Extrapolation spatiale du risque de présence de Xylella fastidiosa basée sur XGBoost

Camille Portes [1] & Dino Ienco [2] & Edith Gabriel [3]

[1] *INRAE, BioSP, 84914 Avignon, France, camille.portes@inrae.fr*
[2] *UMR TETIS, 34000 Montpellier, France, dino.ienco@inrae.fr*
[3] *INRAE, BioSP, 84914 Avignon, France, edith.gabriel@inrae.fr*

Résumé. Nous proposons une méthodologie complète pour évaluer et cartographier le risque de présence de la bactérie *Xylella fastidiosa* en intégrant les composantes spatiales dans le processus de modélisation. Notre approche est basée sur l'apprentissage automatique, tenant compte des particularités des données : hétérogénéité spatiale et déséquilibre. Une pré-sélection de facteurs est réalisée à l'aide d'une approche ensembliste couplée à une validation croisée spatiale. Nous proposons ensuite une adaptation du modèle XGBoost dans laquelle les composantes spatiales sont intégrées au modèle, notamment en considérant des facteurs spatialement pondérés et en basant la sélection de modèle sur une validation croisée par blocs environnementaux. Cette approche nous permet d'obtenir un modèle robuste, généralisable à une zone géographique éloignée de celle utilisée pour l'entraînement du modèle et donc adaptée à l'extrapolation.

Mots-clés. Prédiction, Validation croisée spatiale, XGBoost

Abstract. We propose a methodology to assess and map the risk of presence of the bacterium *Xylella fastidiosa* by integrating spatial components into the modeling process. Our approach is based on machine learning, taking into account the data's specificities: spatial heterogeneity and imbalance. A pre-selection of factors is performed using an ensemble approach coupled with spatial cross-validation. We then propose an adapted XGBoost model where spatial components are integrated, both by considering spatially weighted factors and by basing model selection on block environmental cross-validation. This approach enables us to obtain a robust model that is generalizable to a geographic area distant from the one used for model training and thus suitable for extrapolation.

Keywords. Prediction, Spatial Cross Validation, XGBoost

1 Introduction

Connaître la santé des plantes dans une région permet de confirmer l'absence de la plupart des organismes nuisibles réglementés. La surveillance officielle repose sur des inspections visuelles, des prélèvements et des analyses d'échantillons. En cas de détection, les autorités sanitaires coordonnent des mesures de lutte collective pour empêcher l'établissement de l'organisme

nuisible dans une zone, préservant ainsi le reste du territoire. Nous nous intéressons ici à *Xylella fastidiosa*, bactérie vectée par des insectes (cicadelles) et potentiellement présente dans pas moins de 260 espèces de plantes. Apparue en Italie en 2013 et découverte plus tard en Espagne et au Portugal, la bactérie est désormais observée en Corse, en Provence-Alpes-Côte d'Azur et tout récemment en Occitanie. Dans ce contexte nos objectifs sont d'identifier les facteurs de risque de présence de la bactérie et de proposer un modèle d'extrapolation spatiale pour cartographier ce risque et orienter la surveillance.

Nous développons une méthodologie basée sur l'apprentissage automatique qui prend en compte les spécificités des données à chaque étape de la modélisation : hétérogénéité spatiale et déséquilibre (le nombre d'échantillons négatifs est largement supérieur aux échantillons positifs) et qui permet de faire des prédictions spatiales au-delà des emplacements des échantillons d'entraînement. Nous illustrerons l'approche en utilisant les données de Corse et de PACA comme ensemble d'entrainement et les données d'Occitanie comme ensemble de test.

2 Sélection de facteurs

Nous disposons d'un ensemble de 107 facteurs qui influencent le développement des plantes et donc la propagation des organismes nuisibles. Il s'agit de variables bioclimatiques (calculées à partir des variables climatiques de Siclima), de composition du sol (issues de la base Agroenvgeo), de type de sol (issues de Corine Land Cover), d'orientation du terrain et d'altitude. La base de données est décrite dans Portes *et al.* (2024).

Après élimination des facteurs trop fortement corrélés, nous sélectionnons les facteurs via une approche ensembliste couplée à une validation croisée spatiale (voir Section 3). Afin d'éviter d'éventuels biais d'induction, les modèles considérés sont de différent type : basé sur du boosting (e.g. adaboost, glmboost, xgboost...), sur des arbres (comme le random forest classique ou conditionnel), sur de l'analyse discriminante, sur les support vector machine, sur des modèles de régression logistique, ... Chacun des 29 modèles, adapté à notre contexte de classification binaire (plante testée positive versus négative), sélectionne 20 facteurs parmi nos 107 facteurs + 1 variable aléatoire. Cette sélection est faite en utilisant l'importance basée sur l'erreur "Out-of-Bag". Ainsi, nous avons considéré et comparé les résultats issus de différents critères :

1. Gain d'information

2. Importance (réduite) lorsqu'elle est calculable pour le modèle considéré ou poids uniforme lorsqu'elle ne l'est pas (dans ce cas le modèle sélectionne 20 facteurs mais leurs rangs sont inconnus),

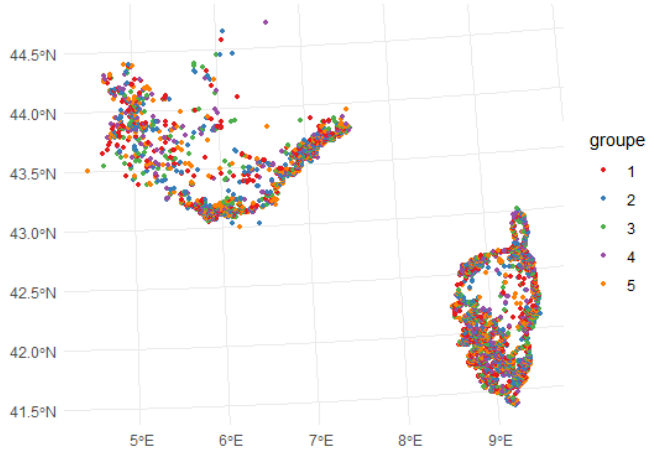
3. Importance (réduite) lorsqu'elle est calculable pour le modèle considéré.

Nous avons écarté le cas 1. car il s'est révélé être assez sensible à l'échelle du support d'observation de certains facteurs, en particulier pour la composition en éléments chimiques du sol. Les cas 2. et 3. mènent sensiblement aux mêmes résultats, avec 83 facteurs retenus (importance positive).

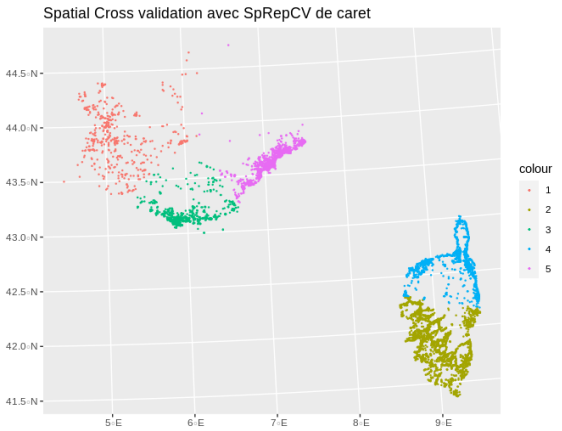
3 Validation croisée

Meyer et Pebesma (2021) mettent en évidence que, dans de nombreux contextes spatiaux, l'utilisation de la validation croisée aléatoire (Figure 1a) donne des résultats trop optimistes près des échantillons d'entrainement et de mauvais résultats en extrapolation et qu'il est préférable d'utiliser la validation croisée spatiale ou par blocs environnementaux (Valavi *et al.*, 2018).

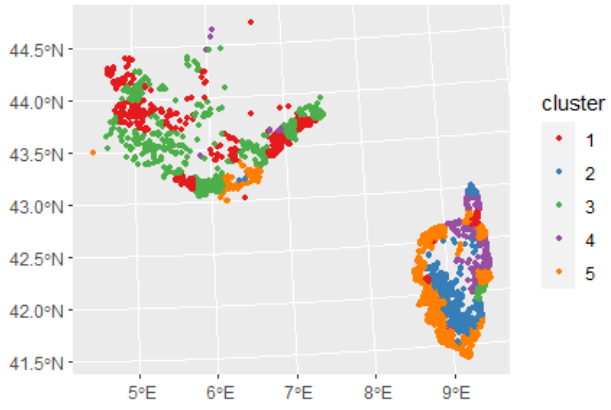
Dans un contexte spatial, la validation croisée valide le modèle sur un ensemble spatialement similaire (Figure 1b) à l'ensemble d'entraînement. Deux observations qui sont proches spatialement sont également proches en valeur pour les variables autocorrélées spatialement. Si le domaine de prédiction est éloigné, le modèle se comportera mal car il sera trop ajusté aux données d'entraînement et donc pas généralisé. La validation croisée spatiale contribue à résoudre ce problème mais peut ne pas être suffisante ou tout à fait adaptée.



(a) Validation croisée aléatoire



(b) Validation croisée spatiale

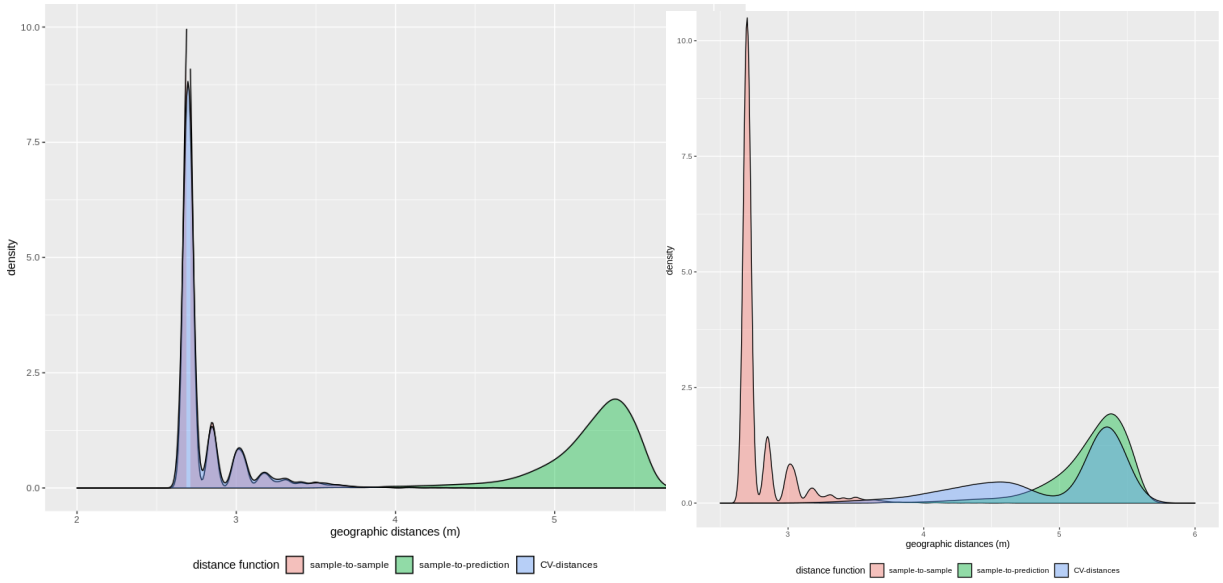


(c) Validation croisée par blocs environnementaux

Figure 1: Repésentation des ensembles d'entrainement et de validation selon le type de validation croisée.

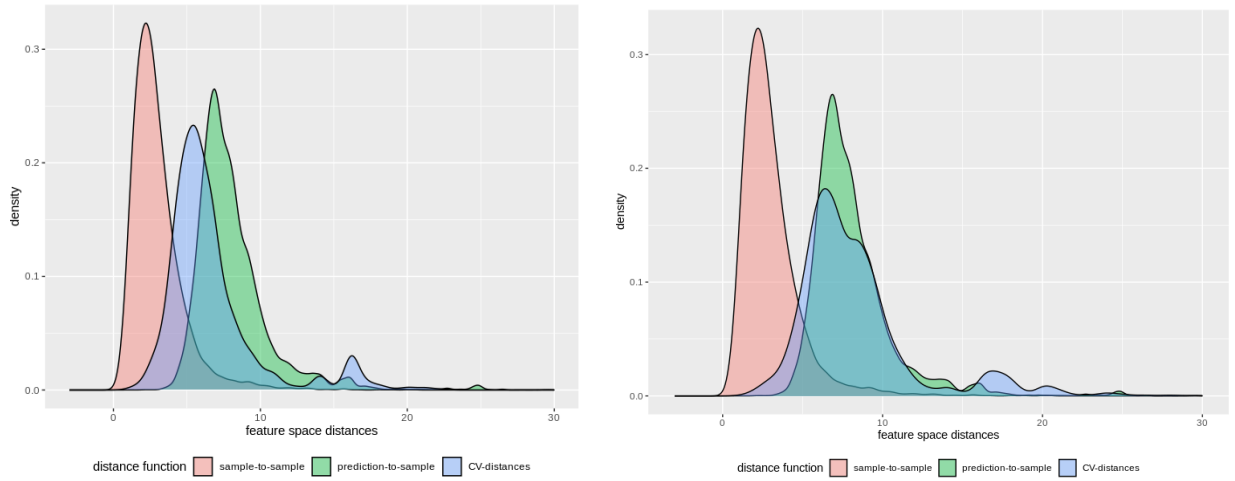
Pour gagner en robustesse sur la prédiction dans des zones éloignées, nous utilisons la validation croisée par blocs environnementaux (Figure 1c). Cette dernière utilise des méthodes de regroupement (k-means) pour spécifier des ensembles de facteurs similaires en fonction des variables d'entrée et du nombre de clusters choisi dans l'espace des facteurs.

La validation croisée par blocs environnementaux est adaptée à notre contexte. En effet, nous avons calculé les distances moyennes entre les points de l'ensemble d'entrainement et les distances entre les points de l'ensemble d'entrainement et de l'ensemble de test. Ces distances sont calculées dans un premier temps avec la position des observations, c'est donc une distance spatiale (Figure 2). Dans un deuxième temps, la distance est calculée à partir des facteurs (indice de dissimilarité) (Figure 3), qui est corrélé à la distance spatiale. La validation croisée spatiale, comme la validation croisée par blocs environnementaux, permet de valider le modèle sur un ensemble plus éloigné de l'ensemble d'entraînement (Figure 2b) comparé à la validation croisée aléatoire (Figure 2a). Cette distance spatiale se traduit par une dissimilarité de l'environnement, c'est-à-dire des facteurs. La validation croisée spatiale (Figure 3a) et la validation croisée par blocs environnementaux (Figure 3b) permettent d'avoir des écarts en terme de similarité du même ordre de grandeur entre les ensembles d'entraînement et de validation et les ensembles d'entraînement et de test, ce qui nous assure de la fiabilité du modèle, choisi sur l'ensemble de validation et sur l'ensemble de test.



(a) Validation croisée aléatoire (b) Validation croisée spatiale

Figure 2: Distances entre les ensembles d'entrainement, de validation et de test



(a) Validation croisée spatiale

(b) Validation croisée par blocs environementaux

Figure 3: Similarités des environnements entre les ensembles d'entrainement, de validation et de test

4 Modélisation

L'autocorrélation spatiale et l'hétérogénéité spatiale sont deux effets spatiaux bien connus dans l'analyse spatiale et la modélisation. L'autocorrélation spatiale fait référence au processus qui crée des amas de valeurs. Dans notre cas, les amas locaux de cas positifs sont à la fois liés au mode de vection de la bactérie et au processus de surveillance (renforcé dans un périmètre étroit autour d'un cas positif). Nous nous affranchissons d'une grande partie de ce biais spatial en agrégeant les données sur une grille de maille 500 $m \times 500$ m comme expliqué dans Martinetti et Soubeyrand (2019). L'hétérogénéité spatiale fait référence au fait que le processus qui génère les données peut différer d'un endroit à un autre. Elle est fréquemment modélisée par des modèles de coefficients variant spatialement (c'est par exemple le cas de la Régression Pondérée Géographiquement).

La littérature de cette dernière décennie montre un progrès visible dans le développement de la modélisation spatiale d'apprentissage automatique. Les modèles rencontrés se déclinent dans les combinaisons "ML classique ou spatial + variables non spatiales et/ou spatiales + validation croisée aléatoire ou spatiale". Le progrès et les innovations sont principalement liés à la formulation des méthodes d'apprentissage automatique pour incorporer des composantes spatiales dans les algorithmes.

Nous utilisons le modèle XGBoost (eXtreme Gradient Boosing). Cette méthode de boosting de gradient assemble séquentiellement des arbres de décision pour minimiser l'erreur du modèle en utilisant un algorithme d'optimisation de descente de gradient (Chen et Guestrin, 2016). La littérature montre qu'un XGBoost correctement ajusté surpasse généralement les méthodes alternatives telles que la forêt aléatoire ou le réseau de neurones profonds pour les problèmes supervisés.

Nous intégrons les composantes spatiales dans ce modèle d'une part en considérant, lorsque cela est pertinent, des facteurs spatialement pondérés et d'autre part en basant la sélection de modèle sur une validation croisée par blocs environnmentaux. Notre démarche est résumée dans la figure 4.

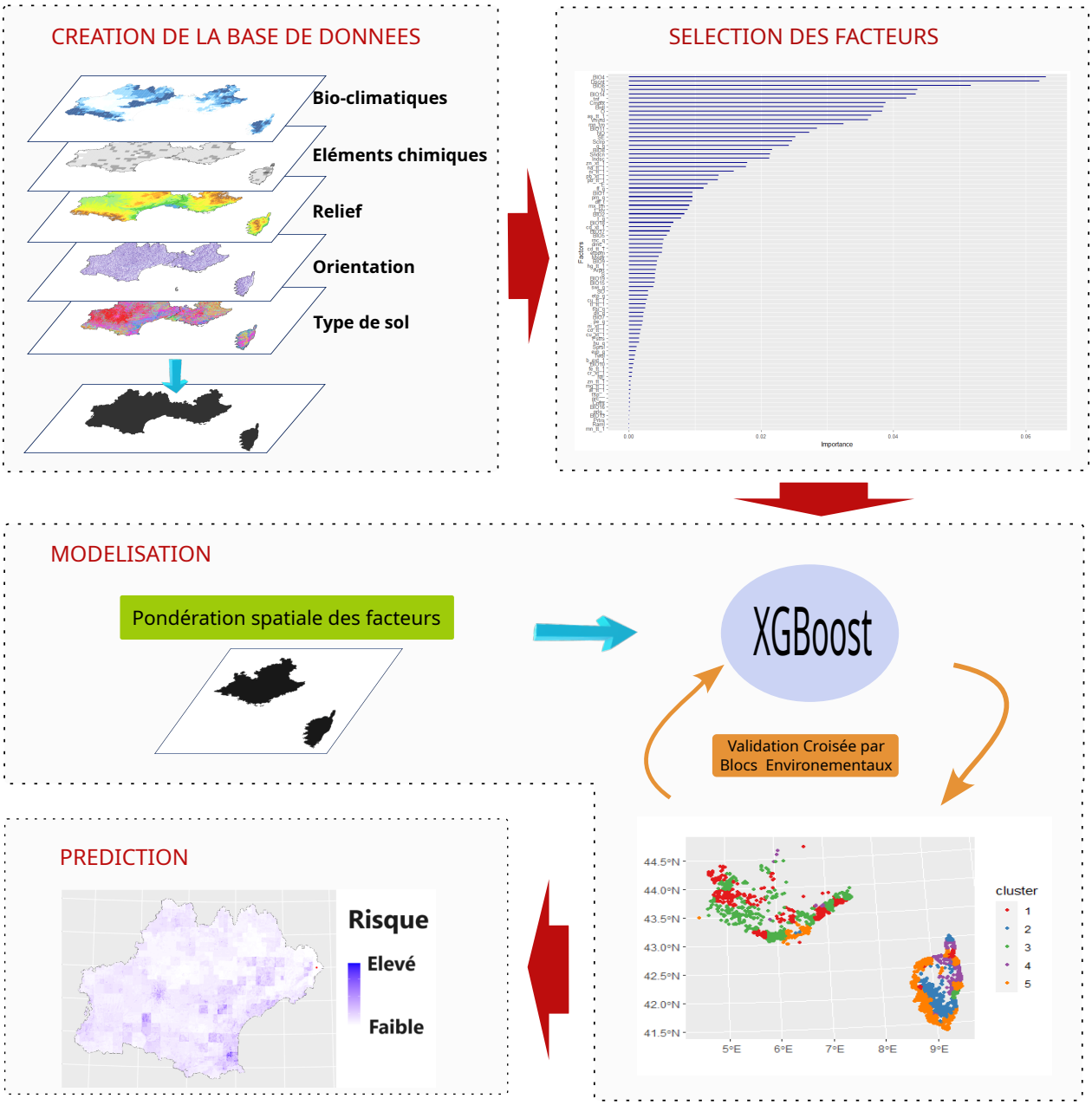


Figure 4: Méthodologie d'extrapolation spatiale basée sur XGBoost.

Ainsi, pour chacun des 83 facteurs sélectionnés nous avons testé l'existence d'une structure spatiale : 8 d'entre eux n'en ont pas. Pour les 75 autres, nous avons réalisé une analyse variographique et estimé la portée (distance au-delà de laquelle on considère qu'il n'y a plus de corrélation). Nous avons ensuite calculé la matrice de poids associée à chacun de ces facteurs à partir d'un noyau de type "triangle", avec une fenêtre égale à la portée, et telle

que seuls les 20 premiers voisins soient pris en compte.

Notre ensemble de facteurs X est donc divisé en deux $X = (X_a, X_s)$ où X_a désigne l'ensemble des facteurs aspatiaux et X_s l'ensemble des facteurs spatiaux. L'algorithme a ensuite été appliqué à l'ensemble $(X_a, W_1X_{1,s}, \ldots, W_{75}X_{75,s})$ que l'on notera WX.

Pour comparer les modèles, nous utilisons des mesures qui prennent en compte le déséquilibre du ratio positifs/négatifs, telles que la balanced accuracy (qui calcule le pourcentage de positifs et négatifs bien prédits parmi tous les positifs et tous les négatifs), le F1-score (qui calcule la moyenne entre le pourcentage de positifs bien prédits parmi les positifs et le pourcentage de positifs parmi ceux prédits comme positifs) ou l'AUC.

Nous illustrerons la comparaison des résultats obtenus à partir des modèles XGBoost(X), XGBoost(WX) et XGBoost(X, WX) et pour les validations croisées aléatoire, spatiale et par blocs environnementaux.

Bibliographie

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. pages 785–794.

Martinetti, D. and Soubeyrand, S. (2019). Identifying lookouts for epidemio-surveillance: application to the emergence of xylella fastidiosa in france. Phytopathology, 109(2):265–276.

Meyer, H. and Pebesma, E. (2021). Predicting into unknown space? estimating the area of applicability of spatial prediction models. Methods in Ecology and Evolution, 12(9):1620–163.

Portes, C., Ienco, D. and Gabriel, E. (2024). Environmental and Bio-climatic Data for Epidemiological Analysis over French Mediterranean areas. (Soumis).

Valavi, R., Elith, J., Lahoz-Monfort, J. J., and Guillera-Arroita, G. (2018). blockcv: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. bioRxiv, page 357798.