

# RAPPORT FINAL

APPRENTISSAGE STATISTIQUE EN ACTUARIAT  
ACT-4114

ÉQUIPE 09

---

## Rapport Inondations en Californie

---

*Par*

Maryjane BASTILLE  
Danny LAROCHELLE  
Henri LEBEL  
ISABELLE LEGENDRE  
Félix-Antoine PARIS

*Numéro d'identification*

111 268 504  
111 174 586  
111 286 185  
536 768 666  
536 776 223

*Travail présenté à  
Monsieur*

OLIVIER CÔTÉ

16 AVRIL 2023



UNIVERSITÉ  
LAVAL

**Faculté des sciences et de génie**  
École d'actuariat

## Table des Matières

<b>Introduction</b>	<b>2</b>
<b>Modèle de base</b>	<b>3</b>
<b>Ajustement des modèles</b>	<b>4</b>
Modèle linéaire (À spécifier) . . . . .	5
Modèle des k plus proches voisins . . . . .	6
Arbre de décision . . . . .	7
Bagging . . . . .	8
Forêt aléatoire . . . . .	9
Boosting . . . . .	11
Gradient Boosting . . . . .	11
Extreme gradient boosting . . . . .	11
<b>Comparaison des modèles</b>	<b>12</b>
<b>Interprétation des meilleurs modèles</b>	<b>13</b>
<b>Conclusion</b>	<b>14</b>
<b>Bibliographie</b>	<b>15</b>

## **Introduction**

## Modèle de base

## Ajustement des modèles

**Modèle linéaire (À spécifier)**

## **Modèle des $k$ plus proches voisins**

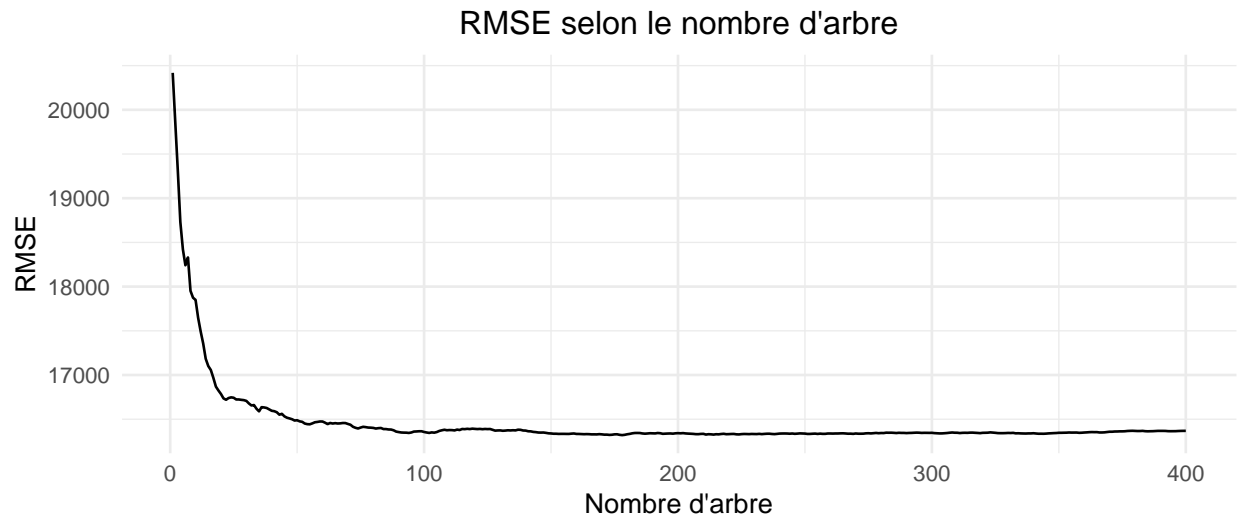
## **Arbre de décision**



## **Bagging**

## Forêt aléatoire

Pour la forêt aléatoire, on commence avec quatre prédicteurs possibles pour chaque séparation, *i.e.*  $m = 4$ , car  $\lfloor 12/3 \rfloor = 4$ . Cette valeur correspond à la “règle du pouce” en régression où l’on utilise la partie entière du nombre de valeurs explicatives divisé par 3. Deplus, en utilisant une proportion de 50% pour les échantillons bootstrap, on aide à diminuer la corrélation entre les arbres.



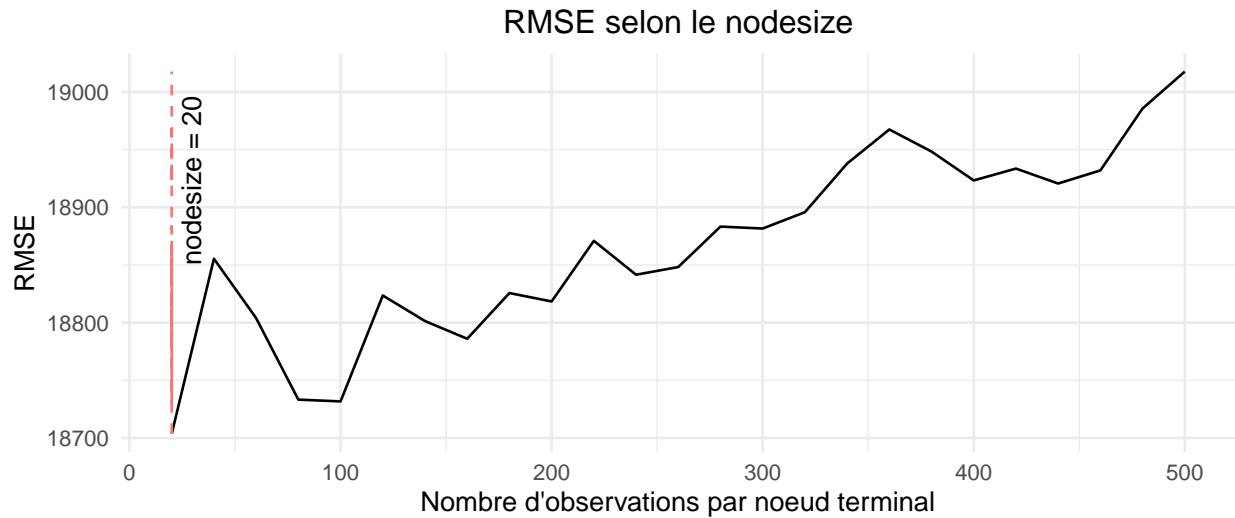
Étant en régression, la racine de l’erreur quadratique moyenne, ou RMSE, sera utilisée comme mesure de comparaison. On remarque ici (Graphique no. #) que la RMSE se stabilise aux alentours de 100-150 arbres, on utilisera alors 200 arbres pour l’optimisation des autres hyperparamètres, puisqu’on ne peut pas surajuster en ayant trop d’arbre avec les forêts aléatoires. Maintenant, on regarde plus en profondeur le nombre de prédicteurs possible à chaque séparation d’un arbre, la variable `mtry`.

Table 1: RMSE par rapport au `mtry`

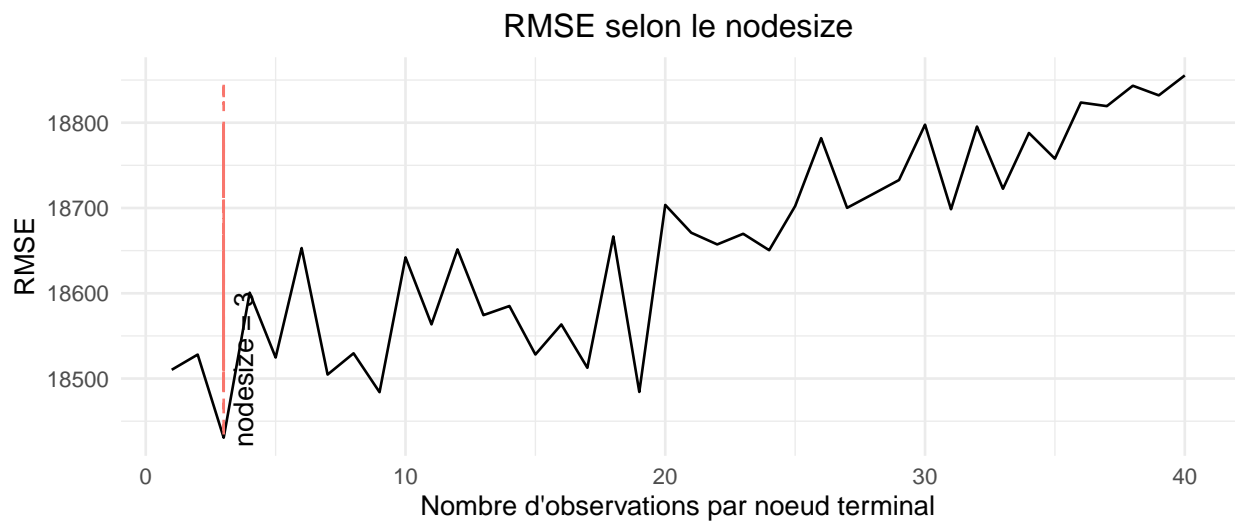
mtry	1	2	3	4	5	6	7	8	9	10	11	12
RMSE	19176	19141	19074	18920	18714	18568	18433	18331	18236	18126	18148	18042

Les résultats de la table no. # ont été obtenus par validation croisée à 5 plis, pour ainsi réduire le biais d’échantillonnage. L’utilisation des 12 choix de variables explicatives à chaque noeud minimise la RMSE.

Pour éviter un surajustement dû à des arbres inutilement trop profonds, on devra ajuster la valeur de `nodesize`, mais il est impossible de le faire directement avec le package `caret`. Puisque le modèle est entraîné sur 8325 observations, les valeurs de 500 et moins seront testées et comparées. Pour limiter le temps de calcul, un premier entraînement sera fait par bond de 20.



Dans le graphique no. #, la valeur minimale de `nodesize` est de 20. Puisque l'analyse précédente a été effectuée par bonds de 20, on la fera à nouveau de manière plus précise entre 1 et 40.



Dans le graphique no. #, la valeur minimale de `nodesize` est de 3.

Par conséquent, les hyperparamètres finaux pour le modèle “Forêt aléatoire” sont ceux décrits dans la table suivante.

Table 2: Valeurs des hyperparamètres du modèle final

Hyperparamètre	Valeur
Nombre d'arbres	200
Nombre de choix de variables à chaque noeud	12
Nombre d'observation dans les noeuds terminaux	3

## **Boosting**

**Gradient Boosting**

**Extreme gradient boosting**

## Comparaison des modèles

## Interprétation des meilleurs modèles

## Conclusion

## **Bibliographie**

The Federal Emergency Management Agency (2023). FIMA NFIP Redacted Claims - v1.

Récupéré de <https://www.fema.gov/openfema-data-page/fima-nfip-redacted-claims-v1>