Prédiction du niveau de nappes phréatiques : comparaison d'approches locale, globale et hybride

Lola Beuchée, Thomas Guyet, Simon Malinowski

Prédiction du niveau de nappes phréatiques : comparaison d'approches locale, globale et hybride

Lola Beuchée*, Thomas Guyet**, Simon Malinowski*

*Univ. Rennes, CNRS, Inria/IRISA
**Inria, Centre Inria de Lyon
thomas.guyet@inria.fr

Résumé. Cet article présente l'exploration d'une méthode autorégressive de prévision d'une série temporelle pour répondre au défi de la prédiction du niveau de nappes phréatiques. Une méthode autorégressive estime une valeur future d'une série temporelle par régression à partir des valeurs historiques de la série. Plusieurs méthodes de régression peuvent alors être employées. Dans cet article, on présente des expérimentations visant à identifier la meilleure configuration pour prédire de manière précise le niveau de nappes phréatiques. On compare pour cela différents prédicteurs, l'apprentissage de modèle par série ou par groupe de séries, et l'utilisation de données exogènes. Des expérimentations intensives ont été menées et nous permettent de conclure sur le choix de la méthode que nous utiliserons pour répondre au défi.

1 Introduction

Le défi EGC 2021 proposait d'explorer le niveau des nappes phréatiques (niveaux piézométriques) au travers des données historiques qui sont mises à disposition par le BRGM[1]. Nous nous sommes plus particulièrement intéressés à l'objectif de la prédiction de l'évolution de ces niveaux à moyen terme (i.e. plusieurs mois).

L'enjeu de la prédiction du niveau des nappes phréatiques est l'aide à la gestion responsable d'une ressource essentielle pour différents usages : alimentation humaine (2/3 des volumes d'eau destinée à l'alimentation humaine proviennent de ressources souterraines), irrigation, usages industriels (refroidissement, lavage, etc.) mais également la gestion des débits des cours d'eau (EGC, 2021; Rodriguez-Galiano et al., 2014).

La prédiction de ces niveaux représente un véritable défi du fait de la complexité des mécanismes hydrologiques à l'œuvre (Brédy et al., 2020). De nombreux modèles numériques ont été développés pour cette tâche. Certains modèles sont basés sur une modélisation physique de la nappe phréatique. Les modèles physiques nécessitent un ajustement des paramètres pour chaque situation (Nayak et al., 2006). Elles offrent une solution précise de prédiction mais difficilement généralisable à grande échelle. Pour les besoins de prédiction de manière plus systématique, l'utilisation de séries temporelles historiques des niveaux a été vue depuis des années comme un outil essentiel dans la planification des ressources en eau (Kisi et al., 2012).

[1] BRGM : Bureau des Recherches Géologiques et Minières

De nombreux modèles d'apprentissage automatique ont ainsi été développés pour prédire le niveau des nappes phréatiques. Par exemple, Brédy et al. (2020) proposent une modélisation à l'aide de forêt aléatoire (RF) ou d'Extreme Gradient Boosting (XGB). Ibrahem Ahmed Osman et al. (2021) comparent également un modèle basé sur XGB mais évaluent également des modèles SVR et des réseaux de neurones. Dans ces deux travaux, les données historiques utilisées comprennent d'une part l'historique des niveaux de nappe, mais également des informations relatives aux pluies et à l'évapotranspiration. Ces informations permettent de tenir compte des flux d'eau entrants et sortants dans une nappe.

Parmi ces nombreux modèles de l'état de l'art, il est difficile d'identifier celui qui est a priori le plus adapté à la prédiction de l'évolution du niveau des nappes phréatiques tel que mesuré par le réseau Hub'eau[2]. Il nous a donc semblé intéressant d'explorer des modèles variés de prévision de l'évolution de séries temporelles pour les données spécifiques mises à disposition et de les comparer pour identifier le meilleur.

Dans cet article nous nous sommes plus particulièrement intéressés à la question suivante : « est-il préférable d'apprendre des modèles par capteur (modèles locaux) ou d'apprendre un modèle global pour prévoir l'évolution des niveaux piézométriques ? ». En effet, les approches usuelles de prévision de séries temporelles sont conçues pour apprendre à prédire, à partir des mesures passées d'un capteur, les mesures à venir de ce même capteur. Mais un modèle appris sur chaque piézomètre nécessite d'avoir un historique de données important pour chacun d'eux. Ce n'est pas le cas pour la mise en place de nouveaux capteurs par exemple et, plus généralement, cela rend le modèle peu robuste aux changements du fonctionnement du système hydrologique mesuré. L'utilisation de modèles appris sur plusieurs piézomètres peut permettre plus de robustesse et une mise en œuvre plus aisée sur de nouveaux capteurs.

Notre exploration de modèles de prévision de l'évolution de séries temporelles est donc structurée par l'objectif de comparer trois types de modèles :

- des modèles *locaux* de prévision de série temporelle qui sont appris indépendamment sur chaque série temporelle (à partir des données historiques) pour prédire son évolution future.

- un modèle *global* de prévision unique appris à partir de l'ensemble des séries temporelles disponibles. Pour qu'un tel modèle puisse tenir compte des conditions spécifiques de chaque région, nous avons aussi pris en compte des informations sur la nature des sols qui influe la dynamique hydrologique.

- un modèle *hybride* qui consiste à identifier des groupes de séries temporelles pour lesquels les dynamiques d'évolutions sont homogènes. Un seul modèle est construit pour chaque groupe de piézomètre ainsi créé.

Pour chacun de ces types de modèles, nous avons proposé plusieurs implémentations possibles qui seront détaillées par la suite. Puis, pour sélectionner le meilleur modèle, nous avons comparé les RMSSE[3] de ces différents modèles soit de manière globale, soit spécifiquement sur les 18 piézomètres qui avaient été identifiés pour le défi.

[2]Hub'eau : `http://hubeau.eaufrance.fr/`.
[3]RMSSE : Root Mean Scaled Square Error.

2 Matériel et méthodes

Dans cette section, nous présentons tout d'abord le jeu de données que nous avons utilisé, ainsi que les prétraitements qui ont été faits. Ensuite, nous présentons les méthodes de prédiction de séries temporelles qui ont été mises en œuvre.

2.1 Constitution des jeux de données

Pour ce travail, nous avons constitué un jeu de données correspondant à la période de janvier 2015 à janvier 2021 (2221 jours) pour tous les piézomètres de France métropolitaine présentant des séries temporelles journalières du niveau de nappe pour lesquelles il y a moins de 50 jours de données manquantes. Les valeurs manquantes ont été imputées par interpolation linéaire. On obtient alors une base de données de 1339 séries temporelles (i.e. piézomètres). Parmi ces piézomètres, on distingue le groupe de 18 séries temporelles auxquelles le défi s'intéresse plus particulièrement.

En complément, nous avons également collecté avec des informations sur la nature du sol issues de la BD LISA[4] (thème, nature, type de milieu et état), ainsi que deux séries temporelles journalières sur la pluviométrie et l'évapotranspiration[5] (ETO) pris sur la maille spatiale du piézomètre (mailles dont la taille[6] est respectivement de 0.25^o et 0.1^o).

Au final, pour chaque piézomètre on dispose de trois séries temporelles journalières (le niveau de nappe, la pluviométrie et l'évapotranspiration), ainsi que quatre caractéristiques de sol.

2.2 Prévision de séries temporelles

La base de données de piézomètres est un ensemble de triplets $\mathcal{Y} = \langle Y^k, \boldsymbol{Z}^k, F^k \rangle$ où $Y^k : y^k_{1\ldots t}$ une série temporelle univariée telle que $y^k_i \in \mathbb{R}$ pour tout i, et $\boldsymbol{z}^k_{1..\infty}$ une série temporelle multivariée ($Z : \boldsymbol{z}^k_i \in \mathbb{R}^m$), dite exogène, connue jusqu'à la date t mais également au delà. La série temporelle Y^k est la série du niveau de nappe pour le k-ième piézomètre tandis que la série de données exogènes correspond à la pluviométrie et à l'évapotranspiration. F correspond aux informations disponibles sur le type de sol (4 informations).

La prédiction d'une série temporelle Y^k à l'horizon h consiste à estimer $y^k_{t+1\ldots t+h}$. Pour cette tâche de prévision de série temporelle, l'approche classique en analyse de données est de construire un modèle autorégressif. Une telle méthode construit une fonction de prédiction de la valeur à la date t_0 à partir des r dernières observations de Y et de $\boldsymbol{Z}$, des valeurs de $\boldsymbol{Z}$ à la date t_0, et éventuellement des caractéristiques de la série temporelle (F). On dénote par $\varphi : \mathbb{R}^{r\times(m+1)+f} \mapsto \mathbb{R}$ une telle fonction de prédiction de la valeur suivante de la série. φ peut alors être vue comme un régresseur : elle prédit une valeur réelle à partir des caractéristiques d'entrée. Pour réaliser une prévision à un horizon h, la fonction de prédiction est récursivement appliquée h fois.

On peut noter que les séries temporelles exogènes sont supposées connues dans le futur. Ce problème est donc différent de celui d'effectuer la prévision d'une série temporelle multivariée pour laquelle la fonction de prédiction aurait pour sortie $m + 1$ valeurs (la valeur de

[4]BD LISA : `https://bdlisa.eaufrance.fr/`

[5]Données issues du Climate Data Store : `https://cds.climate.copernicus.eu`

[6]La taille des mailles est donnée en écart angulaire mesuré en degré, $0.1^o \approx 9\ km$

la série cible ainsi que les m valeurs des séries exogènes). Dans notre problème, nous avons fait l'hypothèse que la pluviométrie et l'évapotranspiration étaient semblables d'une année à l'autre. Par conséquent, les prévisions peuvent être obtenues dans le futur en prenant les valeurs journalières moyennes des années passées. L'erreur d'approximation faite par cette hypothèse nous semble préférable à celle qui serait faite par le cumul des erreurs d'une prédiction récursive multivariée à moyen terme. Pour conforter cette hypothèse, la Section 3.2 compare des résultats obtenus sur des données historiques en utilisant l'approximation à ceux utilisant la valeur réelle des séries exogènes.

2.3 Apprentissage des modèles

Pour définir la méthode d'apprentissage d'une fonction φ telle qu'introduite précédemment, nous avons besoin d'une part de construire un jeu d'apprentissage et, d'autre part, de choisir un type de modèle de régression.

Pour une série temporelle $\langle Y^k, \boldsymbol{Z}^k, F^k \rangle$ dont on cherche à faire la prévision à partir de la date t_0 en utilisant r valeurs dans le passé, le jeu d'apprentissage de la fonction φ est constitué d'exemples

$$(y_{t-r}, \ldots, y_t, z^1_{t-r}, \ldots, z^m_t, f_1, \ldots, f_4), \forall t \in [r, t_0].$$

où y est la variable à prédire et les autres variables sont des variables explicatives.

Pour la construction du jeu d'apprentissage, nous avons considéré deux approches différentes selon que l'on souhaite construire un modèle local, global ou hybride.

- un modèle *local* est construit à partir d'une série temporelle et appliqué sur cette même série pour effectuer une prévision. Il y a donc un modèle local par série temporelle.
- un modèle *global* est construit à partir de l'ensemble des séries temporelles et cet unique modèle est appliqué à toutes les séries dont on souhaite les prévisions.
- un modèle *hybride* est construit pour chaque groupe homogène de séries temporelles. Pour une nouvelle série, on applique le modèle du groupe dont la série est la plus proche. La section suivante détaille comment les groupes homogènes sont constitués.

Pour le choix du type de régression, nous avons exploré les modèles suivants :

- Régression linéaire. C'est la méthode standard pour les problèmes de régression. Cette méthode est alors équivalente à celle qui consiste à utiliser un modèle AR d'ordre r.
- SVR (Awad et Khanna, 2015). Les SVR permettent de traiter des problèmes en grande dimension et proposent des modèles non-linéaires. Nous avons utilisé des noyaux Gaussiens. Après quelques expérimentations, nous avons fixé le paramètre $C = 100$.
- Forêts aléatoires (Breiman, 2001). C'est une méthode d'apprentissage à ensemble (*bagging*). Nous avons opté pour une forêt de 100 arbres.
- *Extreme Gradient Boosting* (Chen et Guestrin, 2016). C'est une méthode à ensemble à base d'arbres qui utilise des techniques d'optimisation pour améliorer l'efficacité calculatoire de l'apprentissage du modèle. Elle peut ainsi traiter de gros jeux de données.

En complément de ces méthodes d'apprentissage, nous avons également exploré le modèle classique de prévision de séries temporelles ARIMA (*Auto-Regressive Integrated Moving Average*) qui peut être vu comme un modèle local dans notre classification.

2.4 Construction de groupes homogènes de séries temporelles

L'approche hybride vise à construire un modèle de prévision pour des groupes de séries temporelles. On souhaite pour cela utiliser une représentation des séries temporelles qui capte une information sur la dynamique de l'évolution de ces séries. Pour cela, on utilise les caractéristiques suivantes :

1. la valeur des paramètres d'un modèle ARMA d'ordre 3 estimé sur la série temporelle Y avant t_0. Ces paramètres servent à représenter la dynamique d'évolution d'une série temporelle.

2. les moyennes et écarts types de la série temporelle Y avant t_0 et

3. les descriptions du sol issues des 4 paramètres de la BD LISA ($f_1, \ldots, f_4$).

Les attributs catégoriels ont été vectorisés de sorte qu'un piézomètre soit représenté par un vecteur réel de dimension 26. Chacun de ces attributs est z-normalisé pour leur donner une égale importance. Les données sont catégorisées par l'algorithme des k-*means*. Les valeurs de k sont testées entre 2 et 15, et le choix du meilleur k est fait selon le critère BIC.

On peut noter que l'approche globale est un cas particulier de l'approche hybride où toutes les séries sont dans le même cluster ($k = 1$).

2.5 Comparaison des modèles

Pour le choix de la meilleure méthode à adopter pour effectuer des prévisions, le défi proposait de comparer les approches en utilisant les RMSSE pour chaque méthode et chaque série pour un horizon de trois mois ($h = 93$) :

$$RMSSE = \sqrt{\frac{\frac{1}{h}\sum_{t=n+1}^{n+h}\left(y_t - \hat{y_t}\right)^2}{\frac{1}{n-1}\sum_{t=2}^{n}\left(y_t - y_{t-1}\right)^2}}$$

Plus précisément, le calcul des RMSSE est fait uniquement sur la période du 15 octobre jusqu'au 15 janvier 2020. En effet, en fonction de la période de l'année, la prédiction est plus ou moins aisée (en période sèche, les nappes phréatiques sont moins impactées par les pluies par exemple). On a donc choisi d'évaluer nos modèles sur la période de l'année que le défi imposait.

À partir du calcul de l'ensemble des RMSSE, nous calculons les moyennes des RMSSE sur l'ensemble des séries et plus spécifiquement sur les 18 séries à prédire (la Section 3 ne présente les résultats que pour l'ensemble des séries, plus de résultats sont disponibles dans le dépôt logiciel du projet). En complément, on réalise également des comparaisons en s'appuyant sur des diagrammes de différences critiques (Demšar, 2006). Ces diagrammes s'appuient sur des statistiques de rang (test de Nemenyi) pour comparer des méthodes.

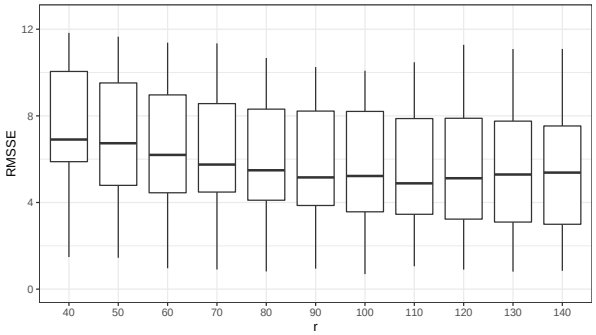

FIG. 1 – *RMSSE moyennes sur les 18 piézomètres à tester en fonction de la taille d'historique.*

3 Expérimentations et résultats

L'ensemble des modèles ont été mis en œuvre et testés à l'aide du logiciel `R`.[7] Pour les modèles ARIMA, forêts aléatoires, SVR et XGBoost, nous avons respectivement utilisé les librairies `forecast`, `randomForest`, `e1071` et `XGBoost`.

Dans une premier temps, nous nous intéressons au choix de la taille de l'historique. Nous détaillons ensuite les résultats obtenus avec les modèles locaux, puis la Section 3.3, nous donnons les résultats obtenus avec les approches globales et hybrides.

3.1 Choix de la taille de l'historique

Le choix de la taille de l'historique (paramètre r) doit tenir compte 1) de l'horizon de prévision, 2) des caractéristiques de la série (notamment des corrélations entre la série à prédire et les séries exogènes) et 3) de la dimension des exemples d'apprentissage. Les analyses de corrélation croisée entre les séries exogènes et la série à prédire montre une forte corrélation pour un décalage de 100 jours : ceci laisse penser que disposer de l'information d'au moins 100 jours ($r > 100$) dans le passé aidera un régresseur à faire des prévisions précises. L'horizon de prévision est fixé à $h = 93$. L'utilisation d'une procédure récursive de prévision pas à pas incite à utiliser une taille d'historique $r > h$ de sorte qu'aucune prédiction ne se fasse sans s'appuyer sur aucune donnée réelle (*i.e.* pas uniquement à partir de données elles-mêmes préalablement prédites). Ces deux contraintes incitent donc à avoir un $r > 100$. Néanmoins, une valeur trop élevé de r conduira à des exemples en très grande dimension, limitant les performances de certaines méthodes d'apprentissage.

Pour finaliser notre décision, nous avons mené une étude expérimentale sur les 18 piézomètres de test en testant des valeurs de r de 40 à 140 pour le modèle local de régression linéaire. Les résultats de cette expérimentation sont illustrés dans la Figure 1. On observe sur cette figure que les meilleurs résultats médians sont obtenus pour $r = 100$ ou 110. Pour limiter la dimensionnalité des exemples, nous avons choisi $r = 100$ pour la suite des expérimentations.

[7] L'ensemble des développements du projet peut être consulté sur le site `https://gitlab.inria.fr/tguyet/projetpiezo`.

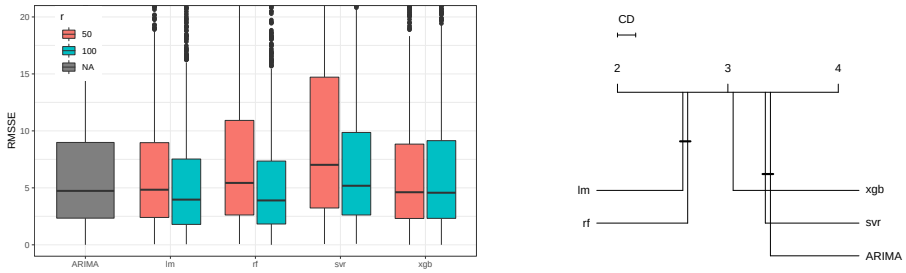

FIG. 2 – *Sur la gauche : RMSSE moyennes sur l'ensemble des piézomètres à tester en fonction des types de régresseur. On compare deux tailles d'historique :* $r = 50$ *(en rouge) et* $r = 100$ *(en bleu). Le modèle ARIMA n'ayant pas de paramètre* r*, seule une valeur est reportée. Sur la droite : diagramme de différence critique comparant les différentes méthodes sur tous les piézomètres (*$r = 100$*).*

3.2 Analyse des résultats des modèles locaux

Comparaison des modèles individuels On commence par comparer les résultats obtenus avec les différents régresseurs sur l'ensemble des piézomètres avec $r = 50$ et $r = 100$.

La Figure 2 illustre les résultats obtenus sous la forme d'un graphique et d'un diagramme de différences critiques avec un niveau de confiance $\alpha = 5\%$ (pour $r = 100$ uniquement).

Ces résultats montrent d'une part que l'utilisation d'un historique de 100 est meilleur qu'un historique de 50 pour tous les régresseurs. D'autre part, on identifie deux régresseurs comme potentiellement plus précis que les autres : la régression linéaire et les forêts aléatoires (non statistiquement différents). Le modèles ARIMA et SVR donnent des RMSSE qui sont significativement supérieures aux autres (selon le test de Nemenyi).

Apport des séries exogènes On cherche maintenant à montrer l'apport de l'utilisation des données exogènes pour améliorer la précision des prédictions. Pour cela, on compare les résultats de RMSSE de différents modèles dans trois configurations : prédiction sans données exogènes, avec la pluie seule, avec l'évapotranspiration (ETO) seule et avec la pluie et l'ETO. Dans ce cas, on a pris $r = 100$. On se place ici la situation idéale où on connaît les séries exogènes réelles.

La Figure 3 illustre les résultats obtenus sur l'ensemble des piézomètres. On constate que l'utilisation de séries exogènes diminue fortement les performances des modèles. L'utilisation de l'information combinée de pluie et évapotranspiration limite cette diminution, mais ne permet pas d'amélioration. Une explication de ces mauvais résultats est l'augmentation de la dimension des exemples qui rend l'apprentissage plus difficile (plus de 300 caractéristiques avec les deux variables exogènes). Les modèles comme les forêts aléatoires et les SVR se montrent plus performants pour ce type de données.

Perte de précision liée à l'utilisation de données exogènes approximées Finalement, bien que l'ajout de données exogènes ne se montre pas utile, nous avons comparé les résultats

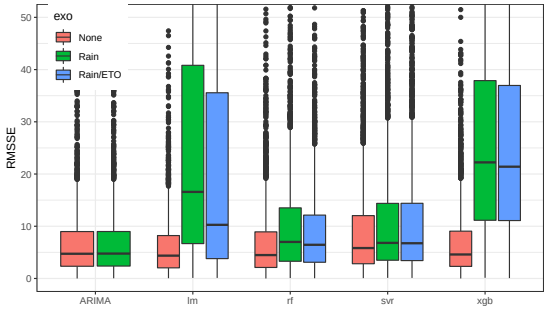

FIG. 3 – *RMSSE moyennes sur l'ensemble des piézomètres en fonction de la taille d'historique.*

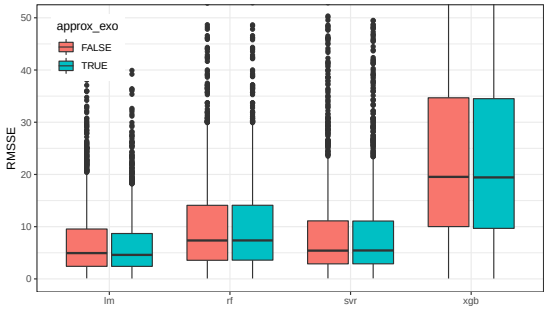

FIG. 4 – *RMSSE moyennes sur l'ensemble des piézomètres avec ou sans approximation des variables exogènes.*

obtenus en utilisant des données exogènes réelles ou bien approximées (valeurs moyennes dans le passé). Ces dernières correspondent mieux aux contraintes pratiques d'utilisation.

La Figure 4 illustre les résultats de RMSSE avec ou sans approximation des données exogènes. On constate que quelque soit le régresseur, les performances sont très semblables avec ou sans approximation. Ce résultat est donc intéressant : il est probable que si une méthode exploite bien les données exogènes, elle puisse indifféremment utiliser les données approximées à la place des données réelles.

3.3 Analyse des résultats des approches hybride ou globale

Dans cette partie, on s'intéresse aux résultats obtenus par les approches globale et hybride. Pour des raisons de ressources de calcul, seuls les modèles basés sur la régression linéaire et XGB ont pu être calculés.[8] Nous nous sommes également limité aux configurations $r = 100$

[8]Les données pour l'ensemble des 1339 piézomètres avec un historique de $r = 100$ dans une série de 2221 jours nécessite environ 20Go de mémoire RAM. L'utilisation de serveurs de calculs disposant de 32Go de RAM ne permettant pas de contenir les informations nécessaires pour les modèles d'arbre de décision ou de SVR.

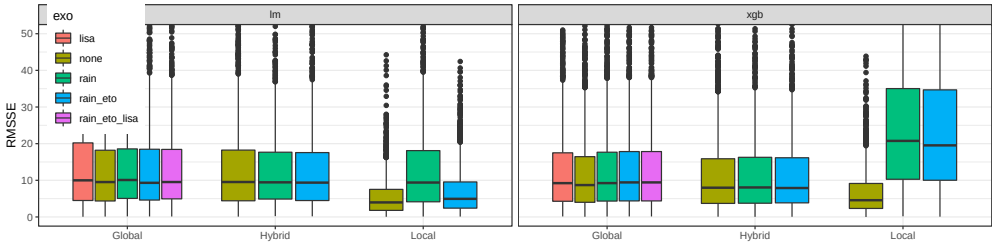

FIG. 5 – *Comparaison des RMSSE en fonction de l'utilisation des données exogènes, selon le type d'approche (globale, hybride ou locale) et selon le classifieur (XGBoost,* `xgb` *ou régression linéaire,* `lm`*).*

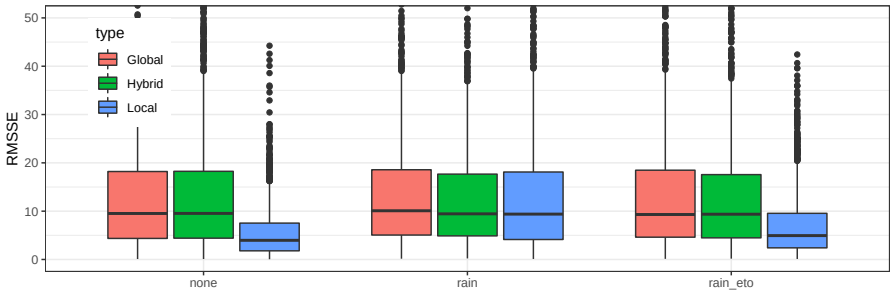

FIG. 6 – *Comparaison des RMSSE en fonction de l'utilisation du type d'approche (globale, hybride ou locale) pour la régression linéaire.*

et sans approximation des valeurs exogènes. Pour les comparaisons avec les modèles locaux, on ne prendra donc que ces configurations en compte.

Pour l'approche hybride, le meilleur nombre de cluster a été estimé à 8 selon le critère BIC. On obtient ainsi 8 groupes de séries temporelles de tailles 426, 266, 181, 132, 106, 65, 11 et 5.

Pour l'approche globale, on utilise l'information des caractéristiques des sols (BD LISA) comme des variables explicatives pour la régression. L'objectif est de voir si il était préférable d'utiliser cette information pour catégoriser les séries temporelles ou bien de l'utiliser directement dans la régression.

La Figure 5 donne les résultats obtenus pour les différentes approches en fonction de l'utilisation de différents jeux de caractéristiques exogènes dans la régression (pluie, pluie+ETO, BD LISA, pluie+ETO+BD LISA). Pour les modèles hybride et global, on constate que quelles que soient les données exogènes disponibles, les performances sont semblables. En particulier, l'utilisation de l'information sur le sol (BD LISA) ne permet pas de réduire les RMSSE. En comparaison avec le modèle local pour lequel l'utilisation d'information exogène était pénalisante (en particulier avec XGBoost), les modèles utilisant plusieurs piézomètres sont plus robustes à l'utilisation de nouvelles caractéristiques. Ceci est probablement dû au plus grand nombre de données disponibles pour ajuster les paramètres additionnels.

La Figure 6 compare les RMSSE obtenues pour les différentes approches avec la régression linéaire. Elle est détaillée pour chaque sous jeu de caractéristiques qui ont été expérimentés en

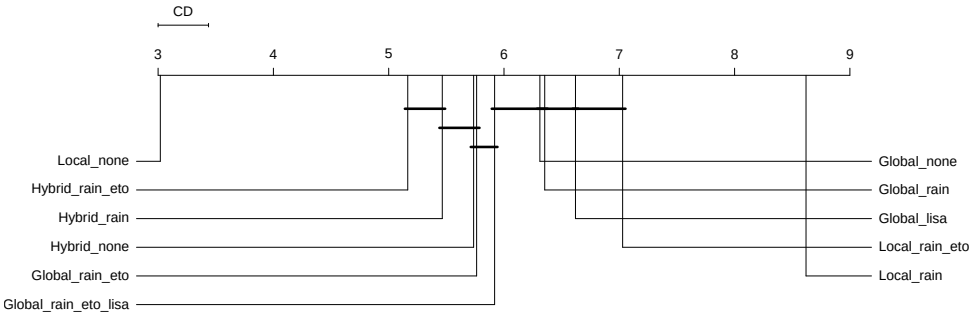

FIG. 7 – *Diagramme de différences critiques comparant les RMSSE en fonction de l'utilisation du type d'approche (globale, hybride ou locale) avec différents jeux de caractéristiques exogènes. La méthode d'apprentissage utilisée est ici la régression linéaire.*

commun. Cette figure illustre le fait qu'en moyenne, les approches globale et hybride ont des précisions semblables. En revanche, avec l'utilisation de données exogènes de pluie et d'évapotranspiration, ou sans données exogènes, on constate que l'approche locale a une meilleure précision.

Pour conforter ce résultat obtenu sur des agrégats de RMSSE, la Figure 7 donne le diagramme de différences critiques obtenus en comparant pair à pair les RMSSE pour chaque couple : type d'approche et jeu de caractéristiques exogènes. On a ici fixé l'utilisation de la régression linéaire, que nous avons vu précédemment comme semblant être la meilleure approche.

Ce diagramme montre clairement la supériorité de l'approche locale, sans données exogènes. Son rang moyen est proche de 3 (sur 11 méthodes comparées). Elle est donc très régulièrement classée parmi les toutes meilleures approches. Le diagramme montre que cette différence est significative statistiquement. Étonnamment, le moins bon modèle est obtenu avec une configuration proche : l'approche locale avec l'utilisation de l'information de pluie. Comme on l'a vu précédemment, cela s'explique probablement par le manque de données suffisantes pour ajuster les nombreux paramètres induits par l'utilisation de cette nouvelle information. Les autres approches offrent globalement des performances similaires entre elles (classements moyens entre 5 et 7).

3.4 Comparaison sur le jeu entier contre les 18 séries de test

Nous concluons cette partie expérimentale en comparant les résultats obtenus sur les 18 séries temporelles de test par rapport aux résultats précédents, obtenus pour l'ensemble du jeu de données. En effet, il faut vérifier que le choix du meilleur modèle sur la base de l'ensemble du jeu de données sera potentiellement le meilleur également pour le sous ensemble de données qui sera utilisé pour le défi.

La Figure 8 illustre les différences de RMSSE obtenues pour l'approche locale sans variable exogène. On constate que lorsqu'on s'intéresse aux 18 piézomètres à prédire, les résultats sont du même ordre que pour toute la base, voire légèrement inférieure, quelque soit

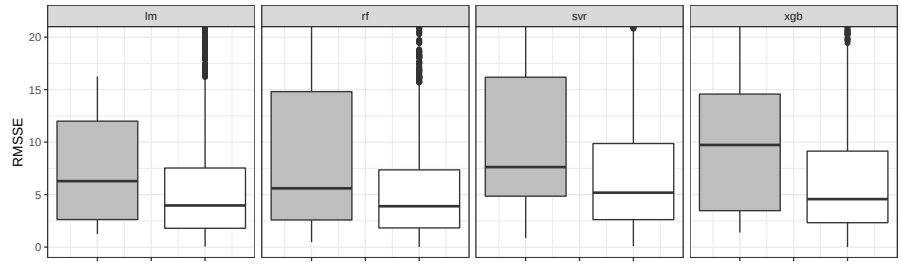

FIG. 8 – *Comparaison des RMSSE pour la régression linéaire des 18 piézomètres à tester (à droite, en blanc) contre tous les piézomètres (à gauche, en gris).*

le régresseur utilisé. Le défi a donc retenu des séries temporelles plutôt plus faciles à prédire que le reste de la base de données. Le meilleur régresseur reste la régression linéaire pour ce sous-ensemble.

4 Conclusion

En conclusion, notre approche du défi a consisté à expérimenter différentes approches de prédiction de séries temporelles en se basant sur une technique d'autorégression : les données historiques de la série temporelle permettent d'apprendre un régresseur pour prédire la valeur suivante de la série. La prévision de la série est menée en appliquant récursivement le régresseur appris.

Pour le choix et l'apprentissage du régresseur, nous avons mis en place un protocole pour identifier le meilleur choix. Nous avons comparé quatre types de classifieur (linéaires, SVR, forêts aléatoires et XGB) et expérimenté la construction d'une base d'apprentissage de trois manières : locale, globale et hybride. Au travers de l'approche hybride, nous souhaitions améliorer la précision des prévisions en créant des sous groupes homogènes de séries temporelles. Finalement, nous avons également expérimenté l'utilisation de variables exogènes (pluie, évapotranspiration et caractéristiques de sol).

En pratique, des expérimentations extensives ont été menées sur les données afin d'identifier la solution que nous emploierons pour répondre au défi. La solution qui est retenue est finalement la plus simple : elle consiste faire un modèle de régression par série temporelle avec un historique de 100 valeurs, et sans variables exogènes.

Le code mis en place est un code flexible et facilement adaptable à l'utilisation et l'expérimentation de nouveaux régresseurs. Une première perspective de ce travail serait donc d'affiner l'utilisation des régresseurs les plus avancés (notamment XGBoost) en ajustant leurs hyper-paramètres de sorte à en tirer les meilleures performances.

Une seconde amélioration possible serait d'améliorer la construction des groupes de séries temporelles dans l'approche hybride. L'absence de différence avec l'approche globale laisse penser que la constitution des groupes n'identifie pas les bonnes caractéristiques pour faciliter l'apprentissage d'un bon modèle commun. Plus d'expérimentations pourraient être menées de ce côté sur les représentations des piézomètres et l'algorithme de clustering.

Finalement, nous avons conduit ici une analyse purement guidée par les données. La littérature montre néanmoins qu'une connaissance applicative peut donner de meilleurs résultats, notamment en tenant compte des délais d'impact entre les variables, ou leur durée d'intégration. L'utilisation de connaissance expertes devrait donc permettre d'améliorer les modèles.

Références

Awad, M. et R. Khanna (2015). Support vector regression. In *Efficient learning machines*, pp. 67–80.

Brédy, J., J. Gallichand, P. Celicourt, et S. J. Gumiere (2020). Water table depth forecasting in cranberry fields using two decision-tree-modeling approaches. *Agricultural Water Management 233*, 106090.

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Chen, T. et C. Guestrin (2016). XGBoost : A scalable tree boosting system. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pp. 785–794.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research 7*(1), 1–30.

EGC (2021). Défi EGC 2022 : prévoir l'évolution du niveau de nos nappes phréatiques.

Ibrahem Ahmed Osman, A., A. Najah Ahmed, M. F. Chow, Y. Feng Huang, et A. El-Shafie (2021). Extreme gradient boosting (XGBoost) model to predict the groundwater levels in Selangor Malaysia. *Ain Shams Engineering Journal 12*(2), 1545–1556.

Kisi, O., J. Shiri, et B. Nikoofar (2012). Forecasting daily lake levels using artificial intelligence approaches. *Computers & Geosciences 41*, 169–180.

Nayak, P. C., Y. S. Rao, et K. Sudheer (2006). Groundwater level forecasting in a shallow aquifer using artificial neural network approach. *Water resources management 20*(1), 77–90.

Rodriguez-Galiano, V., M. P. Mendes, M. J. Garcia-Soldado, M. Chica-Olmo, et L. Ribeiro (2014). Predictive modeling of groundwater nitrate pollution using random forest and multisource variables related to intrinsic and specific vulnerability : A case study in an agricultural setting (southern spain). *Science of the Total Environment 476*, 189–206.

Summary

This paper investigates an auto-regressive method for time series forecasting and applied to the challenge of predicting groundwater levels. An auto-regressive method estimates a next value of a time series by regression with the historical values of the series. Several regression methods can be used. In this paper, experiments are presented to identify the best setting to accurately predict the groundwater levels. Different classifiers, different modes of learning (by series or by group of series), and different usages of exogenous data are compared. Intensive experiments have been conducted and allow us to conclude on the best method we will use to answer the challenge.