INSA – GM 5 Année 2018-2019
ANTOUN Cyril
ROUFFIAC Jean-Eudes

Régression Non Linéaire avec Applications sous R

Rapport de TP n°8

Titre : *CART et Forêts Aléatoires*

L'objectif de ce TP est d'expérimenter quelques techniques statistiques dans le cadre des modèles de régression par arbres CART et des forêts aléatoires.
On souhaite expliquer la pollution par les particules fines PM10 à l'aide des données météorologiques et de mesures de polluants primaires. Les données sont de deux types : des concentrations moyennes journalières de polluants (PM10, NO, NO2 et SO2) et des mesures de paramètres météorologiques (température, pression atmosphérique, humidité relative, vitesse et direction de vent, gradient température). La variable a expliquer est la concentration moyenne journalière en PM10. Les variables explicatives sont :

- des polluants : NO, NO2, SO2 ;
- des variables météorologiques : température (T.min, T.moy et T.max), vitesse de vent (VV.moy et VV.max), humidité relative (HR.min, HR.moy et HR.max), pression atmosphérique (PA.moy), direction de vent (DV.maxvv et DV.dom), gradient de température (GTrouen et GTlehavre), quantité de pluie (PL.som).

Après avoir éliminé les données manquantes, nous construisons un échantillon d'apprentissage constitué de 80% des données et un échantillon test constitué des 20% restantes. Nous souhaitons alors construire l'arbre de régression CART maximal sur notre échantillon d'apprentissage.

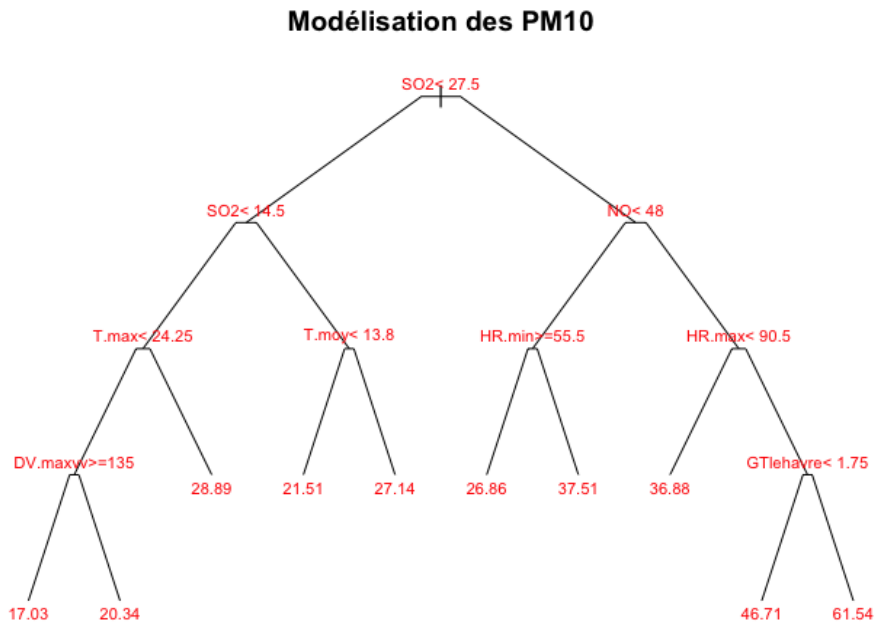


FIGURE 1 – Arbre obtenu grâce au logiciel R

L'arbre ci-dessus qu'on notera "Amax" est obtenu grâce à la commande de R sans modification des paramètres. Cet arbre possède uniquement 10 feuilles, il n'est pas très touffu. De plus, SO2 est la variable la plus importante.

Nous allons maintenant tester les performances de ce modèle à l'aide des fonctions données dans l'énoncé du TP. Dans un premier temps nous calculons les indicateurs d'erreurs classiques en estimation sur notre échantillon d'apprentissage et en prévision sur notre échantillon test. Les résultats sont représentés dans le tableau ci-dessous.

Erreurs	Estimation	Prévision
R	0.82	0.71
EV	0.67	0.47
MAE	4.24	4.38
RMSE	5.68	5.95

Interprétation des résultats :

- R : représente la corrélation entre les valeurs observées et celles estimées/prévues. En estimation, la valeur est de 0,82 pour l'échantillon d'apprentissage. Elle est de 0.71 pour l'échantillon test en prévision. Ces valeurs sont corrects, mais notons tout de même une légère dégradation des performances lorsqu'on passe de l'échantillon d'apprentissage à l'échantillon test ;
- MAE : représente la moyenne de l'erreur absolue. On obtient 4.24 en estimation et 4,38 en prévision. Ce sont des valeurs assez faibles ;
- RMSE : représente la racine de l'erreur quadratique moyenne. Nous avons une valeur de 5.68 en estimation et 5,94 en prévision, ce qui est relativement faible.

Construisons maintenant les tableaux des dépassements et les performances en prévisions qui s'en suivent pour l'estimation de notre échantillon d'apprentissage et la prévision de notre échantillon test.

Tableau des dépassements	Estimation			Prévision		
	Niveau 0	Niveau 1	Niveau 2	Niveau 0	Niveau 1	Niveau 2
Niveau 0	653	6	0	161	1	0
Niveau 1	45	41	2	16	12	1
Niveau 2	1	7	11	0	0	1

Performances	Estimation	Prévision
POD	0.57	0.47
FAR	0.09	0.07
TS	0.54	0.45
SI	0.56	0.46

Nous remarquons grâce aux tableaux de dépassements que le modèle sous-estime légèrement mais reste globalement bon. Pour ce qui est des performances en prévisions :

- POD : représente le taux de bonne détection. Pour ce modèle nous avons un taux de 57% pour l'estimation des données d'apprentissage et 47% pour la prévision des données test. Ces résultats ne sont pas très satisfaisants et encore une fois la légère dégradation des données est mise en évidence lors du passage de l'échantillon d'apprentissage à l'échantillon test ;
- FAR : correspond au taux de fausses alarmes. Un taux de 7% est une valeur raisonnable pour ce modèle ;
- TS : correspond au Threat Score. Ce taux à 47% est relativement moyen.

La dernière fonction permettant d'évaluer les performances du modèle nous donne les graphes observé/estimé et observé/prévu suivant :

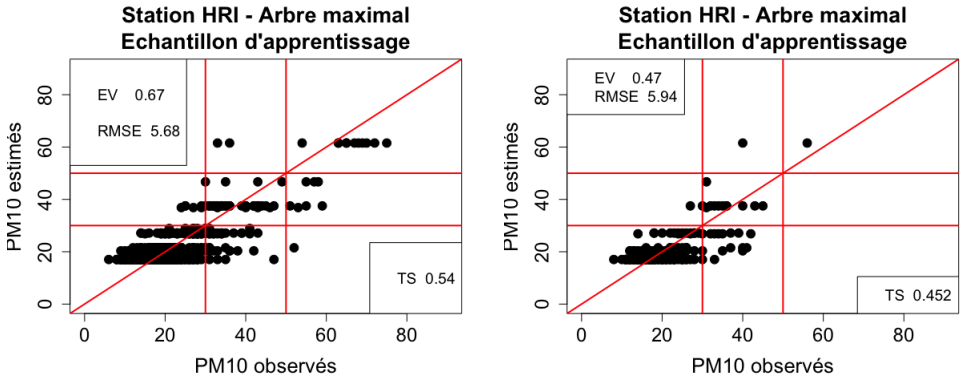


FIGURE 2 – Graphes observé/estimé (à gauche) et observé/prévu (à droite)

Nous voyons ainsi que le modèle a tendance à surestimer les faibles valeurs et à sous-estimer les fortes concentrations.
En conclusion, nous pouvons dire que ce modèle est correct. Nous verrons dans la suite du TP si il est possible d'obtenir de meilleures performances, notamment avec un arbre plus touffu.

Nous cherchons maintenant à comparer ce modèle avec deux autres arbres : l'un plus touffu, que l'on notera "Atouf" et un autre plus élagué, noté "Aela".
Tout d'abord, nous allons construire l'arbre élagué de "Amax". Pour ce faire, nous allons nous baser sur le critère du CP de Mallows. Le graphique représentant l'évolution du Cp de Mallows en fonction du nombre de feuilles, montre que la courbe atteint un minimum aux alentours de 0,021. Le graphique permettant d'obtenir cette valeur est présent dans les annexes. L'arbre élagué obtenu est le suivant :

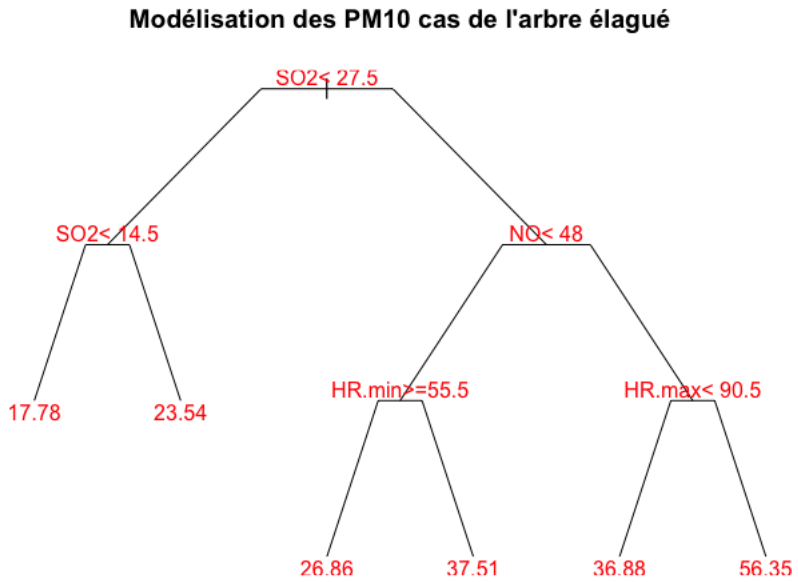


FIGURE 3 – Arbre élagué

Nous obtenons un arbre composé de 6 feuilles. Comme pour le modèle précédent, nous allons étudier les performances en estimation sur l'échantillon d'apprentissage et en prévision sur l'échantillon test. Voici les tableaux de résultat (indicateurs d'erreurs, tableaux de dépassement et indicateurs de performances) ainsi que les graphes observé/estimé et observé/prévu.

Erreurs	Estimation	Prévision
R	0.78	0.69
EV	0.61	0.45
MAE	4.54	4.47
RMSE	6.1	6.08

Tableau des dépassements	Estimation			Prévision		
	Niveau 0	Niveau 1	Niveau 2	Niveau 0	Niveau 1	Niveau 2
Niveau 0	653	6	0	161	1	0
Niveau 1	45	37	6	16	11	2
Niveau 2	1	4	14	0	0	1

Performances	Estimation	Prévision
POD	0.57	0.47
FAR	0.09	0.07
TS	0.54	0.45
SI	0.56	0.46

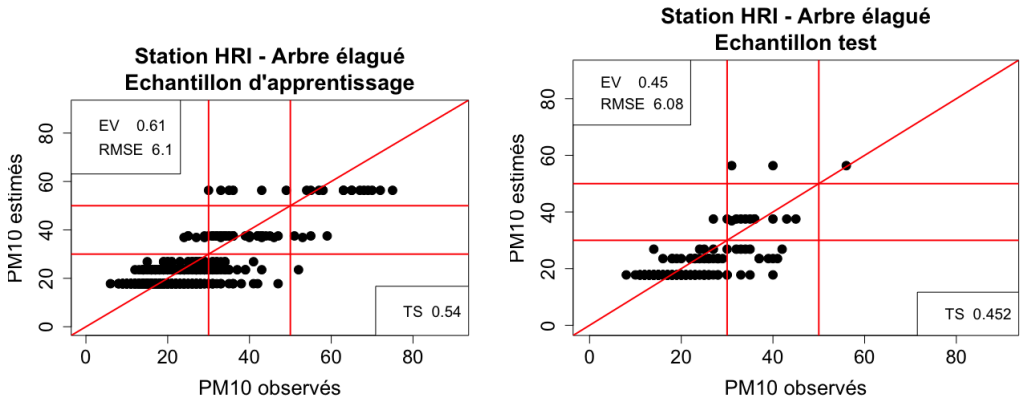


FIGURE 4 – Graphes observé/estimé (à gauche) et observé/prévu (à droite)

Nous remarquons que les résultats sont très légèrement moins bons que les résultats obtenus avec l'arbre maximal. L'indice de corrélation entre les valeurs observées de l'echantillon test et les valeurs prévues est de 0.02 en dessous de celui de l'arbre maximal. Le MAE est de 0.09 et le RMSE de 0.13, ces erreurs sont donc supérieures que celles de l'arbre maximal. De plus les graphes observés/estimés et observés/prévus indiquent aussi une surestimation des faibles valeurs et une sous-estimation des fortes concentrations. L'arbre maximal n'étant pas très touffu, il n'est pas vraiment nécessaire de l'élaguer.

Nous allons maintenant nous intéresser à la construction d'un arbre plus touffu pour voir si les performances du modèle sont meilleures que pour notre arbre maximal. Après modifications des paramètres de la fonction *rpart* de R, nous avons pu obtenir un arbre touffu. L'arbre contient un très grand nombre de feuilles et n'est pas lisible. Nous avons donc décidé de limiter le nombre de feuilles dans un soucis de lisibilité, nous obtenons l'arbre suivant :

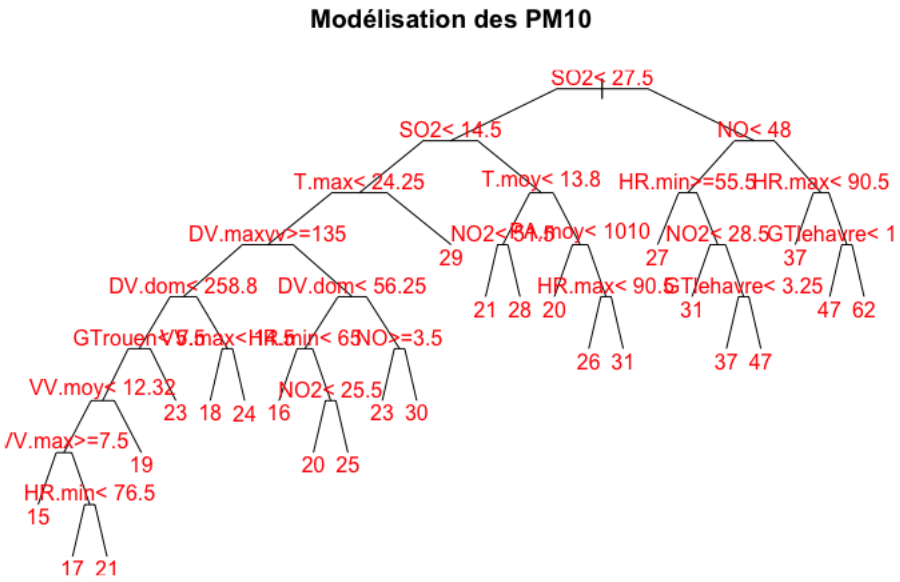


FIGURE 5 – Arbre touffu

Afin de comparer les différents arbres, nous présenterons les résultats dans un tableau. Le premier avec les performances des modèles en estimation sur l'échantillon d'apprentissage et le second en prévision sur l'échantillon test.

Estimation	R	MAE	RMSE	POD	FAR	TS	Nb feuilles
Amax	0,82	4,24	5,68	0,57	0,09	0,54	10
Aela	0,78	4,54	6,1	0,57	0,09	0,54	6
Atouf	0,86	3.72	4.99	0,67	0,16	0,6	26

Prevision	R	MAE	RMSE	POD	FAR	TS	Nb feuilles
Amax	0,71	4,38	5,95	0,47	0,07	0,45	10
Aela	0,69	4,47	6,08	0,47	0,07	0,45	6
Atouf	0,71	4,65	6,17	0,5	0,32	0,41	26

Nous mettons en annexe les graphes observé/estimé et observé/prévu pour l'arbre touffu.
Nous remarquons que l'abre touffu a de bons résultats en estimation sur l'échantillon d'apprentissage mais des résultats équivalents aux autres modèles en prévision sur l'échantillon test. La dégradation des données est plus significative avec ce modèle lors du passage de l'échantillon d'apprentissage à l'échantillon test. Pour conclure, nous voyons que les

résultats sont assez proches en prévision malgré un nombre de feuilles différent. Cependant un arbre avec peu de feuilles est plus simple à comprendre.

Nous allons maintenant étudier à l'aide des forêts aléatoires l'importance des différentes variables explicatives et leur effet marginal sur les PM10. Dans un premier temps, nous cherchons à comprendre l'importance des différentes variables. On peut représenter les résultats à l'aide d'un graphique. Étant donné que les résultats peuvent fluctuer, nous avons relancé plusieurs fois le calcul et effectué la moyenne sur les résultats.

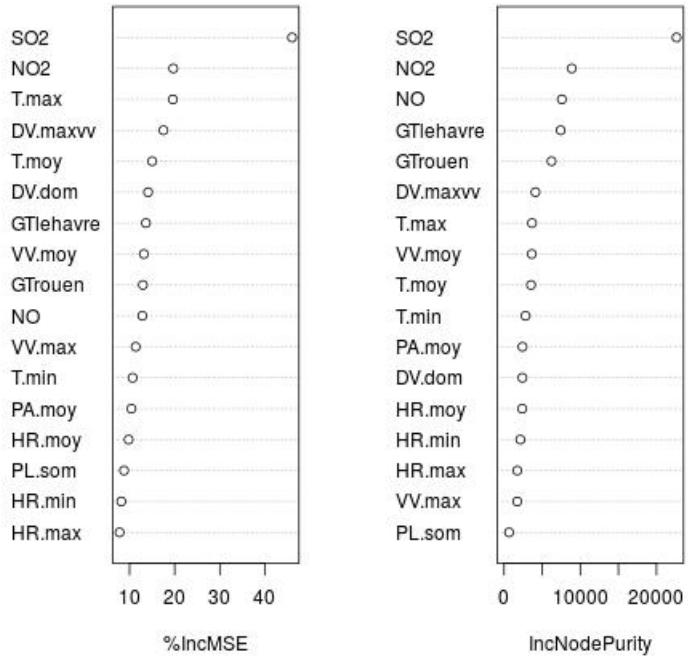


FIGURE 6 – Importance des variables par les Forets Aleatoires

On constate que les polluants sont les régresseurs les plus importants. On cherche maintenant à regarder l'effet de chacune des variables sur la pollution par PM10. On ne conserve qu'une seule des variables explicatives représentant la même donnée (ex : T.moy pour la température). La méthode de sélection est la suivante : on sélectionne la variable ayant une valeur de IncNodePurity la plus élevée.

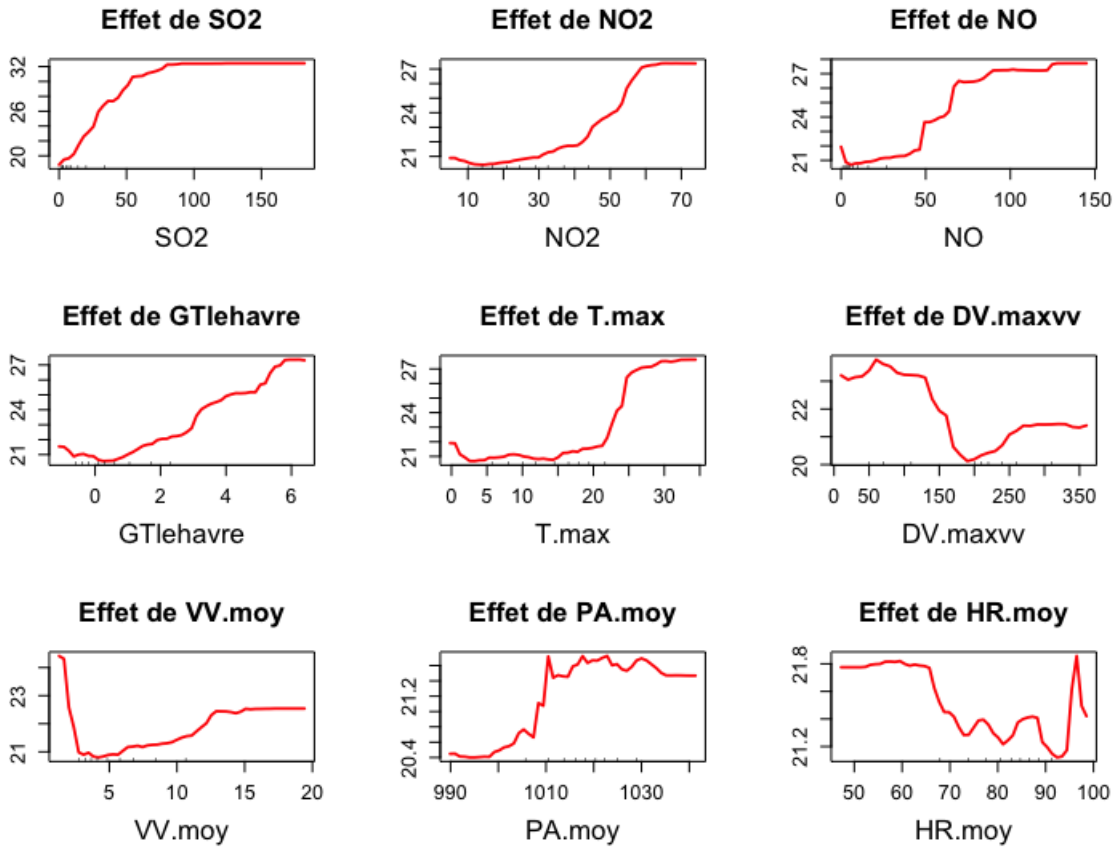


FIGURE 7 – Effet des variables aléatoires

On constate que les trois polluants, la température, le gradient de température au Havre et la pression atmosphérique moyenne ont un effet marginal estimé croissant sur les PM10. En revanche l'humidité relative, la direction du vent et la vitesse du vent ont un effet décroissant puis croissant.

Nous allons maintenant étudier les performances de notre modèle issu de la forêt construite avec nos 9 variables les plus importantes : SO2, NO2, NO, GTlehavre, T.max, DV.maxvv, VV.moy, PA.moy et HR.moy. Nous comparerons les résultats avec un modèle linéaire et un modèle additif généralisé construits avec les neuf variables précédentes après s'être assuré que les variables ne sont pas corrélées deux à deux et après une étude de multicolinéarité avec l'indicateur VIF.
Les résultats sont regroupés dans les deux tableaux ci-dessous. L'un pour les performances en estimation sur l'échantillon d'apprentissage et le deuxième sur les performances en prévision sur l'échantillon test.

Estimation	R	MAE	RMSE	POD	FAR	TS
Forêt	0,81	4,14	5,83	0,62	0,18	0,55
lm	0,75	4,79	6,51	0,59	0,24	0,5
gam	0,82	4,09	5,57	0,64	0,19	0,56

Prevision	R	MAE	RMSE	POD	FAR	TS
Forêt	0,78	3,79	5,17	0,57	0,29	0,47
lm	0,74	4,25	5,58	0,60	0,28	0,56
gam	0,76	4,10	5,57	0,60	0,31	0,47

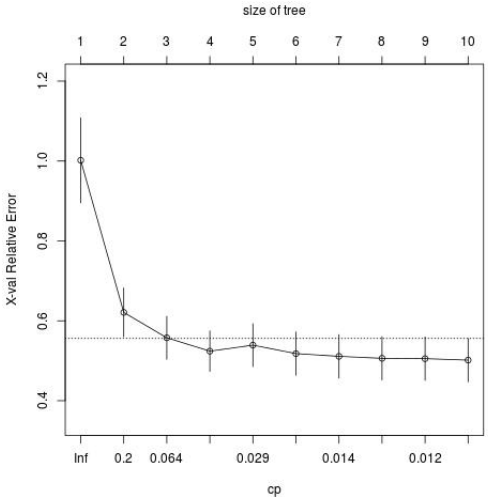
Tous les graphes observés/estimés et observés/prévus se trouvent en annexe. On constate que pour tous les graphes, il y a une surestimation des faibles valeurs et une sous-estimation des fortes concentrations.

D'après les résultats trouvés, nous remarquons que le modèle issu de la forêt construite a les meilleurs résultats. Il minimise le RMSE et le MAE, et il maximimse l'indicateur de corrélation entre les valeurs observées et les valeurs prédites. De plus il n'y a presque pas de dégradation des données lors du passage de l'échantillon d'apprentissage à l'échantillon test. Les résultats sont aussi meilleurs que ceux trouvés pour l'arbre maximal précedemment. Le modèle linéaire obtient de légèrement moins bons résultats. Cependant il met en jeu deux fois moins de variables explicatives que pour l'arbre max. Et enfin, le modèle gam obtient des résultats légèrement moins bons que ceux obtenus avec la forêt construite mais de meilleurs résultats que ceux obtenus avec l'arbre maximal, et lui aussi met en jeu seulement 9 variables explicatives. On peut alors se demander si l'utilisation du modèle gam ne serait pas finalement le plus adapté ici, car il est relativement simple à mettre en oeuvre et n'utilise que 9 variables explicatives.

Annexes

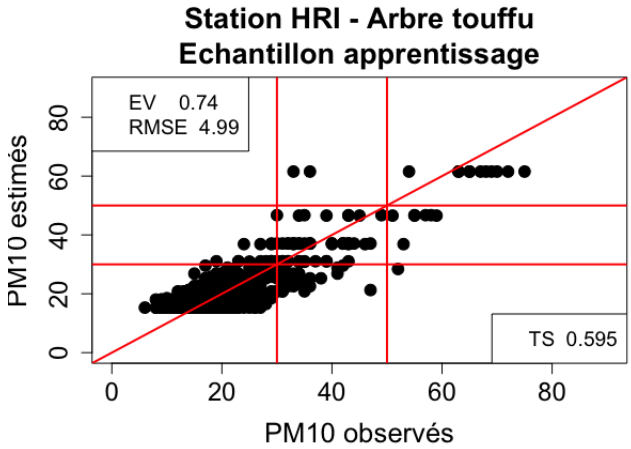
Arbre maximal

Graphique permettant d'obtenir le CP de Mallows :

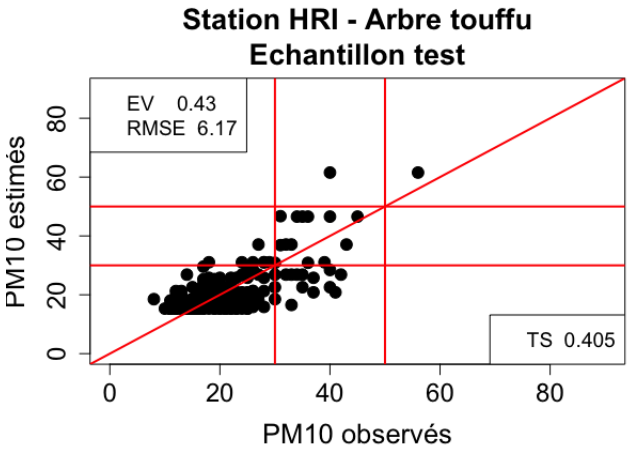


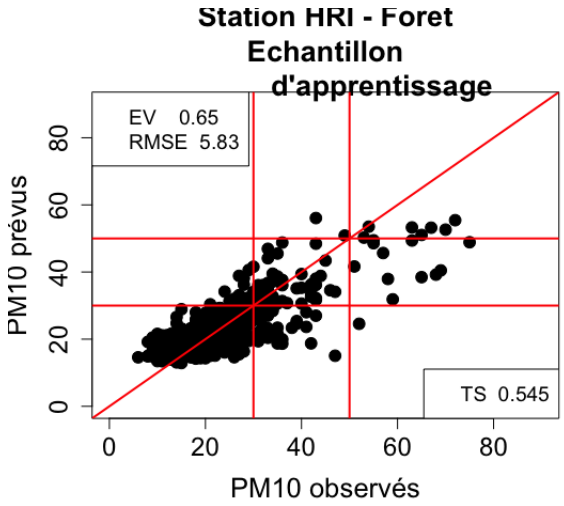
Arbre Touffu

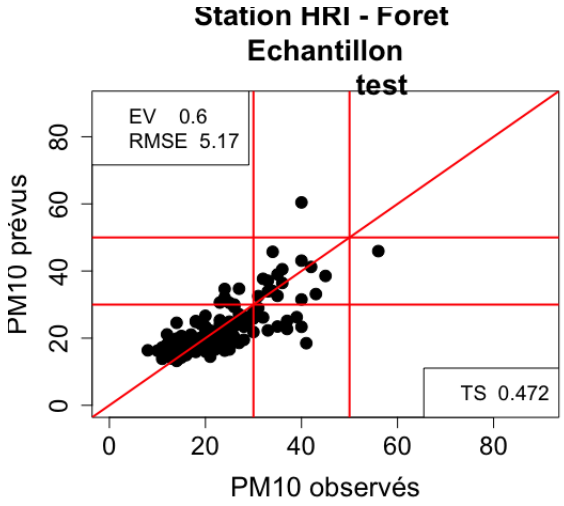
Graphe Observé/Estimé

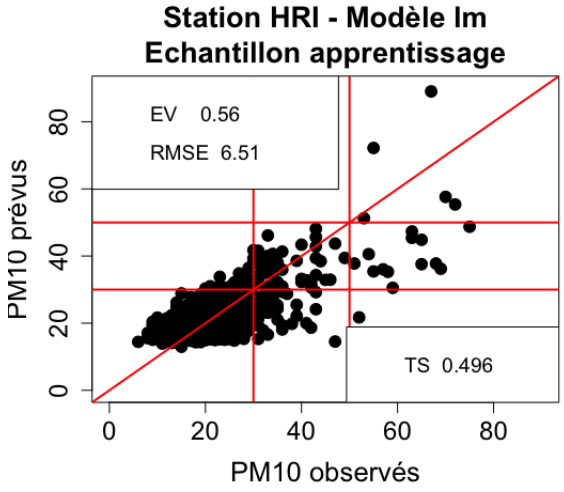


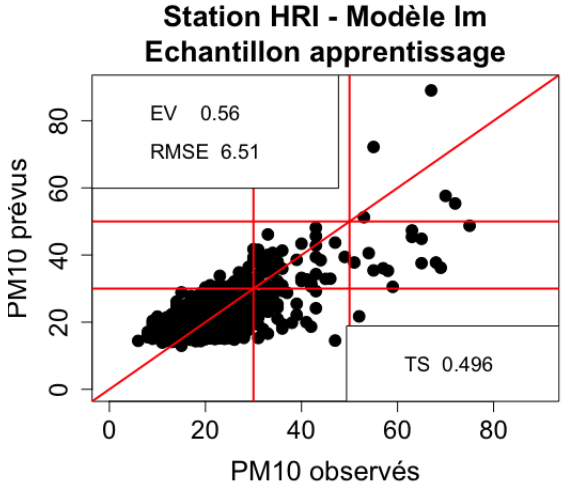
Graphe Observé/Prévu

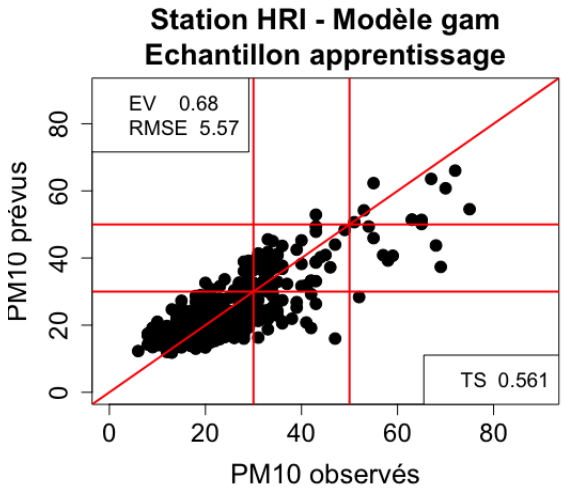

Station HRI - Arbre touffu
Echantillon test
EV 0.43
RMSE 6.17
TS 0.405
PM10 estimés
PM10 observés
0
20
40
60
80

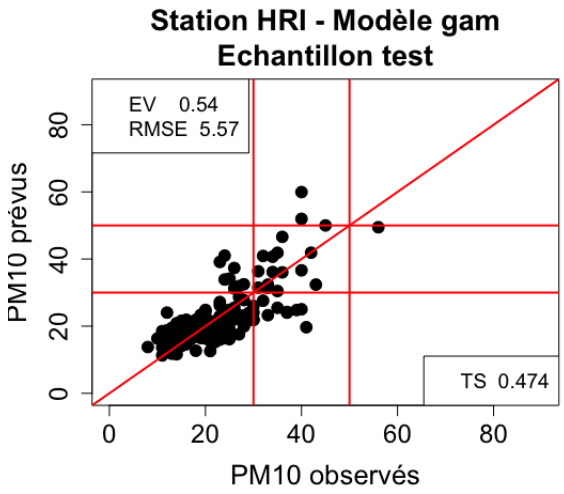

Station HRI - Foret
Echantillon
d'apprentissage
EV 0.65
RMSE 5.83
TS 0.545
PM10 prévus
PM10 observés
0
20
40
60
80


Station HRI - Foret
Echantillon
test
EV 0.6
RMSE 5.17
TS 0.472
PM10 prévus
PM10 observés
0
20
40
60
80


Station HRI - Modèle lm
Echantillon apprentissage
EV 0.56
RMSE 6.51
TS 0.496
PM10 prévus
PM10 observés
0
20
40
60
80


Station HRI - Modèle lm
Echantillon apprentissage
EV 0.56
RMSE 6.51
TS 0.496
PM10 prévus
PM10 observés
0
20
40
60
80


Station HRI - Modèle gam
Echantillon apprentissage
EV 0.68
RMSE 5.57
TS 0.561
PM10 prévus
PM10 observés
0
20
40
60
80


Station HRI - Modèle gam
Echantillon test
EV 0.54
RMSE 5.57
TS 0.474
PM10 prévus
PM10 observés
0
20
40
60
80