Rapport de stage - Étude statistique de la qualité de l'eau

Clément BARCAROLI

Tutrice de stage : Mme Agathe MAUPETIT – Professeur référent : M. Sébastien GADAT

Régie Eau d'Azur, 455 Promenade des Anglais, Immeuble Phoenix, 06200 Nice

2 mai 2025 – 14 août 2025

Remerciements

Je voudrais remercier tout d'abord Agathe Maupetit et Sébastien Gadat pour leur encadrement durant ce stage. Je voudrais aussi remercier Carine Gibowski et Agnalys Michaud pour les retours et conseils qu'elles m'ont apportés. Plus généralement, je me dois de remercier Pierre Roux, Félix Billaud et toute la Direction Projet Hypervision pour l'accueil bienveillant qu'ils m'ont fait. Ensuite, je me dois de remercier ma famille pour tout le soutient qu'elle m'apporte dans les moments plus ou moins faciles que je traverse. Enfin, et non des moindres, je veux remercier Camille, Cléopha, Daphné, Elliot, Lucien, Matéo et tous mes autres amis qui ont toujours été de bonne compagnie.

Table des matières

I. Introduction générale

Le stage s'est déroulé du 2 mai 2025 au 14 août 2025 au sein de la Régie Eau d'Azur (REA) à Nice. La Régie Eau d'Azur est un Etablissement Publique Commercial et Industriel créé en 2013 par la Métropole Nice Côte d'Azur pour assurer un service public de gestion de la ressource en eau. La régie vient remplacer des acteurs privés après l'expiration de leurs contrats de Délégation de Service Public. Depuis 2022, REA est donc chargée de collecter, traiter, distribuer et assainir l'eau de 51 communes du pays niçois mais aussi celle de la Principauté de Monaco.

Le conseil d'administration de REA est composé de vingt-sept membres élus de la Métropole et présidé par M. Hervé Paul, maire de Saint-Martin-du-Var et délégué à l'eau, l'assainissement et l'énergie. Le conseil d'administration a nommé M. Vincent Ponzetto au poste de Directeur Général. M. Ponzetto dirige plus de six cents employés répartis en neuf directions :

Figure 1 : Organigramme de la Régie Eau d'Azur

Le stage est réalisé au sein de la Direction Projet Hypervision (DPH) dirigée par M. Pierre Roux. Cette direction a un rôle de recherche et développement dans le domaine de la supervision de l'activité. La DPH a pour projet à long terme de développer une plateforme baptisée Hypervision, qui centralise toutes les informations logistiques et physiques nécessaires non seulement à l'activité quotidienne mais aussi à l'anticipation des difficultés futures. Pour ce faire, la DPH dispose notamment du service Modélisation Prédictive dirigé par M. Félix Billaud. Ce service produit des modèles physiques et statistiques qui permettent de comprendre le comportement de l'eau, depuis sa chute sous forme de pluie jusqu'à sa distribution à chaque point de livraison du réseau. Par exemple, l'outil AquaVar développé par le service en collaboration avec l'Université Côte d'Azur modélise les bassins versants des trois fleuves majeurs

du réseau pour différents volumes. Ainsi, le territoire peut mieux se prémunir face aux inondations et sécheresse intenses de plus en plus fréquentes du fait du réchauffement climatique et de l'urbanisation.

Le présent stage est mené sous la direction de Mme Agathe Maupetit, qui dirige le Pôle Sciences des données au sein de ce service en tant que scientifique des données. Pour l'accompagner, Mme Carine Gibowski et l'auteur du présent rapport officient eux aussi comme scientifiques des données. Le pôle se voit confier des missions par les autres directions lorsqu'elles ont besoin d'une expertise statistique analytique ou prédictive. Les missions sont traitées en collaboration avec ceux qui les déposent : le demandeur fournit ses données et ses interrogations puis le pôle tente d'y répondre de la meilleure des manières. Des points réguliers entre le Pôle Sciences des données et la personne qui a confié la mission permet d'éclairer les résultats obtenus avec des connaissances métier et d'orienter l'analyse vers les besoins réels. Le livrable prend la forme d'un rapport ou bien d'un modèle statistique construits pour répondre aux interrogations soulevées.

Ce stage est mené en collaboration avec le service des Laboratoires et Expertise Eau Potable de la Direction Technique et Innovation (Figure 1 : Organigramme de la Régie Eau d'Azur). Ce service a pour mission de surveiller de nombreuses variables physiques et bactériologiques dans l'eau du réseau. Cette mission est essentielle au bon fonctionnement de la régie : déjà, comme il s'agit d'un service public, REA doit, statutairement, s'assurer qu'elle livre l'eau de la meilleure qualité possible à tous les utilisateurs. Aussi, REA a des obligations sanitaires légales (Code de la santé publique, art. R. 1321-2, R. 1321-3, R. 1321-7 et R. 1321-38) qui forcent l'entreprise à mener des contrôles réguliers sur son réseau et à mener des actions pour non seulement traiter mais aussi prévenir les éventuels anomalies.

Le premier sujet de ce stage concerne les non-conformités de la qualité de l'eau potable. Il peut y avoir une non-conformité pour deux raisons différentes : d'une part, une mesure réalisée dans l'eau peut être anormale, au sens où elle s'écarte de la valeur de référence fixée par la loi. Dans ce cas, l'eau reste propre à la consommation mais REA doit s'assurer que la variable retourne à sa valeur de référence. D'autre part, une variable peut excéder sa valeur limite de potabilité. Dans ce cas, l'eau est déclarée impropre à la consommation : REA doit avertir la population de la situation, l'assister en distribuant des bouteilles notamment et corriger le plus vite possible le problème qui cause la non-conformité. Pour contrôler le travail de REA, l'Agence Régionale de Santé (ARS) mène en parallèle ses propres tests. Pour éviter toute triche, l'ARS ne prévient pas REA d'où et quand elle mène un contrôle. L'ARS communique cependant à REA les résultats de ses analyses, afin de l'aider dans son travail. Elle peut imposer des amendes si elle estime qu'il y a des manquements répétés au code de la santé publique. Ainsi, la surveillance des non-conformités du réseau présente deux enjeux. D'une part, REA doit s'assurer que la santé et le confort public sont respectés. D'autre part, pour être pérenne, REA ne peut pas payer chaque mois des amendes pour manquement : cela pèserait sur ses finances et nuirait à son image, ce qui pourrait pousser le conseil métropolitain à se tourner de nouveau vers des contrats de délégation de service public. Le réseau d'eau potable est long de 2 973 km sur un territoire couvert de vallées et de montagnes : il est impossible, avec les moyens actuels, de tester tout le réseau chaque jour de l'année. *A priori*, le service des Laboratoires et Expertise Eau Potable suspecte l'été d'être la période avec la plus grande fréquence de non-conformités : l'eau reste plus longtemps dans des tuyaux qui chauffent d'avantage, ce qui est propice au développement de bactéries. En partant de cette intuition, l'analyse statistique s'appuie sur les données de suivi des non-conformité de la qualité de l'eau et de données météorologiques pour analyser la fréquence des non-conformités de la qualité de l'eau relevées par l'ARS et REA, afin de déterminer si la quantité de prélèvements par mois et par secteur est adaptée aux besoins du réseau. Cette question est primordiale car si des manques sont détectés, le service des Laboratoires et Expertise Eau Potable peut adapter son suivi, augmenter son efficacité et donc la qualité de vie des habitants de la Métropole.

Le second sujet s'intéresse à la concentration en sulfate (SO_4^{--}) dans les eaux brute et potables captée sur deux sites différents. Les eaux brutes ne subissent pas de traitement pour modifier leur concentration en sulfate : il n'y a pas de différence avec les eaux potables. Le premier site est la nappe phréatique du

Var, sur le site d'exploitation Joseph Raybaud. Le second est le Canal de la Vésubie, un aqueduc qui transporte l'eau collectée depuis Saint-Jean la Rivière jusqu'au quartier de Gairaut, sur les hauteurs de Nice. Ces deux sites sont donc de nature complètement différentes : le Canal exploite de l'eau de surface, sensible aux changements extérieurs alors que le site J. Raybaud exploite de l'eau souterraine, apportée par infiltration par le fleuve du Var. Le code de la santé publique place la valeur limite de concentration en sulfate à 250 mg/L dans l'eau potable et 200 mg/L dans l'eau brute : au-delà, il y a non-conformité de la qualité de l'eau, qui n'est plus potable. Ainsi, REA instaure un seuil de sureté à 200 mg/L pour toutes les eaux : si la concentration dépasse ce seuil, des actions doivent être mises en place pour s'assurer que l'eau reste potable. Pour mesurer cette concentration, REA a deux possibilités. Soit un agent se déplace sur site et collecte un échantillon, qui doit ensuite être analysé en laboratoire ; cette solution est coûteuse en temps et en ressources humaines, en plus d'être lente (l'analyse est faite sous trois jours). Sinon, REA peut installer un outil d'analyse automatique de l'eau, qui coûte très cher et nécessite un entretien régulier pour que les relevés restent fiables. Ainsi, le service des Laboratoires et Expertise Eau Potable souhaite remplacer ces méthodes d'analyse coûteuses par une mesure indirecte. Cette mesure indirecte doit s'appuyer sur un modèle prédictif auquel il est fourni des données météorologiques et physiques qui coûtent moins cher à collecter.

II. Etude statistique des non conformités de la qualité de l'eau relevées sur le réseau Eau d'Azur

Assurer une qualité de l'eau optimale à tout moment est essentiel pour que la santé et la confiance de la population soient maximales. Prédire des non-conformités sur un réseau aussi vaste que celui de la métropole Nice Côte d'Azur serait extrêmement difficile. C'est pourquoi les contrôles sur le terrain restent la solution privilégiée. Cette méthode est soumise à deux contraintes : d'une part, le nombre d'agents est limité et d'autre part leur volume de travail annuel est fixé. Il est donc impossible de contrôler quotidiennement l'ensemble du réseau, d'où un besoin de rationaliser le processus. Pour ce faire, l'étude répond à deux questions distinctes. D'une part, quelle est la période ayant le risque de non-conformité le plus élevé ? D'autre part, quelles sont les périodes ayant une carence de contrôle ?

1. Présentation des données

Pour mener à bien cette étude, l'étude s'appuie sur trois bases de données distinctes décrites ci-dessous.

A. Tableau des non conformités

Il y a 762 observations faites entre le 30 décembre 2017 et le 31 décembre 2024 sur le territoire métropolitain. Ces données sont fournies par le service des Laboratoires et Expertise Eau Potable d'Eau d'Azur. Chacune de ces observations correspond à un dossier de non-conformité : un dossier est ouvert lorsqu'un incident de non-conformité est détecté. Les recontrôles sont inclus dans ce même dossier et ne sont donc pas comptés comme des incidents à part entière si le problème n'est pas résolu. Chaque observation dispose d'un numéro de dossier (N°), de la date à laquelle le prélèvement qui a mené à l'ouverture du dossier a été réalisé (DATE.PRELEVEMENT), de l'agence à laquelle est rattaché le lieu de l'incident de non-conformité (Agence), de sa commune (COMMUNE), de l'institution ayant réalisé le contrôle (PROGRAMME), du type de la non-conformité, c'est-à-dire la raison pour laquelle un dossier de non-conformité a été ouvert (TYPE.DE.NC(Réf/Limite)) et de sa nature, bactériologique ou physique (Classe). Les variables restantes sont des variables textuelles ou bien trop incomplètes pour

être exploitées. L'étude n'entre donc pas dans le détail de ces dernières. Ci-dessous figurent les fréquences et modalités des différentes variables qualitatives.

Tableau 1 : Fréquences des variables de la base des non-conformités

Variable	Fréquences
Agence	MHP = 0,81 NI_LI = 0,13 RD = 0,05
PROGRAMME	ARS = 0,67 AUTO = 0,33
TYPE.DE.NC.(Réf/Limite)	Réf = 0,62 Limite = 0,38
Classe	BA = 0,45 PHY = 0,54 NA = 0,01

Pour Agence, MHP veut dire Moyen et Haut pays, NI_LI Nice et littoral et RD signifie Rive droite du Var.

Figure 2 : carte du périmètre d'activité de la Régie Eau d'Azur

Sur la carte ci-dessus, l'Agence Rive Droite figure en vert, l'Agence Nice Littoral en bleu et bleu clair et l'Agence Moyen et Haut pays en orange, orange foncé et vert clair. Les communes du sud sont bordées par la mer, celle du nord sont dans un massif montagneux. La Rive Droite n'a qu'un nombre très faible de non conformités relevées, ce qui rend toute démarche statistique difficile à mettre en place.

Pour PROGRAMME, ARS signifie Agence régional de santé et AUTO autocontrôle. L'ARS mène des contrôles aléatoires sur le réseau pour s'assurer que la Régie Eau d'Azur respecte ses engagements légaux en termes de qualité d'eau. Son action est complémentaire du travail de contrôle de la Régie Eau d'Azur, qui inspecte de manière routinière l'eau mais aussi lorsqu'il y a une suspicion de non-conformité. Le matériel et les techniques de contrôles sont les mêmes, peu importe qui les réalise.

Pour TYPE.DE.NC.(Réf/Limite), Réf veut dire référence. Si la non-conformité est de type Réf, cela signifie que la valeur relevée est au-dessus de la valeur de référence fixées par la loi. Par exemple, lorsque la turbidité (transparence de l'eau) est mesurée en sortie d'usine de distribution, elle doit être inférieure ou égale à 0,5 NFU. Si elle dépasse légèrement cette norme, la Régie doit agir pour trouver la source de turbidité et retourner à la normale. Si la non-conformité est de type Limite, cela signifie que la valeur relevée dépasse la valeur limite de qualité de l'eau que la loi tolère. Toujours pour la turbidité, si elle dépasse 1,0 NFU en sortie d'usine de distribution, l'eau est considérée comme impropre à la consommation.

Pour Classe, BA signifie Bactériologique et PHY veut dire physique. Une non-conformité liée à la présence d'E. Coli est de Classe BA. Une non-conformité liée à la détection de plomb est de type PHY. Les valeurs NA correspondent à des évènements exceptionnels et inclassables.

La variable saison est construite à partir de la colonne DATE.PRELEVEMENT. L'hiver inclut tous les mois d'octobre à janvier, le printemps tous ceux compris entre février et mai. L'été comprend les mois restants. Ces saisons sont les saisons météorologiques propres à Nice et son pays.

B. Tableau des prélèvements par Point de Surveillance (PSV)

Cette base contient le nombre de prélèvements réalisés sur le réseau par mois et par agence depuis janvier 2018 jusqu'à décembre 2024. La colonne Agence donne l'agence (MHP, NI_LI ou RD comme ci-dessus). « mois » et « annee » donnent respectivement le mois et l'année de l'observation.

Tableau 2: Statistiques univariées de la variable "Nombre de prélèvements par mois"

	Min	1er quantile	Médiane	Moyenne	3ème quartile	Max
nbr_prlvmt	39	78	122	131,3	168	385

A noter que les contrôles de l'ARS et les autocontrôles ne sont pas distingués.

C. Base de données météorologiques

L'analyse s'appuie sur le cumul de pluie quotidien depuis le 1er janvier 2017 jusqu'au 31 décembre 2024 pour 14 communes de l'Est des Alpes Maritimes (Figure 30). Dans l'ordre alphabétique, il y a Ascros, Carros, Coursegoules, St-Etienne-de-Tinée, Lantosque, Levens, Lucéram, Le Mas, St-Martin-d'Entraunes, St-Martin-Vésubie, Nice, Péone, Puget et Rimplas. Ces données proviennent de Météo France.

Tableau 3 : Statistiques univariées pour les différents cumuls quotidiens de pluie (partie 1)

Variable Encodage	Ascros Ascros	Carros Carros	Coursegoules Coursegoules	St-Etienne-de-Tinée St_Et	Lantosque Lantosque	Levens Levens	Luceram Luceram
Min	0	0	0	0	0	0	0
1er quartile	0	0	0	0	0	0	0
Médiane	0	0	0	0	0	0	0
Moyenne	2,128	2,092	3,213	2,562	2,376	2,481	2,619
3ème quartile	0,2	0	0,2	1	0,2	0,2	0,2
Max	146,5	119	331	109,5	235,6	162,3	205,7

Tableau 4 : Statistiques univariées pour les différents cumuls quotidiens de pluie (partie 2)

Variable Encodage	Le Mas Le_mas	St-Martin-d'Entraunes Saint_martin_dentraunes	St-Martin-Vésubie St_martin	Nice Nice	Péone Peone	Puget Puget	Rimplas Rimplas
Min	0	0	0	0	0	0	0
1er quartile	0	0	0	0	0	0	0
Médiane	0	0	0	0	0	0	0
Moyenne	2,749	2,89	3,018	1,942	2,428	2,25	2,251
3ème quartile	0,4	1,44	0,12	0	0,4	0,2	0,4
Max	314,4	97,1	209,8	162,8	109,8	172,5	188,4

L'abondance de 0 et les valeurs maximales extrêmement loin de la moyenne et de la médiane sont caractéristiques du département. Le climat est d'ordinaire sec mais des épisodes méditerranéens rares et intenses peuvent se produire. La valeur maximum observée en tout point correspond au passage de la tempête Alex au début d'octobre 2020, qui a ravagé une bonne partie des vallées.

2. Méthodologie

Toutes les p valeurs citées sont le résultat d'un test de Student.

A. Isolement de l'effet propre des variables explicatives sur la probabilité d'occurrence des non-conformités de qualité de l'eau

La méthode de matching est utilisée pour isoler l'effet moyen du traitement (ici être durant la saison pour laquelle est suspectée une sur-fréquence de non-conformités) sur le traité (ici un jour donné). Au final, le modèle donne une estimation de la différence de probabilité moyenne d'avoir une non-conformité pour un jour donné, suivant s'il se déroule durant la saison d'intérêt ou pas. Ici, l'algorithme de matching utilisé est l'algorithme de k plus proche voisins, avec k = 1. La pertinence des variables de contrôles est testée par la statistique de synthèse fournie par la fonction MatchBalance du package Matching[1]. Deux modèles sont évalués pour chaque agence. Le premier utilise pour variables de contrôles le cumul de pluie du jour pour les 14 stations météos disponibles. Le second utilise les cumuls de pluie de la veille.

Un modèle logit est aussi évalué pour essayer de mesurer l'effet propre des cumuls de pluie du jour sur la probabilité qu'une non-conformité se produise. Ce modèle utilise comme variables explicatives les 14 cumuls de pluie quotidiens à disposition et la saison.

B. Relation entre la fréquence et le nombre de prélèvements

La fréquence de NC relevées est calculée ainsi : $\frac{Nombre\ de\ NC\ relevées}{Nombre\ de\ prélèvements}$. Le nombre de NC relevées est une fonction croissante du nombre de prélèvements (les non-conformités ne peuvent être trouvées que si elles sont recherchées). Ainsi, mathématiquement, impossible de savoir au préalable comment se comporte la fréquence en fonction du nombre de prélèvements. Cette relation donne des informations sur le risque de non-conformité :

- Si la fréquence augmente avec le nombre de prélèvements, cela signifie que le nombre de NC relevées augmente plus vite que le nombre de prélèvements. Ceci est interprété comme un gain d'efficacité lié à une bonne connaissance du réseau : les campagnes de prélèvements sont adaptées au risque de non-conformité propre à un jour donné.
- Si la fréquence diminue avec le nombre de prélèvements, l'interprétation est inversée : le nombre de prélèvements augmente plus vite que le nombre de NC relevées, il y a une perte d'efficacité de la surveillance. Cela voudrait dire qu'il y a une mauvaise connaissance du réseau : ce qui est su n'est pas seulement faux mais l'inverse de ce qui est réel.
- Si la fréquence est constante quel que soit le nombre de prélèvements, cela signifie que le nombre de NC relevées et le nombre de prélèvements sont colinéaires. Dans ce cas, il n'y a aucune connaissance du réseau à avoir : le risque de NC est le même toute l'année.
- Enfin, s'il n'y a pas de corrélation apparente, cela peut être dû au fait que le nombre de NC détectées par mois varie beaucoup moins et prend des valeurs beaucoup plus faibles que le nombre de prélèvements. Les variations du nombre de NC détectées sont négligeables face à celles du nombre de prélèvement. Par exemple, si le maximum historique de non-conformités pour un mois est de 3 et le nombre de prélèvements n'est jamais inférieur à 95 comme c'est le cas pour l'agence Nice Littoral : la variable fréquence de NC détectées va être très proche de 0 et avec une variance très faible. Ce cas de figure correspond à une situation où le nombre de prélèvements est disproportionné face au risque de non-conformité.

C. Comparaison des contributions de l'ARS et REA aux détections

La variable « nombre de prélèvements » confond les prélèvements réalisés par l'ARS et par REA. Il n'est ainsi pas possible de connaître le nombre de prélèvements effectués par chacun. En revanche, la fréquence $\frac{Nombre\ de\ NC\ relevées\ par\ l'institution}{Nombre\ de\ prélèvements\ total}$ en informe indirectement : c'est la contribution de

[1] Sekhon JS (2011). "Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R." *Journal of Statistical Software*, **42**(7), 1–52.

chaque institution à la détection des NCs. Comme elles ont toutes deux les mêmes méthodes de prélèvement et d'analyse, l'hypothèse est faite qu'elles ont le même taux de détection de NC = $\frac{Nombre\ de\ NC\ relevées\ par\ l'institution}{Nombre\ de\ prélèvements\ réalisés\ par\ l'institution}$ pour une période et une agence donnée. L'idée pour retrouver le nombre de prélèvements de chaque institution est de comparer les contributions définies ci-dessus. Si le nombre de prélèvements évoluait de la même manière pour les deux institutions, leurs contributions devraient évoluer elles aussi de la même manière car elles ont le même taux de détection de NC par hypothèse. Si ce n'est pas le cas, par l'hypothèse de l'égalité des taux de détection, le nombre de prélèvements des deux institutions évolue différemment. Ce test est mené visuellement en comparant les contributions de chaque institution.

3. Réponse statistique

A. Analyse de la fréquence des non-conformités

L'hypothèse à tester est qu'il y a plus de non-conformités en été que le reste de l'année. Les températures élevées de l'été profitent aux colonies de bactéries présentent dans l'eau. Dans un premier temps, le regard se porte sur la fréquence moyenne de non-conformité sur l'ensemble du réseau.

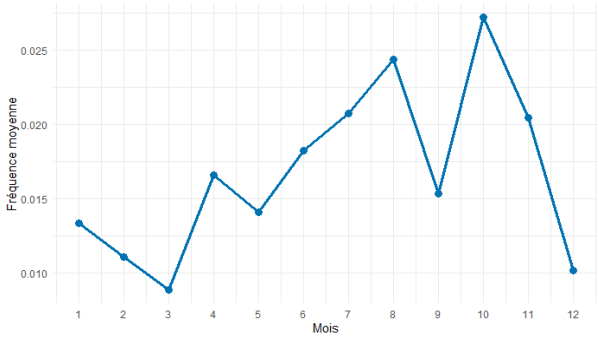


Figure 3 : évolution de la fréquence moyenne de non-conformités détectées en fonction du mois

La fréquence moyenne est particulièrement élevée en juin, juillet et août (respectivement $1{,}8\times10^{-2}$; $2{,}00\times10^{-2}$ et $2{,}4\times10^{-2}$ contre une moyenne de $1{,}6\times10^{-2}$). Cependant, la fréquence la plus élevée est en octobre ($2{,}7\times10^{-2}$) donc en automne et le mois de septembre rompt la tendance de la courbe (Figure 3). Ces premiers résultats vont dans le sens de l'hypothèse de départ.

Voici ce qu'il en est pour les trois agences prises séparément.

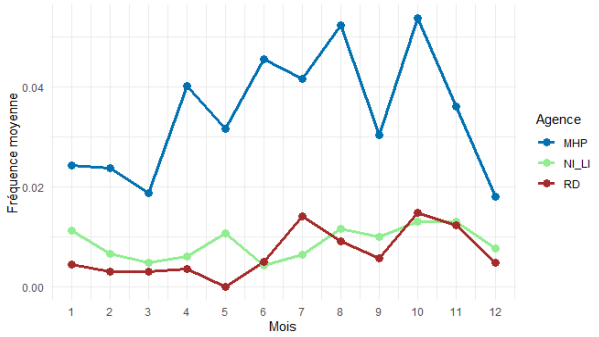


Figure 4 : évolution de la fréquence moyenne de non-conformités détectées en fonction du mois, par agence

Les tendances sont différentes selon l'agence considérée. Déjà, la fréquence dans le Moyen et Haut pays est en moyenne supérieure à celle des autres agences à tout moment de l'année (au minimum, elle vaut $1{,}8\times10^{-2}$ contre $1{,}4\times10^{-2}$ au maximum pour RD). La Rive droite et le Moyen et Haut pays ont les mêmes périodes de crise en août et en octobre. L'agence Nice Littoral présente un profil particulier : les mois

de janvier ($1{,}1\times10^{-2}$) et de mai ($1{,}0\times10^{-2}$) ont des fréquences comparables à août ($1{,}1\times10^{-2}$) et supérieur à septembre ($9{,}9\times10^{-3}$). La fréquence moyenne maximale pour Nice Littoral est atteinte en novembre ($1{,}3\times10^{-2}$) (Figure 4). Ces différences motivent des analyses distinctes agence par agence. Il est impossible de construire un modèle logit qui prend en compte séparément les effets causaux des précipitations, pour l'agence Moyen et Haut pays comme pour les deux autres agences (voir paragraphe A). En effet, les variables à dispositions sont trop fortement corrélées (la corrélation minimale est de 0,48 entre les cumuls quotidiens de Saint-Etienne-de-Tinée et Nice) et pas assez explicatives. Toutes les différentes combinaisons de variables explicatives essayées aboutissent à une non-convergence de l'algorithme.

i. Moyen et Haut pays

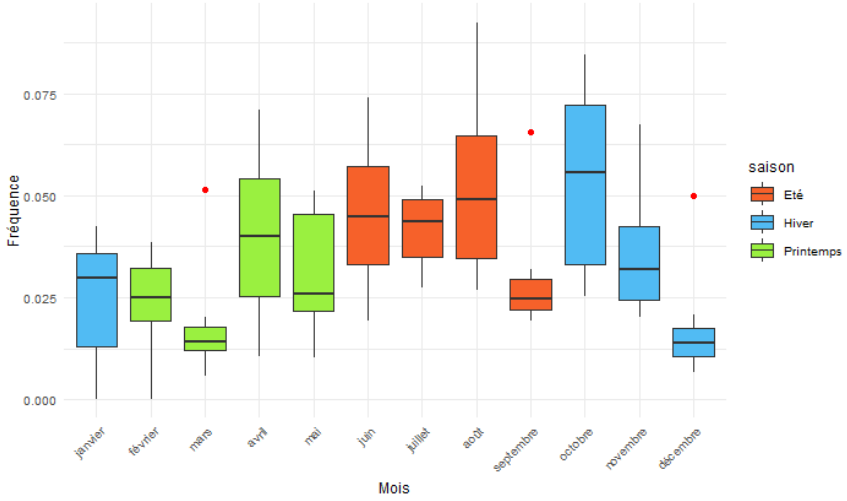


Figure 5 : fréquence de non-conformités détectées par mois pour l'agence Moyen et Haut pays

Pour le Moyen et Haut pays, l'été est la saison avec la fréquence moyenne la plus élevée ($2{,}8\times10^{-2}$ au printemps, $4{,}2\times10^{-2}$ en été et $3{,}3\times10^{-2}$ en hiver) mais c'est octobre qui est le mois avec la plus grande fréquence moyenne ($5{,}3\times10^{-2}$). Mars ($1{,}8\times10^{-2}$) et décembre ($1{,}8\times10^{-2}$) ont des fréquences moyennes particulièrement faibles qui détonnent avec les autres mois : la médiane de la fréquence moyenne pour tous les mois confondus est de $3{,}1\times10^{-2}$. La dispersion faible de la fréquence pour les mois de mars ($2{,}2\times10^{-4}$), septembre ($2{,}5\times10^{-4}$) et décembre ($2{,}1\times10^{-4}$) n'est pas reflétée par la variance empirique ($4{,}1\times10^{-4}$ de variance sur l'ensemble des données) (Figure 5).

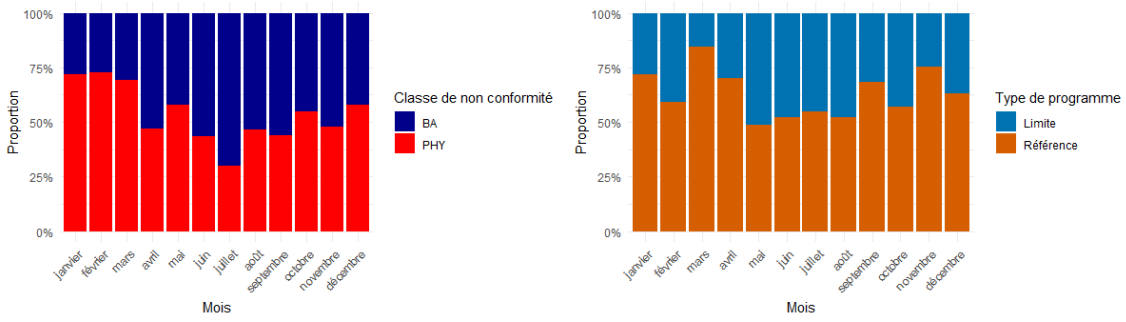


Figure 6 : proportion de chaque classe de non-conformité détectée par mois pour l'agence Moyen et Haut pays

Figure 7 : proportion de chaque type de non-conformité détectée par mois pour l'agence Moyen et Haut pays

Considérer les fréquences par type et par classe permet d'expliquer en partie l'origine des variations de la fréquence moyenne de non-conformités de la qualité de l'eau. L'été est marqué par une grosse

augmentation de la part de non-conformités bactériologiques. En juin, juillet et août, elles deviennent majoritaires alors qu'elles sont largement minoritaires le reste de l'année. Par exemple, en janvier, la fréquence moyenne de NC bactériologiques est de 37% contre 70% en moyenne en juillet (Figure 6). Les non-conformités sont aussi davantage des dépassements de limite en été (41% des cas en moyenne) que le reste de l'année (en moyenne 30% au printemps et 28% en hiver) (Figure 7). Ces résultats vont dans le sens de l'hypothèse selon laquelle l'augmentation de la fréquence viendrait de l'augmentation des températures.

L'analyse se concentre sur l'effet propre de l'été. Premièrement, pour des conditions pluviométriques données (pour connaître la sélection des stations météos, voir annexe), le fait d'être en été augmente la probabilité de relever une non-conformité de 0,13 point : Pour un jour d'hiver avec des conditions pluviométriques données, s'il a un risque de 0,5 de connaître une non-conformité, un jour similaire du point de vue des précipitations a 0,5+0,13 = 0,63 de risque d'avoir une non-conformité. Cette hausse peut être liée à la hausse de températures en été ou à une baisse du temps alloué à chaque prélèvement par manque de main d'œuvre durant l'été.

ii. Rive droite

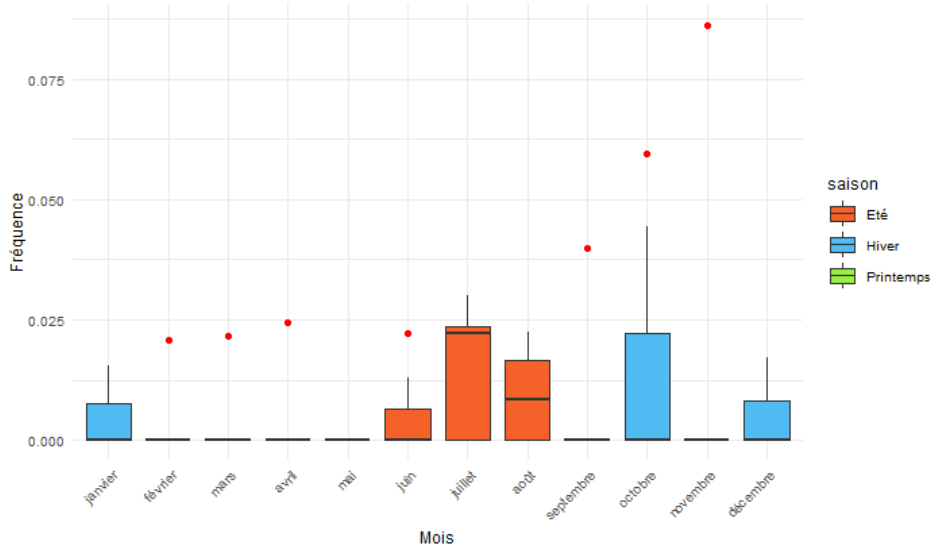
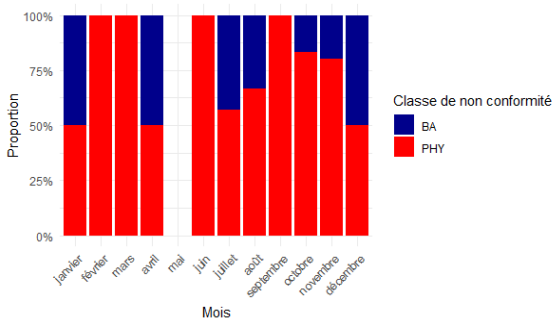


Figure 8 : fréquence de non-conformités détectées par mois pour l'agence Rive Droite

Comme pour MHP, pour la Rive Droite, juillet ($2{,}2\times10^{-2}$ de médiane) et août ($8{,}5\times10^{-3}$ de médiane) sortent du lot avec des médianes de fréquence de NC non nulles. Cependant, si la médiane d'octobre est nulle, sa fréquence moyenne est élevée ($1{,}4\times10^{-2}$ contre $6{,}6\times10^{-3}$ pour la moyenne globale) (Figure 8).



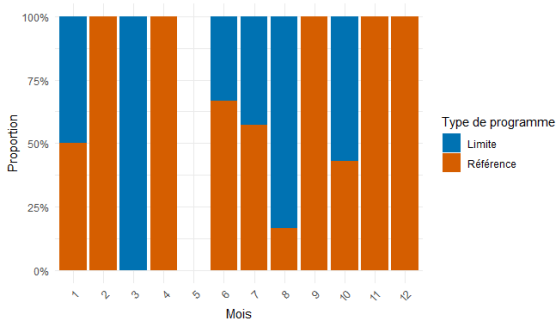


Figure 9 : proportion de chaque classe de non-conformité détectée par mois pour l'agence Rive Droite

Figure 10 : proportion de chaque type de non-conformité détectée par mois pour l'agence Rive Droite

Sur la Rive Droite, aucune non-conformité n'a jamais été détectée en mai (Figure 8). Ce problème est traité en partie ii. Statistiquement parlant, il n'y a aucun lien significatif entre la répartition des classes (p valeur $=5{,}5\times10^{-1}$) et des types (p valeur $=8{,}9\times10^{-1}$) et le mois (Figure 9 et Figure 10). La rareté des non-conformités empêche de tirer des conclusions sur des liens entre les variables. L'effet propre de l'été n'est pas significatif. Comme pour tout ce qui précède concernant la Rive droite, ces résultats sont avant tout le reflet d'un manque de données de NC pour la Rive droite.

iii. Nice Littoral

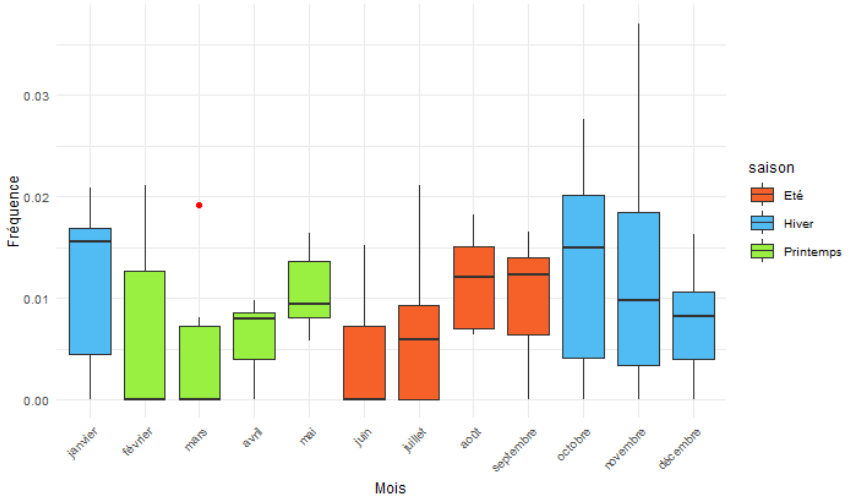


Figure 11 : fréquence de non-conformités détectées par mois pour l'agence Nice Littoral

Sur l'agence Nice-Littoral, l'été n'a pas une moyenne de fréquence significativement différente du reste de l'année (p valeur = $5{,}2\times10^{-1}$). En revanche, l'hiver a une fréquence moyenne significativement supérieure au reste de l'année (p valeur = $3{,}9\times10^{-2}$, en moyenne $1{,}1\times10^{-2}$ contre $7{,}5\times10^{-3}$) et le printemps a une fréquence significativement inférieure à celle du reste de l'année (p = $6{,}3\times10^{-2}$, en moyenne $7{,}0\times10^{-3}$ contre $9{,}6\times10^{-3}$), avec des mois d'octobre ($1{,}2\times10^{-2}$), de janvier ($1{,}1\times10^{-2}$) et surtout de novembre ($1{,}2\times10^{-2}$) particulièrement hauts. Le mois de mai se détache du reste du printemps à cause d'une fréquence moyenne de NC comparable aux mois forts d'été ($1{,}0\times10^{-2}$ en moyenne en mai). Cette fréquence moyenne n'est pas inférieure à celle d'août ($1{,}1\times10^{-2}$ en août avec p = $3{,}6\times10^{-1}$). Elle n'est pas non plus supérieure à celle de septembre ($9{,}9\times10^{-3}$ en septembre avec p = $3{,}9\times10^{-1}$) (Figure 11).

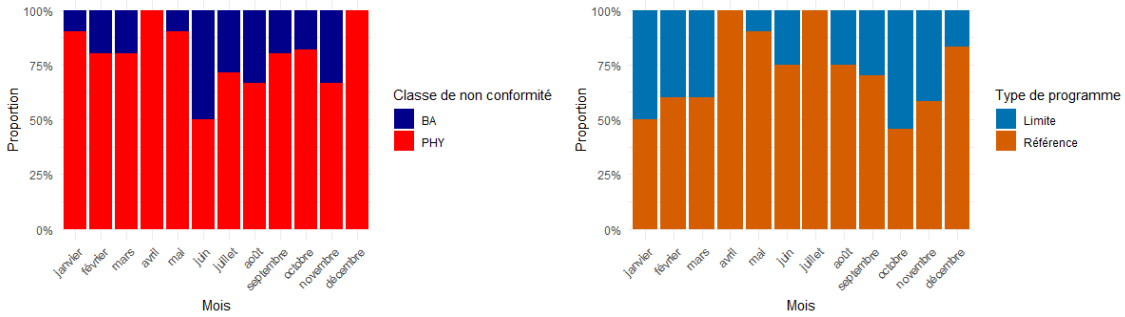


Figure 12 : proportion de chaque classe de non-conformité détectée par mois pour l'agence Nice Littoral

Figure 13 : proportion de chaque type de non-conformité détectée par mois pour l'agence Nice Littoral

Comme pour MHP (voir 0), l'été sur le territoire couvert par l'agence Nice-Littoral connaît une augmentation significative au niveau 10% de la part de non-conformités bactériologiques (35% en moyenne l'été contre 18% le reste de l'année, $p = 6{,}8\times10^{-2}$). Cette augmentation est cependant d'une ampleur moindre : les cas bactériologiques restent en moyenne minoritaires toute l'année. Au maximum, la parité est atteinte en juin (Figure 12). Pour le type de non-conformité, l'été ne se démarque pas du lot non plus : d'avril à juillet, les non-conformités de type limite sont quasiment absentes. Au maximum, elles représentent 25% des NC des mois de juin agrégés. L'hiver est la saison avec la proportion la plus élevée de non-conformités de type limite (60% en moyenne l'hiver contre 18% au printemps et 25% en été) (Figure 13).

L'effet propre de l'hiver est statistiquement non significatif avec pour variables de contrôle les précipitations (p valeur $= 2{,}4\times10^{-1}$).

iv. Synthèse

La fréquence des non-conformités détectées évolue de manière différente selon l'agence observée. Ainsi, pour le Moyen et Haut pays comme pour la Rive droite, l'été présente effectivement des fréquences plus élevées que le reste de l'année. Cette augmentation de la fréquence est corrélée, pour le Moyen et Haut pays, à une hausse des non-conformités bactériologiques. Mais les mois d'octobre et novembre sont comparables à l'été. Pour Nice Littoral, c'est l'hiver qui a la plus haute fréquence de non-conformités. Chaque agence a des mois qui constituent des anomalies. Pour MHP et RD, le mois de septembre connaît un creux de fréquence, alors qu'août et octobre sont les mois avec les plus hautes fréquences. Pour MHP uniquement, le mois de décembre voit une chute importante de la fréquence de non-conformité, qui apparaît anormal au comparaison de novembre et janvier. Pour Nice Littoral, le mois de mai ressort particulièrement du lot.

L'hypothèse selon laquelle l'été produit plus de non-conformité n'est vraie que pour l'agence Moyen et Haut pays. Les agences Rive Droite et Nice Littoral ont davantage de non-conformité en hiver que le reste de l'année. Les températures et précipitations ne suffisent pas pour expliquer le fait d'avoir une non-conformité ou pas.

B. Recherche des lacunes du système de surveillance des non-conformités

L'analyse des non-conformités faite ci-dessus est grandement influencée par le volume d'analyses réalisées. En effet, impossible de trouver des non-conformités sans en chercher.

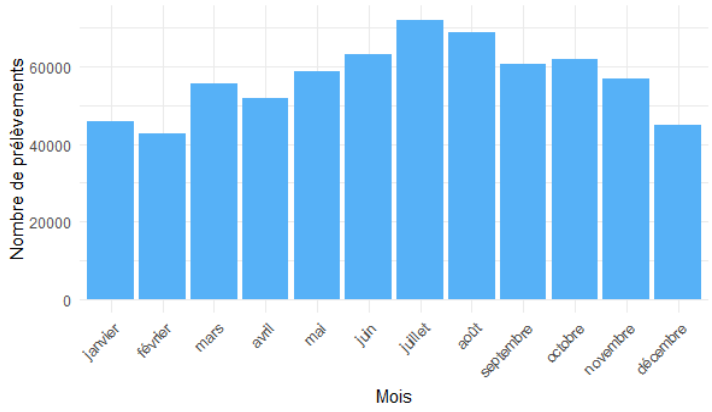


Figure 14 : somme du nombre de prélèvements réalisés par mois entre 2018 et 2024, toutes années confondues

Il y a une saisonnalité claire du nombre de prélèvements : le nombre de prélèvements est croissant de janvier jusqu'à juillet puis décroit jusqu'en décembre. Par exemple, le nombre de prélèvements par mois passe presque du simple au double entre février et juillet (Figure 14).

Lorsque la corrélation entre le nombre de prélèvements et la fréquence de non-conformités détectées est testée, elle est de 58% entre les deux variables avec un niveau de confiance à 99%. Cette corrélation est forte (>50%).

Les moments avec le plus de prélèvements sont ceux qui ont la plus grande fréquence de non-conformité (Figure 15).

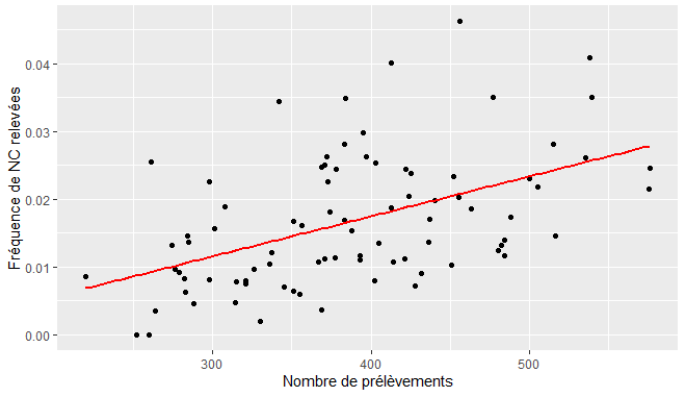


Figure 15 : fréquence de non-conformités détectées en fonction du nombre de prélèvements

Depuis le départ, cette analyse s'intéresse à tous les contrôles réalisés, sans tenir compte de qui les réalise. Or, dans la perspective où REA voudrait remédier aux problèmes relevés jusque-là, il faudrait se pencher sur la fréquence de non-conformités relevées en fonction de si ce sont des autocontrôles (REA) ou bien des contrôles de l'ARS. Ainsi, REA pourrait connaître sa part de responsabilité et agir en conséquence.

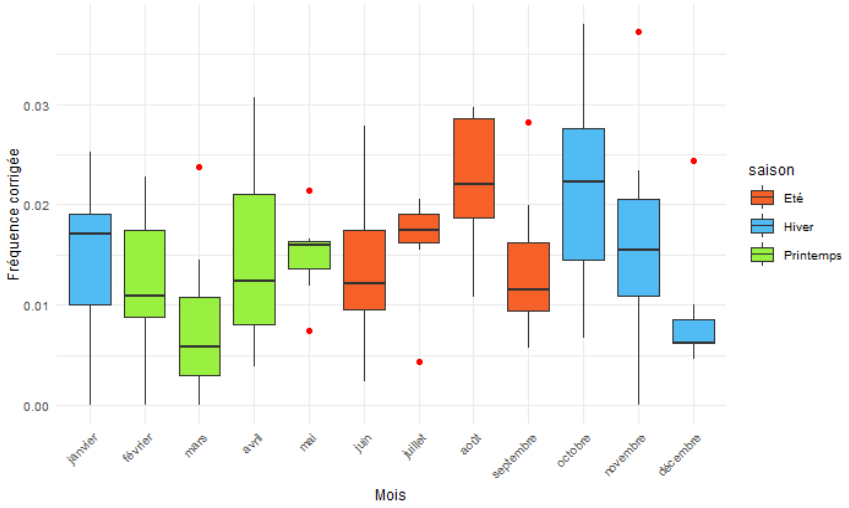


Figure 16 : fréquence de non-conformités détectées par mois relevées par l'ARS, toutes agences confondues

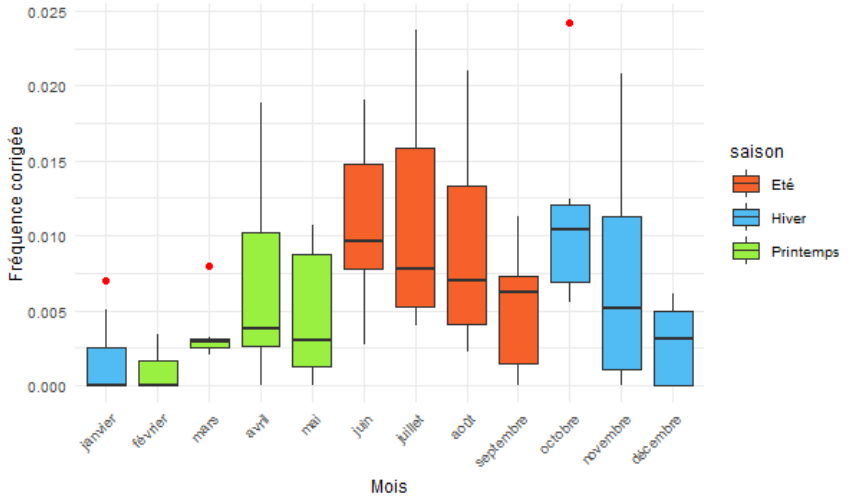


Figure 17: fréquence de non-conformités détectées par mois relevées par REA, toutes agences confondues

Les deux institutions n'ont pas la même évolution de leur contribution à la détection. Les trois premiers mois de l'année sont intéressants : ils ont les trois fréquences les plus basses pour les autocontrôles alors qu'ils se trouvent dans la moyenne pour l'ARS. Il en est déduit que la Régie Eau d'Azur réalise peu de prélèvements durant cette période. (Figure 16 vs Figure 17)

L'analyse se décline pour chaque agence.

i. Moyen et Haut pays

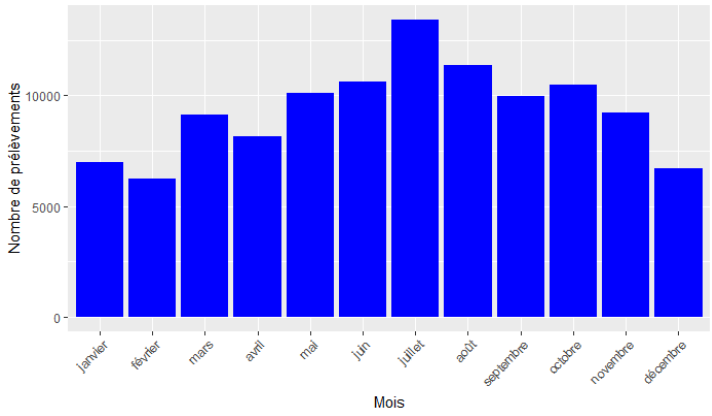


Figure 18 : somme du nombre de prélèvements réalisés par mois pour l'agence Moyen et Haut pays, toute année confondue

Le MHP a une saisonnalité de son nombre de relevés (Figure 18) : il passe du simple au double entre l'hiver et l'été. Le mois de septembre a un nombre de prélèvements proche de celui de ses voisins : la baisse soudaine de la fréquence moyenne en septembre (Figure 3) ne vient pas de là.

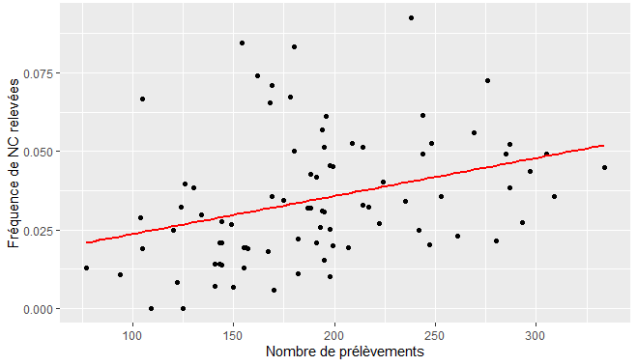


Figure 19 : fréquence de non-conformités détectées en fonction du nombre de prélèvements pour l'agence Moyen et Haut pays

La fréquence de NC relevées est une fonction croissante du nombre de prélèvements (Figure 19). La corrélation est de $3{,}3\times10^{-1}$ avec un niveau de confiance de 99%. La même interprétation est faite que pour les résultats obtenus toutes agences confondues (Figure 15) : le ciblage des prélèvements est bon. Le regard se tourne à présent vers ce qui se passe du point de vue des contributions de l'ARS et d'Eau d'Azur au nombre de prélèvements.

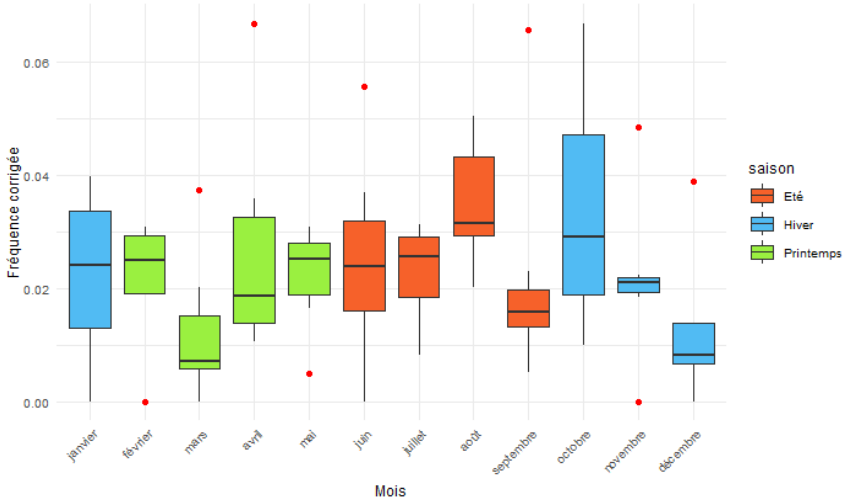


Figure 20 : fréquence de non-conformités détectées par mois relevées par l'ARS pour l'agence Moyen et Haut pays

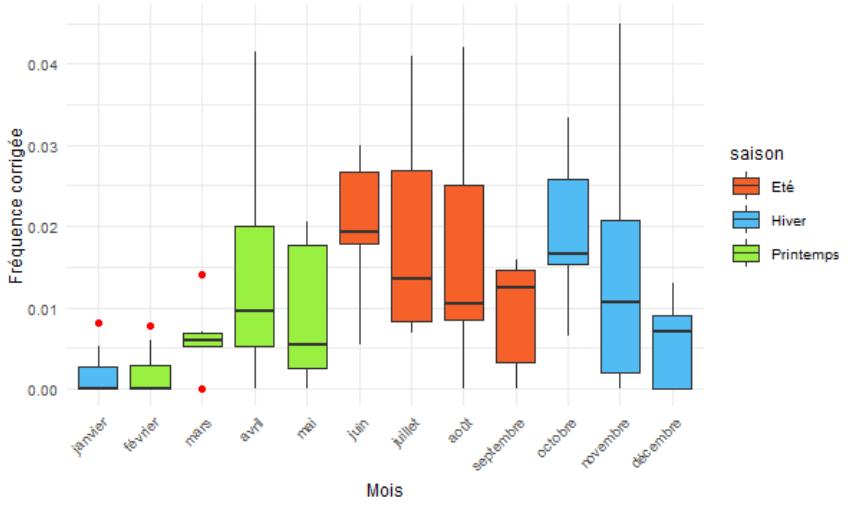


Figure 21: fréquence de non-conformités détectées par mois relevées par REA pour l'agence Moyen et Haut pays

Chronologiquement, il est à noter que les trois premiers mois de l'année sont particulièrement bas pour les autocontrôles (Figure 21), ce qui n'est pas le cas pour l'ARS (Figure 20). Cela peut être dû à une baisse du nombre d'autocontrôles. La fréquence entre le printemps et l'été est constante pour l'ARS mais croissante pour Eau d'Azur. A cela s'ajoute la soudaine augmentation en août pour l'ARS, absente pour Eau d'Azur. Cette fois, il est possible de penser que les autocontrôles sont renforcés sur l'été et que la bonne connaissance du réseau permet à Eau d'Azur d'avoir une fréquence constante sur l'été. Le mois de septembre présente une rupture de continuité pour les deux entités : elle ne peut pas être expliquée par une baisse relative du nombre de contrôles d'Eau d'Azur ou de l'ARS. Dernièrement, il faut remarquer que les variances pour le mois de novembre sont complètement différentes. L'ARS a des résultats très constants et moyens, contrairement aux autocontrôles : le nombre de tests réalisés en novembre par Eau d'Azur est probablement très irrégulier (Figure 20 et Figure 21).

ii. Rive droite

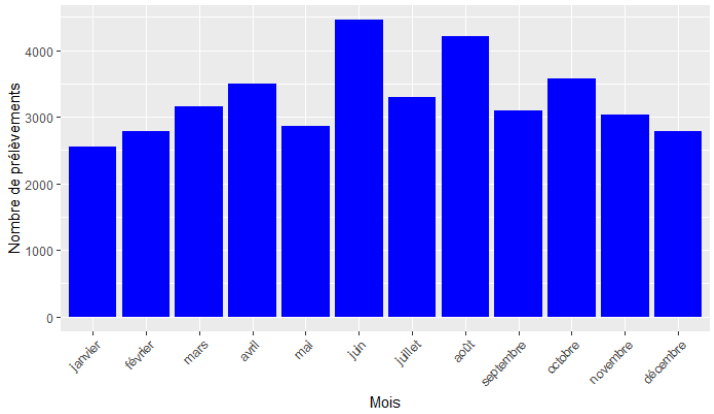


Figure 22 : somme du nombre de prélèvements réalisés par mois pour l'agence Rive Droite, toute année confondue

Pour la Rive droite, le nombre de prélèvements présente une symétrie autour du mois de juillet. Les mois qui voient le plus de contrôle sont le mois de juin et d'août. Mai, juin et septembre rompent une évolution continue du nombre de prélèvements (Figure 22).

Pour mesurer l'efficacité de la stratégie de prélèvements, l'analyse se tourne vers la relation entre la fréquence de NC relevées et la nombre de prélèvements effectués.

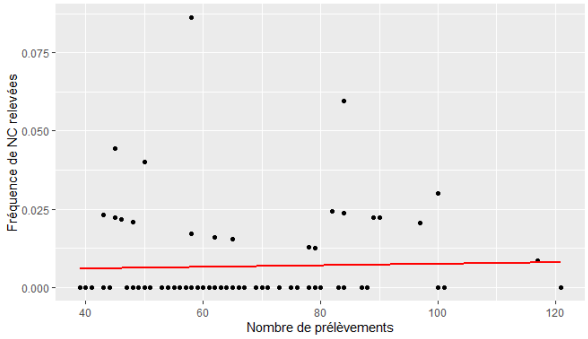


Figure 23 : fréquence de non-conformités détectées en fonction du nombre de prélèvements pour l'agence Rive Droite

La plupart des mois, aucune NC n'est détectée (Figure 23). Il n'existe aucune corrélation entre les deux variables, graphiquement comme statistiquement. Cette non-corrélation indique que le nombre de prélèvements est choisi pour respecter un certain niveau de sécurité.

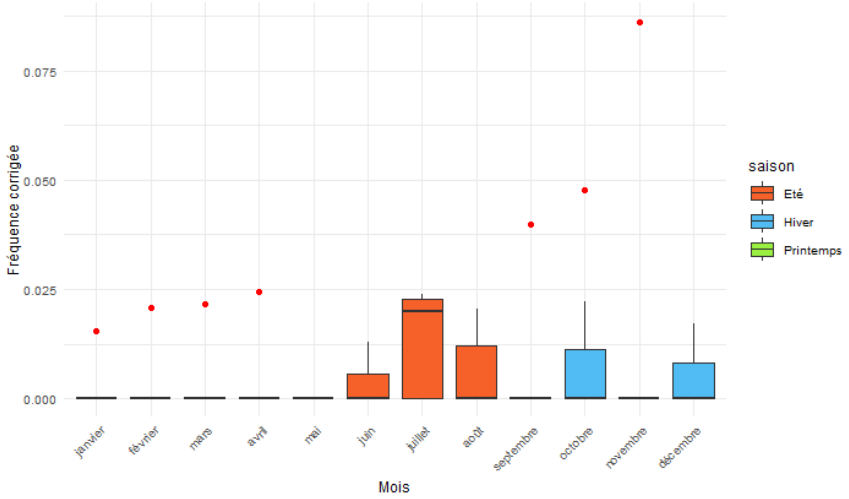


Figure 24 : fréquence de non-conformités détectées par mois relevées par l'ARS pour l'agence Rive Droite

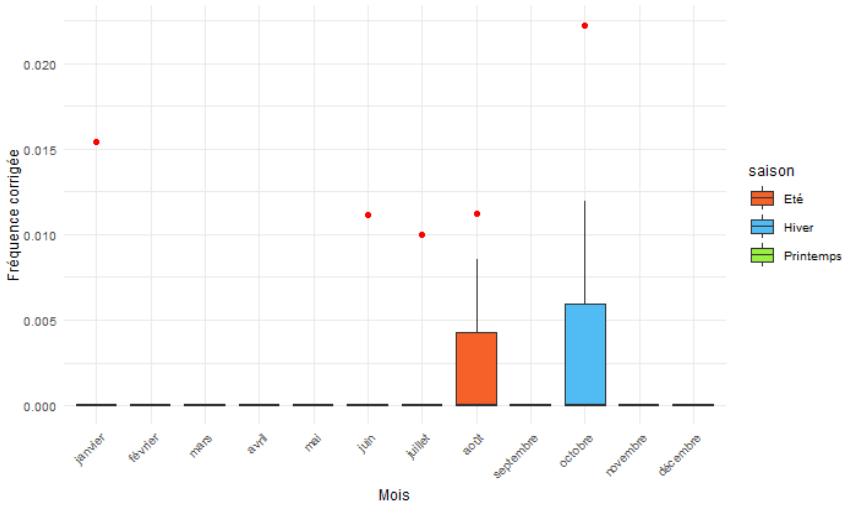


Figure 25: fréquence de non-conformités détectées par mois relevées par REA pour l'agence Rive Droite

Les NCs détectées sur le Rive Droite sont rares (Figure 24 et Figure 25). Cependant, le mois de juillet présente une différence forte selon l'organisme de contrôle (moyennes différentes au niveau 10%). Alors que c'est le mois avec la médiane la plus élevée pour l'ARS (Figure 24), juillet n'a qu'une seule non-conformité détectée par autocontrôle (Figure 25), en juillet 2024. Il est possible qu'Eau d'Azur ne réalise pas assez de contrôles durant ce mois. De même, dans une moindre mesure, pour les mois de juin et décembre. Les autres mois ont un profil similaire pour les deux organismes.

iii. Nice Littoral

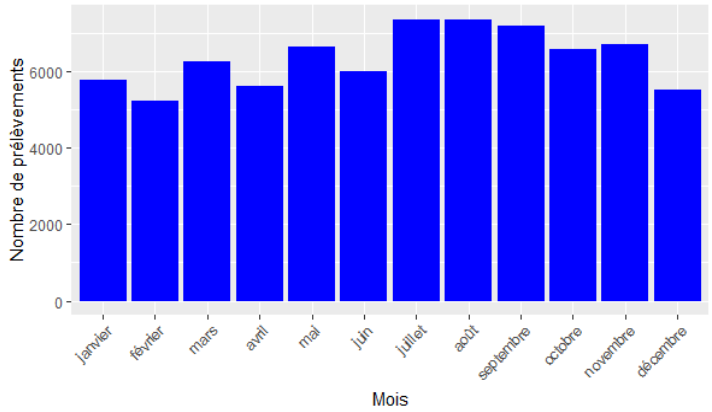


Figure 26 : somme du nombre de prélèvements réalisés par mois pour l'agence Nice Littoral, toute année confondue

Le nombre de prélèvements est stable sur l'année par-rapport aux autres agences (Figure 26). Au maximum, février a 25% de prélèvements en moins que juillet. Les mois de juillet, août et septembre ont le plus grand nombre de prélèvements (respectivement 7357, 7329 et 7168).

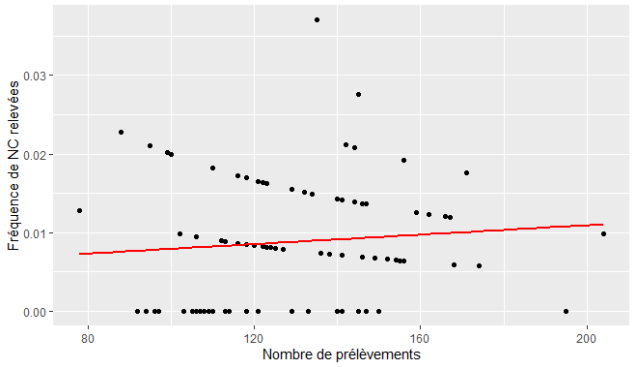


Figure 27 : fréquence de non-conformités détectées en fonction du nombre de prélèvements pour l'agence Nice Littoral

Mathématiquement, il n'y a aucune corrélation entre la fréquence et le nombre de prélèvements (p valeur = $8{,}0\times10^{-1}$ pour le printemps, p = $1{,}9\times10^{-1}$ pour l'été et p = $1{,}0\times10^{-1}$ en hiver). Un artéfact visuel apparaît à cause de la faible variance du nombre de non-conformités (maximum 5 non-conformités pour un mois) : il n'y a pas de relation décroissante entre le nombre de prélèvements et la fréquence de NC relevées (Figure 27). L'analyse est la même que pour la Rive droite : les prélèvements sont surabondants par rapport au risque de non-conformités, afin d'assurer la sécurité du réseau.

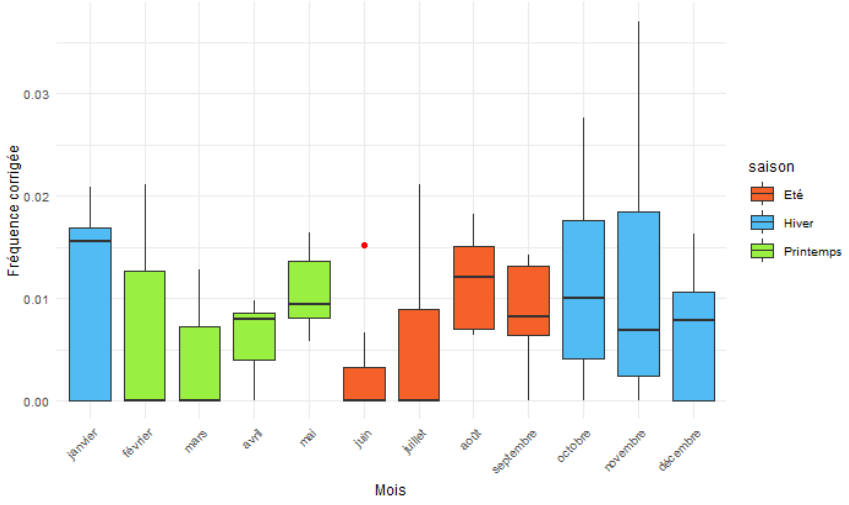


Figure 28 : fréquence de non-conformités détectées par mois relevées par l'ARS pour l'agence Nice Littoral

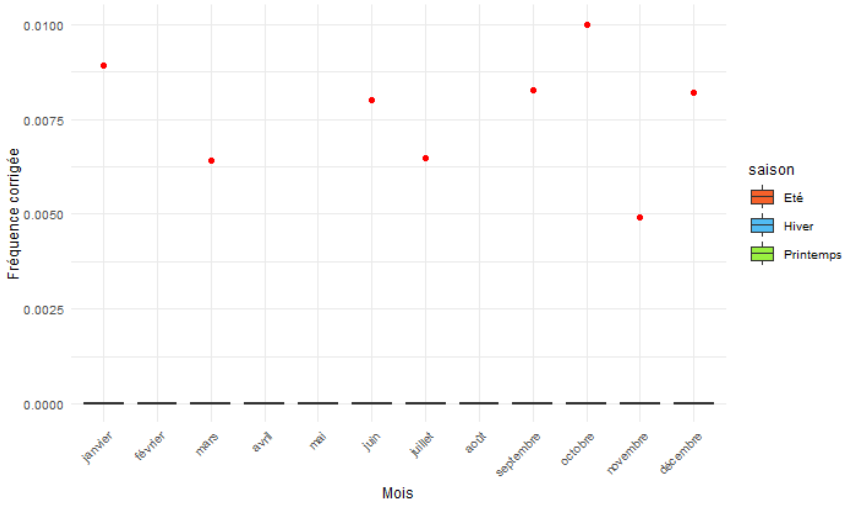


Figure 29 : fréquence de non-conformités détectées par mois relevées par REA pour l'agence Nice Littoral

Les fréquences de NC détectées par l'ARS et par Eau d'Aur à Nice-Littoral sont stables au cours de l'année (Figure 28 et Figure 29). En revanche, Eau d'Azur détecte très rarement des NC (Figure 29). C'est possiblement un signe que REA réalise très peu de tests.

iv. Synthèse

L'agence Rive Droite a quantitativement parlant le meilleur suivi des non-conformités. Seule chose, le nombre de prélèvements dépend de manière assez importante du mois de l'année et la vigilance d'Eau d'Azur semble diminuer pendant les mois de juillet et de décembre. Sinon, d'après les résultats de cette analyse, le secteur Rive Droite est suffisamment surveillée

Pour l'agence Nice Littoral, si le nombre de contrôles est assez stable durant l'année et semble suffire, il apparaît que REA ne contribue quasiment pas à la détection des NCs.

Enfin, le Moyen et Haut pays est l'agence avec les plus gros déséquilibres en terme de nombre de prélèvements. Déjà, le nombre de contrôles passe du simple au double entre l'hiver et l'été. Les données laissent à penser qu'Eau d'Azur réduit sa vigilance les trois premiers mois de l'année, alors qu'ils sont des mois à fréquence assez élevée pour l'ARS. Enfin, il semble que la connaissance des défauts du réseau est bonne et vient compenser l'impossibilité de surveiller en continue l'intégralité du réseau.

4. Conclusion

La problématique de départ est de savoir quelle est la période avec le plus de non-conformités et quels sont les moments où la surveillance faiblie. Il apparaît que la majorité des non-conformités se produisent dans le Moyen et Haut pays. Pour le Moyen et Haut pays et la Rive Droite, la fréquence de détection des non conformités est en moyenne plus élevée en été que le reste de l'année. Pour Nice Littoral en revanche, c'est l'hiver qui a la plus haute fréquence moyenne de NC. Pour chaque agence, octobre est le mois avec la plus haute fréquence de NC ou le second.

Les trois agences sont traitées séparément pour trouver quelles sont les périodes ayant une carence de contrôles. La Rive Droite a un suivi de conformité qui, d'après les données, ne dispose pas de lacunes graves. Les agences Nice Littoral et Moyen et Haut pays ont plus de lacunes. Pour Nice, il semble que le nombre de prélèvements réalisés au cours de l'année est adapté au risque réel mais le nombre de non-conformités détectées par Eau d'Azur est anormalement faible par rapport à l'ARS. Cela porte à croire que le nombre de prélèvements réalisés par REA est lui aussi très faible par rapport à l'ARS. Enfin, le Moyen et Haut pays a une carence en contrôles de la part d'Eau d'Azur les trois premiers mois de l'année et un gros déséquilibre du nombre de contrôles entre l'hiver et l'été.

Pour conclure cette analyse sur les NC, voici les recommandations d'actions à mettre en place pour résoudre les problèmes soulevés :

- Renforcer la surveillance en décembre pour toutes les agences et en janvier, février et mars pour le Moyen et Haut pays en augmentant le nombre de contrôles menés
- Enquêter sur l'origine de la chute de la fréquence de non-conformité en septembre
- S'intéresser davantage au mois d'octobre, qui présente un risque de non-conformité supérieur aux mois d'été pour toutes les agences
- Pour l'agence Nice – Littoral, chercher l'origine du très faible nombre de non-conformités détectées par REA.
- L'analyse pour l'agence Rive Droite est incomplète car trop peu de non-conformités se produisent pour tirer des conclusions
- L'effet propre d'une saison en particulier sur la fréquence de non-conformités détectées n'a pu être mesuré que pour l'agence Moyen et Haut pays.

Pour ce qui est des perspectives :

- Il faudrait trouver des variables de contrôle plus pertinentes que les précipitations et moins corrélées à la saison que les températures pour pouvoir discerner un effet propre des saisons, lié à des problèmes organisationnels.
- Avec davantage de variables, il serait possible de créer un model causal expliquant la présence d'une non-conformité pour un jour donné.
- Il faudrait connaître le nombre exact de prélèvements réalisés par l'ARS et par REA séparément pour ne pas à avoir à extrapoler et pouvoir affirmer plus franchement l'existence de problèmes de suivi.

III. Modélisation de la concentration en sulfate dans les eaux du canal de la Vésubie et de la nappe phréatique du Var sur le site Joseph Raybaud

Les eaux brutes du canal de la Vésubie et de la nappe du Var sur le site Joseph Raybaud sont les deux principales sources d'alimentation en eau potable de la ville de Nice. Mesurer la concentration en sulfate est nécessaire pour des raisons sanitaires mais coûte cher et prend du temps : l'objectif est de mettre en place un système de seuils d'alerte qui notifie l'agent lorsque la concentration en sulfate dépasse les 200 mg/L, le seuil de vigilance réglementaire pour l'eau brute. Le choix d'utiliser le seuil de l'eau brute est doublement motivé : d'une part, une concentration en sulfate à 200 mg/L dans l'eau potable représente déjà une anomalie à surveiller pour REA. D'autre part, les observations supérieures à 250 mg/L sont très rares, ce qui rend très difficile de les prévoir. L'agent averti peut ensuite aviser de la gravité de l'alerte et venir sur site pour confirmer ou infirmer la prévision. La principale contrainte est que, comme il s'agit d'une question de santé publique, il faut minimiser le nombre de faux négatifs en priorité. Rater une alerte est bien plus dommageable que de lancer une fausse alerte. Cependant, trop de fausses alertes peuvent engendrer des surcoûts et de la frustration pour l'utilisateur : le modèle utilisé doit avoir une faible tolérance des faux positifs aussi.

1. Description des données

Les analyses s'appuient sur différentes données hydrométriques et météorologiques : la conductivité de l'eau mesurée en µS/cm, la température de l'eau mesurée en °C et les cumuls de pluie mesurés en mm relevées par 14 stations météorologiques différentes installées sur l'est du département des Alpes Maritimes. Ces stations se situent à Ascros, Carros, Coursegoules, St-Etienne-de-Tinée, Lantosque, Levens, Lucéram, Le Mas, St-Martin-d'Entraunes, St-Martin-Vésubie, Nice, Péone, Puget et Rimplas (Figure 30). Ces données météos sont fournies par Météo France.

La chimie et la physique donnent déjà une idée de l'utilité des deux premières variables : le $SO4^{--}$ est un ion, or les ions sont la cause de la conductivité de l'eau douce. Lorsque la concentration est projetée en fonction de la conductivité, il apparaît une relation linéaire pour les deux sites. Pour J. Raybaud, la corrélation est de $8{,}2\times10^{-1}$. Pour le canal de la Vésubie, elle est de $8{,}9\times10^{-1}$.
Pour la température, une augmentation de la température de l'eau devrait augmenter l'évaporation donc réduire la quantité d'eau disponible pour diluer une quantité constante de sulfate. D'autre part, une eau plus chaude peut être le signe qu'il s'agit d'une période sèche avec des volumes de précipitation faibles. La température de l'eau peut aussi être le reflet de la fonte des neiges. Une relation positive entre concentration et température de l'eau est donc attendue.

Si le cumul de pluie quotidien est observé seul, il n'y a quasiment aucune corrélation avec la concentration. Cela s'explique par l'inertie hydraulique : l'eau doit ruisseler du haut des montagnes avant d'atteindre les vallées où est captée l'eau. La pluie est en grande partie absorbée par les sols, qui sont imperméables à différents degrés et ralentissent le déversement dans les nappes et cours d'eau. Pour tenir compte de tout ce qui précède, le cumul glissant sur x jours est utilisé plutôt que la mesure du jour. Ce x est déterminé pour chaque lieu d'étude et pour chaque station météo comme le nombre de jours qui maximise la corrélation entre cumul glissant sur x jours et concentration. Pour la Vésubie, la corrélation est négative et augmente en valeur absolue en fonction du nombre de jours. Pour J. Raybaud, la corrélation est positive et augmente en valeur absolue en fonction du nombre de jours.

La dernière variable « alerte » est créée en fonction de la concentration en sulfate. Cette variable traduit les seuils d'alerte légaux expliqués en introduction. Elle est encodée de deux manières différentes suivant les cas. Lorsqu'elle est binaire, alerte vaut 0 si la concentration est inférieure à 200 mg/L et 1 sinon. A cause de la réalité métier, une erreur de classification dans un modèle binaire coûte très cher. Au-fur et à mesure des expérimentations, il est apparu qu'une classe intermédiaire de la variable alerte était préférable pour signaler les cas d'incertitude plutôt que de faire une erreur de classification. Une concentration de 180 mg/L est choisie comme « seuil d'incertitude » car les relevés actuels ont une précision à ±10%. Lorsque la variable « alerte » est ternaire, alerte vaut 0 si la concentration est inférieure ou égale à 180 mg/L, 1 si elle est inférieure à 200 mg/L et 2 sinon.

A. Joseph Raybaud

Le premier site d'étude est la nappe phréatique du Var, dans le site d'exploitation Joseph Raybaud. Il y a 83 relevés de la conductivité de l'eau mesurée en µS/cm, de la concentration en sulfate mesurée en mg/L et de la température de l'eau mesurée en °C réalisés entre 2017 et 2024 inclus. Les données proviennent de mesures réalisées par des agents Eau d'Azur. Ces relevés sont réalisés à différents endroits dans l'usine de pompage de la nappe du Var Joseph Raybaud (Figure 30). La majorité des observations sont réalisées lorsqu'il y a suspicion de dépassement. Lorsqu'il y a dépassement, des contrôles de retour à la normale sont programmés. Enfin, certains contrôles sont menés aléatoirement. Comme les contrôles sont surtout menés quand il y a déjà une suspicion de dépassement du seuil de 200 mg/L, l'échantillon est biaisé positivement en terme de concentration. Cependant, étant donnés les objectifs de cette étude, c'est plutôt un avantage qu'un inconvénient : ce biais pousse le modèle à prédire des valeurs plus élevées que dans la réalité, ce qui ajoute une marge de sécurité.

Tableau 5: Statistiques univariées pour le site Joseph Raybaud

Variable	Concentration.mg.L	Conductivité.µS.cm	temperature
Min	150	541	10
1er quartile	180	648	12,4
Médiane	190	687	13,8
Moyenne	193,3	683,4	13,58
3ème quartile	205,5	715,5	14,8
Max	290	885	18,4

Comme il s'agit d'eau souterraine, la température est stable tout au long de l'année. Les concentrations médiane et moyenne sont proches du seuil d'alerte de 200 mg/L (Tableau 5). Justement, 60 observations se trouvent en-dessous de ce seuil contre 23 au-dessus soit 27% de dépassements stricts. Avec l'alerte ternaire, il y a 20 non-alertes (0), 40 cas suspicieux (1) et 23 cas d'alerte (2).
Aucune des trois variables continues mesurées ne présente de valeurs extrêmes (Tableau 5).

B. Canal de la Vésubie

Le second site d'étude est le canal de la Vésubie. Cette fois, il s'agit d'eaux de surface issues de sources de montagne et de la fonte de neige. Il y 892 observations des mêmes trois variables et sur la même

période. Ces données sont aussi des relevés faits en interne. Il y a beaucoup plus d'observations car les prélèvements sont plus systématiques, avec au minimum un par semaine.

Tableau 6 : Statistiques univariées pour le site du Canal de la Vésubie

Variable	Concentration.mg.L	Conductivité.µS.cm	temperature
Min	13	186	4,3
1er quartile	110	404,8	8,8
Médiane	142	488	12,6
Moyenne	143	477,8	12,57
3ème quartile	180	562	16
Max	300	731	24,9

Comme le canal capte de l'eau de surface, les statistiques univariées sur ce site (**Error! Reference source not found.**) sont bien différentes de celle de J. Raybaud (**Error! Reference source not found.**). Déjà, la température de l'eau fait des écarts bien plus importants car elle est exposée à la rudesse des hivers du haut-pays et à la chaleur des étés maralpins. La concentration et la conductivité sont plus basses en moyenne que celles de J. Raybaud. Les *maxima* sont en revanche semblables. Pour ce qui est de la variable « alerte », il y a 82 dépassements stricts soit 9% des observations. En ternaire, ce sont 647 non-alerte (0), 163 cas suspicieux (1) et 82 cas d'alerte (2).

C. Météo quotidienne

La troisième base utilisée résulte de la fusion des cumuls quotidiens (en mm) des précipitations de 14 stations météos différentes de l'Est des Alpes-Maritimes : Ascros, Carros, Coursegoules, Saint-Etienne-de-Tinée, Lantosque, Levens, Lucéram, Le Mas, Saint-Martin-d'Entraunes, Saint-Martin-Vésubie, Nice, Péone, Puget, Rimplas. Chacun de ces relevés est fourni par Météo France (Figure 30).

Tableau 7 : Nombre de jours optimaux et corrélations maximales entre cumul de pluie et la concentration

	Vésubie		Raybaud	
Station météo	Nombre de jours	Corrélation	Nombre de jours	Corrélation
Ascros	54	-0,34	38	0.35
Carros	86	-0.47	54	0.47
Coursegoules	87	-0.47	57	0.48
St-Etienne-de-Tinée	45	-0.41	89	-0.14
Lantosque	86	-0.48	57	0.44
Levens	86	-0.50	54	0.46
Luceram	55	-0.49	57	0.35
Le Mas	54	-0.50	4	0.19
St-Martin-D'Entraunes	43	-0.36	9	0.25
St-Martin-Vésubie	90	-0.55	9	0.36
Nice	87	-0.46	39	0.45
Péone	43	-0.52	9	0.31
Puget	54	-0.51	38	0.34
Rimplas	55	-0.51	38	0.31

Une variable est créée par commune, encodée suivant le modèle cumul_glissant_nomstation_nombrejours. Par exemple pour Nice, la variable correspondante s'appelle cumul_glissant_Nice_87 pour le Canal de la Vésubie car d'après les corrélations obtenues, le nombre de jours optimal est 87.

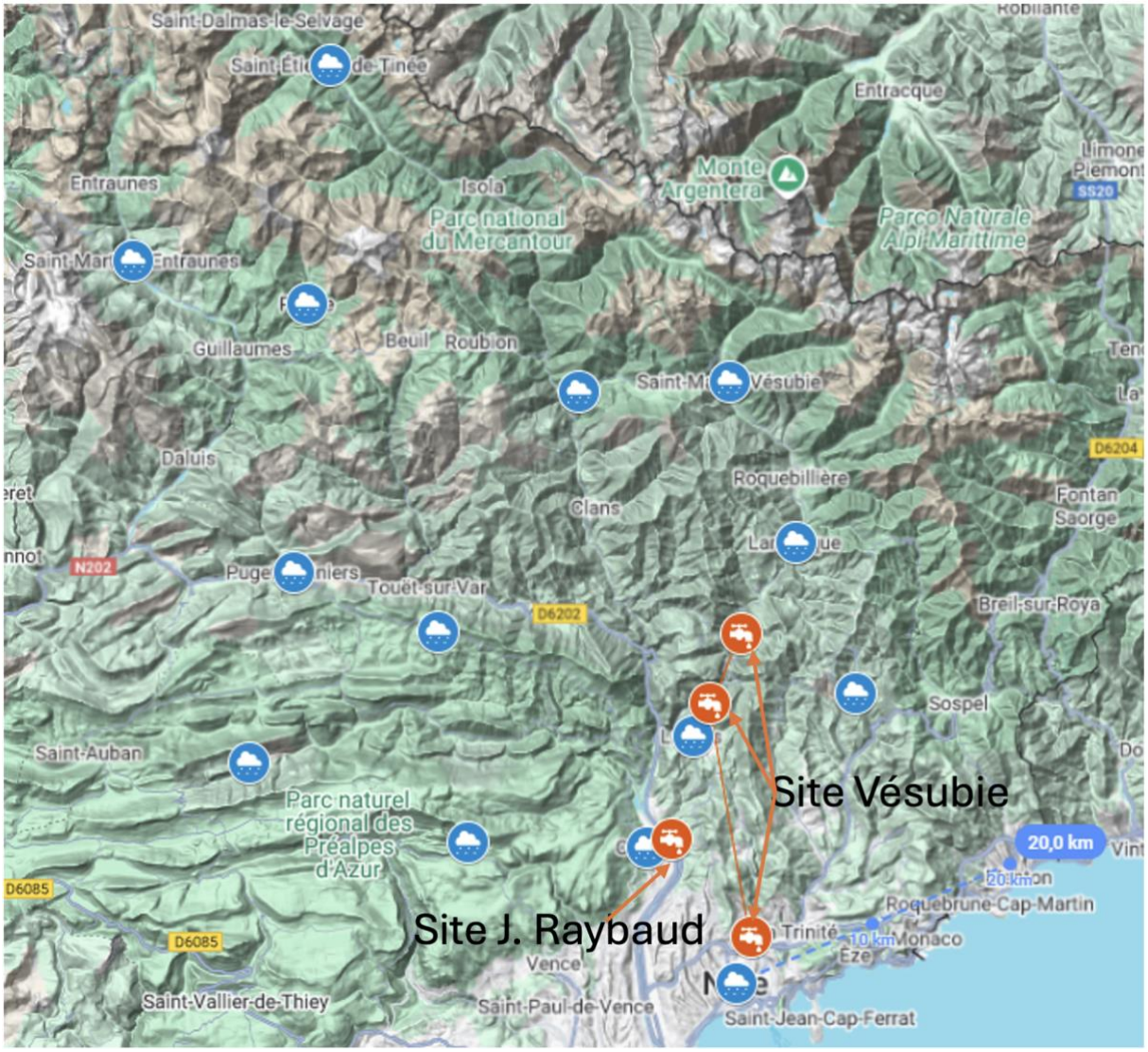


Figure 30 : Carte des stations météos et sites de prélèvement. Chaque point bleu avec une image de nuage correspond à une station météorologique. Chaque point orange correspond à un site dans lequel un ou plusieurs prélèvements de concentration en sulfate, température et conductivité a été réalisé. Le trait orange indique le tracé du Canal de la Vésubie. Le point orange situé sur le Var indique le site J. Raybaud. La distance entre Nice et Menton (20 km) affichée en bleu sert d'échelle.

2. Méthodologie

Sauf mention contraire, chaque modèle utilise l'ensemble des variables sélectionnées ci-dessus. Pour construire les modèles, 80% de l'échantillon d'origine sert de d'échantillon l'entraînement et 20% sert d'échantillon de test des modèles. L'attribution à un échantillon ou l'autre est aléatoire.

A. Choix des variables

Il faut désormais choisir les variables à utiliser dans le modèle : en inclure de trop diminuerait la précision des modèles. Les variables de pluie posent particulièrement problèmes car les stations météo sont suffisamment proches pour qu'il y ait suspicion de colinéarité entre elles. Les variables sont sélectionnées grâce à une Analyses par Composantes Principales (ACP). La concentration n'est pas utilisée pour la construction des dimensions mais elle est tout de même représentée sur les différentes dimensions. Après avoir choisi le nombre de dimensions à garder, des groupes de vecteurs apparaissent sur chaque graphique. De chaque groupe, une seule variable est retenue. Elle sera considérée comme

représentant toutes les autres. La variable gardée est celle qui a la contribution la plus forte de son groupe à la construction de la dimension avec laquelle la concentration est corrélée.

A titre d'exemple, voici le processus en détail pour le site du Canal de la Vésubie. L'ACP utilise deux dimensions qui conservent 77% de l'information. Il faut à présent regarder le cercle d'unité en fonction des dimensions 1 et 2.

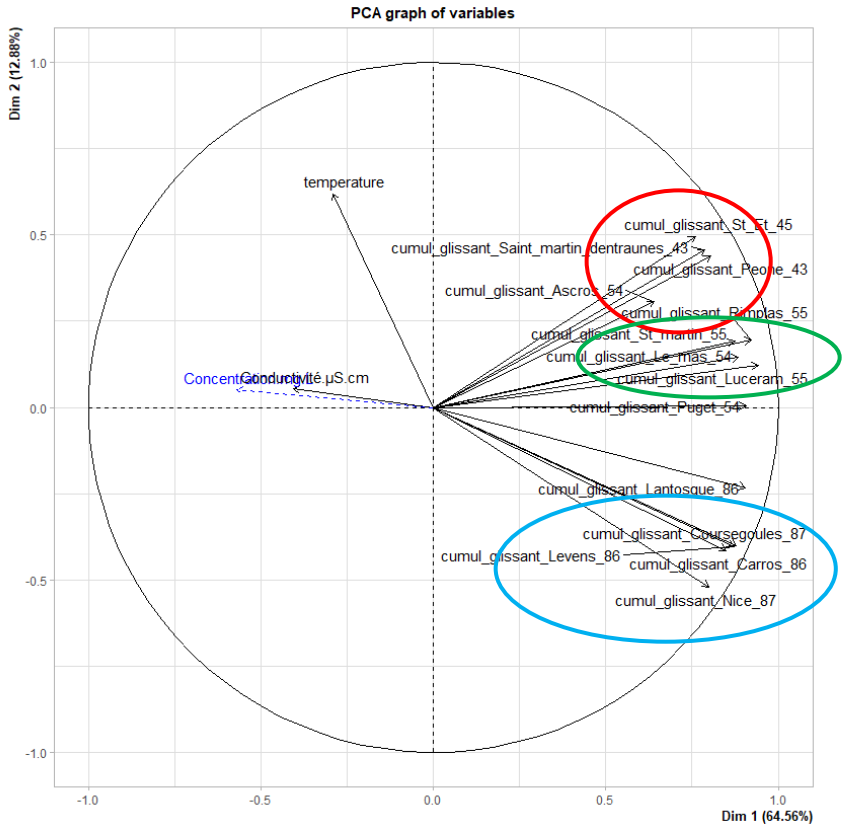


Figure 31 : Cercle d'unité des dimensions 1 et 2 et vecteurs des variables en fonction de ces dimensions construites par ACP

La concentration est bien corrélée avec la dimension 1 (corrélation = -0,58) et quasiment pas avec la dimension 2 (corrélation = 0,02). La conductivité suit de près la concentration : respectivement -0,58 et -0,42 de corrélation avec la dimension 1. De plus, elle est corrélée à 0,89 avec la concentration. La température fait bande à part et est corrélé à -0,37 avec la dimension 1. Puget et Lantosque sortent aussi du lot et sont bien corrélées à la dimension 1 (respectivement 0,90 et 0,91). Pour les autres cumuls météo, trois groupes se détachent, entourés chacun par une couleur différente (Figure 31). Chacune des variables de cumul a sa variance le mieux expliquée par la dimension 1 : ainsi pour chaque groupe, le plus gros contributeur à la dimension 1 représente ce groupe (voir tableau en annexe). Pour le groupe rouge, il s'agit de Péone (contribution de 0,78 à la dimension 1), pour le groupe vert Lucéram (contribution de 0,92 à la dimension 1) et pour le groupe bleu Coursegoules (contribution de 0,89 à la dimension 1). Par le même processus, la conductivité, la température, Péone, St-Martin-Vésubie, Coursegoules, St-Etienne-de-Tinée sont conservées pour J. Raybaud.

B. Algorithmes de classification

Comme l'objectif est de détecter les moments où la concentration dépasse les 200 mg/L, la première idée est d'entraîner un modèle de classification sur l'indicatrice alerte. L'avantage de cette approche est que le résultat est une sortie binaire qui répond directement à la question « Y a-t-il de quoi se déplacer

pour contrôler l'eau ? ». Pour réaliser cette classification, deux algorithmes sont utilisés, le Random Forest et le logit.

L'algorithme Random Forest[2], abrégé RF, s'appuie sur un ensemble d'arbres de décision aléatoires. Ici, 1000 arbres sont créés, chacun utilisant un sous-échantillon aléatoire créé par bagging et un arrangement de variables explicatives aléatoire. Chaque arbre vote pour une classe, la prédiction prend la valeur de la classe pour laquelle il y a eu le plus de vote. Le RF a pour lui qu'il peut découvrir des interactions très complexes entre les variables et donc avoir une meilleure précision globale. En revanche, comme les deux classes sont fortement déséquilibrées pour les deux sites, la précision sur les extrémités de l'échantillon est faible comparée à celle du logit, plus robuste. Pour contrebalancer le déséquilibre entre les classes, il est possible de forcer le bagging à tirer plus probablement la classe minoritaire dans le RF. Les classes sont pondérées par l'inverse de leur fréquence pour deux raisons. D'une part, ces poids ne sont ainsi pas décidés arbitrairement mais dépendent de la distribution d'alerte. D'autre part, prendre un poids plus grand que l'inverse de la fréquence pour la classe minoritaire crée un nombre trop important de faux positifs qui nuisent à l'objectif initial.

Le modèle logit utilise la fonction éponyme $f(p) = \ln\left(\frac{p}{1-p}\right)$ pour prédire la probabilité p d'appartenir à la classe 1 (ici alerte=1). L'idée est que $f(p) = \beta_0 + \beta_1 * X + \epsilon$. Les coefficients β_0 et β_1 sont estimés via la méthode du maximum de vraisemblance. Ces paramètres sont choisis comme ceux qui maximisent la probabilité d'obtenir l'échantillon de départ. Le paramètre de probabilité p est retrouvé par inversion de f.

C. Algorithmes de régression

L'objectif n'a pas changé, il faut encore prédire si oui ou non, la concentration dépasse le seuil des 200 mg/L. Cependant, plutôt que de prévoir l'indicatrice, une étape intermédiaire s'ajoute : la concentration en sulfate est estimée puis la variable de prédiction d'alerte indique si un certain seuil a été dépassé, selon l'estimation de la concentration.

Le premier algorithme utilisé est le RF en mode régression. C'est une généralisation du RF de classification à un espace continu. Cependant, comme dit dans la partie 1, les observations au-dessus de 200 mg/L sont relativement rares : le modèle risque d'être de moins en moins précis à mesure que la concentration réelle augmente.

Comme il s'agit d'une variable continue, il n'est pas possible utiliser de poids pour améliorer les performances de prédiction. A la place, un modèle en trois temps est mis en place. La première étape vise à isoler les observations les plus ambiguës de celles qui ont un niveau d'alerte facilement prédictible. La deuxième consiste à prédire la concentration des observations ambiguës avec un modèle de régression entrainé sur l'ensemble des données. Au final, toutes les observations se voient associer un niveau d'alerte suivant les prédictions réalisées lors des étapes une et deux. Voici la procédure en détail:

[2] Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001)

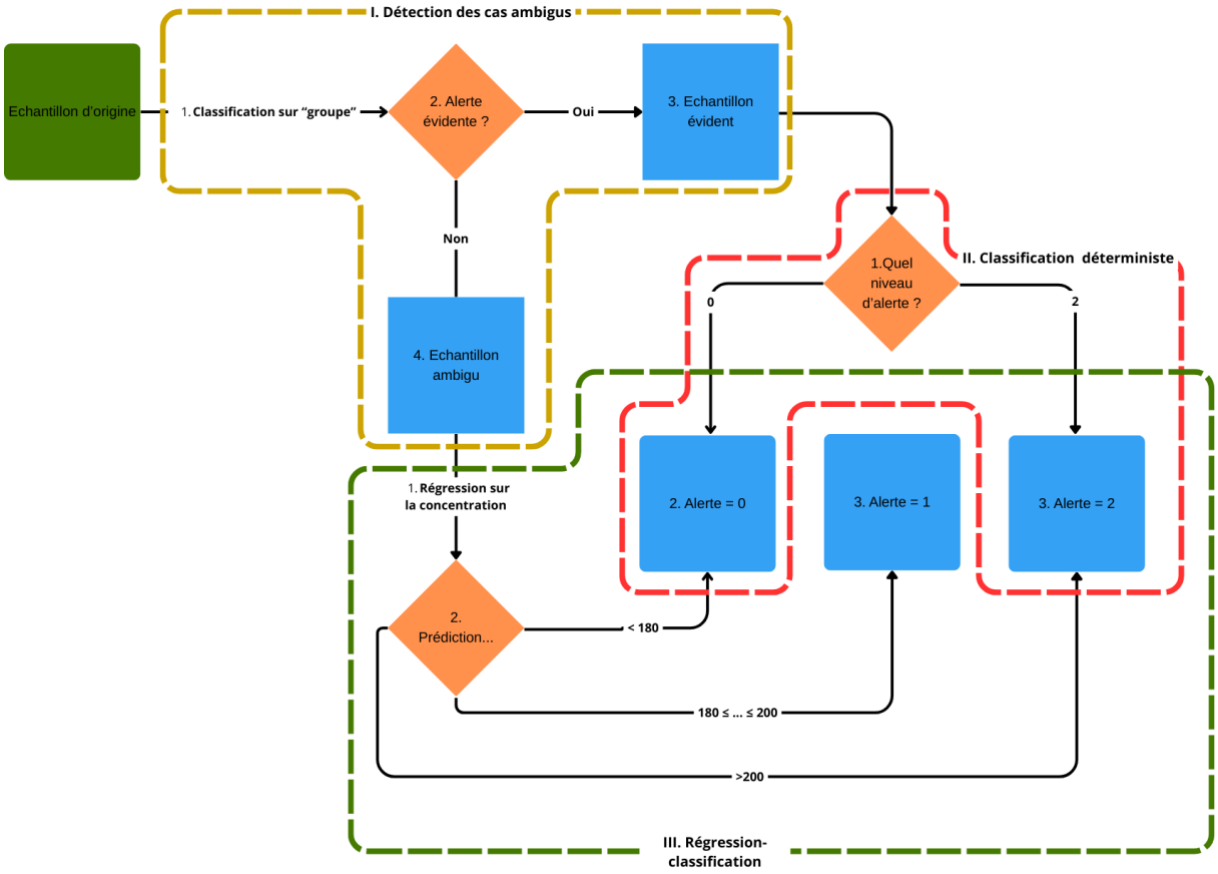


Figure 32 : Schéma explicatif du modèle de classification ternaire avec isolement des observations ambigües

Les nombres entre parenthèses font référence aux étapes de la Figure 32.

Etape préliminaire :

Un clustering est mené sur les observations de la base d'entraînement. Le cluster d'appartenance est enregistré par la variable « groupe ». Tous les cumuls de pluie sélectionnés par l'ACP plus la température, la conductivité et la concentration permettent de regrouper les observations. La fréquence à laquelle la concentration dépasse les 200 mg/L dans chaque groupe est essentielle. Pour chacun des sites, au moins un groupe s'avère contenir uniquement une seule classe d'alerte : une observation qui appartient à l'un de ces groupes n'a aucune ambiguïté sur son niveau d'alerte. Ces groupes sont dits évidents. Les autres groupes sont dits ambigus.

Etape I. Détection des cas ambigus :

Premièrement, chaque observation se voit attribuer à un groupe créé durant l'étape préliminaire grâce à un modèle Random Forest de classification construit sur les observations d'entraînement (I.1). Les résultats de ce modèle permettent de répondre à la question « Le niveau d'alerte associé à cette observation est-il évident ? » (I.2). Cette question sépare l'échantillon en deux sous-échantillons. D'un côté, celui qui contient les observations pour lesquelles le niveau d'alerte ne fait aucun doute (I.3) et de l'autre celui qui contient les observations ambiguës quant à leur niveau d'alerte (I.4).

Etape II. Classification déterministe :

Cette étape ne concerne que l'échantillon évident construit en (I.3). La question est alors « Quel niveau d'alerte est évident pour cette observation, en connaissant la composition du groupe auquel elle a été associée ? » (II.1). Il y a deux réponses possibles à cette question : 0 ou 2. Si une observation appartient à l'échantillon évident, il est impossible de lui attribuer la valeur alerte = 1 car elle encode une incertitude. Si la réponse est 0, elle obtient alerte = 0 pour cette observation (II.2). Sinon, elle obtient alerte = 2 (II.3).

Etape III. Régression-Classification :

Ici, le travail se concentre uniquement sur l'échantillon ambigu construit en (I.3). Cette fois, il faut prédire la concentration de chaque observation. Pour ce faire, un algorithme de régression permet de prédire la concentration à partir des autres variables, entraîné sur l'échantillon d'entraînement. La question est ensuite « Quel est le niveau d'alerte associé à la concentration prédite ? » (III.2). Si elle est strictement inférieure à 180 mg/L, alerte = 0 (II.2). Si elle est strictement supérieure à 200 mg/L, alerte = 2 (II.3). Les observations restantes, qui ont une concentration comprise entre 180 et 200 mg/L, obtiennent le niveau d'alerte 1 (III.3).

Au final, il y a trois sous-échantillons de l'échantillon d'origine. Le premier contient tous les observations pour lesquelles alerte = 0. Dans ce groupe se trouvent toutes les observations pour lesquelles, étant données les conditions météorologiques et hydriques, il est certain que la concentration est inférieure à 200 mg/L et toutes les observations pour lesquelles il est prédit une concentration inférieure à 180 mg/L. Le deuxième contient toutes les observations pour lesquelles alerte = 1. Ce groupe fait office de zone tampon : il est impossible pour l'algorithme de dire si la concentration réelle est supérieure ou inférieure à 200 mg/L. Enfin, le troisième groupe contient toutes les observations avec alerte = 2, c'est-à-dire que la concentration dépasse les 200 mg/L, le seuil réglementaire de vigilance.

Le processus est décliné en trois versions, suivant l'algorithme utilisé pour (III.1) :

- La première déclinaison se fait par un RF de régression.
- La deuxième utilise un modèle linéaire généralisé (GLM). L'idée d'utiliser un modèle linéaire est justifié par la relation observée entre la conductivité et la concentration (Figure 34).
- La dernière déclinaison du processus schématisé en Figure 32 s'appuie sur un réseau de neurones à propagation avant avec une seule couche cachée[3]. Voici en bref son fonctionnement :

  Il faut d'abord introduire des notations. De manière standard, y est la variable expliquée, x_i la variable explicative n°i, w_i le poids associé à cette variable, n le nombre de variables, h_j la sortie du j-ème neurone, b_j le terme constant du j-ème neurone, v_j le poids du j-ème neurone, N le nombre de neurones et c le terme constant de la sortie finale. Pour chaque neurone, n*N poids aléatoires proches de 0 notés w_i sont tirés. Soit $h_j = \sum_i w_i x_i + b_j$ et N poids aléatoires proches de 0 notés v_j pour enfin faire une première estimation de $\hat{y} = \sum_j v_j h_j + c$. De là, l'erreur quadratique moyenne (MSE) fait office de fonction de perte : MSE = $\frac{1}{n}\sum_{k=1}^{n}(y_k - \hat{y}_k)^2$. L'algorithme minimise sa fonction de coût $Coût = MSE + \lambda \sum w^2$, w l'ensemble des poids et λ le facteur de decay. La minimisation est faite avec une méthode de descente de gradient avec

[3] Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*, Fourth edition. Springer, New York. ISBN 0-387-95457-0

comme variables les différents poids. λ peut alors être vu comme une pénalité infligée à un modèle qui donne des poids trop grands.

Au final, lorsque l'algorithme a atteint son nombre limite d'itérations ou son critère de convergence, il retourne l'ensemble des poids finaux.

Cette méthode de séparation entre observations évidentes et ambiguës ne peut être employée ici que pour le site du Canal de la Vésubie. En effet, le site J. Raybaud dispose de trop peu d'observations pour pouvoir entraîner le modèle de (III.1).

D. Correction d'échantillon

Il n'y a que peu d'observations pour le site J. Raybaud (83). Avec le découpage 80%-20% pour construire les ensemble d'entraînement et de test, l'échantillon de test a un cardinal trop petit pour que les modèles puissent être évalués correctement. A la place, le découpage suit les proportions 60%-40%. L'ensemble de test contient alors 34 observations. Pour l'ensemble d'entraînement, l'algorithme Random Over Sampling Examples est utilisé, abrégé ROSE[4]. L'objectif est double : d'une part, il permet d'augmenter le nombre d'observations disponibles pour entraîner le modèle pour améliorer sa précision. D'autre part, il permet de rééquilibrer l'échantillon afin que les observations ayant une concentration supérieure à 200 mg/L soient moins rares. Ce faisant, l'algorithme devient plus performant sur ces valeurs. L'échantillon d'entraînement est composé de 215 observations dont la moitié ont une concentration supérieure à 200 mg/L.

3. Résultats

Les performances des différents modèles évoqués jusque-là sont affichées dans les **Error! Reference source not found.** et **Error! Reference source not found.**. Les résultats sont répartis en deux catégories : d'une part les modèles binaires (**Error! Reference source not found.**) et d'autre part les modèles ternaires (**Error! Reference source not found.**). La spécificité quantifie la capacité du modèle à associer correctement à des individus alerte = 1. Comparer deux rappels n'a de sens que si les modèles ont le même nombre de modalité pour la variable alerte : dans le tableau 1, il s'agit du rappel pour alerte = 0 et dans le tableau 2 pour alerte = 2. La statistique F1 G (détails de son calcul ci-dessous) peut en revanche donner une idée de la puissance relative entre n'importe quel modèle en tenant en compte les objectifs.

A. Joseph Raybaud

Tableau 8 : Comparaison des modèles binaires pour J. Raybaud

Modèle	RF class	Logit	RF reg
F1 G	0.8588	0.8118	0.8353
Recall	0.8182	0.8636	0.8182
Specificity	0.8333	0.5833	0.7500

Le modèle logit est le moins performant d'entre tous en termes de statistique F1 G ($8{,}1\times10^{-1}$). En cause, sa spécificité particulièrement faible ($5{,}8\times10^{-1}$).

Si la statistique F1 G du modèle RF reg est inférieure à celle du modèle RF class, il n'y a qu'une erreur en plus pour RF reg. Leur différence n'est pas significative. Ces trois modèles sont assez peu précis, à cause notamment d'un manque de données.

[4] Lunardon N, Menardi G, Torelli N (2014). "ROSE: a Package for Binary Imbalanced Learning." *R Journal*, **6**(1), 82–92.

Tableau 9 : Comparaison des modèles ternaires pour J. Raybaud

Modèle	RF reg	GLM	NNet
F1 G	0.847	0.8147	0.8559
Recall (classe2)	0.7500	0.4167	0.5833
Specificity (classe2)	0.8182	0.9091	0.9091

Le modèle GLM est le moins performant des modèles ternaires. Son rappel ($4{,}2\times10^{-1}$) est très faible : sur la base de test, l'algorithme prédit pour 7 observations un niveau d'alerte à 1 alors qu'il est en réalité à 2. En parallèle, seules 5 observations sont bien classées dans la catégorie d'alerte 2. Ce modèle ne répond pas aux exigences de réduire au maximum les faux négatifs.

Les modèles RF reg et NNet ont chacun leurs faiblesses et avantages. Le rappel supérieur de RF reg indique que ce modèle capte un plus grand nombre de vrais positifs (alerte = 2 et predict=2) que l'autre modèle. En revanche, la spécificité supérieure de NNet montre que ce modèle capture très peu de faux négatifs. La comparaison de leur MSE tranche en faveur de RF reg : il est de 338 pour le modèle NNet et 286 pour RF reg. Le RF est donc le modèle le plus à même de répondre à la problématique.

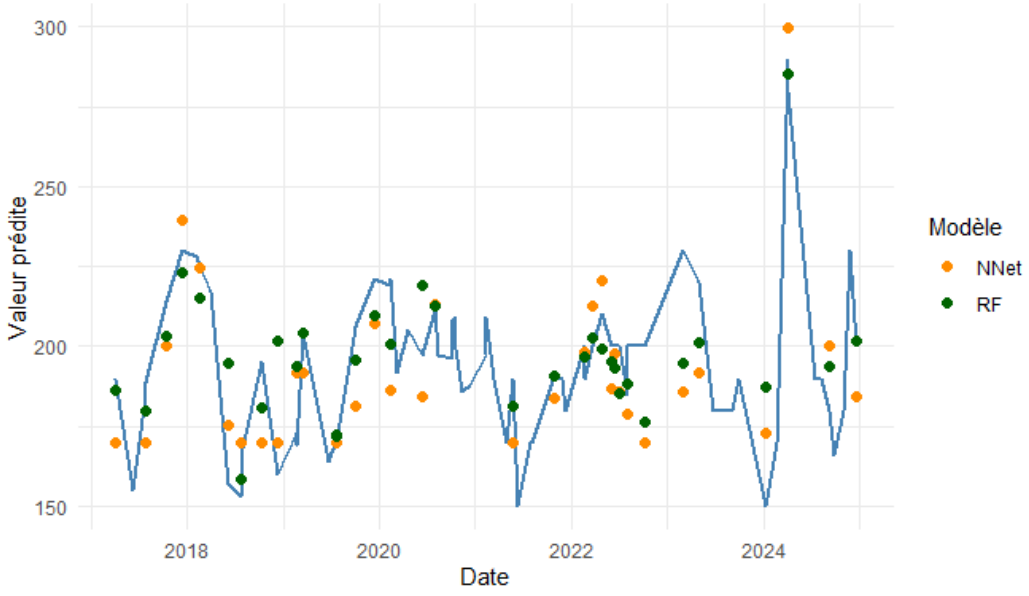


Figure 33 : valeurs prédites des modèles NNet (en orange) et RF reg (en vert) projetées sur les valeurs réelles en fonction du temps

ci-dessus, c'est bien le modèle RF reg (en vert) qui correspond le mieux aux valeurs réelles (en bleu). Cependant, les statistiques calculées pour le modèle RF reg montrent qu'il n'est pas capable de prédire suffisamment précisément la concentration en sulfate pour répondre au problème initial. En effet, l'objectif est de pouvoir précisément estimer si une observation dépasse les 200 mg/L, or la statistique de rappel (recall) vaut $7{,}5\times10^{-1}$: une observation sur quatre qui devrait déclencher une alerte ne le fait pas. Il faudrait davantage de données pour pouvoir estimer des modèles plus précis.

B. Canal de la Vésubie

Tableau 10 : Tableau comparatif des modèles binaires pour le Canal de la Vésubie

Modèle	RF class	Logit	RF reg
F1 G	0,9642	0.9598	0.9648
Recall	0.9819	0.9880	0.9639
Specificity	0.6154	0.4615	0.7692

Les deux classifieurs (RF class et Logit) sous-performent par-rapport à la régression RF (RF reg) sur deux points : premièrement, la régression RF a un meilleur score F1 G que les autres. Deuxièmement, la spécificité du modèle est significativement plus grande, ce qui indique une meilleure performance sur la classe d'intérêt alerte = 1. La classification sur le résultat de la régression répond mieux à l'objectif de minimisation des faux négatifs. La faible spécificité des modèles figurant sur le **Error! Reference source not found.** est la principale motivation pour introduire une classe intermédiaire à la variable alerte.

Tableau 11 : Tableau comparatif des modèles ternaires pour le Canal de la Vésubie

Modèle	RF reg	RF clust	GLM	NNet
F1 G	0.943	0.9637	0.9587	0.9771
Recall (classe2)	0.6923	0.8000	0.8000	0.8000
Specificity (classe2)	0.9698	0.9634	0.9573	0.9878

Déjà, les modèles ternaires (**Error! Reference source not found.**) ont une spécificité au minimum égale à $9{,}5\times10^{-1}$, ce qui est plus grand que le maximum de $7{,}6\times10^{-1}$ pour les modèles binaires : la catégorie alerte = 2 est très pure. Le modèle RF reg est considéré ici comme le modèle de base. Ce modèle est le résultat d'une procédure RF de régression de la concentration puis de l'attribution d'un niveau d'alerte suivant la valeur de la prédiction.

Le modèle RF après filtration (RF clust) performe mieux que le modèle de base sur toutes les statistiques observées : la filtration augmente les performances toutes choses égales par ailleurs. Sur les différents tests menés, ce modèle ne confond jamais les catégories 0 et 2, ce qui répond à l'objectif principal.

La différence de spécificité entre le modèle GLM et le modèle de base disparaît suivant le choix de la graine aléatoire. Cependant, pour plusieurs échantillons de test différents, le modèle a un faux négatif pour la classe alerte = 2, ce qui doit absolument être évité pour répondre à la problématique.
Cependant la relation très linéaire entre la concentration et la conductivité portent à croire que le modèle GLM pourrait être pertinent malgré tout (Figure 34).

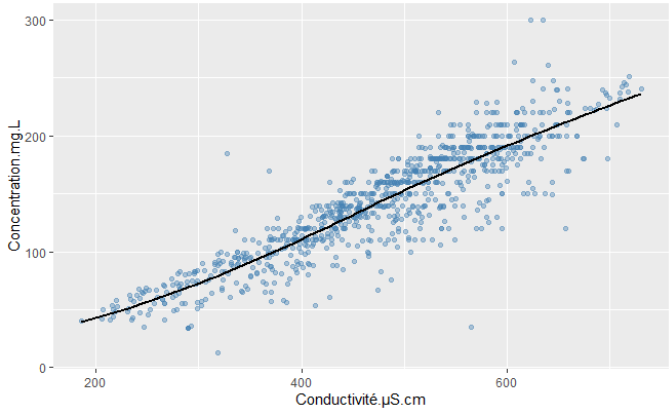


Figure 34 : Conductivité (µS/cm) en fonction de la concentration (mg/L) pour le Canal de la Vésubie

Le dernier modèle NNet est aussi le plus puissant qui est construit ici. Il surpasse en tous points tous les autres modèles vus jusqu'alors. Sur tous les splits train-test réalisés, il ne confond jamais les classes 0 et 2. Les trois alertes qu'il rate sont aussi ratées par tous les autres algorithmes. Pour celles-ci, il prédit des valeurs supérieures ou égales à 189 mg/L.

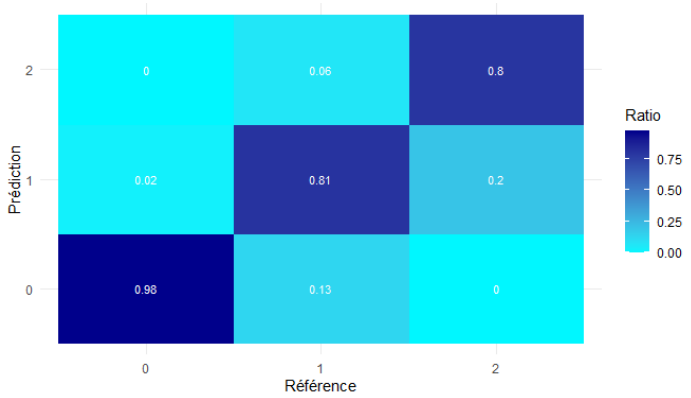


Figure 35 : carte de températures du modèle NNet entraîné pour le site du Canal de la Vésubie. L'intensité de couleur de chaque case montre la propension du modèle à associer des observations à cette case.

Cette carte de températures (Figure 35) montre la performance du modèle. La seconde diagonale ressort très nettement, signe de la bonne précision des prédictions. Cette seconde diagonale correspond aux observations pour lesquelles l'alerte réelle et prédite sont égales. Le groupe d'erreur le plus important proportionnellement parlant est celui pour lequel le modèle prédit un niveau d'alerte à 1 alors qu'il est en réalité à 2, évoqué au paragraphe précédent. Il ne commet aucune erreur grave.

Dans chaque modèle testé, la conductivité est très largement la meilleure variable explicative. Au contraire, la température de l'eau exerce une influence négligeable et ne fait que capter du bruit statistique. Les variables de cumul météo ont un poids particulièrement faible en comparaison à la conductivité. Même après avoir sélectionné plus finement les variables de météo, leur influence individuelle reste faible mais supérieure à la température.

4. Conclusion sur la modélisation des sulfates

A. Discussion des résultats

L'objectif de cette étude est de construire des modèles de prédiction de la concentration en sulfate dans l'eau de deux sites différents, grâce à des variables physiques et météorologiques. Ensuite, il faut classer les prédictions réalisées afin de retourner un signal d'alerte ou non, suivant si la prédiction dépasse les 200 mg/L ou pas. Les contraintes métiers font que le modèle ne peut se permettre de manquer une alerte.

L'idée d'origine de classifier binairement est séduisante mais dans la pratique, il vaut mieux afficher clairement qu'il existe une zone d'ombre dans laquelle l'algorithme hésite. De fait, les modèles les plus performants sont ceux ternaires.

Il faut cependant garder en tête qu'il ne s'agit pas de modèles causaux : la conductivité est une mesure indirecte de la concentration mais c'est bien la concentration qui cause la conductivité et pas l'inverse. Dès lors, la faible contribution des autres variables doit être relativisée : elles ont probablement un effet causal mais caché dans l'ombre de la très forte corrélation conductivité-concentration.

Pour le site Joseph Raybaud, le meilleur modèle est le Random Forest (RF reg). Cependant, il n'est pas suffisamment précis pour pouvoir être utilisé en conditions réelles en l'état.

Pour le site du Canal de la Vésubie, le meilleur modèle est le réseau de neurones avec pré classification des données (NNet). Pour ce site, les prédictions réalisées sont fines et répondent bien à la problématique.

B. Limites

L'une de critiques à émettre sur le travail réalisé est la validité du modèle le plus performant pour le site du Canal de la Vésubie sur d'autre données que celles utilisées ici. Les groupes formés grâce au clustering dépendent de l'échantillon d'entraînement utilisé. Ces groupes dessinent des *scenarii* qui se répètent dans la base de données, mais rien n'indique qu'il n'en existe pas d'autres jamais rencontrés jusqu'ici. Cependant, la diversité des années traitées et le nombre d'observations disponibles pour le Canal de la Vésubie font que le modèle construit pour ce site pourra être utilisé pour d'autres années sur ce même site. Pour le site J. Raybaud, le faible nombre de données affecte non seulement la précision des modèles mais remet aussi sérieusement en question leur validité sur d'autres années.

Pour tous les modèles non linéaires utilisés, il est impossible de connaître en l'état le niveau de confiance que l'algorithme accorde à ses prédictions. La solution longtemps envisagée est d'utiliser une régression quantile. Celle-ci permet d'assurer que la concentration réelle se trouve au-dessus de la valeur estimée par le modèle avec un niveau de confiance personnalisable. Ainsi, le risque pris est chiffré. Cette méthode n'est pas présentée ici car lors des essais, il y a un trop grand nombre de faux positifs pour qu'elle soit applicable.

De même, comme il n'y a pas de valeur de référence pour l'échantillon de test pour la variable groupe, il est impossible d'évaluer la précision de l'étape de classification sur la valeur groupe. En revanche, il se trouve que dans tous les tests menés, il n'y a jamais d'erreur lorsqu'une valeur d'alerte est attribuée manuellement en fonction du groupe (II.1).

Enfin, il faut revenir sur les trois erreurs sur la classe 2 commises par le réseau de neurones entraîné pour le site du Canal de la Vésubie. Tous les autres algorithmes, construits par des méthodes différentes, commettent la même erreur au même endroit, ce qui pointe vers un biais d'échantillon et non pas de modèle. Cependant, après examen, aucun lien évident entre ces observations n'apparaît. De plus, leur faible nombre empêche de trouver un biais systématique qui pourrait être corrigé manuellement.

C. Perspectives

Pour le site J. Raybaud, il faudrait mener une série de prélèvements, décidés aléatoirement et sur une période longue (>1 an), afin d'avoir une base de données suffisamment grande et représentative pour pouvoir construire un modèle robuste. Dans l'idéal, il faudrait avoir à disposition 900 observations représentatives comme pour le Canal de la Vésubie. Ainsi, l'échantillon d'entraînement serait suffisamment grand pour obtenir un modèle précis et l'échantillon de test suffisamment grand pour calculer des statistiques d'évaluation du modèle pertinentes.

Pour le site du Canal de la Vésubie, l'échantillon est de bonne qualité globale. Des données supplémentaires pourrait tout de même être utiles pour mener des tests supplémentaires sur la précision du modèle et donc sa validité externe.

Pour utiliser ces modèles en condition réelle, il faudrait développer une application qui :

- Automatise les appels à l'API Météo France pour récupérer les cumuls quotidiens des stations météos utilisées.
- Automatise le transfert des données de conductivité et de température depuis les capteurs utilisés par REA.
- Ou bien permet le dépôt de données dans un format standardisé à définir.

- Affiche de manière simple, à la manière d'un baromètre, le niveau d'alerte. Si le niveau d'alerte est à 1, il faut que l'utilisateur puisse lire la prédiction faite afin de déterminer s'il est nécessaire ou non de mener un contrôle sur le terrain.

Pour l'instant, les modèles n'ont pas de capacité de projection dans le temps : ils ne prévoient que la concentration du jour. Pour prédire plusieurs jours à l'avance, il est déjà possible de s'appuyer sur les prévisions météorologiques de Météo France. La concentration pourrait être prédite grâce à un modèle de série temporelle. Il faudrait cependant avoir accès aux relevés quotidiens de conductivité. La température de l'eau est relativement stable d'une année sur l'autre pour une période donnée, il ne devrait donc pas être compliqué de la prédire.

Sous réserve d'avoir les bonnes données à disposition, le modèle peut tout à fait être adapté pour des pas de temps différents. Peut-être qu'une estimation toutes les 12h plutôt que chaque 24h serait plus pertinente par exemple.

Le fait que la conductivité soit la meilleure variable explicative du modèle porte à croire que des modèles similaires pourraient expliquer de manière aussi précise la concentration d'autres ions.

IV. Conclusion générale

La qualité de l'eau est un sujet central pour la Régie Eau d'Azur. En tant que service public, la Régie Eau d'Azur a comme raison d'être la satisfaction de ses 174 000 abonnés. Cette satisfaction passe par des standards de qualité élevés et une fiabilité du service. Tout ceci est d'autant plus important qu'il n'existe aucune alternative viable pour les consommateurs : s'ils ne sont pas satisfaits par le service fourni par Eau d'Azur, leur frustration risque d'augmenter très vite, ce qui risque de se répercuter sur les décideurs politiques. Ces derniers pourraient alors décider de mettre fin à la Régie. D'un point de vue légal, REA est aussi contrainte de respecter des normes de qualité, sous peine de sanctions financières et de poursuites judiciaires. Or, pour pouvoir maintenir une bonne qualité, la Régie doit mettre en place des moyens de contrôles qui coûtent cher en temps, en ressources humaines et en ressources financières. Ainsi, le service des Laboratoires et Expertise Eau Potable souhaite rationnaliser son activité de contrôle, afin de maximiser la sécurité et la satisfaction des utilisateurs tout en contenant les coûts d'opérations.

La première problématique traitée durant ce stage a été d'analyser si la stratégie de surveillance des non-conformités de la qualité de l'eau du réseau actuellement en place était adaptée au risque réel de non-conformité. Cette analyse s'est organisée autour de deux axes : d'une part, il a fallu explorer l'évolution de la fréquence de non-conformité détectée en fonction de la période de l'année et du secteur observé ; d'autre part, l'analyse s'est portée sur le nombre de contrôles réalisés par mois et par secteur afin de déterminer les périodes avec le meilleur et le moins bon suivi. Il en ressort que les mois d'été ont une fréquence de non-conformité significativement plus élevée pour la Rive Droite du Var et le Moyen et Haut pays. Pour Nice et la rive gauche du Var, c'est l'hiver qui a une fréquence de non-conformité significativement supérieure au reste de l'année. Les deux rives du Var ont un nombre de contrôles de la qualité de l'eau très satisfaisant tout au long de l'année. Ce qui n'est pas le cas pour le Moyen et Haut pays, pour qui les trois premiers mois de l'année sont une période avec un faible nombre de contrôles.

La seconde problématique était de trouver une méthode qui permettrait d'expliquer la concentration en sulfate par un modèle prédictif. Le modèle obtenu au final est le résultat de nombreuses itérations : déjà, il a fallu s'accorder sur ce qui devait être expliqué : par mesure de sécurité, un agent serait obligé de

mener un contrôle manuel sur place pour confirmer tout relevé automatique. Dès lors, la première approche a été de créer différents modèles binaires qui répondaient à la question « Y a-t-il de quoi se déplacer pour contrôler l'eau ? », autrement dit est-ce que la concentration en sulfate excède les 200 mg/L. Cette première approche ayant de nombreuses lacunes, le modèle final adopte une toute autre approche pour répondre à la même question. Les observations pour lesquelles la réponse est évidente sont immédiatement repérées. Pour les observations restantes, leur concentration est estimée puis, suivant la valeur de cette estimation, la réponse à la question peut être « oui », « non » ou « peut-être ». Le fait de prédire la concentration auparavant et de ne pas forcer le modèle à trancher lorsqu'il hésite permet d'augmenter grandement sa précision et sa fiabilité ainsi que son aisance de lecture.

Pour finir, voici quelques recommandations et perspectives pour poursuivre le travail mené pendant ces trois mois et demi. Déjà, le modèle prédictif construit n'est valide que pour le Canal de la Vésubie, faute de données pour la nappe du Var. Il faudrait donc mener une campagne de collecte de données suffisamment vaste pour obtenir un échantillon représentatif pour ce site. De plus, il serait intéressant d'estimer la concentration en sulfate pour toutes les observations et non pas seulement celles ambiguës, afin d'avoir un modèle plus flexible sur ce que sont considérées des valeurs limites. Le modèle ainsi construit pourrait intégrer une application de supervision, qui permettrait aux agents de connaître la qualité de l'eau sans avoir à se déplacer sur site. Pour le suivi des non-conformités, l'analyse pourrait être approfondie avec une base de données du nombre de prélèvements effectués par mois qui détaille la contribution de l'ARS et de REA à ce nombre de prélèvements.

Les analyses menées au cours de ce stage permettent de répondre en partie aux problèmes de qualité de l'eau qui l'ont motivé. Ces analyses permettent d'une part d'améliorer la traque des non-conformités de la qualité de l'eau mais aussi de mieux les prévoir. Ainsi, la connaissance apportée va aider à garantir une meilleure qualité de l'eau pour tous les consommateurs de la Régie Eau d'Azur.

V. Annexes

Statistique F1 G

La statistique F1 permet de quantifier la performance d'un modèle de classification. Pour ce faire, il faut introduire deux autres scores. Le score de précision, qui mesure pour une classe donnée la proportion d'éléments bien classés (VP) parmi les éléments de cette classe prédite. Les éléments de cette classe sont soit bien classés (VP) soit ce sont des faux positifs, c'est-à-dire des observations qui ne devraient pas être classées ici (FP). Mathématiquement, $Précision = \frac{VP}{VP+FP}$ .Le score de rappel, qui mesure pour une classe donnée la proportion d'éléments bien classés (VP) parmi les éléments attribués à cette classe (VP + FN). Mathématiquement, $rappel = \frac{VP}{VP+FN}$. Ainsi, $F_1 = 2 \times \frac{\text{précision} \times \text{rappel}}{précision + rappel}$. Cette statistique permet de prendre en compte toutes les erreurs commises par le modèle. Or, comme il a été dit, la priorité du modèle est de ne surtout pas manquer d'alerte (prédire 0 pour une valeur réelle d'alerte à 2). En revanche, confondre les niveaux d'alerte 0 et 1 est peu important. Ainsi, la statistique F_1 doit être modifiée pour mieux s'adapter aux besoins de l'étude. Pour ce faire, la matrice

$$W = \begin{matrix} 1 & 0{,}9 & 0{,}2 \\ 0{,}7 & 1 & 0{,}4 \\ 0{,}2 & 0{,}5 & 1 \end{matrix}$$

permet de fixer des poids aux éléments de la matrice de confusion C, en multipliant élément par élément les deux matrices. Ensuite, il suffit de calculer la statistique $F_1\ G$ à partir de la somme des précisions et des rappels calculés pour chaque classe. Les poids sont déterminés de manière inversement proportionnelle à la gravité de l'erreur à laquelle ils sont attachés. Par exemple, confondre les classes 0 et 2 est grave, d'où un poids de 0,2. Au contraire, attribuer à une observation qui a une alerte = 1 une prédiction 0 n'est pas très grave, d'où un poids de 0,9.