



HAL
open science

Utilisation d'une approche par gradient boosting pour l'estimation de modèles autorégressifs spatiaux non linéaires. Application à des données de surveillance de la jaunisse

Ghislain Geniaux

► To cite this version:

Ghislain Geniaux. Utilisation d'une approche par gradient boosting pour l'estimation de modèles autorégressifs spatiaux non linéaires. Application à des données de surveillance de la jaunisse. SEPIM Plénière 2023, INRAE, Mar 2023, paris, France. hal-04229886

HAL Id: hal-04229886

<https://hal.inrae.fr/hal-04229886>

Submitted on 5 Oct 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Utilisation d'une approche par gradient boosting pour l'estimation de modèles autorégressifs spatiaux non linéaires. Application à des données de surveillance de la jaunisse

Ghislain Geniaux, INRAE Ecodéveloppement UR767

PACKAGE R SPBOOST

Ce package vise à permettre d'estimer des modèles autoregressifs spatiaux de type (SAR, SDM, SEM and SARAR) avec des relations non linéaires avec différents estimateurs s'appuyant sur des algorithmes de boosting (Friedman, 2001; Bühlmann et al., 2007).

2 autres méthodes sont aussi proposé pour accélérer les estimations en présence de gros échantillons :

- Closed Form Estimator (CFE, Smirnov 2020) pour les modèles SAR et SEM (CFE, Smirnov 2020),
- Flexible Instrumental Variable Approach (FIVA, Marra and Radice 2010) pour les modèles SAR.

Modèles autoregressifs spatiaux non-linéaires

$$Y = \rho WY + \sum_{j=1}^p h_j(X_j) + \epsilon \quad (SAR)$$

$$Y = \sum_{j=1}^p h_j(X_j) + (I - \lambda M)^{-1} \epsilon \quad (SEM)$$

Modèles autoregressifs spatiaux non-linéaires

$$Y_{it} = \rho \sum_j w_{j,s < t} Y_{j,s < t} + \sum_{j=1}^p h_j(X_j) + \epsilon \quad (SAR)$$

$$Y = (I - \rho W)^{-1} \sum_{j=1}^p h_j(X_j) + \epsilon \quad (SAR)$$

Modèles autoregressifs spatiaux non-linéaires

La logvraisemblance concentré de ce type de modèle s'écrit:

$$\ln L(\rho) = C + |I - \rho W| \quad (1)$$
$$- \frac{n}{2} \ln \left(\frac{(Y - \rho W Y - \sum h_j(X_j))' (Y - \rho W Y - \sum h_j(X_j))}{n} \right)$$

où $C = -(n/2) \ln(2\pi) - (n/2)$.

Functionalités de `spboost`

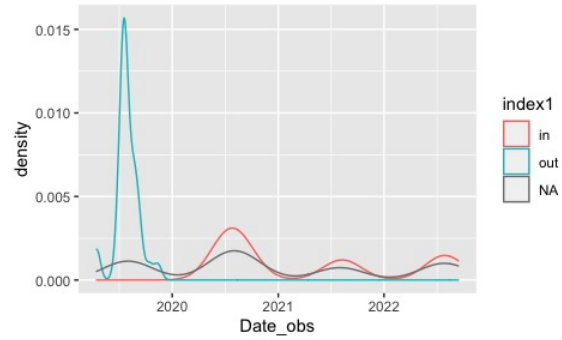
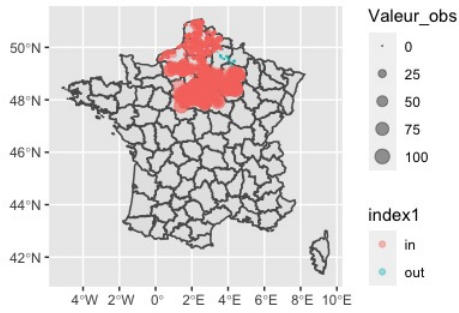
`spboost` permet d'estimer des modèles SAR, SEM et SARAR :

- cas gaussien (Y continu)
- cas probit (Y binomial)
- $h(X) \sim$ fonctions splines (mboost, mgcv/gam)
- $h(X) \sim$ arbres de décisions (xgboost)
- d'identifier la matrice optimale W (in progress)
- Predictions s'appuient sur un estimateur BLUP spatial (Goulard et al. 2017)

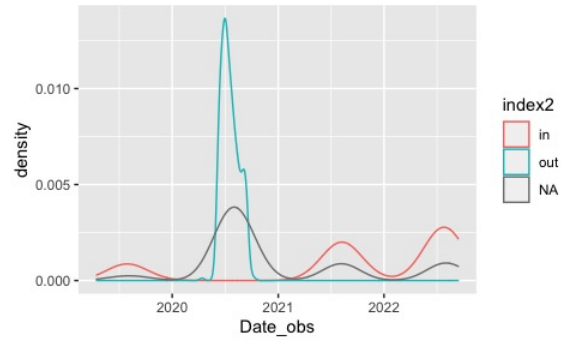
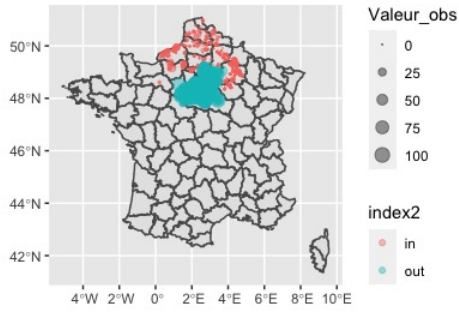
Application aux données *jaunisse*

- Stratégie de cross validation : plus de temporel moins de spatial dans l'échantillonnage. (reset annuel ?)
- Prise en compte de l'information de début d'année
- horizon de prédiction.

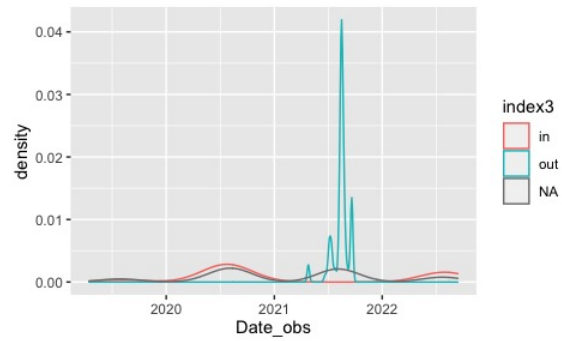
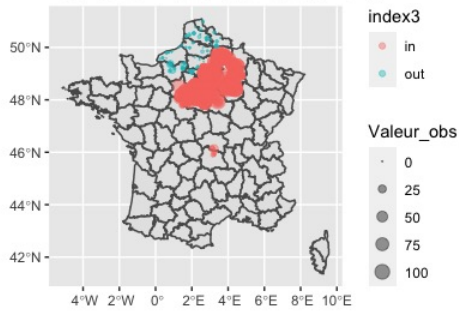
localisation fold 1 (out black,in red)



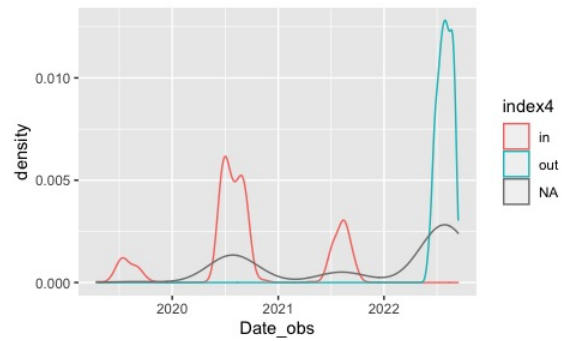
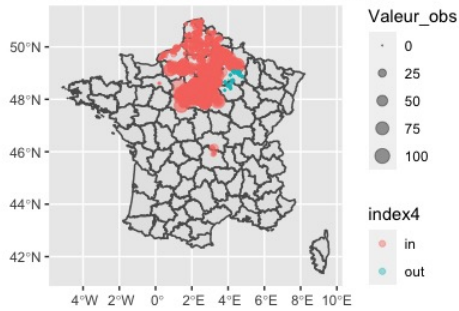
localisation fold 2 (out black,in red)



localisation fold 3 (out black,in red)

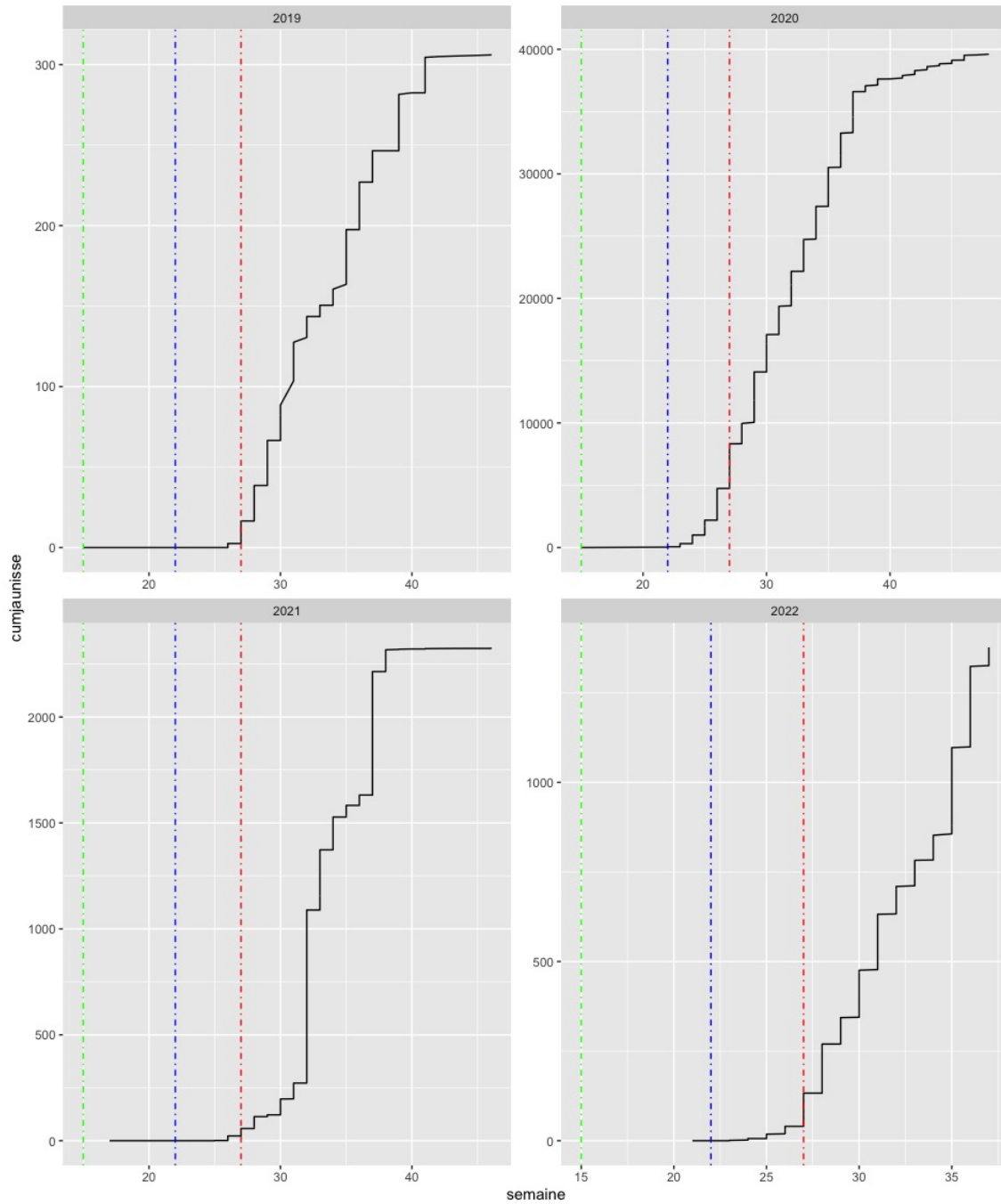


localisation fold 4 (out black,in red)



```
> apply(dist_train,2,function(x) quantile(x,c(0.2,0.8)))
  [,1] [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10]
20%   0   0 0.00000  0.000  0.000  0.000  0.000  0.000  0.00  0.000
80%   0   0 51.00253 1063.318 2536.319 3549.165 4798.226 6430.166 7448.54 8395.293
> apply(dist_test_train,2,function(x) quantile(x,c(0.2,0.8)))
  [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10]
20% 392.1433 400.7523 414.8903 458.1708 544.3642 578.9649 634.1451 786.2675 786.2675 832.5567
80% 2848.7532 3581.4461 4050.9302 4493.7231 4559.7512 4745.8794 6833.6161 7573.0305 8753.5954 9728.5111
```

green debut relevé min by year, red depasse 10 max by year : ref = blue semaine 22

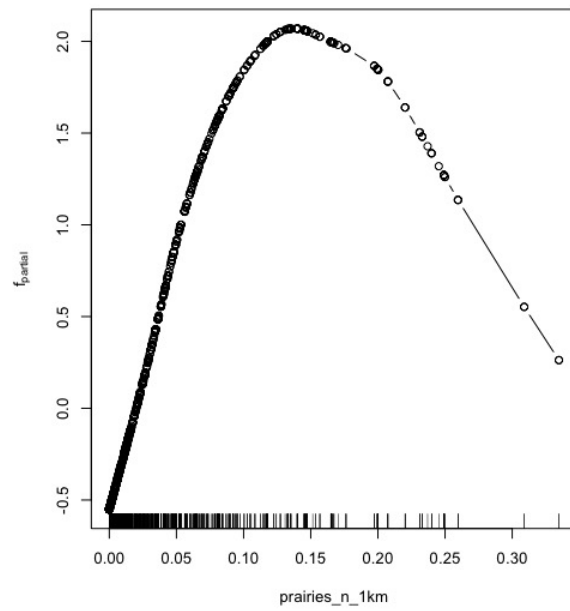
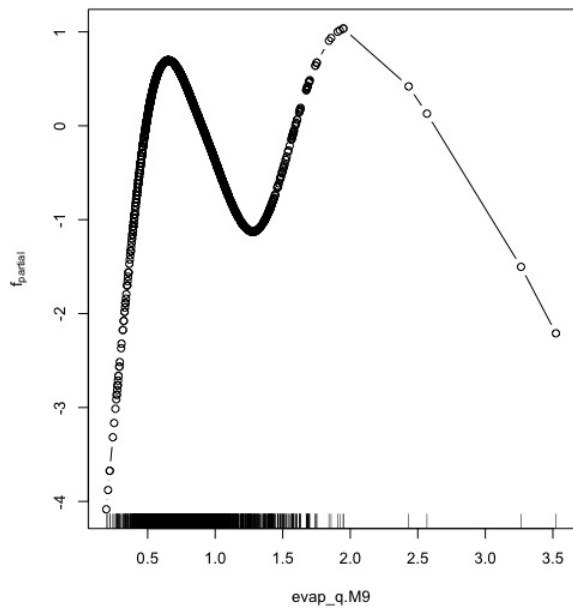
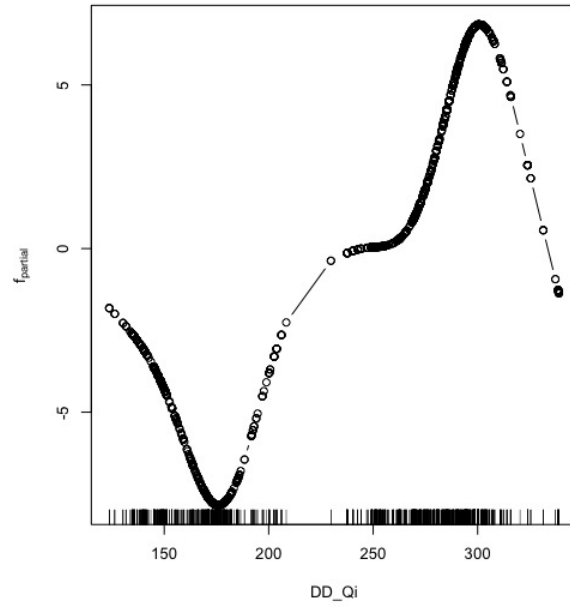
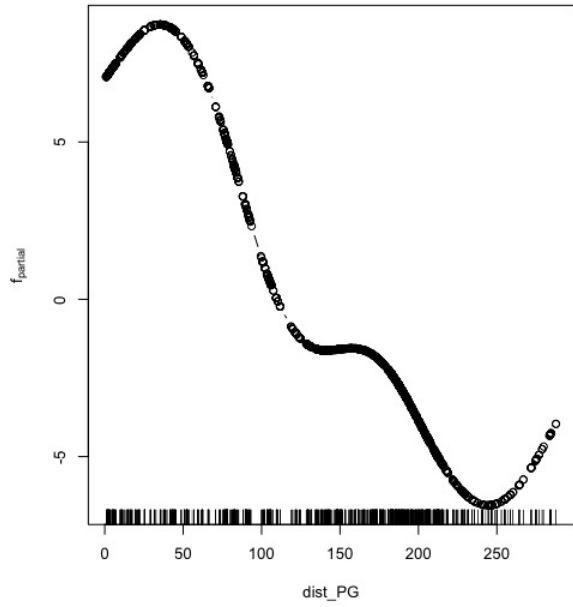


```
s1<-sampling_st(jj,2021,22,36,0)
s2<-sampling_st(jj,2021,22,36,1)
s3<-sampling_st(jj,2021,22,36,2)
s4<-sampling_st(jj,2021,22,36,4)

W1km<-matWpast(coords, TI, k=100, radius=100000
, H=1000, kernels='gauss')
WT5<-matTIME(TI, H=5, kernels='gauss')
## + croisement

model_spboost=spgam(formula=myformula, data=train, W=W_tr
DGP='SAR', method='gamboost_ML',
control=list(control_gamboost=boost_control(mstop=1000,

pred<-predict_spboost(model_spboost, test, train,
W = W_TRAIN_TEST, type = "BPN")
pred[pred<0]<-0
```



Comparaisons : rf, xgboost et spboost sur 2020 2021 et 2022 avec depart du test en semaine 15,20, 22, 24, 26, 28 avec 0,1,2,4 semaine de delais.

- si semaine ≥ 22 spboost $>$ rf $>$ xgboost et particulièrement si la l'horizon de prediction est plus lointain.
- 10 à 20 % d'amélioration des RMSE (max en semaine 30)
- proportional logit/ dirilichet
- W optimal ?