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

Michel Goulard *
INRA, UMR 1201 DYNAFOR, Chemin de Borde Rouge BP52627, F31326 Castanet-Tolosan, FRANCE
Thibault Laurent†
Toulouse School of Economics (CNRS/GREMAQ), 21 allée de Brienne 31042 Toulouse, FRANCE
Christine Thomas-Agnan‡
Toulouse School of Economics (GREMAQ), 21 allée de Brienne 31042 Toulouse, FRANCE

September 25, 2014

**Abstract**

We address the problem of prediction in the classical spatial autoregressive LAG model for areal data. In contrast with the spatial econometrics literature, the geostatistical literature has devoted much attention to prediction using the Best Linear Unbiased Prediction approach. From the methodological point of view, we explore the limits of the extension of BLUP formulas in the context of the spatial autoregressive LAG models for in sample prediction as well as out-of-sample prediction simultaneously at several sites. We propose a more tractable "almost best" alternative. 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

Whereas prediction is a basic concern in geostatistics (Cressie, 1990), it has not been paid as much attention in the econometrics literature. Bivand (2002) recognizes

*e-mail:goulard@toulouse.inra.fr
†e-mail: thibault.laurent@univ-tlse1.fr
‡e-mail: christine.thomas@tse-fr.eu

the importance of the question: "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". Kato (2008) explores the best linear prediction problem in the framework of spatial error models. In the context of spatial lag models, other authors (Bennet et al. (1989), LeSage and Pace (2004, 2008), Kelejian and Prucha (2007)) have addressed some aspects of this question and we will summarize their contribution in section 2.

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. To motivate the need for out-of-sample prediction, 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 10000 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 two first 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 formulae were necessary for spatial models. Figure 1 presents the positions 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 as in Lesne et al. (2008).

We first review the classical prediction formulae encountered in the literature for the spatial simultaneous autoregressive (SAR or LAG depending on authors) models. Then 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 to this formula and finally demonstrate that the simple formulas classically implemented in usual softwares can thus be improved upon substantially.

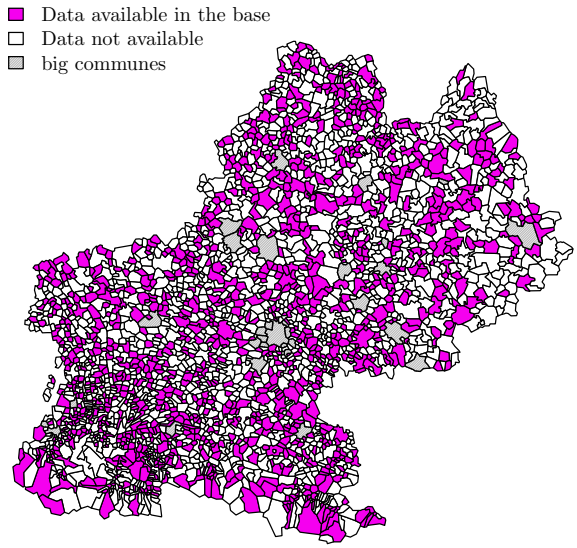


Figure 1: Spatial units where population data was available at the time of this study

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

## 2.1 Models and prediction situations

We consider prediction in the classical homoscedastic spatial autoregressive LAG model (LAG 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 $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ and $\mathbb{E}(\boldsymbol{\epsilon} \mid \mathbf{X}) = \mathbf{0}$. In reduced form, this is equivalent to

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

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

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

and its covariance structure by

$$\boldsymbol{\Sigma} = [(\mathbf{I} - \rho\mathbf{W}')(\mathbf{I} - \rho\mathbf{W})]^{-1}\sigma^2, \tag{3}$$

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

$$\mathbf{Q} = \boldsymbol{\Sigma}^{-\mathbf{1}} = \frac{1}{\sigma^2}(\mathbf{I} - \rho\mathbf{W}')(\mathbf{I} - \rho\mathbf{W}) \tag{4}$$

If $\rho$ is known, the best linear unbiased estimator (BLUE) of $\boldsymbol{\mu} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta}$ is $\hat{\boldsymbol{\mu}} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\hat{\boldsymbol{\beta}}$.

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. This situation is illustrated in the left part of Figure 2. 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 $\mathbf{X_O}$. This situation is illustrated in the right part of Figure 2. 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.

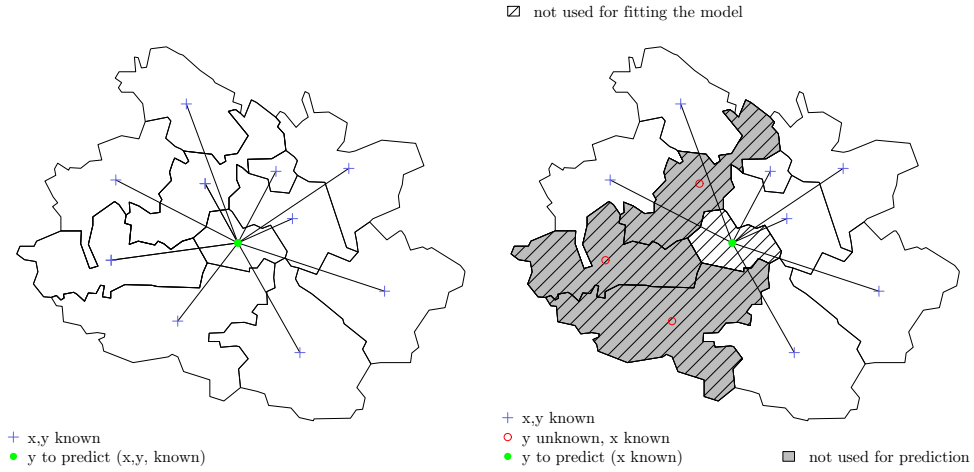


Figure 2: In-sample (left) and and out-of-sample (right) single prediction problem. Shaded areas are sample units which are not used at the model fitting stage. Crosses are in-sample units, empty circles are out-of-sample units and full circle is the point to predict.

## 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 $\mathbf{X} = (\mathbf{X_S}, \mathbf{X_O})$ and $\mathbf{Y} = (\mathbf{Y_S}, \mathbf{Y_O})$ where $\mathbf{X_S}$ (resp $\mathbf{Y_S}$) of dimension $n_S \times p$ (resp $n_S$) denote the matrix of components of $\mathbf{X}$ corresponding to in-sample spatial units and $\mathbf{X_O}$ (resp $\mathbf{Y_O}$) of dimension $n_O \times p$ (resp $n_O$) denote the matrix of components of $\mathbf{X}$ corresponding

to out-of-sample spatial units and $p$ is the number of exogenous variables. Similarly $\boldsymbol{\mu} = (\boldsymbol{\mu_S}, \boldsymbol{\mu_O})$. More generally in this paper, when $J$ denotes a set of indices, the matrix $\mathbf{X_J}$ will denote the matrix of components of $\mathbf{X}$ relative to the indices in $J$. For two sets of indices $I$ and $J$, 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 $I$ and column indices in $J$ and finally $\mathbf{A_{II}} = \mathbf{A_I}$.

Similarly, we partition the spatial weights matrix $\mathbf{W}$ as follows

$$\mathbf{W} = \begin{pmatrix} \mathbf{W_S} & \mathbf{W_{SO}} \\ \mathbf{W_{OS}} & \mathbf{W_O} \end{pmatrix}, \tag{5}$$

where

- $\mathbf{W_S}$ is the $n_S \times n_S$ submatrix corresponding to the neighborhood structure of the $n_S$ in-sample sites,
- $\mathbf{W_O}$ 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.

For out-of-sample prediction, we need to relate the model driving the in-sample units to the out-of-sample ones and we assume there is an overall model driving the in-sample and out-of-sample units. The overall model $M$ is given by (1) with a row-normalized matrix $\mathbf{W}$ for the $n$ observations of $(\mathbf{X}, \mathbf{Y})$. The sub-model $M_S$ driving the vector $\mathbf{X_S}, \mathbf{Y_S}$ corresponding to the sample units follows the same expression (1) but using the submatrix $\mathbf{W_S}$ renormalized (row-normalization after extraction). This natural assumption however leads to two constraints. The compatibility of the two models implies that $((\mathbf{I} - \rho\mathbf{W})^{-\mathbf{1}}\mathbf{X})_\mathbf{S} = (\mathbf{I} - \rho\mathbf{W_S})^{-1}\mathbf{X_S}$ for the mean and that $(\mathbf{var}(\mathbf{Y}))_\mathbf{S} = var(\mathbf{Y_S})$ 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 3.4) in his EM approach. Finally the EM approach proposed in section 3.4 does not require these restrictions and leads to very similar results as the BLUP based on this two models specification.

It is important to note that while a corresponding decomposition of the precision matrix is easily derived from (4), the covariance matrix for sub-model $M_S$ on the other hand is not an extraction of $\mathbf{\Sigma}$ because of the inversion in formula (3).

## 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 $\mathbf{Y} = \boldsymbol{\mu} + \mathbf{V}$ with **known** $V$. The Golberger formula (1962) gives the

BLUP

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

where $\mathbf{Y_O^*} = \lambda' \mathbf{Y_S}$ minimizes $\mathbb{E}(\mathbf{Y_O^*} - \mathbf{Y_O})^2$ under the constraint that $\mathbb{E}(\mathbf{Y_O^*} - \mathbf{Y_O}) = \mathbf{0}$ and where $\hat{\boldsymbol{\mu}}_{\boldsymbol{O}}$ and $\hat{\boldsymbol{\mu}}_{\boldsymbol{S}}$ are estimators of respectively $\mathbb{E}(\mathbf{Y_O})$ and $\mathbb{E}(\mathbf{Y_S})$. Even if the notation does not show, it is understood hereafter that the conditional expectations are also conditional upon the explanatory variables. In practice, one does not know the theoretical variance $\mathbf{V}$ (i.e. $\rho$ in the LAG 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 $O$ reduced to a point but the formula remains true for a set of points $O$ . In that case the problem is to find $\mathbf{Y_O^*} = \Lambda' \mathbf{Y_S}$ minimizing $\mathrm{Tr}(\mathbb{E}(\mathbf{Y_O^*} - \mathbf{Y_O})(\mathbf{Y_O^*} - \mathbf{Y_O})')$ under the constraint that $\mathbb{E}(\mathbf{Y_O^*} - \mathbf{Y_O}) = \mathbf{0}$ where $\Lambda$ 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 LAG model in reduced form.

### 2.3.2 In-sample prediction

In an ordinary linear model which is model (1) for $\rho = 0$, the best linear unbiased predictor (BLUP) of $\mathbf{Y_S}$ coincides with the best linear unbiased estimator (BLUE) of $\boldsymbol{\mu}$ and is given by

$$\hat{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{T}} = \mathbf{X_S}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\mu}}_{\boldsymbol{S}}, \tag{6}$$

where $\hat{\boldsymbol{\beta}}$ is the estimator of $\boldsymbol{\beta}$ calculated by fitting the model with in-sample units.

Based on the equality between BLUE and BLUP for the OLS model, 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

$$\hat{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{TC}} = [(\mathbf{I} - \hat{\rho}\mathbf{W_S})^{-1}]\mathbf{X_S}\hat{\boldsymbol{\beta}}, \tag{7}$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\rho}$ are the estimators of $\boldsymbol{\beta}$ and $\rho$ calculated by fitting the model with in-sample units and $[(\mathbf{I} - \hat{\rho}\mathbf{W})^{-1}]_{\mathbf{S}}$ is the $S$ bloc extraction of the inverse of matrix $(\mathbf{I} - \hat{\rho}\mathbf{W})$. 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

$$\hat{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{TS}} = \mathbf{X_S}\hat{\boldsymbol{\beta}} + \hat{\rho}\mathbf{W}\mathbf{Y_S} \tag{8}$$

Thereafter, we call this predictor the "trend-signal-noise" predictor. This one is used in the Bivand R package **spdep**. Note that if $\hat{\rho} = 0$, then the maximum likelihood

estimator of $\boldsymbol{\beta}$ coincides with the ordinary least squares estimator and thus these three predictors are all equal. If we had $\hat{\rho} = \rho$ and $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, we would get

$$\mathbb{E}(\hat{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{TC}}) = \mathbb{E}(\hat{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{TS}}) = \mathbb{E}(\mathbf{Y_S})$$

Gaetan and Guyon (2008) use another version of the Goldberger formula in the framework of conditional autoregressive CAR models for in-sample prediction

$$\hat{\mathbf{Y}}_{\mathbf{S}} = \hat{\boldsymbol{\mu}}_{\boldsymbol{S}} - \text{Diag}(\mathbf{Q_S})^{-\mathbf{1}}\tilde{\mathbf{Q}}_{\mathbf{S}}(\mathbf{Y_S} - \hat{\boldsymbol{\mu}}_{\boldsymbol{S}}) \tag{9}$$

where $\text{Diag}(\mathbf{Q_S})$ denotes the diagonal matrix containing the diagonal of the precision matrix $\mathbf{Q_S}$ and $\tilde{\mathbf{Q}}_{\mathbf{S}} = \mathbf{Q_S} - \text{Diag}(\mathbf{Q_S})$. This formula remains true for LAG models provided it is applied to the reduced form so that we have $\hat{\boldsymbol{\mu}}_{\boldsymbol{S}} = [(\mathbf{I} - \hat{\rho}\mathbf{W})^{-\mathbf{1}}]_{\mathbf{S}}\mathbf{X_S}\hat{\boldsymbol{\beta}}$. Again, in practice, $\rho$ is unknown and must be substituted by $\hat{\rho}$.

In the framework of the LAG model, the same arguments yield the following version of Goldberger formula

$$\hat{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{BP}} = (\mathbf{I} - \hat{\rho}\mathbf{W_S})^{-1}\mathbf{X_S}\hat{\boldsymbol{\beta}} - \text{Diag}(\mathbf{Q_S})^{-1}\tilde{\mathbf{Q}}_{\mathbf{S}}(\mathbf{Y} - (\mathbf{I} - \hat{\rho}\mathbf{W_S})^{-1}\mathbf{X_S}\hat{\boldsymbol{\beta}}), \tag{10}$$

where $\mathbf{Q_S} = \frac{1}{\hat{\sigma}^2}(\mathbf{I} - \hat{\rho}\mathbf{W}'_{\mathbf{S}})(\mathbf{I} - \hat{\rho}\mathbf{W_S})$. Note that since this second version of Goldberger is based on the precision matrix rather than the covariance matrix, it should be preferred to the first one for the LAG model.

Using a coordinate formulation rather than a matrix form, this formula is equivalent to

$$\hat{Y}_i^{BP} = \hat{\mu}_i - \sum_{j=1, j\neq i}^{n} \frac{q_{ij}}{q_{ii}}(Y_j - \hat{\mu}_j), \tag{11}$$

where $q_{ij}$ is the $(i, j)$ element of matrix $\mathbf{Q_S}$ and $\hat{\mu}_i$ are the components of $\hat{\boldsymbol{\mu}}$ given by (6) which is the formula used in LeSage and Pace (2004).

### 2.3.3 Out-of-sample prediction

The trend-signal-noise predictor $\hat{\mathbf{Y}}^{\mathbf{TS}}$ cannot be defined in 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 on unit $o$, it is possible to compute it because of the zeros on the diagonal of $\mathbf{W}$ which yields

$$\hat{Y}_o^{TS^1} = \mathbf{X_o}\hat{\boldsymbol{\beta}} + \hat{\rho}\mathbf{W_{oS}}\mathbf{Y_S}. \tag{12}$$

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

$$\hat{\mathbf{Y}}^{\mathbf{TC}} = (\mathbf{I} - \hat{\rho}\mathbf{W})^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{TC}} \\ \hat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{TC}} \end{pmatrix} \tag{13}$$

and

$$\begin{aligned} \hat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{TC}} &= -(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1}\mathbf{X_S}\hat{\boldsymbol{\beta}} + (\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{X_O}\hat{\boldsymbol{\beta}} \\ \hat{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{TC}} &= (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{X_S}\hat{\boldsymbol{\beta}} - (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1}\mathbf{X_O}\hat{\boldsymbol{\beta}} \end{aligned} \tag{14}$$

for $(\mathbf{I} - \hat{\rho}\mathbf{W}) = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} = \begin{pmatrix} \mathbf{I_S} - \hat{\rho}\mathbf{W_S} & -\hat{\rho}\mathbf{W_{SO}} \\ -\hat{\rho}\mathbf{W_{OS}} & \mathbf{I_O} - \hat{\rho}\mathbf{W_O} \end{pmatrix}$.
Note that the notation $\hat{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{TC}}$ in (13) denotes something different from (7) because we have here the in-sample prediction which takes into account the out-of-sample units. One can check that the $\hat{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{TC}}$ from (13) coincides with the predictor from 7 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 $O = \{i\}$ and for $\mathbf{W_{i.}Y}$ replacing $\mathbf{Y_S}$. 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 $\hat{Y}_o^{TS^1}$ it is $\{\mathbf{X}, \mathbf{W}, \mathbf{Y_S}\}$.

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

## 3.1 Another formulation of Goldberger formula for LAG models

For out-of-sample best prediction, if we first concentrate on the case of single prediction, formula (11) can be applied with the precision matrix $\mathbf{Q}$ corresponding to the sample units augmented with the point to predict.

In the case of out-of-sample best prediction, Harville (1997) derives a Goldberger formula written in terms of inverse matrix $\mathbf{Q}$, similar to 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 $Cov(\mathbf{Y_O}, \mathbf{Y_S})Var(\mathbf{Y_S})^{-1} = -\mathbf{Q_O}^{-1}\mathbf{Q_{OS}}$, 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

$$\hat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{BP}} = \hat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{TC}} - \mathbf{Q_O}^{-1}\mathbf{Q_{OS}} \times (\mathbf{Y_S} - \hat{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{TC}}) \tag{15}$$

with

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

Let us note that the matrix to invert is $\mathbf{Q_O}$ and has the size of 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 formula instead of the variance one can help.

## 3.2 Extension of the Kelejian-Prucha predictor

We first propose to generalize the Kelejian-Prucha approach to multiple prediction where $\mathbf{Y_O}$ is predicted by linear combination of $\mathbf{W_{OS}Y_S}$ instead of $\mathbf{Y_S}$. In Kelejian

and Prucha (2007), there is only one out-of-sample unit. Indeed, it is easy to extend to the case of several out-of-sample units. The information set is then $\{\mathbf{X}, \mathbf{W}, \mathbf{Y_S}\}$. In that case, Golberger formula gives the best predictor $\hat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{BP_W}} = \mathbb{E}(\mathbf{Y_O} \mid \mathbf{W_{OS}Y_S})$ as

$$\hat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{BP_W}} = \hat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{TC}} + \mathbf{\Sigma_{OS}W'_{OS}}(\mathbf{W_{OS}\Sigma_S W'_{OS}})^{-1}(\mathbf{W_{OS}Y_S} - \mathbf{W_{OS}}\hat{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{TC}}). \quad (16)$$

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, formula (16) is not simpler to compute than the best prediction given by formula (15): the size of the matrix to invert is equal to the number of out-of-sample units.

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

$$\mathbf{W}_{\{\mathbf{J},\mathbf{O}\}} = \left(\begin{array}{c|c} \mathbf{W_J} & \mathbf{W_{JO}} \\ \hline \mathbf{W_{OJ}} & \mathbf{W_O} \end{array}\right).$$

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

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

and thus we get the following predictor

$$\hat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{BP_N}} = \hat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{TC}} - \mathbf{Q_O}^{-1}\mathbf{Q_{OJ}}(\mathbf{Y_J} - \hat{\mathbf{Y}}_{\mathbf{J}}^{\mathbf{TC}}), \quad (17)$$

where $\hat{\mathbf{Y}}_{\mathbf{J}}^{\mathbf{TC}}$ and $\hat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{TC}}$ are obtained by extracting the rows corresponding to units in $J$ from $\hat{\mathbf{Y}}_{\mathbf{J},\mathbf{O}}^{\mathbf{TC}}$. The advantage of this predictor lies in the fact that it reduces the computational burden since the size of the matrix $\mathbf{Q_{OJ}}(\mathbf{Y_J} - \hat{\mathbf{Y}}_{\mathbf{J}}^{\mathbf{TC}})$ is $n_O \times n_J$ instead of $n_O \times n_S$. If we were using the Goldberger formula, the new predictor would be written

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

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 substitute to the ML parameters estimators some approximations similar to the ones we describe in section 2.3. Note that because of the links between

$\mathbf{W}$ and $\mathbf{Q}$, if we consider $\mathbf{W'W}$-neighbouring, that is order 2 $\mathbf{W}$-neighbouring, the predictor will be optimal and is equal to the predictor with $\mathbf{Q}$-neighbours. Indeed the reason is that if we look at prediction with a set of $\mathbf{Q}$-neighbours then it means that $\mathbf{Q}$ can be written :

$$\mathbf{Q} = \left( \begin{array}{c|c} \mathbf{Q_{S\backslash J}} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{Q_{\{J,O\}}} \end{array} \right)$$

and thus $\mathbf{Q_O}^{-1}\mathbf{Q_{OS}} \times (\mathbf{Y_S} - \hat{\mathbf{Y}}_\mathbf{S}^{\mathbf{TC}})$ is equal to $\mathbf{Q_O}^{-1}\mathbf{Q_{JO}} \times (\mathbf{Y_J} - \hat{\mathbf{Y}}_\mathbf{J}^{\mathbf{TC}})$ and therefore is optimal.

## 3.3 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 exists out-of-sample only in the single prediction case. This leads us to defining 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 $\hat{\mathbf{Y}}^{\mathbf{BP_N}}$ which will be denoted respectively by $\hat{Y}^{TC^1}$, $\hat{Y}^{TS^1}$, $\hat{Y}^{BP^1}$, $\hat{Y}^{BP_W{}^1}$ and $\hat{Y}^{BP_N{}^1}$. The precise formulae are detailed in Table 2. These formulae 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.

## 3.4 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,

$$H(\boldsymbol{\theta_1}, \boldsymbol{\theta}) = \mathbb{E}(L(\mathbf{Y}|\boldsymbol{\theta_1})|\mathbf{Y_S}, \boldsymbol{\theta}). \tag{18}$$

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

$$H'(\boldsymbol{\theta_1}, \boldsymbol{\theta}) = L(\mathbb{E}(\mathbf{Y}|\mathbf{Y_S}, \boldsymbol{\theta}))|\boldsymbol{\theta_1}). \tag{19}$$

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 LAG model, we propose an exact EM-algorithm since it is possible to evaluate the expected likelihood.

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

$$\begin{aligned} H(\boldsymbol{\theta_1}, \boldsymbol{\theta}) &= -\frac{n}{2}log(\sigma_1^2) + \log|\mathbf{I} - \rho_1\mathbf{W}| - \frac{1}{2\sigma_1^2}tr(\mathbf{E_{OO}}(\rho_1)\boldsymbol{\Sigma^*}(\boldsymbol{\theta})) \quad (20) \\ &- \frac{1}{2\sigma_1^2}(\mathbf{Y}^* - \mathbf{Z}(\rho_1)\boldsymbol{\beta_1})'\mathbf{A}(\rho_1)(\mathbf{Y}^* - \mathbf{Z}(\rho_1)\boldsymbol{\beta_1}) \quad (21) \end{aligned}$$

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 $\sigma_1$ for given $\rho_1$, we get

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

and

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

Finally the profile expected likelihood as a function of $\rho_1$ which has to be maximized in the M-step is

$$H(\rho_1, \widehat{\sigma_1}, \widehat{\boldsymbol{\beta_1}}) = -\frac{n}{2}log(\widehat{\sigma_1^2}) + \log|\mathbf{I} - \rho_1\mathbf{W}|$$

and the EM predictor is

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

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

Note that this formula differs from the BP formula by the fact that the estimators of the parameters are the ones issued from the EM algorithm whereas in the BP predictor, they are obtained by maximum likelihood from the sample. Hence the EM predictor uses information set $\{\mathbf{Y_S}, \mathbf{X_O}, \mathbf{X_S}, \mathbf{W}\}$ whereas the BP predictor uses $\{\mathbf{Y_S}, \mathbf{X_S}, \mathbf{W_S}\}$. 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. Table 1 summarizes the formulas for the in-sample predictors and Table 2 for the out-of-sample predictors. In Table 2, $\hat{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{TC}}$ (respectively $\hat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{TC}}$) are the extractions corresponding to units in $S$, respectively unit $o$, of

$$\{\mathbf{I_{n_S+1}} - \hat{\rho}\begin{pmatrix} \mathbf{W_S} & \mathbf{W_{So}} \\ \mathbf{W_{oS}} & W_o \end{pmatrix}\}^{-1}.$$

As in Lesne et al. (2008), we use the Midi-Pyrénées region divided into $n = 283$ cantons for our study region. We construct a weight matrix $\mathbf{W}$ using the 10 nearest neighbors scheme (distance is based on the distance between centroids of the cantons).

| Predictor | In-sample predictors formulae |
|---|---|
| $BP$ | $\hat{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{BP}} = (\mathbf{I} - \rho\mathbf{W_S})^{-1}\mathbf{X_S}\hat{\boldsymbol{\beta}} - \mathrm{Diag}(\mathbf{Q_S})^{-1}\tilde{\mathbf{Q}}_{\mathbf{S}}(\mathbf{Y} - (\mathbf{I} - \rho\mathbf{W_S})^{-1}\mathbf{X_S}\hat{\boldsymbol{\beta}})$ |
| $TS$ | $\hat{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{TS}} = \mathbf{X_S}\hat{\boldsymbol{\beta}} + \hat{\rho}\mathbf{W}\mathbf{Y_S}$ |
| $TC$ | $\hat{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{TC}} = (\mathbf{I} - \hat{\rho}\mathbf{W})_{\mathbf{S}}^{-\mathbf{1}}\mathbf{X_S}\hat{\boldsymbol{\beta}}$ |

Table 1: In-sample predictors formulae

| Predictor | Out-of-sample predictors formulae |
|---|---|
| $BP$ | $\hat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{BP}} = \hat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{TC}} - \mathbf{Q_O}^{-1}\mathbf{Q_{OS}} \times (\mathbf{Y_S} - \hat{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{TC}})$ |
| $TC$ | $\hat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{TC}} = [(\mathbf{I} - \hat{\rho}\mathbf{W})^{-1}\mathbf{X}\hat{\boldsymbol{\beta}}]_O$ |
| $TS^1$ | $\hat{Y}_o^{TS^1} = \mathbf{X_o}\hat{\boldsymbol{\beta}} + \hat{\rho}\mathbf{W_{oS}}\mathbf{Y_S}$ |
| $BP_W$ | $\hat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{BP_W}} = \hat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{TC}} + \boldsymbol{\Sigma_{OS}}\mathbf{W'_{OS}}(\mathbf{W_{OS}}\boldsymbol{\Sigma_S}\mathbf{W'_{OS}})^{-1}(\mathbf{W_{OS}}\mathbf{Y_S} - \mathbf{W_{OS}}\hat{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{TC}})$ |
| $BP_N$ | $\hat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{BP_N}} = \hat{\mathbf{Y}}_{\mathbf{O}}^{\mathbf{TC}} - \mathbf{Q_O}^{-1}\mathbf{Q_{OJ}}(\mathbf{Y_J} - \hat{\mathbf{Y}}_{\mathbf{J}}^{\mathbf{TC}})_{\mathbf{J}}$ |
| $TC^1$ | $\hat{Y}_o^{TC^1} = \text{row } o \text{ of } \{\mathbf{I_{n_S+1}} - \hat{\rho}\begin{pmatrix}\mathbf{W_S} & \mathbf{W_{So}} \\ \mathbf{W_{oS}} & W_o\end{pmatrix}\}^{-1}$ |
| $BP^1$ | $\hat{Y}_o^{BP^1} = \hat{Y}_o^{TC^1} - \frac{1}{q_o}(\mathbf{Y_S} - \hat{Y}_S^{TC^1}))$ |
| $BP_W^1$ | $\hat{Y}_o^{BP_W^1} = \hat{Y}_o^{TC^1} + \boldsymbol{\Sigma_{oS}}\mathbf{W'_{oS}}(\mathbf{W_{oS}}\boldsymbol{\Sigma_S}\mathbf{W'_{oS}})^{-1}(\mathbf{W_{oS}}\mathbf{Y_S} - \mathbf{W_{oS}}\hat{\mathbf{Y}}_{\mathbf{S}}^{\mathbf{TC^1}})$ |
| $BP_N^1$ | $\hat{Y}_o^{BP_N^1} = \hat{Y}_o^{TC} - Q_o^{-1}\mathbf{Q_{oJ}}(\mathbf{Y_J} - \hat{\mathbf{Y}}_{\mathbf{J}}^{\mathbf{TC^1}})_J$ for $J$ set of indices of neighbors of $o$ |

Table 2: Out-of-sample predictors formulae

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 LAG regression model to generate the dependent variable

$$\mathbf{Y} = (\mathbf{I} - \rho\mathbf{W})^{-1}(\beta_0 + \mathbf{X_1}\beta_1 + \mathbf{X_2}\beta_2 + \mathbf{X_3}\beta_3 + \boldsymbol{\epsilon}) \text{ where } \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I_n}). \quad (22)$$

The parameter $\boldsymbol{\beta}$ and $\sigma$ are fixed to $\boldsymbol{\beta} = (0, 1/4, 6, 1)$ and $\sigma = 1$. For the in-sample comparison, $\rho$ takes a range of values $\rho = 0.05, 0.2, 0.35, 0.5, 0.65, 0.8, 0.9$. For the out-of-sample comparison, $\rho$ is equal to 0.5.

## 4.2 In-sample prediction simulation results

In this section, the sample contains the 283 initial sites described in section 4.1. For each choice of $\rho$ and $\sigma$, we draw 500 samples of the model and we compute the maximum likelihood estimates of the parameters based on the in-sample locations and the corresponding predictions. We use the total mean square error of prediction $MSE_k = \frac{1}{n}\sum_i^n (y_i - Y_i^k)^2$ for each method $k = TS, TC, BP$ to compare the quality of the predictors. Note that this criterion includes the statistical error due to parameter estimation. The results of the in-sample comparison are in Table 3.

| | $M\bar{S}E_{BP}$ | $M\bar{S}E_{TS}$ | $BP/TS$ | $M\bar{S}E_{TC}$ | $BP/TC$ |
|---|---|---|---|---|---|
| $\rho = 0.05$ | 0.9707 | 0.9720 | 0.9986 | 0.9754 | 0.9952 |
| | (0.0832) | (0.0833) | | (0.0838 ) | |
| $\rho = 0.2$ | 0.9850 | 0.9884 | 0.9966 | 1.0006 | 0.9844 |
| | (0.0832) | (0.0835) | | (0.0852) | |
| $\rho = 0.35$ | 0.9646 | 0.9756 | 0.9897 | 1.0192 | 0.9464 |
| | (0.0847) | (0.0841) | | (0.0896) | |
| $\rho = 0.5$ | 0.9597 | 0.9814 | 0.9779 | 1.0890 | 0.8813 |
| | ( 0.0799) | (0.0803) | | (0.1039) | |
| $\rho = 0.65$ | 0.9494 | 0.9883 | 0.9606 | 1.2531 | 0.7576 |
| | (0.0790) | (0.0799) | | (0.1450) | |
| $\rho = 0.8$ | 0.9308 | 0.9871 | 0.9429 | 1.6571 | 0.5660 |
| | (0.0844) | (0.0848) | | (0.2738) | |
| $\rho = 0.9$ | 0.9152 | 0.9878 | 0.9265 | 2.8981 | 0.3158 |
| | (0.0784) | (0.0812) | | (0.9635) | |

Table 3: MSE for different predictors and comparison with Best predictor when the parameter $\rho$ takes different values from 0.05 (mild correlation) to 0.9 (strong correlation).

The mean error is stable across values of $\rho$ for $TS$, is increasing for $TC$ and decreasing for $BP$. Variances are stable. The efficiency ratio $BP/TS$ is decreasing with spatial correlation but remains close to 1 whereas the efficiency ratio $BP/TC$ decreases dramatically with $\rho$. We do not report results for different values of $\sigma$ because they do not reveal any variation with respect to this parameter.

## 4.3 Out-of-sample prediction simulation results

To evaluate the performance of the different predictors for the out-of-sample case, we use the same model as before to generate the samples. The number of replications is 1000 and we report the average mean square error of prediction over the out-of-sample units.

We choose at random a given number of sites (27 or 54) which will be declared out-of-sample (in $O$). We predict the $\mathbf{Y}$ variable on the out-of-sample locations

based on the sample $S$ 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 Figures 3 and 4 and the level of aggregation is increasing from left to right. Table 4 summarizes the parameter estimates results (by maximum likelihood (ML) and by EM-algorithm (EM)) for configurations 1 and 3 and for 54 out-of-sample units. In general they 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.

The results for the case of 27 out-of-sample units are reported in table 5 and those for the case of 54 out-of-sample units are reported in table 6.

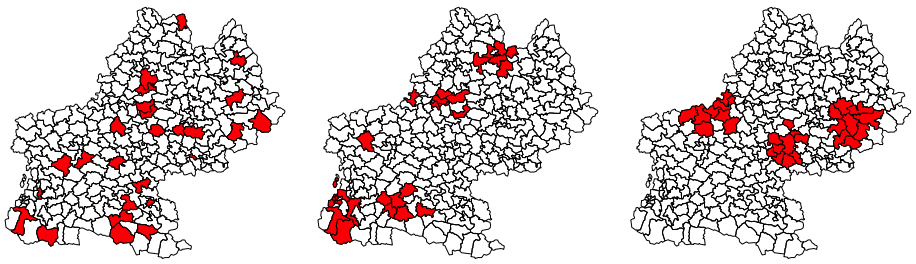

Figure 3: The three configurations for 27 out-of-sample units positions: configuration 1 (left), configuration 2 (center), configuration 3 (right).

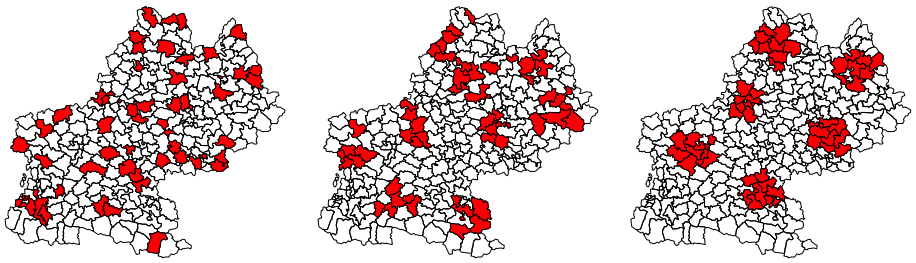

Figure 4: The three configurations for 54 out-of-sample units positions: configuration 1 (left), configuration 2 (center), configuration 3 (right).

Aside $BP^1$ whose rank changes, whatever configurations and number of sites to predict, we obtain the following ranking between methods in decreasing order of efficiency

$$BP \; < \; BP_N \; < \; BP_W \; < \; BP_N^1 \; < \; BP_W^1 \; < \; TS^1 \; < \; TC \; < \; TC^1.$$

| | ML estimates | | EM estimates | |
|---|---|---|---|---|
| | conf. 1 | conf. 3 | conf. 1 | conf. 3 |
| $\hat{\beta}_0$ | 1.575 | 2.401 | 1.566 | 0.581 |
| | (1.900) | (1.774) | (2.530) | (1.993) |
| $\hat{\beta}_1$ | 0.253 | 0.252 | 0.251 | 0.251 |
| | (0.023) | (0.022) | (0.023) | (0.022) |
| $\hat{\beta}_2$ | 5.978 | 5.840 | 5.962 | 5.962 |
| | (1.319) | (1.334) | (1.311) | (1.334) |
| $\hat{\beta}_3$ | 1.004 | 0.997 | 1.009 | 1.005 |
| | (0.073) | (0.072) | (0.076) | (0.0715) |
| $\hat{\rho}$ | 0.428 | 0.395 | 0.428 | 0.473 |
| | (0.078) | (0.074) | (0.109) | (0.084) |
| $\hat{\sigma}$ | 1.004 | 1.007 | 2.048 | 1.193 |
| | (0.098) | (0.095) | (3.098) | (0.388) |

Table 4: Parameter estimation results

Note that the worst ratio is around 0.88. As far as the impact of the level of aggregation is concerned, for predictors including a correction for spatial correlation such as $BP_W$, $BP_{W^1}$, $BP_N$ and $BP_N$ 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 other predictors ($TC$, $TC^1$ and $TS^1$ which do not correct for spatial correlation) since we observe that the prediction error for configuration 2 is higher than the two extreme cases 1 and 3. It seems that the previous effect is compensated at some point by the fact that the level of variability among out-of-sample units is lower in more aggregated configurations leading to an easier prediction problem.

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 7 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 1000 times the following process

- choose a site $i$ at random
- remove $k$ neighbors at random from the neighbors of $i$, their set is $N$
- the in-sample set of sites becomes $S \setminus N$ and the out-of-sample set of sites is $N$
- simulate the vector $\mathbf{Y}$ for all the sites
- predict the $\hat{\mathbf{Y}}$ on the sites in $N$ and compute the prediction error.

The first column of the table contains the predictive mean square error (PMSE) of the $BP$ predictor and the remaining ones report the ratio of the optimal PMSE

| | $BP$ | $BP^1$ | $\frac{BP}{BP^1}$ | $TC$ | $\frac{BP}{TC}$ | $TC^1$ | $\frac{BP}{TC^1}$ | $TS^1$ | $\frac{BP}{TS^1}$ |
|---|---|---|---|---|---|---|---|---|---|
| Conf. 1 | 0.998 | 1.000 | 0.998 | 1.126 | 0.886 | 1.131 | 0.883 | 1.026 | 0.973 |
| | (0.275) | (0.276) | | (0.303) | | (0.305) | | (0.280) | |
| Conf. 2 | 1.031 | 1.048 | 0.983 | 1.145 | 0.900 | 1.184 | 0.870 | 1.069 | 0.964 |
| | (0.296) | (0.302) | | (0.326) | | (0.340) | | (0.305) | |
| Conf. 3 | 1.038 | 1.060 | 0.979 | 1.129 | 0.919 | 1.144 | 0.908 | 1.064 | 0.976 |
| | (0.280) | (0.285) | | (0.307) | | (0.309) | | (0.285) | |

| | $BP_W$ | $\frac{BP}{BP_W}$ | ${BP_W}^1$ | $\frac{BP}{{BP_W}^1}$ | $BP_N$ | $\frac{BP}{BP_N}$ | ${BP_N}^1$ | $\frac{BP}{{BP_N}^1}$ |
|---|---|---|---|---|---|---|---|---|
| Conf. 1 | 1.003 | 0.996 | 1.007 | 0.992 | 0.999 | 1.000 | 1.003 | 0.996 |
| | (0.276) | | (0.277) | | (0.275) | | (0.277) | |
| Conf. 2 | 1.035 | 0.996 | 1.057 | 0.975 | 1.032 | 0.999 | 1.055 | 0.977 |
| | (0.298) | | (0.305) | | (0.297) | | (0.305) | |
| Conf. 3 | 1.041 | 0.997 | 1.065 | 0.975 | 1.039 | 1.000 | 1.063 | 0.977 |
| | (0.281) | | (0.286) | | (0.281) | | (0.285) | |

Table 5: Simulation results for the 27 out-of-sample units case

with the PMSE of all the other methods.

We observe that the $BP$ predictive mean square error indeed slightly increases with the number of missing neighbors. The efficiency of $BP^1$ 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 correction is inefficient). The efficiency of $BP_W$ with respect to $BP$ remains stable.

# 5 Conclusion

At least in the case of this particular model, the performance of $BP_N$, $BP_W$, $BP_N^1$,$BP_W^1$ are very close to that of the best prediction and much better than that of $TC$, $TS$, $TC^1$, $TS^1$. We did not consider a larger variety of parameter values because a few attempts have shown that the results were quite stable.

For the in-sample case, the performance of the trend-signal-noise predictor is not so bad and it is very easy to compute. $BP_N$ 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 $TS$.

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

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

| | $BP$ | $BP^1$ | $\frac{BP}{BP^1}$ | $TC$ | $\frac{BP}{TC}$ | $TC_1$ | $\frac{BP}{TC^1}$ | $TS^1$ | $\frac{BP}{TS^1}$ |
|---|---|---|---|---|---|---|---|---|---|
| Conf. 1 | 1.009 | 1.012 | 0.997 | 1.130 | 0.893 | 1.140 | 0.886 | 1.035 | 0.975 |
| | (0.196) | (0.196) | | (0.221) | | (0.225) | | (0.201) | |
| Conf. 2 | 1.029 | 1.036 | 0.992 | 1.137 | 0.905 | 1.152 | 0.893 | 1.054 | 0.975 |
| | (0.199) | (0.199) | | (0.226) | | (0.230) | | (0.204) | |
| Conf. 3 | 1.037 | 1.061 | 0.978 | 1.136 | 0.913 | 1.158 | 0.896 | 1.069 | 0.970 |
| | (0.205) | (0.213) | | (0.234) | | (0.240) | | (0.214) | |

| | $BP_W$ | $\frac{BP}{BP_W}$ | ${BP_W}^1$ | $\frac{BP}{{BP_W}^1}$ | $BP_N$ | $\frac{BP}{BP_N}$ | ${BP_N}^1$ | $\frac{BP}{{BP_N}^1}$ |
|---|---|---|---|---|---|---|---|---|
| Conf. 1 | 1.012 | 0.997 | 1.017 | 0.992 | 1.010 | 0.999 | 1.015 | 0.994 |
| | (0.196) | | (0.196) | | (0.196) | | (0.196) | |
| Conf. 2 | 1.031 | 0.998 | 1.042 | 0.987 | 1.029 | 1.000 | 1.039 | 0.989 |
| | (0.200) | | (0.201) | | (0.200) | | (0.201) | |
| Conf. 3 | 1.040 | 0.997 | 1.070 | 0.970 | 1.038 | 0.999 | 1.068 | 0.971 |
| | (0.206) | | (0.214) | | (0.206) | | (0.215) | |

Table 6: Simulation results for the 54 out-of-sample units case

with $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ because it can be written as a general linear model with $\boldsymbol{\mu} = (\mathbf{I} - \rho\mathbf{W})^{-1}(\alpha\boldsymbol{\iota} + \mathbf{X}\boldsymbol{\beta} + \mathbf{WX}\boldsymbol{\gamma})$ and variance given by (3). The difference between the LAG 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. The Kato (2008) approach for the SEM however cannot be extended directly for the LAG because the expression of the mean is quite different.

**Acknowledgments**. This work was supported by the French Agence Nationale de la Recherche through the ModULand project (ANR-11-BSH1-005). We wish to thank the referees for very interesting comments which lead us to improve substantially the paper and JP LeSage for helpful discussions about the topic.

# References

[1] Bennet B, Griffith DA, Haining R (1989) Statistical analysis of spatial data in the presence of missing observations: an application to urban census data. *Environment and Planning A* 21: 1511–1523

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

[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

| nb mis. neib. | $P\bar{M}SE_{BP}$ | $BP/BP^1$ | $BP/TC$ | $BP/TC^1$ | $BP/TS^1$ | $BP/BP_W$ |
|---|---|---|---|---|---|---|
| 0 | 0.979 | 1.000 | 0.898 | 0.898 | 0.984 | 0.998 |
| | (1.374) | | | | | |
| 1 | 0.987 | 0.989 | 0.905 | 0.887 | 0.972 | 0.997 |
| | (1.377) | | | | | |
| 2 | 0.986 | 0.987 | 0.902 | 0.884 | 0.973 | 1.000 |
| | (1.364) | | | | | |
| 3 | 1.015 | 0.981 | 0.924 | 0.901 | 0.978 | 0.998 |
| | (1.420) | | | | | |
| 4 | 0.996 | 0.971 | 0.909 | 0.871 | 0.958 | 0.989 |
| | (1.379) | | | | | |
| 5 | 1.002 | 0.959 | 0.913 | 0.850 | 0.947 | 0.991 |
| | (1.409) | | | | | |
| 6 | 1.015 | 0.930 | 0.923 | 0.850 | 0.930 | 0.995 |
| | (1.401) | | | | | |
| 7 | 1.042 | 0.902 | 0.944 | 0.844 | 0.914 | 1.002 |
| | (1.456) | | | | | |
| 8 | 1.036 | 0.798 | 0.939 | 0.756 | 0.806 | 0.989 |
| | (1.435) | | | | | |
| 9 | 1.059 | 0.694 | 0.957 | 0.706 | 0.709 | 0.993 |
| | (1.434) | | | | | |

Table 7: Prediction errors as a function of number of missing neighbors


[5] Gaetan C, Guyon X (2008) *Modélisation et statistique spatiales*. Springer-Verlag Berlin

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

[7] Griffith DA (2003) *Spatial autocorrelation and spatial filtering: gaining understanding through theory and scientific visualization*. Springer-Verlag Berlin

[8] 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*. Springer-Verlag Berlin

[9] Haining R (1990) *Spatial data analysis in the social and environmental sciences*. Cambridge University Press, Cambridge

[10] Harville DA (1997) *Matrix algebra from a statistician's perspective*. Springer-Verlag New-York

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

[12] 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

[13] 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

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

[15] LeSage JP, Pace RK (2008) Spatial Econometric Models, Prediction. In Shekhar S, Xiong H (eds) *Encyclopedia of Geographical Information Science*. Springer-Verlag New-York

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

[17] Rue H, Held L (2005) *Gaussian Markov Random Fields, Theory and Applications*. Chapman & Hall/CRC, Boca Raton, FL