

Rapport du projet d'Analyse de données

Arthur BOCAGE, Guillaume LETELLIER, Corentin PIERRE et Alexandre PIGNARD

22 février 2022

Introduction

Ce travail porte sur une base de données comportant 93 832 observations sur la Touque, un fleuve normand prenant sa source dans le département de l'Orne et se jettant dans la Manche au niveau de Deauville. Il s'agit de températures relevées du 29 mai 2013 au 5 octobre 2018 sur quatre stations classées d'amont en aval, la dernière étant la plus proche de la mer comme on peut le voir sur la figure 1. Les variables que nous utiliserons sont entre autres l'heure et date de l'observation, la température de l'eau, de l'air et la pluviométrie le tout en fonction des sondes qui auront relevées ces données.

De part cette base, une problématique peut d'ores et déjà se former, problématique qui orientera donc nos analyses et interprétations : **Quels sont les facteurs qui pourraient influencer la température de la Touque en fonction de la géologie et du climat près des sondes ?**

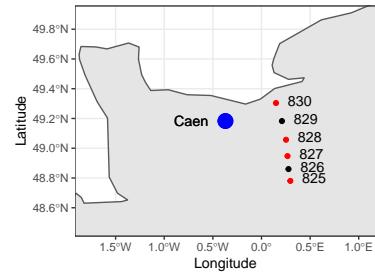


Figure 1: Placement des sondes de la Touques

Analyse

Afin de répondre aux questions que cette problématique soulève, nous allons avoir recours à différentes méthodes d'analyse de données comme l'utilisation de statistiques descriptives, l'analyse en composantes indépendantes et l'analyse en composantes principales. En ajoutant des recherches sur internet, nous pourrons efficacement analyser et interpréter les résultats retournés par ces méthodes.

Statistiques descriptives

Premières observations sur les données

Nous avons tout d'abord décidé de réaliser un bref résumé des données à notre disposition (table 1).

On remarque déjà que les moyennes et médianes des températures de l'eau et de l'air sont très proches et l'on peut d'ores et déjà supposer que ces variables sont peut être corrélées. Pour ce qui est de la volumétrie on ne peut pour l'instant rien supposer compte tenu des valeurs que l'on dispose. On peut noter aussi que la pluviométrie a un coefficient de variation élevé (180 %) mais que celui des températures sont assez basses

Table 1: Résumé des variables

	Min	Max	Q1	50%	Q3	Mean	Geometric mean	Variation coeff
Teau	0.632	22.968	8.817	11.807	14.481	11.671043	11.03716	0.2973760
Tair.EOBS	-5.420	28.950	7.580	11.920	16.230	11.854044	10.71322	0.4674562
Rainf.EOBS	0.000	33.700	0.000	0.000	2.900	2.042345	0.00000	1.7937573

(30% pour l'eau et 47% pour l'air), impliquant qu'il y a peut être des phénomènes distincts à prendre en compte, notamment entre les températures de l'air/eau et la pluviométrie.

Pour en apprendre plus, nous avons choisi de calculer une matrice des corrélations. En effet, elle peut permettre d'affirmer ou d'infirmer s'il y a une corrélation entre deux variables. Ici, nous avons décidé de corriger les données transmises pour les calculs de la matrice des corrélations car, comme dit en introduction, les années 2013 et 2018 ne sont pas complètes, entraînant donc un biais sur certaines valeurs. On peut remarquer qu'il n'y a pas de corrélation entre les températures de l'air/eau et la pluviométrie. Néanmoins, on peut souligner une très légère corrélation entre le mois de l'année et les températures de l'air et de l'eau. En effet, cela est due aux saisons que nous avons sur Terre (chaud en été, froid en hiver dans l'hémisphère Nord). Et enfin, une corrélation est avérée entre la température de l'eau et celle de l'air.

Cela est à néancer car ceux sont des données moyennées sur les quatre sondes. Qu'en est-il vraiment pour chaque sonde ? Pour cela, nous avons effectué des régressions linéaires afin d'éclaircir un peu plus ce point.

Régressions linéaires

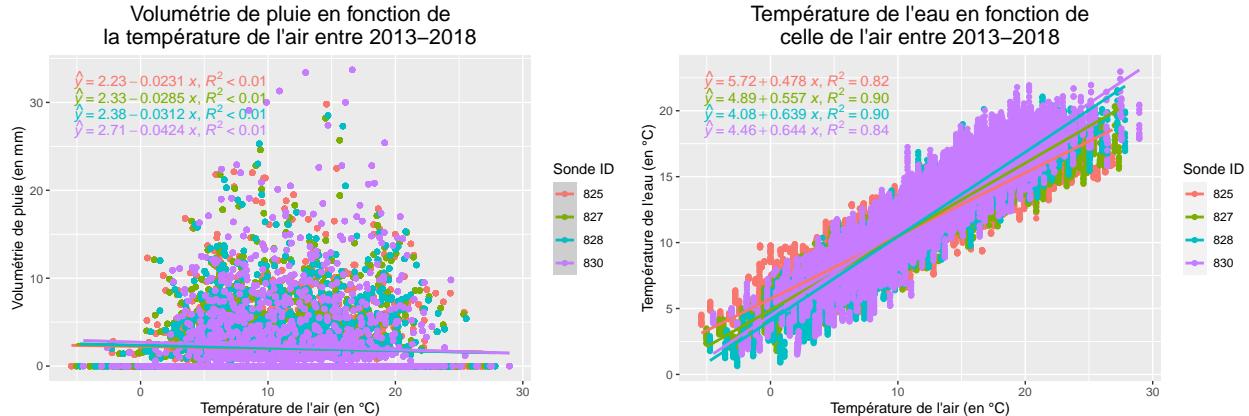
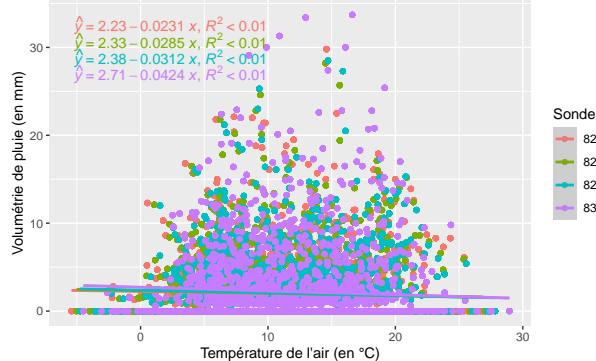
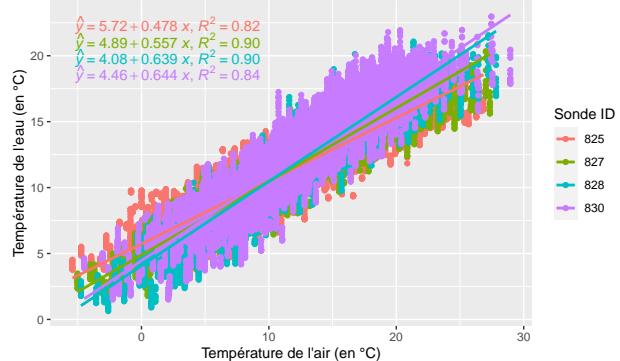


Figure 2: Matrice des corrélations corrigée

Volumétrie de pluie en fonction de la température de l'air entre 2013–2018



Température de l'eau en fonction de celle de l'air entre 2013–2018



Pour commencer, voici un rappel : une régression linéaire est un modèle qui cherche à établir une relation linéaire entre une variable (dite expliquée) et une ou plusieurs autres variables (dites explicatives).

Le constat saute aux yeux et est sans équivoque pour la figure 3 : il n'y a aucune corrélation entre ces variables et comme R^2 est compris entre 0 et 0.01, on peut en conclure que le volume de pluie tombé et la température de l'air sont deux phénomènes distincts (indépendants). Ceci est logique car la pluie est causée par une agrégation de vapeurs d'eau sous forme de nuages en haute atmosphère alors que la température de l'air est plutôt due au temps d'ensoleillement, l'altitude ou encore les saisons (inclinaison de la Terre sur son orbite de rotation).

La figure 4 est plus intéressante car elle permet de mettre en lumière la corrélation entre la température de l'eau et celle de l'air que nous avons évoqué plus tôt et cela pour chaque sonde. Le R^2 est compris entre 0.82 et 0.90 impliquant une certaine dépendance entre les variables et un motif se démarque de façon visible et clair. On voit que plus la température de l'air est élevée, plus la température de l'eau l'est aussi.

On remarque aussi une différence dans la disposition des points associés à chaque sonde. En effet, plus la sonde est en amont et plus la dispersion des points à l'air d'être horizontale. On peut observer cette différence avec les quatre régressions linéaires car en-dessous de 10°C pour l'air et l'eau, l'ordre des droites s'inversent. Cela suggère qu'il y a peut être une autre variable non disponible dans les données qui influence la température de l'eau à un niveau différent pour chacune des sondes.

Table 2: Matrice A retournée par l'algorithme fastICA

	825	827	828	830
C1	-2.433302	-3.061071	-3.512530	-3.464621
C2	-1.768637	-1.048395	-1.141248	-1.151607

ACI

On a pu voir que les variations de températures de l'eau sont majoritairement dues à la température de l'air, mais aussi potentiellement par quelque chose d'autre. Cela peut être la pluviométrie (ce n'est pas le cas ici, voir section sur les statistiques descriptives), la température des affluents de la Touques, le temps d'ensoleillement, la vitesse, la température et la direction du vent ou bien encore l'humidité. Il nous faut donc utiliser un outil permettant d'expliciter les différentes sources qui influent sur la température d'eau de la Touques. L'analyse en composantes indépendantes permet de réaliser cette action. Elle extrait des signaux statistiquement indépendants (décorrélés) à partir de signaux composites permettant ainsi de résoudre le problème de séparation de sources.

Pour cela, nous avons exécuté l'algorithme **fastICA** du package **fastICA** sur les moyennes journalières des températures d'eau des quatre sondes. Avant d'afficher les résultats, il est utile de préciser que les coefficients négatifs rendant plus difficile l'interprétation des signaux sources, nous avons donc décidé d'inverser automatiquement les signaux (multiplié par -1) pour les signaux où les coefficients de A sont négatifs. On obtient ainsi la matrice A (table 2) et les signaux sources (figure 5).

Tout d'abord, on peut rapidement penser que la composante 1 correspond à la température de l'air ou s'en rapproche très grandement. Effectivement, nous pouvons affirmer cela car au niveau climatique dans notre hémisphère, il y a des maximums locaux en été et des minimums locaux en hiver. Le signal correspond effectivement à cette observation. Cependant pour la seconde, l'interprétation est plus difficile. On peut voir ici qu'il y a un déphasage de la seconde composante par rapport à la première car par exemple, quand il y a un minimum très bas ou un maximum très haut dans C1, C2 a tendance à avoir un plateau (bas ou haut) quelques mois après durant plusieurs mois.

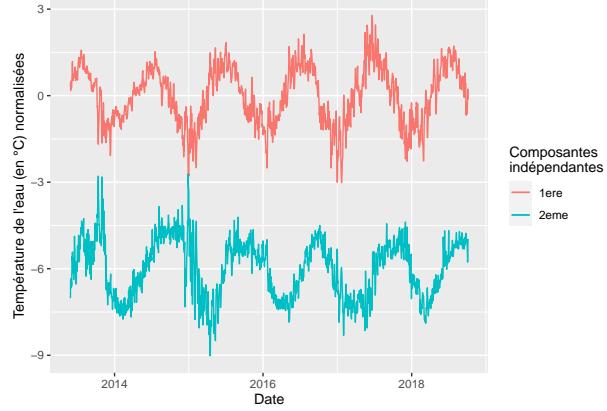


Figure 3: Signaux sources - C2 a été centrée à -6°C par soucis de visualisation et d'interprétation

On peut donc se demander à quoi cela est du ? Il faut donc pousser l'analyse afin de pouvoir bien interpréter cette composante. En affichant les composantes pour chaque sonde (ou les coefficients de A), on voit globalement que plus on va en aval, plus C1 est importante dans la température de l'eau et moins C2 en a. C'est donc une entité qui agit principalement au début de la rivière (et ayant un effet bien moins important par la suite). On peut donc en déduire que cette composante est celle qui représente la ou les nappe(s) phréatique(s) (ou les affluents de la Touques).

Cela peut aussi être du aux températures des affluents de la Touques car on observe une baisse de C1 et une hausse légère de C2. En regardant la représentation de la Touques et de l'Orne, on observe qu'avant la sonde 830, différents affluents rejoignent la Touques dont certaines (et la Touques) passent au-dessus de nappes superficielles. On peut donc en déduire que la seconde représente des nappes phréatiques majoritairement superficielles,

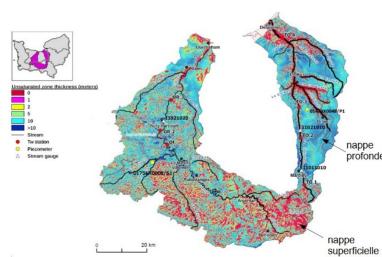
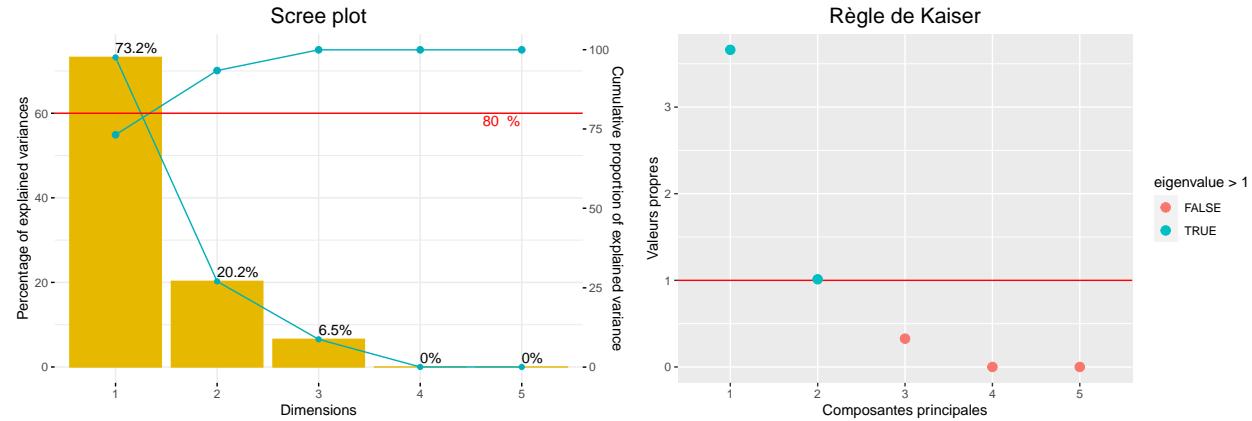


Figure 4: Nappes phréatiques à proximité de l'Orne et de la Touques

nous n'avons aucune données concernant les nappes profondes donc nous pensons qu'elles peuvent jouer un rôle, mais très minimes car comme le nom l'indique, elles sont profondes dans le sol ($> 7\text{-}8$ mètres).

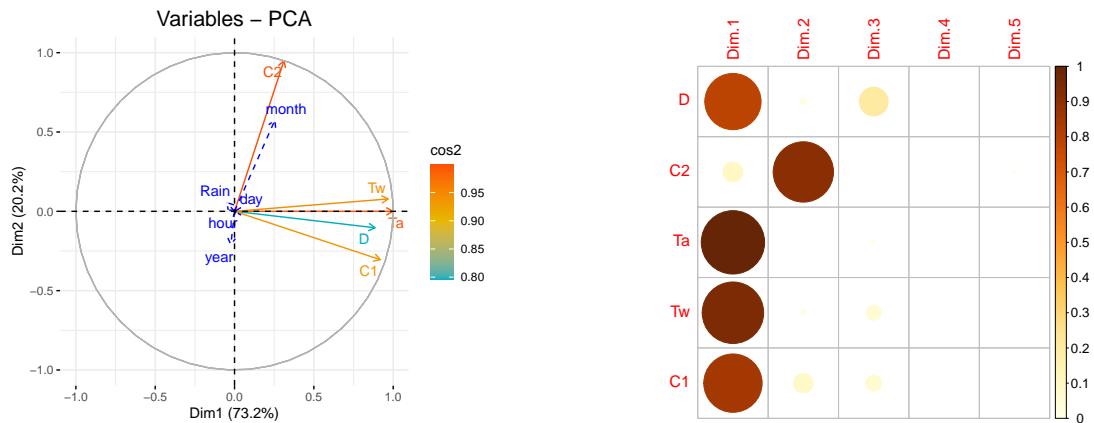
ACP

Nous pouvons désormais réaliser une analyse en composantes principales. Cette méthode d'analyse nous permet de réduire le nombre de dimensions d'un jeu de données tout en gardant un maximum d'informations sur le jeu.



Nous avons réalisé cette ACP sur cet ensemble de variables : les deux composantes indépendantes de l'ACI, les températures de l'air et de l'eau et la différence entre elles (Tair - Teau) le tout en moyenne journalières.

Grâce à la variance expliquée, en choisissant de garder les dimensions expliquant au moins 80% de la variance expliquée cumulée, on peut en garder que deux, simplifiant ainsi le jeu de données et l'interprétation des projections des variables sur le cercle des corrélations. La règle de Kaiser nous confirme aussi ce choix. Ces deux dimensions représentent 93.4% de la variance des données.



Avec le cercle des corrélations, on peut voir des variables supplémentaires (heure, jour, mois, année, pluviométrie) pour nous aider à mieux interpréter les dimensions et variables. On peut remarquer que les quatres variables que l'on souhaite analyser sont très bien représentées. Ensuite, on remarque bien que C1 et C2 sont indépendants car les deux vecteurs les représentant sont orthogonaux (car proviennent de l'ACI réalisée précédemment). De plus, on observe encore une corrélation entre la température de l'air (Ta) et de l'eau (Tw) car l'angle est très petit entre les deux (montrant par ailleurs qu'il n'y a pas de contresens dans notre analyse) et que les variables C1, Ta et D sont relativement proches donc on peut en conclure qu'il existe une corrélation entre ces variables affirmant un peu plus notre déduction sur la nature de la première composante. Puis, grâce à la variable supplémentaire "mois", il y a une grande corrélation entre celle-ci et la C1 (malgré la mauvaise représentation de "mois"). Cela peut venir appuyer l'hypothèse que C2 représente un

ensemble de nappes phréatiques car lors de l'ACI, on avait remarquer un déphasage avec C1 ce qui peut être une corrélation avec les mois.

Enfin, avec la matrice des corrélations, on observe clairement que Dim1 est très corrélée à D, Ta, Tw, et C1 et que Dim2 l'est à C2.

Conclusion

Tout d'abord, rappelons nous quelle était notre problématique : **Quels sont les différents facteurs qui pourraient influencer la température de la Touques en fonction de la géologie et du climat près des sondes ?** À travers les différentes méthodes d'analyse de données, nous avons pu observer que plus on se rapprochait de la côte, plus les températures recueillies par les sondes avaient tendance à être plus hautes. Comme on a pu l'expliquer, cela semble logique d'un point de vu hydrologique et climatique car les températures de l'eau se trouvant aux premières sondes ont tendance à avoir la température des affluents ou des nappes phréatiques sources de la Touques et plus on se rapproche des côtes de la Manche, plus la température de l'air a tendance à prendre le dessus. D'ailleurs, on peut noter que la Touques possède un régime simple et n'a donc qu'un seul mode d'alimentation qui est la nappe superficielle que l'on peut voir en rouge au sud de la carte sur la figure 6, donc l'hypothèse de la nappe phréatique est donc de plus en plus plausible.

Avec davantage de données, que ce soit celles des deux autres sondes de la Touques, ou encore celles de l'ensoleillement, de la température et vitesse du vent ou bien le taux d'humidité à proximité des sondes, cela aurait permis de mieux corroborer les hypothèses formulées. De plus, n'étant pas des géologues, climatologues ou hydrologues, toute cette analyse ne se base que sur des hypothèses qui peuvent être changeantes en fonction de la logique de celles et ceux réalisant cette même analyse.

Pour finir, il est possible d'aller plus loin dans l'analyse en réalisant :

- une analyse canonique des corrélations (méthode permettant de comparer deux groupes de variables quantitatives appliqués sur les mêmes individus afin de savoir s'ils décrivent le même phénomène)
- des analyses prédictives sur les données :
 - classification automatique : essayer de prédire l'ID d'une sonde à partir des données (apprentissage supervisé)
 - prédiction de séries temporelles : prédire l'évolution au cours du temps de la pluviométrie ou bien des températures de l'air et de l'eau (régression, etc)
 - clustering de séries temporelles : faire des regroupements de données afin de prédire l'ID des sondes de manière non supervisée (par exemple un algorithme de K-moyennes basé sur des notions de distances comme la distance euclidienne ou bien la "dynamic time warping")

Références

- Amidon, Alexandra. n.d. "A Brief Survey of Time Series Classification Algorithms." Towards Data Science.
———. n.d. "How to Apply K-Means Clustering to Time Series Data." Towards Data Science.
Burba, Davide. n.d. "An Overview of Time Series Forecasting Models." Towards Data Science.
Faicel Chamroukhi, Bruno Dardaillon et. n.d. "Cours Sur L'ACI." Ecampus - Analyse de données.