1. Description des méthodes comparées

Dans cette étude, nous avons comparé deux algorithmes d'apprentissage supervisé : l'arbre de décision et la forêt aléatoire. L'arbre de régression est un type d'arbre de décision, modélisant des relations non linéaires entre des prédicteurs et une variable cible continue. Il divise les données en nœuds enfants pour minimiser les erreurs de prédiction. Cependant, cette méthode est à risque de faire du surapprentissage, d'où l'intérêt des forêts aléatoires. Celles-ci entraînent plusieurs arbres sur des échantillons aléatoires et combinent leurs prédictions, améliorant ainsi la généralisation et la robustesse du modèle. Cette approche de bagging compense les limites des arbres de régression uniques.

Arbre de décision (rpart):

- *Construction du modèle* : Chaque modèle d'arbre de décision a été entraîné à l'aide de la fonction rpart(). Les données ont été divisées en ensembles d'entraînement (80 %) et de test (20 %).

Forêt aléatoire (randomForest):

- *Construction du modèle* : Les forêts aléatoires ont été entraînées à l'aide de la fonction randomForest(), également sur des ensembles d'entraînement de 80 % des données. Le nombre d'arbres dans la forêt a été fixé à 100 (ntree = 100).

Pour les deux méthodes, le temps d'exécution a été mesuré et l'erreur quadratique moyenne (EQM) a été calculée pour les prédictions sur les ensembles de test. Les hyperparamètres propres à chaque méthode ont été maintenus à leurs valeurs par défaut afin d'évaluer les performances des algorithmes sans ajustements spécifiques.

2. Questions de recherche

La question de recherche consiste donc à déterminer si les forêts aléatoires ont <u>toujours</u> une meilleure performance prédictive que les arbres de régression. Pour évaluer cette différence de performance, nous avons utilisé la simulation Monte-Carlo afin de comparer les deux méthodes à travers divers scénarios, lesquels prennent en compte des facteurs tels que la taille d'échantillon, du nombre de variables indépendantes et des relations entre celles-ci et la variable cible.

3. Simulation Monte-Carlo

La méthode de simulation de Monte-Carlo nous permettra de générer des données synthétiques et d'effectuer les tests de modèles à travers une multitude de types de données illustrés dans le tableau ci-dessous en gardant les hyper paramètres des deux modèles par défauts.

3.1 Plan de génération des données

Tableau 1. Facteurs pour la génération Monte-Carlo de jeux de données

Taille d'observations (n x p**)** X	**Nombre de variables indépendantes (**p**)** β	**Relation entre les variables indépendantes et la cible** y
Petite (n = 500)	Petit (p = 5)	`Linéaire : y = X·β`
Grande (n = 2500)	Élevé (p = 15)	`Quadratique : y = a·X^2 +X·β`

- X : matrice **n** lignes et **p** `colonnes générés à partir d'une normale X~N(0,1)`
- β : vecteur de variables indépendantes générés à partir d'une normale `β~N(0,1)`
- a : vecteur de taille **p** `avec des valeurs tirées d'une normale a~N(0,1)`
- Nombre de répétitions : 1000 répétitions par scénario

Nous avons créé 8 scénarios basés sur le tableau ci-dessus afin de tester les performances : PPL, PPQ, PEL, PEQ, GEQ, GEL, GPL, GPQ. (ex: PPL = Petite taille d'observation, nombre petit de prédicteur et relation linéaire)

3.2 Mesure des résultats

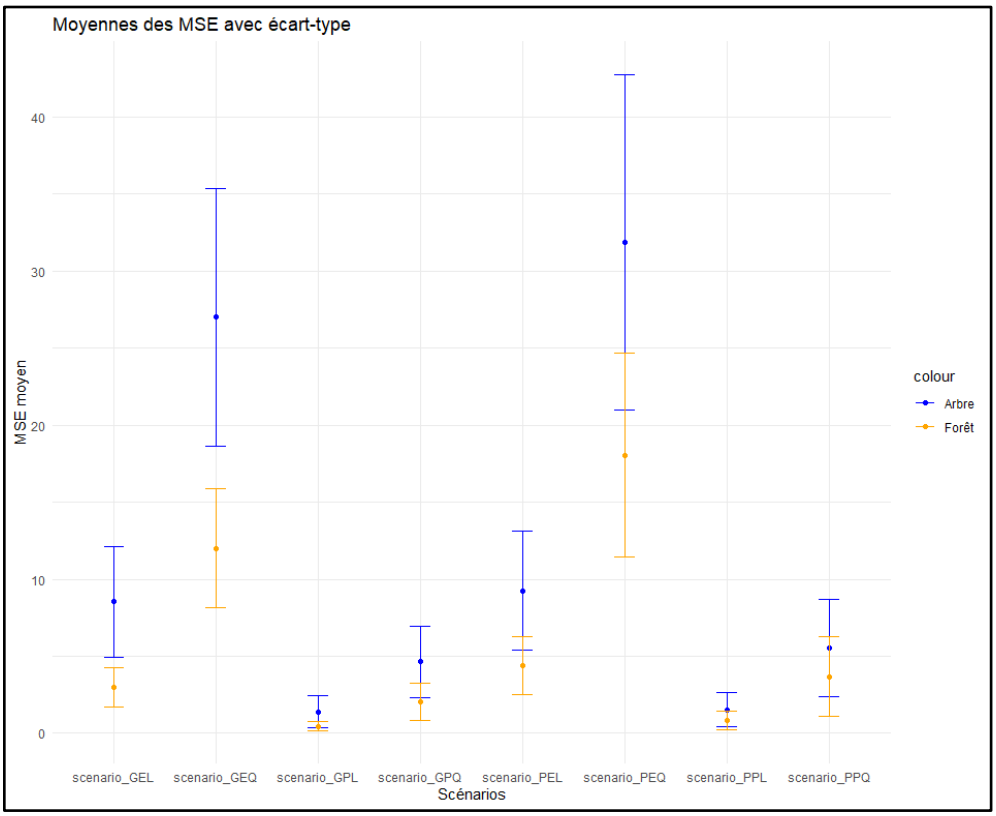
3.2.1 Comparaison des moyennes et écart-type de l'EQM pour chaque scénario

Tableau 2. Moyennes et écart-types de l'EQM par scénario

Scénario	**EQM Arbre (**$\mu \pm \sigma$**)**	**EQM Forêt (**$\mu \pm \sigma$**)**
scenario_GEL	8.53 ± 3.58	2.97 ± 1.29
scenario_GEQ	27.0 ± 8.38	12.0 ± 3.86
scenario_GPL	1.38 ± 1.04	0.445 ± 0.299

scenario_GPQ	4.63 ± 2.35	2.04 ± 1.18
scenario_PEL	9.25 ± 3.86	4.39 ± 1.86
scenario_PEQ	31.9 ± 10.9	18.0 ± 6.61
scenario_PPL	1.53 ± 1.11	0.851 ± 0.594
scenario_PPQ	5.55 ± 3.16	3.67 ± 2.58

Graphique 1. Moyennes et écart-types de l'EQM par scénario



Le graphique 1 montre une performance globalement meilleure de la part des forêts aléatoires par rapport aux arbres de régression. Par contre, nous pouvons observer des chevauchements en raison des écarts-types.

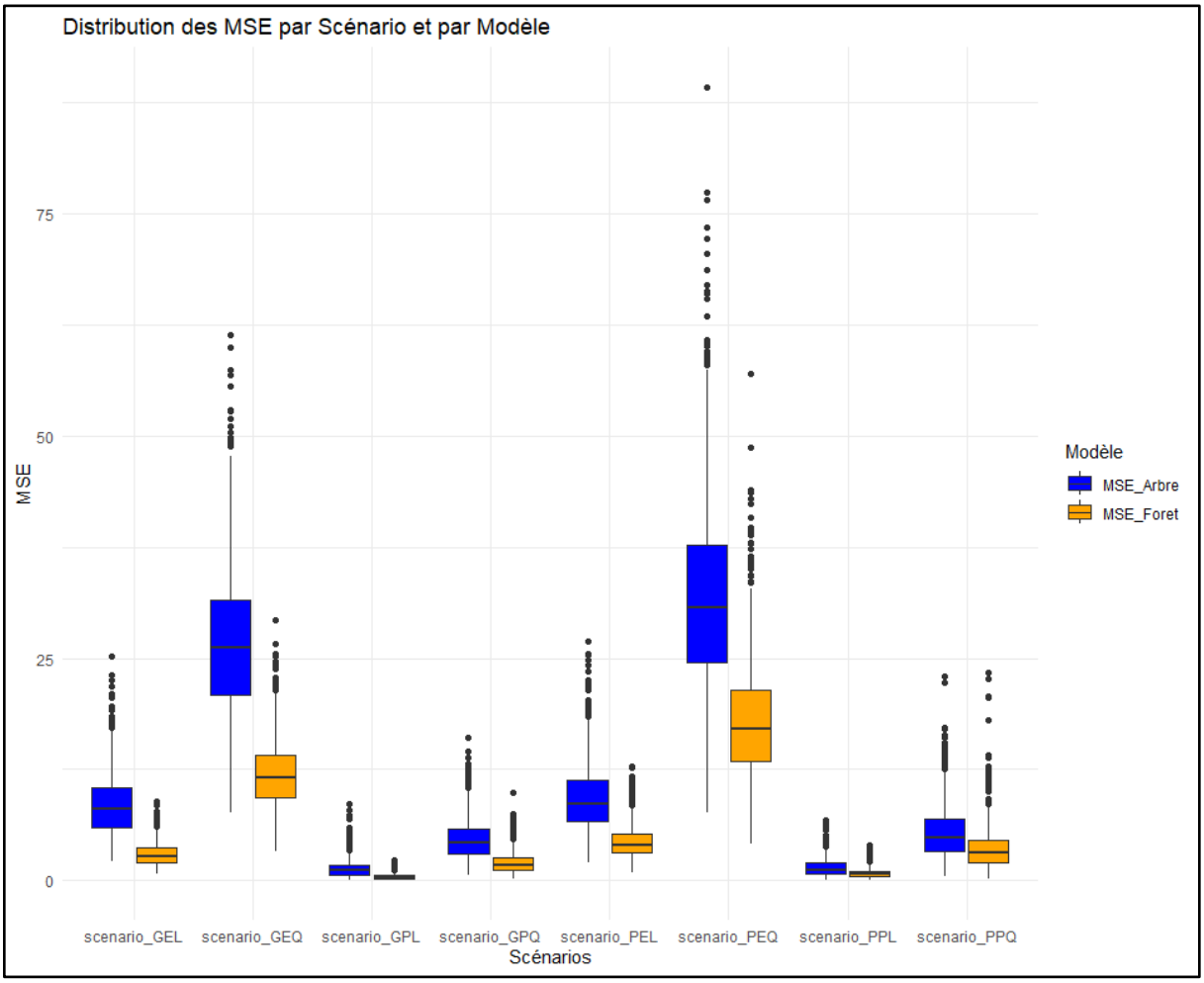
En effet, dans certains scénarios, bien que la forêt aléatoire obtient une EQM moyenne inférieure, l'écart-type est assez large pour que l'intervalle des EQM de l'arbre chevauche celui de la forêt. Nous pouvons voir que c'est le cas pour GPL, GPQ, PPL, et PPQ. Les chevauchements dans ces scénarios sont assez prononcés pour suggérer que pour des jeux de données avec un nombre réduit de prédicteurs, l'avantage des forêts aléatoires en termes de généralisation peut être moins

prononcé, rendant les différences de performance plus sensibles aux variations aléatoires des données.

En revanche, dans les scénarios plus complexes comme GEL et GEQ, les forêts aléatoires montrent une amélioration notable par rapport aux arbres de régression, avec des intervalles n'ayant aucun chevauchement. Ceci indique une bien meilleure robustesse des forêts dans des contextes de grande taille d'échantillon et nombreux prédicteurs, et ce pour les relations linéaires ou quadratiques. Pour les scénarios PEL et PEQ, bien que la forêt aléatoire aie encore une performance supérieure, un léger chevauchement est observable dans les intervalles. Cela suggère que, même si les forêts aléatoires demeurent plus performantes, leur avantage est atténué en situation de petite taille d'observation, même s'il y a un grand nombre de prédicteurs.

3.2.2 Distribution de l'EQM par scénario

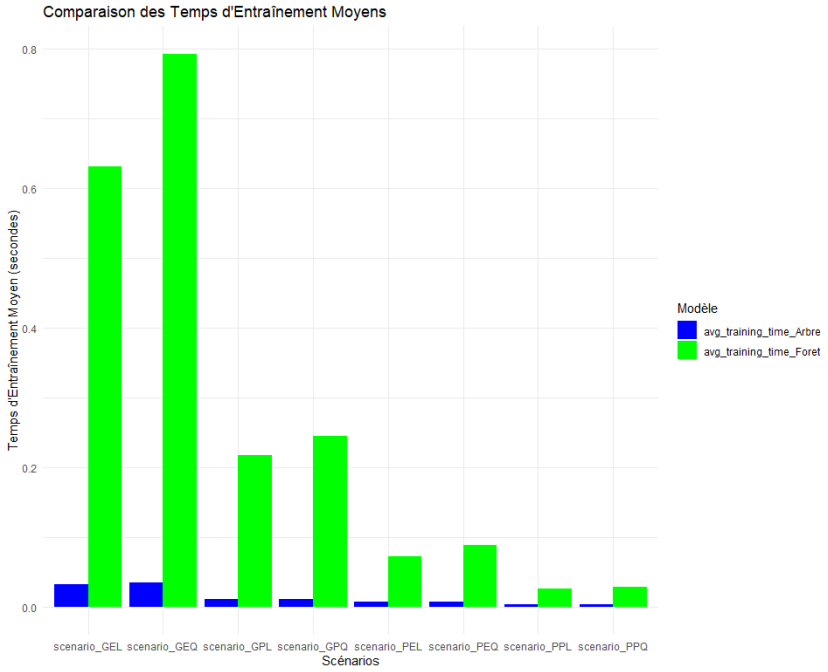
Graphique 2. Distribution de l'EQM par modèle et pour chaque scénario



Dans le graphique 2, on remarque que pour les scénarios GEQ (grande taille d'observation, nombre élevé des prédicteurs et relation quadratique) et PEQ (petite taille d'observation) les deux modèles présentent une distribution plus étendue avec des valeurs élevées aberrantes. Cela indique que les arbres et les forêts sont moins robustes pour prédire des datasets avec beaucoup de prédicteurs et un contexte de relation quadratique. Il est néanmoins important de souligner que pour les 8 scénarios, l'arbre (en bleu) a une plus grande variabilité et une performance généralement moins précise par rapport à la forêt (en orange). À guise d'exemple, pour le scénario GEQ, l'arbre a une médiane EQM proche de 27, avec des valeurs de 10 à plus de 50, démontrant une variabilité très élevée, alors que la forêt a une médiane de 12, avec des valeurs principalement comprises entre 5 et 20. Ainsi, dans ce dernier scénario, la forêt aléatoire permet de réduire l'EQM de 50% de plus que l'arbre de régression. Pour les scénarios plus simples comme PPL et PPQ, les deux modèles ont des EQM faibles et la forêt semble toujours être meilleure que l'arbre, mais l'avantage dans ce cas de figure est moins prononcé. Considérant ce surpassement moins important pour les scénarios simples, il sera important de déterminer si l'avantage de précision de la forêt aléatoire justifie toujours le coût de calcul supplémentaire (voir section 3.2.5).

3.2.4 Comparaison des temps d'exécution pour chaque scénario

Graphique 3. Distribution du temps d'entraînement moyen par modèle et par scénario



Le graphique ci-dessus des temps d'entraînement moyens pour chaque méthode montre les temps d'entraînement moyens pour chaque scénario pour les deux méthodes comparées. Il semble que le temps d'entraînement des forêts aléatoires est généralement plus élevé que celui des arbres de régression pour tous les scénarios testés.

Les scénarios avec un nombre de variables prédictives plus élevé tels que GEQ et GEL semblent avoir les temps d'entraînement les plus élevés pour les arbres et pour les forêts. Ils se distinguent également par le fait qu'ils semblent avoir la plus grande différence de temps d'entraînement entre les deux méthodes avec environ **0,79 secondes** et **0,63 secondes** pour les forêts contre **0,035 secondes** et **0,03 secondes** pour les arbres. Cela montre que la complexité des données a un impact important sur le temps de calcul des forêts aléatoires. Il est cependant également possible de noter que l'écart de temps d'entraînement entre les deux méthodes est beaucoup moins important pour les scénarios comportant moins de variables et de variables prédictives tels que PPL et PPQ par exemple.

Vu que les forêts entraînent plusieurs arbres en parallèle, il est normal que l'entraînement prenne plus de temps et demande davantage de pouvoir computationnel. Bien que les forêts semblent avoir une meilleure précision de prédiction, la décision de la méthode à utiliser dépendra du contexte et du type de données à analyser. En fonction des priorités, un résultat plus rapide ou une plus grande précision, l'arbre de régression pourrait constituer une meilleure alternative que les forêts.

3.2.5 Rapport de l'EQM et du temps d'exécution

Cette section permet de répondre à la question suivante : "Les gains en précision de la forêt aléatoire justifient-ils un temps de calcul supplémentaire par rapport à l'arbre de régression ? "

Tableau 4. Rapports de l'EQM et du temps d'exécution par scénario

Scénario	**Rapport Arbre**	**Rapport Forêt**
scenario_GEL	262.0	4.7
scenario_GEQ	760.0	15.1
scenario_GPL	126.0	2.04
scenario_GPQ	406.0	8,32
scenario_PEL	1170.0	59.9
scenario_PEQ	3848.0	203.0
scenario_PPL	399.0	32.7
scenario_PPQ	1542.0	128.0

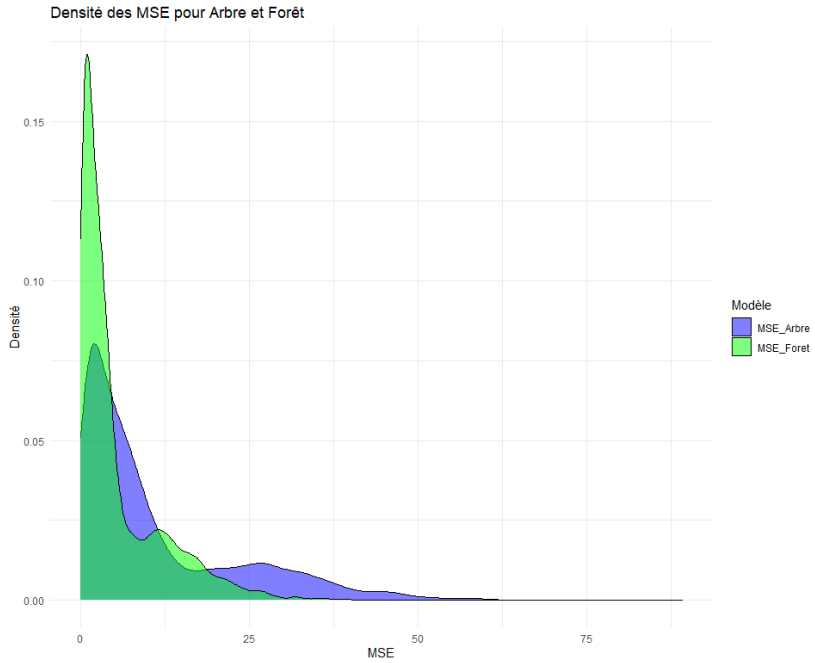
Dans tous les scénarios, la forêt aléatoire présente un rapport EQM/temps d'exécution meilleur que celui de l'arbre de régression. Il est donc justifié d'utiliser un modèle plus complexe comme

la forêt et payer un plus grand coût de calcul afin d'avoir une amélioration significative de la précision (faible EQM). Dans des scénarios plus complexes avec beaucoup de lignes et de colonnes (ex. : GEQ, GEL), le rapport est nettement en faveur de la forêt, ce qui rend préférable son utilisation pour des contextes où la précision est primordiale. Cependant, pour des scénarios simples ou avec peu de prédicteurs (GPL et PPL), l'arbre de régression peut montrer des performances intéressantes dans les cas d'une limitation de ressources de calcul. Enfin, la forêt aléatoire démontre des performances impressionnantes dans tous les cas de figure, tandis que l'arbre de régression permet d'avoir une alternative viable pour des cas plus simples.

4. Conclusion (apprentissage sur la méthode)

Les forêts aléatoires semblent donc généralement avoir une meilleure précision que les arbres de régression avec des MSE plus faibles dans tous les scénarios. Il est également possible de visualiser ceci à travers le graphique de la densité des MSE où la distribution des forêts est plus concentrée sur des faibles valeurs tandis que celle des arbres est plus étendue. Ceux-ci ont donc davantage de variabilité et une performance moins constante ainsi que des valeurs de MSE généralement plus élevées. Il semble également que les relations quadratiques et un nombre de variables plus élevé semble être plus difficile à modéliser pour les deux méthodes, bien que les forêts performent toujours mieux que les arbres, et nécessiter un temps d'entraînement plus élevé que lorsque le nombre de variables est plus petit et la relation est linéaire. Finalement, bien que les forêts ont un meilleur rapport de l'EQM et du temps d'exécution, celles-ci nécessitent malgré tout un plus grand pouvoir computationnel et une durée d'entraînement plus élevée. Le choix de la méthode à utiliser dépend donc du contexte. Si les ressources sont limitées, les données sont plus simples, les arbres de régression demeurent une option viable à considérer, surtout lorsqu'une très grande précision des résultats n'est pas une exigence absolue.

Graphique 4. Distribution de la densité des MSE par modèle



(EN EXTRA)

3.2.3 Analyse de la variance de l'EQM

Tableau 3. Variance de l'EQM par scénario

Scénario	**Variance EQM Arbre (σ^2)**	**Variance EQM Forêt (σ^2)**
scenario_GEL	12.8	1.65
scenario_GEQ	70.2	14.9
scenario_GPL	1.09	0.0894
scenario_GPQ	5.51	1.40
scenario_PEL	14.9	3.47
scenario_PEQ	119.0	43.7
scenario_PPL	1.22	0.353
scenario_PPQ	9.99	6.64

Les arbres semblent avoir, dans les scénarios testés, une variance de l'EQM beaucoup plus élevée que celle des forêts aléatoires. Les différences les plus prononcées entre les deux méthodes

semblent être pour les scénarios avec un grand nombre de variables, surtout en addition d'une relation quadratique pour les variables. C'est notamment le cas pour les scénarios **GEQ** et **PEQ**. Une variance plus élevée de L'EQM peut indiquer une plus grande incertitude quant à la performance du modèle. Les forêts aléatoires pourraient donc avoir davantage de stabilité, surtout lorsqu'il s'agit de scénarios plus complexes. En revanche, pour les scénarios plus simples tels que **PPL** ou **PPQ**, bien que les forêts ont une variance plus petite, l'écart entre les deux variances demeure moins prononcé.

Tableau 5. Temps d'exécution par scénario

Scénario	**Rapport Arbre**	**Rapport Forêt**
scenario_GEL	0,03	0,63
scenario_GEQ	0,035	0,79
scenario_GPL	0,01	0,22
scenario_GPQ	0,011	0,24
scenario_PEL	0,008	0,073
scenario_PEQ	0,008	0,088
scenario_PPL	0,003	0,026
scenario_PPQ	0,0036	0,028