About predictions in spatial autoregressive models: Optimal and almost optimal strategies

Michel Goulard[*], Thibault Laurent[†] and Christine Thomas-Agnan[‡]

December 10, 2016

Abstract

We address the problem of prediction in the spatial autoregressive SAR model for areal data which is classically used in spatial econometrics. With the Kriging theory, prediction using Best Linear Unbiased Predictors is at the heart of the geostatistical literature. From the methodological point of view, we explore the limits of the extension of BLUP formulas in the context of the spatial autoregressive SAR models for out-of-sample prediction simultaneously at several sites. We propose a more tractable "almost best" alternative and clarify the relationship between the BLUP and a proper EM-algorithm predictor. From an empirical perspective, we present data-based simulations to compare the efficiency of the classical formulas with the best and almost best predictions.

JEL classification: C21, C53

Key Words: Spatial simultaneous autoregressive models, out of sample prediction, best linear unbiased prediction

1 Introduction

Prediction is a basic concern in geostatistics (Cressie 1990) and, in the spatial econometrics literature, Bivand (2002) recognizes the importance of the question:

[*]e-mail: michel.goulard@toulouse.inra.fr
INRA, UMR 1201 DYNAFOR, Chemin de Borde Rouge BP52627, F31326 Castanet-Tolosan, France
[†]e-mail: thibault.laurent@tse-fr.eu
Toulouse School of Economics, CNRS, University of Toulouse Capitole, 21 allée de Brienne, 31042 Toulouse, France
[‡]e-mail: christine.thomas@tse-fr.eu
Toulouse School of Economics, University of Toulouse Capitole, 21 allée de Brienne, 31042 Toulouse, France

"Prediction for new data ... is a challenge for legacy spatial econometric models, raising the question of what a BLUP (best linear prediction) would look like".

This question appears either as a prediction problem per say or as a by product of the question of estimation of the parameters in the presence of missing observations. In the context of a linear model with a CAR (conditional autoregressive) error, Griffith et al. (1989) use the BLUP to estimate parameters of the model with missing data. Based on Haining et al. (1984), they propose an iterative process: starting from a classical OLS estimator, the missing data are predicted and then the completed data is used to re-estimate the model, then missing data are predicted again by the BLUP using the observed data and the process is continued until convergence. LeSage and Pace (2004) study the problem of missing data for the SEM and SAR models, and also for the SDM (we use the classical acronyms for spatial models as in LeSage and Pace (2009)). They use a kind of EM algorithm, which maximizes at step k the likelihood of the observed data completed by the prediction of missing data by the BLUP of the missing given the observed calculated from step $k-1$. Griffith (2010) uses a similar EM algorithm to tackle missing data in the context of a SAR. Kelejian and Prucha (2007) study the prediction specifically in the case of the SDM model. They study some classical predictors and they introduce a new one based on the knowledge of weighted sums of the neighboring values, the weights coming from the spatial weights matrix involved in the models. They compare the different predictors by computing the theoretical prediction variances, with two choices of weighted matrices from a circular scheme, one with only first nearest neighbors and a second one with 16 nearest neighbors. Pace and LeSage (2008) give an overview about prediction with the BLUP. Kato (2008, 2013) explores the best linear prediction problem in the framework of spatial error models and using a simulation experiment. He follows a first study of Dubin (2003) comparing the prediction process using either a geostatistical model or a SEM. In the Dubin paper, mispecification of the model is taken into account in terms of estimation and prediction. Note however that Dubin uses the BLUP with a geostatistical model but not with the SEM where an ad-hoc predictor is introduced. Kato (2008, 2013) makes a systematic comparison of the BLUP in the case of SEM and points out the two aspects: estimation with missing data and prediction. Two kinds of spatial estimations are performed: the first one uses maximum likelihood estimation of a marginal SEM model based on the global model for observed data and missing data, and the second uses an EM-like algorithm similar to that of LeSage and Pace (2004).

We first present the different types of prediction situations encountered according to whether we predict at a sample unit or an out-of-sample one and to whether one or several points are predicted simultaneously. In-sample prediction is used as a measure of model fit like in the coefficient of determination and some graphical diagnostics. To motivate the need for out-of-sample prediction in the spatial framework, let us present the context of a case study in Lesne et al. (2008). Until 1999, the French population census was exhaustive and realized by the French statistical institute (INSEE) approximately every ten years. Since 2004, this exhaustive census has been replaced by a census survey which consists in annual samples and delivers an up-to-date information. In particular, the communes with less than 10,000

inhabitants at the 1999 census (called *small communes*) are sampled exhaustively every five year at the rate of one fifth per year. The sampling design of these small communes is stratified by region and inside each region, the small communes are partitioned into five rotational groups by using a balanced sample design taking into account some auxiliary socio-economics variables given by the 1999 census. Between 2004 and 2009, polling organizations needed an estimate of the population for all the small communes and of its evolution since the previous complete census of 1999. The population of all the small communes would not be delivered by the INSEE before 2009 but data sets containing the population of the first two rotational groups, corresponding to 2004 and 2005, were already known and could be used to predict the population of the other three rotational groups. In that case, out-of-sample prediction formulas were necessary for spatial models. Figure 1 presents the locations of the spatial units where population data was available at the time of this case study. We will base the simulations on the same territory.

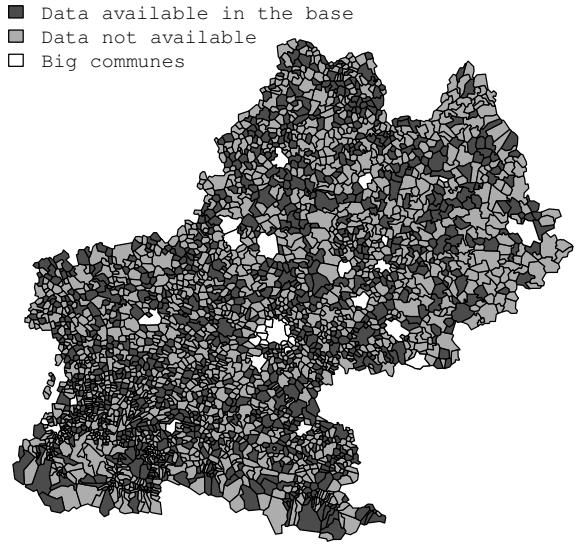

Figure 1: Spatial units where population data was available at the time of the Lesne *et al.* study.

Our objective in the present work is to study the behavior of the BLUP in the context of a SAR model. After a state-of-the-art overview of the spatial predictions in the SAR framework, our first contribution is to gather in the same place the different formulas used in the literature with a unified notation thus facilitating their comparison by the community. In the same spirit, we make our simulation code available as online supplemental material. All the prediction formulas cited in our paper are implemented in the releases later than 0.5-92 of the spdep R package

(Bivand and Piras, 2015), following the GSoC 2015 project by M. Gubri[1]. The second contribution is to present simulation experiment results as in Kato (2008, 2013), which yield a vector of observed values of the dependent variables and a collection of sites where its value is unobserved, in a large set of configurations. The third contribution is to introduce new variants of out-of-sample prediction formulas among which one extends work by Kelejian and Prucha (2007) and to clarify what is the relationship between BLUP and a proper EM approach in the SAR model.

In Section 2, we first review the classical prediction formulas encountered in the literature for the spatial simultaneous autoregressive (SAR or LAG depending on authors) models. Then, in Section 3, we recall how best linear unbiased prediction (BLUP) can be done in the framework of these models using an adapted formulation of the Goldberger formula. We introduce several alternatives and finally demonstrate that the classical formulas can thus be improved upon substantially. Section 4 presents our simulations results.

2 State of the art about best prediction in spatial autoregressive SAR models

2.1 Models and prediction situations

We consider prediction in the classical homoscedastic spatial autoregressive SAR model (SAR model hereafter). Given a spatial weight matrix $\mathbf{W}$ and exogenous variables $\mathbf{X}$, this model can be written

$$\mathbf{Y} = \rho\mathbf{W}\mathbf{Y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbb{E}(\boldsymbol{\epsilon} \mid \mathbf{X}) = \mathbf{0}$. There is no need for a normality assumption of the error term in the sequel, except when we use the gaussian likelihood, for example in the EM algorithm section. In reduced form, this is equivalent to

$$\mathbf{Y} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \rho\mathbf{W})^{-1}\boldsymbol{\epsilon}. \tag{2}$$

Let us recall a few classical facts about this model. The conditional mean of $\mathbf{Y}$ in this model is given by

$$\boldsymbol{\mu} = \mathbb{E}(\mathbf{Y} \mid \mathbf{X}) = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta} \tag{3}$$

and its covariance structure by

$$\boldsymbol{\Sigma} = \mathbb{V}ar(\mathbf{Y} \mid \mathbf{X}) = [(\mathbf{I} - \rho\mathbf{W}')(\mathbf{I} - \rho\mathbf{W})]^{-1}\sigma^2. \tag{4}$$

The precision matrix $\mathbf{Q}$ is then easily derived

$$\mathbf{Q} = \boldsymbol{\Sigma}^{-\mathbf{1}} = \frac{1}{\sigma^2}(\mathbf{I} - \rho\mathbf{W}')(\mathbf{I} - \rho\mathbf{W}). \tag{5}$$

[1] https://www.google-melange.com/gsoc/project/details/google/gsoc2015/framartin/5717271485874176

When ρ is known, the best linear unbiased estimator (BLUE) of $\boldsymbol{\mu} = (\mathbf{I}-\rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta}$ is $\hat{\boldsymbol{\mu}} = (\mathbf{I}-\rho\mathbf{W})^{-1}\mathbf{X}\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I}-\rho\mathbf{W})\mathbf{Y}$ is the best linear unbiased estimator of $\boldsymbol{\beta}$ as well as its maximum likelihood estimator in the gaussian case.

We will distinguish two types of prediction situations: the in-sample and out-of-sample cases. In the in-sample prediction problem, we have n spatial units for which we observe the dependent variable $\mathbf{Y}$ as well as the independent variables $\mathbf{X}$ and we want to predict the value of $\mathbf{Y}$ at the observed sites after fitting the model which is the same as computing the fitted value of $\mathbf{Y}$. These predicted values can be used for example to compute a goodness of fit criterion. In the out-of-sample case, we have two types of spatial units: the in-sample units for which we observe the dependent variable $\mathbf{Y_S}$ as well as the independent variable $\mathbf{X_S}$ and the out-of-sample units for which we only observe the independent variable $\mathbf{X_O}$ and we want to predict the variable $\mathbf{Y_O}$ from the knowledge of $\mathbf{Y_S}, \mathbf{X_S}$ and $\mathbf{X_O}$. In the out-of-sample case, we will further distinguish according to the number of spatial units to be predicted simultaneously: if there is only one such unit, we will talk about a single out-of-sample prediction case, otherwise about a multiple out-of-sample prediction case.

2.2 Submodels for in-sample and out-of-sample units

Let n_O and n_S denote respectively the number of out-of sample and in-sample units with $n = n_O+n_S$. As in Kato (2008), we partition $\mathbf{X}$ and $\mathbf{Y}$ in $\mathbf{X} = (\mathbf{X_S}' \mid \mathbf{X_O}')'$ and $\mathbf{Y} = (\mathbf{Y_S}' \mid \mathbf{Y_O}')'$ where $\mathbf{X_S}$ (resp $\mathbf{Y_S}$) of dimension $n_S \times p$ (resp: $n_S \times 1$) denote the matrix of components of $\mathbf{X}$ (resp the vector of components of $\mathbf{Y}$) corresponding to in-sample spatial units and $\mathbf{X_O}$ (resp $\mathbf{Y_O}$) of dimension $n_O \times p$ (resp: $n_O \times 1$) denote the matrix of components of $\mathbf{X}$ (resp the vector of components of $\mathbf{Y}$) corresponding to out-of-sample spatial units. Similarly $\boldsymbol{\mu} = (\boldsymbol{\mu_S}' \mid \boldsymbol{\mu_O}')'$. More generally in this paper, when $\mathbf{J}$ denotes a set of indices, the vector $\mathbf{Z_J}$ will denote the vector of components of $\mathbf{Z}$ relative to the indices in $\mathbf{J}$. For the case of the spatial weights matrix, variance and precision matrices, we will need a double index for extraction: for two sets of indices $\mathbf{I}$ and $\mathbf{J}$, and a matrix $\mathbf{A}$, the matrix $\mathbf{A_{IJ}}$ will denote the bloc extracted from $\mathbf{A}$ by selecting the rows corresponding to row indices in $\mathbf{I}$ and column indices in $\mathbf{J}$. For example, we partition the spatial weights matrix $\mathbf{W}$ as follows

$$\mathbf{W} = \begin{pmatrix} \mathbf{W_{SS}} & \mathbf{W_{SO}} \\ \mathbf{W_{OS}} & \mathbf{W_{OO}} \end{pmatrix}, \tag{6}$$

where

- $\mathbf{W_{SS}}$ is the $n_S \times n_S$ submatrix corresponding to the neighborhood structure of the n_S in-sample sites,
- $\mathbf{W_{OO}}$ the $n_O \times n_O$ submatrix corresponding to the neighborhood structure of the n_O out-of-sample sites,

- $\mathbf{W_{OS}}$ the $n_O \times n_S$ submatrix indicating the neighbors of the out-of-sample units among the in-sample units
- $\mathbf{W_{SO}}$ the $n_S \times n_O$ submatrix indicating the neighbors of the in-sample units among the out-of-sample units.

Because for out-of-sample prediction, we need to relate the model driving the in-sample units to the out-of-sample ones, we assume there is an overall model driving the in-sample and out-of-sample units. The overall model M is given like (1) for the n observations of $(\mathbf{X}, \mathbf{Y})$:

$$\begin{pmatrix} \mathbf{Y_S} \\ \mathbf{Y_O} \end{pmatrix} = \rho \begin{pmatrix} \mathbf{W_{SS}} & \mathbf{W_{SO}} \\ \mathbf{W_{OS}} & \mathbf{W_{OO}} \end{pmatrix} \begin{pmatrix} \mathbf{Y_S} \\ \mathbf{Y_O} \end{pmatrix} + \begin{pmatrix} \mathbf{X_S} \\ \mathbf{X_O} \end{pmatrix} \boldsymbol{\beta} + \begin{pmatrix} \boldsymbol{\epsilon_S} \\ \boldsymbol{\epsilon_O} \end{pmatrix}.$$

The sub-model M_S driving the data $(\mathbf{X_S}, \mathbf{Y_S})$ corresponding to the sample units follows the equation

$$\mathbf{Y_S} = [(\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta}]_{\mathbf{S}} + [(I - \rho\mathbf{W})^{-1}\boldsymbol{\epsilon}]_{\mathbf{S}}, \tag{7}$$

where the error term has a variance equal to $(Var(\mathbf{Y}))_{\mathbf{SS}}$. However in practice this model cannot be used for estimating the parameters since only in-sample units are available. On the other hand, the following simplified model only based on in-sample units

$$\mathbf{Y_S} = (\mathbf{I} - \rho\mathbf{W_{SS}})^{-1}\mathbf{X_S}\boldsymbol{\beta} + \boldsymbol{\epsilon}_S \tag{8}$$

can be considered as a feasible approximation to (7) after row-normalization of $\mathbf{W_{SS}}$ and corresponds to the natural model people would use in this circumstance. Exact compatibility of the two models would imply the following two constraints $((\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X})_{\mathbf{S}} = (\mathbf{I} - \rho\mathbf{W_{SS}})^{-1}\mathbf{X_S}$ for the mean and $(Var(\mathbf{Y}))_{\mathbf{SS}} = Var(\mathbf{Y_S})$ for the variance. First note that these two restrictions are not so strong as appeared when we tested them on the simulations. Moreover they are very similar to the approximations made by Kato (2013) (see section 3.3) in his EM approach. Finally the EM approach proposed in section 3.3 does not require these restrictions and leads to very similar results as the BLUP based on models (7) and (8).
It is important to note that while a corresponding decomposition of the precision matrix is easily derived from (5) and is an easy combination of extractions from $\mathbf{W}$, the covariance matrix for sub-model M_S on the other hand is not an extraction of $\boldsymbol{\Sigma}$ because of the inversion in formula (4).

2.3 Classical prediction formulas

2.3.1 Goldberger formula

Goldberger (1962) proposed a formula for prediction in the framework of a general linear model $\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$ with **known** variance structure $\boldsymbol{\Sigma} = Var(\boldsymbol{\epsilon})$. The Golberger formula (1962) gives the BLUP (Best Linear Unbiased Predictor) as a linear predictor $\hat{\mathbf{Y}}_{\mathbf{O}}^{BP} := \boldsymbol{\lambda}'\mathbf{Y_S}$, where $\boldsymbol{\lambda} \in \mathbb{R}^n$ minimizes $\mathbb{E}(\hat{\mathbf{Y}}_{\mathbf{O}}^{BP} - \mathbf{Y_O})^2$ under the unbiasedness constraint that $\mathbb{E}(\hat{\mathbf{Y}}_{\mathbf{O}}^{BP} - \mathbf{Y_O}) = \mathbf{0}$ yielding

$$\hat{\mathbf{Y}}_{\mathbf{O}}^{BP} = \hat{\boldsymbol{\mu}}_{\boldsymbol{O}} + Cov(\mathbf{Y_O}, \mathbf{Y_S})Var(\mathbf{Y_S})^{-1}(\mathbf{Y_S} - \hat{\boldsymbol{\mu}}_{\mathbf{S}}),$$

where $\hat{\boldsymbol{\mu}}_{\mathbf{O}}$ and $\hat{\boldsymbol{\mu}}_{\mathbf{S}}$ are the Best Linear Unbiased Estimators respectively of $\mathbb{E}(\mathbf{Y_O})$ and $\mathbb{E}(\mathbf{Y_S})$. In practice, one does not know the theoretical variance $\boldsymbol{\Sigma}$ (i.e. ρ in the SAR model) and one needs to replace it in the formula by an estimator. To simplify, by a slight abuse of language, we will call BLUP as well the predictor obtained by substituting the estimated variance since the real BLUP is not feasible. It is the application of this formula which has given rise to the famous Kriging predictor in geostatistics. In fact Golberger (1962) gave the formula for a set $\mathbf{O}$ reduced to a point but the formula remains true for a set of points $\mathbf{O}$. In that case the problem is to find $\hat{\mathbf{Y}}_{\mathbf{O}}^{BP} = \boldsymbol{\Lambda}'\mathbf{Y_S}$ minimizing $\mathrm{Tr}(\mathbb{E}(\hat{\mathbf{Y}}_{\mathbf{O}}^{BP} - \mathbf{Y_O})(\hat{\mathbf{Y}}_{\mathbf{O}}^{BP} - \mathbf{Y_O})')$ under the constraint that $\mathbb{E}(\hat{\mathbf{Y}}_{\mathbf{O}}^{BP} - \mathbf{Y_O}) = \mathbf{0}$ where $\boldsymbol{\Lambda}$ is a matrix. Note that the matrix formulation is equivalent to applying the Goldberger formula one point at a time. Let us emphasize the fact that the Goldberger formula applies as soon as a model can be written in a classical general linear model form which is the case for the SAR model in reduced form with $\boldsymbol{\mu}$ given by (3) and $\boldsymbol{\Sigma}$ given by (4).

2.4 Another formulation of Goldberger formula for SAR models

Using algebra results from Harville (1997), the Goldberger formula can be written in terms of the precision matrix $\mathbf{Q}$, as in the prediction formula for Markov gaussian vector field of Rue and Held (2005, page 31). As LeSage and Pace (2008) point out, it is based on the fact that $Cov(\mathbf{Y_O}, \mathbf{Y_S})Var(\mathbf{Y_S})^{-1} = -\mathbf{Q_{OO}}^{-1}\mathbf{Q_{OS}}$, which arises from expressing that the partitioned matrix $\mathbf{Q}$ is the inverse of the partitioned matrix $Var(\mathbf{Y})$. The Goldberger formula can thus be expressed in terms of precision matrices as follows

$$\hat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{BP}} = \hat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{TC}} - \mathbf{Q_{OO}}^{-1}\mathbf{Q_{OS}} \times (\mathbf{Y_S} - \hat{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{TC}}), \tag{9}$$

with

$$\mathbf{Q} = \frac{1}{\sigma^2}(\mathbf{I} - \rho(\mathbf{W}' + \mathbf{W}) + \rho^2\mathbf{W}'\mathbf{W}) = \begin{pmatrix} \mathbf{Q_{SS}} & \mathbf{Q_{SO}} \\ \mathbf{Q_{OS}} & \mathbf{Q_{OO}} \end{pmatrix}.$$

Let us note that the size of the matrix to invert, $\mathbf{Q_{OO}}$, is the number of out-of-sample units whereas in the first version of the Goldberger formula, the size of the matrix to invert is equal to the number of in-sample units. If the size of the matrix to be inverted is a crucial point, then using the precision based formula instead of the based variance one can help. Moreover, this formulation is to be preferred in the SAR model since the precision matrix is a quadratic function of the weight matrix whereas the covariance matrix requires an inversion.

2.4.1 In-sample prediction

For the in-sample prediction problem, we consider that the sample units are driven by equation (8). To emphasize the fact that in-sample predictions for $\mathbf{Y_S}$ will be different from out-of-sample predictions for the same vector, we will denote $\check{\mathbf{Y}}_{\mathbf{S}}$ the

in-sample predictions and $\hat{\mathbf{Y}}_\mathbf{S}$ the out-of-sample ones. In an ordinary linear model which is model (1) for $\rho = 0$, the BLUE of $\boldsymbol{\mu}_\mathbf{S}$ is given by $\mathbf{X}_\mathbf{S}\hat{\boldsymbol{\beta}}$ and the classical in-sample predictor of $\mathbf{Y}$ is

$$\check{\mathbf{Y}}_\mathbf{S}^\mathbf{T} = \mathbf{X}_\mathbf{S}\hat{\boldsymbol{\beta}}, \tag{10}$$

where $\hat{\boldsymbol{\beta}}$ is the classical estimator of $\boldsymbol{\beta}$ calculated by fitting the ordinary linear model with in-sample units $\mathbf{Y}_\mathbf{S} = \mathbf{X}_\mathbf{S}\beta + \boldsymbol{\epsilon}_\mathbf{S}$.

It is then easy and natural to imagine a predictor for the general case $\rho \neq 0$ which we will call the "trend corrected predictor" given by

$$\check{\mathbf{Y}}_\mathbf{S}^\mathbf{TC} = (\mathbf{I} - \hat{\rho}\mathbf{W}_\mathbf{SS})^{-1}\mathbf{X}_\mathbf{S}\hat{\boldsymbol{\beta}} := \hat{\boldsymbol{\mu}}_\mathbf{S}, \tag{11}$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\rho}$ are the estimators of $\boldsymbol{\beta}$ and ρ calculated by fitting the model by gaussian maximum likelihood (ML) with in-sample units. This predictor is used for example in the LeSage matlab toolbox for computing the in-sample predicted values. Note however that this one does not possess any kind of optimality property.

Another predictor introduced by Haining (1990) and detailed by Bivand (2002) is given by

$$\check{\mathbf{Y}}_\mathbf{S}^\mathbf{TS} = \mathbf{X}_\mathbf{S}\hat{\boldsymbol{\beta}} + \hat{\rho}\mathbf{W}_\mathbf{SS}\mathbf{Y}_\mathbf{S}. \tag{12}$$

Thereafter, we call this predictor the "trend-signal-noise" predictor. This one is used in the Bivand R package **spdep**.

Inspired by an idea found in Gaetan and Guyon (2010) for the case of CAR models and LeSage and Pace (2004) for the case of SEM models, we could propose an alternative in-sample predictor for the SAR model. This idea consists in using the BLUP of each in-sample unit value Y_i based on the knowledge of the remaining in-sample units and yields in matrix form the following formula

$$\check{\mathbf{Y}}_\mathbf{S}^\mathbf{BP} = (\mathbf{I} - \hat{\rho}\mathbf{W}_\mathbf{SS})^{-1}\mathbf{X}_\mathbf{S}\hat{\boldsymbol{\beta}} - \mathrm{Diag}(\hat{\mathbf{Q}}_\mathbf{SS})^{-1}\tilde{\hat{\mathbf{Q}}}_\mathbf{SS}(\mathbf{Y}_\mathbf{S} - (\mathbf{I} - \hat{\rho}\mathbf{W}_\mathbf{SS})^{-1}\mathbf{X}_\mathbf{S}\hat{\boldsymbol{\beta}}). \tag{13}$$

where $\mathrm{Diag}(\hat{\mathbf{Q}}_\mathbf{SS})$ denotes the diagonal matrix containing the diagonal of the precision matrix $\hat{\mathbf{Q}}_\mathbf{SS}$ of the SAR model given by $\hat{\mathbf{Q}}_\mathbf{SS} = \frac{1}{\hat{\sigma}^2}(\mathbf{I} - \hat{\rho}\mathbf{W}_\mathbf{SS}')(\mathbf{I} - \hat{\rho}\mathbf{W}_\mathbf{SS})$, $\hat{\sigma}^2$ is the gaussian maximum likelihood estimate of the variance, and $\tilde{\hat{\mathbf{Q}}}_\mathbf{SS} = \hat{\mathbf{Q}}_\mathbf{SS} - \mathrm{Diag}(\hat{\mathbf{Q}}_\mathbf{SS})$. The index BP in the notation is to recall that this formula is based on some kind of best prediction practice. Using a coordinate formulation rather than a matrix one, this formula is equivalent to

$$\check{Y}_i^{BP} = \hat{\mu}_i - \sum_{j=1, j\neq i}^{n_S} \frac{\hat{q}_{ij}}{\hat{q}_{ii}}(Y_j - \hat{\mu}_j), \tag{14}$$

where $\hat{q}_{ij}$ is the (i, j) element of matrix $\hat{\mathbf{Q}}_\mathbf{SS}$ and $\hat{\mu}_i$ are the components of $\hat{\boldsymbol{\mu}}_\mathbf{S}$ given by (11) which is the formula used in LeSage and Pace (2004) for the case of the SEM model with a $\hat{\boldsymbol{\mu}}$ adapted to the SEM model. Table 1 summarizes the different formulas.

Table 1: In-sample predictors formulas

Predictor	Formula
BP	$\check{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{BP}} = (\mathbf{I} - \rho\mathbf{W_{SS}})^{-1}\mathbf{X_S}\hat{\boldsymbol{\beta}} - \mathrm{Diag}(\hat{\mathbf{Q}}_{\mathbf{SS}})^{-1}\tilde{\hat{\mathbf{Q}}}_{\mathbf{SS}}(\mathbf{Y} - (\mathbf{I} - \rho\mathbf{W_{SS}})^{-1}\mathbf{X_S}\hat{\boldsymbol{\beta}})$
TS	$\check{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{TS}} = \mathbf{X_S}\hat{\boldsymbol{\beta}} + \hat{\rho}\mathbf{W_{SS}}\mathbf{Y_S}$
TC	$\check{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{TC}} = (\mathbf{I} - \hat{\rho}\mathbf{W})^{-1}_{SS}\mathbf{X_S}\hat{\boldsymbol{\beta}}$

2.4.2 Out-of-sample prediction

The trend-signal-noise predictor $\check{\mathbf{Y}}^{\mathbf{TS}}$ cannot be extended to the case of out-of-sample prediction since it requires some values of $\mathbf{Y_O}$ which are unobserved. However in the case of a single prediction for unit o, it is possible to compute it because of the presence of zeros on the diagonal of $\mathbf{W}$, which yields

$$\hat{Y}_o^{TS^1} = \mathbf{X_o}\hat{\boldsymbol{\beta}} + \hat{\rho}\mathbf{W_{oS}}\mathbf{Y_S}. \tag{15}$$

The trend-corrected strategy can be extended here because it only involves the values of $\mathbf{X}$ (and not $\mathbf{Y}$) for the out-of-sample units

$$\hat{\mathbf{Y}}^{\mathbf{TC}} := \begin{pmatrix} \hat{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{TC}} \\ \hat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{TC}} \end{pmatrix} = (\mathbf{I} - \hat{\rho}\mathbf{W})^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} \tag{16}$$

and we get

$$\begin{array}{rcl} \hat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{TC}} & = & -(\mathbf{D} - \mathbf{CA}^{-1}\mathbf{B})^{-1}\mathbf{CA}^{-1}\mathbf{X_S}\hat{\boldsymbol{\beta}} + (\mathbf{D} - \mathbf{CA}^{-1}\mathbf{B})^{-1}\mathbf{X_O}\hat{\boldsymbol{\beta}} \\ \hat{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{TC}} & = & (\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C})^{-1}\mathbf{X_S}\hat{\boldsymbol{\beta}} - (\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C})^{-1}\mathbf{BD}^{-1}\mathbf{X_O}\hat{\boldsymbol{\beta}} \end{array} \tag{17}$$

for $(\mathbf{I} - \hat{\rho}\mathbf{W}) = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} = \begin{pmatrix} \mathbf{I_{n_S}} - \hat{\rho}\mathbf{W_{SS}} & -\hat{\rho}\mathbf{W_{SO}} \\ -\hat{\rho}\mathbf{W_{OS}} & \mathbf{I_{n_O}} - \hat{\rho}\mathbf{W_{OO}} \end{pmatrix}$.

Note that $\hat{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{TC}}$ in (16) denotes something different from $\check{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{TC}}$ (11) because here the in-sample prediction takes into account the out-of-sample units. One can check that the $\hat{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{TC}}$ from (16) coincides with the predictor from (11) when there is no out-of-sample unit.

Kelejian and Prucha (2007) use Goldberger formula for single out-of-sample prediction in the particular case when $O = \{o\}$ and replacing $\mathbf{Y_S}$ by $\mathbf{W_{o.}Y}$, where $\mathbf{W_{o.}}$ is row o of matrix $\mathbf{W}$. Griffith (2010) proposes an EM procedure combining estimation of spatial parameters and imputation of missing values in the framework of the spatial filtering method (Griffith 2003). Let us mention that the information set associated to these predictors are different: for $\hat{\mathbf{Y}}^{\mathbf{TC}}$, it is $\{\mathbf{X}, \mathbf{W}\}$, for $\hat{\mathbf{Y}}_{\mathbf{o}}^{\mathbf{TS^1}}$ it is $\{\mathbf{X}, \mathbf{W}, \mathbf{Y_S}\}$.

3 Out-of-sample prediction: extensions and new proposals

3.1 Extension of the Kelejian-Prucha predictor for the SAR model

In the framework of the SAR model we first propose to generalize the Kelejian-Prucha (2007, see equation (8) page 367) approach to multiple prediction where $\mathbf{Y_O}$ is predicted by a linear combination of $\mathbf{W_{OS}Y_S}$ instead of $\mathbf{Y_S}$. Compared to our present framework, the approach of Kelejian and Prucha (2007) also includes the spatially correlated error term model (SEM) but there is a single out-of-sample unit and the comparisons are made in dimension one based on theoretical prediction errors. Indeed, it is easy to extend to the case of several out-of-sample units. The information set is then $\{\mathbf{X}, \mathbf{W}, \mathbf{W_{OS}Y_S}\}$. In that case, the classical formula for gaussian conditional expectation gives

$$\mathbb{E}(\mathbf{Y_O} \mid \mathbf{W_{OS}Y_S}) = \mathbb{E}(\mathbf{Y_O}) + \mathbf{\Sigma_{OS}W_{OS}}'(\mathbf{W_{OS}\Sigma_{SS}W_{OS}}')^{-1}(\mathbf{W_{OS}Y_S} - \mathbf{W_{OS}}\mathbb{E}(\mathbf{Y_S}). \tag{18}$$

When we substitute ρ by $\hat{\rho}$, σ by $\hat{\sigma}$ and finally $\mathbb{E}(\mathbf{Y_O})$ and $\mathbb{E}(\mathbf{Y_S})$ by their natural estimators, we obtain

$$\hat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{BPw}} = \hat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{TC}} + \mathbf{\Sigma_{OS}W_{OS}}'(\mathbf{W_{OS}\Sigma_{SS}W_{OS}}')^{-1}(\mathbf{W_{OS}Y_S} - \mathbf{W_{OS}}\hat{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{TC}}) \tag{19}$$

. However we believe that it is unlikely in practical situations that one has the information about the linear combination of neighboring values $\mathbf{W_{OS}Y_S}$ without having the entire knowledge of $\mathbf{Y_S}$. Using the linear combination $\mathbf{W_{OS}Y_S}$ instead of the full vector $\mathbf{Y_S}$ can only result in a useless loss of information. Moreover, if we compare formula (19) to (9), the size of the matrix to invert is equal to the number of out-of-sample units in both cases but (19) uses the variance matrix $\mathbf{\Sigma}$ which has to be computed by inversion from the precision matrix $\mathbf{Q}$ whereas (9) directly uses $\mathbf{Q}$ which is easily computed from $\mathbf{W}$, ρ and σ (see (5)).

For this reason, we propose the following alternative which consists in using the second version of the Golberger formula for a case where the set $\mathbf{S}$ is replaced by $\mathbf{N}$, where $\mathbf{N}$ is the set of all sites in $\mathbf{S}$ which are neighbors in the sense of $\mathbf{W}$ of at least one site in $\mathbf{O}$. The idea is to only use among the sample locations the neighbors of the out-of-sample sites in order to predict: let $\mathbf{J}$ be the set of indices of such neighbors and n_J its size. When necessary, $\mathbf{J}$ will be denoted $\mathbf{J}(\mathbf{O})$ to indicate the dependence upon $\mathbf{O}$. Let $\mathbf{W}_{\{\mathbf{J},\mathbf{O}\}}$ be the neighborhood matrix for sites which are in $\mathbf{J}$ or $\mathbf{O}$.

$$\mathbf{W}_{\{\mathbf{J},\mathbf{O}\}} = \begin{array}{c} \begin{array}{cc} \overset{n_J}{\longleftrightarrow} & \overset{n_O}{\leftrightarrow} \end{array} \\ \left(\begin{array}{c|c} \mathbf{W_{JJ}} & \mathbf{W_{JO}} \\ \hline \mathbf{W_{OJ}} & \mathbf{W_{OO}} \end{array}\right) \end{array} \begin{array}{l} \\ \updownarrow n_J \\ \\ \updownarrow n_O \end{array} .$$

The corresponding partition of the precision matrix corresponding to sites in $\{\mathbf{J}, \mathbf{O}\}$ is

$$\hat{\mathbf{Q}}_{\{\mathbf{J},\mathbf{O}\}} = \frac{1}{\hat{\sigma}^2}(\mathbf{I}_{\mathbf{n_J+p}} - \hat{\rho}(\mathbf{W}_{\{\mathbf{J},\mathbf{O}\}} + \mathbf{W}_{\{\mathbf{J},\mathbf{O}\}}{}') + \hat{\rho}^2(\mathbf{W}_{\{\mathbf{J},\mathbf{O}\}}{}'\mathbf{W}_{\{\mathbf{J},\mathbf{O}\}})) = \begin{pmatrix} \hat{\mathbf{Q}}_{\mathbf{JJ}} & \hat{\mathbf{Q}}_{\mathbf{JO}} \\ \hat{\mathbf{Q}}_{\mathbf{OJ}} & \hat{\mathbf{Q}}_{\mathbf{OO}} \end{pmatrix}$$

and thus we get the following feasible predictor

$$\hat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{BP_N}} = \hat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{TC}} - \hat{\mathbf{Q}}_{\mathbf{OO}}^{-1}\hat{\mathbf{Q}}_{\mathbf{OJ}}(\mathbf{Y_J} - \hat{\mathbf{Y}}_{\mathbf{J}}^{\mathbf{TC}}), \tag{20}$$

where $\hat{\mathbf{Y}}_{\mathbf{J}}^{\mathbf{TC}}$ is obtained by extracting the rows corresponding to units in J from $\hat{\mathbf{Y}}^{\mathbf{TC}}$. The advantage of this predictor lies in the fact that it reduces the computational burden since the size of the matrix $\hat{\mathbf{Q}}_{\mathbf{OO}}^{-1}\hat{\mathbf{Q}}_{\mathbf{OJ}}$ is $n_O \times n_J$ instead of $n_O \times n_S$ for the matrix $\hat{\mathbf{Q}}_{\mathbf{OO}}^{-1}\hat{\mathbf{Q}}_{\mathbf{OS}}$. The improvement is particularly interesting when W is sparse. When W is dense, as J is closest to S, the advantage is relatively small. In that case, an alternative for a larger gain in computation, but with a risk of loss of efficiency, would be to only keep in J the neighbors of O with the highest weights (for a given threshold).

A variance based version of this new predictor can be written

$$\hat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{BP_N}} = \hat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{TC}} + \hat{Cov}(\mathbf{Y_O}, \mathbf{Y_J})\hat{Var}(\mathbf{Y_J})^{-1}(\mathbf{Y_J} - \hat{\mathbf{Y}}_{\mathbf{J}}^{\mathbf{TC}}).$$

However note that, because of the simple expression of the precision matrix as a function of the weight matrix in the SAR model, extractions of the precision matrix are directly linked to extractions of the weight matrix. Hence using a precision matrix version of Goldberger as we do is much easier than using a covariance matrix version of Goldberger as in Kelejian and Prucha (2007).

Clearly the new predictor is not optimal, but one can hope it has some almost optimality behavior. Our proposition can be related to the classical "kriging with moving neighborhood" which is often used in geostatistics. In the framework of spatial error models (herefater SEM models), Kato (2008) uses the same best prediction approach but substitutes to the ML parameters estimators some approximations similar to the ones we describe in section 2.2. Note that because of the links between $\mathbf{W}$ and $\mathbf{Q}$, if we now replace $\mathbf{J}$ by $\mathbf{J'}$, where $\mathbf{J'}$ denotes the set of indices of sites in $\mathbf{S}$ which are neighbors of at least one site in $\mathbf{O}$ in the sense of $\mathbf{W'W}$, i.e. second order $\mathbf{W}$ neighbors, then the predictor $\hat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{BP_N}}$ will be optimal.

Indeed if $S \setminus J$ is the set notation for S deprived from $S \cap J$, we have that $\hat{\mathbf{Q}}_{\mathbf{OS}\setminus\mathbf{J'}} = 0$ and thus

$$\hat{\mathbf{Q}}_{\mathbf{OO}}^{-1}\hat{\mathbf{Q}}_{\mathbf{OS}} \times (\mathbf{Y_S} - \hat{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{TC}}) = \hat{\mathbf{Q}}_{\mathbf{OO}}^{-\mathbf{1}}\hat{\mathbf{Q}}_{\mathbf{OJ'}} \times (\mathbf{Y_{J'}} - \hat{\mathbf{Y}}_{\mathbf{J'}}^{\mathbf{TC}})$$

and therefore the predictor $\hat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{BP_N}}$ based on the second order $\mathbf{W}$ neighbors is exactly optimal.

3.2 Alternative: back to single prediction

Because the single prediction formulas are simpler, when p out-of-sample units have to be predicted, we propose to apply the "single out-of-sample" formula to each of

the out-of-sample unit separately, ignoring at each stage the remaining $p-1$ units. This allows also to include the Trend-signal strategy which only exists in the single prediction case. This leads us to define alternatives of each of the five predictors $\hat{\mathbf{Y}}^{\mathbf{TC}}$, $\hat{\mathbf{Y}}^{\mathbf{TS}}$, $\hat{\mathbf{Y}}^{\mathbf{BP}}$, $\hat{\mathbf{Y}}^{\mathbf{BP_W}}$ and $\hat{\mathbf{Y}}^{\mathbf{BP_N}}$. The resulting predictors for location o will be denoted by $\hat{Y}_o^{TC^1}$, $\hat{Y}_o^{TS^1}$, $\hat{Y}_o^{BP^1}$, $\hat{Y}_o^{BP_W{}^1}$ and $\hat{Y}_o^{BP_N{}^1}$. The precise formulas are detailed in Table 2. These formulas of course do not apply if an out-of-sample point has no neighbors among the sample units but in that situation a non-spatial formula is doing just as well.

3.3 EM approach

The EM algorithm (Dempster et al. 1977) is meant for implementing maximum likelihood in the case of incomplete data which is our case since $\mathbf{Y_S}$ is observed whereas $\mathbf{Y_O}$ is not. Let us briefly recall that the original EM algorithm (Dempster et al. 1977) involves two steps called E-step and M-step. For incomplete observations ($\mathbf{Y_S}$ observed and $\mathbf{Y_O}$ not observed) and parameter $\boldsymbol{\theta}$, the E-step is the computation of the expected likelihood function,

$$H(\boldsymbol{\theta_1}, \boldsymbol{\theta}) = \mathbb{E}(L(\mathbf{Y}|\boldsymbol{\theta_1})|\mathbf{Y_S}, \boldsymbol{\theta}). \tag{21}$$

The M-step then involves maximizing $H(\boldsymbol{\theta_1}, \boldsymbol{\theta_0})$ with respect to $\boldsymbol{\theta_1}$, where $\boldsymbol{\theta}$ is the previous value of the parameter. After an initialization of the parameter $\boldsymbol{\theta}$, the overall algorithm consists in alternating between an E-step and an M-step. Kato (2013) uses an EM algorithm approach in the framework of the SEM model. Kato's (2013) implementation of the EM algorithm involves an approximation in the E-step replacing H by

$$H'(\boldsymbol{\theta_1}, \boldsymbol{\theta}) = L(\mathbb{E}(\mathbf{Y}|\mathbf{Y_S}, \boldsymbol{\theta}))|\boldsymbol{\theta_1}). \tag{22}$$

This procedure would be exact if $\mathbb{E}(\mathbf{Y}|\mathbf{Y_S}, \boldsymbol{\theta})$ were a sufficient statistic which is not the case. For the SAR model, we propose an exact EM-algorithm since it is possible to evaluate the expected likelihood in the gaussian case.

Indeed let $\mathbf{E} = \sigma^2\mathbf{Q}$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}, \rho, \sigma^2)$. The conditional distribution of $\mathbf{Y_O}$ given $\mathbf{Y_S}$ is gaussian with mean $\boldsymbol{\mu^*}(\boldsymbol{\theta}) = \boldsymbol{\mu_O} + \boldsymbol{\Sigma_{OS}}\boldsymbol{\Sigma_{SS}}^{-1}(\mathbf{Y_S} - \boldsymbol{\mu_S}) = \boldsymbol{\mu_O} - \mathbf{E_{OO}}^{-1}\mathbf{E_{OS}}(\mathbf{Y_S} - \boldsymbol{\mu_S})$ and with variance covariance matrix

$$\boldsymbol{\Sigma^*}(\boldsymbol{\theta}) = \boldsymbol{\Sigma_{OO}} - \boldsymbol{\Sigma_{OS}}\boldsymbol{\Sigma_{SS}}^{-1}\boldsymbol{\Sigma_{S0}} = \boldsymbol{\Sigma_{OO}} + \boldsymbol{\Sigma_{OS}}\boldsymbol{Q_{SO}}\boldsymbol{Q_{OO}}^{-1} = \boldsymbol{\Sigma_{OO}} + \boldsymbol{\Sigma_{OS}}\mathbf{E_{SO}}\mathbf{E_{OO}}^{-1}.$$

We then get the expected likelihood (up to a constant term)

$$\begin{aligned} H(\boldsymbol{\theta_1}, \boldsymbol{\theta}) &= -\frac{n}{2}log(\sigma_1^2) + \log|\mathbf{I} - \rho_1\mathbf{W}| - \frac{1}{2\sigma_1^2}tr(\mathbf{E}(\rho_1)_{\mathbf{OO}}\boldsymbol{\Sigma^*}(\boldsymbol{\theta})) & (23) \\ &- \frac{1}{2\sigma_1^2}(\mathbf{Y}^* - \mathbf{Z}(\rho_1)\boldsymbol{\beta_1})'\mathbf{A}(\rho_1)(\mathbf{Y}^* - \mathbf{Z}(\rho_1)\boldsymbol{\beta_1}) & (24) \end{aligned}$$

where $\mathbf{Y}^{*'} = (\mathbf{Y}_\mathbf{S}', \mu^{*'})$, $\mathbf{Z}(\rho_1) = (\mathbf{I} - \rho_1\mathbf{W})^{-1}\mathbf{X}$, $\mathbf{A}(\rho_1) = (\mathbf{I} - \rho_1\mathbf{W}')(\mathbf{I} - \rho_1\mathbf{W})$. Optimizing with respect to $\boldsymbol{\beta_1}$ and σ_1 for given ρ_1, we get

$$\hat{\boldsymbol{\beta_1}} = (\mathbf{Z}(\rho_1)'\mathbf{A}(\rho_1)\mathbf{Z}(\rho_1))^{-1}\mathbf{Z}(\rho_1)'\mathbf{A}(\rho_1)\mathbf{Y}^*$$

and

$$\hat{\sigma_1^2} = \frac{1}{n}(\mathbf{tr}(\mathbf{A}(\rho_1)_{\mathbf{OO}}\mathbf{\Sigma^*}(\boldsymbol{\theta})) + (\mathbf{Y}^* - \mathbf{Z}(\rho_1)\hat{\boldsymbol{\beta}}_\mathbf{1})'\mathbf{A}(\rho_1)(Y^* - \mathbf{Z}(\rho_1)\hat{\boldsymbol{\beta}}_\mathbf{1}))$$

Finally the profile expected likelihood as a function of ρ_1 which has to be maximized in the M-step is

$$H(\rho_1, \widehat{\sigma_1}, \widehat{\boldsymbol{\beta}_\mathbf{1}}) = -\frac{n}{2} log(\widehat{\sigma_1^2}) + \log|\mathbf{I} - \rho_1\mathbf{W}|,$$

and the EM predictor is

$$\hat{\mathbf{Y}}_\mathbf{O}^\mathbf{EM} = \boldsymbol{\mu^*}(\hat{\boldsymbol{\theta}}_\mathbf{1}) = \hat{\boldsymbol{\mu}}_\boldsymbol{O} - \hat{\mathbf{E}}_{\mathbf{OO}}^{-1}\hat{\mathbf{E}}_{\mathbf{OS}}(\mathbf{Y_S} - \hat{\boldsymbol{\mu}}_\boldsymbol{S}) = \hat{\boldsymbol{\mu}}_\boldsymbol{O} - \hat{\mathbf{Q}}_{\mathbf{OO}}^{-1}\hat{\mathbf{Q}}_{\mathbf{OS}}(\mathbf{Y_S} - \hat{\boldsymbol{\mu}}_\boldsymbol{S})$$

where $\hat{\boldsymbol{\mu}} = \mathbf{Z}(\hat{\rho}_1)\hat{\boldsymbol{\beta}}_\mathbf{1}$.

Note that this formula only differs from the BP formula (9) by the fact that the estimators of the parameters are the ones issued from the EM algorithm (hence involve out-of-sample values) whereas in the BP predictor, they are obtained by gaussian maximum likelihood from the sample values only. Hence, in the case of the BP predictor, the EM version uses information set $\{\mathbf{Y_S}, \mathbf{X_O}, \mathbf{X_S}, \mathbf{W}\}$ whereas the ML version uses $\{\mathbf{Y_S}, \mathbf{X_S}, \mathbf{W_{SS}}\}$. The impact of this difference depends upon the parameter estimation difference which we evaluate by simulation later.

4 Comparing the predictors by simulation

4.1 Simulation framework

In order to compare the different predictors, we design a simulation study. As in Lesne et al. (2008), we use the Midi-Pyrénées region divided into $n = 283$ cantons for our study region. We consider two well-known spatial weights matrix specifications: $\mathbf{W_1}$ based on the 10 nearest neighbors scheme (distance is based on the distance between centroids of the cantons) and $\mathbf{W_2}$ based on inverse distance with decay of influence based on a cut-off beyond $125Km$, such that $\mathbf{W_1}$ is relatively sparse (96.5% of null values) whereas $\mathbf{W_2}$ is denser (40% of null values). We use row-stochastic scalings for both $\mathbf{W_1}$ and $\mathbf{W_2}$.

We simulate three explanatory variables as follows. $\mathbf{X_1}$ follows a gaussian distribution $\mathcal{N}(15, 3)$, $\mathbf{X_2}$ follows (up to a constant) a binomial distribution $\mathcal{B}(100, 0.45)/100$ and $\mathbf{X_3}$ follows a log-uniform distribution $\log(\mathcal{U}_{[0,283]})$. In order not to restrict attention to gaussian distributions, the choice of the second distribution is motivated by its bounded support and the choice of the third by its right skewness. We use the following spatial autoregressive SAR data generating processes (DGP) to generate the dependent variable:

$$\mathbf{Y} = (\mathbf{I} - \rho\mathbf{W})^{-1}(\beta_0\mathbf{1_n} + \mathbf{X_1}\beta_1 + \mathbf{X_2}\beta_2 + \mathbf{X_3}\beta_3 + \boldsymbol{\epsilon}) \text{ where } \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I_n}), \quad (25)$$

where $\mathbf{1_n} = (1, \cdots, 1)' \in \mathbb{R}^n$. The parameter $\boldsymbol{\beta}$ is fixed to $\boldsymbol{\beta} = (5, 1/4, 6, 1)$. We consider two values of σ: $\sigma = 1$ which gives an R^2 of approximately 90% and $\sigma = 3$ which gives 55%. We consider two values of ρ, one positive $\rho = 0.35$ and one negative $\rho = -0.2$. Negative values are interesting to include because they correspond to situations with competition effects (Elhorst and Zigova, 2014).

4.2 Out-of-sample prediction simulation results

To evaluate the performance of the different predictors, the number of replications is 1,000 and we report the average mean square error of prediction over the out-of-sample units. Table 2 summarizes the formulas for the out-of-sample predictors. In Table 2, $\hat{Y}_O^{TC^1}$ is the extraction corresponding to unit o, of

$$\{\mathbf{I_{n_S+1}} - \hat{\rho}\begin{pmatrix}\mathbf{W_{SS}} & \mathbf{W_{So}}\\ \mathbf{W_{oS}} & W_{oo}\end{pmatrix}\}^{-1}\begin{pmatrix}\mathbf{X_S}\\ \mathbf{X_o}\end{pmatrix}\hat{\boldsymbol{\beta}}$$

and $\hat{\boldsymbol{\beta}}$ is the parameter estimated with sample units in $\mathbf{S}$.

Table 2: Out-of-sample predictors formulas

Predictor	Formula
BP	$\hat{\mathbf{Y}}_\mathbf{O}^\mathbf{BP} = \hat{\mathbf{Y}}_\mathbf{O}^\mathbf{TC} - \hat{\mathbf{Q}}_\mathbf{OO}^{-1}\hat{\mathbf{Q}}_\mathbf{OS} \times (\mathbf{Y_S} - \hat{\mathbf{Y}}_\mathbf{S}^\mathbf{TC})$
TC	$\hat{\mathbf{Y}}_\mathbf{O}^\mathbf{TC} = [(\mathbf{I} - \hat{\rho}\mathbf{W})^{-1}\mathbf{X}\hat{\boldsymbol{\beta}}]_O$
TS^1	$\hat{Y}_o^{TS^1} = \mathbf{X_o}\hat{\boldsymbol{\beta}} + \hat{\rho}\mathbf{W_{oS}}\mathbf{Y_S}$
BP_W	$\hat{\mathbf{Y}}_\mathbf{O}^\mathbf{BPw} = \hat{\mathbf{Y}}_\mathbf{O}^\mathbf{TC} + \hat{\boldsymbol{\Sigma}}_{\boldsymbol{OS}}\mathbf{W}'_\mathbf{OS}(\mathbf{W_{OS}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{SS}}\mathbf{W}'_\mathbf{OS})^{-1}(\mathbf{W_{OS}}\mathbf{Y_S} - \mathbf{W_{OS}}\hat{\mathbf{Y}}_\mathbf{S}^\mathbf{TC})$
BP_N	$\hat{\mathbf{Y}}_\mathbf{O}^\mathbf{BP_N} = \hat{\mathbf{Y}}_\mathbf{O}^\mathbf{TC} - \hat{\mathbf{Q}}_\mathbf{OO}^{-1}\hat{\mathbf{Q}}_\mathbf{OJ}(\mathbf{Y_J} - \hat{\mathbf{Y}}_\mathbf{J}^\mathbf{TC})_\mathbf{J}$ for $\mathbf{J} = \mathbf{J(O)}$
TC^1	$\hat{Y}_o^{TC^1} =$ row o of $\{\mathbf{I_{n_S+1}} - \hat{\rho}\begin{pmatrix}\mathbf{W_{SS}} & \mathbf{W_{So}}\\ \mathbf{W_{oS}} & \mathbf{W_{oo}}\end{pmatrix}\}^{-1}\begin{pmatrix}\mathbf{X_S}\\ \mathbf{X_o}\end{pmatrix}\hat{\boldsymbol{\beta}}$
BP^1	$\hat{Y}_o^{BP^1} = \hat{Y}_o^{TC^1} - \hat{Q}_{oo}^{-1}\hat{\mathbf{Q}}_\mathbf{oS}(\mathbf{Y_S} - \hat{\mathbf{Y}}_\mathbf{S}^{\mathbf{TC^1}})$
BP_W^1	$\hat{Y}_o^{BP_W^1} = \hat{Y}_o^{TC^1} + \hat{\boldsymbol{\Sigma}}_\mathbf{oS}\mathbf{W}'_\mathbf{oS}(\mathbf{W_{oS}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{SS}}\mathbf{W}'_\mathbf{oS})^{-1}(\mathbf{W_{oS}}\mathbf{Y_S} - \mathbf{W_{oS}}\hat{\mathbf{Y}}_\mathbf{S}^{\mathbf{TC^1}})$
BP_N^1	$\hat{Y}_o^{BP_N^1} = \hat{Y}_o^{TC} - \hat{Q}_{oo}^{-1}\hat{\mathbf{Q}}_\mathbf{oJ}(\mathbf{Y_J} - \hat{\mathbf{Y}}_\mathbf{J}^{\mathbf{TC^1}})$ for $\mathbf{J} = \mathbf{J(o)}$

We choose at random a given number of sites (27 or 54) which will be declared out-of-sample (in $\mathbf{O}$). We predict the $\mathbf{Y}$ variable on the out-of-sample locations based on the sample $\mathbf{S}$ constituted by the remaining sites. We consider several situations depending upon the number of out-of-sample units and upon the aggregation level of the out-of-sample units. The corresponding configurations of out-of-sample units are shown in Figure 2 and the level of aggregation is increasing from left to right.

Table 3 summarizes the parameter estimates results (by gaussian maximum likelihood (ML) and by EM-algorithm (EM)) for configurations 1 and 3 and for 54 out-of-sample units, for $\rho = 0.35$ and $\sigma = 1$. We do not report the case $\sigma = 3$ (see supplemental material) because it does not lead to a different interpretation. In the ML case, the estimation of the coefficients is only based on the sample units and a renormalization of the in-sample weight matrix is performed. When $\mathbf{W}$ is sparse ($\mathbf{W_1}$), in general ML and EM parameter estimators are very similar but in some cases, they differ: the intercept for configuration 3 is better for EM whereas the variance for configuration 1 is better for ML. For some simulations, the EM estimates yield outliers. When $\mathbf{W}$ is dense ($\mathbf{W_2}$), ML and EM parameter estimators are also very similar. This explains why the predictors behave quite similarly whatever the method used for estimating the parameters (ML or EM).

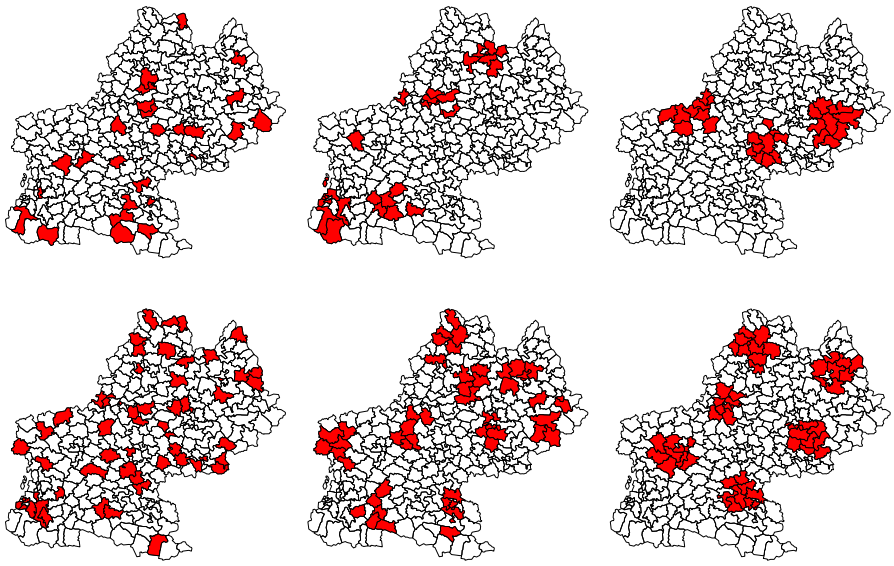

Figure 2: The three configurations for 27 (1st row) and 54 (2nd row) out-of-sample units positions: configuration 1 (left), configuration 2 (center), configuration 3 (right).

Table 3: Parameter estimation results. Standard deviations in parentheses.

	Sparse matrix $\mathbf{W_1}$				Dense matrix $\mathbf{W_2}$			
	Conf. 1 ($n_O = 54$)		Conf. 3 ($n_O = 54$)		Conf. 1 ($n_O = 54$)		Conf. 3 ($n_O = 54$)	
	ML	EM	ML	EM	ML	EM	ML	EM
$\hat{\beta}_0$	6.663	5.732	6.508	5.532	11.185	10.191	11.113	10.054
	(2.167)	(2.226)	(2.319)	(2.419)	(8.253)	(6.752)	(7.679)	(6.886)
$\hat{\beta}_1$	0.25	0.251	0.25	0.251	0.251	0.251	0.25	0.251
	(0.024)	(0.023)	(0.024)	(0.024)	(0.023)	(0.023)	(0.024)	(0.024)
$\hat{\beta}_2$	5.874	5.958	5.972	5.961	5.99	5.986	5.973	5.975
	(1.481)	(1.479)	(1.474)	(1.472)	(1.476)	(1.475)	(1.479)	(1.481)
$\hat{\beta}_3$	1.004	1	1.008	0.999	1	0.999	0.999	0.999
	(0.062)	(0.062)	(0.07)	(0.069)	(0.062)	(0.062)	(0.07)	(0.069)
$\hat{\rho}$	0.284	0.32	0.287	0.328	0.097	0.138	0.1	0.144
	(0.084)	(0.085)	(0.089)	(0.094)	(0.335)	(0.273)	(0.309)	(0.278)
$\hat{\sigma}^2$	0.991	1.202	0.991	1.104	0.982	1.791	0.98	1.575
	(0.097)	(3.238)	(0.095)	(0.33)	(0.096)	(6.449)	(0.093)	(3.923)

We report in Tables 4 and 5 the average over 1,000 replicates of the total mean square error of prediction $MSE_k = \frac{1}{n_O}(\mathbf{Y_O} - \hat{\mathbf{Y}}_\mathbf{O}^\mathbf{k})'(\mathbf{Y_O} - \hat{\mathbf{Y}}_\mathbf{O}^\mathbf{k})$ for each method $k = BP$, TC, TS^1, BP_W, BP_N, TC^1, BP^1, BP_W^1, BP_N^1 with $\rho = 0.35$ and $\sigma = 1$.

As before, we do not report the case $\sigma = 3$ because it does not lead to different interpretations. Similarly, we only give comments about the case $\rho = -0.2$. Finally, the predictors obtained using parameters estimated by ML or by EM being very similar, we only report the results using ML estimation. Complete results about these parameter choices can be found in the supplemental material. We compare the quality of the predictors when $\mathbf{W}$ is sparse (Table 4) and dense (Table 5).

Table 4: Simulation results for the 27 and 54 out-of-sample units case with $\mathbf{W_1}$ (sparse). Standard deviations in parentheses.

	27 out-of-sample units			54 out-of-sample units		
	config 1	config 2	config 3	config 1	config 2	config 3
BP	1.011	1.038	1.033	1.016	1.028	1.035
	(0.278)	(0.295)	(0.28)	(0.196)	(0.198)	(0.203)
BP_N	1.011	1.039	1.033	1.016	1.028	1.035
	(0.278)	(0.296)	(0.28)	(0.196)	(0.198)	(0.203)
$\frac{BP}{BP_N}$	100%	100%	100%	100%	100%	100%
BP_W	1.013	1.04	1.034	1.017	1.029	1.036
	(0.279)	(0.296)	(0.279)	(0.196)	(0.199)	(0.203)
$\frac{BP}{BP_W}$	99.8%	99.8%	99.9%	99.9%	99.9%	99.9%
BP^1	1.012	1.05	1.057	1.018	1.037	1.053
	(0.278)	(0.3)	(0.284)	(0.196)	(0.201)	(0.208)
$\frac{BP}{BP^1}$	99.9%	98.9%	97.9%	99.8%	99.2%	98.4%
${BP_N}^1$	1.013	1.054	1.058	1.02	1.038	1.056
	(0.278)	(0.301)	(0.284)	(0.196)	(0.201)	(0.209)
$\frac{BP}{{BP_N}^1}$	99.8%	98.5%	97.8%	99.6%	99.1%	98.1%
${BP_W}^1$	1.015	1.054	1.058	1.021	1.039	1.056
	(0.278)	(0.3)	(0.283)	(0.196)	(0.202)	(0.208)
$\frac{BP}{{BP_W}^1}$	99.6%	98.5%	97.8%	99.5%	99%	98.1%
TS^1	1.023	1.057	1.059	1.028	1.043	1.055
	(0.279)	(0.299)	(0.283)	(0.198)	(0.203)	(0.208)
$\frac{BP}{TS^1}$	98.8%	98.2%	97.6%	98.8%	98.6%	98.1%
TC	1.063	1.081	1.069	1.061	1.067	1.07
	(0.289)	(0.304)	(0.288)	(0.206)	(0.208)	(0.211)
$\frac{BP}{TC}$	95.1%	96%	96.6%	95.8%	96.3%	96.7%
TC^1	1.065	1.095	1.098	1.068	1.08	1.087
	(0.288)	(0.308)	(0.293)	(0.207)	(0.212)	(0.216)
$\frac{BP}{TC^1}$	94.9%	94.7%	94.1%	95.1%	95.2%	95.2%

When $\mathbf{W}$ is sparse (Table 4), whatever configurations and number of sites to predict, we obtain the following ranking between methods in decreasing order of efficiency

$$BP < BP_N < BP_W < BP^1 < BP_N^1 < BP_W^1 < TS^1 < TC < TC^1$$

Note that the worst ratio is around 94%. As far as the impact of the level of aggregation is concerned, predictors including a correction for spatial correlation such as BP_N, BP_W, BP^1, BP_{N^1} and BP_{W^1} tend to perform better when the level of aggregation is low which is understandable since for high aggregation, the neighborhood of an out-of-sample unit will contain few in-sample units. This effect is not the same for the predictor TC which does not correct for spatial correlation since we observe that the ratio BP/TC is now increasing with the level of aggregation.

When $\mathbf{W}$ is dense (Table 5), all the predictors seem to be very similar. This can be explained by the fact that in that case, there is a lot more information about a given out-of-sample point in the neighboring sample points and hence a good spatial correction is less necessary. For $\rho = -0.2$, the comparison between predictors efficiency yields the same ranking for the first three BP, BP_N, BP_W and some ranking inversions appear for the last five but with very small efficiency differences. The value of the worst ratio in that case, 98.4%, is higher than for the positive ρ value, meaning less difference between the performances, which is understandable because this particular choice of negative ρ corresponds to a smaller amount of spatial autocorrelation. We also did some mispecification tests for the weight matrix using a dense matrix to predict whereas the DGP involved a sparse one and reversely. Detailed results included in the supplemental material show that it does not hurt to use for predicting a matrix which is sparser than the true one whereas the reverse may be detrimental.

Because the reported prediction errors are averages over out-of-sample units, we suspected it may hide different situations depending on the number of missing neighbors of a given out-of-sample unit. Table 6 reports the prediction errors as a function of the number of missing neighbors for the following simulation framework. This number k ranges from 0 to 9 and for each k, we repeat 1,000 times the following process

- choose a site i_0 at random
- remove k neighbors at random from the neighbors of i_0, and let $\mathbf{O}$ contain i_0 and the indices of these missing neighbors. The in-sample set of sites $\mathbf{S}$ is constituted by the remaining sites
- simulate the vector $\mathbf{Y}$ for all the sites
- predict the $\mathbf{Y}$ on the sites in $\mathbf{O}$ and compute the prediction error of Y_{i_0} denoted by PE_{i_0}

The first column of the table contains the mean over the 1,000 draws of PE_{i_0} for the BP predictor and the remaining ones report the ratio of this quantity the mean PE_{i_0} of all the other methods.

When $\mathbf{W}$ is sparse, we observe that the BP predictive mean square error indeed slightly increases with the number of missing neighbors. The efficiency of BP^1, BP_N^1, BP_W^1, TS^1 and TC^1 with respect to BP decreases with the number of missing neighbors. The efficiency of TC with respect to BP increases with the number of missing neighbors which we interpret as revealing the fact that when the information gets poor in the neighborhood, it is just as well to use the mean to predict (the

Table 5: Simulation results for the 27 and 54 out-of-sample units case with $\mathbf{W_2}$ (dense). Standard deviations in parentheses.

	27 out-of-sample units			54 out-of-sample units		
	config 1	config 2	config 3	config 1	config 2	config 3
BP	1.02	1.042	1.021	1.023	1.022	1.022
	(0.28)	(0.293)	(0.271)	(0.197)	(0.196)	(0.198)
BP_N	1.02	1.042	1.021	1.023	1.022	1.022
	(0.28)	(0.293)	(0.271)	(0.197)	(0.196)	(0.198)
$\frac{BP}{BP_N}$	100%	100%	100%	100%	100%	100%
BP_W	1.02	1.042	1.021	1.023	1.022	1.022
	(0.28)	(0.293)	(0.271)	(0.197)	(0.196)	(0.198)
$\frac{BP}{BP_W}$	100%	100%	100%	100%	100%	100%
BP^1	1.02	1.041	1.021	1.023	1.022	1.023
	(0.28)	(0.293)	(0.271)	(0.197)	(0.196)	(0.199)
$\frac{BP}{BP^1}$	100%	100%	100%	100%	99.9%	99.9%
${BP_N}^1$	1.02	1.041	1.021	1.023	1.022	1.023
	(0.28)	(0.293)	(0.271)	(0.197)	(0.196)	(0.199)
$\frac{BP}{{BP_N}^1}$	100%	100%	100%	100%	99.9%	99.9%
${BP_W}^1$	1.02	1.041	1.021	1.023	1.022	1.023
	(0.28)	(0.293)	(0.271)	(0.197)	(0.196)	(0.198)
$\frac{BP}{{BP_W}^1}$	100%	100%	100%	100%	100%	99.9%
TS^1	1.019	1.04	1.022	1.022	1.022	1.022
	(0.279)	(0.292)	(0.271)	(0.197)	(0.196)	(0.198)
$\frac{BP}{TS^1}$	100%	100%	100%	100%	100%	100%
TC	1.022	1.042	1.022	1.023	1.023	1.023
	(0.279)	(0.292)	(0.27)	(0.197)	(0.196)	(0.197)
$\frac{BP}{TC}$	99.8%	100%	99.9%	100%	99.9%	99.9%
TC^1	1.021	1.041	1.023	1.023	1.023	1.023
	(0.278)	(0.291)	(0.27)	(0.197)	(0.196)	(0.197)
$\frac{BP}{TC^1}$	99.9%	100%	99.8%	99.9%	99.9%	99.9%

correction is inefficient). The efficiency of BP_W with respect to BP remains stable. Finally the almost-optimal predictor BP^N is clearly very close to the best predictor. When $\mathbf{W}$ is dense, all the predictors seem to be very similar whatever the number of missing values.

5 Conclusion

At least in the case of this particular model, when $\mathbf{W}$ is sparse, the performance of BP_N, BP_W, BP^1, BP_N^1,BP_W^1 are very close to that of the best prediction and much better than that of TC, TS, TC^1, TS^1. Attempts to consider a larger variety of

Table 6: Prediction errors as a function of number of missing neighbors nb.

	nb	BP	SD	$\frac{BP}{BP_N}$	$\frac{BP}{BP_W}$	$\frac{BP}{BP^1}$	$\frac{BP}{{BP_N}^1}$	$\frac{BP}{{BP_W}^1}$	$\frac{BP}{TS^1}$	$\frac{BP}{TC}$	$\frac{BP}{TC^1}$
	0	0.992	(1.443)	100%	99.7%	100%	100%	99.7%	98.7%	94.4%	94.4%
	1	0.991	(1.445)	100%	99.4%	99.8%	99.7%	99.3%	98.3%	94.4%	94%
	2	0.991	(1.445)	100%	99.7%	99.3%	99.3%	99.1%	98%	94.4%	93.4%
	3	1.01	(1.47)	99.9%	99.7%	97.1%	97.1%	97.1%	96.5%	95.8%	92.2%
	4	1.013	(1.487)	99.9%	99.9%	93.8%	93.8%	93.6%	93.2%	96%	89.3%
$\mathbf{W_1}$	5	1.023	(1.495)	99.9%	99.9%	92.9%	92.8%	92.8%	93.3%	97.1%	90.6%
	6	1.029	(1.5)	99.9%	99.8%	92.4%	92.2%	92.2%	93.5%	97.6%	92.5%
	7	1.026	(1.495)	100%	100%	91%	91%	90.9%	92.2%	97.3%	91.4%
	8	1.037	(1.502)	100%	100%	82.6%	82.4%	82.4%	84.2%	98.3%	86.3%
	9	1.037	(1.503)	100%	100%	82.4%	82.4%	82.4%	84.7%	98.3%	90%
	0	0.999	(1.45)	100%	100%	100%	100%	100%	100%	99.7%	99.7%
	1	0.999	(1.449)	100%	100%	100%	100%	100%	100%	99.8%	99.7%
	2	0.999	(1.449)	100%	100%	100%	100%	100%	100%	99.8%	99.7%
	3	1.002	(1.453)	100%	100%	99.9%	99.9%	99.9%	100%	99.9%	99.8%
	4	1.002	(1.457)	100%	100%	99.9%	99.9%	99.9%	99.9%	99.8%	99.7%
$\mathbf{W_2}$	5	1.002	(1.458)	100%	100%	99.9%	99.9%	100%	100%	99.9%	99.8%
	6	1.002	(1.457)	100%	100%	100%	100%	100%	100%	99.9%	99.9%
	7	1.001	(1.455)	100%	100%	100%	100%	100%	100%	99.7%	99.7%
	8	1.002	(1.455)	100%	100%	100%	100%	100%	100%	99.8%	99.8%
	9	1.001	(1.454)	100%	100%	100%	100%	100%	100%	99.7%	99.7%

parameter values (different values of σ and ρ) and alternative method for estimating the parameters (EM) have shown that the results are quite stable. We also tested a variant of this model replacing the explanatory variables by their spatially lagged version and the results were very similar with the same conclusions in terms of ranking of the methods.

For the out-of-sample case, BP_N is better than BP_W in terms of efficiency but BP_W is closer to BP in terms of projection coefficients. BP_W is better than TC, less good than TS. When $\mathbf{W}$ is dense, all predictors behave quite similarly.

We developed our study on the case of the SAR model. For the case of the spatial error model SEM which is a linear model with SAR residuals, we refer the reader to Kato (2008). Our conclusions apply for the Spatial Durbin model :

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \alpha\boldsymbol{\iota} + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

with $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ because it can be written as a general linear model with $\boldsymbol{\mu} = (\mathbf{I} - \rho\mathbf{W})^{-1}(\alpha\boldsymbol{\iota} + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\gamma})$ and variance given by (4). The difference between the SAR and the Durbin stands only in the mean $\boldsymbol{\mu}$ and it is the same expression but with additional explanatory variables. Hence the same arguments apply but the formulas have to be adapted. The Kato (2008) approach for the SEM however cannot be extended directly to the SAR because the expression of the mean

is quite different.

References

[1] Bivand R (2002) Spatial econometrics functions in R: classes and methods. *Journal of Geographical Systems* 4: 405–421.

[2] Bivand R, Piras G (2015). Comparing implementations of estimation methods for spatial econometrics. *Journal of Statistical Software*, 63(18): 1–36.

[3] Cressie N (1990) The origins of kriging. *Mathematical Geology* 22: 239–252.

[4] Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39: 1–38.

[5] Dubin R (2003) Robustness of spatial autocorrelation specifications: some Monte Carlo evidence. *Journal of Regional Science* 43(2): 221–248.

[6] Elhorst JP, Zigova K (2014) Competition in research activity among economic departments: evidence by negative spatial autocorrelation. *Geographical Analysis* 46(2): 104–125.

[7] Gaetan C, Guyon X (2010) *Spatial Statistics and Modeling*. Berlin, Springer-Verlag.

[8] Goldberger AS (1962) Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association* 57: 369–375.

[9] Griffith DA (2003) *Spatial Autocorrelation and Spatial Filtering: Gaining Understanding through Theory and Scientific Visualization*. Berlin, Springer-Verlag.

[10] Griffith DA (2010) Spatial filtering and missing georeferenced data imputation: a comparison of the Getis and Griffith methods. In Anselin L, Rey SJ (eds) *Perpectives on Spatial Data Analysis*. Berlin, Springer-Verlag.

[11] Griffith DA, Bennet R, Haining R (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.

[12] Haining R (1990) *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge, Cambridge University Press.

[13] Haining R, Griffith DA, Bennett R (1984) A statistical approach to the problem of missing spatial data using a first-order Markov model. *Professional Geographer* 36: 338–45.

[14] Harville DA (1997) *Matrix Algebra from a Statistician's Perspective*. New-York, Springer-Verlag.

[15] Kato T (2008) A further exploration into the robustness of spatial autocorrelation specifications. *Journal of Regional Science* 48: 615–638.

[16] Kato T (2013) Usefulness of the information contained in the prediction sample for the spatial error model. *Journal of Real Estate Finance and Economics* 47: 169–195.

[17] Kelejian HH, Prucha IR (2007) The relative efficiencies of various predictors in spatial econometric models containing spatial lags. *Regional Science and Urban Economics* 37: 363–374.

[18] LeSage JP, Pace RK (2004) Models for spatially dependent missing data. *Journal of Real Estate Finance and Economics* 29: 233–254.

[19] LeSage JP, Pace RK (2008) Spatial econometric models, prediction. In Shekhar S, Xiong H (eds) *Encyclopedia of Geographical Information Science*. Springer-Verlag New-York.

[20] LeSage JP, Pace RK (2009) *Introduction to Spatial Econometrics*. Boca Raton, Taylor & Francis.

[21] Lesne JP, Tranger H, Ruiz-Gazen A, Thomas-Agnan C (2008) Predicting population annual growth rates with spatial models. *Preprint*.

[22] Rue H, Held L (2005) *Gaussian Markov Random Fields, Theory and Applications*. Boca Raton, Taylor & Francis.