



OPTIMISATION ET INCERTITUDE DE LA PREDICTION DES CONTAMINATIONS MICROBIOLOGIQUES DES EAUX DE SURFACE LORS DES EVENEMENTS PLUVIEUX

CONTEXTE :

Depuis quelques années, l'attention des municipalités s'oriente vers les fleuves, les rivières, les canaux, les bras morts, et les plans d'eau, avec une volonté de reconquête de la baignade en ville. De nombreuses villes européennes favorisent l'ouverture de zones de baignade et organisent des compétitions de natation en eau libre dans leurs rivières (Kistemann et al., 2016 ; Mouchel et al., 2020). À l'échelle mondiale, les épisodes de canicule ont récemment accéléré le développement des activités récréatives aquatiques dans les grandes villes, augmentant ainsi la fréquentation des zones de baignade en milieu urbain (Jang, 2016 ; Houtman, 2010).

Cette dynamique s'inscrit dans un contexte d'amélioration progressive de la qualité des eaux de surface, liée à l'évolution des réglementations et à la modernisation des infrastructures (Schreiber et al., 2015). En Île-de-France, malgré une interdiction historique de baignade dans la Seine et la Marne, un fort engagement politique et social vise la réhabilitation de ces rivières pour les usages récréatifs (Noury et al., 2018). Dans le cadre des Jeux Olympiques et Paralympiques de Paris 2024, les collectivités locales se sont engagées à rendre possible la baignade dans la Seine et la Marne d'ici 2025 (Bouleau et al., 2024), avec comme objectif principal l'amélioration continue de la qualité de l'eau à des fins récréatives.

Cependant, en milieu urbain dense, ces activités posent un risque sanitaire réel lié à l'exposition à des microorganismes pathogènes d'origine hydrique. Ces contaminations peuvent être amplifiées par l'intensification attendue des usages récréatifs dans les années à venir (Schijven & de Roda Husman, 2005 ; Islam et al., 2018). Plusieurs sources animales ou humaines peuvent contribuer à ces flux pathogènes (Guérineau et al., 2014). Les bactéries indicatrices fécales (BIF), telles qu'*Escherichia coli* et les entérocoques intestinaux, sont des marqueurs classiques du risque sanitaire, car elles sont excrétées dans les fèces des humains et des animaux à sang chaud (Paruch & Mæhlum, 2012 ; Boehm & Sassoubre, 2014). Ces indicateurs sont utilisés dans le cadre du suivi réglementaire défini par la directive européenne 2006/7/CE (Wade et al., 2003 ; Borja et al., 2020), car leur présence est relativement bien corrélée au risque de gastroentérite (Payment & Locas, 2011). Néanmoins, ce suivi repose sur des analyses coûteuses et longues (au mieux 24 h) après prélèvement, ce qui limite fortement la réactivité des gestionnaires (Manjakkal et al., 2021).

Pour améliorer la gestion quotidienne des sites de baignade en cas de pollution temporaire, une surveillance en temps réel de la qualité microbiologique devient indispensable. L'Organisation mondiale de la santé (OMS, 2018) recommande le recours à la modélisation pour prédire les niveaux de contamination microbiologique. Divers modèles sont disponibles (Mälzer et al., 2016 ; Chen et al., 2020), mais leurs performances varient selon le jeu de données et le contexte. La forte variabilité spatio-temporelle des concentrations en microorganismes d'origine fécale, couplée à la complexité des interactions entre bassin versant et contamination, rend la prédiction des niveaux de pollution particulièrement difficile (Cha et al., 2016). Or, la dynamique des pathogènes hydriques en lien avec les événements pluvieux reste encore peu documentée (Curriero et al., 2001).

OBJECTIFS

Les objectifs menés dans le cadre de la thèse de Manel Naloufi (2025), étaient :

- Optimiser la prédiction des concentrations en *E. coli* à l'aide de modèles d'apprentissage automatique.

- Explorer des approches complémentaires comme l'apprentissage par transfert et la collecte ciblée de données pour améliorer les performances des modèles prédictifs.
- Développer et tester des capteurs à faible coût afin d'assurer une surveillance en continu et en temps réel des paramètres physico-chimiques de l'eau.
- Définir l'incertitude associée à la mesure des BIF et des marqueurs de contamination fécale humaine et animale.
- Étudier la dynamique temporelle des concentrations en *E. coli* après des événements pluvieux, afin d'évaluer la résistance et la résilience des futurs sites de baignade.
- Proposer un cadre conceptuel et des outils pour la surveillance de la qualité microbiologique des eaux urbaines et la gestion quotidienne des sites de baignade.

MÉTHODOLOGIE

1. Optimisation de la collecte de données pour la modélisation de la qualité microbiologique des eaux de surface

Cette thèse s'est dans un premier temps appuyée sur plusieurs expérimentations complémentaires visant à :

- prédire la qualité microbiologique des eaux de surface,
- améliorer la surveillance grâce à des capteurs à faible coût,
- et intégrer l'incertitude dans la prise de décision.

Des modèles d'apprentissage automatique ont été utilisés pour prédire les concentrations en *E. coli* à partir de données physico-chimiques (turbidité, conductivité, température...) et météorologiques (pluie, débit). Trois algorithmes classiques (KNN, arbres de décision, SVM), ainsi que trois approches ensemblistes (Bagging, Random Forest, AdaBoost), reposant sur la combinaison de plusieurs modèles de base, ont été sélectionnés pour cette étude. Les jeux de données utilisés provenaient de deux rivières urbaines, la Seine et la Marne, et les performances des modèles ont été évaluées par validation croisée à l'aide d'indicateurs statistiques tels que le RMSE, le MAE et le RPD.

À partir de l'analyse des erreurs de prédiction, une stratégie de collecte ciblée a été développée pour identifier les situations dans lesquelles les modèles ne parvenaient pas à produire des prédictions fiables, permettant ainsi d'optimiser les campagnes de prélèvement sur le terrain. Le MAPE (Mean Absolute Percentage Error) a été utilisé comme critère principal d'évaluation, en quantifiant l'écart relatif moyen entre les valeurs mesurées et prédites. Cette analyse a permis de distinguer les prédictions fiables des estimations inexactes selon les plages de valeurs des paramètres d'entrée, et ainsi d'identifier les conditions nécessitant des mesures supplémentaire.

Par ailleurs, une approche d'apprentissage par transfert a été mise en œuvre afin d'examiner la possibilité de réutiliser les modèles entre différents bassins versants, dans une logique de mutualisation des jeux de données d'entraînement.

Enfin, un dispositif open-source basé sur Arduino a été développé, intégrant des capteurs physico-chimiques low-cost pour mesurer la température, le pH, la turbidité, la conductivité et l'oxygène dissous. Le système a été d'abord testé en laboratoire, puis déployé sur deux sites parisiens (le Bassin de la Villette). Une évaluation de la stabilité à long terme des capteurs a été réalisée, ainsi qu'une analyse des performances de transmission via le protocole LoRa.

2. Incertitudes et variabilité des dynamiques bactériologiques dans la surveillance des eaux de surface

Cette thèse s'est appuyée par la suite sur plusieurs expérimentations complémentaires visant à :

- analyser l'incertitude liée aux mesures microbiologiques dans l'évaluation de la qualité des eaux de surface,

- intégrer cette incertitude pour une prise de décision réglementaire plus robuste,
- et étudier la dynamique temporelle de la disparition des bactéries indicatrice fécales après une pollution pluviale.

Les prélèvements d'eau de surface ont été réalisés sur quatre sites en Île-de-France, couvrant un gradient de contamination allant d'eaux de baignade à eaux usées, complétés par des données régulières issues de la Ville de Paris et du système automatisé ColiMinder. Différents équipements de prélèvement ponctuel (bécher, pompe, seau) et automatiques (Bühler 2000) ont été testés, avec des protocoles de nettoyage évalués pour limiter la contamination croisée. L'impact du transport, du stockage et du temps d'incubation sur la mesure des bactéries indicatrices fécales (*E. coli* et entérocoques) a été étudié selon les normes ISO. Des analyses moléculaires par qPCR ont ciblé des marqueurs spécifiques de contamination fécale liés aux animaux (oies bernaches, chiens, mouettes et goélands), ainsi que des marqueurs humains et des pathogènes *Campylobacter*. L'incertitude analytique a été estimée à partir de répliquats, et différentes méthodes statistiques ont permis d'évaluer l'incertitude associée à chaque étape, de l'échantillonnage à l'analyse.

Ces incertitudes ont été intégrées dans une classification innovante basée sur la logique floue, permettant de prendre en compte les marges d'erreur autour des seuils réglementaires. Pour cela, des ensembles flous définis par des fonctions sigmoïdes ont été utilisés, attribuant à chaque mesure un degré d'appartenance à plusieurs classes simultanément.

Cinq méthodes de défuzzification (centre de gravité, bisecteur, moyenne des maxima, maximum à gauche, maximum à droite) ont été utilisées pour obtenir une classification numérique précise. Les données ont été analysées sur plusieurs intervalles de temps (4 à 24 h) afin d'évaluer la classification avec une résolution temporelle fine.

Cette méthode a été appliquée aux données issues du système automatisé ColiMinder et comparée aux méthodes classiques et à celle employée lors des Jeux Olympiques 2024.

Parallèlement, six sites en Île-de-France ont fait l'objet d'un suivi intensif pour étudier la disparition de *E. coli* après des événements pluvieux. Ces sites comprennent trois stations en Marne (SMV1, SMV10, SMV14) et trois en Seine à Paris (pont de l'Alma, Tolbiac rive droite et gauche). Les données utilisées combinent des mesures réglementaires historiques et des données à haute fréquence fournies par ColiMinder.

Afin de mesurer la disparition des bactéries en conditions réelles. Un modèle de décroissance exponentielle inverse a été appliqué. En plus du taux de disparition estimé à partir des courbes de décroissance, trois indicateurs ont également été estimés pour caractériser la pollution et la résilience des sites : l'amplitude maximale de contamination, le temps de retour à un état de qualité acceptable (T90) et l'amplitude de récupération après pollution.

Les données ont été analysées statistiquement avec des modèles linéaires, mixtes et tests non paramétriques afin d'identifier les facteurs environnementaux influençant ces dynamiques.

RÉSULTATS

1. Optimisation de la collecte de données pour la modélisation de la qualité microbiologique des eaux de surface

Permettre la baignade en rivière est un enjeu à plusieurs niveaux : i) améliorer la qualité de l'eau, ii) favoriser le cadre de vie des habitants, et iii) diminuer les risques pour les futurs baigneurs. Un suivi de la qualité bactériologique est nécessaire pour autoriser la baignade et assurer une surveillance continue.

Différents modèles d'apprentissage automatique ont été testés comme outils de prédiction des concentrations en *E. coli* dans les eaux de surface. Le modèle Random Forest a obtenu les meilleures performances, avec un RPD moyen de $1,91 \pm 1,65$ pour la Marne (modèle RF avec 11 variables) et $1,47 \pm 0,23$ pour la Seine.

Par ailleurs, l'approche d'apprentissage par transfert a montré que les modèles entraînés et testés sur une même rivière restent globalement plus performants, même si la similarité entre les deux rivières est limitée malgré leur appartenance à la même hydroécocorégion. Ceci confirme

l'importance d'une base de données adaptée et suffisamment fournie pour garantir la fiabilité des prédictions.

Sur la base du meilleur modèle identifié pour la Seine et la Marne (RF), nous avons analysé le pourcentage de prédictions inexactes, estimé à $53,20 \pm 3,50$ % pour la Marne, contre $63,25 \pm 3,11$ % pour la Seine, indiquant une meilleure précision sur le jeu de données de la Marne, qui comportait un nombre plus important de mesures ainsi que trois paramètres supplémentaires par rapport à la Seine.

L'ensemble des paramètres à optimiser en priorité a été identifié en comparant les prédictions raisonnables aux prédictions inexactes, ce qui a permis de déterminer les mesures à ajouter à la base de données. Ces plages de données serviront à définir un système d'alerte pour déclencher des prélèvements ciblés, afin d'enrichir la base de données de manière efficace et à moindre coût.

Un système de mesure continue basé sur des capteurs à bas coût (Arduino) a également été testé. L'étude a confirmé la pertinence de ces capteurs pour le suivi de la qualité de l'eau, à l'exception notable du capteur de turbidité, trop instable et fortement sensible à la lumière. Les capteurs de température et de pH ont montré de très bonnes performances, suivis par le capteur de conductivité, qui a toutefois nécessité une correction par un facteur de compensation de $0,0265\text{ }^{\circ}\text{C}^{-1}$ pour prendre en compte les effets de la température. Sur le terrain, les capteurs Arduino ont présenté des écarts-types faibles entre mesures répétées pour la température ($0,010 \pm 0,004\text{ }^{\circ}\text{C}$), la conductivité ($0,004 \pm 0,001\text{ mS}\cdot\text{cm}^{-1}$), et les deux sondes de pH (pH-1 : $0,028 \pm 0,012$; pH-2 : $0,010 \pm 0,007$). En revanche, la turbidité reste peu fiable (écart-type $42,4 \pm 43,9\text{ NTU}$). Le dispositif a été validé pour une utilisation in situ avec un réseau IoT, couplé à des modules LoRa ayant atteint des portées de 170 à 200 m.

Enfin, un cadre méthodologique flexible a été proposé pour tester et intégrer différents capteurs selon les objectifs de mesure. Ces dispositifs alimenteront à terme des modèles d'apprentissage automatique permettant de prédire la qualité de l'eau et de déclencher des alertes, notamment sur les sites de baignade urbains.

2. Incertitudes et variabilité des dynamiques bactériologiques dans la surveillance des eaux de surface

L'intégration de l'incertitude dans les modèles de gestion de la qualité de l'eau est cruciale pour l'évaluation de la qualité microbiologique des eaux de baignade, afin de permettre une prise de décision éclairée concernant l'ouverture ou la fermeture des zones de baignade. Cette incertitude résulte de plusieurs facteurs, notamment les modalités de prélèvement, le stockage des échantillons, les méthodes analytiques, ainsi que la variabilité naturelle des microorganismes.

Si la variabilité liée aux modes de prélèvement (seau, pompe, bécet) reste relativement faible, un stockage prolongé à température ambiante peut fortement accroître l'incertitude. L'incertitude globale a ainsi été estimée à 40 %, en tenant compte d'une étape de stérilisation des équipements, d'un stockage à froid et d'une analyse le jour même et de l'incertitude analytique liée aux méthodes miniaturisées en microplaques. De même, une incertitude d'environ 42 % a pu être estimée pour le dispositif ColiMinder, intégrant à la fois les incertitudes analytiques et temporelles.

Afin d'intégrer cette incertitude dans la prise de décision, une approche fondée sur la logique floue a été testée et validée. La méthode du centre de gravité s'est révélée la plus robuste, permettant d'obtenir 78 ± 2 % de vrais positifs par rapport à la méthode réglementaire. Cette approche prend en compte la variabilité, en particulier pour les concentrations proches des seuils réglementaires, et s'avère compatible avec une surveillance en temps réel via des capteurs tels que le ColiMinder, facilitant ainsi une prise de décision plus réactive et adaptée.

En complément de l'analyse de l'incertitude et de son intégration dans la prise de décision, la dynamique de dégradation d'*E. coli* après un événement pluvial constitue un facteur essentiel à considérer, car elle permet de refléter le taux effectif de disparition au fil du temps.

Les précipitations entraînent des pics de pollution bactériologique, dus aux débordements des réseaux d'assainissement et au ruissellement. La résolution temporelle fine offerte par les capteurs ColiMinder a permis de mieux détecter ces pics de pollution après les épisodes pluvieux, avec une bien meilleure précision grâce à des mesures toutes les 2 heures, comparées aux mesures

règlementaires ponctuelles journalières. En effet, les concentrations en *E. coli* peuvent être multipliées jusqu'à 22 fois en Seine et 32 fois en Marne après la pluie selon les mesures réglementaires, et jusqu'à plus de 100 fois avec le ColiMinder. Cette meilleure résolution a également renforcé la fiabilité de l'estimation des taux de disparition, ainsi que des indicateurs de résistance et de résilience, dont la robustesse dépend fortement de la fréquence des mesures, notamment en présence de dynamiques de décroissance rapides.

D'après l'analyse des données réglementaires, les taux de disparition d'*E. coli* et les temps de retour sont comparables entre la Seine ($0,47 \pm 0,32$ jours⁻¹ ; $3,50 \pm 1,43$ jours) et la Marne ($0,44 \pm 0,35$ jours⁻¹ ; $3,68 \pm 1,18$ jours).

Cependant, le temps de retour à une qualité conforme pour la baignade varie fortement selon les sites et les événements. Il est souvent conforme à la directive européenne (72 h), mais peut être dépassé, notamment sur le site SMV10 (Marne), contrairement aux sites SMV14 (Marne) et Pont de l'Alma (Seine), qui présentent une meilleure résilience. Le ColiMinder permet une évaluation plus précise, avec un temps de retour moyen estimé à 31 ± 22 h en Seine, contre 87 ± 32 h avec les mesures réglementaires.

La contamination varie également d'une année à l'autre, avec des pics souvent corrélés aux cumuls de précipitations. Après un épisode pluvial, la concentration en *E. coli* augmente rapidement avant de décroître progressivement, soulignant l'importance de suivre la dynamique post-pluie pour une gestion efficace de la qualité des eaux de baignade.

CONCLUSIONS

Ce travail a mis en évidence les enjeux complexes de la gestion de la qualité des eaux de surface en milieu urbain, notamment pour les baignades en rivière (Figure.1). Il a démontré l'intérêt d'approches intégrées alliant technologies avancées, modélisation prédictive et méthodologies innovantes pour mieux surveiller et gérer ces milieux. L'installation de capteurs en continu et l'utilisation d'outils d'aide à la décision, comme la logique floue intégrant l'incertitude des mesures, permettent une classification plus fiable de la qualité microbiologique en temps quasi réel, facilitant la gestion opérationnelle des sites. Par ailleurs, l'association de modèles hydrodynamiques et d'apprentissage automatique offre des perspectives prometteuses pour prédire les événements de pollution et guider les décisions d'ouverture ou de fermeture. Ce travail souligne aussi l'importance d'une réduction continue de l'incertitude liée aux méthodes d'échantillonnage et d'analyse, ainsi que la nécessité d'outils d'optimisation pour l'installation, la maintenance et la validation des capteurs. En combinant ces avancées technologiques et méthodologiques, il est possible d'améliorer la surveillance durable des eaux urbaines, renforcer la sécurité sanitaire des usagers, et contribuer à la préservation des écosystèmes aquatiques dans un contexte de forte urbanisation. Cette approche transversale et interdisciplinaire constitue un levier essentiel pour le développement futur de réseaux intelligents de gestion de la qualité des eaux et de leurs usages récréatifs en milieu urbain.

-
- The diagram illustrates the architecture of a river management system, centered around a 'Rivière' (River) represented by a blue wavy line. The system is divided into several interconnected components:
- Measurement and Data Collection:**
 - Mesure:** The central point of data collection, influenced by 'Capteurs à bas coût' (low-cost sensors) and 'Capteurs haute résolution' (high-resolution sensors). It also receives data from 'Quasi temps réel' (quasi real-time) and 'ColiMinder'.
 - Capteurs à bas coût:** Includes 'Guide installation / validation' and 'Maintenance et traitement des données ?'.
 - Capteurs haute résolution:** Includes 'Réseau de capteurs' and 'Automatisation'.
 - Quasi temps réel:** Includes 'Couplage' and 'Customisation du kit'.
 - Data Processing and Analysis:**
 - BD* (Base de données):** Stores data from 'Mesure' and 'Analyse microbiologique'.
 - Analyse microbiologique:** Processes data from 'Mesure' and 'BD*'. It leads to 'Sources de contaminations' and 'Incertitude'.
 - BIF (Biosensing Interface):** Processes data from 'Analyse microbiologique' and 'BD*'. It leads to 'Incertitude'.
 - Incertitude:** A central concept that influences 'Spatial' and 'Temporel' dynamics.
 - Modeling and Prediction:**
 - Modèle hydrodynamique:** A central model that receives input from 'Optimisation' and 'ML*'. It leads to 'Prédiction'.
 - Optimisation:** Involves 'ML*' (Machine Learning) and 'FL* TL*' (Fuzzy Logic / Temporal Logic).
 - Prédiction:** The output of the hydrodynamic model, leading to 'Surveillance'.
 - Surveillance and Decision Making:**
 - Surveillance:** Monitors the river's state, leading to 'Prise de décision'.
 - Prise de décision:** Involves 'Logique floue' (Fuzzy Logic) and 'Logique floue' (Fuzzy Logic). It leads to 'Gestion des baignades'.
 - Gestion des baignades:** The final outcome, leading to 'Fermeture baignade' (Swimming closure).
 - Alert and Forecasting:**
 - Système d'alerte:** Receives input from 'Optimisation de la BD' and 'Prise de décision'. It leads to 'Alerte pollution' and 'Nowcasting'.
 - Optimisation de la BD:** Involves 'Réduire l'incertitude' and 'Optimisation'.
 - Alerte pollution:** Leads to 'Fermeture baignade'.
 - Nowcasting:** Leads to 'Forecasting'.
 - Forecasting:** Leads to 'Planification des ouvertures' (Planning of openings).
 - Other Key Elements:**
 - Adaptation au site:** Influenced by 'Caractéristiques du nouveau site' and 'Métamodèle'.
 - Caractéristiques du nouveau site:** A key input for the system.
 - Caractéristiques du site:** Influences 'Dynamique' and 'Résilience'.
 - Dynamique:** Influenced by 'Spatial' and 'Temporel' dynamics.
 - Résilience:** Influenced by 'Dynamique' and 'Caractéristiques du site'.
 - Profil de baignade:** Influenced by 'Pollution fécale' (Fecal pollution) and 'Mesure'.
 - Pollution fécale:** A key input for the system.
 - Intégration de incertitude des mesures bactériologiques:** Influences 'Logique floue'.
 - Mesure en continu (Intervalle large):** Influences 'Logique floue'.
- The diagram is created using VISME software, as indicated by the text 'Made with VISME' in the bottom right corner.

- Timothy J Wade, Nitika Pai, Joseph NS Eisenberg, and John M Colford Jr. Do us environmental protection agency water quality guidelines for recreational waters prevent gastrointestinal illness ? a systematic review and meta-analysis. *Environmental health perspectives*, 111 (8) :1102–1109, 2003.
- Angel Borja, Mathew P White, Elisa Berdalet, Nikolaj Bock, Claire Eatock, Peter Kristensen, Anne Leonard, Josep Lloret, Sabine Pahl, Mariluz Parga, et al. Moving toward an agenda on ocean health and human health in europe. *Frontiers in Marine Science*, 7 :37, 2020
- Pierre Payment and Annie Locas. Pathogens in water : value and limits of correlation with microbial indicators. *Groundwater*, 49(1) :4–11, 2011.
- Libu Manjakkal, Srinjoy Mitra, Yvan R Petillot, Jamie Shutler, E Marian Scott, Magnus Willander, and Ravinder Dahiya. Connected sensors, innovative sensor deployment, and intelligent data analysis for online water quality monitoring. *IEEE Internet of Things Journal*, 8(18) : 13805–13824, 2021.
- OMS. Organisation mondiale de la santé. <https://www.who.int/docs/default-source/wash-documents/who-recommendations-on-ec-bwd-august-2018.pdf>, 2018. Accessed on 16 July 2021.
- Hans-Joachim Mälzer, Tim aus der Beek, Silke Müller, and Jörg Gebhardt. Comparison of different model approaches for a hygiene early warning system at the lower ruhr river, germany. *International journal of hygiene and environmental health*, 219(7) :671–680, 2016. ISSN 1438-4639.
- Kangyang Chen, Hexia Chen, Chuanlong Zhou, Yichao Huang, Xiangyang Qi, Ruqin Shen, Fengrui Liu, Min Zuo, Xinyi Zou, Jinfeng Wang, Yan Zhang, Da Chen, Xingguo Chen, Yongfeng Deng, and Hongqiang Ren. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water research (Oxford)*, 171 :115454–115454, 2020. ISSN 0043-1354.
- YoonKyung Cha, Mi-Hyun Park, Sang-Hyup Lee, Joon Ha Kim, and Kyung Hwa Cho. Modeling spatiotemporal bacterial variability with meteorological and watershed land-use characteristics. *Water research*, 100 :306–315, 2016.
- Frank C Curriero, Jonathan A Patz, Joan B Rose, and Subhash Lele. The association between extreme precipitation and waterborne disease outbreaks in the united states, 1948–1994. *American journal of public health*, 91(8) :1194–1199, 2001.