Projet d’étude

Intelligence Artificielle

Téo SAMUEL, Alice SPACCESI, Pierre CHEN et Maximilien BOURCIER

GE5 FISE

Février 2025

Table des matières

[Part 4 - Modeling river runoff using Artificial Neural Networks 3](#_Toc191489168)

[Question 1 3](#_Toc191489169)

[Question 2 3](#_Toc191489170)

[Question 3 3](#_Toc191489171)

[Question 4 3](#_Toc191489172)

[Question 5 3](#_Toc191489173)

[Question 6 4](#_Toc191489174)

[Part 5 - Modeling river flow using decision trees and their ensembles 4](#_Toc191489175)

[Question annexe 1 4](#_Toc191489176)

[Question annexe 2 4](#_Toc191489177)

[Question 1 4](#_Toc191489178)

[Question 2 4](#_Toc191489179)

[Question 3 4](#_Toc191489180)

[Question 4 5](#_Toc191489181)

[Question 5 5](#_Toc191489182)

[Part 6 - Modeling river runoff using Long Short-Term Memory (LSTM) 5](#_Toc191489183)

[Question 1 5](#_Toc191489184)

[Question 2 5](#_Toc191489185)

[Question 3 5](#_Toc191489186)

[Question 4 5](#_Toc191489187)

[Question 5 5](#_Toc191489188)

# Part 4 - Modeling river runoff using Artificial Neural Networks

## Question 1

La courbe rouge (MLP) représente les données prédit par le modèle tandis que la courbe bleue (observed) affiche les données « brutes » récupérées et directement affichées. A la lecture du graphique, on s’aperçoit que l’intervalle des données prédites est beaucoup moins conséquent. En effet, l’intervalle des données observées s’étend de zéro à plus de 50mm de pluie, tandis que les données prédites n’excèdent pas 20mm d’écart, soit plus d’un facteur deux de différence. De manière générale, les pics de millimètres de pluie sont sous-estimés dans le jeu de données prédit. Ils apparaissent sur la courbe, relativement aux dates correspondantes, mais affichent toujours des valeurs moindres comparées à celles observées, avec parfois des atténuation trois fois inférieurs à celles enregistrées : c’est le cas en 2009, lorsque plus de 60 mm de pluie étaient observés, mais seulement 20 mm étaient prédits. Ce modèle n’est donc pas optimal lorsque la plage de valeur à prédire est hétérogène et varie beaucoup.

## Question 2

Le modèle Random Forest (RF) est un modèle simple, tout comme notre modèle actuel, connu pour sa fiabilité et sa robustesse. Il est principalement apprécié pour sa facilité d’utilisation. Pour autant, il nécessite un jeu de données important, ce qui n’est pas forcément notre cas, et est donc chronophage. A l’inverse, notre modèle est non linéaire, donc capture très bien les relations entre les variables. Le temps de disponible et la quantité de données sont donc des critères de choix essentiels. A noter que ce modèle, au même titre que notre modèle actuel, est sujet aux mêmes « limites » et comportements indésirables cités question 3.

## Question 3

Plusieurs raison pourrait expliquer la faible efficacité du réseau de neurones. Tout d’abord on peut évoquer le surapprentissage, un comportement indésirable du machine learning. Il se produit lorsque le modèle fonctionne correctement sur les données d'entraînement mais pas sur les données de test, car le modèle ne généralise pas bien les nouvelles données. On peut aussi évoquer l’architecture du modèle. Bien que le modèle ait une seule couche cachée de 500 neurones, on peut considérer que le squelette n’est pas optimal pour la tache demandée. Aussi, il peut aussi s’agir simplement de la qualité insuffisante des données, empêchant une bonne corrélation des variables (température…). D’autres raisons peuvent intervenir (temps de convergence…).

## Question 4

Pour améliorer le modèle, et donc lutter contre les problèmes évoqués plus haut tel que le surapprentissage, on peut typiquement citer la validation croisée. On divise les données d'entraînements en plusieurs groupes pour optimiser le modèle. En créant des nouvelles variables qui connecteraient mieux celles déjà existantes on endiguerait la potentielle qualité médiocre des données. On pourrait peut-être aussi instaurer un système de validation progressive, pour checker l’efficacité du modèle pas à pas.

## Question 5

On utilise généralement les modèles profonds lorsque les données à traiter sont complexes, de par leur nature, leur relation ou leur présentation. Ici il s’agit de nombre (et non d’image typiquement), regroupé de manière tubulaire, simple. Théoriquement, un modèle peu profond devrait donc suffire, avec seulement une couche achée comme c’est le cas actuellement. Pour autant, étant donné que nous avons dit plus haut que le modèle n’était pas très performant tel quel, il serait peut-être judicieux de tester d’autre méthode, dont les modèles profonds.

## Question 6

Il faudrait opter pour un modèle simple, rapide et dont la généralisation est efficace. Etant donné que nous sommes face à une série de donnée temporelle, il serait surement judicieux d’utiliser un modèle Long Short-Term Memory (LSTM), qui est particulièrement adapté aux dépendances chronologiques. Pour autant, pour être efficace, le LSTM nécessite un grand volume de donnée, ce qui peut n’est potentiellement pas être notre cas. Malgré cet inconvénient, étant donné qu’il s’agit d’un modèle « sûr » et assez répandu, LSTM reste une des options les plus pertinentes à suivre.

# Part 5 - Modeling river flow using decision trees and their ensembles

## Question annexe 1

On peut conclure des deux résultats obtenus que la performance du modèle sur les données d’entrainement est plutôt bonne avec 81% de réussite alors que les résultats sur les données de test sont mauvais avec un résultat négatif de -0.5154. Ces résultats sont généralement caractéristiques d’un « overfitting » (surapprentissage) du modèle.

## Question annexe 2

De la même façon que pour la question 1, on obtient un résultat correct sur les données d’entrainement avec une réussite de 70%, alors que ceux des données de test sont négatifs (proche de 0). Ceci indique que ce score est mauvais pour notre cas d’usage.

## Question 1

D’après le graphique, la courbe d’apprentissage (en rouge) est correcte avec environ 60% même si cette valeur reste assez faible. Cependant, la courbe de validation (en bleu) reste négative, ce qui indique que le modèle généralise très mal les données. Les courbes restant assez constantes avec l’augmentation du nombre d’arbre indique que la performance ne s’améliore pas significativement avec le nombre d’arbre.

## Question 2

Non, ce modèle Random Forest n'est pas meilleur qu'un réseau de neurones (ANN) pour ce problème spécifique car le score de validation croisée reste négatif malgré l'optimisation. Un bon modèle doit avoir un score de validation positif.

## Question 3

Pour améliorer les performances du modèle, il est possible d’optimiser les hyperparamètres, c’est-à-dire tester pour différentes valeurs de « max-depth », « min\_samples », « sample\_lead » ou « max\_features ». Il pourrait également être possible de l’améliorer en normalisant les données ou en utilisant des techniques de régularisation tel que limiter la profondeur de l’arbre (réduire le nombre maximum de niveau de l’arbre décisionel).

## Question 4

Non, un modèle profond n'est probablement pas nécessaire pour ce problème. En effet, le problème semble venir d’un surapprentissage. Un modèle plus simple mais bien paramétré pourrait offrir de meilleures performances.

## Question 5

Si un modèle devait être utilisé, cela pourrait être soit un modèle de type Gradient Boosting ou de type LTSM.

En effet, les modèles Gradient Boosting sont souvent plus performants que Random Forest pour les données temporelles, ce qui est le cas ce lab. De la même manière, un modèle LTSM simple pourrait être plus efficace pour une approche profonde si l’on considère que la dimension temporelle est importante.

# Part 6 - Modeling river runoff using Long Short-Term Memory (LSTM)

## Question 1

Les scores R² obtenus sont très faibles : R² vaut 0,1347 pour les données d'entraînement, et 0,1369 sur les données de test. Ces valeurs sont très basses et indiquent que le modèle n'explique pas bien la variation des données. Un bon modèle devrait avoir un R² proche de 1. Ici, l'LSTM ne semble pas bien prédire le débit des rivières.

## Question 2

Non, le modèle LSTM ne semble pas être le meilleur pour ce problème. Il faudrait comparer avec d'autres modèles, comme Random Forest (RF), qui fonctionne bien sur les données tabulaires et hydrologiques, car il peut capturer des relations non linéaires.

Les réseaux de neurones classiques (ANNs) peuvent également être efficaces, mais ils nécessitent un bon réglage des hyperparamètres pour obtenir de bonnes performances. Quant au LSTM, bien qu'il soit adapté aux données séquentielles, la prédiction ici ne concerne que le mois suivant, ce qui limite son avantage en matière de mémoire à long terme.

## Question 3

Pour améliorer ce modèle, plusieurs approches sont possibles, comme améliorer les données (plus de variables), optimiser le modèle (optimiser le taux d’apprentissage, et éviter le surapprentissage) ou encore tester d’autres métriques pour comprendre l’erreur.

## Question 4

R² étant très bas, nous n’avons pas besoin d’un modèle profond. LTSM est utile pour les prédictions à long terme, ici on a besoin de modèles plus simples et rapides.

## Question 5

Le LSTM ne donnant pas de bons résultats, d’autres alternatives peuvent être envisagées. Parmi les modèles classiques, Random Forest (RF) capture efficacement les tendances, tandis que XGBoost et LightGBM sont souvent plus performants sur les données tabulaires. Pour les séries temporelles, ARIMA et SARIMA sont adaptés aux prévisions à court terme, et Prophet (Facebook) permet de modéliser les tendances saisonnières. Enfin, une approche hybride, combinant LSTM et Random Forest ou XGBoost, pourrait améliorer les performances.