Université Paris Nanterre

200 av. de la République
92001 Nanterre Cedex
www.parisnanterre.fr

ICIMAF

Membre de l'université Paris Lumières
École doctorale 139 : Connaissance, langage, modélisation
MODAL'X UMR 9023

Cristina Olimpia CHÁVEZ CHONG

Méthodes de validation croisée pour des données spatialement dépendantes

Thèse présentée et soutenue publiquement le 11/04/2025

en vue de l'obtention du doctorat de Mathématiques appliquées et applications des mathématiques de l'Université Paris Nanterre

sous la direction de Mme Cécile HARDOUIN (Université Paris Nanterre),

de Mme. Ana Karina FERMIN RODRIGUEZ (Université Paris Nanterre)

et de M. Jesús Eladio SÁNCHEZ GARCÍA (ICIMAF)

Jury * **:**

Rapporteur·e :	Mme. Anne GÉGOUT-PETIT	Prof., Université de Lorraine, Institut Elie Cartan de Lorraine
Rapporteur·e :	M. Éric MATZNER-LØBER	Prof., Centre de Recherche en Economie et Statistique (CREST), ENSAE
Membre du jury :	Mme. Liliane BEL	Prof., AgroParisTech/INRAE UMR 518 Math. Info. Appli.
Membre du jury :	M. Thierry DUMONT	MCF, Université Paris Nanterre
Président du jury :	M. Jean-Marc BARDET	Prof., Université Paris 1 Panthéon Sorbonne SAMM (UR 4543)

UNIVERSITÉ PARIS NANTERRE

Méthodes de validation croisée pour données spatialement dépendantes

Cristina Olimpia Chávez Chong

Encadrantes : Cécile Hardouin, Ana Karina Fermín Rodríguez
Co-encadrant : Jesús E. Sánchez García

Soutenance publique : 11 avril 2025

Introduction

La généralisation des données géoréférencées, portée par le progrès technologique, le mouvement *open data* et des politiques incitatives, ouvre des perspectives sans précédent en sciences de l'environnement, écologie, épidémiologie ou aménagement urbain. Or, l'autocorrélation spatiale très marquée de ces jeux de données contredit l'hypothèse d'indépendance qui sous-tend la majeure partie de la théorie classique de l'apprentissage statistique. Ainsi, des outils réputés fiables en cadre indépendant (optimisation du risque, procédures de validation croisée, etc.) deviennent biaisés dans un contexte spatial.

La validation croisée (VC) illustre ce problème : en régression indépendante, la procédure *leave-one-out* est asymptotiquement non biaisée et de variance minimale. Dès que « les choses proches sont plus reliées que les choses lointaines » (1^{er} loi de Tobler (1970)), un simple découpage aléatoire transfère de l'information entre ensembles d'apprentissage et de validation, aboutissant à des estimations de risque excessivement optimistes.

Une littérature récente a donc proposé des variantes « spatialisées » de la VC — tampons, blocs réguliers, partitions par clusters, procédures hybrides, etc. Cette thèse se place dans un cadre de régression spatiale à design fixe et cherche à rendre l'estimation de risque fiable face à la dépendance spatiale.

Modèle général. On observe, pour $\mathbf{s} \in \mathcal{S}_n \subset \mathbb{R}^d$,

$$Y(\mathbf{s}) = \mathbf{X}(\mathbf{s})\boldsymbol{\beta} + \varepsilon(\mathbf{s}), \qquad \varepsilon(\mathbf{s}) \text{ auto-corrélée},$$

où $\mathbf{X}(\mathbf{s})$ est non aléatoire (*design fixe*). Tous les estimateurs de risque, schémas de VC et expériences de simulation sont formulés dans ce cadre.

Validation croisée pour données spatiales

1. Estimation du risque : cadre à design fixe

Nous travaillons dans le modèle linéaire à design fixe

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \overset{\text{i.i.d.}}{\sim} (0, \sigma^2),$$

où les lignes de $\mathbf{X}$ sont considérées non aléatoires. Pour un estimateur $\hat{\boldsymbol{\beta}}$, le *risque prédictif* est la perte quadratique moyenne obtenue sur un réplicat $\mathbf{Y}^{\text{new}}$ indépendant mais observé sur les mêmes sites :

$$\mathcal{R}_{\text{pred}}(\hat{\boldsymbol{\beta}}) \;=\; \mathbb{E}\Big[n^{-1}\big\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Y}^{\text{new}}\big\|^2\Big].$$

Ne disposant jamais de $\mathbf{Y}^{\text{new}}$, on emploie la validation croisée (VC) : on ré-estime $\hat{\boldsymbol{\beta}}$ sur un *train*, puis on évalue la perte sur un *test* indépendant (hold-out, K-fold, LOO). Si les deux sous-échantillons partagent des observations corrélées spatialement, l'hypothèse d'indépendance est violée et la VC classique devient systématiquement optimiste.

2. Versions spatiales des procédures de VC

Pour garantir l'indépendance on sépare géographiquement train et test à l'aide d'une *zone tampon* de rayon h ; on distingue alors :

— **SLOO** (Spatial Leave-One-Out). À l'itération m, on retire le point

$$\mathcal{S}_n^{\mathrm{v}_m} = \{\mathbf{s}_m\},$$

and

$$\mathcal{D}_n^{\mathrm{v}_m} = \{(\mathbf{X}(\mathbf{s}_m), Y(\mathbf{s}_m))\}$$

forms the validation set. The "dead zone" around $\{\mathbf{s}_m\}$ is defined as its neighbourhood of radius h,

$$\mathcal{S}_{n,h}^{\mathrm{D}_m} = \{\mathbf{s}_j \in \mathcal{S}_n : j \neq m \text{ and } d(\mathbf{s}_j, \mathbf{s}_m) \leq h)\}.$$

The remaining points constitute the training domain,

$$\mathcal{S}_{n,h}^{\mathrm{t}_m} = \{\mathbf{s}_j \in \mathcal{S}_n : d(\mathbf{s}_j, \mathbf{s}_m) > h\},$$

and the corresponding training set is

$$\mathcal{D}_n^{\mathrm{t}_m} = \{(\mathbf{X}(\mathbf{s}_j), Y(\mathbf{s}_j)) | \mathbf{s}_j \in \mathcal{S}_{n,h}^{\mathrm{t}_m}\}.$$

Très peu biaisé mais coûteux (1 modèle/site) et de variance élevée.

$$\mathcal{R}_{n,h}^{\mathrm{SLOO}} = \frac{1}{n}\sum_{m=1}^{n} \mathcal{R}_{n,h}^{\mathrm{v}_m} = \frac{1}{n}\sum_{m=1}^{n} (Y(\mathbf{s}_m) - \mathbf{X}(\mathbf{s}_m)\hat{\boldsymbol{\beta}}_h^{\mathrm{t}_m})^2, \tag{1}$$

— **SKCV** (Spatial K-fold). On tire d'abord K plis aléatoires, puis on élimine, autour de chaque pli, les points situés à moins de h. L'échantillon d'apprentissage fond rapidement : jusqu'à 98 % des sites supprimés pour $h = 20$ (Tab. **??**).

— **SBCV** (Spatial Blocked CV). Les plis sont des blocs réguliers entourés d'un tampon h. La zone morte est bien plus réduite que sous SKCV, donc la variance reste contenue. The risk estimate for SBCV is given by :

$$\mathcal{R}_{n,h}^{\mathrm{SBCV}} = \frac{1}{B}\sum_{m=1}^{B} \mathcal{R}_{n,h}^{\mathrm{v}_m}. \tag{2}$$

Here, each validation domain $\mathcal{S}_{n,h}^{V_m}$ is a block. Variantes : blocs irréguliers, k-means, « leave-one-disc-out ».

Le paramètre critique est h : trop petit ⇒ biais, trop grand ⇒ variance et parfois train vide.

3. Étude de simulation : principaux résultats

Cadre. Lattice 80×60 ($n = 4800$). Processus $Y(\mathbf{s}) = \mathbf{X}(\mathbf{s})\boldsymbol{\beta} + \varepsilon(\mathbf{s})$ avec covariance exponentielle $C(h) = \exp(-h/\theta)$, $\theta \in \{3, 6, 9, 12, 15, 20\}$ (dépendance croissante). Pour la l^{e} réalisation, on estime $\hat{\boldsymbol{\beta}}^{(l)}$ (ici par OLS) et on définit le *risque idéal empirique*

$$\mathcal{R}_n^{I(l)} = \frac{1}{M-1}\sum_{j \neq l} \frac{1}{n}\sum_{\mathbf{s} \in \mathcal{S}_n} \left(\hat{Y}^{(l)}(\mathbf{s}) - Y^{(j)}(\mathbf{s})\right)^2. \tag{3}$$

Pour chaque θ : $M = 100$ réplicats, estimation OLS et calcul de

— LOO, KCV ($K = 200$), BCV (200 blocs) ;

— SLOO, SKCV, SBCV pour $h \in \{5, 10, 15, 20\}$;

— le *risque idéal* $\mathcal{R}_n^{I(l)}$ (prédiction croisée entre réplicats indépendants).

(i) Biais de la VC classique. Dans tous les cas LOO, KCV et BCV sous-estiment le risque, de plus en plus lorsque θ augmente (Fig. 1).

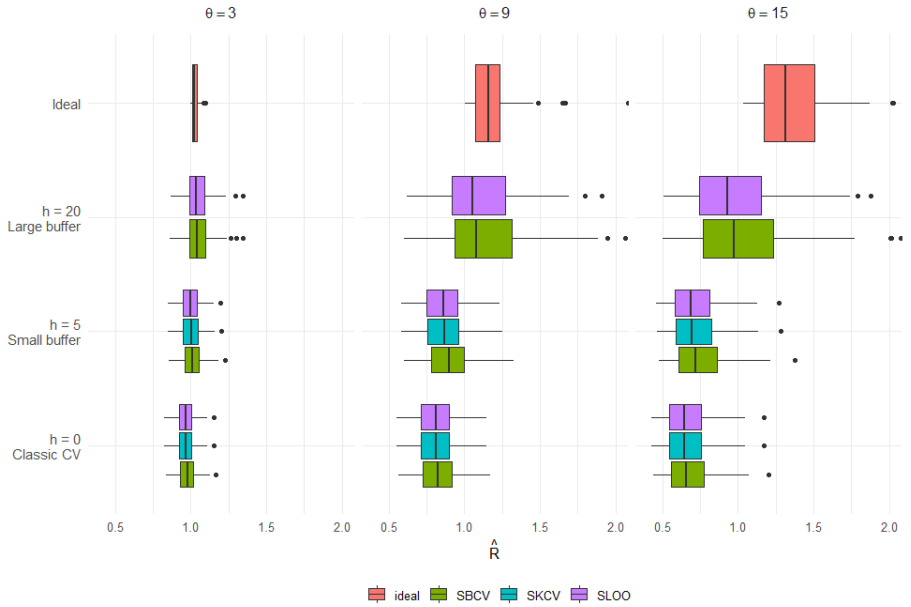


FIGURE 1 – Ideal risk and cross-validation risk estimates derived from 100 replications of the response variable Y across different cross-validation methods. For $h = 0$, SLOO, SKCV and SBCV become LOO, KCV and BCV, respectively.

(ii) SKCV impraticable. À $h = 10$ il ne conserve que 25 % des points, et le train peut être vide pour $h \geq 15$.

(iii) SLOO et SBCV les plus proches du risque idéal. Pour $\theta \leq 9$ ils atteignent l'idéal avec un tampon modeste ; au-delà, accroître h réduit le biais mais fait exploser la variance. SLOO est légèrement plus précis, SBCV 24 fois plus rapide.

(iv) Taille des blocs en SBCV. Des blocs trop grands (20×20) gonflent inutilement le risque et sa variance, en particulier lorsque la dépendance est faible.

Conclusion. SKCV est déconseillé à cause de la perte massive d'observations. SLOO et SBCV fournissent les estimations les moins biaisées ; leur qualité dépend presque exclusivement du bon choix du rayon tampon h (et, pour SBCV, du gabarit des blocs), problématique qui sera traitée au chapitre suivant. Les schémas SLOO et SBCV, moyennant un *tampon* adéquat, sont les seuls à fournir un risque quasi non biaisé et donc une sélection de modèles fiable.

Résumé du Chapitre 2 : Détermination de la taille optimale du buffer

1. Risque idéal et écart de validation croisée

Dans le cadre spatial à design fixe, on dispose de M réalisations indépendantes $\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(M)}$ sur le même domaine $\mathcal{S}_n$. Pour la l^{e} réalisation, on estime $\hat{\boldsymbol{\beta}}^{(l)}$ (ici par OLS) et on définit le *risque idéal empirique*

$$\mathcal{R}_n^{I(l)} = \frac{1}{M-1} \sum_{j \neq l} \frac{1}{n} \sum_{\mathbf{s} \in \mathcal{S}_n} \left(\hat{Y}^{(l)}(\mathbf{s}) - Y^{(j)}(\mathbf{s}) \right)^2,$$

qui mesure la perte moyenne lorsque le modèle entraîné sur $\mathbf{Y}^{(l)}$ prédit toutes les autres réalisations.

Pour une procédure de validation croisée spatiale CV (SLOO ou SBCV) utilisant un buffer de rayon h, le risque estimé sur la l^{e} réalisation s'écrit $\mathcal{R}_{n,h}^{(l)}$ (équations (1) ou (2)). On évalue alors, pour une grille $\mathcal{H} = \{1, \ldots, H_{\max}\}$, l'écart quadratique moyen

$$\hat{\mathcal{L}}(h) = \frac{1}{M} \sum_{l=1}^{M} \left(\mathcal{R}_{n,h}^{(l)} - \mathcal{R}_n^{I(l)} \right)^2,$$

et on appelle **buffer optimal empirique**

$$h_I \;=\; {}_{h \in \mathcal{H}} \hat{\mathcal{L}}(h).$$

Décomposition de $\hat{\mathcal{L}}(h)$. En posant $\overline{\mathcal{R}_n^I} = \frac{1}{M} \sum_l \mathcal{R}_n^{I(l)}$ et $\overline{\mathcal{R}_{n,h}} = \frac{1}{M} \sum_l \mathcal{R}_{n,h}^{(l)}$, on obtient la décomposition :

$$\hat{\mathcal{L}}(h) = T_1 + T_2 + T_3 + T_4 + T_5,$$

où T_1 est la variance empirique du risque idéal, T_2 la variance du risque CV, $T_3 = (\overline{\mathcal{R}_n^I} - \overline{\mathcal{R}_{n,h}})^2$ (le biais ^{2}), et T_4, T_5 des termes de covariance (faibles). Les simulations montrent que T_1, T_4, T_5 sont quasi constants, alors que

$$\boxed{T_2 \uparrow (h) \quad \text{et} \quad T_3 \downarrow (h)}$$

de sorte que leur somme réalise un compromis biais–variance, avec un minimum net en $h = h_I$ (Fig.2).

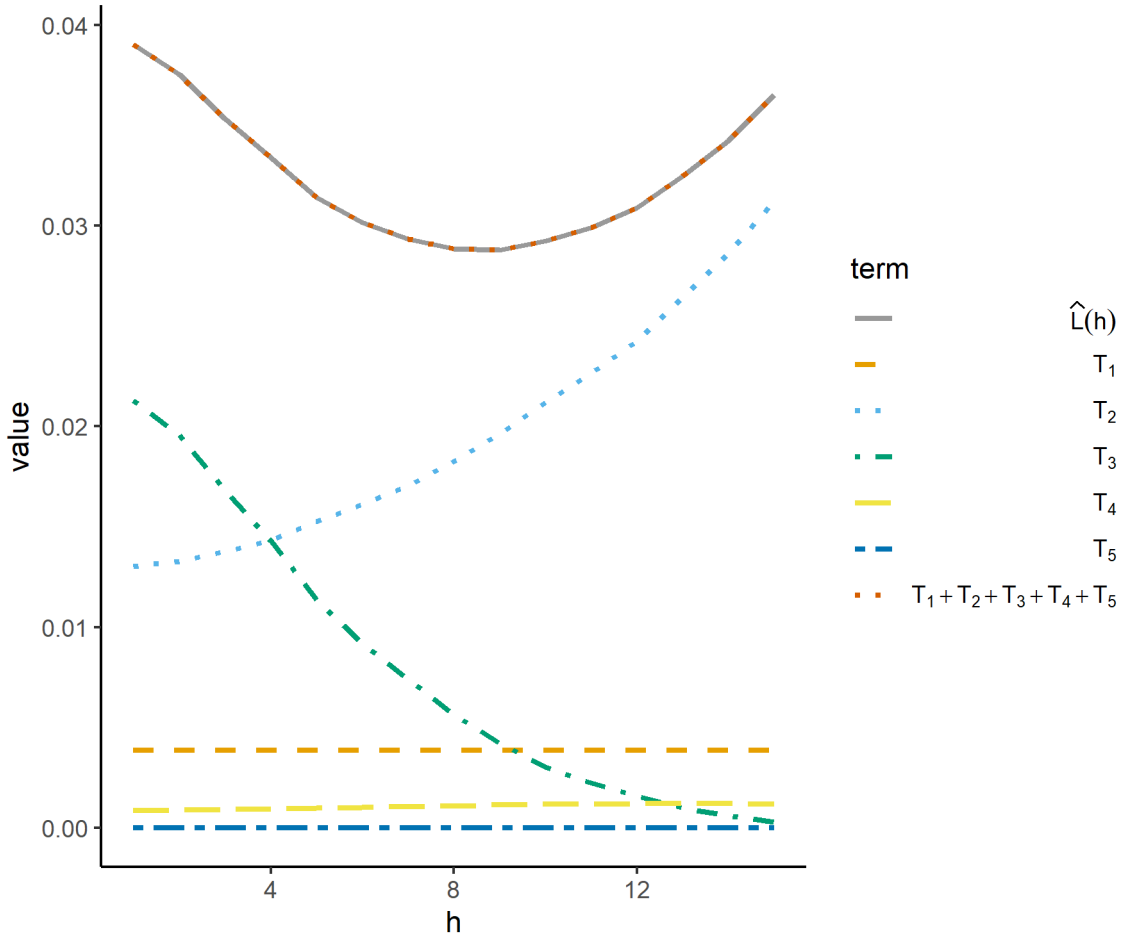


FIGURE 2 – Decomposition of $\hat{\mathcal{L}}(h)$ into the five components T_1–T_5 for $\theta = 5$.

2. Stratégies « naïves » pour choisir h

(i) *Portée du variogramme* (Le Rest *et al.*, 2014) : on ajuste un variogramme aux résidus $\hat{\varepsilon}(\mathbf{s})$ et l'on prend h_{range} = portée (distance du palier). Méthode sensible au modèle retenu.

(ii) *Indice de Moran* (Karasiak *et al.*, 2022) : on calcule $I_M(h)$ par classes de distance et l'on choisit

$$h_{\text{Moran}} = \min\{h : \ I_M(h) \text{ n'est plus significatif}\}, \quad \text{ou} \quad h_{0.05} \text{ tel que } I_M(h) = 0.05.$$

Méthode non-paramétrique mais coûteuse.

(iii) *SAC-0.1* (proposée ici) : on calcule l'autocorrélation empirique $\hat{\rho}(h)$ des résidus (direction horizontale ou verticale) et l'on retient

$$h_{0.1} =_{h \in \mathcal{H}} |\hat{\rho}(h) - 0.1|.$$

Aucune hypothèse de forme, calcul rapide.

3. Étude de simulation

Plan d'expérience. $\mathcal{S}_n = 80 \times 60$, $n = 4800$. Deux tendances : *constante* ($Y = \varepsilon$) ou *non stationnaire* $Y(\mathbf{s}) = \mathbf{X}(\mathbf{s})\boldsymbol{\beta} + \varepsilon(\mathbf{s})$. Bruit gaussien spatial exponentiel $C(h) = \exp(-h/\theta)$ avec $\theta \in \{0.5, 1, 2, \ldots, 20\}$. Pour chaque ($\theta$, scénario) : $M = 100$ réplications, estimation OLS, risques idéaux et CV (SLOO / SBCV, blocs 4×6 ou 8×12), puis évaluation de h_I, $h_{0.1}$, h_{range}, h_{Moran}, $h_{0.05}$.

Comportement de $\hat{\mathcal{L}}(h)$. Quelle que soit la méthode, $\hat{\mathcal{L}}(h)$ présente un minimum net. Pour SLOO, h_I croît de 1 (faible dépendance) à ≈ 25 (très forte dépendance) ; pour SBCV il est un peu plus petit, l'effet *blocs* cassant déjà la dépendance.

Performance des règles naïves. Tables `tab1`–`tab2` (citées dans le texte) et figure 3 montrent :

— **SAC-0.1**. – Dans 90% des cas $\hat{\mathcal{L}}(h_{0.1})$ est à moins de 10% de $\hat{\mathcal{L}}(h_I)$, et souvent $<3\%$. Lorsque θ devient très grand (≥ 16) la règle tend à sous-estimer h_I, mais la perte de risque reste modérée.

— **Variogramme**. – Bonne précision tant que le modèle ajusté est correct et la dépendance modérée. Dès que θ est fort ou que le variogramme est mal spécifié, h_{range} est instable et peut conduire à des risques nettement supérieurs.

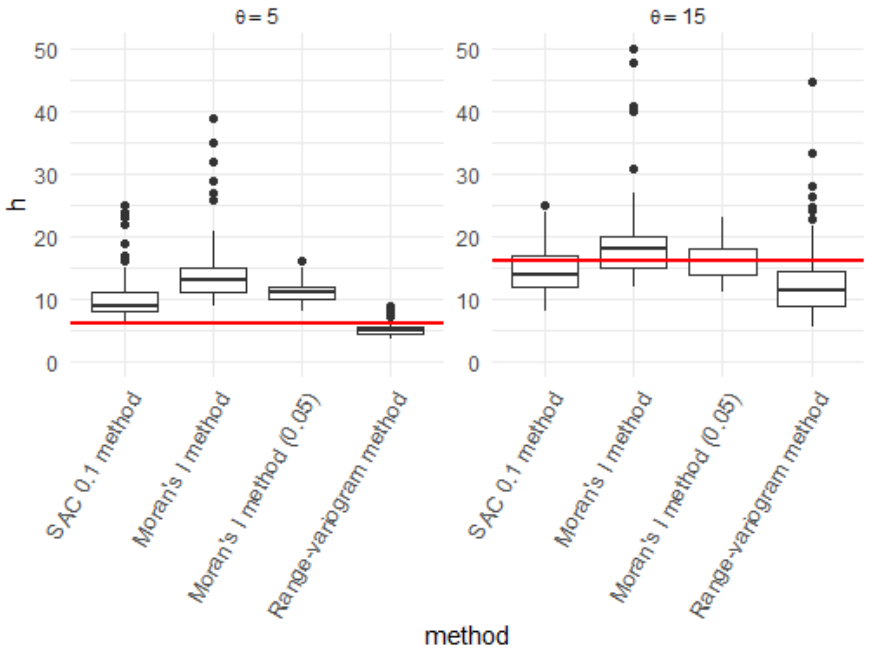


FIGURE 3 – Boxplots of buffer radius selected using the different estimation methods. The horizontal red line represents h_I

— **Moran**. – Méthode prudente : h_{Moran} (ou $h_{0.05}$) est souvent plus grand que h_I, surtout quand les blocs sont larges (SBCV 8×12) ; la validation devient alors très pessimiste. Pour dépendance moyenne ($\theta \approx 7$–10) les performances peuvent toutefois dépasser SAC-0.1.

Effet de la taille des blocs (SBCV). Avec des blocs 4×6, SAC-0.1 reste proche de l'optimum. Avec des blocs 8×12, la règle SAC-0.1 reste néanmoins régulière.

4. Conclusions opérationnelles

1. Un **buffer optimal existe** et équilibre biais (T_3) et variance (T_2) du risque CV ; il dépend de la force de la dépendance spatiale.
2. **Heuristique pratique** : choisir le rayon h pour lequel l'autocorrélation résiduelle atteint 0.1 fournit dans la plupart des cas un risque CV très proche du risque idéal, sans ajustement paramétrique ni test lourd.
3. La règle *SAC-0.1* est robuste aux formes de tendance, aux tailles de blocs SBCV et à un large éventail de dépendances, là où les approches variogramme ou Moran peuvent être trop sensibles (modèle mal spécifié, buffer sur-dimensionné, ...).

4. Pour des dépendances *extrêmes*, $h_{0.1}$ peut légèrement sous-estimer h_I ; augmenter le seuil (e.g. 0.05) ou pénaliser la perte de données sont alors des pistes, discutées en perspectives.

La prochaine étape (Chapitre 3) vise à *approcher directement* le risque idéal via un bootstrap spatial paramétrique, afin d'estimer h_I sans disposer de multiples réalisations du champ spatial.

Résumé du Chapitre 3 : Méthodes affinées de détermination de la taille optimale du buffer

1. Objectif

Soit une unique réalisation d'un champ spatial $\mathbf{Y}$. On veut choisir le rayon h qui rapproche au mieux le *risque CV spatial* du *risque idéal*. Dans le cadre bootstrap ($b = 1, \ldots, B$) :

$$\mathcal{R}_{n,h}^{\star(b)} = \frac{1}{n} \sum_{m=1}^{n} \left(Y^{\star(b)}(\mathbf{s}_m) - \mathbf{X}(\mathbf{s}_m) \hat{\boldsymbol{\beta}}_h^{t_m(b)} \right)^2, \tag{4}$$

$$\mathcal{R}_{n,h}^{\star(b)} = \frac{1}{K} \sum_{m=1}^{K} \frac{1}{|\mathrm{v}_m|} \sum_{\mathbf{s} \in \mathrm{v}_m} \left(Y^{\star(b)}(\mathbf{s}) - \mathbf{X}(\mathbf{s}) \hat{\boldsymbol{\beta}}_h^{t_m(b)} \right)^2 \quad \text{(version SBCV)}, \tag{5}$$

$$\mathcal{R}_{n}^{I\star(b)} = \frac{1}{B-1} \sum_{\substack{j=1 \\ j \neq b}}^{B} \frac{1}{n} \sum_{\mathbf{s} \in \mathcal{S}_n} \left(\hat{Y}^{\star(b)}(\mathbf{s}) - Y^{\star(j)}(\mathbf{s}) \right)^2. \tag{6}$$

Le **critère de discrépance** vaut

$$\hat{\mathcal{L}}^{\star}(h) \;=\; \frac{1}{B} \sum_{b=1}^{B} \left(\mathcal{R}_{n,h}^{\star(b)} - \mathcal{R}_{n}^{I\star(b)} \right)^2, \tag{7}$$

et l'on retient $h_I^{\star} = \arg\min_h \hat{\mathcal{L}}^{\star}(h)$.

2. Accélération « Sample SLOO » (SSLOO)

Plutôt que de laisser un point à la fois sur les n sites, on ne considère qu'un sous-échantillon aléatoire $\mathcal{S}_{np} \subset \mathcal{S}_n$ ($p \in (0, 1]$). Le risque empirique devient

$$\mathcal{R}_{np,h}^{(b)} = \frac{1}{np} \sum_{\mathbf{s}_m \in \mathcal{S}_{np}} \left(Y^{\star(b)}(\mathbf{s}_m) - \mathbf{X}(\mathbf{s}_m) \hat{\boldsymbol{\beta}}_h^{t_m(b)} \right)^2, \tag{8}$$

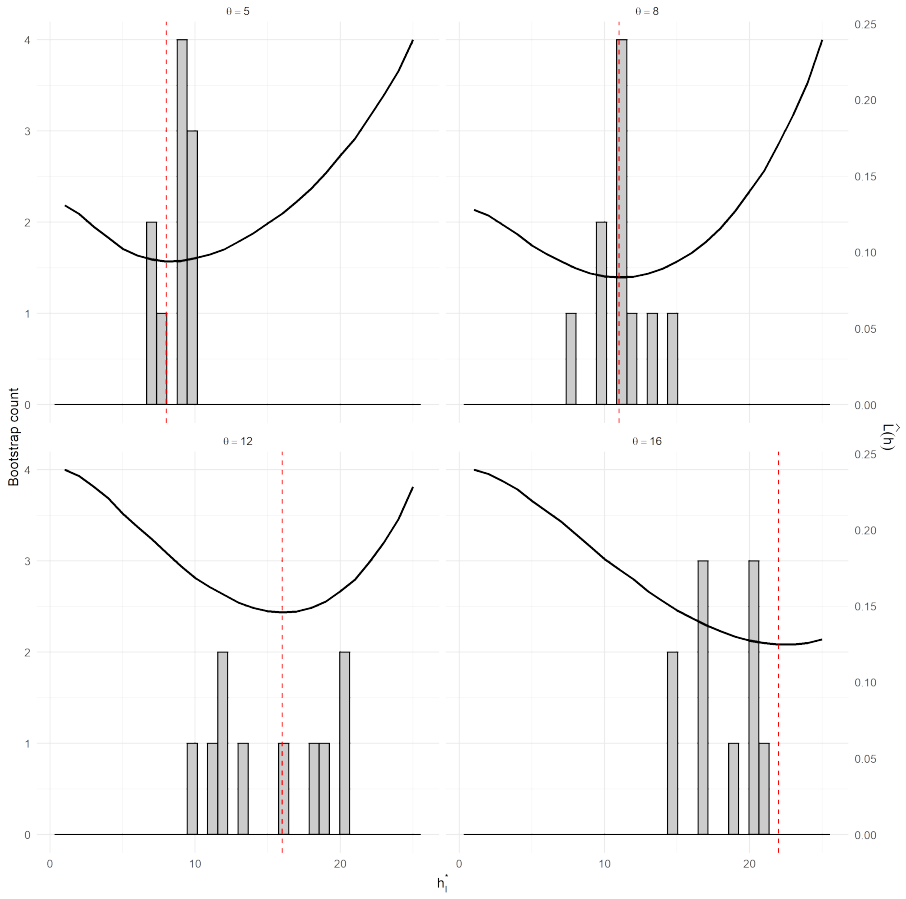


FIGURE 4 – Bootstrap distribution of $h^{\star}_{I,p}$ under the non-stationary trend scenario.

et l'on minimise

$$\hat{\mathcal{L}}^{\star}(h,p) = \frac{1}{B}\sum_{b=1}^{B}\Big(\mathcal{R}^{(b)}_{np,h} - \mathcal{R}^{I\star(b)}_{n}\Big)^2.$$

3. Heuristique SAC-0.1 et intervalle réduit

On calcule l'autocorrélation empirique des résidus $\hat{\rho}(h)$ et on fixe

$$h_{0.1} = \arg\min_{h} |\hat{\rho}(h) - 0.1|.$$

La recherche du minimum de $\hat{\mathcal{L}}^{\star}$ est alors limitée au créneau $[\,h_{0.1} - 3,\ h_{0.1} + 5\,]$, ce qui divise le temps de calcul par 3.

4. Principaux constats

— **Bootstrap** : pour $B = 499$ et $p \simeq 0.5$, $h^{\star}_{I}$ se concentre autour du vrai h_I même lorsque la dépendance spatiale est forte, Fig. 4.

— **SSLOO** : prendre seulement 50 % des sites conserve le même minimum et réduit le temps de 35 h à 5,5 h (83 %).

— **SBCV** : blocs 4×6 précis jusqu'à dépendance modérée ; pour θ élevé, blocs 8×12 stabilisent $h_I^\star$.

— **Application Roberts *et al.* (2017)** : malgré un modèle linéaire partiellement faux, SAC-0.1 et le bootstrap donnent des buffers proches du véritable optimum.

5. Conclusion

Le bootstrap paramétrique, conjugué (i) à l'heuristique SAC-0.1 pour borner l'intervalle et (ii) au sous-échantillonnage SSLOO, fournit une estimation robuste et peu coûteuse du buffer optimal, garantissant des risques CV proches du risque idéal même en présence de forte autocorrélation spatiale.

Résumé du Chapitre 4
Ridge regularization for spatial autoregressive models with multicollinearity issues

1. Contexte et objectifs

Les modèles autorégressifs spatiaux—**SAR** (mixed spatial–lag) et **SEM** (spatial–error)—sont devenus incontournables pour capturer l'autocorrélation spatiale (Anselin, 1988; LeSage, 2008). Or, dans les applications réelles, les covariables $\mathbf{X}$ sont souvent fortement collinéaires, ce qui rend la matrice $\mathbf{X}^\top\mathbf{X}$ (quasi-)singulière, gonfle la variance des estimateurs et détruit l'interprétabilité.

Ce chapitre propose :

— une *extension complète* de la régression de Ridge aux modèles SAR (9) et SEM (10) ;

— un *choix automatique* du paramètre de pénalisation γ par **SLOO** (Spatial Leave-One-Out) ;

— des *tests d'importance* (permutation F et t) adaptés au contexte régularisé ;

— une validation par simulation et une application aux hospitalisations COVID-19 en France.

2. Modèles et notation

Soit un vecteur–réponse $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))^\mathsf{T}$ observé sur $S_n \subset \mathbb{R}^d$ et $W = (w_{ij})_{i,j}$ une matrice de voisinage (normalisée par son rayon spectral).

SAR (*mixed spatial–lag*)

$$\mathbf{Y} = \rho W\mathbf{Y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}\Big(\mathbf{0}, \sigma^2\mathbf{I}_n\Big). \tag{9}$$

SEM (*spatial–error*)

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \qquad \mathbf{u} = \lambda W\mathbf{u} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}\Big(\mathbf{0}, \sigma^2\mathbf{I}_n\Big). \tag{10}$$

3. Estimateurs Ridge pour SAR et SEM

Pour des variables centrées–réduites, l'estimateur Ridge OLS classique est

$$\hat{\boldsymbol{\beta}}_{\mathrm{R}}(\gamma) = (\mathbf{X}^\mathsf{T}\mathbf{X} + \gamma\mathbf{I}_p)^{-1}\mathbf{X}^\mathsf{T}\mathbf{Y}, \qquad \gamma > 0.$$

Algorithme RRSAR

1. *Estimation initiale.* $\hat{\boldsymbol{\beta}}_{0,R}$ et $\hat{\boldsymbol{\beta}}_{L,R}$ via deux régressions Ridge séparées ($\mathbf{Y} \sim \mathbf{X}$ et $W\mathbf{Y} \sim \mathbf{X}$).
2. ρ par maximisation du *log-vraisemblance concentré*

$$\ell^C_{\mathrm{SAR}}(\rho) = -\tfrac{n}{2}\log\widehat{\sigma^2}(\rho) + \operatorname{logdet}(\mathbf{I}_n - \rho W), \quad \widehat{\sigma^2}(\rho) = \frac{\|e_O - \rho e_L\|^2}{n}. \tag{11}$$

3. *Filtrage.* $\mathbf{Y}_{\hat\rho} = (\mathbf{I}_n - \hat\rho W)\mathbf{Y}$.
4. *Ridge global.*

$$\hat{\boldsymbol{\beta}}^{\mathrm{RRSAR}}(\gamma) = (\mathbf{X}^\mathsf{T}\mathbf{X} + \gamma\mathbf{I}_p)^{-1}\mathbf{X}^\mathsf{T}\mathbf{Y}_{\hat\rho}. \tag{12}$$

Algorithme RRSEM idem, mais avec itérations $\mathbf{Y}_{\hat\lambda} \to \hat{\boldsymbol{\beta}} \to \hat\lambda$ jusqu'à convergence.

4. Sélection de γ : SLOO spatial

Lors du m^{e} repli-test SLOO ($\mathbf{s}_m$ validation) :

$$\hat\beta_{R,m}(\gamma) = (\mathbf{X}_m^\mathsf{T}\mathbf{X}_m + \gamma\mathbf{I}_p)^{-1}\mathbf{X}_m^\mathsf{T}\mathbf{Y}_{\hat\rho,m},$$

$$\ell_m(\gamma) = -\tfrac{1}{2}\log\hat{\sigma}^2 - \frac{1}{2\hat{\sigma}^2}\Big(Y(\mathbf{s}_m) - \hat{\rho}\sum_{\mathbf{s}_j\in\partial\mathbf{s}_m} w_{mj}Y(\mathbf{s}_j) - \mathbf{x}(\mathbf{s}_m)^\top\hat{\beta}_{R,m}(\gamma)\Big)^2.$$

Le paramètre est alors

$$\hat{\gamma} = \arg\max_{\gamma\in\Gamma}\widehat{\mathcal{L}}(\gamma), \qquad \widehat{\mathcal{L}}(\gamma) = \frac{1}{n}\sum_{m=1}^{n}\ell_m(\gamma), \tag{13}$$

où Γ est une grille log-décroissante de K valeurs ($K=100$, $\gamma_{\min} = 10^{-3}\gamma_{\max}$).

5. Tests d'importance après Ridge

Statistique « t » de Halawa

$$T_j = \frac{\hat{\beta}_{R,j}}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_{R,j})}}, \qquad \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_R) = \hat{\sigma}^2(\mathbf{X}^\top\mathbf{X}+\gamma\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X}+\gamma\mathbf{I}_p)^{-1}.$$

Permutation F-test (variables échangeables)

$$F_j^{(b)} = \frac{\|\mathbf{Y}-\hat{\mathbf{Y}}_0\|^2 - \|\mathbf{Y}-\hat{\mathbf{Y}}_{j(b)}\|^2}{\|\mathbf{Y}-\hat{\mathbf{Y}}_{j(b)}\|^2/(n-p)}, \quad \text{p-value} = \frac{1}{B}\sum_{b=1}^{B}\mathbf{1}\{F_j^{(b)}\geq F_j\}. \tag{14}$$

6. Principaux résultats de simulation

— **Gain de stabilité** : par rapport aux estimateurs ML classiques, RRSAR/RRSEM divisent la variance moyenne des β par 50 à 100 (scénarios $\rho,\lambda\in[0.1,0.9]$).

— **Biais** légèrement accru (penalisation) mais MSE global inférieur à la méthode « SFRR » de Fan *et al.*.

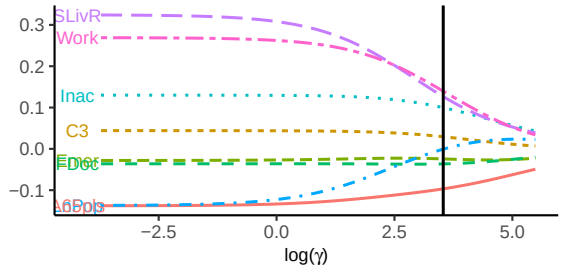
— **Choix** γ : la maximisation de $\widehat{\mathcal{L}}(\gamma)$ coïncide systématiquement avec la stabilisation des trajectoires Ridge (Fig. 5).

— **Tests** : le permutation F-test se révèle plus conservateur que les t-tests ; tous confirment la significativité des variables réellement pertinentes dans les scénarios simulés.

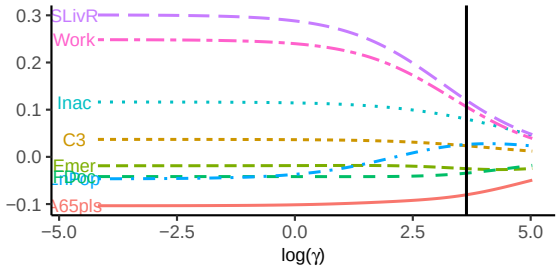
7. Application : hospitalisations COVID-19 (France, 2020)

— **Données** 96 départements, variable réponse : `LnHosp` (log-taux d'hospitalisation). 8 covariables socio-économiques très corrélées (Fig. 6).

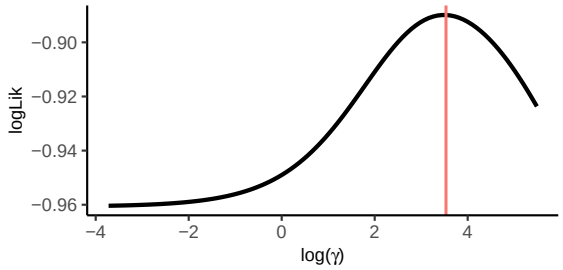
— **Multicolinéarité** sévère : $\kappa(\mathbf{X}^\top\mathbf{X}) = 1.4\times10^5$.



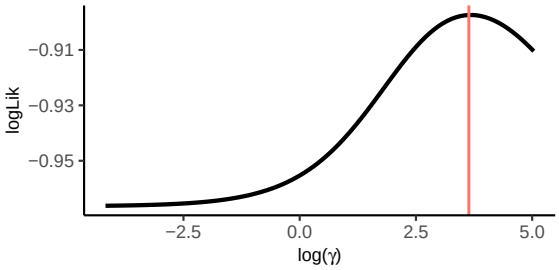
(a) RRSAR coefficient paths



(b) RRSEM coefficient paths



(c) RRSAR SLOO log-likelihood



(d) RRSEM SLOO log-likelihood

FIGURE 5 – Ridge-path diagrams (top) and SLOO cross-validation profiles (bottom). The vertical line marks the optimal γ.

— $\gamma_{\text{opt}} \simeq 34$ (RRSAR) / 38 (RRSEM).

— Variables significatives (permutation F à 5 `SLivR` (inégalités de revenu), `Work` (% ouvriers), `Inac` (% inactifs) et, à 10 Les variables sanitaires `Emer` et `FDoc` sont non significatives.

— **Interprétation** : au-delà de la densité, ce sont surtout les facteurs socio-économiques défavorisés qui expliquent la sévérité locale de la première vague COVID-19.

8. Conclusion

La procédure RRSAR/RRSEM :

— traite simultanément la dépendance spatiale *et* la multicolinéarité ;

— choisit γ de manière totalement automatique via SLOO ;

— fournit des estimateurs stables et interprétables ;

— s'appuie sur des tests permutationnels pour l'importance des variables.

Le code est disponible sur `https://github.com/c0ra/RRSARMMI`. Des extensions vers le Lasso, l'Elastic-Net et les modèles spatiaux plus généraux (Durbin, SDEM, etc.) sont envisagées.

The four chapters of this thesis form a coherent line of work that moves from reliable performance assessment of spatial algorithms to robust estimation of spatial-

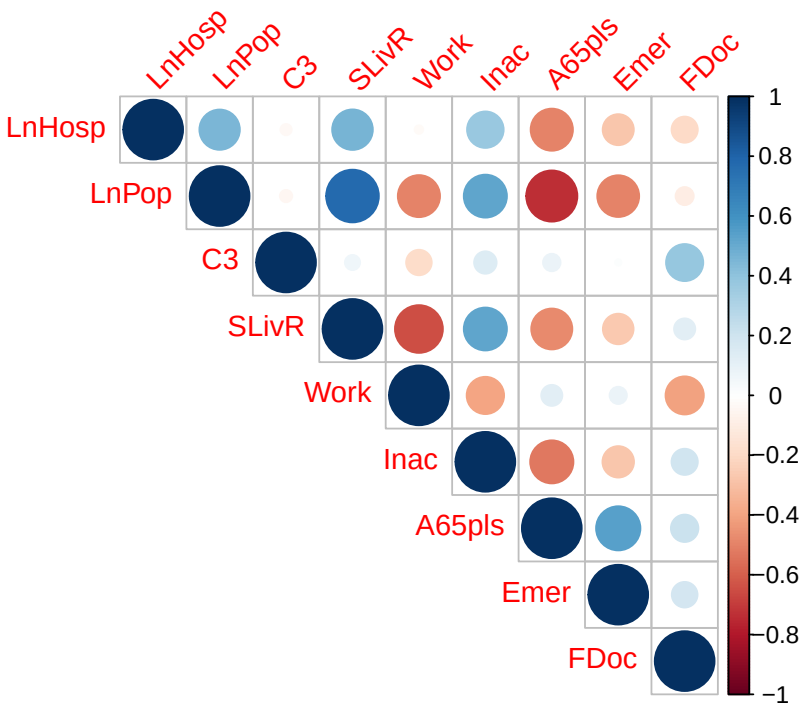


FIGURE 6 – Correlation matrix for the eight socio-economic covariates and the response (`LnHosp`).

autoregressive models under collinearity. Chapter 1 establishes that classical CV protocols are not trustworthy when residuals are spatially correlated and shows—through controlled experiments—that only CV schemes that enforce an explicit spatial gap (SBCV and SLOO) provide nearly unbiased risk estimates. Building on that insight, Chapter 2 turns the vague notion of a "large enough buffer" into a data–driven quantity : the buffer that minimises the squared distance between the unobtainable "ideal" risk and the empirical CV risk. A simple operational rule—stopping the buffer at the lag where residual SAC 0.1—emerges as a surprisingly effective proxy across a wide spectrum of dependence strengths and block layouts. Chapter 3 removes the need for multiple independent realisations by replacing them with parametric bootstrap copies that honour the fitted covariance. The bootstrap criterion, optionally combined with a subsampled SLOO (SSLOO), consistently rediscovers the true optimal buffer while cutting computing time by more than 80 %. Finally, Chapter 4 shows that the same spatial-CV machinery can be exploited to tune a Ridge penalty in SAR/SEM models, giving an estimator that is simultaneously spatially coherent and multicollinearity-robust. Simulations and a COVID-19 case study confirm that the Ridge–SAR/SEM

estimator outperforms both classical spatial regression and the previously proposed "spatially-filtered Ridge" in terms of bias–variance trade-off and interpretability.

Together, these contributions deliver a practical toolbox for spatial statisticians : (i) use SLOO or SBCV, not random CV ; (ii) set the buffer either at the SAC-0.1 lag or, when higher precision is required, via the bootstrap discrepancy criterion ; (iii) combine that spatial CV with regularisation to stabilise inference in the presence of strong covariate collinearity. While the work was conducted under linear-Gaussian assumptions and regular lattices, the methodology—and the software that accompanies it—lays the groundwork for future extensions to irregular sampling designs, random-$\mathbf{X}$ settings, non-Gaussian responses, and other penalties such as Lasso or Elastic Net. In short, this thesis bridges the gap between how we validate spatial models and how we fit them when real-world complications—dependence and multicollinearity—cannot be ignored.

Bibliographie

Anselin, L. (1988). *Spatial econometrics : methods and models*. Dorddrecht : Kluwer Academic Publishers, Dordrecht.

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

Karasiak, N., Dejoux, J.-F., Monteil, C. et Sheeren, D. (2022). Spatial dependence between training and test sets : another pitfall of classification accuracy assessment in remote sensing. *Machine Learning*, 111(7):2715–2740.

Le Rest, K., Pinaud, D., Monestiez, P., Chadoeuf, J. et Bretagnolle, V. (2014). Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. *Global Ecology and Biogeography*, 23(7):811–820.

LeSage, J. P. (2008). An introduction to spatial econometrics. *Revue d'économie industrielle*, 123:19–44.

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.* (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic geography*, 46(sup1):234–240.