Ridge regularization for spatial auto-regressive models with multicollinearity issues

Chavez-Chong Cristina, Cécile Hardouin, Ana Karina Fermin

Ridge regularization for spatial auto-regressive models with multicollinearity issues

Cristina Olimpia Chavez-Chong[1,2*†], Cécile Hardouin[2†] and Ana Karina Fermin[2†]

[1*]Department of Mathematics, Institute of Cybernetics, Mathematics and Physics, ICIMAF, 13th, Plaza de la Revolución, 10400, Havana, Cuba.

[2]MODAL'X, UPL, Univ. Paris Nanterre, CNRS, Av. de la République, Nanterre, 92000, Île-de-France, France.

*Corresponding author(s). E-mail(s): cristina@icimaf.cu;
Contributing authors: hardouin@parisnanterre.fr;
aferminrodriguez@parisnanterre.fr;
†These authors contributed equally to this work.

Abstract

This work proposes a new method for building an explanatory spatial autoregressive model in a multicollinearity context. We use Ridge regularization to bypass the collinearity issue. We present new estimation algorithms that allow for the estimation of the regression coefficients as well as the spatial dependence parameter. A spatial cross-validation procedure is used to tune the regularization parameter. In fact, ordinary cross-validation techniques are not applicable to spatially dependent observations. Variable importance is assessed by permutation tests since classical tests are not valid after Ridge regularization. We assess the performance of our methodology through numerical experiments conducted on simulated synthetic data. Finally, we apply our method to a real dataset and evaluate the impact of some socio-economic variables on the COVID-19 intensity in France.

Keywords: Spatial autoregressive models, Multicollinearity, Ridge regularization, Spatial cross-validation, Variable importance, Permutation tests

1 Introduction

Statistical models intend to represent, understand and interpret the information lying in a data set. Among others, linear regression analysis is a popular tool used to model underlying relationships between the response variable and covariates (or explanatory variables). This method is widely used in practice due to its advantages, namely the simplicity of the model structure and the consequent ease of interpretation of its results, in addition to the good properties of the estimator, which is unbiased and efficient. However, the assumptions of this model are violated in the case of spatial data, particularly the independence of residuals. There are many ways to take into account spatial dependence in linear regression models. Among them, we shall consider the class of spatial autoregressive models which arose in spatial econometrics. Spatial Autoregressive models, also called Simultaneous Autoregressive models, are widely used, see for instance the seminal works of Anselin, LeSage and their co-authors, mainly (Anselin, 1988) and (LeSage, 2008); their success relies on their intuitive writing, similar to ordinary regression models with the addition of a "spatial lag" term. We shall deal in this work with two well-known key models, the mixed spatial lag model and the spatial error model.

A common trait of real data is the presence of multicollinearity, i.e. some covariates are approximately linear combination of others. This phenomenon causes high variability of parameter estimators and biased inference statistics. Also, for model interpretation, the effects of the variables cannot be distinguished and extrapolation is likely to be misleading (Alin, 2010). Regularization techniques for regression are frequently applied to solve the issue. A quite general theory on the properties of the regularization techniques has been developed over the last decades, see for example (Hoerl and Kennard, 1988), (Zou and Hastie, 2005) and (Tibshirani, 1996). The most classical techniques, Ridge and Lasso, constrain the norm of the vector $\boldsymbol{\beta}$ of regression coefficients by adding a regularization term to the function to be optimized. This introduces some bias, but can greatly reduce the variance. The difference between Ridge and Lasso lies in the regularization term; the Ridge imposes an L^2-penalty which is $\gamma \times \sum_{i=1}^{p} \beta_i^2$ while Lasso considers a L^1-penalty equal to $\gamma \times \sum_{i=1}^{p} |\beta_i|$. In both cases, γ is an hyperparameter to be tuned. Ridge regression always keeps all the covariates in the model. On the contrary, Lasso produces a parsimonious model by allowing some coefficient estimates to be set to 0. In the situation of highly correlated covariates, Lasso may keep only one of them.

In this work, our objective is to build an explanatory model, which is different from building a predictive model. Indeed, when one wants to build a predictive model, the main goal is predictive accuracy. The objective is to obtain good predictions for the outcome, without any further consideration about the significance of the predictors and even their collinearity. It may happen that statistically significant variables are not included in a predictive model because their addition adds no predictive benefit. On the other hand, in an explanatory model objective, we want to identify the variables which are statistically significant to express their relationship with the response variable. Moreover, the knowledge of non-significant variables also provides valuable information for the practitioners. Thus, we aim at including all explanatory variables in the model

and keeping them all. To this aim, Lasso technique is not suitable since it performs variable selection, while the Ridge regularization method is in line with our objective.

There has been a tremendous amount of research about Ridge adaptation, see (McDonald, 2009) for a general review. Ridge regularization has been adapted to the spatial setting; Wheeler and Páez (Wheeler, 2009) consider Ridge adaptation in the framework of geographically weighted regression; more interesting in our framework, Fan et al. (2017) considered the context of spatial autoregressive models. They propose "Spatially Filtered Ridge Regression", which we describe in section 3.2. However, if the authors use a Ridge procedure for computing the coefficients estimates, they estimate the spatial dependence parameter as if there were no collinearity issues. In this work, we propose new estimation algorithms which take into account that matter for all the parameters.

Furthermore, Ridge regression involves the delicate choice of the regularization parameter γ. Indeed, this parameter plays a crucial role since it controls the strength of the regularization. The appropriate choice of this parameter is a difficult problem. A quite general theory on the properties of the various regularization methods and different parameter choice procedures has been developed over the last decades, starting from the seminal work of Tikhonov and Arsenin (1977). Fan et al. (2017) choose γ according to the Ridge trace criterion proposed by Hoerl and Kennard (1970), that is plotting Ridge coefficients versus γ values, and select the minimum value of γ for which the coefficients start to stabilize. Despite being commonly used in applications, this method suffers from being user-dependent, in the sense that the final choice is done by the user according to what he observes in the plot. Cross-validation is another common re-sampling method used to tune model parameters. But special care needs to be taken in the presence of spatial dependence; indeed, classical cross-validation techniques are no longer suitable since the assumption of independence between the training and test sets is violated, see Roberts et al. (2017). Here, we want to use an adapted spatial cross-validation procedure to select automatically the regularization parameter; we will make use of spatial leave-one-out (SLOO) which is especially well adapted for spatial autoregressive models.

Building an explanatory model, a crucial point is to determine which variables significantly affect the behaviour of the outcome. Unfortunately, one consequence of Ridge regression is that Student tests or classical F-tests are not appropriate any more. Halawa and El Bassiouni (2000) proposed a t-test based on Ridge estimators; however, other authors showed in intensive simulation studies that it may fail depending on the value of the regularization parameter. We propose permutation F-tests and t-tests.

In Section 2, we present the two main spatial autoregressive models that we are going to consider. We develop our new estimation procedures in Section 3. The crucial step of selecting the regularization parameter is achieved through a coherent spatial leave-one-out procedure. Finally, we propose in Section 4 to run permutation F-tests to assess the importance of the explanatory variables.

To validate the performance of our new procedure, we conduct comprehensive simulations whose results are summarise in Section 5; we simulate spatial autoregressive

models with 8 highly correlated variables, and estimate their parameters using classical spatial and non-spatial methods, the Spatially Filtered Ridge Regression proposed by Fan et al. (2017), and our procedure.

Finally, for purpose of illustration, we conduct a thorough study on a real data set in Section 6. We consider the hospitalization rate due to Covid-19 pandemic in metropolitan France, to be explained by socio-economic covariates. First we resume the data and include an exploratory analysis which highlights the presence of spatial dependence and multicollinearity. Then we run our estimation procedure and finally determine the core variables affecting the epidemic indicator. Our concluding remarks are provided in Section 7.

2 Spatial autoregressive models

Let us consider a finite set of sites $S = \{\boldsymbol{s}_i, \quad i = 1, \cdots, n\}$ on a spatial domain $D \subset \mathbb{R}^d$. We assume that we observe some data $\mathbf{Y} = (Y(\boldsymbol{s}_1), \cdots, Y(\boldsymbol{s}_n))^T$ on S and p explanatory variables $\mathbf{X}_1, \ldots, \mathbf{X}_p$.

Simultaneous spatial autoregressive models are common in spatial econometrics; they take into account spatial dependence structures by using a neighbourhood graph on S. In this work, we will consider two main models; the first one, that we shall denote SAR, is a mixed regressive-spatial autoregressive model defined by

$$\begin{aligned} \mathbf{Y} &= \rho W\mathbf{Y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon} &\sim N(\mathbf{0}, \sigma^2\mathbf{I}_n) \end{aligned} \tag{1}$$

where $\mathbf{I}_n$ is the identity matrix of order n and $\mathbf{X} = [\mathbf{X}_1 \cdots \mathbf{X}_p]$ is a $n \times p$ matrix with $\mathbf{X}_i = (X_i(\boldsymbol{s}_1), \cdots, X_i(\boldsymbol{s}_n))^T$, $i = 1, \ldots, p$. W is a deterministic spatial weights matrix; though there's no direct counterpart, it is often considered as the equivalent of the backshift operator B for time series, like a spatial lag operator acting as a shift over space. The spatial weights depend on the definition of a neighbourhood set for each observation. We set $w_{ii} = 0$ and $w_{ij} = 0$ if sites $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ are not neighbours. $(WY)_i$ is interpreted as a weighted average of the neighbouring values of $y(\boldsymbol{s}_i)$. Let us note that there are many specifications of the spatial weights, the usual way is to consider geographical distances (at a negative power), but one can consider economic or social distances. The parameter ρ in model (1) above reflects the strength of the spatial dependence between the elements of $\mathbf{Y}$.

The second widely used model is the Spatial Error Model, denoted by SEM, and defined by:

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \\ \mathbf{u} &= \lambda W\mathbf{u} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n) \end{aligned} \tag{2}$$

In this model $\mathbf{u}$ corresponds to a noise that is spatially correlated and λ characterizes the strength of the dependence.

These two models are well-defined under assumptions relying on the eigenvalues of the spatial weight matrix W that we will not discuss here; see for instance the works of Anselin (1988) or LeSage (2008).

Let us recall that our objective is to obtain explanatory SAR or SEM models under the assumption of multicollinearity between the covariates $\boldsymbol{X}_i$. Thus, the matrix $[\mathbf{X}^T\mathbf{X}]$ has very small eigenvalues, leading to numerical instability of its inverse. The poor behaviour of the estimator of the coefficients $\boldsymbol{\beta}$ in the presence of multicollinearity is inherited from ordinary linear regression by spatial autoregressive models. Furthermore, the estimation of the parameter ρ or λ is also computed indirectly from matrix $[\mathbf{X}^T\mathbf{X}]$. The Ridge regularization is a way to overcome the problem, bypassing the multicollinearity issue while preventing from variable removal. But it has to be adapted to the spatial framework. We present new algorithms to derive Ridge estimates for SAR and SEM models in the next paragraph.

3 Estimation

To exhibit the novelty of our procedure, we start by recalling the usual estimation procedure to estimate SAR and SEM models defined in (1) and (2).

3.1 Spatial autoregressive models estimation

Let us recall that the least squares estimation of these models produces biased and inconsistent estimates. Therefore, one considers the following procedure, which is described for example in Anselin (1988). The estimation is based on the so-called concentrated likelihood and is achieved in a few steps.

Let us start with the SAR model. The log-likelihood has the following expression,

$$\begin{aligned} l_{SAR}(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2, \rho) = & -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) + \ln|\mathbf{I}_n - \rho W| \\ & - \frac{1}{2\sigma^2}\big((\mathbf{I}_n - \rho W)\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\big)^T\big((\mathbf{I}_n - \rho W)\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\big) \end{aligned} \tag{3}$$

We first focus on parameter $\boldsymbol{\beta}$. Its maximum likelihood estimator is

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{ML} = & \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{I}_n - \rho W)\mathbf{Y} \\ = & \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} - \rho\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T W\mathbf{Y} \\ = & \hat{\boldsymbol{\beta}}_O - \rho\hat{\boldsymbol{\beta}}_L \end{aligned} \tag{4}$$

where $\hat{\boldsymbol{\beta}}_O = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$ is the ordinary least squares (OLS) estimator of the coefficients of the regression of $\mathbf{Y}$ on $\mathbf{X}$ and $\hat{\boldsymbol{\beta}}_L = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T W\mathbf{Y}$ is the OLS estimator of the coefficients of the regression of the spatial lag $(W\mathbf{Y})$ on $\mathbf{X}$. So, we obtain $\hat{\boldsymbol{\beta}}_{ML}$ as soon as ρ is known. The idea is then to perform OLS on the two regressions, and compute the residuals $e_O = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_O$ and $e_L = W\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_L$. Then, it

also can be shown that the ML estimate of σ^2 is given by

$$\widehat{\sigma^2} = \frac{1}{n}(e_O - \rho e_L)^T(e_O - \rho e_L). \tag{5}$$

Again, we obtain this estimate as soon as ρ is known. Substituting σ^2 and $\boldsymbol{\beta}$ in the log-likelihood (3) by their expressions in equations (5) and (4), we obtain the concentrated log-likelihood function,

$$l^C_{SAR}(\mathbf{y} \mid \rho) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\{\frac{1}{n}(e_O - \rho e_L)^T(e_O - \rho e_L)\} + \ln|\mathbf{I}_n - \rho W| \tag{6}$$

We then find $\hat{\rho}$ maximising (6), which resumes in a one-parameter non-linear optimization problem. The final step consists in plugging $\hat{\rho}$ in (4) and (5) to obtain the final estimators.

Let us now turn to the SEM model. The estimation procedure is analogous to the previous one. We write the log-likelihood as,

$$\begin{aligned} l_{SEM}(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2, \lambda) = & -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) + \ln|\mathbf{I}_n - \lambda W| \\ & - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{I}_n - \lambda W)^T(\mathbf{I}_n - \lambda W)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned} \tag{7}$$

Let us note the filtered variables $\mathbf{X}_\lambda = \mathbf{X} - \lambda W\mathbf{X}$ and $\mathbf{Y}_\lambda = \mathbf{Y} - \lambda W\mathbf{Y}$. Then, for a fixed λ, we get the maximum likelihood estimators $\widehat{\boldsymbol{\beta}_\lambda} = \left(\mathbf{X}_\lambda^T\mathbf{X}_\lambda\right)^{-1}\mathbf{X}_\lambda^T\mathbf{Y}_\lambda$ and $\widehat{\sigma^2_\lambda} = \frac{1}{n}\mathbf{e}_\lambda^T\mathbf{e}_\lambda$ with $\mathbf{e}_\lambda = \mathbf{Y}_\lambda - \mathbf{X}_\lambda\widehat{\boldsymbol{\beta}_\lambda}$. One can notice that these estimators are those obtained from writing again the SEM model (2) as

$$\mathbf{Y}_\lambda = \mathbf{X}_\lambda^T\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n) \tag{8}$$

Substituting the expressions of $\widehat{\boldsymbol{\beta}_\lambda}$ and $\widehat{\sigma^2_\lambda}$ in the log-likelihood (7), we obtain the concentrated log-likelihood,

$$l^C_{SEM}(\mathbf{y} \mid \lambda) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\{\frac{1}{n}\mathbf{e}_\lambda^T\mathbf{e}_\lambda\} + \ln|\mathbf{I}_n - \lambda W| \tag{9}$$

which is a non-linear function of λ. Again, optimizing this function gives $\hat{\lambda}$, then the final estimates $\widehat{\boldsymbol{\beta}_\lambda}$ and $\widehat{\sigma^2_\lambda}$.

In both models, the procedures will suffer from collinearity issues and produce unstable numerical solution of the estimate of $\boldsymbol{\beta}$. Therefore, the residuals derived from $\widehat{\boldsymbol{\beta}}$ inherit from its instability. Finally, $\widehat{\sigma^2}$ as well as the estimates of ρ or λ can't be trusted. The impact of collinearity is felt all the way to all the parameters. It seems then important to consider all parameters in the new estimation algorithms to release them from the collinearity burden.

3.2 Ridge regression for spatial autoregressive models

In ordinary non-spatial regression, the Ridge estimator is defined by

$$\hat{\boldsymbol{\beta}}_R = (\mathbf{X}^T\mathbf{X} + \gamma\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{Y}, \tag{10}$$

$\gamma > 0$ being the so-called regularization parameter. Let us note that in all this paragraph, we consider centred variables, since the regularization concerns the coefficients of the variables without the intercept. It is also very common to scale the variables.

Spatially filtered Ridge regression (SFRR) as named by the authors who first proposed it (Fan et al., 2017) integrates Ridge regression and spatial autoregressive models. This method follows three fundamental steps. For the SAR model, [resp. the SEM model],

1. Estimate ρ [resp. λ].
2. Consider the new response vector to be $\mathbf{Y}_{\hat{\rho}} = (\mathbf{I} - \hat{\rho}W)\mathbf{Y}$ [resp. $\mathbf{Y}_{\hat{\lambda}} = (\mathbf{I} - \hat{\lambda}W)\mathbf{Y}$ and $\mathbf{X}_{\hat{\lambda}} = (\mathbf{I} - \hat{\lambda}W)\mathbf{X}$].
3. Select γ following Ridge trace criterion and estimate $\hat{\boldsymbol{\beta}}_R$ in (10) with the new response vector (and current or new design matrix).

Fan et al. do not raise it, but the first step is, in fact, an issue, since, as we mentioned above, the ML estimator of ρ (or λ) is obtained using residuals, derived on their own from the ill-conditioned matrix $[\mathbf{X}^T\mathbf{X}]$. We bypass the problem in proposing an iterative algorithm which takes into consideration both parameters ρ (or λ) and $\boldsymbol{\beta}$. The estimation of σ^2 automatically follows.

The algorithms we propose for estimating parameters of the SAR and SEM models are similar.

Ridge Regularization for SAR models (RRSAR)

1. Initialization. Consider the ordinary linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and estimate $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}_{0,R} = (\mathbf{X}^T\mathbf{X} + \gamma_0\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{Y}$ defined in (10) for $\gamma_0 > 0$. Similarly, for $\gamma_L > 0$, compute $\hat{\boldsymbol{\beta}}_{L,R} = (\mathbf{X}^T\mathbf{X} + \gamma_L\mathbf{I}_p)^{-1}\mathbf{X}^TW\mathbf{Y}$ the ridge estimate of $\boldsymbol{\beta}$ in the ordinary regression of $(W\mathbf{Y})$ on $\mathbf{X}$.
2. Compute $\mathbf{e}_0 = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{0,R}$ and $\mathbf{e}_L = W\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{L,R}$ and estimate ρ by using the concentrated ML (6).
3. Consider the filtered $\mathbf{Y}_{\hat{\rho}} = (\mathbf{I} - \hat{\rho}W)\mathbf{Y}$ and compute $\hat{\boldsymbol{\beta}}_R^{SAR} = (\mathbf{X}^T\mathbf{X} + \gamma\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{Y}_{\hat{\rho}}$ for $\gamma > 0$.

Ridge Regularization for SEM models (RRSEM)

1. Initialization. Consider the ordinary linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and estimate $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}_R$ defined in (10).
2. Consider $\mathbf{Y}_\lambda = \mathbf{X}_\lambda\hat{\boldsymbol{\beta}}_R + \boldsymbol{\varepsilon}$ and $\mathbf{e}_\lambda = \mathbf{Y}_\lambda - \mathbf{X}_\lambda\hat{\boldsymbol{\beta}}_R$ and estimate λ by using the concentrated ML (9).
3. Consider the filtered $\mathbf{Y}_{\hat{\lambda}} = (\mathbf{I} - \hat{\lambda}W)\mathbf{Y}$, the filtered matrix $\mathbf{X}_{\hat{\lambda}} = (\mathbf{I} - \hat{\lambda}W)\mathbf{X}$, and compute $\hat{\boldsymbol{\beta}}_R^{SEM} = (\mathbf{X}_{\hat{\lambda}}^T\mathbf{X}_{\hat{\lambda}} + \gamma\mathbf{I}_p)^{-1}\mathbf{X}_{\hat{\lambda}}^T\mathbf{Y}_{\hat{\lambda}}$ for $\gamma > 0$.

4. Consider $\mathbf{Y}_\lambda = \boldsymbol{X}_\lambda \hat{\boldsymbol{\beta}}_R^{SEM} + \boldsymbol{\varepsilon}$ and $\mathbf{e}_\lambda = \mathbf{Y}_\lambda - \mathbf{X}_\lambda \hat{\boldsymbol{\beta}}_R^{SEM}$ and estimate new λ by using the concentrated ML (9).
5. Repeat step 3 to obtain the final $\hat{\boldsymbol{\beta}}_R^{SEM}$.

Let us note that the RRSAR algorithm follows the original estimation algorithm of SAR models; Ridge estimates replace OLS estimates in the first step, and $\hat{\boldsymbol{\beta}}^{SAR}$ is changed to $\hat{\boldsymbol{\beta}}_R^{SAR}$. Let us point out that the regularization parameters γ in the first step are not identical, we get different values whether considering the dependent variable to be $\mathbf{Y}$ or $W\mathbf{Y}$. This is crucial to get a correct estimation of ρ. Then we can't just plug-in the obtained estimates in (4) but we need to regularize globally $\boldsymbol{\beta}$ considering the new dependent variable $\mathbf{Y}_{\hat{\rho}}$.

The RRSEM algorithm has two more steps than the RRSAR because the first estimate of λ in step 2 is obtained after an OLS procedure and has to be refined.

In these algorithms, the computation of the Ridge estimates necessitate substeps to determine a good regularization parameter γ. This is done by spatial cross-validation. This step is the subject of the next section 3.3.

3.3 Spatial cross-validation for selecting the Ridge parameter

The parameter γ has a key role in Ridge regularization ; assigning γ to 0 is considering ordinary least squares (OLS); on the contrary, high values of γ increase the penalty term and thus drags down the regression coefficients β_i. Each value leads to different estimates of $\boldsymbol{\beta}$, crushing them more or less towards zero. In their algorithm, Fan et al. (2017) follow the Ridge trace criterion (Hoerl and Kennard, 1970) to choose γ. They plot the coefficients estimates versus γ and look for the value of γ for which the coefficients stabilize. Another common method is selecting the parameter γ that minimizes the mean square error estimated using cross-validation techniques.

Spatial data, and more accurately, spatial auto-correlation, challenges classical cross-validation techniques. In this situation, random splitting of the data into training and testing sets does no longer simulates the original structure of the data and therefore the key assumption of independent data samples behind cross-validation is violated. There exist some extensions of classical cross-validation to deal with spatial dependence. Detailed works and applications have been conducted on spatial leave-one-out (SLOO) (Le Rest et al., 2013), spatial k-fold cross-validation (Pohjankukka et al., 2017) and blocked cross-validation (Brenning, 2012). The majority of these modifications correspond to the idea of achieving independence between the training set and the test set by the means of "point separation". This is done by deleting points from the training set within a distance h of the test set. The area of deleted points is called "buffer" or dead zone. The selection criterion for h is still under debate. In the case of spatial autoregressive models, SLOO rises naturally because the determination of h is beneficially replaced by taking out the first order neighbours designed by W.

Spatial leave-one-out follows the usual leave-one-out procedure; at the $m-$th iteration of SLOO, one observation becomes the validation set $\mathcal{V}_m = \{(y(\boldsymbol{s}_m), \boldsymbol{x}(\boldsymbol{s}_m))\}$, as in the classical method. Then, we define the buffer around $\boldsymbol{s}_m$ as the set of its neighbours $\partial \boldsymbol{s}_m = \{\boldsymbol{s}_i \in S : \ w_{mi} \neq 0\}$. The training set is then $\mathcal{T}_m = \{(y(\boldsymbol{s}_j), \boldsymbol{x}(\boldsymbol{s}_j)) :$

$w_{mj} = 0\}$. If needed, one can easily extend the buffer to take out the k-th order neighbouring observations, k being chosen for instance following a Moran test (Moran, 1950) based on the k-th order neighbours.

We describe now the SLOO algorithm that we use to select the regularization parameter γ. Let us consider, for example, the SAR model; the procedure is similar for the SEM model. Our criterion is based on the maximisation of the likelihood rather than MSE; indeed, MSE is mostly used for prediction purposes, while we are interested in an explicative model.

For a fixed γ, we run the following procedure. For $m = 1$ to n, we compute $\hat{\boldsymbol{\beta}}^{SAR}_{R,m}(\gamma) = (\mathbf{X}_m^T\mathbf{X}_m + \gamma\mathbf{I}_p)^{-1}\mathbf{X}_m^T\mathbf{Y}_{\hat{\rho},m}$; the subscript m states for the calculations are made on the training set $\mathcal{T}_m$. Note that $\hat{\rho}$ has been obtained from computations on the whole data set at the previous step of the estimation algorithm and is fixed at this point. The same applies for $\hat{\sigma}^2$.

Next we compute the conditional log-likelihood on the validation set, conditionally to the neighbouring values,

$$l_m(\gamma) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}\left(y(\boldsymbol{s}_m) - \hat{\rho}\sum_{\boldsymbol{s}_j\in\partial\boldsymbol{s}_m} w_{mj-}y(\boldsymbol{s}_j) - \boldsymbol{x}(\boldsymbol{s}_m)^T\hat{\boldsymbol{\beta}}^{SAR}_{R,m}(\gamma)\right)^2.$$

Finally, the regularization parameter $\hat{\gamma}$ selected in the estimation algorithm is defined by

$$\hat{\gamma} := \underset{\gamma\in\{\gamma_k\}_{k=1}^K}{\arg\max}\ \frac{1}{n}\sum_{m=1}^{n} l_m(\gamma). \tag{11}$$

where $\{\gamma_k\}_{k=1}^K$ is a sequence of possible values of γ; this sequence is created using the path-wise coordinate descend method (Friedman et al., 2010); it first selects γ_{max}, the smallest value that crushes down $\hat{\boldsymbol{\beta}}$ to $\mathbf{0}$; then the sequence of K values decreases on the log-scale from γ_{max} to $\gamma_{min} = c * \gamma_{max}$. The most common choices are $K = 100$ and $c = 0.001$.

4 Importance of the explanatory variables

When building an explanatory model, one major point is to determine how important is the influence of the explanatory variables on the dependent variable $\boldsymbol{Y}$. For spatial autoregressive models, the interpretation of coefficients is not classical, it is made in terms of "impacts" rather than exploiting coefficients values. Indeed, if we consider a SAR model, for example, the presence of the spatially lagged dependent variable entails, as a logical consequence, that a change in an explanatory variable at a single location can affect several values of $\boldsymbol{Y}$ in other sites. These impacts depend directly on ρ and the components of $\boldsymbol{\beta}$, see LeSage (2008). Furthermore, in the context of Ridge regression, the coefficients suffer a steep fall due to the regularization; then, classic tests are not valid any more.

Halawa and El Bassiouni (2000) propose to use a statistics that imitates the classic Student ratio to test individual regression coefficients in the case of ordinary regression (without spatial dependence).

Let us recall that the bias and variance of the Ridge estimator of the regression coefficients $\boldsymbol{\beta}$ are defined by respectively,

$$\begin{aligned} bias(\hat{\boldsymbol{\beta}}_R) &= E[\hat{\boldsymbol{\beta}}_R] - \boldsymbol{\beta} = -\gamma \left(\mathbf{X}^T\mathbf{X} + \gamma\mathbf{I}\right)^{-1} \boldsymbol{\beta} \\ Var(\hat{\boldsymbol{\beta}}_R) &= \sigma^2 \left(\mathbf{X}^T\mathbf{X} + \gamma\mathbf{I}\right)^{-1} \mathbf{X}^T\mathbf{X} \left(\mathbf{X}^T\mathbf{X} + \gamma\mathbf{I}\right)^{-1} \end{aligned} \tag{12}$$

To test the null hypothesis $(H_0) : \beta_{R,j} = 0$, Halawa and El Bassiouni (2000) propose to consider $T_j = \dfrac{\hat{\beta}_{R,j}}{S(\hat{\beta}_{R,j})}$, where $\hat{\beta}_{R,j}$ is the j-th element of $\hat{\boldsymbol{\beta}}_R$ and $S(\hat{\beta}_{R,j})$ is the square root of the j-th diagonal element of $Var(\hat{\boldsymbol{\beta}}_R)$. They state that this statistics follows approximately a Student distribution. In practice, σ^2 in (12) is replaced by its estimation based on the residuals. However, Perez-Melo and Kibria (2020) show that the test behaves differently according to the regularization parameter; in his thesis (available at https://theses.hal.science/tel-01326950v2) Bécu also states that the test fails especially for large values of γ.

Permutation tests offer a good alternative to determine statistical significance. They are based on the assumption that under the hypothesis of no relationship between the dependent variable $\boldsymbol{Y}$ and an explanatory variable X_j, the observations are exchangeable and the joint probability distribution of the permuted samples coincides with the joint probability distribution of the original sample. There are numerous works proposing permutation tests for OLS with the permutation done to the response (Manly, 2006), or the residuals (Kennedy, 1995; Anderson and Robinson, 2001), or the predictor (Hastie and Tibshirani, 1995). Thus, Bécu et al. (2017) propose to replace the classic F-test by a permutation F-test, which is asymptotically exact when the tested variable is independent of the other explanatory variables, and is approximate in the general case. Though the assumption of independence is violated, this test might perform well in the case of not so high correlations.

The permutation F-test is defined as follows. First, let us note $\mathcal{M}_0$ the small model (without the variable to be tested) nested in $\mathcal{M}_1$ the larger model (including all the variables), $\hat{\mathbf{Y}}_0$ and $\hat{\mathbf{Y}}_1$ the respective predictors of $\mathbf{Y}$ estimated in these models. The classic F-statistic is defined by

$$F = \frac{\|\boldsymbol{Y} - \hat{\mathbf{Y}}_0\|^2 - \|\boldsymbol{Y} - \hat{\mathbf{Y}}_1\|^2}{\|\boldsymbol{Y} - \hat{\mathbf{Y}}_1\|^2/(n-p)} \tag{13}$$

Now, let us permute B times the n values of X_j, and let us note $X_{j(b)}$ the b-th permutation. We consider the estimator of the parameters of the spatial autoregressive model with the explanatory variables $(X_1, \cdots, X_{j-1}, X_{j(b)}, X_{j+1}, \cdots, X_p)$. $\hat{\boldsymbol{Y}}_{j(b)}$ is the corresponding predictor of $\boldsymbol{Y}$. We define the permutation test statistic $F_{j(b)}$ by

$$F_{j(b)} = \frac{\|\boldsymbol{Y} - \hat{\mathbf{Y}}_0\|^2 - \|\boldsymbol{Y} - \hat{\mathbf{Y}}_{j(b)}\|^2}{\|\boldsymbol{Y} - \hat{\mathbf{Y}}_{j(b)}\|^2/(n-p)} \tag{14}$$

The estimated p-value is

$$\mathrm{p_v} = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\left\{F_{j(b)} \geq F_j\right\} \tag{15}$$

where F_j is the usual F-statistic defined in equation (13) above. This resumes to

$$\mathrm{p_v} = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\left\{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}_1\|^2 \geq \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}_{j(b)}\|^2\right\} \tag{16}$$

Similarly, we can derive a permutation t-test considering the following statistics,

$$T_{j(b)} = \frac{\hat{\beta}_{R,j}}{S(\hat{\beta}_{R,j})} \tag{17}$$

associated to the p-value

$$\mathrm{p_v} = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\left\{T_{j(b)} \geq T_j\right\} \tag{18}$$

Note that these permutation tests are considered to be more conservative tests than parametric tests; moreover in our case, they may fail in case of very large values of γ too, and most of all, they do not take into account spatial dependence in the model and potentially spatially correlated covariates. We point out that in all the tests above, γ is chosen by the user and in our case, results from SLOO (each permutation gives its own optimal value).

5 Simulation experiments

In this section, we present simulation results to assess the performance of our methodology. We consider various scenarios, always in a multicollinearity framework. We simulate SAR and SEM models, with 8 highly correlated covariates, as described further, and for various values of ρ or λ in order to consider weak to strong spatial dependence. For each model, we explore two settings; in the first one, we simulate the covariates once for all and only the error term is renewed at each simulation. In the second scenario, we simulate the covariates each time. Then, each resulting dataset is estimated according to different methods depending on whether or not the spatial feature and the Ridge regularization are taken into consideration. The code for our simulations is available in our Github repository https://github.com/c0ra/RRSARMMI along with the full results.

We consider a 30x30 grid over which we generate 8 covariates with multicollinearity issues. Here, we use the function RFsimulate from the package RandomFields in R.

The first two variables are generated as Gaussian random fields; X_1 is generated with an exponential covariance function with variance equal to 1 and scale set to 0.5, while X_2 has a Gaussian covariance function with variance 1 and scale set to 0.4. The remaining six variables are generated as follows,

$$X_3 = \exp(X_1) - |X_2| \qquad X_4 = |X_2| + (\frac{X_1 + X_2}{2} - 4)^2$$

$$X_5 = X_1 + X_1X_2 \qquad X_6 = \log(X_4) - \frac{X_1}{12} + X_1^2$$

$$X_7 = X_1 + 2X_2 + \sqrt{X_4} \qquad X_8 = X_5 + \frac{X_2}{2} + X_2^2$$

Afterwards, all covariates are standardized to build the matrix $\mathbf{X}$. In the deterministic scenario, this matrix is the same for all simulations; the high multicollinearity is expressed by very large values of matrix $\mathbf{X}^T\mathbf{X}$'s condition number, 138 473.1, and pf the variance inflation factor (VIF) for the covariates, given in Table 1; moreover we plot the correlation matrix of the covariates in Figure 1.

Table 1: Variance inflation factor (VIF) of the simulated covariates

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
4821.923	16507.146	13.756	3804.250	13.649	29.921	10690.157	74.898

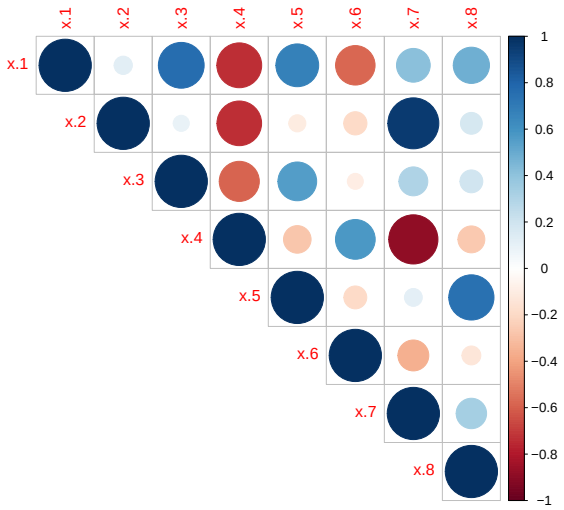


Fig. 1: Correlations between the 8 simulated covariates

The spatial weight matrix W was constructed using a row-standardized contiguity matrix with the rook neighbourhood (which is similar on a regular grid to the four nearest neighbours system). Then we simulate the dependent variable $\mathbf{Y}$ according to the SAR or the SEM model, with 5 different values of the spatial autoregressive coefficient reflecting weak to strong dependence, ρ $(resp.\lambda) \in (0.1, 0.3, 0.5, 0.7, 0.9)$. The dependent variable $\mathbf{Y}$ was obtained by either

$$\mathbf{Y} = (I_n - \rho W)^{-1}\mathbf{X}\boldsymbol{\beta} + (I_n - \rho W)^{-1}\boldsymbol{\varepsilon}, \tag{19}$$

or

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + (I_n - \lambda W)^{-1}\boldsymbol{\varepsilon}, \tag{20}$$

where $\mathbf{X}$ is the covariates matrix, $\boldsymbol{\beta}$ is a vector of ones, and $\boldsymbol{\varepsilon}$ follows a Gaussian distribution with mean 0 and standard deviation 1. Then we centre $\mathbf{Y}$ but we do not scale it.

Each model is simulated 500 times, and then estimated. We consider the following estimation procedures, whose abbreviation that we shall use in the summaries is given between parenthesis: Ordinary least squares regression (OLS), ordinary Ridge regression (RR), ordinary SAR or SEM (SAR or SEM) without regularization, Spatially Filtered Ridge Regression (SFRR) proposed by Fan et al. (2017), and our estimation procedure, named as Ridge Regression for SAR models (RRSAR) or Ridge Regression for SEM models (RRSEM). In this regular grid setting, the buffer around each site $\mathbf{s}$ during the SLOO procedure is defined by the rook neighbourhood.

For comparability purpose we compute the bias, variance and mean square error (MSE) of the regression coefficients β_1 to β_8 as well as for ρ or λ.

We summarize the results in the tables hereafter. We only display the results for the case of "stochastic" covariates simulated each time; those obtained for the case where we simulate them once for all are very similar. Moreover, for sake of place, we gathered the results for the regression coefficients in presenting the average of their bias, variance and MSE. We do not present the results between SFRR, RRSAR and RRSEM for the dependence parameters ρ and λ. Indeed, the bias, variances, and MSE of λ estimates are identical up to the third digit for SFRR and RRSEM, equal to ± 0.000; they are similar for the SAR framework, with a small difference in the bias; we observe a small bias of -0.001 in our procedure for ρ equal to 0.3 to 0.7, while it is -0.000 following the SFRR procedure. Detailed results for each regression coefficient and for each value of ρ or λ are retrievable from our Github repository.

Let us note that if we are going to compare the MSE between the different estimation procedures, one has to keep in mind that other diagnostics are ignored here, but obviously, SAR and SEM lead to poor estimates due to the multi-collinearity issue, we can get low variances from RR but it suffers from the absence of spatial dependence in the model, and OLS endures both inconveniences. Our real challenge is to compare our procedure with SFRR.

When examining the average bias, variance and MSE of regression coefficient estimates (Tables 2, 3, 4), a clear pattern emerges as we vary the value of the dependence parameter. Bias and variance (and thus MSE) constantly increase with the spatial dependence parameter for OLS, leading sometimes to crazy values (we get an average variance of 656.4 for $\rho = 0.9$). This illustrates the failure of this model in our framework and the strong need of regularization. SAR and SEM estimation procedures help to reduce the bias, which is almost constant whatever the strength of the spatial dependence; on the other hand, they lead to high (stable) variances of order 5 or 6. Unsurprisingly, the RR estimation procedure helps a lot to reduce the variances but may lead to a large bias (2.5 for $\rho = 0.9$); it is interesting to note that the variances increase with ρ and λ denoting the lack of spatial feature inclusion. To summarize, as we transition from 0.1 to 0.9, OLS experiences a significant surge in MSE, multiplied by more than 100. In contrast, RR's MSE average is multiplied by "only" 25 for the SAR simulation, and increase by 40% over the same range for the SEM one.

As expected, the best results are obtained for SFRR and our algorithms. In both cases, the bias is more important than for ordinary SAR and SEM estimation, but this is expected since these techniques force the coefficients towards zero. Let us note that in all cases, the bias is more important after SFRR estimation than after our procedures.

Especially in the SEM simulation, SFRR leads to increasing larger bias (-1 for $\lambda = 0.9$ and SFRR versus -0.4 for RRSEM). SFRR estimation leads to very low variances, lower than ours. Interestingly, though they stay constant for the SAR simulation over the ρ variation, they increase with λ in the SEM framework for both methods, SFRR and RRSEM. Considering the MSE, our method is superior to SFRR, for both models. Our conclusion is that SFRR crushes down the coefficients too far, raising the bias; indeed our regularization procedure is sufficient to stabilise the coefficients, leading to lower bias.

Table 2: Average regression coefficient bias

	SAR simulation					SEM simulation				
ρ/λ	OLS	SAR	RR	SFRR	RRSAR	OLS	SEM	RR	SFRR	RRSEM
0.1	0.044	-0.236	-0.521	-0.551	-0.445	0.164	0.161	-0.533	-0.576	-0.422
0.3	0.746	-0.233	-0.360	-0.550	-0.444	0.166	0.165	-0.533	-0.703	-0.413
0.5	1.803	-0.229	-0.112	-0.547	-0.443	0.176	0.168	-0.547	-0.810	-0.406
0.7	3.714	-0.224	0.441	-0.548	-0.443	0.199	0.168	-0.559	-0.933	-0.402
0.9	9.092	-0.220	2.567	-0.543	-0.443	0.280	0.167	-0.551	-1.051	-0.400

Table 3: Average regression coefficient variance

	SAR simulation					SEM simulation				
ρ/λ	OLS	SAR	RR	SFRR	RRSAR	OLS	SEM	RR	SFRR	RRSEM
0.1	6.739	5.788	0.053	0.053	0.109	6.277	6.201	0.068	0.070	0.131
0.3	11.702	5.778	0.078	0.053	0.109	6.968	6.227	0.074	0.079	0.143
0.5	26.719	5.769	0.083	0.052	0.109	8.484	6.126	0.075	0.110	0.155
0.7	83.771	5.765	0.167	0.052	0.109	12.301	5.909	0.086	0.141	0.166
0.9	656.366	5.776	1.519	0.054	0.109	32.018	5.598	0.194	0.209	0.170

Table 4: Average regression coefficient MSE

	SAR simulation					SEM simulation				
ρ/λ	OLS	SAR	RR	SFRR	RRSAR	OLS	SEM	RR	SFRR	RRSEM
0.1	6.741	5.844	0.323	0.356	0.307	6.304	6.226	0.352	0.401	0.309
0.3	12.258	5.832	0.207	0.355	0.307	6.996	6.254	0.358	0.573	0.313
0.5	29.971	5.822	0.095	0.351	0.306	8.514	6.154	0.374	0.766	0.320
0.7	97.565	5.815	0.361	0.353	0.306	12.341	5.938	0.399	1.013	0.328
0.9	739.036	5.824	8.107	0.349	0.305	32.096	5.625	0.497	1.314	0.330

6 Application

We apply our methodology to a real data set. The data is related to the Covid-19 epidemic in metropolitan France in 2020. The observations are collected from the 96 "départements", which are administrative units. Their geographic distribution is illustrated in Figure 2, including the Corsica island, and we will call them "departments" from now. We consider the hospitalization rate due to the Covid, reflecting the strength of the epidemic, and a set of socio-economic covariates, which are highly correlated. The aim of the study is to determine whose covariates have a significant impact on the dependent variable. First, we conduct an exploratory analysis of the data to bring out spatial dependence as well as multi-collinearity issues. Then we consider both SAR and SEM models; we estimate the coefficients of each model using Ridge regularization, according to one or the other of the iterative algorithms RRSAR and RRSEM derived in section 3.2; the Ridge parameter is chosen following a spatial leave-one-out procedure, as described in paragraph 3.3. Once the model is estimated, the significance of the explanatory variables is determined by running permutation F-tests presented in section 4. Finally, we compare our results with those obtained by classical estimation of SAR and SEM models. Our code and results are available on our Github repository https://github.com/c0ra/RRSARMMI

6.1 Data description

The COVID-19 epidemic in France, upon its arrival at the end of February 2020, did not affect the entire territory equally. The "Grand Est", "Hauts-de-France" and "Ile-de-France" regions, corresponding roughly to East France, North, and Paris and its suburbs, concentrated the highest number of cases in the initial months of 2020. The spatial study of the state of the epidemic before the introduction of vaccination campaigns is a crucial tool to provide information on the behaviour of future epidemics, as well as to suggest in which sites more resources should be invested to prepare facilities and personnel capacities to face this kind of health emergencies.

Amdaoud et al. (available at EconomiX Working Papers https://ideas.repec.org/p/drm/wpaper/2020-4.html) studied the spatial distribution of three indicators of the intensity of the COVID-19 epidemic. These indicators are the hospitalization rate, the mortality rate, and the excess of mortality rate; the data are collected from 19 March 2020 to 12 May 2020 and aggregated at the scale of the department. Even though we studied the three indicators, we will focus here on the hospitalization rate; the results for the two others are analogous. The exploratory analysis revealed an heterogeneous spatial distribution as can be seen in Figure 2, and the authors proposed to use spatial autoregressive models involving some socio-economic covariates to explain the spatial behaviour of the epidemic. We extend their study, including more updated data and more importantly, a larger number of socio-economic variables; actually, Amdaoud et al. removed some of the covariates in order to avoid multi-collinearity problems.

For better linearisation fitting, we consider the logarithm of some variables; thus, the response variable, denoted by `LnHosp`, is the logarithm of the hospitalization rate, attributed to the covid. We consider 8 covariates; some are related to the general population characteristics: population density (`LnPop`), proportion of people aged 65

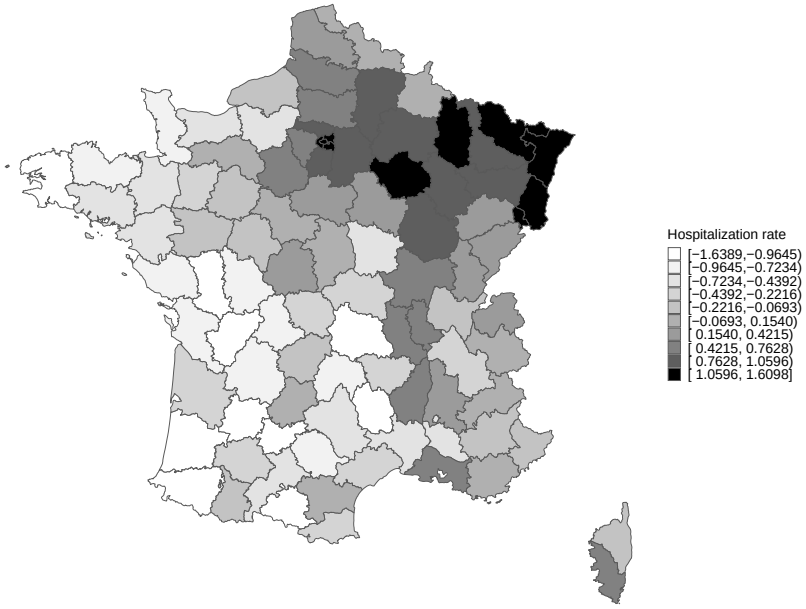


Fig. 2: Hospitalization rate by French department (by decile)

and over (`A65pls`), share of the three main municipalities in the population (`C3`). Three variables are linked to social aspects: the proportions of workers (`Work`), inactive people (`Inac`) and the disparities of income (`SLiveR`); finally two variables reflect the professional health care facilities, the rate of doctors (`FDoc`) and the rate of emergency services (`Emer`). The exact definition of the variables are listed in table A1 in the Appendix. Let us note that these covariates are considered as explanatory factors of the health status of a population and its mortality rate (Link and Phelan, 1995).

6.2 Exploratory analysis

According to the map of french departments, we choose the classic Queen neighbourhood, displayed in Figure 3, and in a standard way we choose the weights of the associated matrix W to be equal to the inverse distance between adjacent departments and 0 elsewhere. However, contrary to the usual row-standardization, we normalize W by its spectral radius. The Moran's one-sided randomization tests (see (Moran, 1950)) confirms the presence of a strong positive spatial autocorrelation for the hospitalization rate, for neighbours up to the second order, see Table 5.

Table 5: Moran's one sided randomization tests with alternative "greater"

Variable	Neigh.Order	Moran I	Mean	SD	Statistic	p-value
`LnHosp`	1	0.988	-0.0105	0.0069	12.13	3.65e-34
	2	0.576	-0.0107	0.0023	11.98	2.20e-33

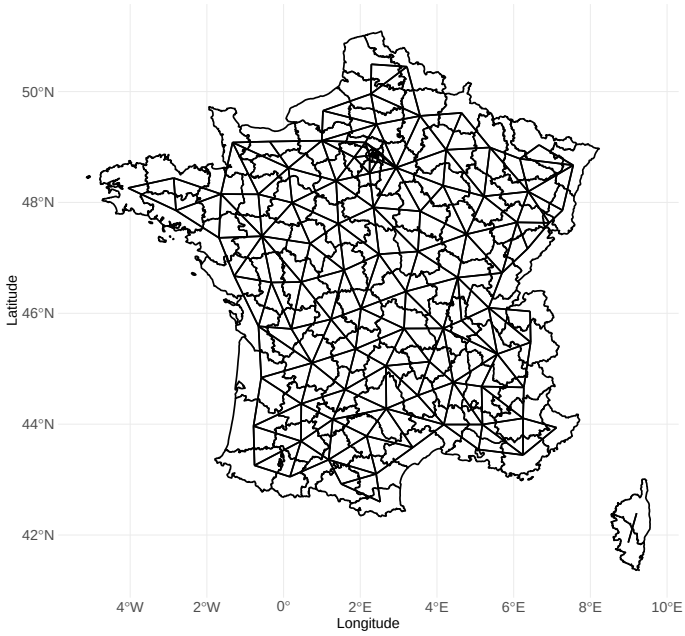


Fig. 3: Queen neighbourhood graph for French departments

Multi-collinearity

We compute the correlation matrix between the variables; the correlation matrix plot displayed in Figure 4 shows that the pairwise correlation is high for several pairs of variables, with the highest values obtained for the pairs `LnPop - SLivR`, `LnPop - A65pls`, `Work - SLivR` and `LnPop - Emer`.

Furthermore, some multi-collinearity tests and individual diagnostic tests were performed to detect problematic variables. When performing the overall multi-collinearity diagnostic tests, a small value of the determinant of the correlation matrix ($|\mathbf{X}^T\mathbf{X}| = 0.0199$) was found, accompanied by a high value of the Farrar's statistic (358.3260). The individual diagnostic tests for multi-collinearity point to the variables `LnPop`, `SLivR`, `Work`, `Inac` and `A65pls` as probable sources of multi-collinearity.

Principal Component Analysis on the covariates

Finally, we conducted a principal component analysis on the explanatory variables considering the hospitalization rate as a supplementary variable.

As shown in Figure 5, the first two principal components accumulated 65.1% of the variance present in the data set. The first component is positively correlated with `LnPop, SLivR, Inac,` and negatively correlated with `A65pls, Emer` and `Work`.

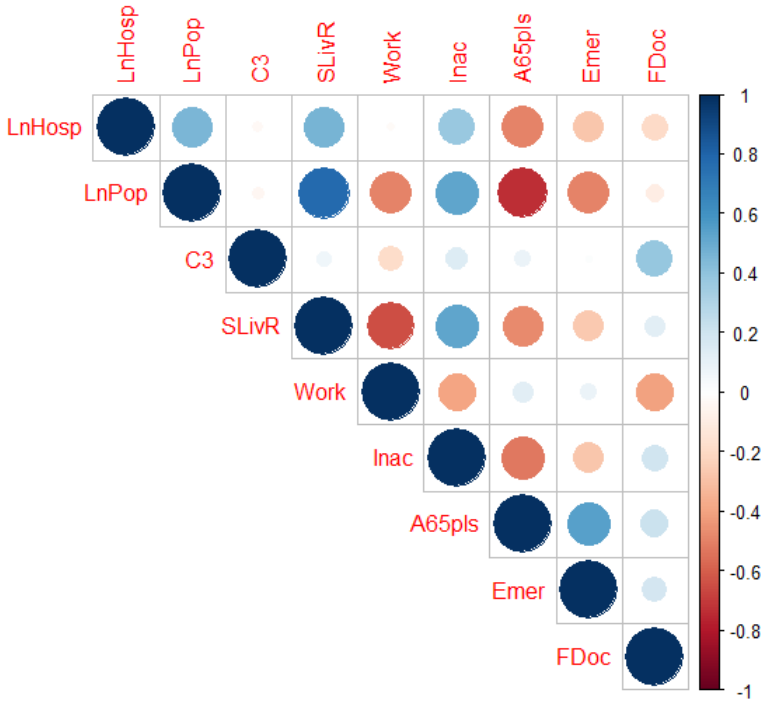


Fig. 4: Correlation matrix

The location of the epidemic indicator `LnHosp` suggests that excess hospitalization rate is mostly characterized by the variables featuring the first component; its high values correspond to densely populated departments, with large economic inequalities, large portions of inactive population and low proportion of population older than 65 years; the opposite corresponds to its low values. This is related to a urban / rural segmentation of the departments. Indeed, the representation of individuals, see Figure 6, clearly illustrates this feature, with mostly urban departments on the right side and rural ones on the left side.

Fig. 5: PCA biplot

The second component is mainly dominated by the variable `FDoc` with a notable contribution from the variable `C3`. Interestingly, this component is negatively correlated with the hospitalization rate.

In addition, PCA provides evidence of spatial correlation in the data because many departments located in the same region are close to each other in the representation of individuals; this latter plot is displayed in Figure B1 in the Appendix because it is difficult to interpret for non-experts in the geography of France.

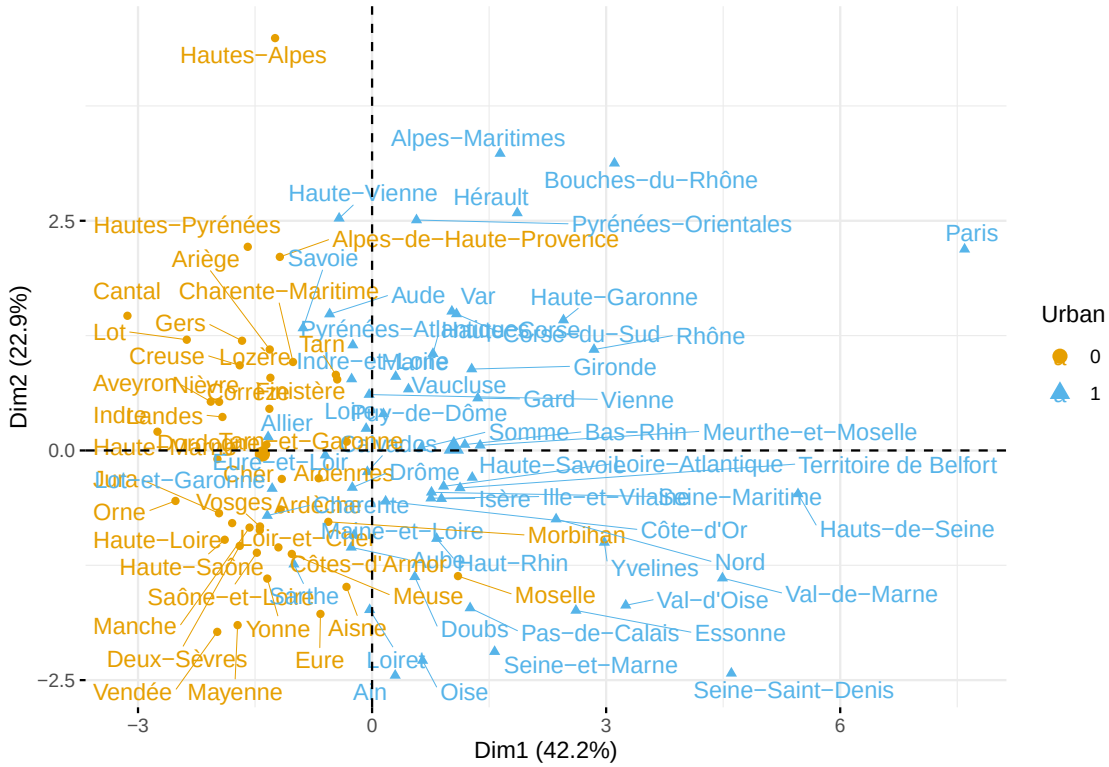
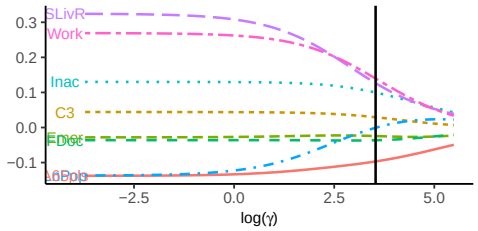


Fig. 6: Plot of departments in the space of the first two principal components, labelled according to the variable `Urban`

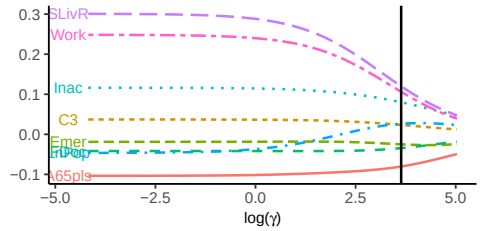
6.3 Results

Let us recall that the goal of the study is to determine which of the explanatory variables play a significant role in the spread of the Covid-19, represented by the hospitalization rate (when hospitalization is attributed to the disease). In all what follows, the explanatory variables are centred and scaled, and `LnHosp` is centred. SAR and SEM models are estimated under their ordinary procedure (without regularization) and following our algorithms RRSAR and RRSEM derived in Section 3.2. Let us recall that in each Ridge estimation procedure, the regularization parameter γ is determined by SLOO; to this aim, the dead zone is defined for each department by its first neighbours according to the Queen neighbourhood graph.

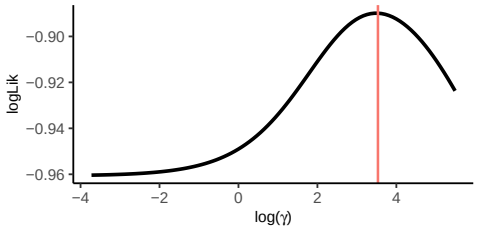
Figure 7 displays the evolution of the coefficients values (7a and 7b) as well as the log-likelihood (7c and 7d) in function of $\log(\gamma)$. The coloured lines are paths of regression coefficients, and the vertical line gives the set of coefficients corresponding to the value of γ selected via SLOO. We observe as expected the classic behaviour of the coefficients crushing down to zero as the regularization parameter γ increases, but



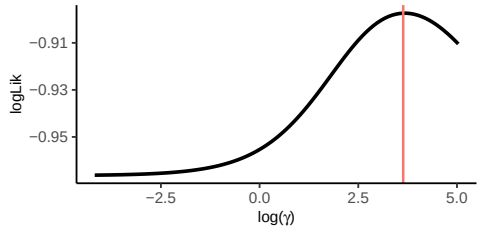
(a) SAR β_{ridge} coefficients paths as a function of $\log(\gamma)$



(b) SEM β_{ridge} coefficients paths as a function of $\log(\gamma)$



(c) SAR Ridge log-likelihood as a function of $\log(\gamma)$



(d) SEM Ridge log-likelihood as a function of $\log(\gamma)$

Fig. 7: Coefficients paths and log-likelihood as a function of the logarithm of the regularization parameter. Optimal regularization parameter for Ridge SAR and Ridge SEM is represented by a vertical line.

the behaviour is different between the variables. While the coefficients of `Inac, C3, Emer` are very stable, those of `SLivR, Work, A65pls` head towards zero; the coefficient of `LnPop` is very particular as it changes its sign in the SEM framework. We also note that the selected γ values according to one or the other model are similar; indeed, the values are about 34.26 for the SAR model and 38.02 for the SEM (see Table 7). Finally, we emphasize that the maximum log-likelihood coincides with the selected γ.

Once the model is estimated, we want to assess the importance of the explanatory variables. When the estimation has been performed under ordinary SAR and SEM modelling, we report the p-values corresponding to a two-sided z-test. For the Ridge versions, we conduct different tests: the t-test for Ridge estimators (Halawa and El Bassiouni, 2000), the permutation t-test, and the permutation F-test, as described in Section 4. Here, we chose to run 100 permutations for each of the 8 variables in the permutation versions. We present the results in Table 6 and Table 7. Note that the values of the regression coefficients in this table have been multiplied each one by the sample standard deviation of the corresponding covariate (but the tests statistics are computed with the original coefficient).

Results are quite coherent. Before analysing them, we point out that here of course we do not know the true model. The absolute value of the regression coefficients are without interest while we are mostly interested in their p-value.

We see how collinearity affects the results for the ordinary SAR and SEM models; for instance in the SAR and SEM estimation, the coefficients of `LnPop` is negative, though this variable is positively correlated with `LnHosp`. The most correlated variables with `LnHosp` are `LnPop, SLivR, Inac` and `A65pls,` but `LnPop` and `A65pls` appear to be non-significant; `Inac` is also non significant in the SEM model.

The regularized estimation of those models suppress most of these hassles.

	SAR		RRSAR			
				R. Perm. F-test	R. t-test	R. Perm. t-test
	Coef	p-value	Coef	p-value	p-value	p-value
`LnPop`	-0.1256	0.2297	-0.0006	0.8900	0.4910	0.4000
`C3`	0.5337	0.4365	0.3333	0.6500	0.5157	0.2700
`SLivR`	0.6277	0.0029	0.2593	0.0000	0.0031	0.0000
`Work`	12.8074	0.0019	6.7131	0.0000	0.0027	0.0100
`Inac`	6.7543	0.0742	5.1264	0.0800	0.0184	0.0300
`A65pls`	-0.0357	0.1713	-0.0257	0.0800	0.0108	0.9900
`Emer`	-6.6582	0.6206	-4.8059	0.7300	0.2132	0.7300
`FDoc`	-0.1584	0.7012	-0.2118	0.9700	0.4542	0.7200
γ			34.2557			
ρ	0.8489	0.0001	0.7282	0.0003		

Table 6: Coefficients and p-values of variable importance tests for hospitalization rate models. For SAR, we report the p-values corresponding to a two-sided z-test. For RRSAR, importance is determined via tests as described in Section 4 with 100 permutations for each of the 8 variables in the case of the permutation tests.

	SEM		RRSEM			
				R. Perm. F-test	R. t-test	R. Perm. t-test
	Coef	p-value	Coef	p-value	p-value	p-value
`LnPop`	-0.0363	0.7775	0.0211	0.5300	0.2019	0.2400
`C3`	0.4144	0.5608	0.2672	0.7500	0.5964	0.2900
`SLivR`	0.6293	0.0036	0.2455	0.0000	0.0065	0.0000
`Work`	12.0400	0.0101	5.1042	0.0100	0.0218	0.0100
`Inac`	6.0018	0.1553	4.1157	0.2100	0.0586	0.0300
`A65pls`	-0.0284	0.3336	-0.0213	0.1300	0.0393	0.9800
`Emer`	-3.4274	0.8062	-4.9290	0.7300	0.2242	0.7100
`FDoc`	-0.2583	0.5795	-0.2010	0.6300	0.5110	0.7800
γ			38.0189			
λ	0.8549	0.0047	0.8996	0.0000		

Table 7: Coefficients and p-values of variable importance tests for hospitalization rate models. For SEM, we report the p-values corresponding to a two-sided z-test. For RRSEM, importance is determined via tests as described in Section 4 with 100 permutations for each of the 8 variables in the case of the permutation tests.

First, the coefficient of `LnPop` becomes positive in RRSEM, and is very close to zero in RRSAR (we have seen that it would become positive with a larger γ in Figure 7a)

which is an improvement. We note that the variables `SlivR, Work` that were significant stay significant. On the other side and very suprinsingly, we see that variables `Emer, FDoc` do not play any role as explanatory variables. Then, very interestingly, the three tests do not select the same variables to be important. Globally, the permutation F-test seems to be more conservative than the t-tests. The variable `Inac` which was barely significant in the SAR model (p-value equal to 0.0742) keeps the same level of significance for the F-test but becomes important according to the two t-tests; it also becomes significant in the SEM framework according to the same t-tests. One other interesting variable is `A65pls`; it is not significant in the SAR and SEM models after the ordinary estimation procedure; but it becomes significant after the Ridge regularization following the t-test in both models, and barely significant (at level 10%) for the F-test in the SAR framework (note that its p-value has jumped from 33% down to 13% in the SEM model). The result of the permutation t-test for this variable is surprising, giving a very large p-value of about 98% for both models. To summarize, it is clear that variables `Work` and `SLivR` are very significant to explain the hospitalization rate due to Covid-19, which reflects the importance of inequalities and low income, before all. The relative selection of `Inac` supports this idea. Then, the proportion of elder people is also determinant, illustrated by the p-values of `A65pls`.

7 Conclusions

Multi-collinearity is a common feature of real life data; it also affects the class of simultaneous spatial autoregressive models in spatial econometrics.

We propose estimation algorithms that take into account multi-collinearity in the estimation of all the parameters using a regularization technique of type Ridge; then, the regularization parameter is obtained via a spatial ad-hoc cross-validation procedure; SLOO is particularly well adapted to spatial autoregressive models since there's no issue defining the size of the buffer surrounding the validation set, it is naturally determined by the first order neighbours. A drawback of SLOO is that it is computationally expensive in the case of a large number of observations. Some users propose to choose a large proportion of observations to serve as validation sets in the SLOO, instead of all the observations. These points can be chosen either randomly or regularly spaced. After Ridge regularization, the question of importance of the covariates rises. We considered three different tests adapted to ridge regression; ran on the application they provided coherent and different results. The question of a test adapted to both regularization and spatial dependence, including also the spatial dependence of the covariates themselves, is still an open problem.

A R package containing the whole method is a work in progress, but our code is already available on our Github project.

Our methodology can be extended in several directions. First, for variable selection purpose, the procedure can be easily adapted to Lasso regression and Elastic net. Next, it can also be further developed to consider other models like the general spatial model (which roughly speaking integrates both SAR and SEM in one model) or the Durbin model which extends the SAR model to include spatially lagged explanatory variables.

8 Acknowledgments

The authors thank Nadine Levratto, Mounir Amdaoud and Giuseppe Arcuri for providing part of the data used in the Application section.

This research has been conducted within the FP2M federation (CNRS FR 2036), and as part of the project Labex MME-DII (ANR11-LBX-0023-01).

Declarations

- Funding : FP2M federation (CNRS FR 2036), and as part of the project Labex MME-DII (ANR11-LBX-0023-01).
- Competing interests : The authors declare no conflict of interest.
- Ethics approval : Not applicable.
- Consent to participate : Not applicable.
- Consent for publication : All authors consent on the publication of the manuscript.
- Availability of data and materials : Data available at the Github repository https://github.com/c0ra/RRSARMMI.
- Code availability : Code available at the Github repository https://github.com/c0ra/RRSARMMI
- Authors' contributions : All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by all authors. All authors contributed to the writing and read and approved the final manuscript.

Appendix A Dictionary of variables

Appendix B PCA Individuals plot colored by region

Table A1: Dictionary of variables

Wording	**Acronym**	**Definition**	**Date/Year**	**Source**
Response variables				
Logarithm of hospitalization rate	`LnHosp`	$\log((\frac{Number\ of\ Covid-19\ hospitalizations}{Population}) * 1000)$	19 March 12 May 2020	Public Health France
Explanatory covariates				
Logarithm of population density	`LnPop`	$\log(\frac{Number\ of\ inhabitants}{Km^2})$	2016	INSEE
Share of people aged 65 and over	`A65pls`	$\frac{Number\ of\ people\ aged\ 65\ and\ over}{total\ population}$	2020	INSEE
Percentage of inhabitants in the three largest cities of the department	`C3`	Share of the 3 main municipalities in the population of the department	2018	INSEE
Rate for workers	`Work`	Share of the population aged 15 years or more by CSP Worker (2016)	2016	INSEE
Inactive rate	`Inac`	Share of the population aged 15 years or more by CSP Other + persons not in employment	2016	INSEE
Standard of living gap	`SLivR`	Interdecile income ratio (9e decile/1er decile)	2016	INSEE
Rate of emergency services per 1000 habitants	`Emer`	$\log(Number\ of\ Emergency\ Department)$	2018	INSEE
Rate of doctors per 1000 habitants	`FDoc`	$\frac{Number\ of\ doctors}{total\ population}$	2016	INSEE

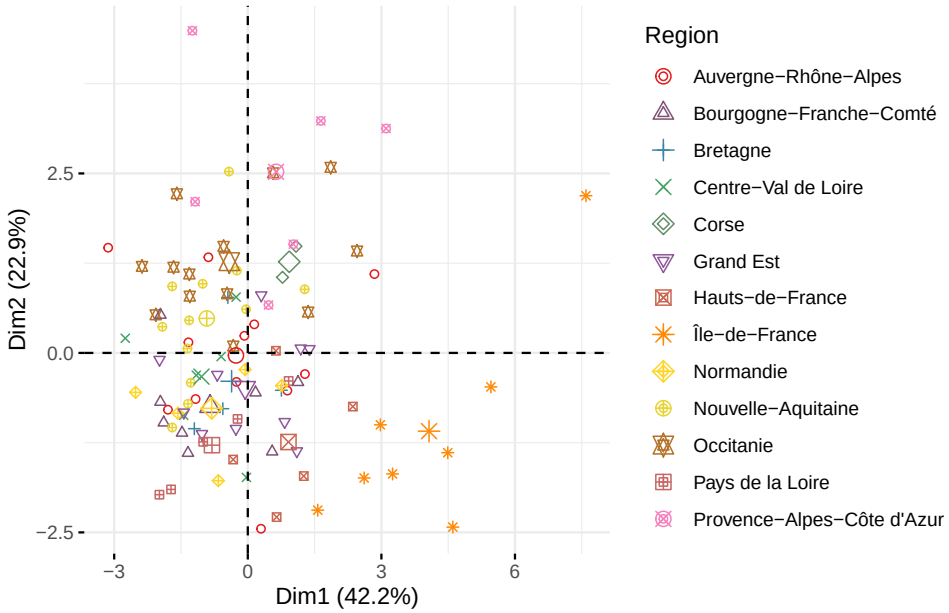


Fig. B1: Plot of departments in the space of the first two principal components, labelled according to their region

References

Anselin, L.: Spatial Econometrics: Methods and Models. Dorddrecht: Kluwer Academic Publishers, Dordrecht (1988)

LeSage, J.P.: An introduction to spatial econometrics. Revue d'économie industrielle **123**, 19–44 (2008) https://doi.org/10.4000/rei.3887

Alin, A.: Multicollinearity. WIREs Computational Statistics **2**, 370–374 (2010) https://doi.org/10.1002/wics.84

Hoerl, A., Kennard, R.: Ridge regression. In: Kotz, S., Johnson, N.L., Read, C.B. (eds.) Encyclopedia of Statistical Sciences vol. 42, pp. 129–136. Wiley, New York (1988)

Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society **67**, 301–320 (2005) https://doi.org/10.1111/j.1467-9868.2005.00503.x

Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) **58**(1), 267–288 (1996)

McDonald, G.C.: Ridge regression. WIREs Computational Statistics **1**, 93–100 (2009)

https://doi.org/10.1002/wics.14

Wheeler, D.C.: Simultaneous coefficient penalization and model selection in geographically weighted regression: The geographically weighted lasso. Environment and Planning A: Economy and Space **41**, 722–742 (2009) https://doi.org/10.1068/a40256

Fan, C., Rey, S.J., Myint, S.W.: Spatially filtered ridge regression (sfrr): A regression framework to understanding impacts of land cover patterns on urban climate. Transactions in GIS **21**, 862–879 (2017) https://doi.org/10.1111/tgis.12240

Tikhonov, A.N., Arsenin, V.I.: Solutions of Ill-posed Problems. Wiley, New York (1977)

Hoerl, A.E., Kennard, R.W.: Ridge regression: Biased estimation for nonorthogonal problems. Technometrics **12**(1), 55–67 (1970)

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., *et al.*: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography **40**, 913–929 (2017) https://doi.org/10.1111/ecog.02881

Halawa, A.M., El Bassiouni, M.Y.: Tests of regression coefficients under ridge regression models. Journal of Statistical Computation and Simulation **65**(1-4), 341–356 (2000) https://doi.org/10.1080/00949650008812006

Le Rest, K., Pinaud, D., Bretagnolle, V.: Accounting for spatial autocorrelation from model selection to statistical inference: application to a national survey of a diurnal raptor. Ecological informatics **14**, 17–24 (2013) https://doi.org/10.1016/j.ecoinf.2012.11.008

Pohjankukka, J., Pahikkala, T., Nevalainen, P., Heikkonen, J.: Estimating the prediction performance of spatial models via spatial k-fold cross validation. International Journal of Geographical Information Science **31**, 2001–2019 (2017) https://doi.org/10.1080/13658816.2017.1346255

Brenning, A.: Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. Paper presented at the 2012 IEEE international geoscience and remote sensing symposium, IEEE, Munich, 5372–5375 July 2012 (2012). https://doi.org/10.1109/IGARSS.2012.6352393

Moran, P.A.P.: A test for the serial independence of residuals. Biometrika **37**, 178–181 (1950) https://doi.org/https://www.jstor.org/stable/2332162

Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. Journal of statistical software **33**(1), 1–22 (2010)

Perez-Melo, S., Kibria, B.M.G.: On some test statistics for testing the regression coefficients in presence of multicollinearity: A simulation study. Stats **3**(1), 40–55 (2020) https://doi.org/10.3390/stats3010005

Manly, B.F.J.: Randomization, Bootstrap and Monte Carlo Methods in Biology vol. 70. CRC press, Boca Raton (2006)

Kennedy, F.E.: Randomization tests in econometrics. Journal of Business & Economic Statistics **13**, 85–94 (1995) https://doi.org/10.1080/07350015.1995.10524581

Anderson, M.J., Robinson, J.: Permutation tests for linear models. Australian & New Zealand Journal of Statistics **43**, 75–88 (2001) https://doi.org/10.1111/1467-842X.00156

Hastie, T., Tibshirani, R.: Generalized additive models for medical research. Statistical Methods in Medical Research **4**, 187–196 (1995) https://doi.org/10.1177/096228029500400302

Bécu, J.M., Grandvalet, Y., Ambroise, C., Dalmasso, C.: Beyond support in two-stage variable selection. Statistics and Computing **27**, 169–179 (2017) https://doi.org/10.1007/s11222-015-9614-1

Link, B.G., Phelan, J.: Social conditions as fundamental causes of disease. Journal of Health and Social Behavior, 80–94 (1995) https://doi.org/10.2307/2626958