ModStatSAP 19/11/2023

Peut-on améliorer les prédictions d'une épiphytie en combinant les résultats de plusieurs méthodes de machine learning ? Leçons d'une expérience numérique

César Martinez[1], Edith Gabriel[1], Ghislain Géniaux[2], Dorian Chauvin[1], Samuel Soubeyrand[1]

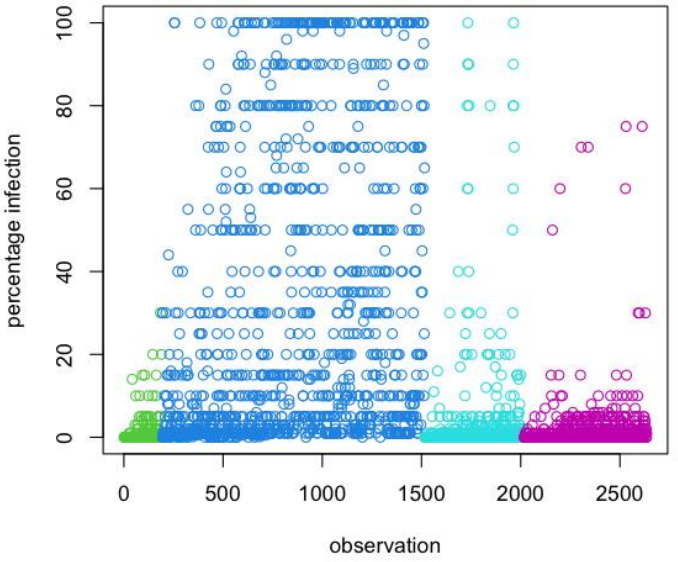
[1] INRAE Avignon BioSP
[2] INRAE Avignon Ecodev

PLAN

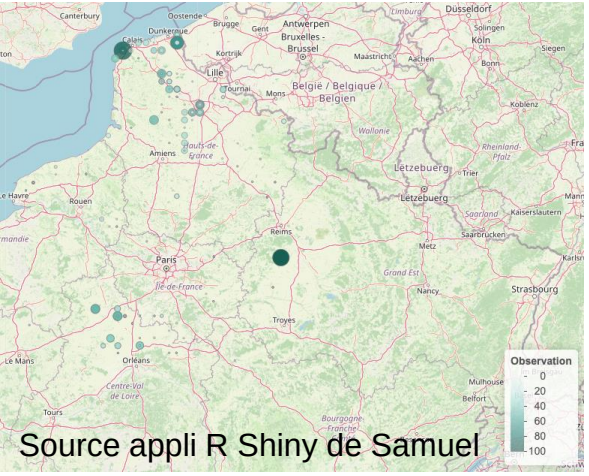
1) Contexte
2) Expérience numérique
3) Conclusion

Jaunisse de la betterave et données

Y_obs = %
parcelle jaune

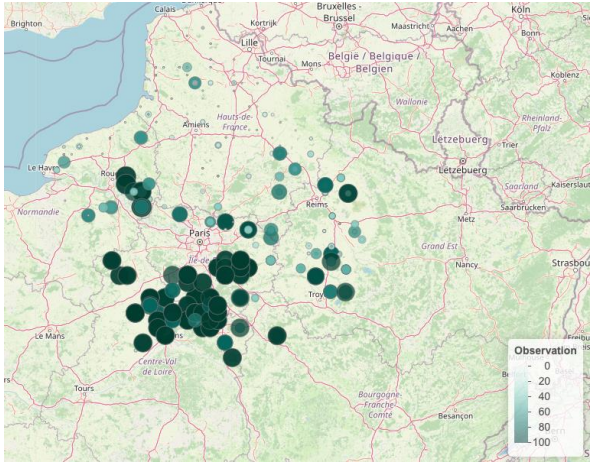


Observations 2019, 2020, 2021,2022



Source appli R Shiny de Samuel

2019



2020

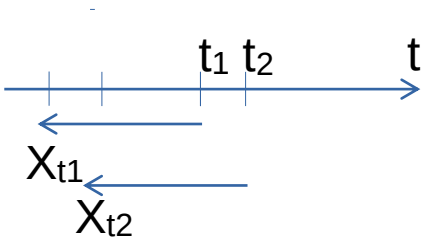
Dans SEPIM (travail de Dorian)

Bases de données

- Variables agronomiques
- Variables environnementales

Variables fixes

Variables « glissantes » ex :



Modèles de machine learning
(prédire la jaunisse)

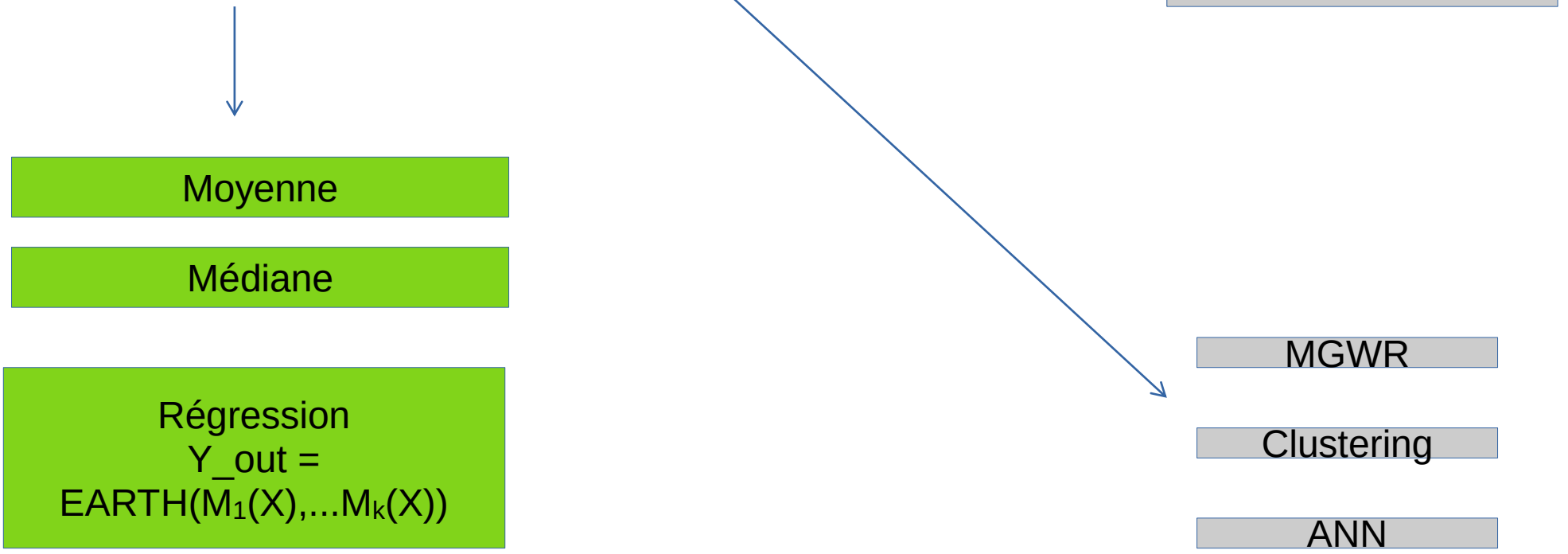
Earth
Rf
Cubist
XgBoost

$M_1(X)$, $M_2(X)$, …, $M_k(X)$

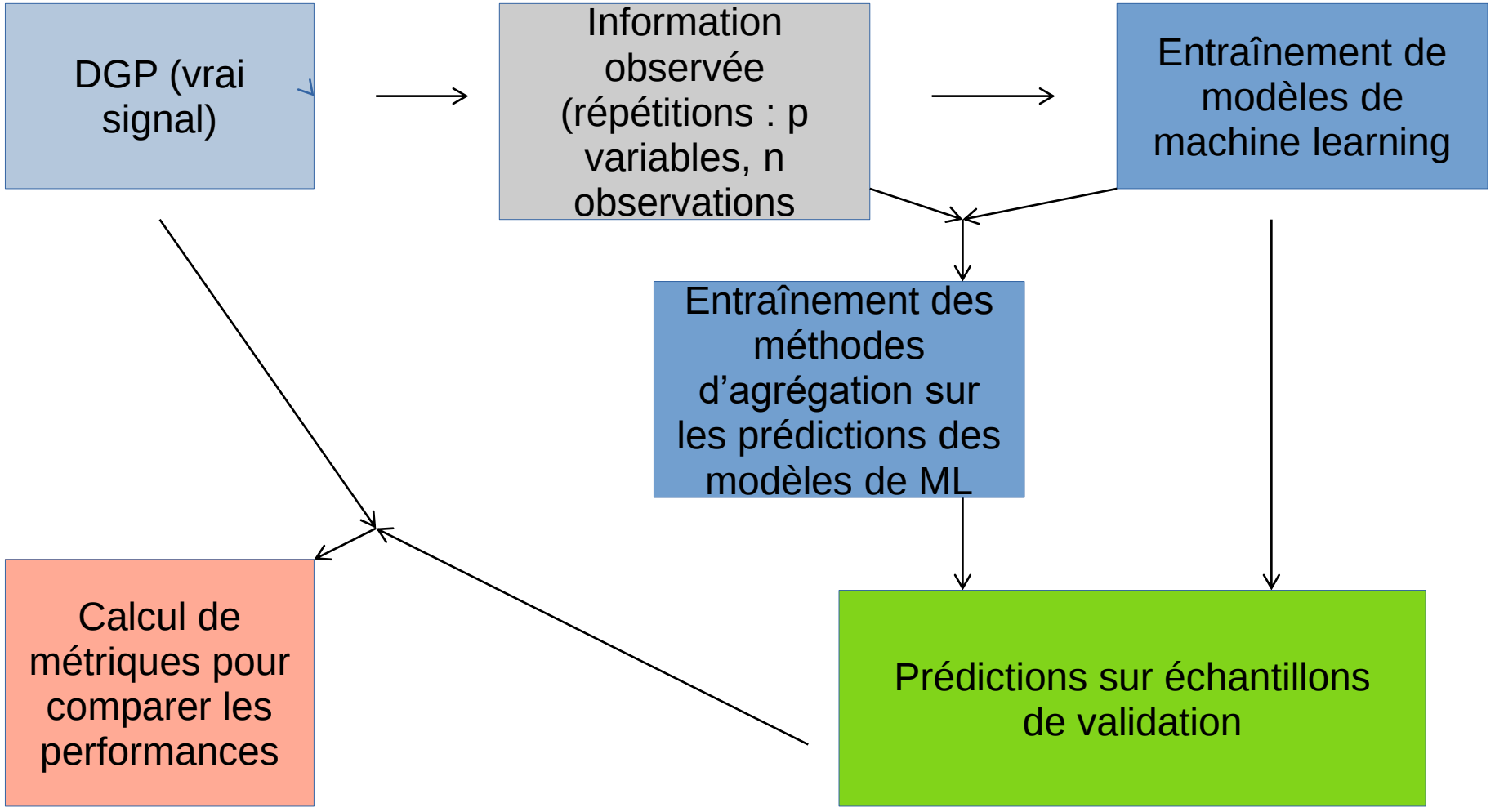
Agrégation ?

$Y = f(M_1(X), M_2(X),\ldots,M_k(X))$?

Méthodes d'agrégation
(doivent fonctionner avec des modèles de ML)

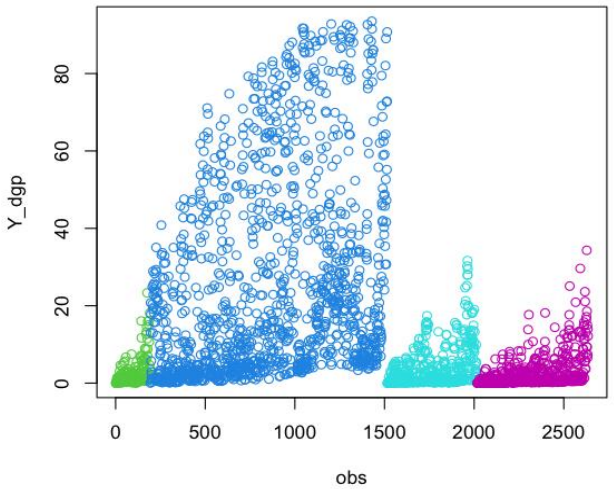


Mise en place d'une expérience numérique

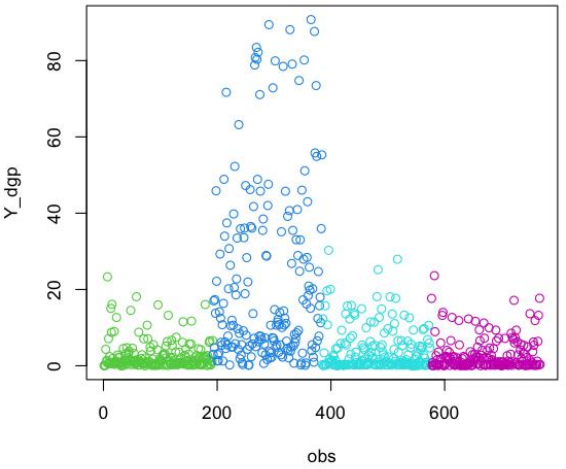


DGP, échantillonnage

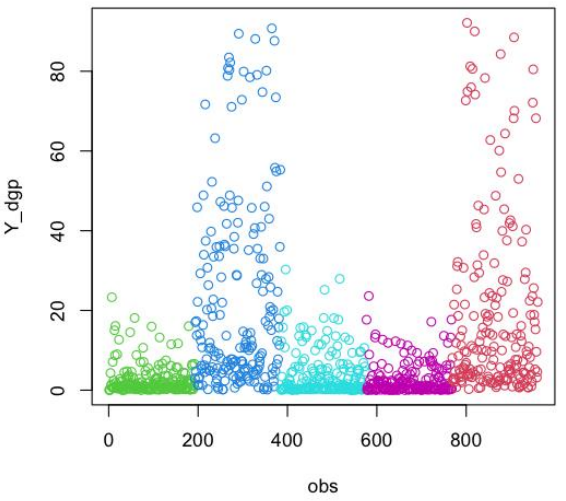
Y_DGP = f(X(SEPIM))
→Modèle de régression logistique
→Prédictions in sample



192 observations par an



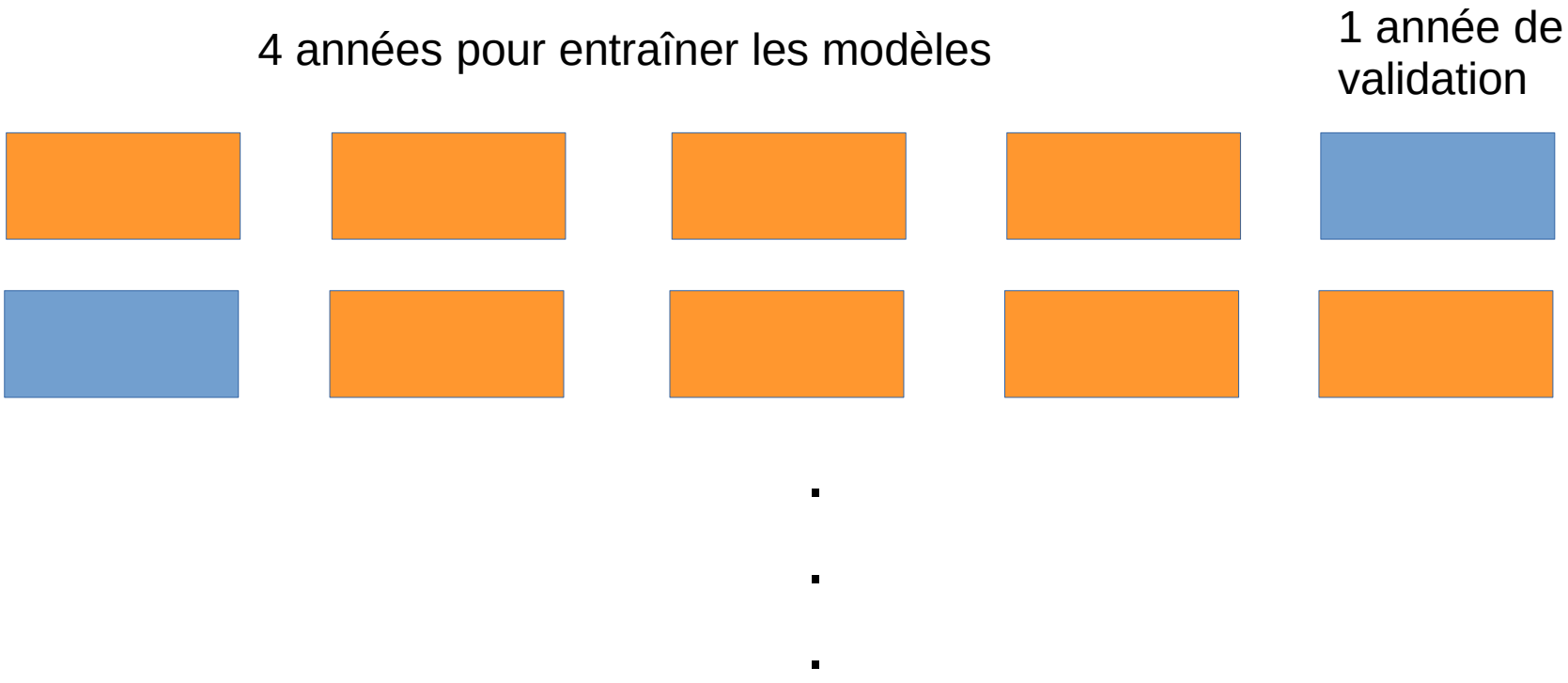
Ajout d'un deuxième tirage de 192 dans 2020



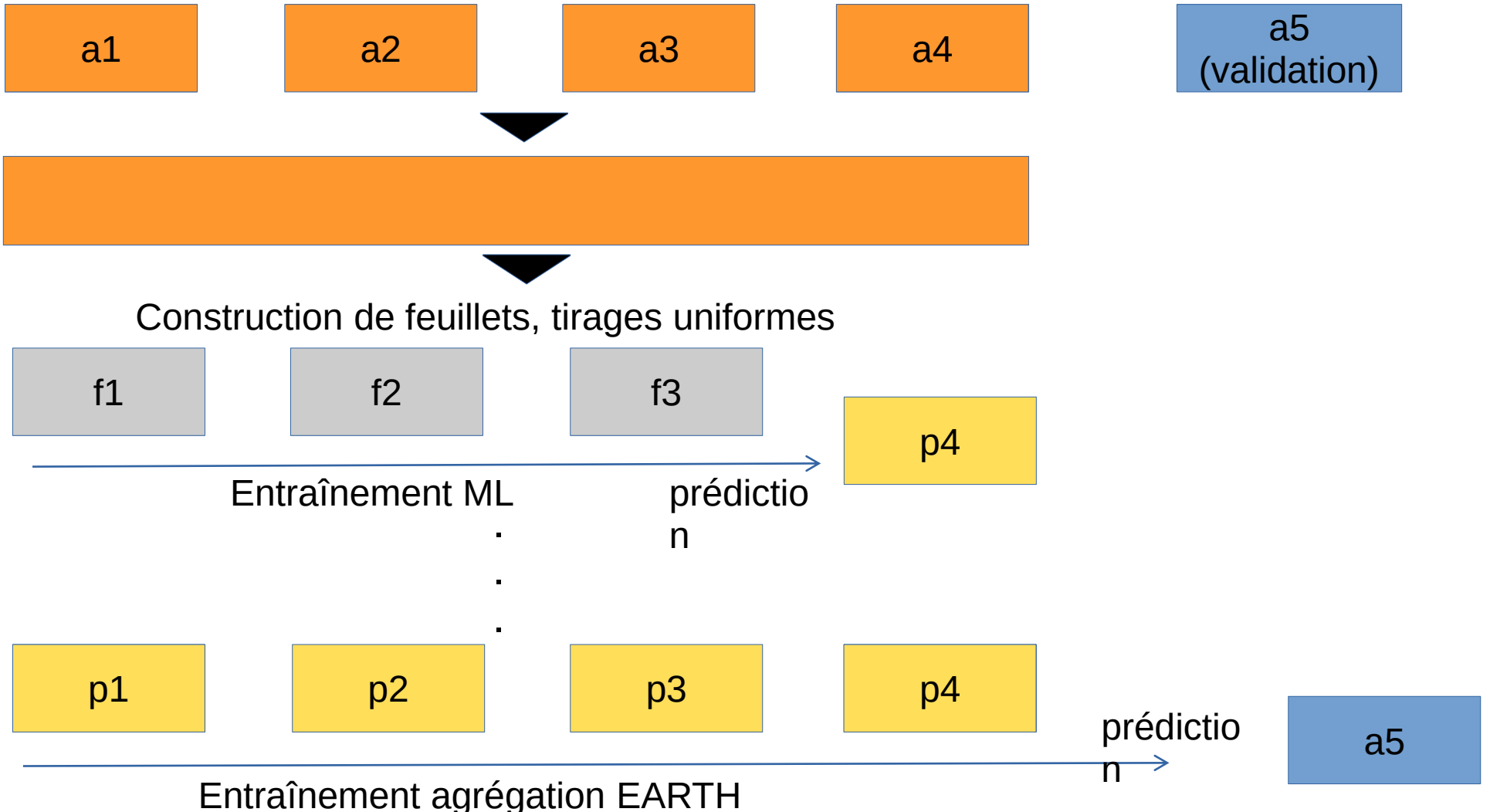
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}
$Y_{1,1}$																	
...																	
$Y_{1,192}$																	
$Y_{2,1}$																	
...																	
$Y_{2,192}$																	
$Y_{3,1}$																	
...																	
$Y_{3,192}$																	
$Y_{4,1}$																	
...																	
$Y_{4,192}$																	
$Y_{5,1}$																	
...																	

Cross-validation

Données sur 5 années qui permettront de définir les feuillets



Entraînement de la méthode de régression pour l'agrégation



Mesure des performances

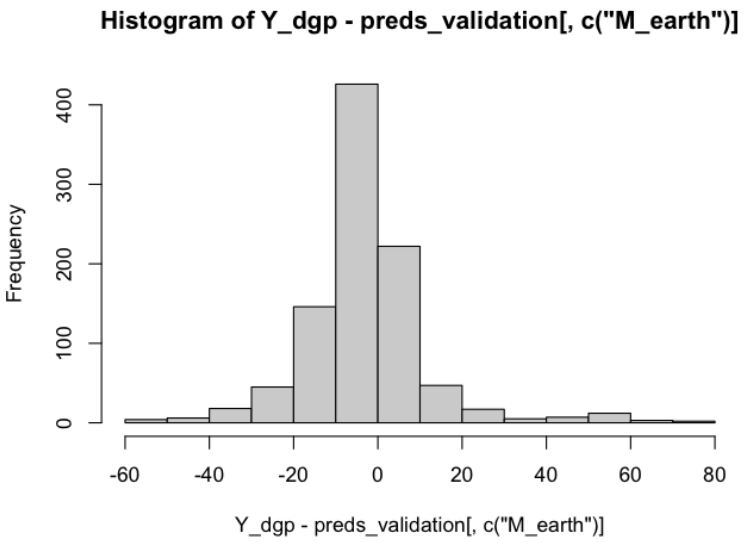
Mean absolute error (MAE) par rapport au DGP, sur l'ensemble des données

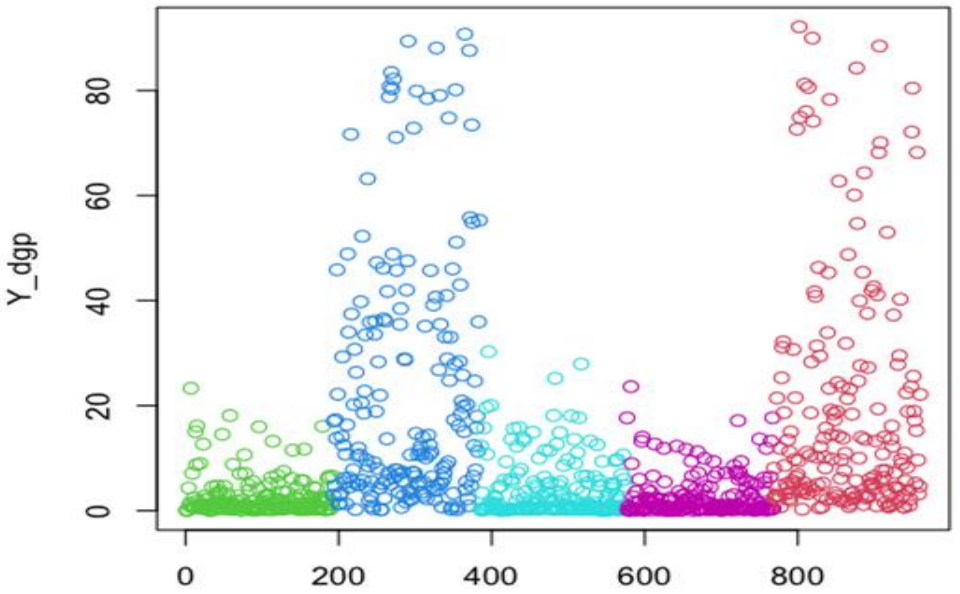
Pour des infections > 20 %
TPR = TP/P

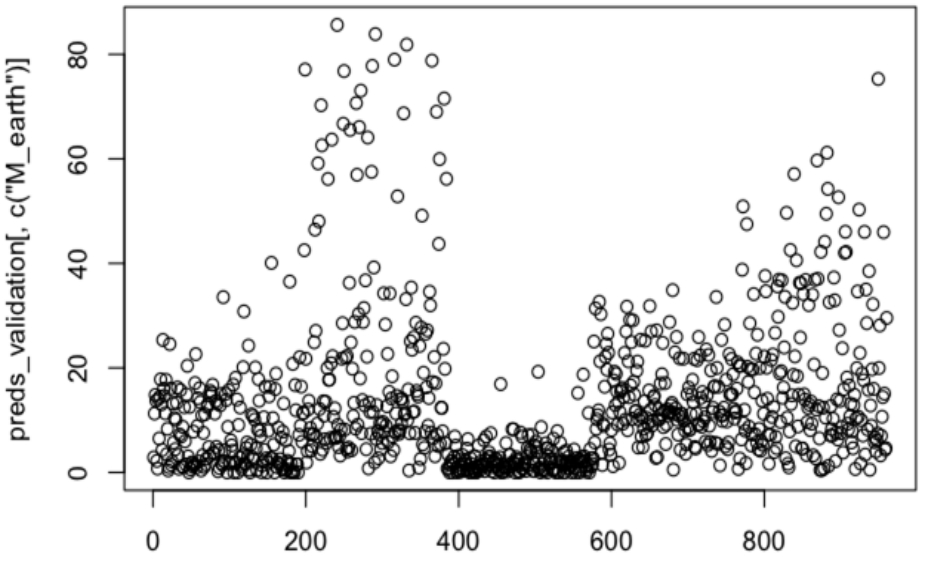
TPR pour des infections < 3 %

Expérience :

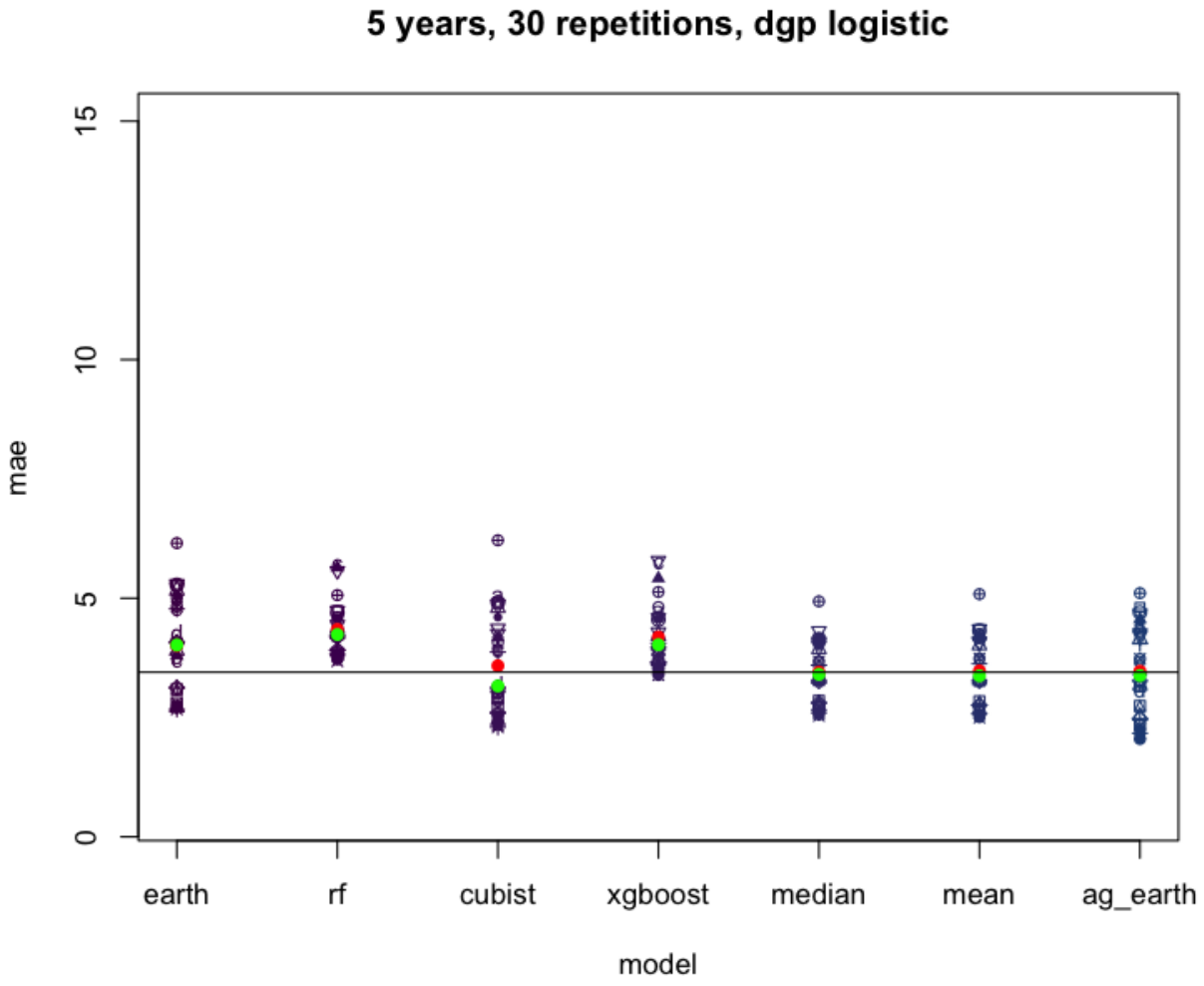
5 échantillons : 2019,
2020_1, 2021, 2022, 2020_2



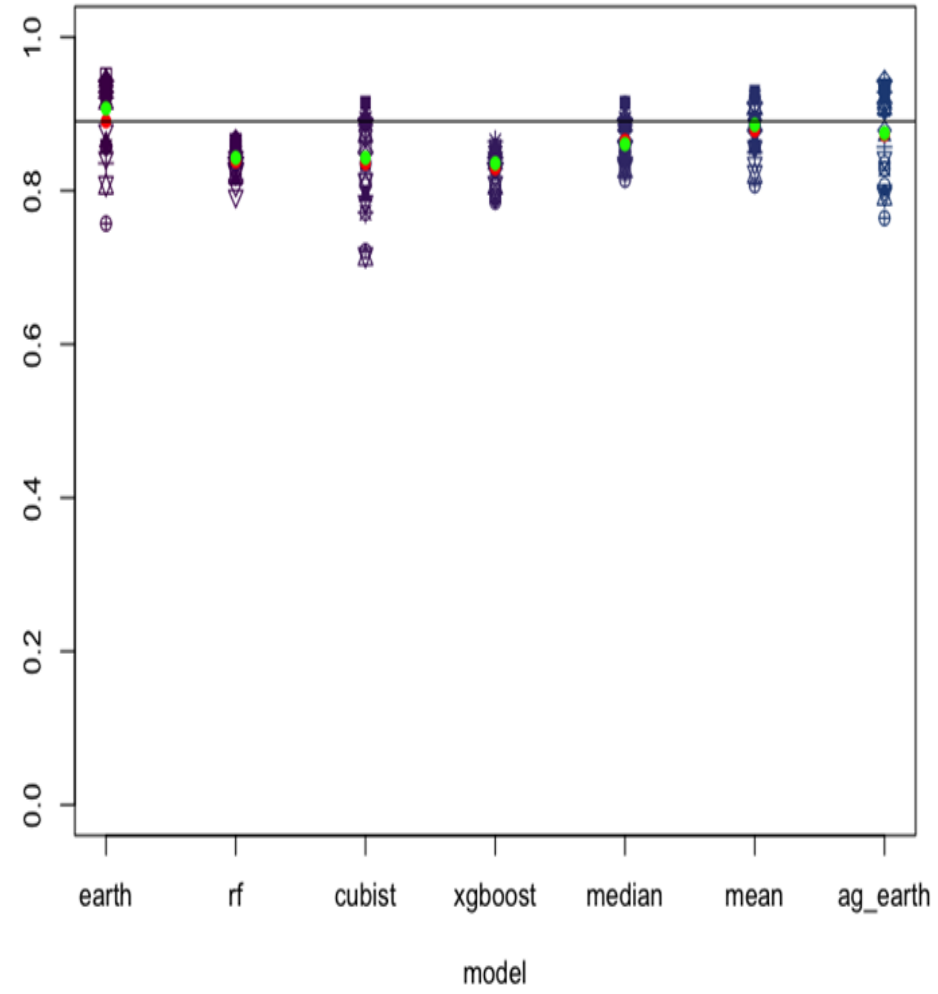




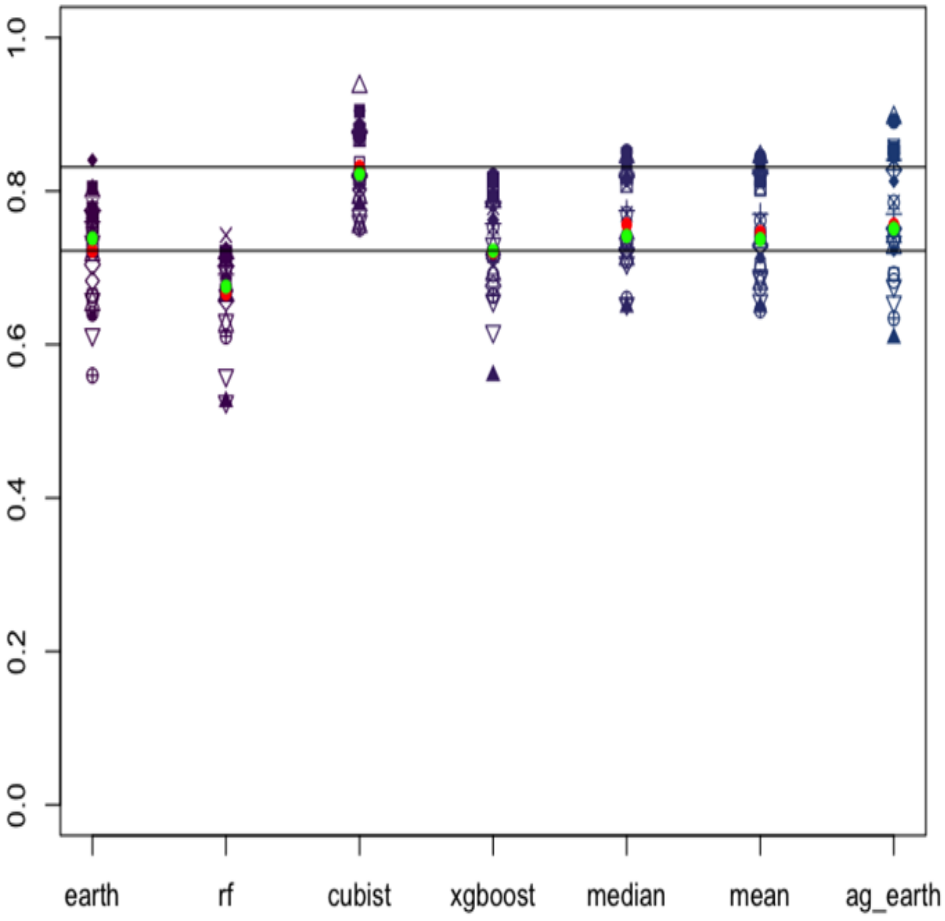
Expérience :
mae

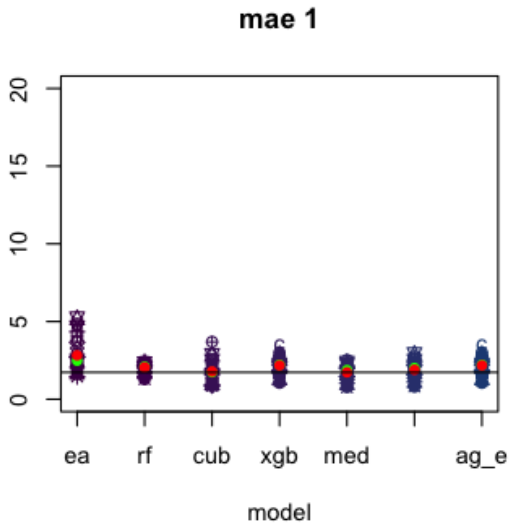


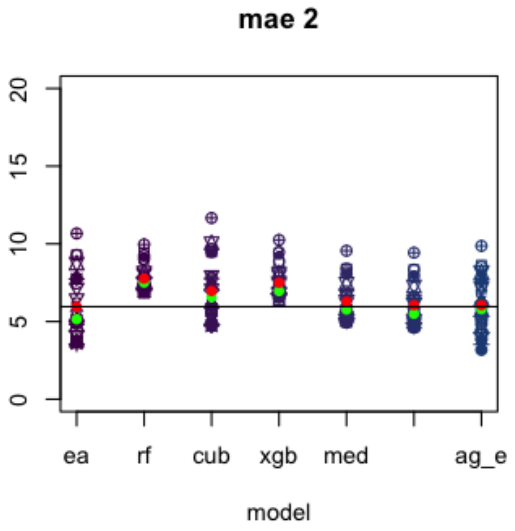
>20 % bien identifiés

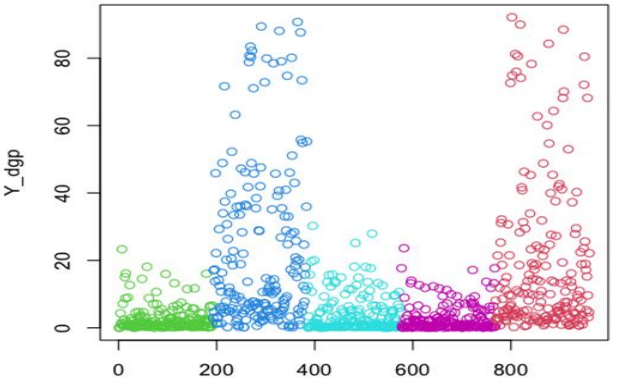


<3 % bien identifiés

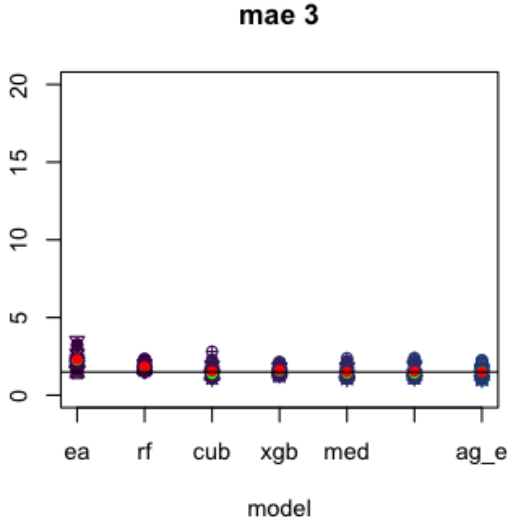


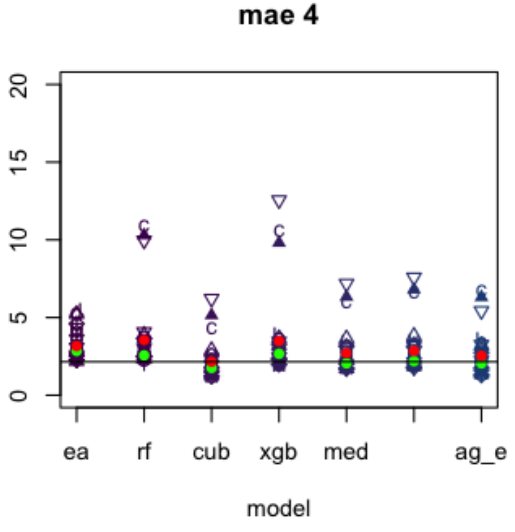


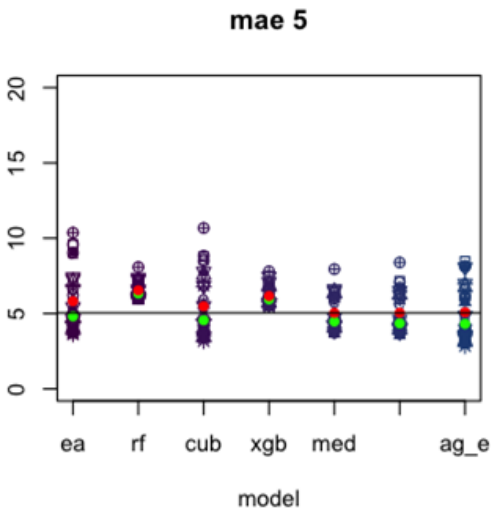




Observations dgp







Conclusion

Expérience préliminaire

1) Tenir compte des effets d'échantillonnage
2) Choisir des métriques adaptées à la question
3) Agrégation de modèles intéressante dans le cas étudié

Pour poursuivre :

Finesse des méthodes d'agrégation :
→ comparaison des méthodes
→ prendre en compte des variations spatiales

Enrichir le jeu de modèles :
Couplage des modèles statistiques avec des modèles « physiques »
→ date d'arrivée des pucerons
→ modèle de développement de la betterave
→ modèle de propagation épidémique (spatio-temporel)

MERCI POUR VOTRE ATTENTION !