J Geograph Syst (2002) 4:405–421

Journal of Geographical Systems

Spatial econometrics functions in R: Classes and methods

Roger Bivand

Economic Geography Section, Department of Economics, Norwegian School of Economics and Business Administration, Breiviksveien 40, N-5045 Bergen, Norway;
e-mail: Roger.Bivand@nhh.no

Received: 26 August 2002 / Revised version: 15 January 2003

Abstract. Placing spatial econometrics and more generally spatial statistics in the context of an extensible data analysis environment such as R exposes similarities and differences between traditions of analysis. This can be fruitful, and is explored here in relation to prediction and other methods usually applied to fitted models in R. Objects in R may be assigned a class attribute, including fitted model objects. Such fitted model objects may be provided with methods allowing them to be displayed, compared, and used for prediction, and it is of interest to see whether fitted spatial models can be treated in the same way.

Key words: Spatial econometrics, spatial statistics, spatial data analysis, open source software

JEL classification: C12, C13, C80, C88

1 Introduction

Developments in the R implementation of the S data analysis language are providing new and effective tools needed for writing functions for spatial analysis. The release of an R package for constructing and manipulating spatial weights, and for testing for global and local dependence during 2001 has been followed by work on functions for spatial econometrics (package **spdep**[1]). The paper gives an introduction to some of the issues faced in writing this package in R, to the use of classes and object attributes, and to class-based method dispatch. In particular, attention will be paid to the question of how prediction

I would like to thank anonymous referees, and participants of the European Regional Science Association Congress in Dortmund, August 2002, for their helpful comments. This work was funded in part by EU Contract Number Q5RS–2000–30183.

[1] source package: `spdep_0.1-6.tar.gz` or later, Windows binary package: `spdep.zip` may be retrieved from: `http://cran.r-project.org`.

should be understood in relation to the most commonly employed spatial econometrics simultaneous autoregressive models. Prediction is of importance because fitted models may reasonably be expected to be used to provide predictions of the response variable using new data – both attribute and position – that may not have been available when the model was fitted.

Class-based features are important because they encapsulate information about the data in a generic way, also when the data is for example a model formula, an object describing spatial neighbourhood relationships, or the results of fitting a model to data. This permits the flexible handling of subsetting, missing data, dummy variables, and other issues, based on existing classes that are extended to handle spatial econometrics functions. For the analyst, it is convenient if generic access functions can be applied to spatial analysis classes, such as making a summary or plotting a spatial neighbours structure. The same applies to the use of model formulae, describing the model to be estimated, for a range of estimating functions. In this setting, a spatial linear model should build on the classes of the arguments of the underlying linear model. There should be no difference in the syntax of shared arguments between the aspatial linear model, spatial econometrics models, or geographically weighted regression models, although of course function-specific arguments should be introduced.

It is also of interest to compare spatial econometric formulations with other related model structures, such as those for mixed effects models, and to explore other alternative approaches. These may include extensions to repeated measurements, to spatial time series, and to generalised linear models, although here the spatial case is often currently unresolved. However, the underlying classes are important in that their implementation may make the flexible extension of spatial analysis tools more or less difficult, and consequently that they should admit the quick prototyping of experimental new modelling techniques rather than hinder it.

It is clear that different disciplines and data analysis communities do not approach the writing of code, or the use of command line interfaces, in the same ways, and have varying expectations regarding the concerns of users. It is however arguable that language environments such as `S`, and its implementations `R` and `S-PLUS`, are instrumental in reducing barriers between users, who are not supposed to meddle with the software, but who can be expected to know about their data and methods, and developers. When these qualities of the `S` language environment for data analysis are coupled to free access to source code, opportunities for mutual peer-review and exchange between and among users and developers arise that are otherwise very difficult to create.

The development of the paper is, after sketching the position of the `R` project and the **spdep** package, first to review some open problems in spatial econometrics. Next, the experience of the class/method mechanisms in `S` and `R` is drawn on, including a discussion of the use of classes in **spdep** at present. This leads to an extended discussion of how prediction might be approached in spatial econometrics, since `predict()` is a method typically implemented for classes of fitted models. This is exemplified using the revised Harrison and Rubinfeld Boston house price data set, which is also distributed with **spdep**.

1.1 The R *project*

As summarised in brief in Bivand and Gebhardt (2000), R is a language and environment for statistical computing and graphics, and is similar to S. The S language is described and documented in Becker et al. (1988), Chambers and Hastie (1992), and more recently in Chambers (1998). There are differences between implementations of S: S-PLUS – which is a well-supported commercial product with many enhancements – manages both memory and data object storage in different ways from R. The chief syntactic differences are described in Ihaka and Gentleman (1996). Perhaps the most comprehensive introduction to the use of current versions of S-PLUS and R is Venables and Ripley (2002); a simpler alternative for R is Dalgaard (2002).

R is available as source, and as binaries for Unix/Linux, Windows, and Macintosh platforms[2]. Contributed code is distributed from mirrored archives following control for adherence to accepted standards for coding, documentation and licensing. The contributed packages are distributed as source, and for some platforms – including Windows – as binaries, which can in addition be updated on-line using the `update.packages()` function within R. As usual in Free Software projects, there is no guarantee that the code does what it is intended to do, but since it is open to inspection and modification, the analyst is able to make desired changes and fixes, and if so moved, to contribute them back to the community, preferably through the package maintainer.

1.2 The **spdep** *package*

The current version of the **spdep** package is a collection of functions to create spatial weights matrix objects from polygon contiguities, from point patterns by distance and tessellations, for summarising these objects, and for permitting their use in spatial data analysis[3]; a collection of tests for spatial autocorrelation, including global Moran's I, Geary's C, Hubert/Mantel general cross product statistic, and local Moran's I and Getis/Ord G, saddlepoint approximations for global and local Moran's I; and functions for estimating spatial simultaneous autoregressive (SAR) models. It contains contributions including code and/or assistance in creating code and access to legacy data sets from quite a number of spatial data analysts; full details are in the licence file installed with the package. It is indeed central to the dynamics of free software/open source software projects such as R and its contributed packages, that communities are brought into being and fostered, leading where appropriate to collaborative development, and indeed to the replacement of code or class structures found by users in the community to be unsatisfactory or limiting.

---

[2] Both R itself and contributed packages may be downloaded from `http://cran.r-project.org`.

[3] the treatment of spatial weights matrices has been discussed in greater length in Bivand and Portnov (forthcoming).

2 Spatial models and spatial statistics

It often seems to be the case that spatial statistical analysis, including spatial econometrics, finds it challenging to give insight into general relationships guiding a data generation process. It is quite obvious that inference to general relationships from cross-section spatial data using aspatial techniques raises the question of whether the positions of the observations in relation to each other should not have been included in the model specification. We now have quite a range of tests for examining these kinds of potential mis-specifications. We can also offer tools for exploring and fitting local and global spatial models, so that perhaps better supported inferences may be drawn for the data set in question, under certain assumptions.

These assumptions are not in general easy or convenient to handle, and constitute a major part of the motivation for further work on inference for spatial data generation processes. As Ripley (1988, p. 2) suggests and Anselin (1988, p.9) confirms, they remove hope that spatial data are a simple extension of time series to a further dimension (or dimensions). The assumptions of concern here include (Ripley 1988) those affecting the edges of our chosen or imposed study region, how to perform asymptotic calculations and how this doubt impacts the use of likelihood inference, how to handle inter-observational dependencies at multiple scales (both short-range and long-range), stationarity, and discretisation and support. Ripley (1988, p. 8) concludes: "(T)he above catalogue of problems may give rather a bleak impression, but this would be incorrect. It is intended rather to show why spatial problems are different and challenging".

Although many of these challenges are intractable in the point-process part of spatial statistics, more has been done to address them here. In particular, it has been recognised for some time that if we have a simple null hypothesis to simulate the spatial process model, we can generate exchangeable samples permitting us to test how well the model fits the data. As Ripley (1992) notes, an early example of this approach for the non-point-process case is the use of Monte Carlo simulation by Cliff and Ord (1973, p. 50–2). Substantial advances have also been taking place in geostatistics (Cressie 1993, Diggle et al. 1998). In addition, the implications of large volumes of data from remote sensing and geographical information systems, including data with differing support, have been recognised in a recent review by Gotway and Young (2002).

One of the characteristics of treatments of the statistical modelling of spatial data – especially lattice data – is that changes in techniques occur slowly, despite radical changes in data acquisition and computing speed. Haining's discussion of the research agenda twenty years ago (1981, pp. 88–89), focusing on spatial homogeneity and stationarity, is taken up again by him ten years later (1990, pp. 40–50), and remains relevant. Apart from the actual difficulty of the problems, it may be argued that exploring feasible solutions has been hindered by poor access to toolboxes combining both the specificity needed for handling spatial dependence between observations and general numerical and statistical functions. The coming first of SpaceStat (Anselin 1995), then James LeSage's Econometrics Toolbox for MATLAB[4], have created important opportunities, which the `R` **spdep** package attempts to

[4] `http://www.spatial-econometrics.com/`

follow up and build upon. In addition, code by Griffith (1989) for MINITAB, and by Griffith and Layne (1999) for SAS and SPSS has been made available. Finally, the spatial statistics module for S-PLUS provides additional and supplementary analytical techniques in a somewhat different form (Kaluzny et al. 1996).

To concentrate attention on the problem at hand, it may help to express the relationship between data and model in a number of parallel ways:

$$\text{data} = \left\{ \begin{array}{c} \text{model} \\ \text{fit} \\ \text{smooth} \end{array} \right\} + \left\{ \begin{array}{c} \text{error} \\ \text{residuals} \\ \text{rough} \end{array} \right\}$$

where our general grasp of the spatial data generation process on the data is incorporated in the first term on the right hand side, while the second term comprises the difference between this understanding and the observed data for our possibly unique region of study (Haining 1990, p. 29, and p. 51; cf. Hartwig and Dearing 1979, p. 10; Cox and Jones 1981, p. 140).

The model term may be made up of say fixed and random effects, of global and local smooths, of aspatial and spatial component models, of trend surface and variogram model components, or of locally or geographically weighted parts. The distribution of the error term is assumed to be known, and should be such that as much as possible of the predictable regularity is taken up in the model.

In general, the model term should give a parsimonious description of the process or processes driving the data, and techniques used to choose between alternative models should take this requirement into account. It is also not necessarily the case that the model should be fitted using all of the data to hand; indeed many model forms may be compared by partitioning the available data into training and testing subsets. This position in fact reaches back to fundamental questions regarding the application of statistical estimation methods to spatial data, especially when the goals of such application may include inference, generalisation to a wider domain than the data used for calibration (Olsson 1970, Gould 1970). In particular, Olsson's comment that: "If the ultimate purpose is prediction, then it also follows that specification of the functional relationships is more urgent than specification of the geometric properties of a spatial phenomenon" (1968, p. 131) continues to point up the question of what is being inferred to in spatial statistical analyses, also known as the geographical inference problem.

3 Classes and methods in modelling using R

Three main programming paradigms underly S: object-oriented programming, functional languages, and interfaces (Chambers and Hastie 1992, pp. 455–480). Classes and methods were introduced to S at the time of this 1992 "White" book, and were not part of the 1988 "Blue" book (Becker et al. 1988) defining the fundamentals of the language. This step was, for practical reasons, incremental, and was intended to assist in the further development of modelling functions. For this reason, language objects may, but do not have to, have a class attribute – all objects may have attributes with name

strings, and class is simply one such string with specific consequences for the way that functions in the system handle objects.

This established form of class and method use in S and hence R is the one which will be covered here. It should however be noted that a new class/method formalism has been introduced to S in the 1998 "Green" book (Chambers 1998), and is being introduced to R, as well as underlying S-PLUS 6.x. Programming using both styles of classes and methods is described in detail in Venables and Ripley (2000, pp. 75–121). From the point of view of the user, however, the differences are either few or beneficial, and now require that each object shall have a class, and that each object of a given class shall have the same structure, requirements which were not present before.

The class/method formalisms in S have been adopted in the spirit of object-oriented programming, that evaluation should be data-driven. Functions for generic tasks, such as `print()`, `plot()`, `summary()`, or `logLik()`, are constructed as stubs that pass their own arguments through to the `UseMethod()` function. In the following code snippets, > is the R command line prompt – entering the name of a function causes its body to be printed:

```
> print
function (x, ...)
UseMethod ("print")
```

Within `UseMethod()`, the first argument object is examined to see if it has an attribute named "`class`". If it does, and a function named, say, `print.`"`class`" `()` exists, the arguments are passed to this function. If it has no class attribute, or if no generic function qualified with the class name is found, the object is passed to, say, `print.default()`. If we have estimated a spatial error model for the Columbus data set, and wish to display the log likelihood value of the object, we might do the following:

```
> COL.err <- errorsarlm(CRIME ~ INC + HOVAL, data = COL.OLD,
+nb2listw(COL.nb))
> class(COL.err)
[1] "sarlm"
> ll.COL.err <-logLik(COL.err)
> class(ll.COL.err)
[1] "logLik"
> ll.COL.err
'log Lik.' - 183.3805 (df = 4)
```

The model object `COL.err` has class `sarlm`, so the function used by method dispatch from `logLik()` is `logLik.sarlm()`, yielding a resulting object with class `logLik`. If an object with class `logLik`, is to be printed, `UseMethod()` will look for `print.logLik()`. As can be seen, this function expects the `logLik` object to be a scalar value, with an attribute named "`df`", the value of which is also printed.

```
>print.logLik
function(x, digits=getOption("digits"),...)
{
    cat("logLik.' ", format(c(x), digits=digits), "(df=",
        format(attr(x, "df")), ")\n", sep="")
    invisible(x)
}
```

This brief example shows both the convenience of the class/method mechanism, and the reason for moving to the new style, since in the old style there are no barriers to prevent the class attribute of an object being changed or removed, nor are there any structures to ensure that class objects have the same properties. It could be argued that software code, and by extension the formalisms employed in writing software, such as class/method formalisms in object oriented programming described briefly above, are not of importance for advancing spatial data analysis.

A response to this position is that, for computable applications, abstractions and conjectures are enriched by being implemented in structured code, especially where the code is available, documented, and open to peer review, as in R and other community supported software projects and repositories. Further, formalisms such as class/method mechanisms also provide useful standards through which the assumptions and customs underlying computing practises may be exposed and compared. Finally, class/method mechanisms, in particular care in constructing classes, are associated with concern for data modelling as also understood for example in geographical information systems. In this case, it is important that classes support data types, structures, and metadata components adequately and in a robust way.

At present the key classes in **spdep** are written in the old style, and are "nb", "listw", "sarlm", and the generic class "htest" for hypothesis tests. The first is for lists of neighbours, the second for sparse neighbour weights lists, and the third for the object returned from the fitting of SAR (simultaneous autoregressive) linear models of three types: lag, mixed, and error (corresponding to LeSage's sar(), sar() including spatially lagged independent variables, and sem() functions; there is no equivalent to his sac() function). The "htest" class is used to report the results of hypothesis tests, not least because `print.htest()` already existed, and conveniently standardised the displaying of test results.

The "sarlm" class is still under development, not least because writing methods leads to changes in components that need to be in the object itself, or can conveniently be computed at a later stage by functions such as `summary.sarlm()`, `logLik.sarlm()`, `residuals.sarlm()`, and so on. Migration to new-style classes will occur when the requirements have been refined following further exploration – old-style classes can be augmented without breaking existing code more easily than can new-style classes.

The function that has prompted the most thought is however `predict.sarlm()` – essentially all the fitted model classes in S (and R and its contributed packages) have methods for prediction, including prediction from new data. It is to this problem we will turn to show that class/method formalisms are more than a programming convenience, but

also establish baselines for what analysts should expect from model fitting software.

4 Issues in prediction in spatial econometrics

Prediction may be subdivided into several similar kinds of tasks: calculating the fitted values when the values of the response variable observation are known and are those used in fitting the model, the same scenario, but when the predictions are not for observations used to fit the model, and finally predictions for observations for which the value of the response variable is unknown. Here we choose to measure the difference between the predicted and observed values of the response variable using the root mean square error of prediction. In the aspatial linear model, predictions are a function of the fitted coefficients and their standard errors and confidence intervals may be obtained using the fitted residual standard error. Extensions to the linear model can be furnished with prediction mechanisms in generally similar ways, although expressing standard errors and confidence intervals may become more difficult.

Work on filling in missing values (Bennett et al. 1984, Haining et al. 1989, Griffith et al. 1989) has not been followed up in the spatial econometrics literature, and was focused on the case when the position of an observation was known, but where one or more attribute values was missing (see also Martin 1990). This differs from prediction using new data where there is no contiguity between the positions of the data used to fit the model and the new data, where both the positions of the observations are new, and only explanatory variable values are available for making the prediction. Where contiguity between the data sets' positions is present, predicting missing values can be accommodated in the present approach; the main thrust of this literature has been to explore the consequences for parameter estimation of the absence of some data values. Given the provision noted by Martin (1984, p. 1278) that data should be missing at random, it is not clear how to proceed when the new data adjoin the data used for fitting, for instance in one direction.

4.1 'Trend', 'signal' and 'noise'

Prediction for spatial data may be seen as the core of geostatistics; most applications of kriging aim to interpolate from known data points to other points within or adjacent to the study area, or to other support. Interpolation of this kind also underlies the use of modern statistical techniques, such as local regression or generalised additive models among many others. As pointed out above, it is usual for prediction functions to accompany each new variety of fitted model object in S, not least because the comparison of prediction errors for in-sample and out-of-sample data give insight into how well models perform. Some model fitting techniques can be found to perform very well in relation to in-sample data, but do very poorly on out-of-sample data, that is, they are 'over-fitted'. While they may exhaust the training data, they will be very restricted to that particular region of data-space, and may perform worse than other, less 'over-fitted' models, on unseen test data.

The three terms: 'trend', 'signal' and 'noise', are taken from Haining (1990, p. 258), and the S-PLUS spatial statistics module (Kaluzny et al. 1996, pp. 154–156), in which Haining's comment is followed up. In Haining (1990), the underlying linear model was a trend surface model, so that it was logical to partition the data into 'trend' and 'noise':

$$\underbrace{\mathbf{y}}_{\text{data}} = \underbrace{\mathbf{X}\beta}_{\text{trend}} + \underbrace{\varepsilon}_{\text{noise}}$$

where $E[\varepsilon] = \mathbf{0}$ and $E[\varepsilon\varepsilon^T] = \sigma^2\mathbf{I}$. If we generalise this model to the error autoregressive form, we get:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$$

with $E[\mathbf{u}] = \mathbf{0}$ and $E[\mathbf{u}\mathbf{u}^T] = \mathbf{V}$. If we write $\mathbf{V} = \sigma^2\mathbf{L}\mathbf{L}^T$, and $\mathbf{L}^{-1} = (\mathbf{I} - \lambda\mathbf{W})$, we can rewrite the relationship:

$$(\mathbf{I} - \lambda\mathbf{W})\mathbf{y} = (\mathbf{I} - \lambda\mathbf{W})\mathbf{X}\beta + \varepsilon$$

$$\underbrace{\mathbf{y}}_{\text{data}} = \underbrace{\mathbf{X}\beta}_{\text{trend}} + \underbrace{\lambda\mathbf{W}(\mathbf{y} - \mathbf{X}\beta)}_{\text{signal}} + \underbrace{\varepsilon}_{\text{noise}}$$

To predict $\mathbf{y}$, we could pre-multiply by $(\mathbf{I} - \lambda\mathbf{W})^{-1}$:

$$\mathbf{y} = \mathbf{X}\beta + (\mathbf{I} - \lambda\mathbf{W})^{-1}\varepsilon$$

which can yield the trend component, but for which the signal and noise components are combined. Cliff and Ord (1981, p. 152, cf. pp. 146–147) give $\mathbf{u} = \sigma(\mathbf{I} - \lambda\mathbf{W})^{-1}\varepsilon$ as the simultaneous autoregressive generator from ε independent identically distributed random deviates, yielding $\mathbf{u} \sim \mathrm{N}(\mathbf{0}, \mathbf{V})$. If normality is assumed for ε, then $\mathbf{u}$ is multivariate normal. Here, predictions from error autoregressions are restricted to the trend component.

Kaluzny et al. (1996, pp. 158–160) use Haining's results (1990, p. 116) to suggest that a simulation of the unobservable autocorrelated error term may be used to attempt to predict the signal, but this necessarily depends on the assumption of normality. In the SAR case, they suggest computing $\mathbf{V} = \sigma^2[(\mathbf{I} - \lambda\mathbf{W})^T(\mathbf{I} - \lambda\mathbf{W})]^{-1}$, next computing $\mathbf{L}$ as the lower triangular matrix of the Cholesky decomposition of $\mathbf{V}$, and finally simulating $\mathbf{u}$ by $\mathbf{u} = \mathbf{L}\varepsilon$, where ε is a random deviate as above.

A further alternative based on work by Martin (1984, see also modifications by Haining et al. 1989, Griffith et al. 1989, and comment by Martin 1990) is to base the approximation of the unobservable autocorrelated signal on the projection of the residuals of the fitted process through a covariance matrix expressing the spatial dependence of the positions used to fit the model and the positions of the new data (using the spatial parameter from the fitted model). If the data used for fitting the model and the new data are not contiguous in position, this term is zero.

This alternative may be compared to the case of for time series with autocorrelated errors, since the estimate of the autoregressive coefficient is needed to make an estimate of the one-period forecast error (Stewart and Wallis 1981, pp. 239–241; Johnson and DiNardo 1997, pp. 192–193). Johnson and DiNardo term this the feasible forecast, and note that there is no closed form expression for the forecast variance in this case. Suppose we have $y_t = \mathbf{x}_t^T\beta + u_t$, where $u_t = \lambda u_{t-1} + \varepsilon_t$. The same model can be written:

$$y_t - \lambda y_{t-1} = \mathbf{x}_t^T\beta - \lambda\mathbf{x}_{t-1}^T\beta + \varepsilon_t$$

Assuming λ known, β can be estimated, and substituting and rearranging, we can make a forecast of y_{t+1} by:

$$\hat{y}_{t+1} = \underbrace{\mathbf{x}_{t+1}^T\hat{\beta}}_{\text{trend}} + \underbrace{\lambda(y_t - \mathbf{x}_t^T\hat{\beta})}_{\text{signal}}$$

for which the forecast variance is also available; the terms 'trend' and 'signal' here describe the non-autoregressive and the autoregressive components of the forecast by analogy with Haining's description. When we only have an estimate of λ, the feasible forecast becomes:

$$\hat{y}_{t+1} = \mathbf{x}_{t+1}^T\hat{\beta} + \hat{\lambda}(y_t - \mathbf{x}_t^T\hat{\beta})$$

that is the sum of products of the new $\mathbf{x}_{t+1}$ values and the $\hat{\beta}$ fitted using observations $1, \ldots, t$, plus $\hat{\lambda}$ times the residual at time t, representing the temporal dependency of the series, the forecast error for the one-step-ahead forecast.

Since t and $t + 1$ are contiguous, it is possible to use the residual value from the fitted model in prediction in the time series case. In the simultaneous autoregressive spatial error model, when the new data positions coincide with, or are contiguous to, the positions of data used for fitting, it may be possible to calculate a signal component on the basis of the residuals of the fitted model and a rectangular matrix expressing the correlation structure of the original and new data positions – this approach has however not been attempted here, although Martin (1984, p. 1279) provides a solution. To accommodate this, modifications to the current spatial weights list class in **spdep** are required, but have not yet been implemented. Consequently, for the simultaneous autoregressive error model, the prediction currently implemented in `predict.sarlm()` for the newdata case is the trend, and the signal is set to zero.

Haining's approach may be extended to the spatial lag model, in which dependence is not present in the error term, but rather in the dependent variable. Here we have:

$$\underbrace{\tilde{\mathbf{y}}}_{} = \underbrace{\mathbf{X}\tilde{\beta}}_{\text{trend}} + \underbrace{\tilde{\rho}\mathbf{W}\mathbf{y}}_{\text{signal}} + \underbrace{\varepsilon}_{\text{noise}}$$

Rewriting, we have:

$$(\mathbf{I} - \rho\mathbf{W})\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

Once again, to predict $\mathbf{y}$, we could pre-multiply by $(\mathbf{I} - \rho\mathbf{W})^{-1}$:

$$\mathbf{y} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\beta + (\mathbf{I} - \rho\mathbf{W})^{-1}\varepsilon$$

The second term on the right hand side is equivalent to that in the error autoregressive case, and combines signal and noise components, while the first term combines trend and signal components.

As a first approximation, the `predict.sarlm()` function assumes that the trend can be expressed by $\mathbf{X}\hat{\beta}$, and part of the signal by $\hat{\rho}\mathbf{W}(\mathbf{I} - \hat{\rho}\mathbf{W})^{-1}\mathbf{X}\hat{\beta}$. The rationale is that if:

$$(\mathbf{I} - \hat{\rho}\mathbf{W})\mathbf{y} = \mathbf{X}\hat{\beta}$$

$$\hat{\mathbf{y}} = (\mathbf{I} - \hat{\rho}\mathbf{W})^{-1}\mathbf{X}\hat{\beta}$$

then the signal may be approximated by:

$$\hat{\rho}\mathbf{W}\hat{\mathbf{y}} = \hat{\rho}\mathbf{W}(\mathbf{I} - \hat{\rho}\mathbf{W})^{-1}\mathbf{X}\hat{\beta}$$

While this yields an estimate of part of the signal component, it is not complete, for new data missing the part combined with the noise component. This is clearly less than adequate, and more work is required here, as with the completely missing signal component for the error model.

Finally, it has been assumed that the weights matrix used for fitting the model is furnished with attributes detailing its construction: whether it is row standardised, and which type of underlying binary or general neighbourhood representation has been used (contiguity, distance, triangulation, *k*-nearest neighbours, etc.). Consequently, in predicting from new data, it is expected that the new attribute data will be accompanied by a suitable spatial weights list. This is not used in the error model predictions, but is used for the lag model, in the approximation to the part of the signal component described above.

Even if prediction for new data is as yet less well grounded, the partition of spatial model fitted values into trend and signal allows us to use alternative diagnostic plots. Examples of such plots for the data set discussed in Sect. 4.2 below are shown in Fig. 1. Tracts lying in towns in Boston city are distinguished in the plot, since their patterns seem to indicate different behaviour both in relation to the aspatial trend, and the spatial autoregressive error signal. It may be remarked that the fit of the spatial error model (AIC = −508.85) is better than that of the spatial lag model (AIC = −498.02), than the aspatial linear model (AIC = −283.96), but worse than the mixed spatial lag model (AIC = −545.23). The full results may be obtained by executing `example(boston)` after loading **spdep** into R, in which the sphere of influence row standardised weighting scheme is also presented.

4.2 Boston housing values case

The data set chosen here is that described by Gilley and Pace (1996), a revision of the Harrison and Rubinfeld Boston hedonic house price data, relating median house values to a range of environmental and social variables over 506 tracts. It is chosen because it is easily available, it has been used in a range of spatial econometric studies, including particularly LeSage's online materials on spatial econometrics[5]. The original data set is also featured as one of a corpus of machine learning datasets[6], and as such is well suited to applications such as the present. Most use of this dataset in machine learning research also seems to ignore the spatial nature of the data. Here, two prediction settings will be used.

In the first, the data are divided into northern and southern parts at UTM zone 19 northing 4,675,000 m (dividing the tracts into two almost equal

[5] `http://www.rri.wvu.edu/WebBook/LeSage/etoolbox/index.html`

[6] `ftp://ftp.ics.uci.edu/pub/machine-learning-databases`

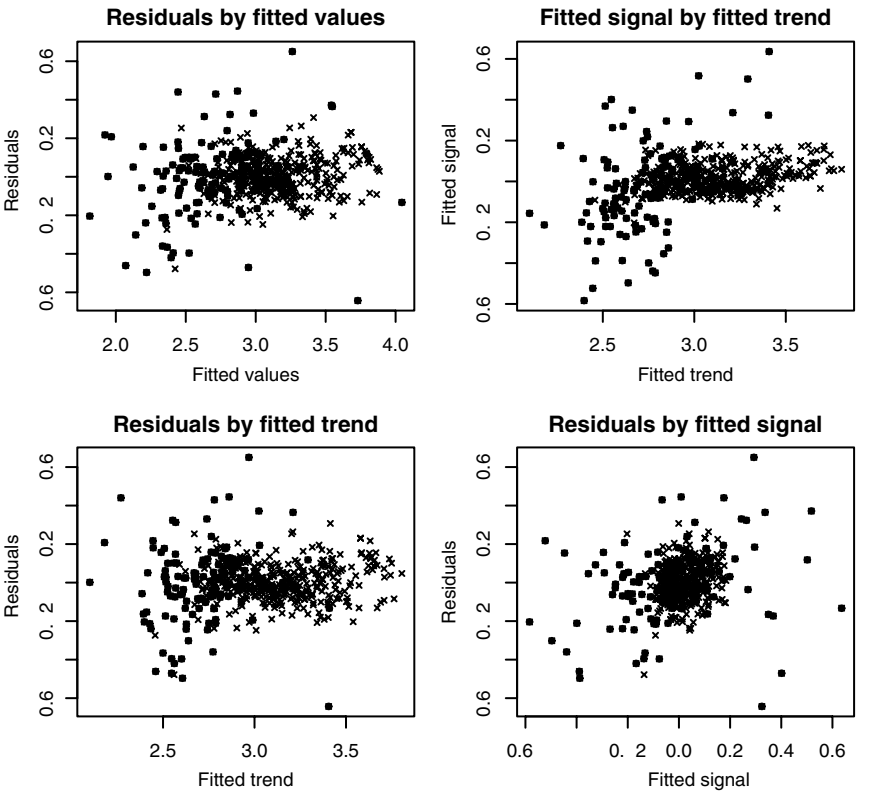

Fig. 1. Boston tract log median house price data: plots of spatial autoregressive error model fit components and residuals for all 506 tracts; tracts in towns in Boston plotted with •

groups, with the dividing line running through the Boston city tracts). The data frame is subsetted by a logical variable expressing whether the centre point of the tract is north or south of the dividing line. The spatial weights used are constructed using the sphere of influence approach based on a triangulation of the UTM zone 19 projected tract centres, subsetted using the same north/south logical variable. An ordinary least squares model was fitted to each of the parts of the city, and predictions were made with the data used for fitting the models, and then using the model fitted on the southern data with the northern data, and vice-versa. The same procedure was repeated for the spatial lag model, the spatial error model, and the spatial mixed model (the spatial lag model augmented with the spatial lags of the explanatory variables – also known as the Common Factor model).

Although it can be seen from Fig. 2 that the spatial models are better fitted to the data, especially in the south, the cross-predictions are no better than, and often worse than those for the aspatial linear model (lm). The linear model gives the best prediction of the southern median house prices using the fitted coefficient values from the northern data. At least part of the reason for this is that the fits of the models, both aspatial and spatial coefficient values, differed between the two parts of the metropolitan area, suggesting that

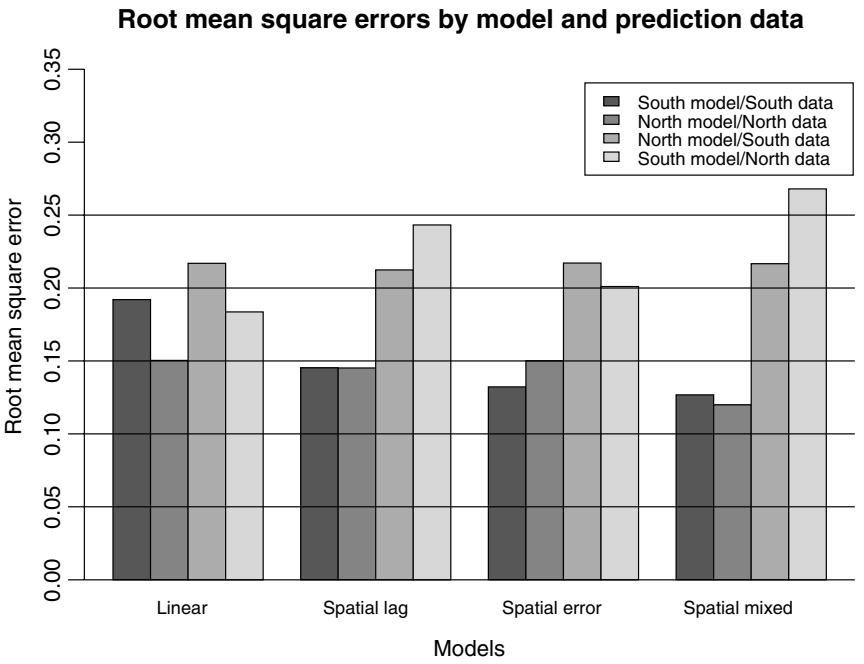

Fig. 2. Comparison of model prediction root mean square errors for four models divided north/south, Boston house price data

spatial regimes and/or non-stationarity are present. This could be held to justify the abandonment of methods not accommodating this lack of stability in parameter estimates across the chosen data set, for example by comparing the fit of a geographically weighted regression with the baseline model. This will however not be pursued here, although some indication is given of the specific behaviour of Boston city tracts is given in Fig. 1.

In the second approach, 100 samples of 250 in-sample tracts were chosen, leaving 256 tracts out-of-sample. The samples were replicated in order to get a feeling for the variations in predictions which could result. Here, the spatial weights matrices were prepared for each data set as row standardised schemes for the six nearest neighbours of each tract centre (UTM zone 19). In addition, use was made of the `gam()` function in package **mgcv** to fit a generalised additive model (see Kelsall and Diggle (1998) for a similar use of GAM). In this specification, the model fitted was:

$$\mathbf{y} = \mathbf{X}\beta + s(\text{lon},\ \text{lat}) + \varepsilon$$

where $s(\text{lon}, \text{lat})$ is a smoothing function using a penalised thin plate regression spline basis in 12 dimensions to incorporate spatial dependence. Alternative modern statistical fitting techniques could have been used, and here the joint smoothing of longitude and latitude was chosen after inspecting the results of smoothing each of them and their interaction separately. Although such fitting techniques are not typically used in spatial econometric analyses, it may be of interest to compare prediction results across such analysis-community boundaries. It can be noted that GAM

predictions in the first setting, with the data set divided into Northern and Southern parts, were very poor when predicting for new data.

Figure 3 reinforces the results of testing model predictions after dividing Boston into two parts. The linear model (lm) has the least satisfactory fit within the sample from which the model was fitted, but performs as well or better than all the spatial econometrics models when predicting for other data than those used to fit the model. The mixed spatial lag model (the Common Factor model) does best in predicting on the training set – the data it was fitted with, but worst on the test – excluded – data. This may be taken as an indication of over-fitting, capturing too much of the specificity of the spatial dependencies of the training data set. The performance of the generalised additive model is better than that of the linear model both on the training and the test data sets, despite the 'black-box' nature of the specification of the spatial pattern in this case as penalised thin plate regression spline.

5 Concluding remarks

Among the opportunities and challenges posed by trying to implement spatial econometric techniques in R in the **spdep** package have been issues raised by the object-oriented data-driven approach implicit in classes and methods. So far, old style classes and methods have been used for spatial neighbour objects, spatial weights objects, and for spatial simultaneous

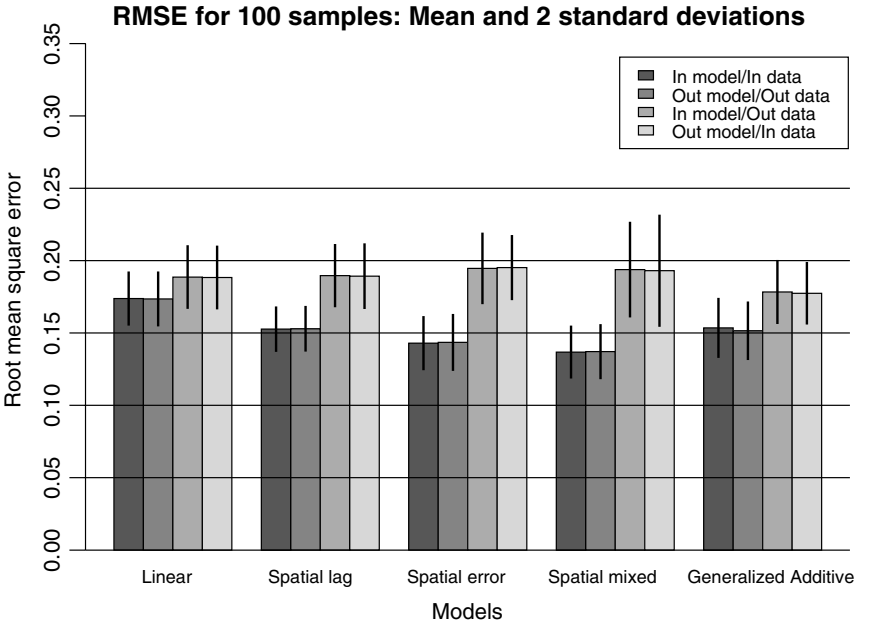

Fig. 3. Comparison of model prediction root mean square errors means and standard deviations for 100 random samples of 250 in-sample tracts and 256 out-of-sample tracts, for five models, Boston house price data

autoregressive model objects. Many of the methods usually accompanying fitted model objects are simple to write, but `predict.sarlm()` revealed areas of spatial econometrics which perhaps have received little attention hitherto. The current implementation does however need to be augmented to handle situations in which the dependencies between the locations of observations from which the model to be used for prediction was fitted, and the locations of new data observations, can be represented as a correlation structure of some kind, thus better capturing the signal component.

It does seem that Haining's partitioning of the fitted values of spatial models is of interest in itself, as indicated by the diagnostic plots in Fig. 1. It may well be that such diagnostic plots, perhaps dynamically linked to maps, will help us in establishing which further misspecification problems are present in our spatial models, shifting focus from criticising the misspecification of aspatial models to trying to construct spatial models with better properties. Haining's proposals for more general regression diagnostics for models in which spatial dependence is present do not seem to have as yet met with the acceptance they deserve (1990, 1994). Prediction for new data and new spatial weights matrices is a challenge for legacy spatial econometric models, raising the question of what spatial predictions should look like. Can for example spatial econometrics models be recast as mixed effects models, since as Pinheiro and Bates (2000) show, spatial correlation structures can "plugged" into such models?

A further consequence of examining fitted model classes and methods, in particular with regard to prediction, is to question whether we need to fit models on very large data sets. Can we not rather fit and refine them on smaller data sets and predict or interpolate to larger data sets? Housing values are not infrequently the subject of analysis, and would perhaps be an attractive target for prediction. An advantage of fitting on moderate sized data sets, maybe training sets from larger data collections, is that the use of sparse matrix techniques in some circumstances would become unnecessary. Standard errors of prediction remain open.

It also seems that a relaxation of single data set fitting of spatial econometrics models may also help to lower barriers between geostatistics and legacy spatial econometrics models when using distance criteria for representing dependence. It appears that some movement is already taking place in this regard, given the use of spatial covariance in Ord and Getis (2001) in the development of the $O_i(d)$ local spatial autocorrelation statistic allowing for global dependence. In addition, the Getis filtering approach (Getis 1995, Getis and Griffith 2002) is distance based, and seems to admit prediction to new data locations using the distance criteria and filtering functions recorded in the fitted model. The Griffith eigenfunction decomposition approach discussed in Getis and Griffith (2002), and described in detail in Griffith (2000a, 2000b), does not, however, seem to open for prediction to new locations not contiguous with the locations on which the estimated model was fitted, because of its clear focus on the eigenvectors of the spatial weights matrix of the training data set. In addition, the selection of the eigenvectors to use for filtering may not transfer between geographical settings.

Finally, focusing on prediction using spatial econometric models does concentrate attention on assumptions about spatial homogeneity, including

stationarity, support, multi-scale issues, and edge effects. Approaching modern statistical techniques as it were from the other side, we find work on geographically weighted regression (Brunsdon et al. 1996) and geographically weighted summary statistics (Brunsdon et al. 2002), in which many of these assumptions are addressed directly. In this context, it would be worthwhile to be able to test a geographically weighted regression fit against say a spatial error model fit, for instance by implementing a model comparison function like `anova(gwr.fit, sarlm.fit)`. But it is the flexibility of a language environment such as R, and the fruitfulness of class and method formalisms, that give rise to such projects for future research and implementation.

References

Anselin L (1988) *Spatial econometrics: Methods and models*. Kluwer, Dordrecht

Anselin L (1995) *SpaceStat version 1.80 user's guide*. Morgantown, WV: Regional Research Institute, West Virginia University

Becker RA, Chambers JM, Wilks AR (1988) *The New S Language*. Chapman & Hall, London

Bennett RJ, Haining RP, Griffith DA (1984) The problem of missing data on spatial surfaces. *Annals of the Association of American Geographers* 74:138–156

Bivand RS, Gebhardt A (2000) Implementing functions for spatial statistical analysis using the R language. *Journal of Geographical Systems* 2:307–317

Bivand RS, Portnov BA (2003) Exploring spatial data analysis techniques using R: the case of observations with no neighbours. In: Anselin L, Florax R, Rey S (eds) *Advances in spatial econometrics*, Springer, Berlin, Heidelberg and New York

Brunsdon C, Fotheringham AS, Charlton M (1996) Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis* 28:281–289

Brunsdon C, Fotheringham AS, Charlton M (2002) Geographically weighted summary statistics – a framework for localised exploratory data analysis. *Computers, Environment and Urban Systems* 26:501–524

Chambers JM (1998) *Programming with data*. Springer, Berlin, Heidelberg and New York

Chambers JM, Hastie TJ (1992) *Statistical models in S*. Chapman & Hall, London

Cliff, AD, Ord JK (1973) *Spatial autocorrelation*. Pion, London

Cox NJ, Jones K (1981) Exploratory data analysis. In: Wrigley N, Bennett RJ (eds) *Quantitative geography: A British view*. London, Routledge & Kegan Paul, pp. 135–143

Cressie NAC (1993) *Statistics for spatial data*. Wiley, New York

Dalgaard P (2002) *Introductory statistics with R*. Springer, Berlin, Heidelberg and New York

Diggle PJ, Tawn JA, Moyeed RA (1998) Model-based geostatistics. *Applied Statistics* 47:299–350

Getis A (1995) Spatial filtering in a regression framework: Examples using data on urban crime, regional inequality, and government expenditure. In: Anselin L, Florax RJGM (eds) *New directions in spatial econometrics*. Springer, Berlin, Heidelberg and New York, pp. 172–185

Getis A, Griffith DA (2002) Comparative spatial filtering in regression analysis. *Geographical Analysis* 34:130–140

Gilley OW, Pace RK (1996) On the Harrison and Rubinfeld data. *Journal of Environmental Economics and Management* 31:403–405

Gotway CA, Young LJ (2002) Combining incompatible spatial data. *Journal of the American Statistical Association* 97:632–648

Griffith DA (1989) Spatial regression analysis on the PC: Spatial statistics using MINITAB. Discussion Paper, Institute of Mathematical Geography, Ann Arbor, Michigan

Griffith DA (2000a) A linear regression solution to the spatial autocorrelation problem. *Journal of Geographical Systems* 2:141–156

Griffith DA (2000b) Eigenfunction properties and approximations of selected incidence matrices employed in spatial analyses. *Linear Algebra and Its Applications* 321:95–112

Griffith DA, Bennett RJ, Haining RP (1989) Statistical analysis of spatial data in the presence of missing observations – a methodological guide and an application to urban census data. *Environment and Planning A* 21:1511–1523

Griffith DA, Layne LJ (1999) *A casebook for spatial statistical data analysis: A compilation of analyses of different thematic data sets*. Oxford University Press, New York

Gould P (1970) Is *Statistix Inferens* the geographical name for a wild goose? *Economic Geography* 46:439–448

Haining RP (1981) Spatial and temporal analysis: Spatial modelling. In: Wrigley N, Bennett RJ (eds) *Quantitative geography: A British view*, London, Routledge & Kegan Paul, pp. 86–91

Haining RP (1990) *Spatial data analysis in the social and environmental sciences*. Cambridge University Press, Cambridge

Haining RP (1990) The use of added variable plots in regression modeling with spatial data. *Professional Geographer* 42:336–344

Haining RP (1994) Diagnostics for regression modeling in spatial econometrics. *Journal of Regional Science* 34:325–341

Haining RP, Griffith DA, Bennett RJ (1989) Maximum-likelihood estimation with missing spatial data and with an application to remotely sensed data. *Communications in Statistics – Theory and Methods* 18:1875–1894

Hartwig F, Dearing BE (1979) *Exploratory data analysis*. Beverly Hills, Sage

Ihaka R, Gentleman R (1996) R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics* 5:299–314

Johnson J, NiDardo J (1997) *Econometric methods*. Mc Graw Hill, New York

Kaluzny SP, Vega SC, Cardoso TP, Shelly AA (1996) *S+SPATIALSTATS users manual version 1.0*. MathSoft Inc., Seattle

Kelsall JE, Diggle PJ (1998) Spatial variation in risk of disease: A nonparametric binary regression approach. *Journal of the Royal Statistical Society, Series C* 47:449–473

Martin RJ (1984) Exact maximum likelihood for incomplete data from a correlated Gaussian process. *Communications in Statistics: Theory and Methods* 13:1275–1288

Martin RJ (1990) The role of spatial statistical processes in geographical modelling. In: Griffith DA (ed) *Spatial statistics: Past, present, and future*, Ann Arbor, Michigan, Institute of Mathematical Geography, pp. 109–127

Olsson G (1968) Complementary models: A study of colonization maps. *Geografiska Annaler B* 50:115–132

Olsson G (1970) Explanation, prediction, and meaning variance: An assessment of distance interaction models. *Economic Geography* 46:223–233

Ord JK, Getis A (2001) Testing for local spatial autocorrelation in the presence of global autocorrelation. *Journal of Regional Science* 41:411–432

Pinheiro JC, Bates DM (2000) *Mixed-effects models in S and S−PLUS*. Springer, Berlin, Heidelberg and New York

Ripley BD (1988) *Statistical inference for spatial processes*. Cambridge University Press, Cambridge

Ripley BD (1992) Applications of Monte-Carlo methods in spatial and image analysis. In: Jöckel KJ, Rothe G, Sendler W (eds) *Bootstrapping and related techniques*, Springer, Berlin, Heidelberg and New York, pp. 47–53

Stewart MB, Wallis KF (1981) *Introductory econometrics*. Blackwell, Oxford

Venables WN, Ripley BD (2000) *S programming*. Springer, Berlin, Heidelberg and New York

Venables WN, Ripley BD (2002) *Modern applied statistics with S−PLUS*. 4th ed, Springer, Berlin, Heidelberg and New York