

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

Introduction	2
Modèle de base	3
Ajustement des modèles	5
Modèle des k plus proches voisins	5
Arbre de décision	6
<i>Bagging</i>	7
Forêt aléatoire	10
<i>Gradient Boosting</i>	13
Comparaison des modèles	15
Interprétation des meilleurs modèles	16
Conclusion	23
Bibliographie	24

Introduction

Modèle de base

Le choix d'un modèle linéaire généralisé Tweedie est justifié dans notre étude pour plusieurs raisons. Tout d'abord, cette distribution est particulièrement adaptée pour modéliser la fréquence et la sévérité combinées des données de sinistres d'inondation, tout en prenant en compte leur nature asymétrique. De plus, cette distribution est paramétrée, ce qui permet de modéliser des relations complexes entre les variables explicatives et les montants de sinistres. Le modèle Tweedie retenu peut facilement intégrer des variables continues et catégorielles, ce qui est pertinent pour caractériser le jeu de données considéré. Enfin, l'interprétation des résultats est facilitée par la structure du modèle linéaire généralisé Tweedie, permettant de mieux comprendre les effets des variables explicatives sur les montants de sinistres. Ainsi, ce modèle servira de base efficace pour établir la prime de référence, dans le but d'évaluer la performance d'autres modèles.

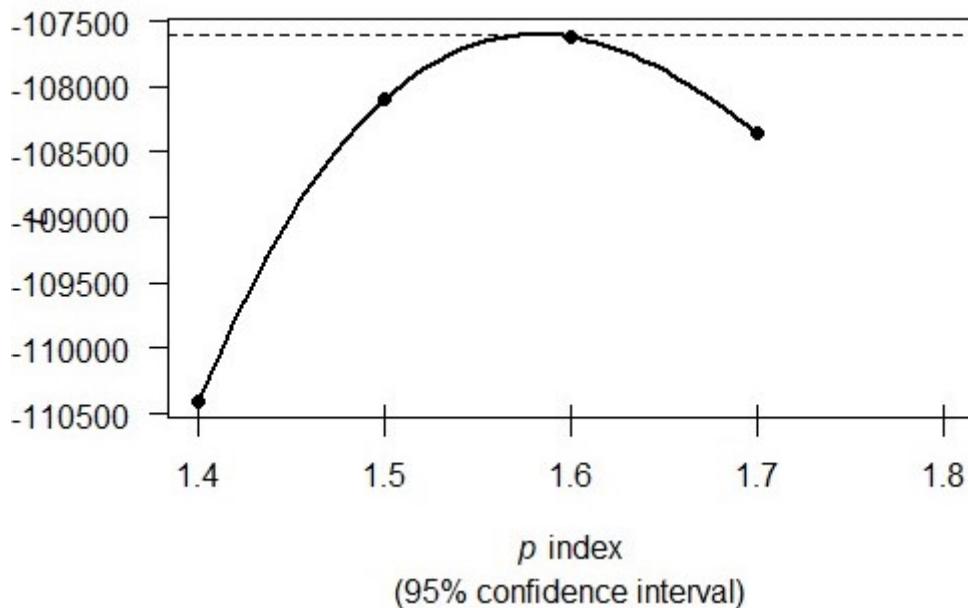


Figure 1: Valeur de p maximal pour Tweedie

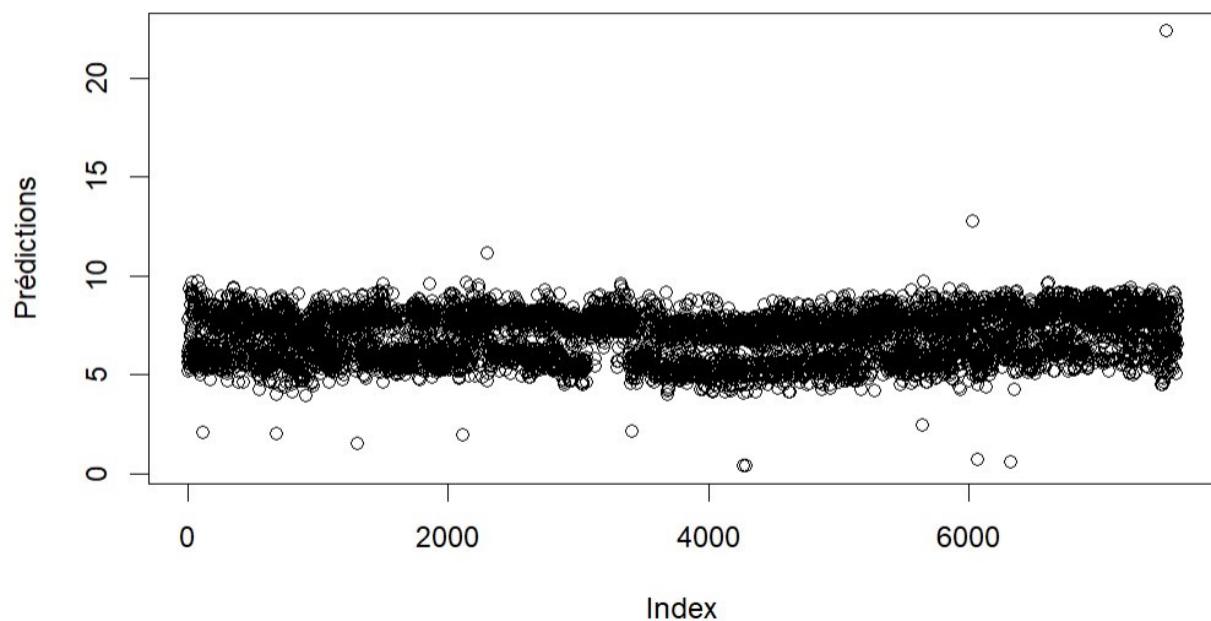


Figure 2: Prédictions selon le GLM Tweedie

Ajustement des modèles

Modèle des k plus proches voisins

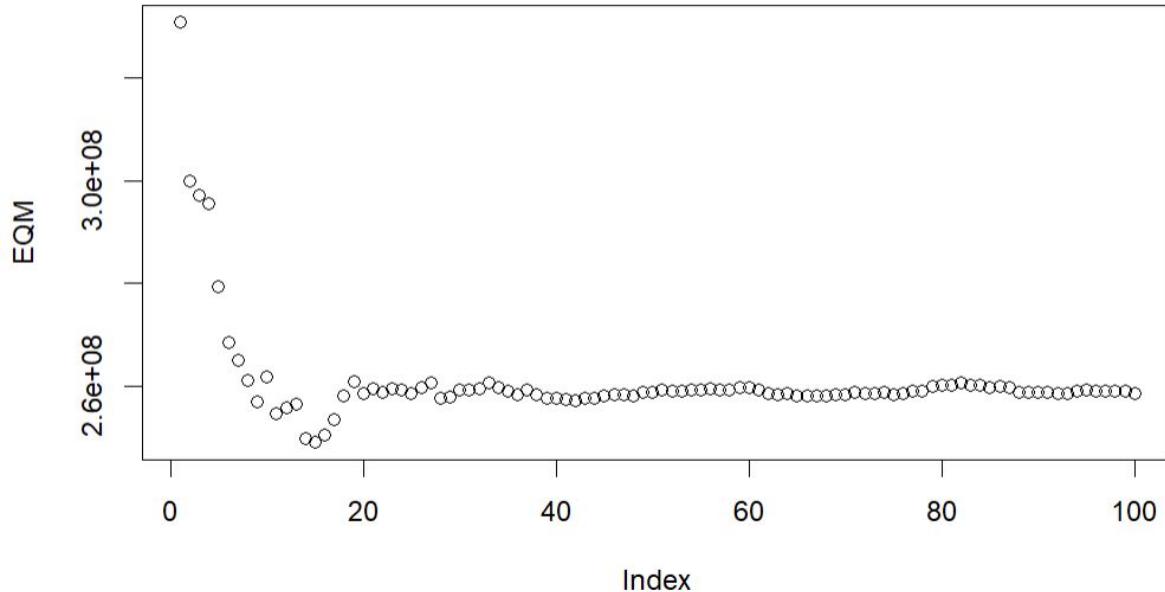


Figure 3: Erreur quadratique moyenne selon le k

Pour construire un modèle de régression k-plus proche voisin, nous avons du tout d'abord transformé nos variables catégorielles en valeurs numériques. Cette étape fut nécessaire, car nous voulions obtenir des chiffres plutôt que des classes. Après cette transformation, nous avons cherché à déterminer le facteur "k" optimal pour obtenir le meilleur modèle possible. Nous avons utilisé la méthode de l'erreur quadratique moyenne (EQM) pour évaluer la performance de notre modèle. Cette mesure est estimée à l'aide de la formule suivante :

$$\frac{1}{n} \sum_{i=1}^n [y_i - \hat{f}(x_i)]^2.$$

Nous avons testé différentes valeurs de k allant de 1 à 100 et avons observé que l'EQM augmente après un certain point. Après avoir analysé le graphique, nous avons constaté que le k optimal pour notre modèle était de 15, car c'est à ce point que l'EQM atteint son minimum.

Arbre de décision

Dans le contexte de la modélisation de montants de réclamation en assurance inondation, il est important de trouver un modèle prédictif à la fois précis et interprétable. Les arbres de décision sont très utiles dans ce contexte, car ils permettent une interprétation facile des résultats et la visualisation simple des règles de décision. Cependant, les arbres non élagués peuvent être très complexes et sujets à un surajustement.

Nous avons donc optimisé le paramètre de complexité pour l'élagage en utilisant une validation croisée LGOCV (*Leave Group Out Cross Validation*) à 10 ensembles, puis en minimisant la racine carrée de l'erreur quadratique moyenne (RMSE). L'optimisation du paramètre de complexité permet de trouver le bon niveau d'élagage pour éviter un surajustement tout en maintenant la précision prédictive du modèle.

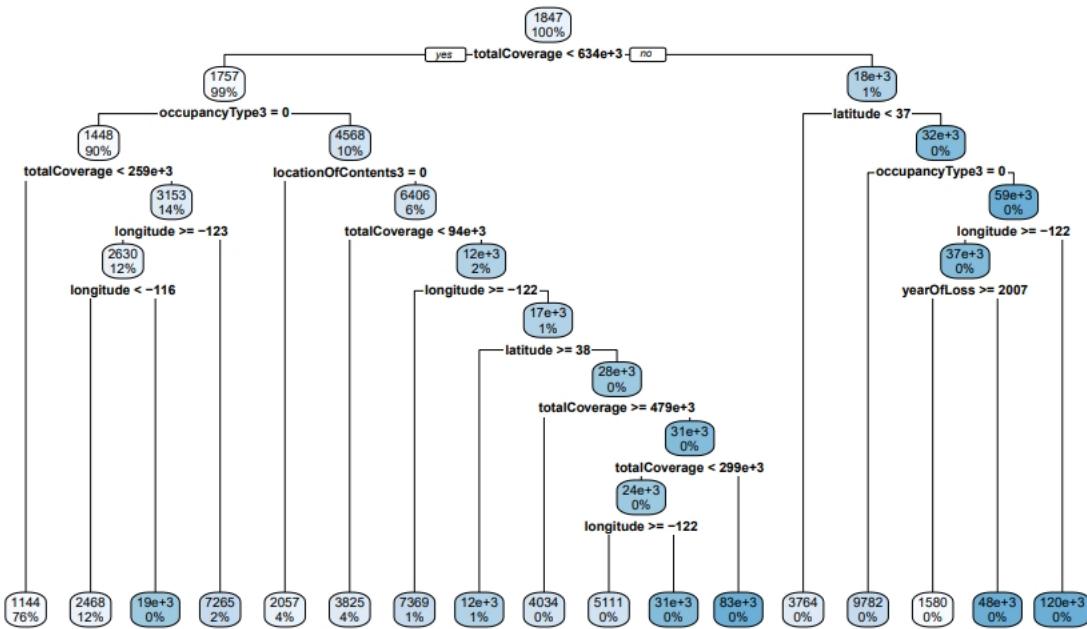


Figure 4: Arbre élagué

Bagging

Pour le *bagging*, nous utilisons le nombre de variables explicatives pour déterminer le nombre de variables prises en compte à chaque division de l'arbre de décision, soit 12. L'algorithme utilise également un échantillon avec remplacement de la même taille que l'échantillon des données d'entraînement.

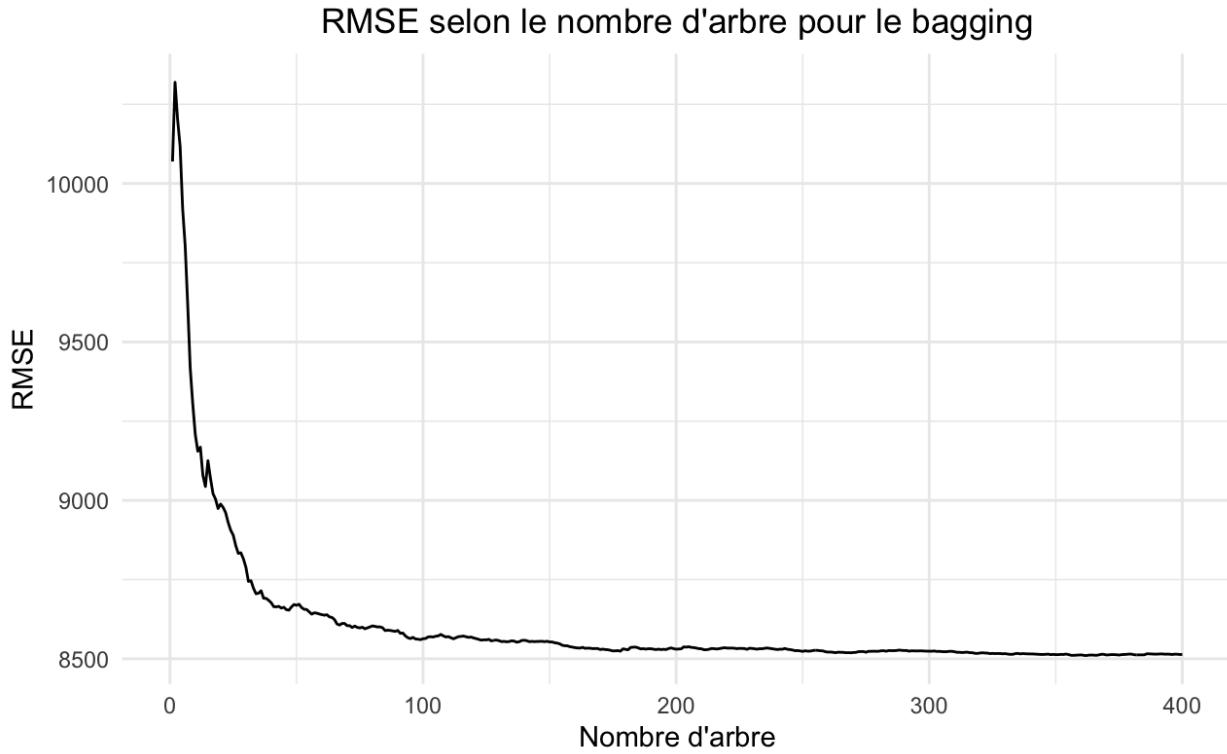


Figure 5: REQM selon le nombre d'arbres

Le nombre d'arbres pour le *bagging* est déterminé à l'aide de la racine de l'erreur quadratique moyenne (RMSE), car il s'agit d'un problème de régression. En observant la figure ci-dessus (figure 5), nous constatons que le nombre d'arbres se stabilise vers 200 arbres. Pour diminuer le temps de calcul, nous prenons 200 arbres même si un grand nombre d'arbres ne conduit pas à du surajustement.

Ensuite, l'hyperparamètre `nodesize` sera exploré pour prévenir le surajustement causé par des arbres trop profonds. Pour éviter un temps de calcul excessif, les valeurs manuellement testées pour `nodesize` vont de 0 à 500, en augmentant par bonds de 20.

La figure ci-dessus (figure 6) montre que le `nodesize` qui minimise l'RMSE se trouve autour de 80. Nous allons donc tester toutes les valeurs de 60 à 100 pour trouver la valeur optimale de l'hyperparamètre.

Dans la figure ci-dessus (figure 7), nous pouvons voir que la valeur de `nodesize` optimale est 97.

Les hyperparamètres optimaux pour le *bagging* sont donc les suivants :

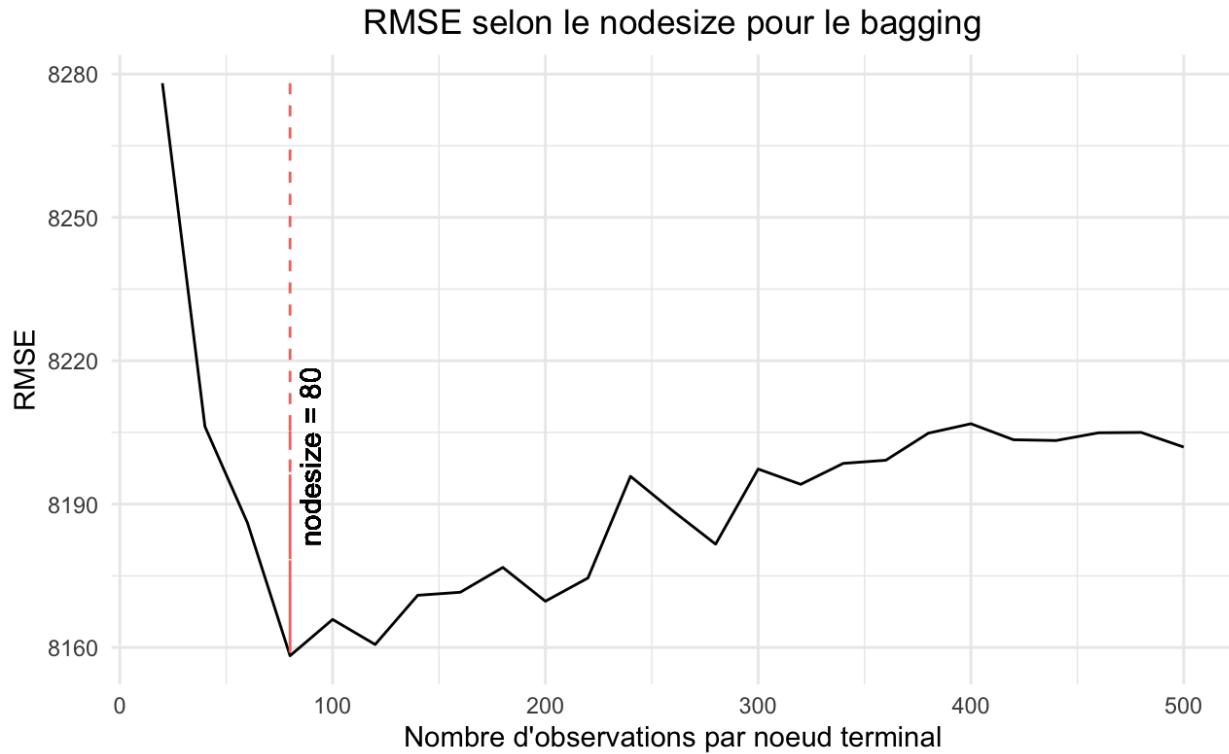


Figure 6: REQM selon le nombre minimal d'observations par feuille

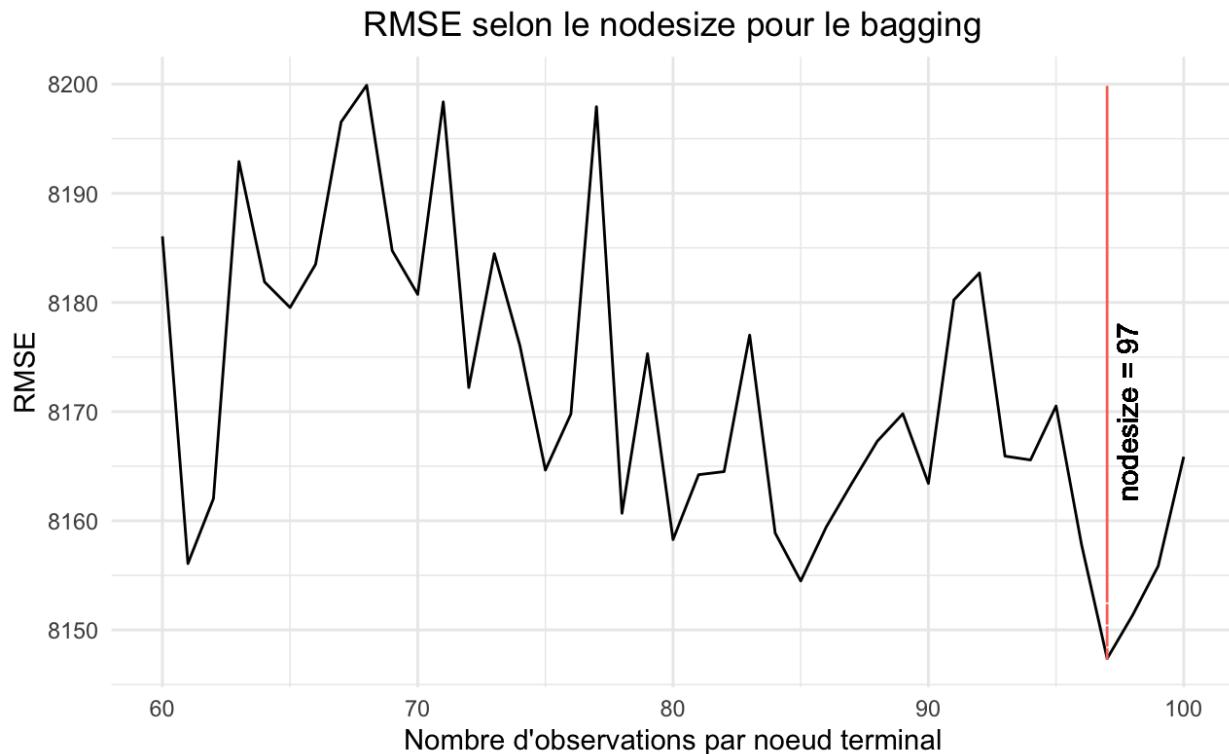


Figure 7: REQM selon le nombre minimal d'observations par feuille de 60 à 100

Table 1: Valeurs des hyperparamètres du modèle de bagging final

Hyperparamètre	Valeur
Nombre d'arbres	200
Nombre d'observation dans les noeuds terminaux	97

Forêt aléatoire

Pour la forêt aléatoire, nous commençons par 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. De plus, en utilisant une proportion de 50% pour les échantillons *bootstrap*, nous aidons à diminuer la corrélation entre les arbres.

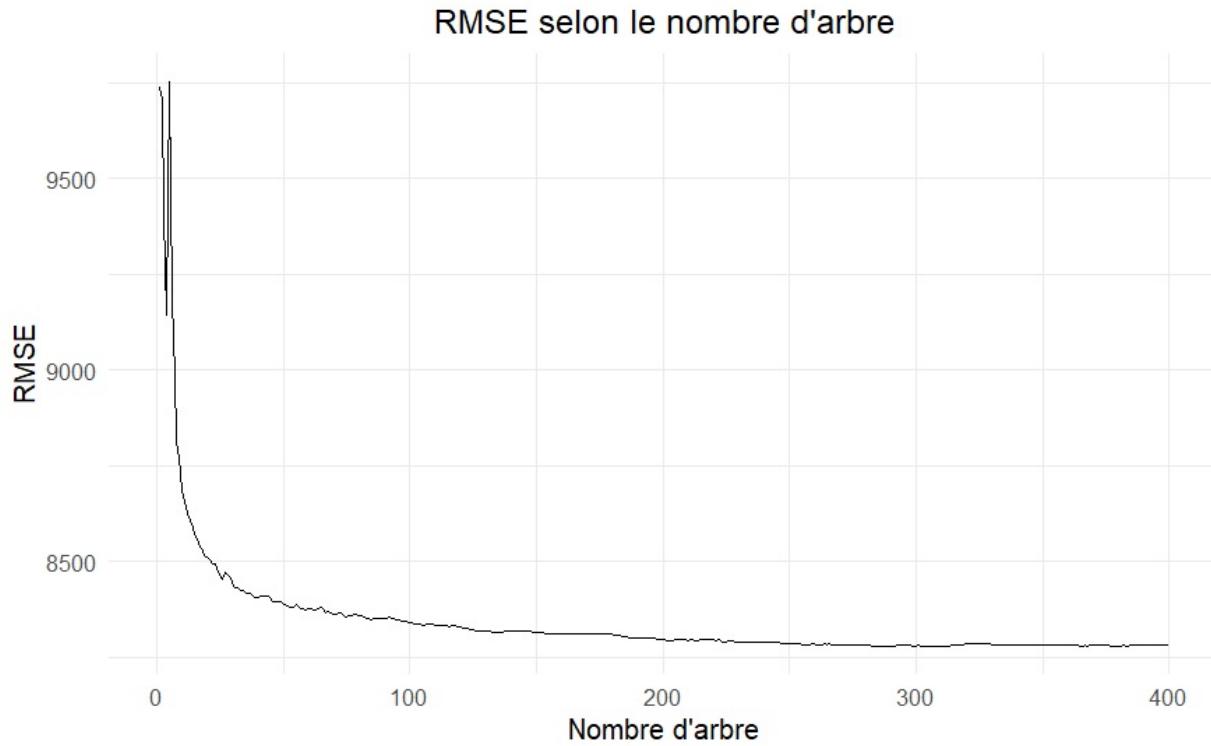


Figure 8: REQM selon le nombre d'arbres

Étant en régression, la racine de l'erreur quadratique moyenne (REQM) sera utilisée comme mesure de comparaison. Nous remarquons ici (figure 8) que la REQM se stabilise aux alentours de 100-150 arbres, nous utiliserons alors 200 arbres pour l'optimisation des autres hyperparamètres, puisque nous ne pouvons pas surajuster en ayant trop d'arbre avec les forêts aléatoires. Maintenant, nous regardons plus en profondeur le nombre de prédicteurs possible à chaque séparation d'un arbre, la variable `mtry`.

Table 2: RMSE par rapport au `mtry`

<code>mtry</code>	1	2	3	4	5	6	7	8	9	10	11	12
RMSE	8848	8660	8527	8457	8408	8379	8366	8352	8367	8356	8360	8361

Les résultats de la table 2 ont été obtenus par validation croisée à 5 plis, pour ainsi réduire le biais d'échantillonage. L'utilisation des 8 choix de variables explicatives à chaque noeud minimise la REQM.

Pour éviter un surajustement dû à des arbres inutilement trop profonds, nous devrons ajuster la valeur de

`nodesize`, mais il est impossible de le faire directement avec le package `caret`. Puisque le modèle est entraîné sur 43061 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.

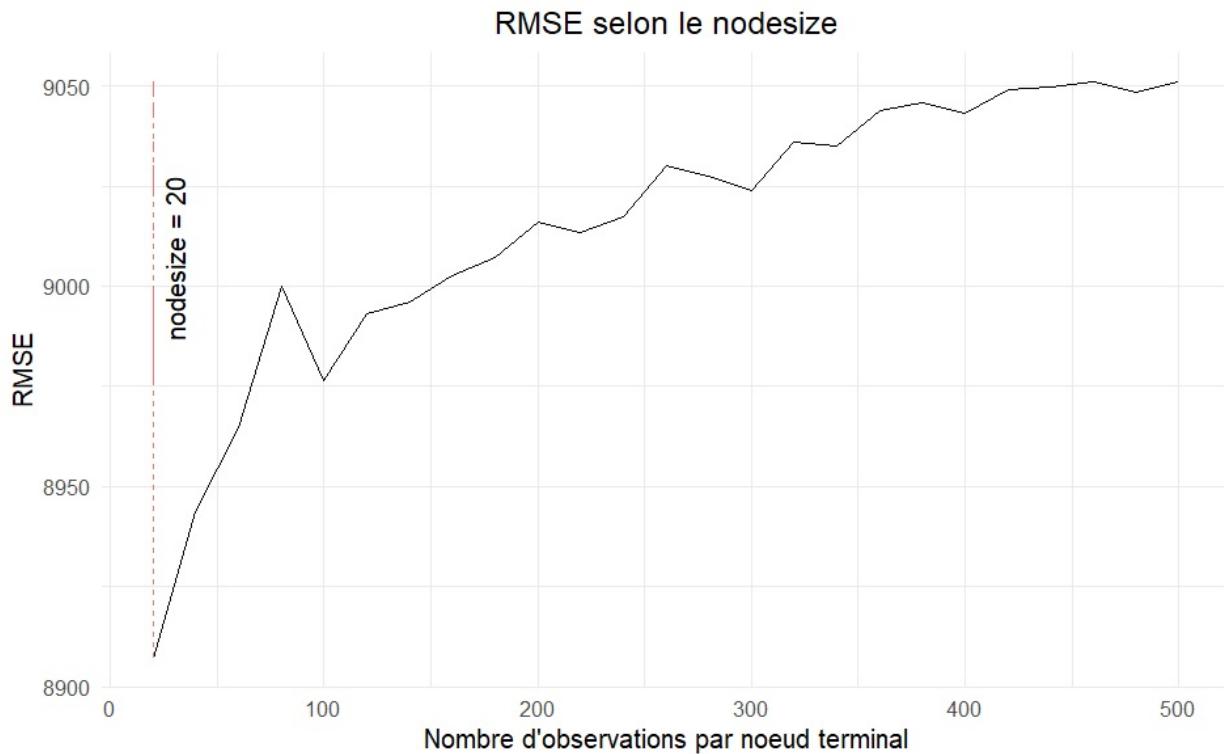


Figure 9: REQM selon le nombre minimal d'observations par feuille

Dans la figure 9, la valeur minimale de `nodesize` est de 20. Puisque l'analyse précédente a été effectuée par bonds de 20, nous la ferons à nouveau, de manière plus précise, de 1 à 40.

Dans la figure 10, la valeur minimale de `nodesize` est de 27.

Par conséquent, les hyperparamètres finaux pour le modèle de forêt aléatoire sont ceux décrits dans la table suivante.

Table 3: Valeurs des hyperparamètres du modèle final

Hyperparamètre	Valeur
Nombre d'arbres	200
Nombre de choix de variables à chaque noeud	8
Nombre d'observation dans les noeuds terminaux	27

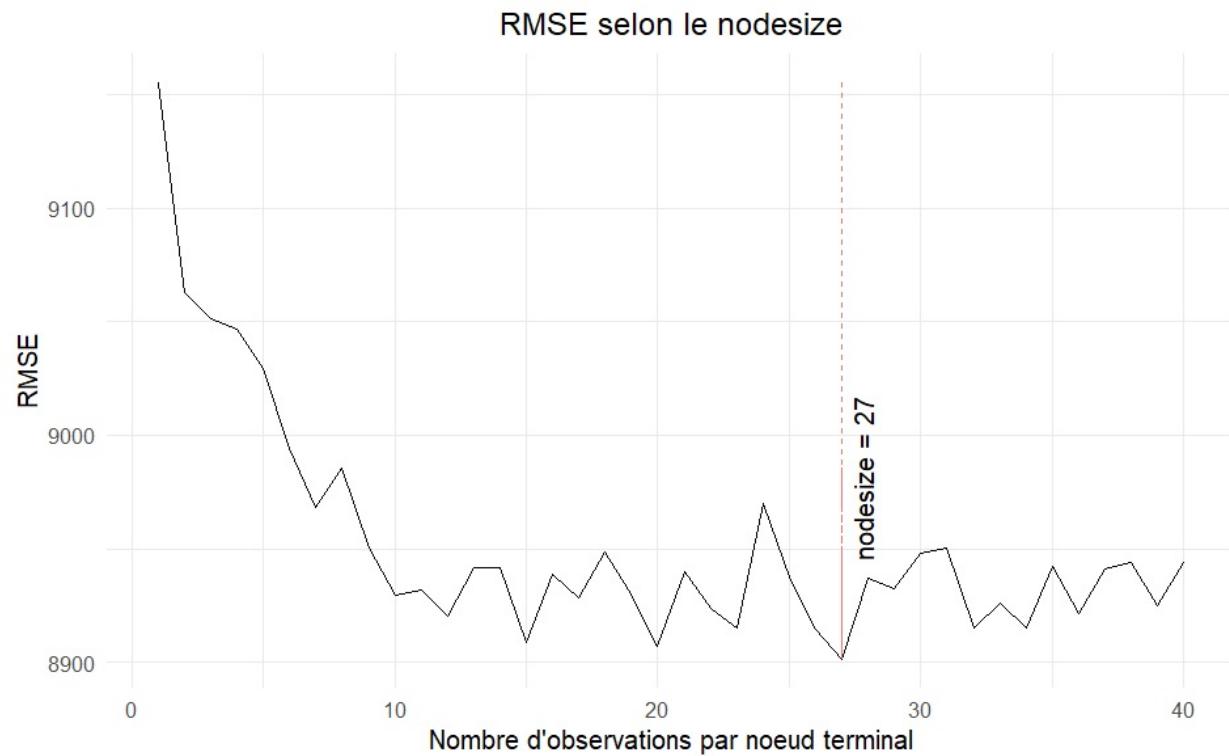


Figure 10: REQM selon le nombre minimal d'observations par feuille de 1 à 40

Gradient Boosting

Nous commençons par entraîner un modèle GBM par validation croisée à 5 plis. Nous testons les valeurs de la taille maximale de chaque arbre $\{5, 10, 15\}$ et un nombre d’itérations de 1000 à 6000 par bonds de 1000, tout en utilisant un paramètre d’apprentissage, $\lambda = 0.01$.

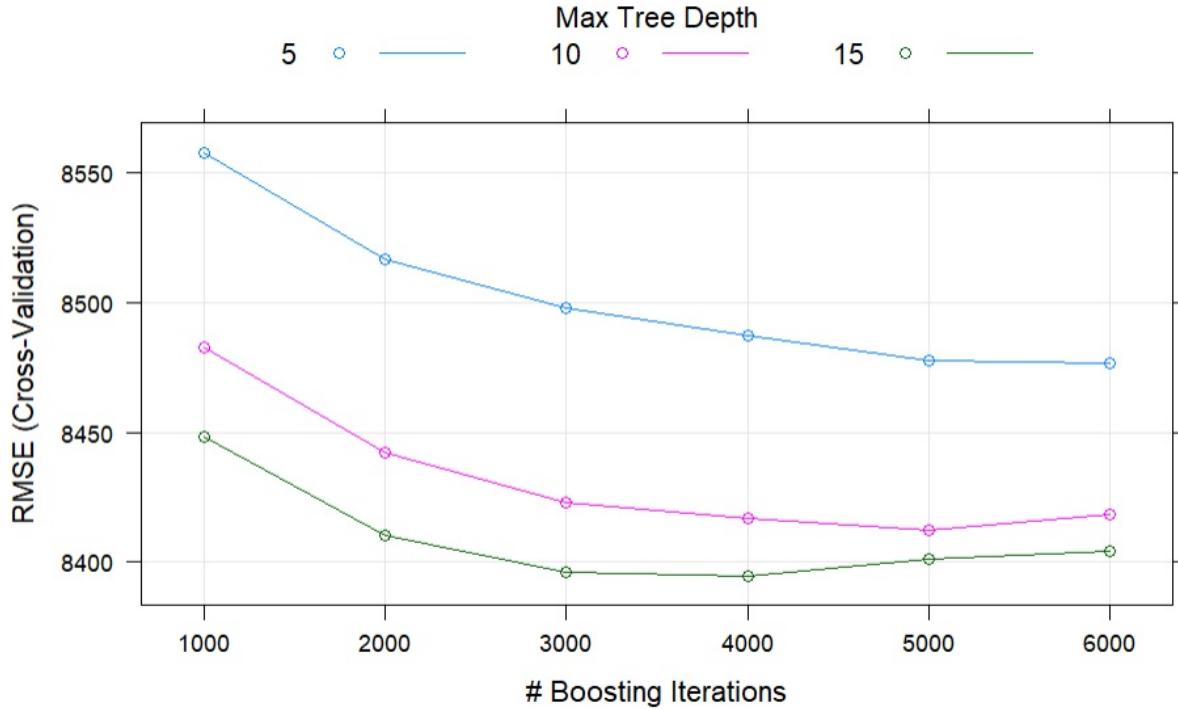


Figure 11: REQM selon la taille maximale de chaque arbre et le nombre d’itérations

Nous voyons, dans la figure 11, qu’une profondeur de 15 et qu’environ 4000 arbres minimisent la REQM. Une profondeur de 15 semble suffisante pour capter les itérations entre les variables, sans trop faire exploser le temps de calcul. Il est difficile de voir le nombre idéal.

Après avoir fait une analyse plus précise, nous remarquons que 3000 itérations est optimal, et les paramètres finaux seront ceux décrits à la table suivante.

Table 4: Valeurs des hyperparamètres du modèle final

Hyperparamètre	Valeur
Nombre d’arbres	3000.00
Profondeur de chaque arbre	15.00
Taux d’apprentissage	0.01

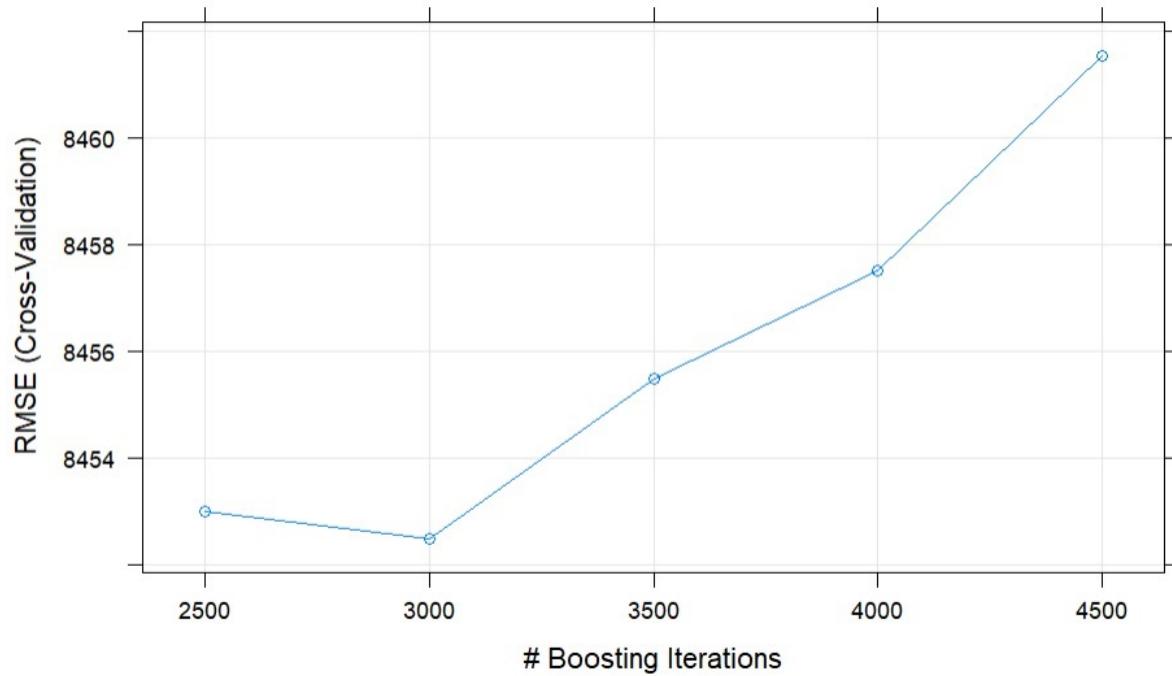


Figure 12: REQM selon le nombre d'itérations

Comparaison des modèles

Nous avons décidé d'utiliser la racine carrée de l'erreur quadratique moyenne(REQM) afin de comparer la performance prédictive des différents modèles discutés dans le rapport. Les deux modèles retenus selon cette métrique sont le *bagging* et la forêt aléatoire puisque ce sont ceux qui ont la plus petite valeur de REQM.

Table 5: Valeurs des racines des erreurs quadratiques moyennes des modèles finaux

Modèle	Valeur
GLM Tweedie	9306.878
K plus proches voisins	17926.944
Arbre de décision	9302.756
Bagging	8679.762
Forêt aléatoire	8664.902
Gradient boosting	9204.148

Afin d'être sûr de nos choix, nous avons aussi mesuré l'erreur absolue moyenne(EAM).

Table 6: Valeurs des erreurs absolues moyennes des modèles finaux

Modèle	Valeur
GLM Tweedie	1921.988
K plus proches voisins	8395.884
Arbre de décision	3038.017
Bagging	2709.397
Forêt aléatoire	2705.454
Gradient boosting	2890.436

Les modèles ayant les petites valeurs sont le GLM Tweedie et la forêt aléatoire. Cependant, le GLM Tweedie avait la deuxième valeur plus grande pour le REQM. Puisque l'EAM est vraiment sensible aux valeurs aberrantes et que le *bagging* a encore une des plus petites valeurs cette métrique, nous avons décidé de choisir les modèles *bagging* et forêt aléatoire comme la REQM nous l'indiquait.

Interprétation des meilleurs modèles

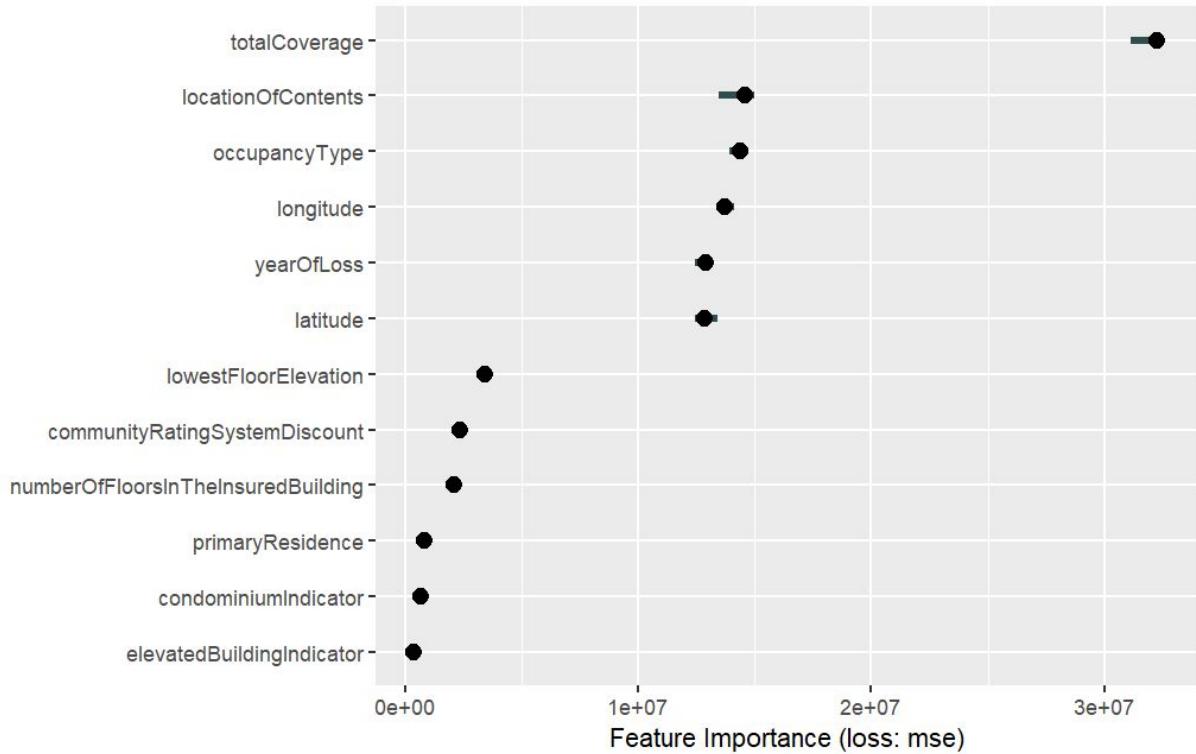
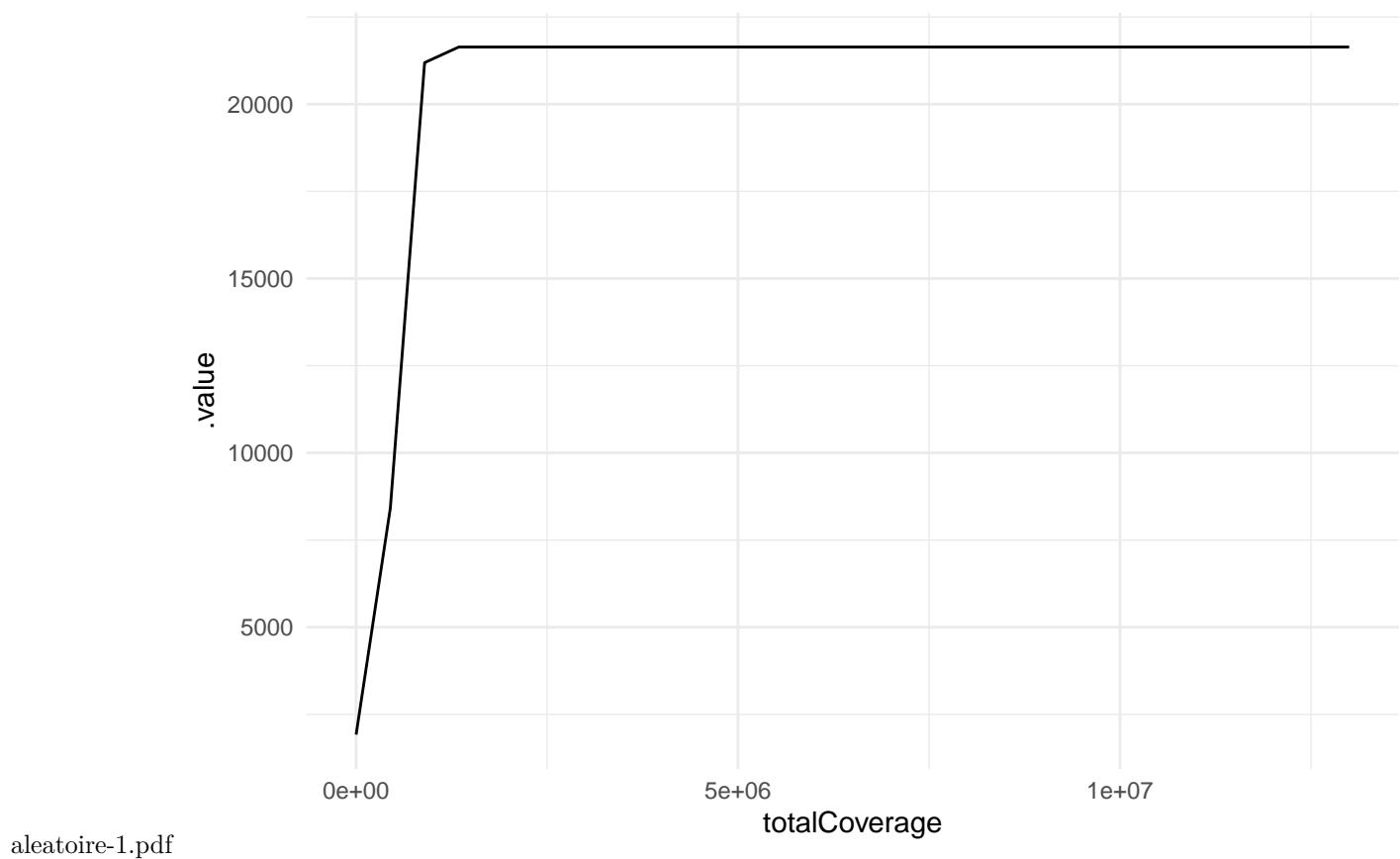
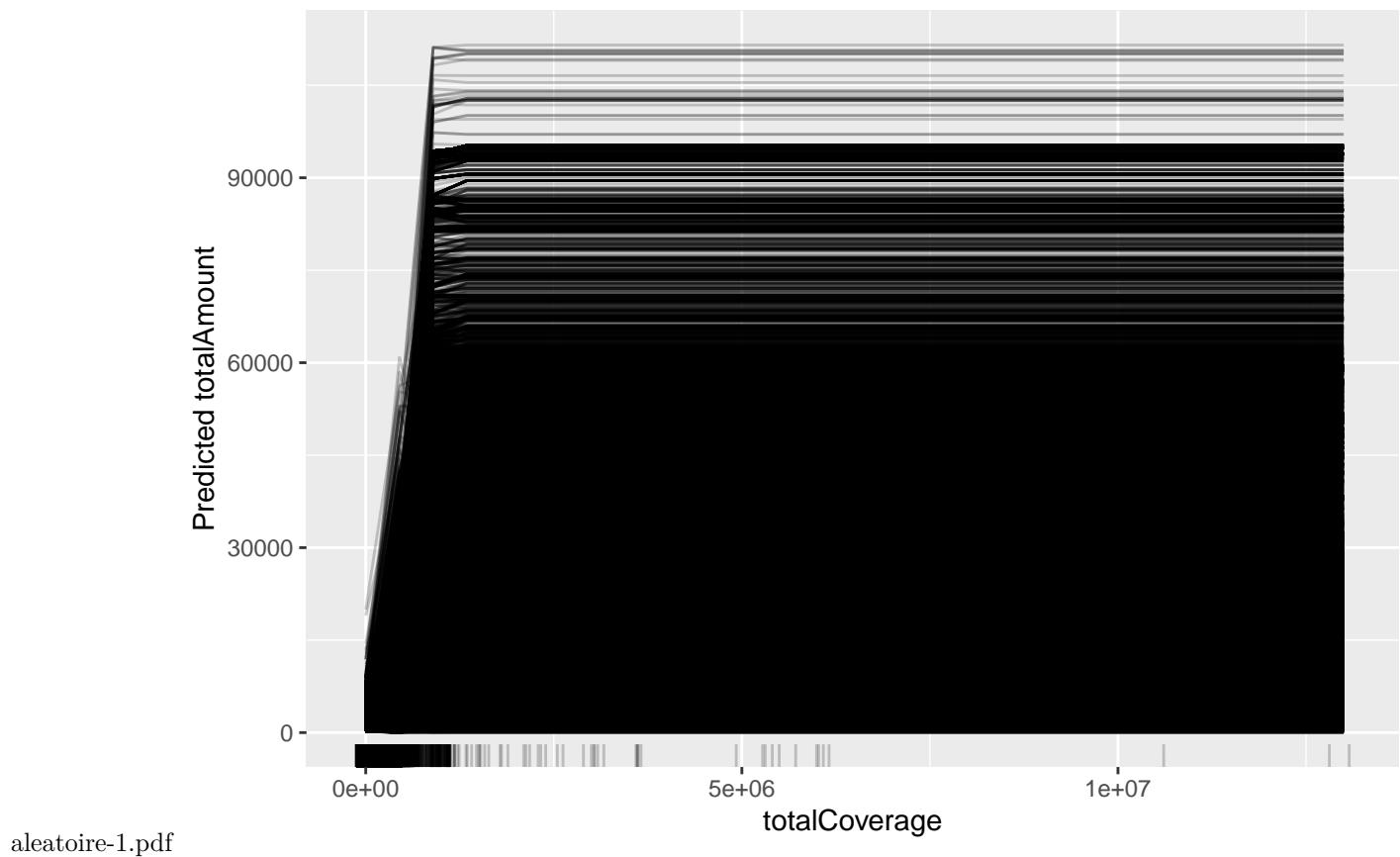


Figure 13: Importances des variables pour le bagging

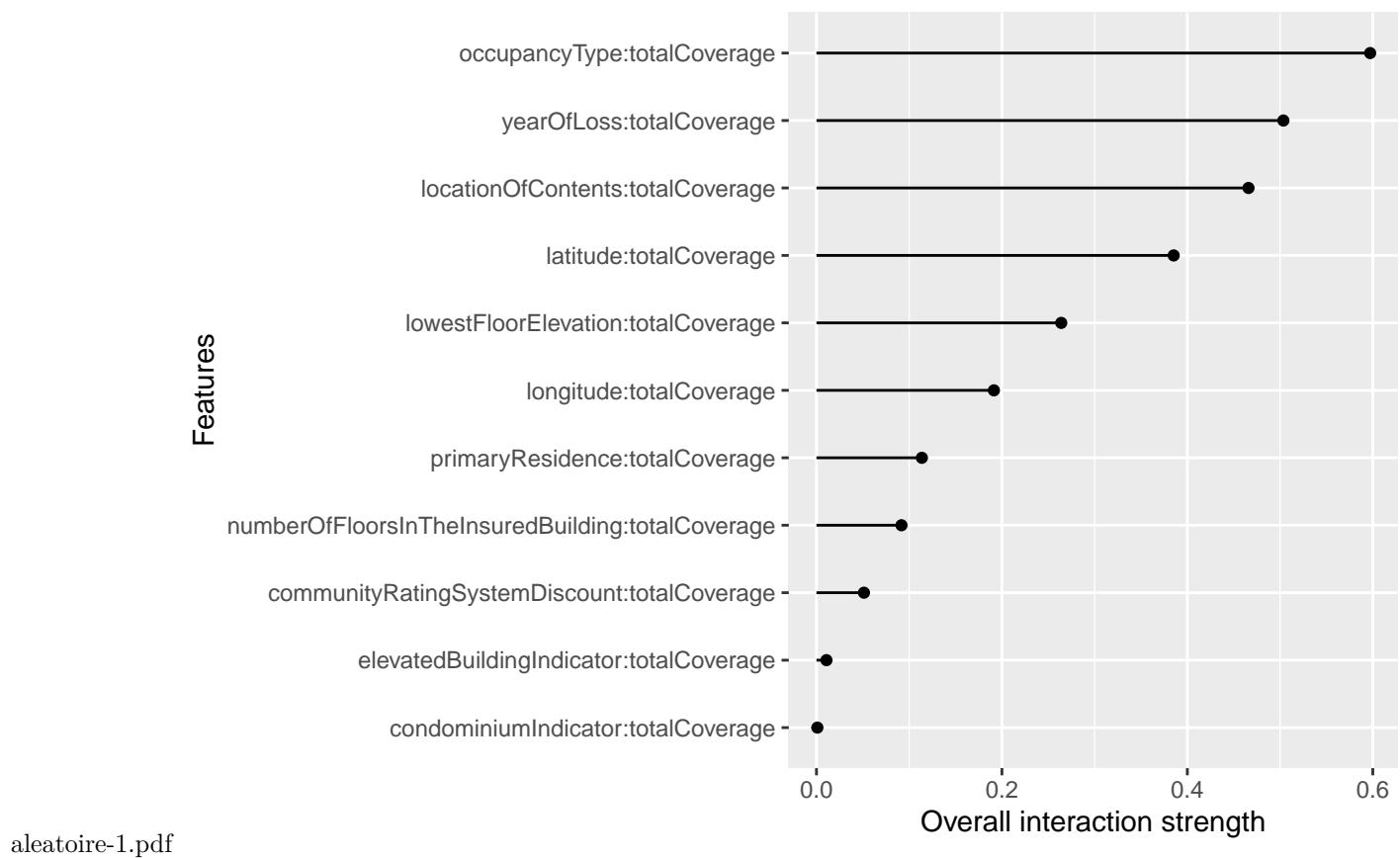
Selon la figure 12, la variable la plus importante pour le modèle *bagging* est `totalCoverage`. Pour mieux comprendre son effet marginal sur la prévision, nous visualisons le graphique de dépendance partielle.



aleatoire-1.pdf



aleatoire-1.pdf



aleatoire-1.pdf

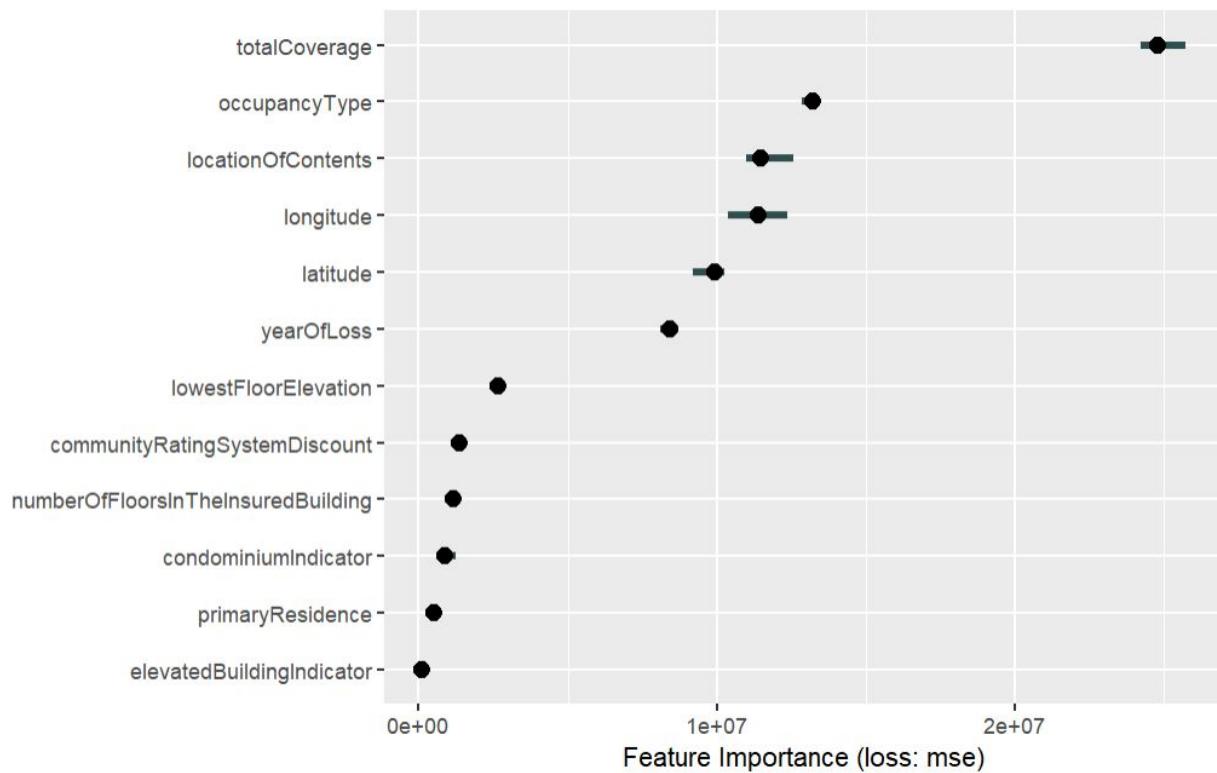
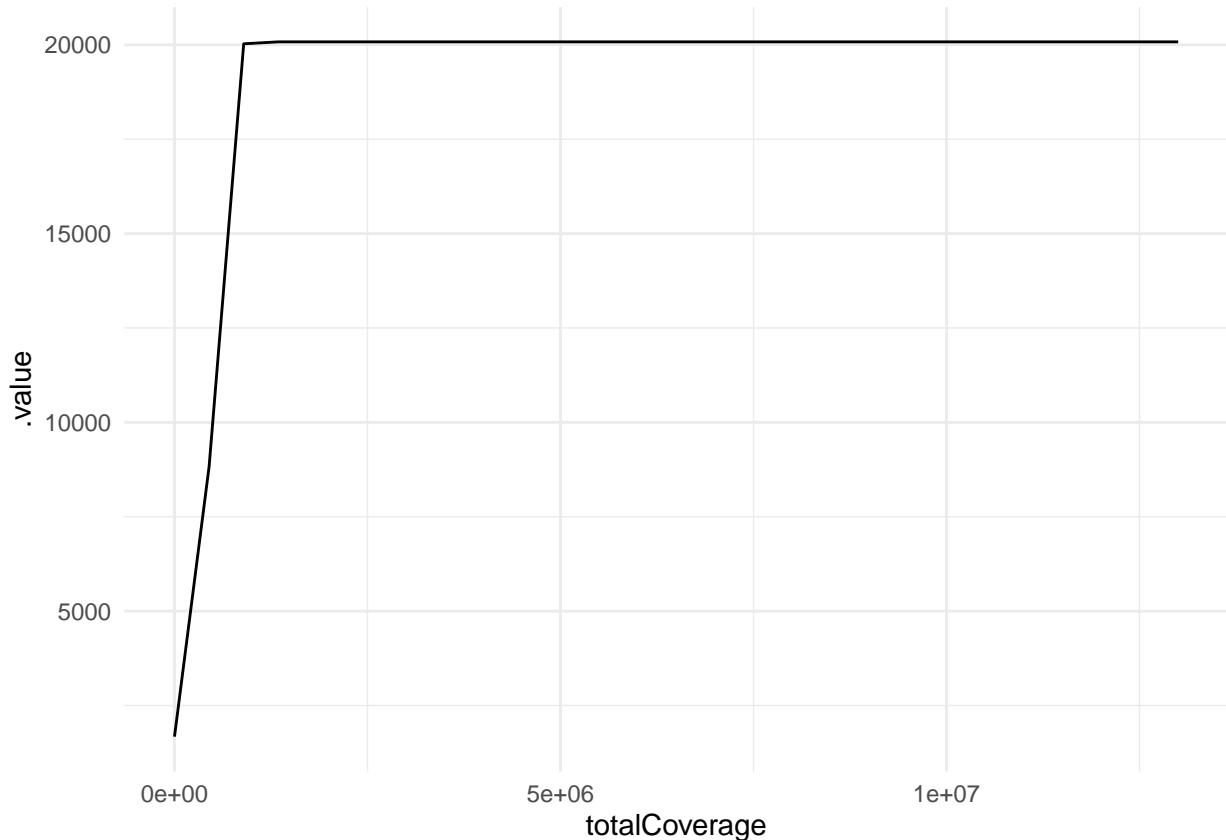
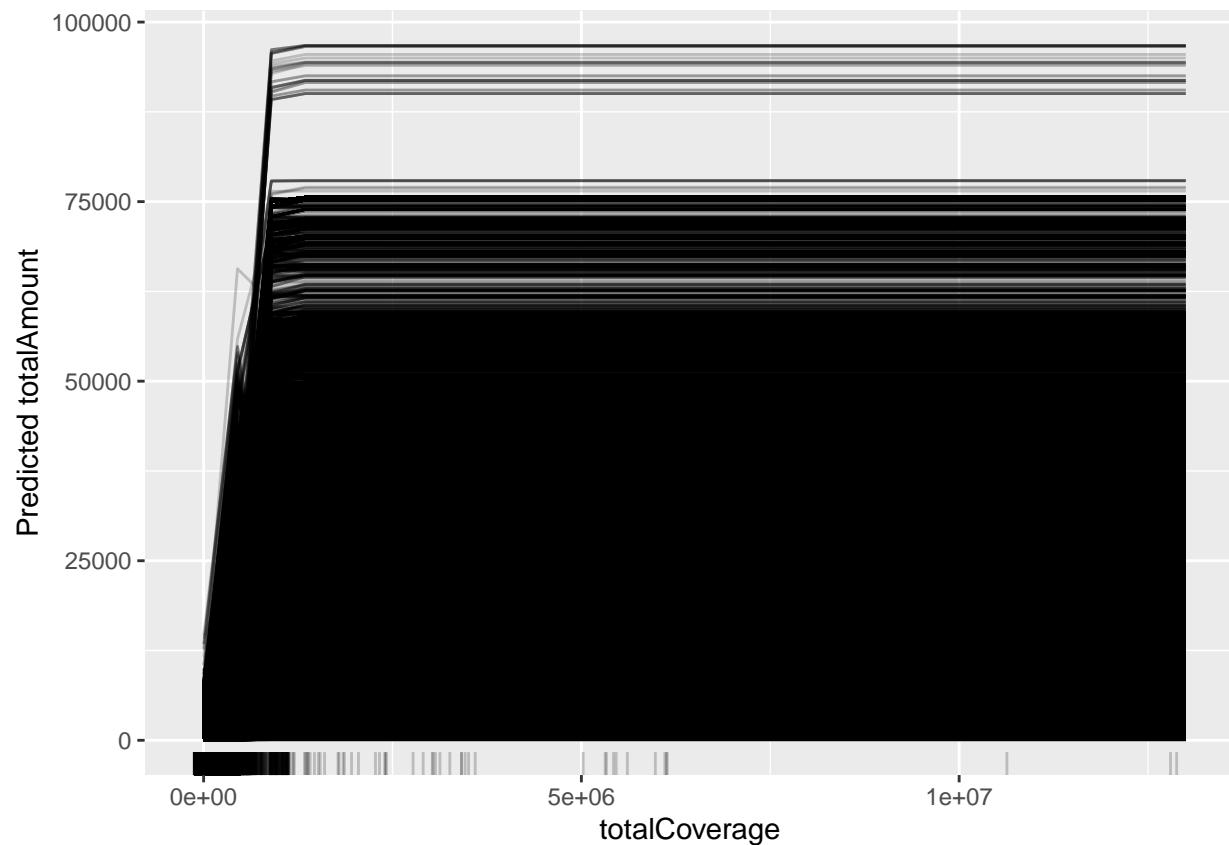
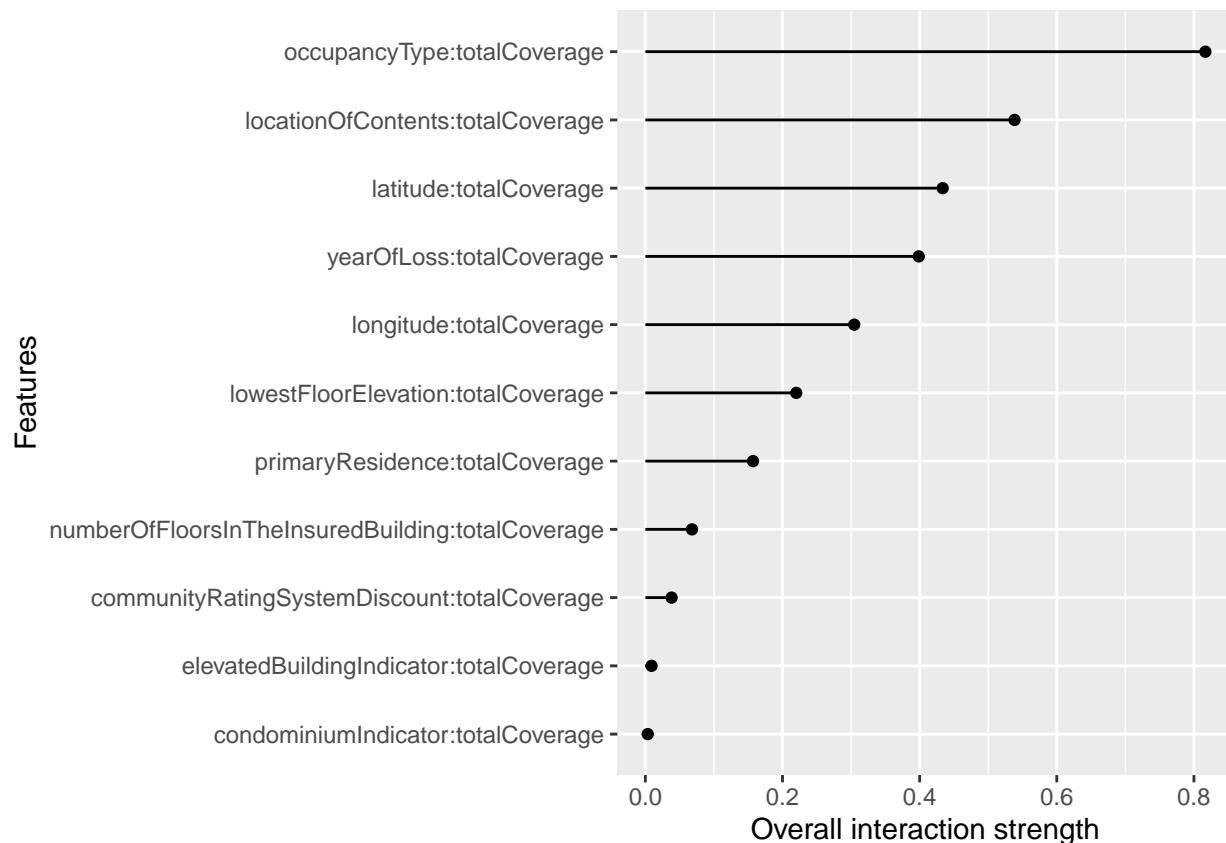


Figure 14: Importances des variables pour le bagging







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>