Développement d'un outil de déconnexion des eaux pluviales à l'aide de méthodes de Machine Learning

Tool development for stormwater disconnection using Machine Learning methods

R. Baudet[1], C. Pothier[2], V.A. Montoya-Coronado[3], D. Tedoldi[1], G. Lipeme Kouyi[1]

[1]INSA Lyon, DEEP, UR7429 | robinson.baudet@insa-lyon.fr ; damien.tedoldi@insa-lyon.fr ; gislain.lipeme-kouyi@insa-lyon.fr

[2]INSA Lyon, CNRS, Universite Claude Bernard Lyon 1, LIRIS, UMR5205 | catherine.pothier@insa-lyon.fr

[3]HSM, Univ Montpellier, CNRS, IRD | violeta.montoya-coronado@umontpellier.fr

RÉSUMÉ

Les rejets des déversoirs d'orage dégradent les milieux aquatiques. En outre, du fait des événements extrêmes liés au dérèglement climatique, les volumes déversés vont augmenter. Le déploiement des solutions de déconnexion des eaux de ruissellement permet d'atténuer les déversements, mais les performances de ces stratégies dépendent de l'endroit où elles sont opérées. L'objectif de cet article est d'analyser la pertinence des approches de Machine Learning pour identifier quels facteurs hydrologiques – associés au fonctionnement des sous-bassins versants – expliquent les déversements. Une méthode d'apprentissage non supervisé a été testée sur des précipitations pour les différencier selon leurs caractéristiques. Une méthode d'apprentissage supervisé – arbre de décision – a ensuite été appliquée sur une base de données globale générée grâce à la modélisation hydrologique du bassin versant d'Écully. Les résultats, obtenus avec une *accuracy* de plus de 80%, montrent le rôle majeur des précipitations. Ils indiquent également les secteurs à déconnecter en priorité et les paramètres hydrologiques influençant les déversements, tels que F_{imp} et le *lag time* K. Ces procédés d'apprentissage nécessitent d'être appliqués sur plus de données diversifiées pour continuer à développer un outil de déconnexion des eaux pluviales à l'aide de méthodes de Machine Learning.

ABSTRACT

Combined Sewer Overflows (CSOs) deteriorate waterbodies. Furthermore, due to extreme events related to climate change, water discharges will even increase. The implementation of stormwater disconnection solutions helps to mitigate discharges, but their effectiveness depend on where they are conducted. This article aims to study the relevance of Machine Learning approaches to identify which factors related to the urban sub-catchments operating mode explain CSOs. One unsupervised learning method has been evaluated on precipitations to distinguish them according to their properties. One supervised learning method – decision tree – has then been applied on a general database created using the hydrological modelling of the Écully catchment. The first results, provided with an accuracy larger than 80%, show the role of precipitations data. They also indicate the first sectors to disconnect and the hydrological parameters influencing CSOs, such as F_{imp} or the lag time K. These processes need to be applied on more diversified data so that a tool for disconnecting stormwater using Machine Learning methods can be further developed.

MOTS CLÉS

Déversoir d'orage, gestion des eaux pluviales, Machine Learning, processus hydrologiques, classification

1 INTRODUCTION

Le système classique de gestion unitaire des eaux usées est remis en question depuis de nombreuses années. Les débits de pointe générés par des précipitations contribuent à accroître les risques de débordement du réseau (Zhou, 2014), et les volumes déversés vont augmenter du fait du changement climatique (Gogien et al., 2023). Le réseau unitaire donne également un accès rapide aux eaux urbaines à de nombreux contaminants. Launay et al. (2016) expliquent que les rejets de déversoir d'orage à l'échelle d'un bassin versant constituent une faible part du volume total annuel déversé (18%), mais que ces mêmes rejets forment entre 30 et 95% du flux annuel de plus de 20 micropolluants – caféine, ibuprofène, benzo[a]pyrène par exemple. Des stratégies de réduction des eaux de ruissellement à la source doivent être proposées (Montoya et al., 2022). Pour déconnecter les eaux de ruissellement du réseau, des ouvrages de gestion intégrée des eaux pluviales (OGIEP) permettent de réduire la fréquence des déversements et les volumes déversés.

Calibré, évalué sur des données observées, le modèle semi-distribué TONIC a été construit pour modéliser le fonctionnement hydrologique du bassin versant urbain d'Ecully. Plusieurs scénarios de déconnexion ont également été simulés à l'aide de cet outil dans ce même but de réduire les volumes déversés.

Cet article vise à étudier en quoi un modèle à base d'intelligence artificielle semble être une bonne alternative pour permettre de détecter les secteurs à déconnecter, en proposant des algorithmes de regroupement adaptés (Tan et al., 2018). À l'aide du Machine Learning, il est possible d'analyser les données hydrologiques d'un bassin pour indiquer les critères causant un déversement et les solutions pour s'en prémunir. Différentes méthodes d'apprentissage, supervisé ou non, sont présentées dans cet article et appliquées sur le bassin versant d'Écully.

2 MATERIEL ET METHODES

2.1 Présentation du modèle et de la base de données

Une base de données répertorie 900 simulations du modèle TONIC. Chacune correspond à une situation de déversement et est réalisée à partir d'une pluie choisie parmi les 288 événements mesurés sur Écully de 2007 à 2010. Ces pluies sont décrites avec 3 paramètres – hauteur totale précipitée H (mm), durée D_{pluie} (min) et intensité moyenne maximale I_{Mmax} calculée sur 30 minutes (mm/h). À ces données s'ajoutent les 13 propriétés hydrologiques des 6 Sous-Bassins Versants (notés SBVi, i variant de 1 à 6) composant le bassin versant d'Écully. Parmi ces 13 paramètres, cinq ont été conservés dans l'étude car ils peuvent être influencés par les aménagements urbains qui pourraient y être réalisés. Il s'agit de IL (mm), les pertes initiales, c'est-à-dire les premiers millimètres de pluie infiltrés à la source ; PII (L/s), le débit d'infiltration des eaux claires parasites permanentes ; K (min), le *lag time* du bassin versant ; F_{imp}, la fraction de ruissellement due aux surfaces imperméables ; F_{per}, la fraction de ruissellement due aux surfaces perméables. Cette base de données contient également le débit de pointe Q_{maxDO} et le volume total V_{totDO} calculés au niveau du déversoir d'orage. Grandeurs déterminantes pour connaître la situation de déversement, elles serviront d'étiquettes aux modèles supervisés.

2.2 Analyse préalable des données par méthode non supervisée

Les codes utilisés pour le Machine Learning ont été développés en Python 3.12.12, avec la librairie « sklearn ».

Le but du clustering non supervisé de type K-means – méthode de clustering par partition – est de segmenter des données en créant des groupes, ou clusters, basés sur les valeurs des variables explicatives. Cette méthode est appliquée sur les 288 pluies tombées sur Écully de 2007 à 2010, plus précisément sur les valeurs des trois paramètres hydrologiques – H, D_{pluie}, I_{Mmax}. L'espace des observations pluviométriques se retrouve ainsi discrétisé, en maximisant l'homogénéité intra-cluster et minimisant l'hétérogénéité inter-cluster. Le nombre optimal de clusters a été déterminé empiriquement en appliquant la méthode du coude.

Les méthodes d'apprentissage supervisé, comme l'arbre de décision, nécessitent d'avoir une étiquette, une classe de sortie à prédire. De ce fait, un clustering de type K-means a aussi été effectué sur les données du débit de pointe Q_{maxDO} et du volume total V_{totDO}. Ces deux paramètres sont calculés par TONIC sur le principe du réservoir linéaire et permettent de déterminer si un déversement a lieu, et dans quelle ampleur. Ainsi, le clustering réalisé sur ces deux grandeurs permet de générer une unique classe de sortie à prédire.

2.3 Analyse supervisée

L'arbre de décision est une méthode d'apprentissage supervisé qui a été utilisée sur le jeu de données, qui

comporte donc 900 objets, avec un total de 33 paramètres d'entrée – ou attributs – et une classe de sortie déterminée avec Q_{maxDO} et V_{totDO} – ou étiquette. L'algorithme opère en deux phases : il s'entraîne d'abord en étudiant 80% des objets, avec leurs attributs et étiquettes correspondantes. Puis il travaille sur les 20% restants en connaissant seulement les attributs. Il doit tenter de prédire correctement les valeurs des étiquettes, puis comparer ses prédictions aux véritables étiquettes des 20% restants, qui ne lui ont pas été données au préalable. Pour cela, des métriques évaluant ses performances sont calculées – *accuracy,* précision*, recall*.

L'arbre de décision est un modèle d'apprentissage supervisé qui opère une discrétisation de l'espace des données en utilisant une série de tests sur les paramètres. Guidée par un critère d'homogénéité, ici l'indice Gini, l'objectif de cette segmentation est de maximiser la pureté des nœuds finaux – ou feuilles – pour que ces derniers contiennent des données de la même classe de sortie – nœud pur, ou feuille pure. Dans le cas de cette analyse, une feuille est considérée pure si elle contient par exemple uniquement des données pour lesquelles la valeur de sortie du débit de pointe QmaxDO est particulièrement élevée. La profondeur de l'arbre a été optimisée pour éviter un sous-apprentissage ou un surapprentissage. Cependant, l'étude a été faite pour un arbre unique, dont la segmentation et le critère d'homogénéité peuvent être localement biaisés par le tirage de graine.

3 RESULTATS ET DISCUSSIONS

3.1 Etude des pluies et définition des étiquettes de sortie par la méthode K-means

Les résultats d'apprentissage non supervisé pour les pluies sont présentés Figure 3.1 pour 6 clusters. L'algorithme ne sépare pas les précipitations ne causant pas de déversements des autres, car chaque cluster contient les deux types de pluies. En revanche, il propose des clusters distinguant correctement les pluies selon leurs propriétés. Par exemple, une pluie a été isolée dans le cluster 3 en raison de sa hauteur H élevée. Cette analyse de données pluviométrique est à coupler avec l'analyse des propriétés intrinsèques du bassin d'Écully.

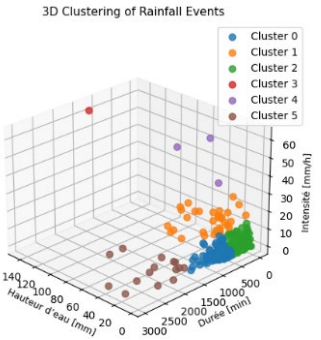

Figure 3.1 : clustering K-means appliqué sur les pluies d'Écully

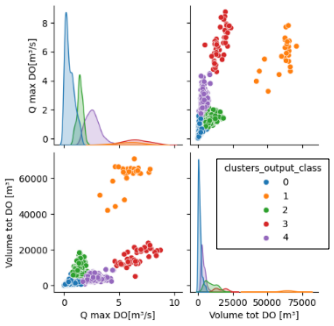

Figure 3.2 : clustering K-means appliqué sur [Q_{maxDO}; V_{totDO}]

De même, les couples de valeurs du débit de pointe Q_{maxDO} et du volume total V_{totDO} calculées par TONIC ont été correctement regroupées en 4 clusters selon leurs amplitudes respectives – voir Figure 3.2. Les 4 clusters générés fournissent directement des étiquettes, ou classes de sortie, pour l'apprentissage supervisé qui suit.

3.2 Résultats de l'apprentissage supervisé : arbre de décision

Dans cette étude, la profondeur idéale de l'arbre présenté Figure 3.3 correspond à 6 branches. L'*accuracy* moyenne, mesurant le taux général de bonnes étiquettes prédites, est de 86.5 ± 5.3 %. La précision moyenne, mesurant la fréquence à laquelle l'algorithme est bon lorsqu'il prédit une étiquette cible, est de 87.2 ± 4.8 %. Le *recall* moyen, mesurant si l'algorithme peut trouver l'ensemble des étiquettes cibles, est de 86.5 ± 5.3 %. Ces valeurs indiquent une bonne correspondance entre les résultats prédits par l'arbre et les résultats attendus.

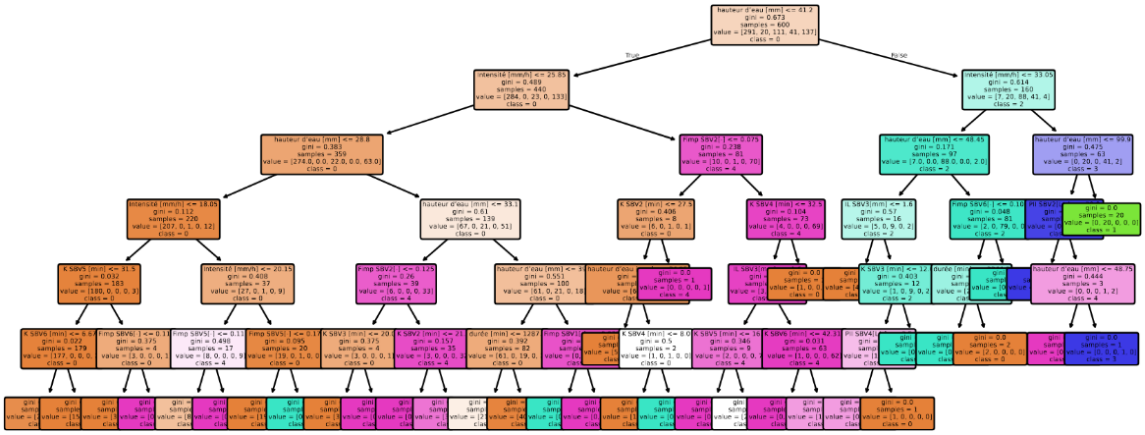

Figure 3.3 : arbre de décision de la base de données d'Écully – méthode supervisée

3.3 Discussion

L'apprentissage supervisé vise à développer une méthode pour identifier selon quels critères sont regroupées les données, mais aussi de voir sur quel paramètre de quel sous-bassin il est possible de jouer pour éviter des déversements : en somme d'étudier les sous-bassins problématiques et les propriétés correspondantes.

Le rôle majeur des propriétés des précipitations est confirmé pour l'arbre de décision, avec des contributions en pourcentage de 52% pour H, 33% pour I_{Mmax} et 3% pour D_{pluie}. Les premiers critères de séparation de l'arbre sont des paramètres de pluie, il est nécessaire de descendre à une profondeur de 2 au moins pour que les critères de séparation commencent à dépendre d'un paramètre hydrologique d'un sous-bassin.

Les contributions des paramètres des pluies et des sous-bassins versants sont présentées dans la partie bleue du Tableau 1. Seules les contributions majeures sont représentées, celles des autres paramètres ne le sont pas car, proches de 0, négligeables. Les contributions par ensemble – sous-bassins 1 à 6, et pluies – sont disponibles dans la partie violette du Tableau 1. Ces résultats sont préliminaires, la base de données étant non exhaustive (Alwosheel et al., 2018) et ne couvrant pas l'ensemble des cas, entre autres ceux de non-déversement.

Les considérations suivantes sont faites en mettant à part les contributions des données de pluie. Au vu des contributions fournies à l'arbre de décision, jouer sur les SBV ou Sous-Bassins Versants 2, 3 et 5 en priorité pourrait avoir un impact. Il est également possible de jouer sur les fractions F_{imp} des SBV 2, 5, 1, et 6, ou les *lag times* K des SBV 3 et 4. D'autres paramètres comme les pertes initiales IL du SBV 3 se démarquent. Des résultats supplémentaires, indiquant les sous-bassins à déconnecter selon l'arbre de décision, mais aussi selon d'autres méthodes d'apprentissage supervisé, comme la forêt aléatoire, seront montrés lors de la conférence.

Tableau 1 : contributions en % à l'arbre de décision des paramètres en bleu, et sommées par ensemble en violet

H (mm)	*I_{Mmax} (mm/h)*	F_{imp} SBV2 (-)	*D_{Pluie} (min)*	F_{imp} SBV5 (-)	SBV1	SBV2	SBV3	*Pluie (H, D_{pluie}, I_{Mmax})*
52,30	*32,82*	3,22	*2,95*	1,39	0.98	4.58	2.43	*88.07%*
IL SBV3 (mm)	K SBV3 (min)	F_{imp} SBV1 (-)	K SBV4 (min)	F_{imp} SBV6 (-)	SBV4	SBV5	SBV6	
1,37	1,06	0,98	0,78	0,76	1.10	1.82	1.02	

4 CONCLUSION

La méthodologie du travail de l'apprentissage non supervisé et supervisé a été développée sur les données actuelles, il est possible de l'appliquer sur une base de données plus conséquente. Le projet TONIC s'intéressait à l'analyse du bassin versant de Figeac : il serait pertinent d'étudier comment les méthodes d'apprentissage, une fois consolidées sur Écully, s'adapteraient à ce nouveau terrain. Tester d'autres méthodes d'intelligence artificielle comme le réseau neuronal est possible. Mais ce dernier pourrait être moins adapté à la quantité de données que l'apprentissage supervisé, qui fournit déjà de bons résultats avec une base de données améliorable. L'outil de déconnexion des eaux pluviales assisté par intelligence artificielle, présenté dans cet article, mérite d'être développé, afin de proposer aux collectivités une analyse accessible de l'évitement des déversements.

BIBLIOGRAPHIE

Alwosheel, A., Van Cranenburgh, S., and G Chorus, C. (2018). *Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis*. Journal of Choice Modelling, Vol. 28, 167-182, ISSN 1755-5345.

Gogien, F., Dechesne, M., Martinerie, R., and Lipeme Kouyi, G. (2023). *Assessing the impact of climate change on Combined Sewer Overflows based on small time step future rainfall timeseries and long-term continuous sewer network modelling*. Water Research, Vol. 230, 119504, ISSN 0043-1354.

Launay, M., Dittmer, U., Steinmetz., H. (2016). *Contribution of combined sewer overflows to micropollutant loads discharged into urban receiving water*. Novatech 2016 - 9ème Conférence internationale sur les techniques et stratégies pour la gestion durable de l`Eau dans la Ville, Lyon, France.

Montoya-Coronado, V.A., Bret, P., Molle, P., Castebrunet, H., Tedoldi, D., and Lipeme Kouyi, G. (2022). *Stratégies de déconnexion des eaux pluviales à l'échelle d'un bassin versant pour réduire les déversements*. TSM, Numéro 4, 27–37.

Roy, S. (2025). *Automated Detection of Sub-Catchment to Disconnect stormwater from Combined Sewer Overflow (CSO) Structure*. Lyon: INSA Lyon, Centrale Lyon, Water and Wind engineering WWE Master, 10p.

Tan, P.-N., Steinbach, M., Karpatne A., and Kumar, V. (2018). *Introduction to Data Mining, Second Edition*. Pearson, ISBN: 9780133128901.

Zhou, Q. (2014). *A Review of Sustainable Urban Drainage Systems Considering the Climate Change and Urbanization Impacts*. Water, 6(4), 976-992.