The use of R for spatial econometrics

Roger Bivand

19 April 2023, 16:00 CEST

Webinar aims

- To provide participants with an overview of one of the three kinds of modelling with spatial data, namely areal or lattice data modelling
- It will become clear that these kinds of data may be characterised by spatial autocorrelation
- On the one hand, information leaking between neighbouring spatial entities needs to be taken into account
- On the other, we will see that looking carefully at spatial entities, and understanding that such spillovers can occur, may lead us to clearer analysis

Webinar learning outcomes

- Being able to place spatial econometrics in a broader context of modelling with spatial data
- Knowing the most common models proposed by spatial econometrics
- Knowing which R packages provide these models
- Understanding the concepts of support, spatial autocorrelation, and how they may interact when modelling with spatial data

Webinar contents

- What is spatial econometrics? How does it relate to econometrics and to other fields modelling with spatial data?
- Which estimation methods are used in spatial econometrics, which are specific to spatial econometrics, and which shared with proximate fields?
- How are spatial (and spatio-temporal) data represented in R packages, and which packages provide implementations of relevant estimation methods?
- Boston housing value data set: case of trying to study a problem when the support of the data probably does not match the problem

Spatial econometrics

Spatial econometrics

- Anselin (2010) indicates clearly and repeatedly (Anselin 1988) that we should acknowledge *Spatial Econometrics* by Paelinck and Klaassen (1979) of the Netherlands Economic Institute as our starting point (see also Hordijk (1974) and Hordijk and Paelinck (1976))
- In a short commentary, Paelinck (2013) recalls his conviction, expressed in 1967, that "early econometric exercises ... relating only variables possessing the same regional index ... were inadequate to represent the correct spatial workings of the economy, which would then be reflected in the policy outcomes."
- Central government expenditure in region i could spill over into income and consumption in other regions, through labour market and interregional trade channels

Statistical maps

- Two statisticians, Moran (1948) and Geary (1954), had proposed measures that began to address the need to infer from maps
- Geary's measure was followed up by Duncan, Cuzzort, and Duncan (1961) in *Statistical geography: Problems in analyzing areal data*, where they point to issues raised by the modifiable nature of spatial units used for collecting and analysing information (modifiable areal unit problem, MAUP)
- "Sooner or later in a study of areal variation the investigator runs up against the fact that areal units situated close to each other are more likely to be similar in their characteristics than are areal units which are some distance apart ..." (pp. 128–9)
- and heterogeneity — for example upper level units "breaking up" the smooth surface of lower level units

Spatial autocorrelation

- A. D. Cliff and Ord (1969) generalised the way in which neighbours could be defined as a spatial weights matrix (see Ch. 14), and in *Spatial Autocorrelation* (A. D. Cliff and Ord 1973) set out the framework for global measures of spatial autocorrelation
- Ord (2010) reflects on their legacy, expressing doubt that the serious points raised by Granger (1969) (and noted by Ripley (1988)) had been addressed adequately
- Another early summary (Hepple (1974)) shows how much had already been grasped, including the impact of spatial autocorrelation on multivariate analysis
- Finally, Tobler (1970) proposed a “first law” of geography, immediately criticised by Olsson (1970) for over-reaching (see Ch. 15)

Spatial autocorrelation in regression residuals

- Using the tools created to examine spatial autocorrelation, it became possible to extend to regression residuals
- A. Cliff and Ord (1972) provided an extension of Moran's I to regression residuals, followed by Hordijk (1974)
- It was long felt that the omission of special (spatial) treatment for models using spatial data invalidated inferences made
- In a careful study, Smith and Lee (2012) show that inferences are not affected only when covariates are not spatially autocorrelated

Spatial econometrics or spatially structured random effects?

- From the mid 1970's, two traditions developed, one handling the effects of spatial autocorrelation in modelling in ways analogous to time series, the other adding spatially structured random effects to models
- The latter was proposed by Besag (1974), and has been widely adopted in spatial epidemiology (disease mapping) and spatial ecology, as an effective way of including the unobserved spatial process
- Both Besag (1974) and A. D. Cliff and Ord (1973) reach back to Whittle (1954), but the subsequent developments of conditional autoregression models (CAR, spatially structured random effects) and simultaneous (joint) autoregression models (SAR, spatial econometrics), have diverged. Ord takes this up in his discussion of Besag (1974), page 229 (see also Ch. 16)

We can represent a simple modelling situation in the following way:

$$\textbf{data} = \textbf{smooth} + \textbf{rough}$$

where the $\textbf{rough}$ are taken to have no remaining patterning information. If, on the other hand, useful information remains in the $\textbf{rough}$, for example with discernable spatial patterning, we can try to retrieve it:

$$\textbf{data} = \textbf{smooth} + \textbf{spatial smooth} + \textbf{rough}$$

This is useful both for predictions from $\textbf{smooth} + \textbf{spatial smooth}$, and possibly less biassed inference from the $\textbf{smooth}$.

Spatial smooth: spatially structured random effects

- The spatially structured random effects literature is very rich, and now expresses the **spatial smooth** in the context of linear mixed models (LMM)
- This can be extended to generalised linear mixed models (GLMM) and to multi-level models
- The spatial structuring is typically described as by a Markov Random Field (MRF) term added to the model, either with a parametric or intrinsic conditional autoregressive form; the MRF is expressed through a graph of 0/1 neighbours
- The output includes an estimate of the random effect for each observation — which may be mapped, and an expression of the distribution around those estimates

Spatial smooth: spatial lag model

In spatial econometrics, the **spatial smooth** term is not as simple.

The spatial lag model (SLM, a.k.a SAR) is the most frequently encountered specification in spatial econometrics:

$$\mathbf{y} = \rho_{\text{Lag}} \mathbf{W}\mathbf{y} + \mathbf{X}\beta + \varepsilon,$$

where $\mathbf{y}$ is an $(N \times 1)$ vector of observations on a dependent variable taken at each of N locations, $\mathbf{W}$ is a fixed $(N \times N)$ spatial weights matrix, $\mathbf{X}$ is an $(N \times k)$ matrix of exogenous variables, β is an $(k \times 1)$ vector of parameters, ε is an $(N \times 1)$ vector of independent and identically distributed disturbances and ρ_{Lag} is a scalar spatial lag parameter.

The **spatial smooth** term is $\rho_{\text{Lag}} \mathbf{W}\mathbf{y}$.

Spatial smooth: spatial Durbin model

In the spatial Durbin model (SDM), the spatially lagged exogenous variables are added to the model; spatial Durbin models are reviewed by Mur and Angulo (2006):

$$\mathbf{y} = \rho_{\mathrm{Lag}} \mathbf{W}\mathbf{y} + \mathbf{X}\beta + \mathbf{W}\mathbf{X}\gamma + \varepsilon,$$

where γ is an $((k-1) \times 1)$ vector of parameters where $\mathbf{W}$ is row-standardised (all rows sum to unity), and a $(k \times 1)$ vector otherwise.

The $\textbf{spatial smooth}$ term is $\rho_{\mathrm{Lag}} \mathbf{W}\mathbf{y} + \mathbf{W}\mathbf{X}\gamma$.

LeSage and Pace (2009) show that these models share a complicated data generation process:

$$\mathbf{y} = (\mathbf{I} - \rho_{\mathrm{Lag}}\mathbf{W})^{-1}(\mathbf{X}\beta + \mathbf{W}\mathbf{X}\gamma) + \varepsilon.$$

in which ρ_{Lag} and β (and possibly γ) interact. These measures of the effects of each included covariate need to be estimated in addition to fitting the model

The spatial error model (SEM) may be written as Ord (1975) or Hepple (1976):

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}, \qquad \mathbf{u} = \rho_{\mathrm{Err}}\mathbf{W}\mathbf{u} + \varepsilon,$$

and $\mathbf{u}$ is a spatially autocorrelated disturbance vector with constant variance σ^2 and covariance terms specified:

$$\mathbf{u} \sim N(0, \sigma^2(\mathbf{I} - \rho_{\mathrm{Err}}\mathbf{W})^{-1}(\mathbf{I} - \rho_{\mathrm{Err}}\mathbf{W}^\top)^{-1})$$

What to do about time?

Spatio-temporal models in the spatially structured random effects branch are just (G)LMM with a added temporal random effect. Non-separability between time and space remains a problem, but a lot can be achieved, see Blangiardo and Cameletti (2015) and Gómez-Rubio (2020)

In the spatial econometrics branch, Elhorst (2003) presents the extension of panel econometrics to spatial panel data (see also Elhorst (2014)). In extending to time panels, a range of combined models has also come into being, a general nested model (GNM) nesting all the others, a model without spatially lagged covariates (SARAR). If neither the residuals nor the response are modelled with spatial processes, spatially lagged covariates may be added to a linear model, as a spatially lagged X model (SLX) (LeSage 2014; Halleck Vega and Elhorst 2015). We can write the GNM as (here a cross-sectional model for simplicity):

$$\mathbf{y} = \rho_{\mathrm{Lag}}\mathbf{W}\mathbf{y} + \mathbf{X}\beta + \mathbf{W}\mathbf{X}\gamma + \mathbf{u}, \qquad \mathbf{u} = \rho_{\mathrm{Err}}\mathbf{W}\mathbf{u} + \varepsilon.$$

Which branch?

- The literature covering both the development and especially the use of spatially structured random effects (SSRE) is vast, and they have few problems with limited dependent variables
- The literature on the specification and development of spatial econometrics models (including spatial panel models) is large, but usage is limited, not least because of the need to choose between model specifications; only this branch may open for instrumenting endogeneous covariiates
- Both use the same specifications defining neighbours of observations, but spatial econometrics models most often use row standardised spatial weights, and SSRE most often use binary spatial weights (and require symmetric weights in a single graph)

Estimation methods for spatial econometrics

Estimation methods for spatial econometrics

- The estimation methods first introduced by Ord (1975) and Hepple (1976) used maximum likelihood (ML); this was followed up by Anselin (1988)
- Bayesian methods are reviewed by LeSage and Pace (2009)
- Generalised method of moments (GMM) methods are reviewed by Kelejian and Piras (2017), based on earlier work by (Kelejian and Prucha 1998, 1998)
- Other methods like the Conley (1999) approach are clearly spatial econometrics, but are not often discussed in the same way (no spatial weights matrices are usd)
- Please see R. Bivand, Millo, and Piras (2021) and R. Bivand and Piras (2015) for summaries of ML and GMM approaches; the former also covers spatial panel models

Maximum likelihood

The log-likelihood function for the spatial lag model is:

$$\ell(\beta, \rho_{\text{Lag}}, \sigma^2) = -\frac{N}{2}\ln 2\pi - \frac{N}{2}\ln \sigma^2 + \ln|\mathbf{I} - \rho_{\text{Lag}}\mathbf{W}|$$
$$-\frac{1}{2\sigma^2}[((\mathbf{I} - \rho_{\text{Lag}}\mathbf{W})\mathbf{y} - \mathbf{X}\beta)^\top((\mathbf{I} - \rho_{\text{Lag}}\mathbf{W})\mathbf{y} - \mathbf{X}\beta)]$$

and by extension the same framework is used for SDM when $[\mathbf{X}(\mathbf{W}\mathbf{X})]$ are grouped together. The sum-of-squared errors (SSE) term in the square brackets is found using auxilliary regressions $\mathbf{e} = \mathbf{y} - (\mathbf{X}^\top\mathbf{X})\mathbf{X}\mathbf{y}$ and $\mathbf{u} = \mathbf{W}\mathbf{y} - (\mathbf{X}^\top\mathbf{X})\mathbf{X}\mathbf{W}\mathbf{y}$, and $SSE = \mathbf{e}^\top\mathbf{e} - 2\rho_{\text{Lag}}\mathbf{u}^\top\mathbf{e} + \rho_{\text{Lag}}^2\mathbf{u}^\top\mathbf{u}$. The cross-products of $\mathbf{u}$ and $\mathbf{e}$ can conveniently be calculated before line search (univariate non-linear optimisation) begins.

Log determinant

The first published versions of the eigenvalue method for finding the log determinant Ord (1975) is:

$$\ln(|\mathbf{I} - \rho\mathbf{W}|) = \sum_{i=1}^{N} \ln(1 - \rho\zeta_i)$$

where ζ_i are the eigenvalues of $\mathbf{W}$. One specific problem addressed by Ord (1975) is that of the eigenvalues of the asymmetric row-standardised matrix $\mathbf{W}$ with underlying symmetric neighbour relations $c_{ij} = c_{ji}$. If we write $\mathbf{w} = \mathbf{C1}$, where $\mathbf{1}$ is a vector of ones, we can get: $\mathbf{W} = \mathbf{CD}$, where $\mathbf{D} = \mathrm{diag}(1/\mathbf{w})$; by similarity, the eigenvalues of $\mathbf{W}$ are equal to those of: $\mathbf{D}^{\frac{1}{2}}\mathbf{C}\mathbf{D}^{\frac{1}{2}}$. From R. K. Pace and Barry (1997), sparse Cholesky and sparse LU alternatives were available for cases in which finding the eigenvalues of a large weights matrix would be impracticable. R. Bivand, Hauke, and Kossowski (2013) describe the available alternatives.

Bayesian methods

- LeSage and Pace (2009) and their earlier and later work form the foundation for Markov chain Monte Carlo (MCMC) approaches
- Griddy Gibbs sampling from a spline smooth of values of LU decomposition-based log determinants are used for spatial process coefficients
- Gómez-Rubio, Bivand, and Rue (2021) describe the use of a new experimental latent model `"slm"` in INLA (integrated nested Laplace approximation), complementing many existing latent models for spatial regression
- The work presented by LeSage and Pace (2009) is further documented by Matlab code, which is often used for comparison http://www.spatial-econometrics.com/

- R. Bivand, Millo, and Piras (2021) and R. Bivand and Piras (2015) review and summarise GMM approaches to estimation
- These methods handle the spatially lagged response $\mathbf{Wy}$ by taking $\mathbf{WX}$ and $\mathbf{WWX}$ as instruments
- The spatially lagged error term is handled by non-linear optimisation; both of these choices remove the need to handle the log determinant term
- GMM can also handle RHS endogeneous covariates by the use of instrumental variables

R packages implementing spatial econometrics methods

R packages implementing spatial econometrics methods i

- Navigating through the R package ecosystem is not easy; Joo et al. (2020) make a thorough attempt to track down packages for time-space movement data
- Task Views are the mechanism proposed twenty years ago when there were many fewer contributed packages
- Zeileis, McDermott, and Tappe (2023) maintain the Econometrics task view and mention spatial regression
- R. Bivand and Nowosad (2023) maintain the Spatial task view, which covers handling, mapping and analysing spatial data, and also mention spatial regression

R packages implementing spatial econometrics methods ii

- Pebesma and Bivand (2022) maintain the SpatioTemporal task view and mention spatio-temporal regression
- None of the task views concentrates on spatial econometrics, so perhaps review and comparison articles may assist
- Also note that acceptance on the Comprehensive R Archive Network (CRAN) only certifies that the package meets general standards for packages (licence declared, code runs examples and tests, functions are minimally documented), it does not confirm that packages do what they claim to do
- If there is a JSS or other published article subject to substantive peer review, one can be more confident on this point

R packages implementing spatial econometrics methods iii

- R. Bivand, Millo, and Piras (2021) summarise and present central R packages for spatial econometrics: **spatialreg** for ML (R. Bivand and Piras 2022, bold are R package names), **sphet** for GMM (see Piras (2010), Piras (2022)), and **splm** (Millo and Piras 2012, 2022), building on **plm** (Croissant, Millo, and Tappe 2022), for spatial panel models, (see Millo and Piras (2012), Croissant and Millo (2008), Millo (2017) and Croissant and Millo (2018))
- These packages are also tightly integreted in the use of the same estimation methods for the log determinant in ML estimation, and sharing infrastructure to estimate impacts; see also chapter 17 in Pebesma and Bivand (2023)

R packages implementing spatial econometrics methods iv

- R. Bivand, Millo, and Piras (2021) and chapter 16 in Pebesma and Bivand (2023) also follow R. Bivand et al. (2017), which was provoked by work with Osland, Thorsen, and Thorsen (2016) on multi-level models, and the now-archived **HSAR** package (Dong, Harris, and Mimis (2020))
- In chapter 14 in Pebesma and Bivand (2023), the use of **spdep** (R. Bivand 2023), used to create spatial neighbour objects, and from these spatial weights objects is presented
- If we see which other packages use this functionality in **spdep**, we can extend the scope of packages engaging with broadly understood spatial econometrics

R packages implementing spatial econometrics methods v

- There are six packages in small area estimation: **emdi** (Harmening et al. 2022), **saeRobust** (Warnholz 2023), **saeSim** (Warnholz and Schmid 2022), **tipsae** (De Nicolò and Gardini 2023), **mcmcsae** (Boonstra 2023) and **SUMMER** (Li et al. 2022)
- Some of these are also mentioned in the Official Statistics task view (Templ, Kowarik, and Schoch 2022)
- Apart from these, there are many other relevant packages in application areas close to spatial econometrics; note overlaps between package authors showing something of the contributed package ecosystem network. For references to underlying methods, see the packages' documentation

- **ssfa** (Fusco and Vidoli 2022) provides functions for spatial stochastic frontier analysis among a number of SFA packages noted in the Econometrics task view
- Heterogeneity is approached in **SpatialRegimes** (Vidoli and Benedetti 2022) and **hspm** (Piras and Sarrias 2022); **conleyreg** (Düben 2022; Conley 1999) provides a selection of high-performance spatially-clustered residual methods
- **spsur** (Angulo et al. 2022) and **pspatreg** (Minguez et al. 2022) contain spatial seemingly unrelated and semiparametric regression models; **spqdep** (Lopez et al. 2022) is from some of the same team and implements a number of tests for categorical data
- **SDPDmod** (Simonovska 2022) is a recent package for spatial dynamic panel data extending **splm**

R packages implementing spatial econometrics methods vii

- **spmoran** (Murakami 2022) provides modern extensions to `spatialreg::SpatialFiltering()` for spatial filtering, the addition of selected eigenvectors of the doubly-centred spatial weights matrix to "wash" spatial dependence from the residuals
- **McSpatial** (McMillen 2013a) will hopefully re-appear on CRAN and provides code for McMillen (2013b), for quantile regression for spatial data, and early GMM methods for limited dependent variables. **spldv** (Sarrias and Piras 2022) is a recent package for limited dependent variables, while **spatialprobit** (Wilhelm and de Matos 2022) fits models for limited dependent variables using MCMC following LeSage and Pace (2009), and **ProbitSpatial** (Martinetti and Geniaux 2021) uses the approximate value of the true likelihood of spatial probit models for fast estimation

- **spflow** (Dargel and Laurent 2021) provides origin-destination spatial models and **spnaf** (Y. Lee et al. 2022) spatial network models
- There are very many simulation-based (MCMC and other sampling schemes) packages, both specialised: **CARBayes** (D. Lee 2022) for conditional autoregressive models typically for disease mapping, and general packages permitting the use of MRF spatially structured random effects: **geostan** (Donegan 2022), **R2BayesX** (Umlauf, Kneib, et al. 2022), **brms** (Bürkner 2022) and **bamlss** (Umlauf, Klein, et al. 2022), using models stemming from WinBUGS and GeoBUGS; many are listed in the Bayesian task view (Jong Hee Park 2022)

R packages implementing spatial econometrics methods ix

- The **INLA** package is maintained outside CRAN, but can be installed and updated using similar mechanaisms (Rue, Lindgren, and Teixeira Krainski 2022). CRAN packages including **INLABMA** (Gómez-Rubio and Bivand 2018), **bigDM** (Adin, Orozco-Acosta, and Ugarte 2022), **inlabru** (Lindgren and Bachl 2022) and **DClusterm** (Gomez-Rubio, Serrano, and Rowlingson 2020) use INLA models for fitting spatial and spatio-temporal models (Gómez-Rubio and Palmí-Perales 2019; Blangiardo and Cameletti 2015; Gómez-Rubio 2020)
- Spatial generalised additive models of various kinds can also be estimated using **gamlss.spatial** (De Bastiani, Stasinopoulos, and Rigby 2018), and the MRF smooth in **mgcv** (Wood 2022)
- **lagsarlmtree** (Wagner and Zeileis 2019) inserts `spatialreg::lagsarlm()` into a **partykit** tree-structured regression model framework

R packages implementing spatial econometrics methods x

- In the training/testing paradigm, **waywiser** (Mahoney 2022) provides a number of ways of assessing predictive models of spatial data, among others using **spatialsample** (Silge and Mahoney 2023) for spatial resampling **mlr3spatiotempcv** (Schratz and Becker 2022); **blockCV** (Valavi et al. 2023) also provides spatial resampling, and **CAST** (Meyer, Milà, and Ludwig 2023) uses **caret** models incorporating very important recent results reported by Milà et al. (2022)

Support case: willingness to pay for air pollution mitigation

Is the choice of model specification the only problem? i

- In practical introductions to spatial econometrics, such as Arbia (2014), Anselin and Rey (2014), Elhorst (2014), and recently Kopczewska (2020), it may appear to the reader that the choice of model specification is the key step between data and results
- I have no excuse, having also many convictions for stressing model specification since R. S. Bivand (1984); it does remain vital
- However, the data on which model estimation are based are equally vital, as some common steps may unwittingly create problems that we subsequently seem to need special methods to overcome
- The analysis of areal aggregates are particularly prone to a range of entitation problems (Wilson 2000, 2002)

Is the choice of model specification the only problem? ii

- not only the dreaded MAUP (Gelfand 2010)
- the ecological fallacy (Wakefield and Lyons 2010)
- and change of support more generally (Gotway and Young 2002)
- see Do, Thomas-Agnan, and Vanhems (2015) and Do, Laurent, and Vanhems (2021) for reviews of areal interpolation methods

Boston housing values hedonic model

- Harrison and Rubinfeld (1978b) made a serious and thorough attempt to use census data observed at the census tract level to try to establish willigness to pay (WTP) for air pollution abatement in Boston (Harrison and Rubinfeld 1978a, 1978c)
- Their data set was published in Belsley, Kuh, and Welsch (1980), a book on regression diagnostics, and began to be used widely, including provision from Newman et al. (1998), available as R package **mlbench** (Leisch and Dimitriadou 2021); it is also avaailable from Statlib http://lib.stat.cmu.edu/datasets/
- Gilley and Pace (1996) provided a corrected dataset, pointing out that the median housing value variable is, in fact, censored
- R. Kelley Pace and Gilley (1997) added coordinates giving the relative locations of the tracts, and established that the residuals of the original hedonic regression were autocorrelated, affecting the willingness to pay estimates

Acronym soup and SAS

- The acronym soup of SLX/SLM/SAR/SEM/SDM/SDEM/SARAR/SAC/SADC/... also reaches SAS documentation, in two blogs from 2021, https://blogs.sas.com/content/subconsciousmusings/2021/03/02/spatial-econometric-modeling-unleashes-the-geographic-potential-of-your-data/ and https://blogs.sas.com/content/subconsciousmusings/2021/08/09/automate-spatial-regression-model-selection-using-proc-cspatialreg/
- Both of these use the Boston data set, but just focus on mapping and fitting standard spatial econometrics models to a subset of the covariates (omitting the air pollution measure
- They also use 1970 tract boundaries without describing how they were generated, and without taking up the challenges of the data set, not even mentioning that the response is censored

Boston housing values hedonic model

- R. Bivand (2017) is based on access to historical online census data, both for the boundaries of the tracts, and for analysis of the census-based covariates and response variable
- The response was the weighted median of counts of responses to a self-assessed item in the 1970 census: If you live in a one-family house which you own or are buying - what is the value of this property? That is, how much do you think this property (house and lot) would sell for if it were for sale?

Censoring and exclusion

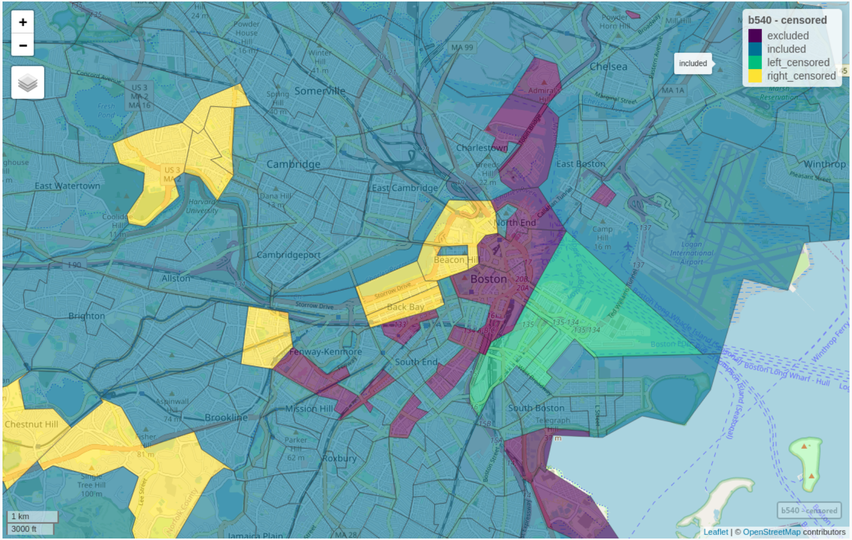
- The values were < $ 5000, … >= $ 50000, with 9 intervening unequal intervals; this is why a weighted median was used to calculate reported tract median values
- Some urban tracts had no such properties and were omitted, others had median values of $ 5000 (left censored) and $ 50000 (right censored)
- Even for tracts with assessed properties, the property counts varied greatly between tracts (minimum 5, median 511, maximum 3031); case weights were considered but not used

Starting the examples

```r
library(sf)
b540 <- st_read("data/bo_540_df4.shp", quiet=TRUE)
b540$censored <- rep("included", nrow(b540))
b540$censored[is.na(b540$CMEDV)] <- "excluded"
b540$censored[b540$CMEDV == 5 & is.na(b540$median)] <- "left_censored"
b540$censored[b540$CMEDV == 50 & is.na(b540$median)] <- "right_censored"
table(b540$censored)
```

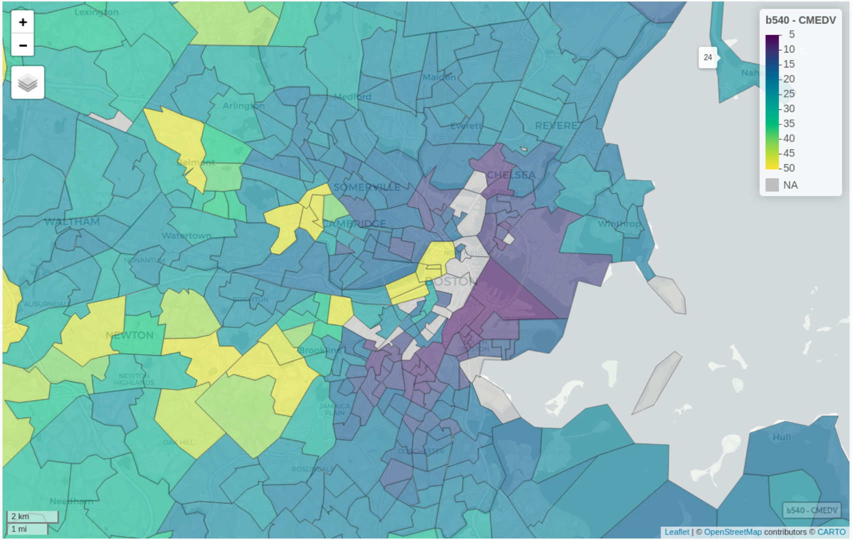
```
##
##       excluded       included  left_censored right_censored
##             34            489              2             15
```

Where are the drop-outs?



Harbor area Labor Day 1978

Naive tract median housing values



How was air pollution measured to use as a covariate?

- Use was made of the Transportation and Air Shed SImulation Model (TASSIM) (Ingram and Fauth 1974; Ingram, Fauth, and Kroch 1974)
- This generated output not from measurement of actual air pollution in 1970, but rather predictions from point-source polluters (mostly near the port), and from major highways through meteorological models
- The predictions were reported for 122 model output zones extending beyond the parts of the Boston SMSA used for the WTP study
- The model output zones appear to roughly coincide with towns - administrative districts, of which there are 92 in the 506 tract dataset, 15 in Boston itself

What is the support of the key WTP covariate?

We'll reconstruct the data objects used in Pebesma and Bivand (2023) chapters 16-17 (refer to these for details), and the data set as provided in the **spData** package (R. Bivand, Nowosad, and Lovelace 2022), and use them here.

The number of unique values of the NOX variable in the data set is well below 506, the number of tracts in the original data set

```
length(unique(boston_506$NOX))
```

```
## [1] 81
```

Spatial autocorrelation: tracts vs. model output zones

This indicates that the tract values were copied to tracts intersecting the model output zones; however, strong positive spatial autocorrelation was present in the model output zones already, as is only reasonable:

Tract level NOX spatial autocorrelation

```
spdep::moran.test(boston_506$NOX, lw_q_506) |> glance_htest()
```

```
## Moran I statistic       Expectation          Variance        Std deviate
##      9.065225e-01     -1.980198e-03      7.271085e-04       3.369199e+01
##           p.value
##     3.786993e-249
```

Spatial autocorrelation: tracts vs. model output zones

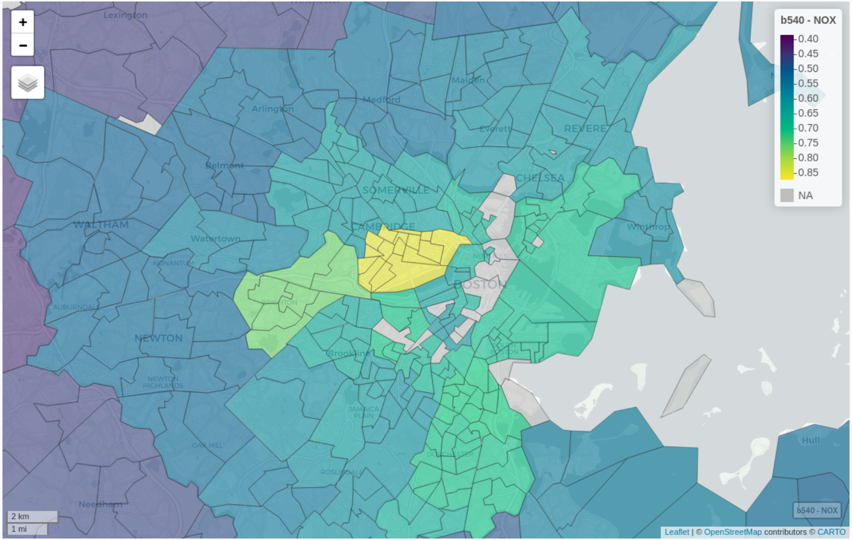
Model output zone level NOX spatial autocorrelation

```
spdep::moran.test(boston_93$NOX, lw_q_93) |> glance_htest()
```

```
## Moran I statistic       Expectation          Variance        Std deviate
##      7.578171e-01     -1.086957e-02      4.669481e-03      1.124903e+01
##           p.value
##      1.170779e-29
```

The model output forms an uneven cone declining with distance from the central business district and harbour, so autocorrelation when reported for neighbouring entities on the surface of the cone is to be expected

Naive tract NOX air pollution values



The SAS blogs use a subset of the actual covariates used by Harrison and Rubinfeld (1978b). For the chosen covariates and 506 census tracts, the coefficient values agree in the linear model case:

```
coef(lm(log(CMEDV) ~ log(PTRATIO) + log(LSTAT), data=boston_506))
```

```
##  (Intercept) log(PTRATIO)   log(LSTAT)
##    6.0247877   -0.6122500   -0.5102814
```

and in the spatial error model case (using `spatialreg::errorsarlm()` and pre-computing the spatial weights object eigenvalues):

Are our results the same as SAS? ii

```r
e <- spatialreg::eigenw(lw_q_506)
coef(spatialreg::errorsarlm(log(CMEDV) ~ log(PTRATIO) + log(LSTAT),
    data=boston_506, listw=lw_q_506, control=list(pre_eig=e)))
```

```
##        lambda   (Intercept) log(PTRATIO)   log(LSTAT)
##     0.7630700     5.1150187   -0.3880317   -0.4100675
```

Finally, the AIC of the general nested model also agrees (listed as SDAC in the blog):

```r
AIC(spatialreg::sacsarlm(log(CMEDV) ~ log(PTRATIO) + log(LSTAT),
    data=boston_506, listw=lw_q_506, Durbin=TRUE,
    control=list(pre_eig1=e, pre_eig2=e)))
```

```
## [1] -377.6966
```

Are our results the same as SAS? iii

Using the SAS covariates, the 506 tract data set, and ignoring censoring, the GNM would also be chosen as the best alternative by AIC.

Omitting the censored tracts i

We pre-compute the eigenvalues of the 487 tract dataset, and specify the same covariates as were used in the original article (any changes are noted in R. Bivand (2017))

```
e <- spatialreg::eigenw(lw_q_487)
f <- formula(log(median) ~ I(RM^2) + AGE + log(DIS) + log(RAD) + TAX +
        PTRATIO + I(BB/100) + log(I(LSTAT/100)) + CRIM + ZN + INDUS +
        CHAS + I((NOX*10)^2))
```

Omitting the censored tracts ii

Omitting the censored tracts creates no-neighbour observations, which can be accommodated here using the `zero.policy=` argument; the GNM, SDEM and SLX are estimated (`CHAS` is a categorical variable, for which the spatial lag is not well understood, and is here omitted from the Durbin term):

```
GNM_487 <- spatialreg::sacsarlm(f, data=boston_487, listw=lw_q_487,
    zero.policy = TRUE, Durbin=update(f, ~ . - CHAS),
    control=list(pre_eig1=e, pre_eig2=e))
SDEM_487 <- spatialreg::errorsarlm(f, data=boston_487, listw=lw_q_487,
    zero.policy = TRUE, Durbin=update(f, ~ . - CHAS),
    control=list(pre_eig=e))
SLX_487 <- spatialreg::lmSLX(f, data=boston_487, listw=lw_q_487,
    zero.policy = TRUE, Durbin=update(f, ~ . - CHAS))
```

Omitting the censored tracts iii

Performing likelihood ratio tests, the most complex model is preferred:

```
options(show.signif.stars=FALSE)
o <- lmtest::lrtest(GNM_487, SDEM_487)
attr(o, "heading")[2] <- "GNM_487 vs. SDEM_487"
o
```

```
## Likelihood ratio test
##
## GNM_487 vs. SDEM_487
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1  29 311.45
## 2  28 307.65 -1 7.5901   0.005869
```

while the SDEM is clearly preferred before SLX:

```
o <- lmtest::lrtest(SDEM_487, SLX_487)
attr(o, "heading")[2] <- "SDEM_487 vs. SLX_487"
o
```

```
## Likelihood ratio test
##
## SDEM_487 vs. SLX_487
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1  28 307.65
## 2  27 226.96 -1 161.38  < 2.2e-16
```

R. Kelley Pace and LeSage (2008) propose a test for SEM/SDEM models to check that the fitted coefficient values are close enough to the equivalent linear models; here they are not, SDEM is not well specified:

```
spatialreg::Hausman.test(SDEM_487)
```

```
##
##  Spatial Hausman test (asymptotic)
##
## data:  NULL
## Hausman test = 52.257, df = 26, p-value = 0.001674
```

Fitting models for the model output zones i

Once again we fit three models including the spatially lagged continuous covariates:

```
e <- spatialreg::eigenw(lw_q_94)
GNM_94 <- spatialreg::sacsarlm(f, data=boston_94, listw=lw_q_94,
    zero.policy = TRUE, Durbin=update(f, ~ . - CHAS),
    control=list(pre_eig1=e, pre_eig2=e))
SDEM_94 <- spatialreg::errorsarlm(f, data=boston_94, listw=lw_q_94,
    zero.policy = TRUE, Durbin=update(f, ~ . - CHAS),
    control=list(pre_eig=e))
SLX_94 <- spatialreg::lmSLX(f, data=boston_94, listw=lw_q_94,
    zero.policy = TRUE, Durbin=update(f, ~ . - CHAS))
```

and test GNM versus SDEM (GNM does not fit better than SDEM):

```
o <- lmtest::lrtest(GNM_94, SDEM_94)
attr(o, "heading")[2] <- "GNM_94 vs. SDEM_94"
o
```

```
## Likelihood ratio test
##
## GNM_94 vs. SDEM_94
##   #Df LogLik Df Chisq Pr(>Chisq)
## 1  29 81.164
## 2  28 81.163 -1 5e-04     0.9831
```

Fitting models for the model output zones iii

then SDEM versus SLX (SDEM does not fit better than SLX):

```
o <- lmtest::lrtest(SDEM_94, SLX_94)
attr(o, "heading")[2] <- "SDEM_94 vs. SLX_94"
o
```

```
## Likelihood ratio test
##
## SDEM_94 vs. SLX_94
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1  28 81.163
## 2  27 81.106 -1 0.1149     0.7347
```

and finally SLX versus a linear model without spatially lagged continuous covariates (SLX does fit better than LM):

```
LM_94 <- lm(f, data=boston_94)
o <- lmtest::lrtest(SLX_94, LM_94)
attr(o, "heading")[2] <- "SLX_94 vs. LM_94"
o
```

```
## Likelihood ratio test
##
## SLX_94 vs. LM_94
##   #Df LogLik  Df  Chisq Pr(>Chisq)
## 1  27 81.106
## 2  15 58.452 -12 45.308  9.124e-06
```

The linear model does show some spatial autocorrelation in its residuals:

```
spdep::lm.morantest(LM_94, listw=lw_q_94) |> glance_htest()
```

```
## Observed Moran I      Expectation         Variance       Std deviate
##      0.087773954     -0.051758503      0.003826674       2.255612673
##          p.value
##      0.012047449
```

Fitting models for the model output zones vi

but this is reduced in the residuals of the SLX model:

```
spdep::lm.morantest(SLX_94, listw=lw_q_94) |> glance_htest()
```

```
## Observed Moran I       Expectation          Variance       Std deviate
##       0.018425326      -0.077139644       0.003767882       1.556861818
##           p.value
##       0.059751648
```

The Hausman test does not find differences between the regression coefficients of the SLX and SDEM models:

```
spatialreg::Hausman.test(SDEM_94)
```

```
##
##  Spatial Hausman test (asymptotic)
##
## data:  NULL
## Hausman test = 3.175, df = 26, p-value = 1
```

and with weights: i

We repeat the exercise using weights (the counts of houses used to calculate the response variable):

```
SDEM_94w <- spatialreg::errorsarlm(f, weights=units, data=boston_94,
    listw=lw_q_94, zero.policy = TRUE, Durbin=update(f, ~ . - CHAS),
    control=list(pre_eig=e))
SLX_94w <- spatialreg::lmSLX(f, weights=units, data=boston_94,
    listw=lw_q_94, zero.policy = TRUE, Durbin=update(f, ~ . - CHAS))
```

and with weights: ii

Again, the weighted SDEM model does not fit better than the weighted SLX model:

```
o <- lmtest::lrtest(SDEM_94w, SLX_94w)
attr(o, "heading")[2] <- "SDEM_94w vs. SLX_94w"
o
```

```
## Likelihood ratio test
##
## SDEM_94w vs. SLX_94w
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1  28 97.997
## 2  27 97.527 -1 0.9401     0.3323
```

and with weights: iii

but the weighted SLX model with spatially lagged continuous coordinates included is clearly better than the weighted linear model:

```
LM_94w <- lm(f, weights=units, data=boston_94)
o <- lmtest::lrtest(SLX_94w, LM_94w)
attr(o, "heading")[2] <- "SLX_94w vs. LM_94w"
o
```

```
## Likelihood ratio test
##
## SLX_94w vs. LM_94w
##   #Df LogLik  Df  Chisq Pr(>Chisq)
## 1  27 97.527
## 2  15 81.038 -12 32.978  0.0009758
```

The weighted linear model shows substantial residual autocorrelation:

```
spdep::lm.morantest(LM_94w, listw=lw_q_94) |> glance_htest()
```

```
## Observed Moran I      Expectation         Variance       Std deviate
##     2.084688e-01    -4.779942e-02     3.952739e-03      4.076108e+00
##          p.value
##     2.289786e-05
```

and with weights: v

and the weighted SLX model has some residual spatial autocorrelation:

```
spdep::lm.morantest(SLX_94w, listw=lw_q_94) |> glance_htest()
```

```
## Observed Moran I      Expectation         Variance       Std deviate
##       0.074962529     -0.078790019      0.003792667       2.496605785
##           p.value
##       0.006269413
```

The impacts for SLX and SDEM models do not involve the coefficient on the spatially lagged response, so can be created with their standard errors by linear combination:

```
o_SLX <- summary(spatialreg::impacts(SLX_94))
```

Impacts: ii

Tabulating for the SLX variable for the air pollution variable, we see that the direct and indirect (local spillovers) are both sizable, as are their total:

```
cn <- c("impacts", "se", "z-value", "p-value")
sapply(o_SLX[3:6], function(x) x["I((NOX * 10)^2)",]) |>
    as.data.frame() |> magrittr::set_names(cn)
```

```
##              impacts          se   z-value      p-value
## Direct   -0.01284041 0.002774153 -4.628589 3.681652e-06
## Indirect -0.01917370 0.005432929 -3.529164 4.168741e-04
## Total    -0.03201411 0.005954414 -5.376534 7.593317e-08
```

Impacts: iii

In the weighted case, the local spillovers are greater than the direct impacts, and the total impacts are reduced compared to the unweighted model:

```
o_SLXw <- summary(spatialreg::impacts(SLX_94w))
sapply(o_SLXw[3:6], function(x) x["I((NOX * 10)^2)",]) |>
    as.data.frame() |> magrittr::set_names(cn)
```

```
##                impacts          se   z-value     p-value
## Direct   -0.006225692 0.003235771 -1.924021 0.054351892
## Indirect -0.011927052 0.005859945 -2.035352 0.041815445
## Total    -0.018152745 0.005921712 -3.065456 0.002173386
```

Impacts: iv

The outcomes for the SDEM and weighted SDEM models are very similar:

```
o_SDEM <- summary(spatialreg::impacts(SDEM_94))
sapply(o_SDEM[3:6], function(x) x["I((NOX * 10)^2)",]) |>
    as.data.frame() |> magrittr::set_names(cn)
```

```
##                 impacts          se   z-value      p-value
## Direct   -0.01286931 0.002352776 -5.469842 4.504381e-08
## Indirect -0.01903733 0.004635196 -4.107126 4.006126e-05
## Total    -0.03190665 0.005162277 -6.180731 6.380549e-10
```

Impacts: v

```
o_SDEMw <- summary(spatialreg::impacts(SDEM_94w))
sapply(o_SDEMw[3:6], function(x) x["I((NOX * 10)^2)",]) |>
    as.data.frame() |> magrittr::set_names(cn)
```

```
##                 impacts          se   z-value     p-value
## Direct   -0.005916509 0.002693549 -2.196548 0.028052726
## Indirect -0.010703048 0.005070825 -2.110712 0.034797114
## Total    -0.016619558 0.005373664 -3.092780 0.001982914
```

If we go back to the original census tract level models, and examine the direct/total impacts, they are substantially smaller, both for the linear model:

```
LM_506 <- lm(update(f, log(MEDV) ~ .), data=boston_506)
printCoefmat(coef(summary(LM_506))["I((NOX * 10)^2)",, drop=FALSE])
```

```
##                  Estimate Std. Error t value Pr(>|t|)
## I((NOX * 10)^2) -0.0065753  0.0011240 -5.8499 8.98e-09
```

and the spatial error model with added trend surface covariates:

```
SEM_506 <- spatialreg::errorsarlm(update(f, log(CMEDV) ~ . +
        poly(LON, LAT, degree=2)), data=boston_506, listw=lw_q_506)
printCoefmat(coef(summary(SEM_506))["I((NOX * 10)^2)",, drop=FALSE])
```

```
##                   Estimate Std. Error z value Pr(>|z|)
## I((NOX * 10)^2) -0.0047442  0.0015085  -3.145 0.001661
```

The weakest weighted SDEM total impact of the air pollution covariate is still 2.5 times greater than the original calculation.

The willingness to pay for a one part per hundred milliom (pphm) reduction in NOX in 1970 USD in the original article are taken as the mean difference between prediction from the base model using the original data, and prediction with NOX reduced by 0.1 parts per ten million (1 pphm; the formula expression is `I((NOX*10)^2))`):

```
boston_506_1 <- boston_506
boston_506_1$NOX <- boston_506_1$NOX - 0.1
```

WTP ii

Since the response was taken as the logarithm of median housing value per tract or model output zone, we take the exponents of the mean predictions (in the original article USD 1613 was reported when all variables apart from NOX were set at their mean values):

```
p0 <- predict(LM_506, newdata=boston_506)
p1 <- predict(LM_506, newdata=boston_506_1)
1000*(exp(mean(p1)) - exp(mean(p0)))
```

```
## [1] 1426.712
```

WTP iii

This is reduced when using the NOX coefficient from the all-tracts spatial error model:

```
p0 <- predict(SEM_506, newdata=boston_506, listw=lw_q_506)
p1 <- predict(SEM_506, newdata=boston_506_1, listw=lw_q_506)
1000*(exp(mean(p1)) - exp(mean(p0)))
```

```
## [1] 1009.061
```

WTP iv

Repeating the exercise for the 94 air pollution model output zones dataset:

```
boston_94_1 <- boston_94
boston_94_1$NOX <- boston_94_1$NOX - 0.1
```

we se an apparently much larger WTP in the SLX model:

```
p0 <- predict(SLX_94, newdata=boston_94, listw=lw_q_94)
p1 <- predict(SLX_94, newdata=boston_94_1, listw=lw_q_94)
exp(mean(p1)) - exp(mean(p0))
```

```
## [1] 8168.437
```

Taking the SLX model weighted by the number of reported housing units per model output zone, varying from a minimum of 25 to a maximum of 12411, and a median of 3020, and with the lowest unit counts seen where NOX values are highest:

```
p0 <- predict(SLX_94w, newdata=boston_94, listw=lw_q_94)
p1 <- predict(SLX_94w, newdata=boston_94_1, listw=lw_q_94)
exp(mean(p1)) - exp(mean(p0))
```

```
## [1] 4291.622
```

For safety's sake, the SDEM WTP is:

```
p0 <- predict(SDEM_94, newdata=boston_94, listw=lw_q_94,
    legacy.mixed=TRUE)
p1 <- predict(SDEM_94, newdata=boston_94_1, listw=lw_q_94,
    legacy.mixed=TRUE)
exp(mean(p1)) - exp(mean(p0))
```

```
## [1] 8136.228
```

WTP vii

and the weighted SDEM:

```
p0 <- predict(SDEM_94w, newdata=boston_94, listw=lw_q_94,
    legacy.mixed=TRUE)
p1 <- predict(SDEM_94w, newdata=boston_94_1, listw=lw_q_94,
    legacy.mixed=TRUE)
exp(mean(p1)) - exp(mean(p0))
```

```
## [1] 3891.738
```

differing very little from the comparably specified SLX model outcomes. These suggest that an average WTP of about USD 1500 in the original article could have been increased by a factor of three had the analysis been conducted on more appropriate support, and using the indirect local spillovers given by the spatially lagged covariates.

We could think that adding IID or MRF terms at the level of the model output zones, in addition to copying out upper level covariates to lower level tract entities, might help:

```
library(Matrix)
library(lme4)
MLM <- lmer(update(f, . ~ . + (1 | NOX_ID)),
    data = boston_487, REML = FALSE)
summary(MLM)$coefficients["I((NOX * 10)^2)",]
```

```
##      Estimate   Std. Error      t value
## -0.003479827  0.002260249 -1.539577132
```

We can see that the NOX coefficient is relatively small, something that is reflected in the very moderate WTP estimates in the IID random effects case:

```r
boston_487_1 <- boston_487
boston_487_1$NOX <- boston_487_1$NOX - 0.1
p0 <- predict(MLM, newdata=boston_487)
p1 <- predict(MLM, newdata=boston_487_1)
(exp(mean(p1)) - exp(mean(p0)))
```

```
## [1] 726.7937
```

The estimates and WTP outcomes in the IID case are very similar using `mgcv::gam()`:

```
suppressPackageStartupMessages(library(mgcv))
GAM_iid <- gam(update(f, . ~ . + s(NOX_ID, bs = "re")),
    data = boston_487, method = "REML")
summary(GAM_iid)$p.table["I((NOX * 10)^2)",]
```

```
##      Estimate    Std. Error       t value      Pr(>|t|)
## -5.787834e-03  1.075113e-03 -5.383464e+00  1.153104e-07
```

How do multi-level models fit into the picture? iv

```r
p0 <- predict(GAM_iid, newdata=boston_487)
p1 <- predict(GAM_iid, newdata=boston_487_1)
(exp(mean(p1)) - exp(mean(p0)))
```

```
## [1] 1223.08
```

If we include a spatially structured random effect expressed as an Markov random field, the results are even more depressing:

```
names(nb_q_93) <- attr(nb_q_93, "region.id")
boston_487$NOX_ID <- as.factor(boston_487$NOX_ID)
GAM_MRF <- gam(update(f, . ~ . +
        s(NOX_ID, bs = "mrf", xt = list(nb = nb_q_93))),
    data = boston_487, method = "REML")
summary(GAM_MRF)$p.table["I((NOX * 10)^2)",]
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## -0.002085996  0.002185312 -0.954553195  0.340364383
```

How do multi-level models fit into the picture? vi

```
p0 <- predict(GAM_MRF, newdata=boston_487)
p1 <- predict(GAM_MRF, newdata=boston_487_1)
(exp(mean(p1)) - exp(mean(p0)))
```

```
## [1] 432.6188
```

Unfortunately, the coefficient estimates for the air pollution variable for these multilevel models are not helpful. All are negative as expected, but the inclusion of the model output zone level effects, IID or spatially structured, makes it is hard to disentangle the influence of the scale of observation from that of covariates observed at that scale rather than at the tract level.

Conclusions so far

- Entitation, that is using spatial entities that match the aims of the study being undertaken, is as important as the technical specification of the estimation model
- In addition to the aims of the study, the entities should try to match the spatial footprint of known spatial processes avoiding unnecessary or avoidable leakage or spillover between entities
- Sensitivity to assumptions concerning functional form in (generalised) linear models
- So spatial econometrics isn't as simple as the SAS blogs, is it?

Aftermatter

Adin, Aritz, Erick Orozco-Acosta, and Maria Dolores Ugarte. 2022. *bigDM: Scalable Bayesian Disease Mapping Models for High-Dimensional Data*. https://CRAN.R-project.org/package=bigDM.

Angulo, Ana, Fernando A Lopez, Roman Minguez, and Jesus Mur. 2022. *spsur: Spatial Seemingly Unrelated Regression Models*. https://CRAN.R-project.org/package=spsur.

Anselin, Luc. 1988. *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer Academic Publishers.

———. 2010. "Thirty Years of Spatial Econometrics." *Papers in Regional Science* 89: 3–25.

Anselin, Luc, and Sergio Rey. 2014. *Modern Spatial Econometrics in Practice*. Chicago: GeoDa Press.

Arbia, Giuseppe. 2014. *A Primer for Spatial Econometrics: With Applications in R*. London: Palgrave Macmillan UK. https://doi.org/10.1057/9781137317940.

Belsley, David A., Edwin Kuh, and Roy E. Welsch. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley & Sons.

Besag, Julian. 1974. “Spatial Interaction and the Statistical Analysis of Lattice Systems.” *Journal of the Royal Statistical Society. Series B (Methodological)* 36: pp. 192–236.

Bivand, R. S. 1984. “Regression Modeling with Spatial Dependence: an Application of Some Class Selection and Estimation Methods.” *Geographical Analysis* 16: 25–37.

Bivand, Roger. 2017. “Revisiting the Boston Data Set - Changing the Units of Observation Affects Estimated Willingness to Pay for Clean Air.” *REGION* 4 (1): 109–27. https://doi.org/10.18335/region.v4i1.107.

———. 2023. *spdep: Spatial Dependence: Weighting Schemes, Statistics*. https://CRAN.R-project.org/package=spdep.

Bivand, Roger, Jan Hauke, and Tomasz Kossowski. 2013. "Computing the Jacobian in Gaussian Spatial Autoregressive Models: An Illustrated Comparison of Available Methods." *Geographical Analysis* 45 (2): 150–79.

Bivand, Roger, Giovanni Millo, and Gianfranco Piras. 2021. "A Review of Software for Spatial Econometrics in R." *Mathematics* 9 (11). https://doi.org/10.3390/math9111276.

Bivand, Roger, and Jakub Nowosad. 2023. *CRAN Task View: Analysis of Spatial Data.* https://cran.r-project.org/view=Spatial.

Bivand, Roger, Jakub Nowosad, and Robin Lovelace. 2022. *spData: Datasets for Spatial Analysis.* https://CRAN.R-project.org/package=spData.

Bivand, Roger, and Gianfranco Piras. 2015. "Comparing Implementations of Estimation Methods for Spatial Econometrics." *Journal of Statistical Software* 63 (1): 1–36. https://doi.org/10.18637/jss.v063.i18.

———. 2022. *spatialreg: Spatial Regression Analysis*. https://CRAN.R-project.org/package=spatialreg.

Bivand, Roger, Zhe Sha, Liv Osland, and Ingrid Sandvig Thorsen. 2017. "A Comparison of Estimation Methods for Multilevel Models of Spatially Structured Data." *Spatial Statistics*. https://doi.org/10.1016/j.spasta.2017.01.002.

Blangiardo, Marta, and Michela Cameletti. 2015. *Spatial and Spatio-Temporal Bayesian Models with R-INLA*. Chichester, UK: John Wiley & Sons.

Boonstra, Harm Jan. 2023. *mcmcsae: Markov Chain Monte Carlo Small Area Estimation*. https://CRAN.R-project.org/package=mcmcsae.

Bürkner, Paul-Christian. 2022. *brms: Bayesian Regression Models Using 'Stan'*. https://CRAN.R-project.org/package=brms.

References v

Cliff, A. D., and J. K. Ord. 1969. "The Problem of Spatial Autocorrelation." In *Studies in Regional Science*, edited by A. J. Scott, 25–55. London Papers in Regional Science. London: Pion.

———. 1973. *Spatial Autocorrelation*. London: Pion.

Cliff, A., and J. K. Ord. 1972. "Testing for Spatial Autocorrelation Among Regression Residuals." *Geographical Analysis* 4: 267–84.

Conley, T. G. 1999. "GMM Estimation with Cross Sectional Dependence." *Journal of Econometrics* 92 (1): 1–45. https://doi.org/https://doi.org/10.1016/S0304-4076(98)00084-0.

Croissant, Yves, and Giovanni Millo. 2008. "Panel Data Econometrics in R: The plm Package." *Journal of Statistical Software* 27 (2): 1–43. https://doi.org/10.18637/jss.v027.i02.

———. 2018. *Panel Data Econometrics with R*. Chichester, UK: John Wiley. https://doi.org/10.1002/9781119504641.

Croissant, Yves, Giovanni Millo, and Kevin Tappe. 2022. *plm: Linear Models for Panel Data*. https://CRAN.R-project.org/package=plm.

Dargel, Lukas, and Thibault Laurent. 2021. *spflow: Spatial Econometric Interaction Models*. https://CRAN.R-project.org/package=spflow.

De Bastiani, Fernanda, Mikis Stasinopoulos, and Robert Rigby. 2018. *gamlss.spatial: Spatial Terms in Generalized Additive Models for Location Scale and Shape Models*. https://CRAN.R-project.org/package=gamlss.spatial.

De Nicolò, Silvia, and Aldo Gardini. 2023. *tipsae: Tools for Handling Indices and Proportions in Small Area Estimation*. https://CRAN.R-project.org/package=tipsae.

Do, Van Huyen, Thibault Laurent, and Anne Vanhems. 2021. "Guidelines on Areal Interpolation Methods." In *Advances in Contemporary Statistics and Econometrics: Festschrift in Honor of Christine Thomas-Agnan*, edited by Abdelaati Daouia and Anne Ruiz-Gazen, 385–407. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-73249-3_20.

Do, Van Huyen, Christine Thomas-Agnan, and Anne Vanhems. 2015. "Accuracy of Areal Interpolation Methods for Count Data." *Spatial Statistics* 14: 412–38. https://doi.org/10.1016/j.spasta.2015.07.005.

Donegan, Connor. 2022. *geostan: Bayesian Spatial Analysis*. https://CRAN.R-project.org/package=geostan.

Dong, Guanpeng, Richard Harris, and Angelos Mimis. 2020. *HSAR: Hierarchical Spatial Autoregressive Model*. https://cran.r-project.org/src/contrib/Archive/HSAR/.

Düben, Christian. 2022. *conleyreg: Estimations Using Conley Standard Errors*. https://CRAN.R-project.org/package=conleyreg.

Duncan, O. D., R. P. Cuzzort, and B. Duncan. 1961. *Statistical Geography: Problems in Analyzing Areal Data*. Glencoe, IL: Free Press.

Elhorst, J. Paul. 2003. "Specification and estimation of spatial panel data models." *INTERNATIONAL REGIONAL SCIENCE REVIEW* 26 (3): 244–68. https://doi.org/%7B10.1177/0160017603253791%7D.

———. 2014. *Spatial Econometrics: From Cross-Sectional Data to Spatial Panels*. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-40340-8.

Fusco, Elisa, and Francesco Vidoli. 2022. *ssfa: Spatial Stochastic Frontier Analysis*. https://CRAN.R-project.org/package=ssfa.

Geary, R. C. 1954. “The Contiguity Ratio and Statistical Mapping.” *The Incorporated Statistician* 5: 115–45.

Gelfand, A. E. 2010. “Misaligned Spatial Data: The Change of Support Problem.” In *Handbook of Spatial Statistics*, edited by Alan E. Gelfand, Peter Diggle, Peter Guttorp, and Montserrat Fuentes, 517–39. Boca Raton: Chapman & Hall/CRC.

Gilley, Otis W., and R. Kelley Pace. 1996. “On the Harrison and Rubinfeld Data.” *Journal of Environmental Economics and Management* 31 (3): 403–5.

Gomez-Rubio, Virgilio, Paula Esther Moraga Serrano, and Barry Rowlingson. 2020. *DClusterm: Model-Based Detection of Disease Clusters*. https://CRAN.R-project.org/package=DClusterm.

Gómez-Rubio, Virgilio. 2020. *Bayesian Inference with INLA*. Boca Raton, FL: CRC Press.

Gómez-Rubio, Virgilio, and Roger Bivand. 2018. *INLABMA: Bayesian Model Averaging with INLA*. https://CRAN.R-project.org/package=INLABMA.

Gómez-Rubio, Virgilio, Roger Bivand, and Håvard Rue. 2021. "Estimating Spatial Econometrics Models with Integrated Nested Laplace Approximation." *Mathematics* 9 (17). https://doi.org/10.3390/math9172044.

Gómez-Rubio, Virgilio, and Francisco Palmí-Perales. 2019. "Multivariate posterior inference for spatial models with the integrated nested Laplace approximation." *Journal of the Royal Statistical Society Series C* 68 (1): 199–215. https://doi.org/10.1111/rssc.12292.

Gotway, C. A., and L. J. Young. 2002. "Combining Incompatible Spatial Data." *Journal of the American Statistical Association* 97: 632–48.

Granger, Clive W. J. 1969. "Spatial Data and Time Series Analysis." In *Studies in Regional Science*, edited by A. J. Scott, 1–25. London Papers in Regional Science 1. London: Pion.

Halleck Vega, Solmaria, and J. Paul Elhorst. 2015. "The SLX Model." *Journal of Regional Science* 55 (3): 339–63. https://doi.org/10.1111/jors.12188.

Harmening, Sylvia, Ann-Kristin Kreutzmann, Soeren Pannier, Felix Skarke, Natalia Rojas-Perilla, Nicola Salvati, Timo Schmid, Matthias Templ, Nikos Tzavidis, and Nora Würz. 2022. *emdi: Estimating and Mapping Disaggregated Indicators*. https://CRAN.R-project.org/package=emdi.

Harrison, David, and Daniel L. Rubinfeld. 1978a. "Distribution of benefits from improvements in urban air-quality." *Journal of Environmental Economics and Management* 5: 313–32. https://doi.org/%7B10.1016/0095-0696(78)90017-7%7D.

———. 1978b. "Hedonic Housing Prices and the Demand for Clean Air." *Journal of Environmental Economics and Management* 5: 81–102.

———. 1978c. "The Air Pollution and Property Value Debate: Some Empirical Evidence." *The Review of Economics and Statistics* 60: 635–38.

Hepple, Leslie W. 1974. "The Impact of Stochastic Process Theory Upon Spatial Analysis in Human Geography." *Progress in Geography* 6: 89–142.

———. 1976. "A Maximum Likelihood Model for Econometric Estimation with Spatial Series." In *Theory and Practice in Regional Science*, edited by I. Masser, 90–104. London Papers in Regional Science. London: Pion.

Hordijk, Leen. 1974. "Spatial Correlation in the Disturbances of a Linear Interregional Model." *Regional and Urban Economics* 4: 117–40.

Hordijk, Leen, and Jean H. P. Paelinck. 1976. "Some Principles and Results in Spatial Econometrics." *Recherches Economiques de Louvain* 42: 175–97.

Ingram, G. K., and G. R. Fauth. 1974. *TASSIM: A Transportation and Air Shed SIMulation Model, Volume 1. Case Study of the Boston Region*. Department of City; Regional Planning, Harvard University.

Ingram, G. K., G. R. Fauth, and E. A. Kroch. 1974. *TASSIM: A Transportation and Air Shed SIMulation Model, Volume 2: Program User's Guide*. Department of City; Regional Planning, Harvard University.

Jong Hee Park, Xun Pang, Michela Cameletti. 2022. *CRAN Task View: Bayesian Inference*. https://cran.r-project.org/view=Bayesian.

Joo, Rocío, Matthew E. Boone, Thomas A. Clay, Samantha C. Patrick, Susana Clusella-Trullas, and Mathieu Basille. 2020. "Navigating Through the R Packages for Movement." *Journal of Animal Ecology* 89 (1): 248–67. https://doi.org/10.1111/1365-2656.13116.

Kelejian, Harry, and Gianfranco Piras. 2017. *Spatial Econometrics*. Academic Press.

Kelejian, Harry, and Ingmar Prucha. 1998. "Generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances." *Journal of Real Estate Finance and Economics* 17 (1): 99–121.

Kopczewska, Katarzyna. 2020. *Applied Spatial Statistics and Econometrics: Data Analysis in R.* Abingdon, UK: Routledge.

Lee, Duncan. 2022. *CARBayes: Spatial Generalised Linear Mixed Models for Areal Unit Data.* https://CRAN.R-project.org/package=CARBayes.

Lee, Youngbin, Hui Jeong Ha, Sohyun Park, Kyusik Kim, and Jinhyung Lee. 2022. *spnaf: Spatial Network Autocorrelation for Flow Data.* https://CRAN.R-project.org/package=spnaf.

Leisch, Friedrich, and Evgenia Dimitriadou. 2021. *mlbench: Machine Learning Benchmark Problems.*

LeSage, James P. 2014. "What Regional Scientists Need to Know about Spatial Econometrics." *Review of Regional Studies* 44: 13–32. https://journal.srsa.org/ojs/index.php/RRS/article/view/44.1.2.

LeSage, James P., and R. Kelley Pace. 2009. *Introduction to Spatial Econometrics.* Boca Raton, FL: CRC Press.

Li, Zehang R, Bryan D Martin, Yuan Hsiao, Jessica Godwin, John Paige, Peter Gao, Jon Wakefield, Samuel J Clark, Geir-Arne Fuglstad, and Andrea Riebler. 2022. *SUMMER: Small-Area-Estimation Unit/Area Models and Methods for Estimation in R.* https://CRAN.R-project.org/package=SUMMER.

Lindgren, Finn, and Fabian E. Bachl. 2022. *inlabru: Bayesian Latent Gaussian Modelling Using INLA and Extensions.* https://CRAN.R-project.org/package=inlabru.

Lopez, Fernando, Roman Minguez, Antonio Paez, and Manuel Ruiz. 2022. *spqdep: Testing for Spatial Independence of Qualitative Data in Cross Section.* https://CRAN.R-project.org/package=spqdep.

Mahoney, Michael. 2022. *waywiser: Methods for Assessing Spatial Models*. https://CRAN.R-project.org/package=waywiser.

Martinetti, Davide, and Ghislain Geniaux. 2021. *ProbitSpatial: Probit with Spatial Dependence, SAR, SEM and SARAR Models*. https://CRAN.R-project.org/package=ProbitSpatial.

McMillen, Daniel. 2013a. *McSpatial: Nonparametric Spatial Data Analysis*. https://cran.r-project.org/src/contrib/Archive/McSpatial/.

———. 2013b. *Quantile Regression for Spatial Data*. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-31815-3.

Meyer, Hanna, Carles Milà, and Marvin Ludwig. 2023. *CAST: 'Caret' Applications for Spatial-Temporal Models*. https://CRAN.R-project.org/package=CAST.

Milà, Carles, Jorge Mateu, Edzer Pebesma, and Hanna Meyer. 2022. "Nearest Neighbour Distance Matching Leave-One-Out Cross-Validation for Map Validation." *Methods in Ecology and Evolution* 13 (6): 1304–16. https://doi.org/10.1111/2041-210X.13851.

Millo, Giovanni. 2017. "Robust Standard Error Estimators for Panel Models: A Unifying Approach." *Journal of Statistical Software* 82 (3): 1–27. https://doi.org/10.18637/jss.v082.i03.

Millo, Giovanni, and Gianfranco Piras. 2012. "splm: Spatial Panel Data Models in R." *Journal of Statistical Software* 47 (1): 1–38.

———. 2022. *splm: Econometric Models for Spatial Panel Data.* https://CRAN.R-project.org/package=splm.

Minguez, Roman, Roberto Basile, Maria Durban, and Gonzalo Espana-Heredia. 2022. *pspatreg: Spatial and Spatio-Temporal Semiparametric Regression Models with Spatial Lags.* https://CRAN.R-project.org/package=pspatreg.

Moran, P. A. P. 1948. "The Interpretation of Statistical Maps." *Journal of the Royal Statistical Society, Series B (Methodological)* 10 (2): 243–51.

Mur, Jesús, and Ana Angulo. 2006. "The Spatial Durbin Model and the Common Factor Tests." *Spatial Economic Analysis* 1 (2): 207–26. https://doi.org/10.1080/17421770601009841.

Murakami, Daisuke. 2022. *spmoran: Fast Spatial Regression Using Moran Eigenvectors*. https://CRAN.R-project.org/package=spmoran.

Newman, D. J., S. Hettich, C. L. Blake, and C. J. Merz. 1998. "UCI Repository of Machine Learning Databases." University of California, Irvine, Dept. of Information; Computer Sciences. http://www.ics.uci.edu/~mlearn/MLRepository.html.

Olsson, Gunnar. 1970. "Explanation, Prediction, and Meaning Variance: An Assessment of Distance Interaction Models." *Economic Geography* 46: 223–33. https://doi.org/10.2307/143140.

Ord, J. Keith. 1975. "Estimation Methods for Models of Spatial Interaction." *Journal of the American Statistical Association* 70: 120–26.

———. 2010. "Spatial Autocorrelation: A Statistician's Reflections." In *Perspectives on Spatial Data Analysis*, edited by L. Anselin and S. J. Rey, 165–80. Berlin: Springer-Verlag.

Osland, Liv, Ingrid Sandvig Thorsen, and Inge Thorsen. 2016. "Accounting for Local Spatial Heterogeneities in Housing Market Studies." *Journal of Regional Science* 56 (5): 895–920. https://doi.org/DOI:10.1111/jors.12281.

Pace, R. K., and R. P. Barry. 1997. "Fast spatial estimation." *Applied Economics Letters* 4 (5): 337–41.

Pace, R. Kelley, and O. W. Gilley. 1997. "Using the Spatial Configuration of the Data to Improve Estimation." *Journal of the Real Estate Finance and Economics* 14: 333–40.

Pace, R. Kelley, and James P. LeSage. 2008. "A Spatial Hausman Test." *Economics Letters* 101 (3): 282–84. https://doi.org/10.1016/j.econlet.2008.09.003.

Paelinck, Jean H. P. 2013. "Some Challenges for Spatial Econometricians." In *Spatial Econometrics and Regional Economic Analysis*, edited by Bogdan Suchecki, 292:11–20. Acta Universitas Lodziensis, Folia Oeconomica. Wydawnictwa Uniwersytetu Łódzkiego.

Paelinck, Jean H. P., and Leo H. Klaassen. 1979. *Spatial Econometrics*. Farnborough: Saxon House.

Pebesma, Edzer, and Roger Bivand. 2022. *CRAN Task View: Handling and Analyzing Spatio-Temporal Data*. https://cran.r-project.org/view=SpatioTemporal.

———. 2023. *Spatial Data Science with Applications in R*. Chapman & Hall. https://www.routledge.com/Spatial-Data-Science-With-Applications-in-R/Pebesma-Bivand/p/book/9781138311183.

Piras, Gianfranco. 2010. "sphet: Spatial Models with Heteroskedastic Innovations in R." *Journal of Statistical Software* 35 (1): 1–21.

———. 2022. *sphet: Estimation of Spatial Autoregressive Models with and Without Heteroskedastic Innovations*. https://CRAN.R-project.org/package=sphet.

Piras, Gianfranco, and Mauricio Sarrias. 2022. *hspm: Heterogeneous Spatial Models*. https://CRAN.R-project.org/package=hspm.

Ripley, B. D. 1988. *Statistical Inference for Spatial Processes*. Cambridge: Cambridge University Press.

Rue, Håvard, Finn Lindgren, and Elias Teixeira Krainski. 2022. *INLA: Full Bayesian Analysis of Latent Gaussian Models Using Integrated Nested Laplace Approximations*. https://www.r-inla.org/.

Sarrias, Mauricio, and Gianfranco Piras. 2022. *spldv: Spatial Models for Limited Dependent Variables*. https://CRAN.R-project.org/package=spldv.

Schratz, Patrick, and Marc Becker. 2022. *mlr3spatiotempcv: Spatiotemporal Resampling Methods for 'Mlr3'*. https://CRAN.R-project.org/package=mlr3spatiotempcv.

Silge, Julia, and Michael Mahoney. 2023. *spatialsample: Spatial Resampling Infrastructure*. https://CRAN.R-project.org/package=spatialsample.

Simonovska, Rozeta. 2022. *SDPDmod: Spatial Dynamic Panel Data Modeling*. https://CRAN.R-project.org/package=SDPDmod.

Smith, T. E., and K. L. Lee. 2012. "The effects of spatial autoregressive dependencies on inference in ordinary least squares: a geometric approach." *Journal of Geographical Systems* 14 (January): 91–124. https://doi.org/10.1007/s10109-011-0152-x.

Templ, Matthias, Alexander Kowarik, and Tobias Schoch. 2022. *CRAN Task View: Official Statistics & Survey Statistics*. https://cran.r-project.org/view=OfficialStatistics.

Tobler, W. R. 1970. "A Computer Movie Simulating Urban Growth in the Detroit Region." *Economic Geography* 46: 234–40. https://doi.org/10.2307/143141.

Umlauf, Nikolaus, Nadja Klein, Achim Zeileis, and Thorsten Simon. 2022. *bamlss: Bayesian Additive Models for Location, Scale, and Shape (and Beyond)*. https://CRAN.R-project.org/package=bamlss.

Umlauf, Nikolaus, Thomas Kneib, Stefan Lang, and Achim Zeileis. 2022. *R2BayesX: Estimate Structured Additive Regression Models with 'BayesX'*. https://CRAN.R-project.org/package=R2BayesX.

Valavi, Roozbeh, Jane Elith, José Lahoz-Monfort, Ian Flint, and Gurutzeta Guillera-Arroita. 2023. *blockCV: Spatial and Environmental Blocking for k-Fold and LOO Cross-Validation*. https://CRAN.R-project.org/package=blockCV.

Vidoli, Francesco, and Roberto Benedetti. 2022. *SpatialRegimes: Spatial Constrained Clusterwise Regression*. https://CRAN.R-project.org/package=SpatialRegimes.

Wagner, Martin, and Achim Zeileis. 2019. *lagsarlmtree: Spatial Lag Model Trees*. https://CRAN.R-project.org/package=lagsarlmtree.

Wakefield, J. C., and H. Lyons. 2010. "Spatial Aggregation and the Ecological Fallacy." In *Handbook of Spatial Statistics*, edited by Alan E. Gelfand, Peter Diggle, Peter Guttorp, and Montserrat Fuentes, 541–58. Boca Raton: Chapman & Hall/CRC.

Warnholz, Sebastian. 2023. *saeRobust: Robust Small Area Estimation*. https://CRAN.R-project.org/package=saeRobust.

Warnholz, Sebastian, and Timo Schmid. 2022. *saeSim: Simulation Tools for Small Area Estimation*. https://CRAN.R-project.org/package=saeSim.

Whittle, P. 1954. "On Stationary Processes in the Plane." *Biometrika* 41 (3-4): 434–49. https://doi.org/10.1093/biomet/41.3-4.434.

Wilhelm, Stefan, and Miguel Godinho de Matos. 2022. *spatialprobit: Spatial Probit Models*. https://CRAN.R-project.org/package=spatialprobit.

Wilson, Alan G. 2000. *Complex Spatial Systems: The Modelling Foundations of Urban and Regional Analysis*. Harlow: Pearson Education.

———. 2002. "Complex Spatial Systems: Challenges for Modellers." *Mathematical and Computer Modelling* 36: 379–87.

Wood, Simon. 2022. *mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation*. https://CRAN.R-project.org/package=mgcv.

Zeileis, Achim, Grant McDermott, and Kevin Tappe. 2023. *CRAN Task View: Econometrics*. https://CRAN.R-project.org/view=Econometrics.