_0+lnoccgr+net
linear <- lm(formlin, data = prod_it)
summary(linear, vcov = function(x) vcovHC(x, type = "HC1"))
```

[4]:
```
##
## Call:
## lm(formula = formlin, data = prod_it)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0085734 -0.0019501 -0.0000671  0.0021081  0.0089063
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.131008   0.039651   3.304 0.001312 **
## lnPROD_0    -0.010650   0.003747  -2.842 0.005402 **
## lnoccgr     -0.153775   0.072837  -2.111 0.037173 *
## net          0.107752   0.027962   3.854 0.000203 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.003064 on 103 degrees of freedom
## Multiple R-squared:    0.2,  Adjusted R-squared:  0.1767
## F-statistic: 8.585 on 3 and 103 DF,  p-value: 3.854e-05
```

[5]:
```
beta_conv <- as.numeric(-log(linear$coefficients[2]*16 + 1)/16)
beta_conv
```

[5]:
```
## [1] 0.01167548
```

The results indicate a positive correlation between the growth rate of labor productivity and the net migration rate of working-age population. Nevertheless, this linear specification of the model is characterized by a number of potential mis-specification biases. First, there can be a reverse causality problem between migration and productivity growth, so that the net migration variable should be instrumented. A second source of endogeneity could be the presence of omitted variables (or unobserved heterogeneity) correlated with the observed covariates. Indeed, we do not control for human and physical capital accumulation rates in the estimation above, due to the lack of information on these variables at the province level in Italy. Additionally, we cannot exclude a correlation between these omitted terms and the coviarates introduced in the model. Third, substantive spatial dependence effects can emerge due to the network structure of Italian provinces, which are strongly connected via trade or other kinds of links. A wrong functional form (due to non-linearities) can represent a further source of model mis-specification. For the sake of simplicity, we disregard the reverse causality issue and focus on the other sources of bias (unobserved heterogeneity, spatial dependence, and nonlinearities) in what follows. In particular, we show that controlling for unobserved heterogeneity is a fundamental challenge in cross-sectional analysis (where we cannot include spatial fixed effects). Moreover, we

should also consider that spatial dependence may simply be the consequence of (spatially correlated) omitted variables rather than being the result of spillovers. If this is the case, there are no compelling reasons for using traditional parametric models, like the SAR or SEM. As McMillen (2012) shows, a simple semiparametric model, with a smooth interaction between latitude and longitude (the so-called Geoadditive Model), can remove unobserved heterogeneity.

5.1 The Parametric SAR Model

Following a step-by-step procedure, we first extend the linear classical model by including a spatial autoregressive term, i.e. by estimating a SAR model[5]:

$$\gamma_i = \alpha + \rho \sum_{j=1}^{N} w_{ij,N}\gamma_j + \beta \ln y_{i,0} + \delta m_i + \tau \ln(n_i) + \epsilon_i.$$

We estimate this model using the function `pspatfit()` of the package *pspatreg* and the function `impactspar()` to compute direct, indirect, and total marginal effects. The results show a significant spatial autoregressive parameter ρ of 0.365. The average direct effect of net migration (0.11) is similar to the coefficient estimated with OLS, but we also observe an indirect (spillover) impact of 0.06 and thus a total average effect of 0.17. The same results are obviously obtained using the package *spatialreg*.

[6]:
```
linsar <- pspatfit(formlin, data = prod_it,
                   listw = lwsp_it,
                   method = "eigen",
                   type = "sar")
```

[6]:
```
##
## Fitting Model...
##
##  Time to fit the model:  0.88 seconds
```

[7]:
```
summary(linsar)
```

[7]:
```
##
##  Call
## pspatfit(formula = formlin, data = prod_it, listw = lwsp_it,
##     type = "sar", method = "eigen")
##
##  Parametric Terms
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  0.1310379  0.0368036  3.5605 0.0005640 ***
## lnPROD_0    -0.0111188  0.0034779 -3.1970 0.0018495 **
## lnoccgr     -0.1367972  0.0676062 -2.0234 0.0456430 *
## net          0.1087253  0.0259539  4.1892 5.961e-05 ***
## rho          0.3654309  0.0977673  3.7378 0.0003065 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Goodness-of-Fit
##
##  EDF Total:      5
##  Sigma: 0.00310498
##  AIC:  -1138.08
##  BIC:  -1124.72
```

[8]:
```
imp_parvar_sar <- impactspar(linsar, list_varpar)
summary(imp_parvar_sar)
```

[8]:
```
##
##  Total Parametric Impacts (sar)
##            Estimate Std. Error    t value Pr(>|t|)
## lnPROD_0 -0.0179094  0.0066196 -2.7055239   0.0068
```

[5] Preliminary diagnostic tests (using likelihood ratio statistics) work in favor of the SAR model, rather that the SDM and the SEM. Spatial autoregressive models are estimated using a standardized inverse distance W matrix combined with a binary minimum threshold distance matrix.

```
## lnoccgr  -0.2227227  0.1107963 -2.0101994   0.0444
## net       0.1767814  0.0510990  3.4595896   0.0005
##
##  Direct Parametric Impacts (sar)
##             Estimate Std. Error    t value Pr(>|t|)
## lnPROD_0 -0.0116030  0.0037606 -3.0854441   0.0020
## lnoccgr  -0.1450955  0.0691365 -2.0986808   0.0358
## net       0.1147453  0.0275071  4.1714852   0.0000
##
##  Indirect Parametric Impacts (sar)
##             Estimate Std. Error    t value Pr(>|t|)
## lnPROD_0 -0.0063064  0.0034608 -1.8222652   0.0684
## lnoccgr  -0.0776272  0.0493192 -1.5739763   0.1155
## net       0.0620361  0.0299719  2.0698063   0.0385
```

5.2 *Including the Spatial Trend*

As already mentioned, McMillen (2012) and McMillen (2003) stress the importance of considering whether apparent spatial dependence is in fact engendered by model mis-specifications, such as the erroneous inclusion or omission of covariates and the inappropriate functional form of included covariates. Therefore, we extend the SAR model by first including a smooth spatial trend (thus estimating a semiparametric geoadditive SAR model):

$$\gamma_i = \alpha + \rho \sum_{j=1}^{N} w_{ij,N}\gamma_j + \beta \ln y_{i,0} + \delta m_i + \tau \ln(n_i) + \tilde{f}(s_{1i}, s_{2i}) + \epsilon_i.$$

We use the function `pspt()` with 10 knots for each each variable (latitude and longitude of the centroid) to estimate the spatial trend. A model with a smooth spatial trend can also be estimated in R using alternative packages, such as *mgcv*. The novelty of *pspatreg* is to combine this model with the SAR or any other spatial model. The introduction of the spatial trend in the model has some relevant consequences on the parameters of the linear terms. First, the spatial lag parameter ρ decreases from 0.365 (estimated with the linear SAR) to 0.202. Therefore, there is a clear trade-off between controlling for unobserved heterogeneity and the extent of spatial spillover. Also, the parameter associated to the net migration variable diminishes from 0.109 to 0.072 and becomes less significant. This evidence suggests that omitted variables could have generated a bias in the estimates of both OLS linear and pure SAR linear models, which do not include any control for unobserved heterogeneity. Moreover, the marginal impacts do not reveal any more evidence of indirect (spatial spillover) effects of the covariates.

[9]:
```
formgeo <- growth_PROD ~ lnPROD_0+lnoccgr+ net +
  pspt(longitude,latitude, nknots = c(10, 10), psanova = FALSE)
geosar <- pspatfit(formgeo, data = prod_it,
                   listw = lwsp_it,
                   method = "eigen",
                   type = "sar")
```

[9]:
```
##
## Fitting Model...
##
##  Time to fit the model:  8.81 seconds
```

[10]:
```
summary(geosar)
```

[10]:
```
##
##  Call
## pspatfit(formula = formgeo, data = prod_it, listw = lwsp_it,
##     type = "sar", method = "eigen")
##
##  Parametric Terms
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  0.2169096  0.0473358  4.5824 1.430e-05 ***
## lnPROD_0    -0.0191931  0.0045088 -4.2568 4.958e-05 ***
## lnoccgr     -0.0451729  0.0761514 -0.5932   0.55449
## net          0.0727907  0.0415870  1.7503   0.08337 .
```

```
## Xspt.2      -0.0116569  0.0155548 -0.7494   0.45551
## Xspt.3       0.0119923  0.0174543  0.6871   0.49375
## Xspt.4      -0.0149656  0.0204262 -0.7327   0.46561
## rho          0.2017864  0.1127590  1.7895   0.07679 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Non-Parametric Spatio-Temporal Trend
##               EDF
## f(sp1, sp2) 6.228
##
##  Goodness-of-Fit
##
##  EDF Total: 14.2276
##  Sigma: 0.00271168
##  AIC:  -1122.71
##  BIC:  -1084.69
```

[11]:
```
list_varpar <- as.character(names(summary(geosar)$bfixed)[2:4])
eff_parvar <- impactspar(geosar, list_varpar)
summary(eff_parvar)
```

[11]:
```
##
##  Total Parametric Impacts (sar)
##            Estimate Std. Error    t value Pr(>|t|)
## lnPROD_0 -0.0246683  0.0069725 -3.5379541   0.0004
## lnoccgr  -0.0591070  0.0969102 -0.6099146   0.5419
## net       0.0938277  0.0555063  1.6903981   0.0910
##
##  Direct Parametric Impacts (sar)
##            Estimate Std. Error    t value Pr(>|t|)
## lnPROD_0 -0.0196565  0.0045438 -4.3259872   0.0000
## lnoccgr  -0.0478782  0.0766887 -0.6243182   0.5324
## net       0.0748585  0.0424452  1.7636519   0.0778
##
##  Indirect Parametric Impacts (sar)
##            Estimate Std. Error    t value Pr(>|t|)
## lnPROD_0 -0.0050118  0.0038085 -1.3159346   0.1882
## lnoccgr  -0.0112288  0.0240919 -0.4660817   0.6412
## net       0.0189692  0.0184880  1.0260276   0.3049
```

We can plot the estimated spatial trend using the function `plot_sp2d`.

[12]:
```
plot_sp2d(geosar, data = prod_it)
```

[12]:
```
For the output see Figure 2
```

5.3 *Including Other Univariate Smooth Terms*

As a last step in our empirical application, we extend the model by allowing the variables $lnPROD_0$ and *net* to enter smoothly as non-parametric terms. Specifically, we use the function `pspl` with 9 knots for each univariate term:

$$\gamma_i = \alpha + \rho \sum_{j=1}^{N} w_{ij,N}\gamma_j + g_1(\ln y_{i,0}) + g_2(m_i) + \tau \ln(n_i) + \tilde{f}(s_{1i}, s_{2i}) + \epsilon_i.$$

[13]:
```
formgam <- growth_PROD ~ pspl(lnPROD_0, nknots = 9)+
  lnoccgr+ pspl(net, nknots = 9)+
    pspt(longitude,latitude, nknots = c(10, 10), psanova = FALSE)

gamsar <- pspatfit(formgam, data = prod_it,
                   listw = lwsp_it,
                   method = "eigen",
                   type = "sar")
```

[13]:
```
##
## Fitting Model...
##
##  Time to fit the model:  8.76 seconds
```

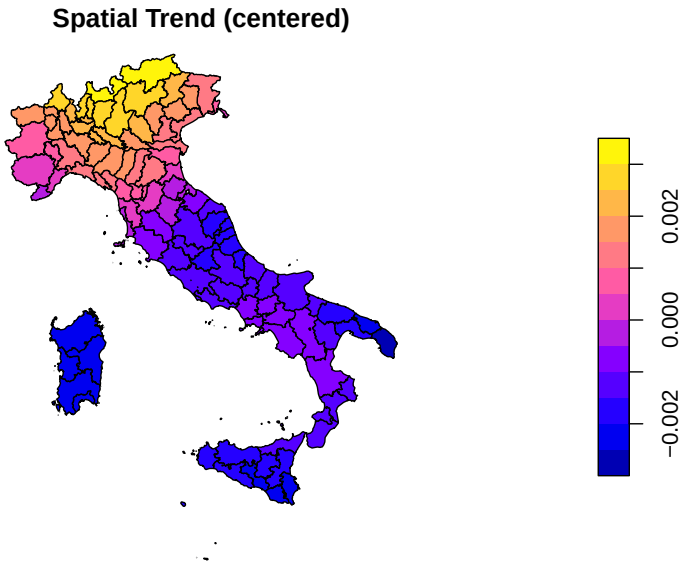


Figure 2: Output from codebox 12

[14]:
```
summary(gamsar)
```

[14]:
```
##
##  Call
## pspatfit(formula = formgam, data = prod_it, listw = lwsp_it,
##     type = "sar", method = "eigen")
##
##  Parametric Terms
##                                 Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                    0.0108138  0.0021183  5.1050 1.754e-06 ***
## lnoccgr                       -0.0349342  0.0752188 -0.4644 0.6434249
## Xspt.2                        -0.0137355  0.0118824 -1.1560 0.2506681
## Xspt.3                         0.0172283  0.0137329  1.2545 0.2127989
## Xspt.4                        -0.0136286  0.0158405 -0.8604 0.3918063
## pspl(lnPROD_0, nknots = 9).1   0.0200607  0.0054356  3.6906 0.0003772 ***
## pspl(net, nknots = 9).1       -0.0052598  0.0031396 -1.6753 0.0972432 .
## rho                            0.1922019  0.1113342  1.7264 0.0876119 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Non-Parametric Terms
##                                EDF
## pspl(lnPROD_0, nknots = 9) 1.0295
## pspl(net, nknots = 9)      1.4016
##
##  Non-Parametric Spatio-Temporal Trend
##               EDF
## f(sp1, sp2) 3.769
##
##  Goodness-of-Fit
##
##  EDF Total: 14.2005
##  Sigma: 0.00268938
##  AIC:  -1126.26
##  BIC:  -1088.31
```

[15]:
```
list_varnopar <- c("lnPROD_0","net")
terms_nopar <- fit_terms(gamsar, list_varnopar)
plot_terms(terms_nopar, prod_it, alpha = 0.10)
```

[15]:
```
For the output see Figure 3
```

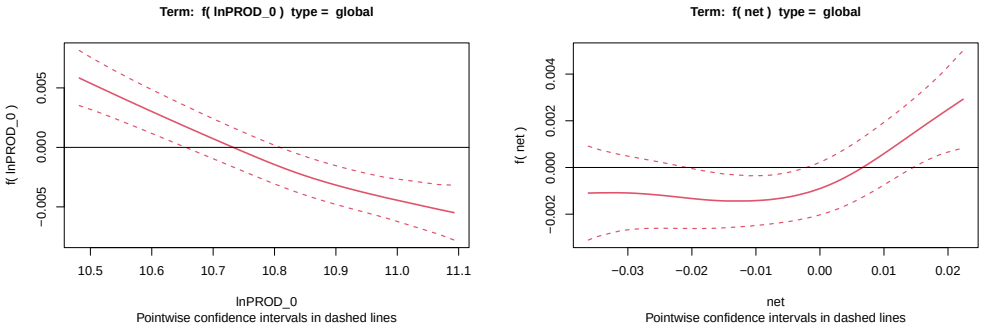


Figure 3: Ouput from codebox 15

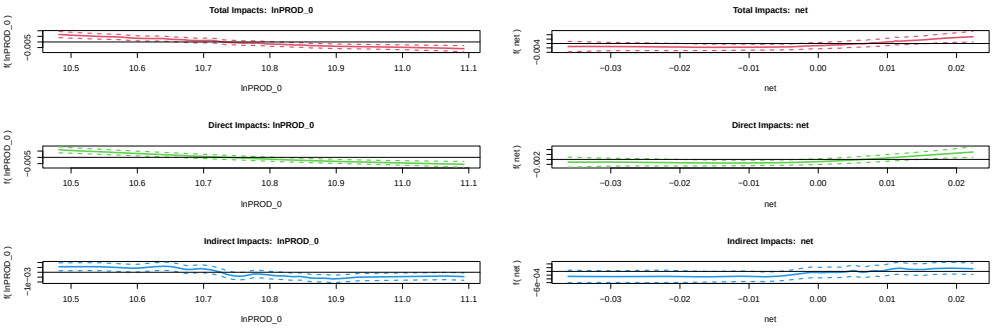


Figure 4: Output from codebox 16

Then, we compute the direct and indirect (or spillover) effects of the two smooth terms in the PS-SAR using the function `impactsnopar`:

[16]:
```
gamsar_impnopar <- impactsnopar(gamsar, listw = lwsp_it,
                                viewplot = FALSE,
                                smooth = FALSE, alpha = 0.1)
plot_impactsnopar(gamsar_impnopar, data = prod_it, smooth = TRUE)
```

[16]:
```
For the output see Figure 4
```

6 Conclusions

This article has highlighted several advantages of using a semiparametric approach over a purely parametric approach to space-time data modeling. Additionally, it has provided a brief introduction to a new `R` package (`psatreg`) that allows estimating this class of models.

The article has also demonstrated the use of this package by using spatial cross-sectional data. This simple application has illustrated the existence of a strong interference between the various problems of mis-specification that characterize the models for spatial data. Specifically, it highlighted the existence of a strong trade-off between spatial dependence and spatial heterogeneity. The inclusion of a spatial trend within a simple SAR model for cross-sectional data (where the lack of degrees of freedom prevents the inclusion of spatial fixed effects) has a strong impact on the magnitude of the spatial spillover parameter (ρ), as well on the magnitude of the other model parameters (β). Other important examples, also for spatio-temporal (i.e. panel) data, are provided by the vignettes included in the package.

We also recognize the existence of limitations of the semiparametric approach for dealing with spatio-temporal data proposed here. An obvious limitation is the difficulty of

these kinds of models to fit data that are characterized by a weak spatial pattern. In this case, while a fixed-effect approach (applied to spatial panel data) is capable of capturing spatial heterogeneity, the inclusion of a spatial trend surface on the r.h.s. of the model hardly captures the effects of omitted variables. However, we also observe that most of the standard economic and social variables show a relevant spatial trend.

We would also like to point out some practical problems associated with the implementation of Spatial Autoregressive Semiparametric Models. In particular, it is well known that nonparametric estimates may be spurious due to outliers, although in the case of penalized splines the effect of the extreme values is often mitigated. In practice, it might be necessary to trim extreme values at the edge of the data domain.

Regarding the problem of model selection, it seems preferable to simply compare the performance of the different models in terms of some Information Criterion. We do not yet provide a battery of diagnostic tests for Spatial Autoregressive Semiparametric Models like the Lagrange Multiplier tests widely used in the traditional parametric spatial econometric literature (LM-SEM,LM-SAR,LM-SARSAR). Indeed, the use and abuse of LM tests for the spatial autocorrelation of the residuals has been largely criticized, as it may lead to a mechanical selection process.

Finally, for future considerations, it is planned to include some functionalities in the package to allow the estimation and inference of spatio-temporal regression models with varying coefficients using P-spline metodology. These models can be seen as an alternative to the usual Geographically Weighted Regression (GWR) models.

References

Bai J, Li K (2013) Spatial panel data models with common shocks. MPRA Paper 52786, University of Munich, Germany

Bailey N, Holly S, Pesaran MH (2016) A two-stage approach to spatio-temporal analysis with strong and weak cross-sectional dependence. *Journal of Applied Econometrics* 31: 249–280. CrossRef

Basile R, Durbán M, Mínguez R, Montero JM, Mur J (2014) Modeling regional economic dynamics: spatial dependence, spatial heterogeneity and nonlinearities. *Journal of Economic Dynamics and Control* 48: 229 – 245. CrossRef

Bivand R (2022) R packages for analyzing spatial data: A comparative case study with areal data. *Geographical Analysis* 54: 488–518. CrossRef

Bivand R, Millo G, Piras G (2021) A review of software for spatial econometrics in R. *Mathematics* 9: 1276. CrossRef

Bivand R, Yu D (2022) *spgwr: Geographically Weighted Regression*. R package version 0.6-35

Bivand RS, Pebesma E, Gomez-Rubio V (2013) *Applied spatial data analysis with R* (Second ed.). Springer, New York

Blundell R, Powell JL (2003) Endogeneity in nonparametric and semiparametric regression models. In: Dewatripont M, Hansen LP, Turnovsky SJ (eds), *Advances in Economics and Econometrics Theory and Application*, Volume II. Cambridge University Press, 312–357. CrossRef

Eilers PH, Marx BD, Durbán M (2015) Twenty years of p-splines. *SORT-Statistics and Operations Research Transactions* 39: 149–186

Elhorst J (2014) *Spatial Econometrics. From Cross-Sectional Data to Spatial Panels*. SpringerBriefs in Regional Science. Springer, Berlin-Heidelberg-New York. CrossRef

Hoshino T (2018) Semiparametric spatial autoregressive models with endogenous regressors: With an application to crime data. *Journal of Business & Economic Statistics* 36: 160–172. CrossRef

Juhl S (2021) spfilteR: An R package for semiparametric spatial filtering with eigenvectors in (generalized) linear models. *The R Journal* 13: 450–459. CrossRef

Kuhn M (2020) *AmesHousing: The Ames Iowa Housing Data*. R package version 0.0.4

Lee DJ, Durban M, Eilers P (2013) Efficient two-dimensional smoothing with P-spline ANOVA mixed models and nested bases. *Computational Statistics & Data Analysis* 61: 22–37. CrossRef

LeSage J, Pace K (2009) *Introduction to Spatial Econometrics*. CRC Press, Boca Raton. CrossRef

Lopez F, Minguez R, Mur J (2020) ML versus IV estimates of spatial SUR models: Evidence from the case of Airbnb in Madrid urban area. *The Annals of Regional Science* 64: 313–347. CrossRef

McMillen DP (2003) Spatial autocorrelation or model misspecification? *International Regional Science Review* 26: 208–217. CrossRef

McMillen DP (2012) Perspectives on spatial econometrics: Linear smoothing with structured models. *Journal of Regional Science* 52: 192–209. CrossRef

Millo G, Piras G (2012) splm: Spatial panel data models in R. *Journal of Statistical Software* 47: 1–38. CrossRef

Mínguez R, Basile R, Durbán M (2020) An alternative semiparametric model for spatial panel data. *Statistical Methods & Applications* 29: 669–708. CrossRef

Montero J, Mínguez R, Durbán M (2012) SAR models with nonparametric spatial trends. A P-spline approach. *Estadística Española* 54: 89–111

Pesaran MH, Tosetti E (2011) Large panels with common factors and spatial correlation. *Journal of Econometrics* 161: 182–202. CrossRef

Piras G (2010) sphet: Spatial models with heteroskedastic innovations in R. *Journal of Statistical Software* 35: 1–21. CrossRef

Shi W, Lee Lf (2018) A spatial panel data model with time varying endogenous weights matrices and common factors. *Regional Science and Urban Economics* 72: 6–34. CrossRef

Vega SH, Elhorst JP (2016) A regional unemployment model simultaneously accounting for serial dynamics, spatial dependence and common factors. *Regional Science and Urban Economics* 60: 85–95. CrossRef

Wood SN (2017) *Generalized Additive Models: An Introduction with R* (2 ed.). Chapman and Hall/CRC. CrossRef

© 2022 by the authors. Licensee: REGION – The Journal of ERSA, European Regional Science Association, Louvain-la-Neuve, Belgium. This article is distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).