Prédiction de surverses et catégorisation d'événements pluviaux à Montréal

Alice Breton - alice.breton@polymtl.ca
Encadré par Jonathan Jalbert
https://github.com/AliceB08/prediction_surverses

Août 2019

Contents

1 Introduction

Une surverse est un rejet des eaux usées et des eaux pluviales dans un milieu naturel sans traitement préalable, souvent à la suite de fortes pluies. Des microbes, diverses particules toxiques et débris se retrouvent ainsi dans nos cours d'eau, posant un problème à la fois sanitaire et environnemental. Dans un contexte d'augmentation d'épisodes pluviaux intenses, il est important de comprendre les liens existants entre précipitations et surverses pour être en mesure de prédire de façon précise l'occurrence de débordements.

Le but de ce projet est d'explorer les données de surverses sur l'île de Montréal. A la différence d'études déjà effectuées, nous nous intéressons au gain de précision sur la prédiction de surverses, suite à l'utilisation de variables explicatives différentes par ouvrage.

Les données publiques de surverse sur l'île de Montréal sont disponibles de 2013 à 2017. Pour avoir un lien entre les surverses et les précipitations uniquement, nous excluons les mois de fonte des neige. Nous nous plaçons donc entre les mois de mai et d'octobre pour chaque année de l'étude.

2 Préparation des données

2.1 Données pluviométriques

Les données pluviométriques proviennent d'Environnement Canada. Nous considérons les 3 stations de l'Île de Montréal : APET (aéroport Trudeau), MT (McTavish) et SADB (Sainte-Anne-de-Bellevue). La carte 1 montre leur emplacement sur l'île. Nous divisons les données originales par 10 pour avoir des précipitation en mm.

Pour avoir une représentation de la météo près de l'ouvrage, nous utilisons en priorité les données de la station la plus proche. Si celle-ci a des données manquantes, nous les complétons avec la deuxième station la plus proche, et ainsi de suite. Si aucune donnée n'est disponible pour une certaine heure, nous fixons la précipitation à 0 (entre 2013 et 2017 14 heures sont manquantes simultanément dans les 3 stations).

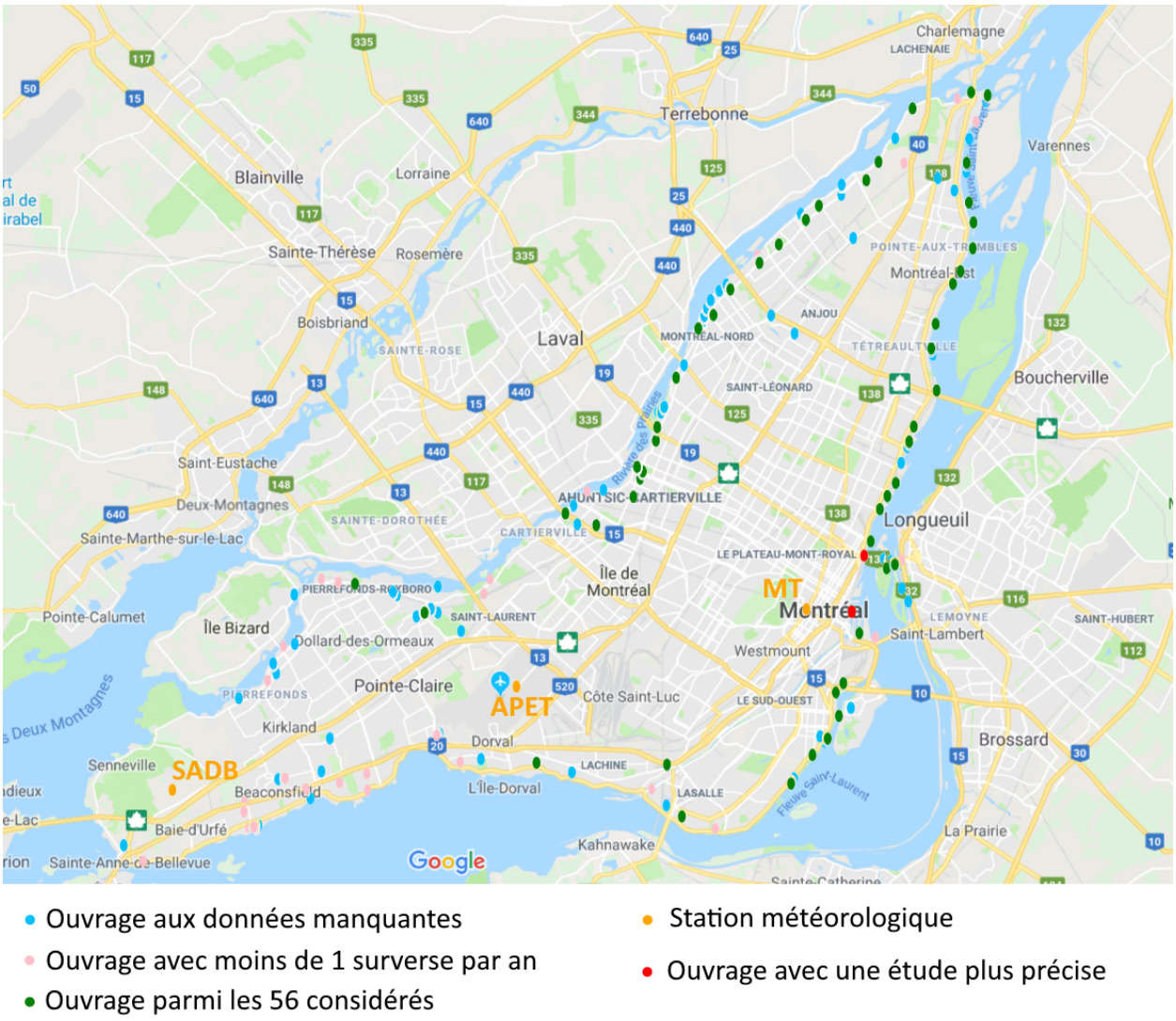

Figure 1: Emplacement des stations météo et des ouvrages à Montréal

2.2 Création d'événements pluviaux

Dans la littérature, il est courant de créer des "événements pluviaux" (**EP**) pour exploiter de manière plus méthodique les données pluviométriques [3], [2]. Celles-ci sont regroupées telles que deux EP soient séparés par une période sèche (pluie horaire inférieure à 0.3mm/h) de minimum X heures consécutives. Cette valeur peut être de 6h [2] ou 12h [3] par exemple. Nous choisissons une durée de 12h pour éviter d'avoir un trop grand nombre d'événements pluvieux par jour ce qui permet d'être cohérent avec le comptage quotidien des surverses.

Entre le 1er mai et le 31 octobre (184 jours) de chaque année de 2013 à 2017, nous avons 297 EP avec une période sèche de 6H et 247 EP avec une période sèche de 12h.

2.3 Surverses

Nous obtenons les surverses à partir des données publiques du Québec. La liste des ouvrages de surverse à Montréal est disponible ici http://donnees.ville.montreal.qc.ca/dataset/ouvrage-surverse et les mesures de débordements pour la ville de Montréal ici https://www.donneesquebec.ca/recherche/fr/dataset/vmtl-debordement. Pour chaque jour entre 2013 et 2017 nous avons accès à la durée de la surverse. Malheureusement nous ne connaissons pas la quantité d'eau déversée, dans ce projet nous nous concentrons donc sur la présence (1) ou absence (0) de surverse, pour chaque jour et pour chaque ouvrage.

La ville de Montréal compte environ 170 ouvrages de surverse, répartis sur le fleuve Saint Laurent et la rivière des Prairies (voir fig 1). La majorité des ouvrages sont suivis dynamiquement, c'est à dire que les mesures sont automatiques et quotidiennes. Certains ouvrages sont surveillés hebdomadairement et manuellement, nous ne connaissons alors pas le jour exact du débordement. Nous ne garderons donc que les ouvrages contrôlés dynamiquement.

L'importation des données de surverse est plus complexe que celle des données pluviométriques puisqu'il n'y a pas de normalisation d'une année à l'autre. Un pré-traitement des fichiers csv est nécessaire et est assez long, les détails sont donnés dans le notebook.

Parmi les 170 ouvrages nous sélectionnons les 87 ouvrages qui n'ont pas de données manquantes entre mai et octobre de 2013 à 2017. De plus, un grand nombre d'ouvrages n'ont peu ou pas de surverses (figure 2) donc nous gardons les 56 ouvrages qui ont strictement plus de 1 surverse par an.

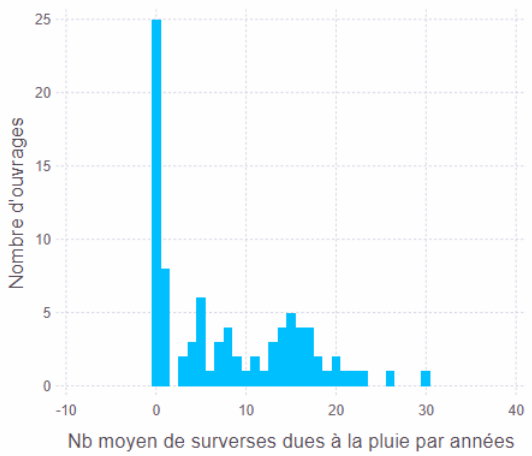

Figure 2: Nombre moyen de surverses causés par la pluie par an

Dans la suite du projet nous ne garderons que les surverses qui sont dues à la pluie. Il peut y avoir par exemple des surverses pour cause de travaux, urgence, ...

2.4 Association événement pluvial - surverse

Pour illustrer notre travail nous utiliserons les graphiques et données de deux ouvrages se situant dans le bassin Alexandra au Vieux Port : **4350-01D** et **4340-03D**. Dans le notebook il est néanmoins possible de sélectionner n'importe quel autre des 56 ouvrages.

Une fois les EP (événements pluviaux) créés, nous pouvons regarder s'ils ont généré ou non une surverse dans un ouvrage particulier.

Pour chaque surverse de l'ouvrage 4350-01D regardons les EP qui finissent la veille ou le jour même de la surverse. Si un seul événement est trouvé alors nous lui associons la surverse. 11 des 38 surverses de cet ouvrage ont deux EP qui suivent cette condition. La figure 3 montre les caractéristiques de ces doublons, pour chaque débordement.

Les barres en orange correspondent à l'événement précédent directement la surverse et les barres vertes à l'événement avant celui-là. L'événement le plus proche est très souvent plus intense, plus long, a une profondeur de pluie plus grande. Nous choisissons donc d'associer la surverse à l'EP le plus proche en cas de doublon. Dans d'autres études une surverse est parfois rattachée à plusieurs EP avec un score associé à la probabilité de déclenchement. Pour ce court projet nous simplifions l'association avec les deux critères suivants

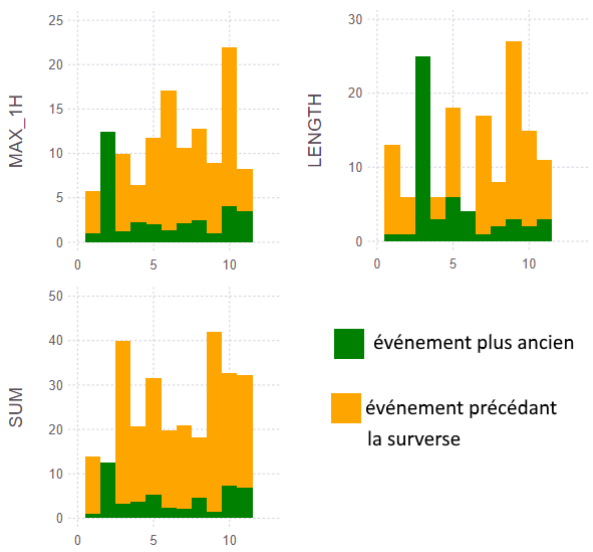

Figure 3: Comparaison des EP précédent une surverse pour l'ouvrage 4350-01D

- L'EP prend fin le jour même ou la veille de la surverse
- Si plusieurs EP correspondent au critère au-dessus, nous choisissons l'EP le plus proche temporellement

Il arrive qu'un EP qui dure plusieurs jours génère plus d'une surverse. Puisque nous étudions la présence ou l'absence de surverse après un EP, nous considérerons une variable binaire et non le nombre de surverses déclenchés par un EP.

Certaines surverses ne sont pas associées avec les critères précédents, même si la labélisation indique qu'elle est due à la pluie. Il faudrait dans un projet futur mieux comprendre d'où proviennent ces surverses.

Voici les variables explicatives que nous obtenons pour chaque EP, et qui pourraient expliquer la présence d'une surverse :

- **SUM** La profondeur totale de l'événement pluvial en mm
- **MAX_1H ou MAX_2H** L'intensité maximale en mm/h de l'événement pluvial (pour 1H ou 2H). L'intensité prise sur 2h est divisée par 2 pour être ramenée à un taux horaire
- **LENGTH** La durée en heure de l'événement pluvial. Il peut y avoir des périodes sèches au milieu, si elles ne durent pas plus de 5h ou 11h d'affilée
- **MEAN** Moyenne de précipitation de l'événement pluvial en mm/h
- **PTS_1H ou PTS_2H** Peak-to-sum ratio avec IM prise sur 1H ou 2H
- **MEAN_PREC** Moyenne de précipitation depuis la dernière surverse en mm/h

2.5 Caractéristiques des EP générant ou non une surverse

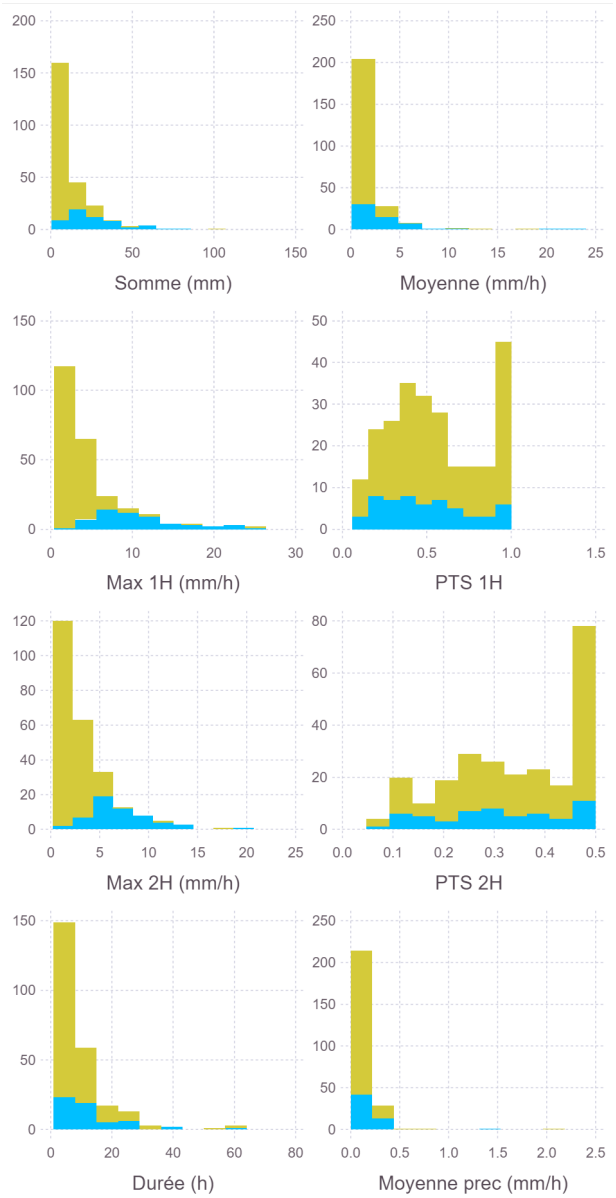

Figure 4: Répartition des caractéristiques d'un EP selon la génération (bleur) ou non (jaune) d'une surverse pour l'ouvrage 4350-01D.

Une fois les EP labélisés, regardons la répartition de leurs caractéristiques en fonction de lé présence ou absence de surverse. Voici les données pour l'ouvrage 4350_01D (fig 4).

A première vue les variables les plus discriminatives pour cet ouvrage sont SUM, MAX_1H, MAX_2H, MEAN. Les autres variables LENGTH, MEAN_PREC, PTS_1H et PTS_2H permettent moins bien de trancher.

3 Régression logistique

Pour prédire des surverses la précipitation totale est une variable très utilisée [3], avec un modèle probabiliste (pour une hauteur de précipitation on associe une probabilité de surverse). L'intensité maximale peut également être exploitée[1]. Dans les deux cas, les modèles sont ajustés pour chaque ouvrage, mais c'est toujours la même variable explicative qui est utilisée pour tous les ouvrages.

Dans ce projet nous nous intéressons au potentiel gain de précision en choisissant pour chaque ouvrage une variable explicative différente en plus de seuils différents.

Dans un premier temps nous utiliserons des modèles de régression logistique à 1 variable explicative puis des modèles avec plusieurs variables. Les paramètres seront estimés par maximum de vraisemblance. Nous regardons la puissance de prédiction de ces modèles (précision, sensibilité et spécificité).

3.1 Modèles à 1 variable

Pour chacune de nos 8 variables nous créons un modèle de régression logistique avec des paramètres ajustés sur les données d'un ouvrage. Pour illustrer le projet avec des figures et tableau nous utiliserons l'ouvrage 4350-01D puis 4340-03D.

Les figures 5 et 6 présentent les modèles de régression logistique à 1 variable, classés par précision décroissante.

Pour l'ouvrage 4350-01D la variable SUM donne les meilleurs résultats, et pour l'ouvrage 4340-03D c'est la variable MAX_2H (précision améliorée de 6% par rapport à la variable SUM). Nous voyons que pour cet ouvrage, il y a une amélioration significative de la précision en changeant de variable explicative.

Pour nos 56 ouvrages voici les variables explicatives qui permettent d'avoir la meilleure précision. SUM : 33 ouvrages, MAX_2H : 13 ouvrages, MAX_1H : 5 ouvrages, MEAN_PREC : 4 ouvrages, LENGTH : 1 ouvrage.

	VARIABLE	SENSIBILITE	SPECIFICITE	PRECISION	SIGNFICATIF
	Symbol	Float64	Float64	Float64	Bool
1	SUM	0.454545	0.976636	0.906883	true
2	MAX_2H	0.333333	0.96729	0.882591	true
3	LENGTH	0.0909091	0.990654	0.870445	true
4	MEAN	0.0606061	0.990654	0.866397	true
5	PTS_1H	0.0	1.0	0.866397	false
6	PTS_2H	0.0	1.0	0.866397	false
7	MEAN_PREC	0.030303	0.995327	0.866397	true
8	MAX_1H	0.212121	0.962617	0.862348	true

Figure 5: Modèles à 1 variable classés par précision décroissante pour l'ouvrage 4350-01D

	VARIABLE	SENSIBILITE	SPECIFICITE	PRECISION	SIGNFICATIF
	Symbol	Float64	Float64	Float64	Bool
1	MAX_2H	0.642857	0.963351	0.890688	true
2	MAX_1H	0.607143	0.963351	0.882591	true
3	SUM	0.446429	0.95288	0.838057	true
4	MEAN	0.178571	0.979058	0.797571	true
5	PTS_1H	0.0	1.0	0.773279	false
6	PTS_2H	0.0	1.0	0.773279	false
7	MEAN_PREC	0.0178571	0.989529	0.769231	true
8	LENGTH	0.0178571	0.984293	0.765182	true

Figure 6: Modèles à 1 variable classés par précision décroissante pour l'ouvrage 4340-3D

En utilisant la "bonne" variable explicative pour chaque ouvrage, nous améliorons en moyenne de 0.8% la précision.

3.2 Modèles à plusieurs variables

Regardons si une combinaison de plusieurs variables explicatives permet d'améliorer encore cette précision. Nous testons toutes les combinaisons de variables possibles parmi les 8 disponibles puis gardons seulement les modèles où toutes les variables utilisées ont un pouvoir prédictif significatif. Voici le classement par ordre de précision des meilleurs modèles obtenus pour nos deux ouvrages du bassin Alexandra (figure 7 et figure 8).

Nous remarquons que pour l'ouvrage 4350-01D la précision reste la meilleure pour la variable explicative SUM, mais que pour notre deuxième ouvrage dans le bassin d'Alexandra la précision est améliorée de 8% en combinant le maximum d'intensité sur 1H et la moyenne de l'événement pluvial par rapport à l'utilisation de la précipitation totale.

Quand on compare la précision maximale obtenue

	COMBINATION	SENSIBILITE	SPECIFICITE	PRECISION
	Array…	Float64	Float64	Float64
1	[:SUM]	0.454545	0.976636	0.906883
2	[:MAX_2H, :PTS_2H]	0.454545	0.976636	0.906883
3	[:SUM, :MAX_1H]	0.484848	0.96729	0.902834
4	[:MAX_1H, :PTS_1H]	0.454545	0.971963	0.902834
5	[:MAX_2H, :PTS_1H]	0.424242	0.976636	0.902834
6	[:MAX_1H, :PTS_2H]	0.454545	0.96729	0.898785
7	[:MAX_2H]	0.333333	0.96729	0.882591
8	[:MAX_2H, :LENGTH]	0.333333	0.96729	0.882591
9	[:MAX_1H, :LENGTH]	0.363636	0.957944	0.878543
10	[:LENGTH, :MEAN, :PTS_1H]	0.212121	0.981308	0.878543
11	[:LENGTH, :MEAN]	0.181818	0.981308	0.874494
12	[:LENGTH, :MEAN_PREC]	0.151515	0.985981	0.874494
13	[:LENGTH]	0.0909091	0.990654	0.870445
14	[:MEAN]	0.0606061	0.990654	0.866397
15	[:MEAN_PREC]	0.030303	0.995327	0.866397
16	[:MEAN, :MEAN_PREC]	0.0909091	0.985981	0.866397
17	[:PTS_1H, :MEAN_PREC]	0.030303	0.995327	0.866397
18	[:MAX_1H]	0.212121	0.962617	0.862348

Figure 7: Modèles à plusieurs variable classés par précision décroissante pour l'ouvrage 4350-1D

	COMBINATION	SENSIBILITE	SPECIFICITE	PRECISION
	Array…	Float64	Float64	Float64
1	[:MAX_1H, :MEAN]	0.696429	0.973822	0.910931
2	[:MAX_2H]	0.642857	0.963351	0.890688
3	[:SUM, :MAX_1H]	0.642857	0.963351	0.890688
4	[:MAX_1H]	0.607143	0.963351	0.882591
5	[:SUM, :PTS_1H]	0.553571	0.979058	0.882591
6	[:SUM, :PTS_2H]	0.535714	0.979058	0.878543
7	[:SUM, :MEAN]	0.517857	0.963351	0.862348
8	[:SUM, :LENGTH]	0.5	0.958115	0.854251
9	[:SUM]	0.446429	0.95288	0.838057
10	[:LENGTH, :MEAN]	0.25	0.963351	0.801619
11	[:MEAN]	0.178571	0.979058	0.797571
12	[:MEAN, :PTS_1H]	0.178571	0.968586	0.789474
13	[:MEAN_PREC]	0.0178571	0.989529	0.769231
14	[:LENGTH]	0.0178571	0.984293	0.765182

Figure 8: Modèles à plusieurs variable classés par précision décroissante pour l'ouvrage 4340-3D

en utilisant toutes les variables à notre disposition et seulement la précipitation totale, nous gagnons en moyenne 1.9% de précision sur les 56 ouvrages.

3.3 Prévision de surverses

Un second aspect de ce projet est la prédiction de surverses en fonction des données pluviométriques. Pour ce faire, nous divisons les EP en deux groupes : un ensemble d'entraînement avec 80% des données et un ensemble de test avec les 20% restants. Nous nous assurons que les deux groupes aient environ la même proportion de surverses. L'ensemble d'entraînement est utilisé pour calculer les coefficients du modèle de régression, puis les surverses sont prédites sur l'ensemble de test. Nous partitionnons les données quelques milliers de fois pour avoir une prédiction moyenne.

La précision baisse de 2% pour la plupart des ouvrages, mais l'ordre des variables explicatives reste le même.

3.4 Comparaison des différents modèles

Pour nos 56 ouvrages comparons la précision des différents modèles.

- **Modèle naïf**. Nous considérons qu'aucun EP ne génère de surverse : précision de 79,6%.
- **Modèle avec précipitation totale**. La variable SUM est utilisée pour chacun des ouvrages : précision de 84,4% soit un gain de précision de **4,8%**.
- **Modèle avec intensité maximale**. La variable MAX_2H est utilisée pour chacun des ouvrages : précision de 83,5% soit un gain de précision de **3,9%**.
- **Modèle avec la meilleure variable explicative**. Pour chaque ouvrage, la variable donnant la meilleure précision est utilisée : précision de 85,3% soit un gain de précision de **5,7%**.
- **Modèle avec la meilleure combinaison de variables explicatives**. Pour chaque ouvrage, on choisit la combinaison de variables explicatives qui donne la meilleure précision : précision de 86,3% soit un gain de précision de **6,7%**.

Le gain de précision est important, si l'on change la variable explicative d'un ouvrage à l'autre.

4 Catégorisation d'événements pluviaux pour la prédiction de surverses

Il est intéressant de regrouper les EP par catégories, en fonction de leur risque à déclencher une surverse. En appliquant des modèles climatiques futurs il serait alors possible de voir comment évoluent ces catégories et prédire d'une autre façon l'évolution du nombre de surverses par ouvrage.

4.1 Catégorisation avec SUM et MAX_2H

Dans son rapport sur l'évolution des régimes de précipitation en climat futur pour la région de Montréal [2] Alain Mailhot définit 5 catégories pour les EP, en fonction de l'intensité maximale sur 2H et de la précipitation totale de l'EP. Les seuils ont été choisis grâce à l'expertise des ingénrieurs de la ville de Montréal

- **cat 1** : Faible-peu intense (SUM $<$ 10mm)
- **cat 2** : Modéré-peu intense (10mm $\leq$ SUM $\leq$ 20 et MAX_2H $<$ 5mm/h)
- **cat 3** : Modéré-intense (10mm $\leq$ SUM $\leq$ 20mm et MAX_2H $\geq$ 5mm/h)
- **cat 4** : Fort-peu intense (SUM $>$ 20mm et MAX_2H $<$ 5mm/h)
- **cat 5** : Fort-intense (SUM $>$ 20mm et MAX_2H $\geq$ 5mm/h)

Voici la répartition des EP dans ces catégories (bleu = EP ne générant pas de surverse, rouge = EP générant une surverse) pour l'ouvrage 4340-03D (figure 9) :

Ici la catégorisation des EP se fait indépendamment des ouvrages, et la proportion d'EP générant une surverse dans chaque catégorie peut varier d'un ouvrage à l'autre. Autrement dit, il se peut que certains ouvrages soient beaucoup plus sensibles à des événements faible-peu intenses (par exemple) et que la proportion de surverses soit alors très différente.

De plus, cette définition utilise seulement les deux variables explicatives **SUM** et **MAX_2H**, alors que nous avons une multitude d'autres à notre disposition.

Le but de la suite du projet est de définir de nouvelles catégories d'EP, dont les seuils et variables explicatives peuvent varier d'un ouvrage à l'autre. Nous

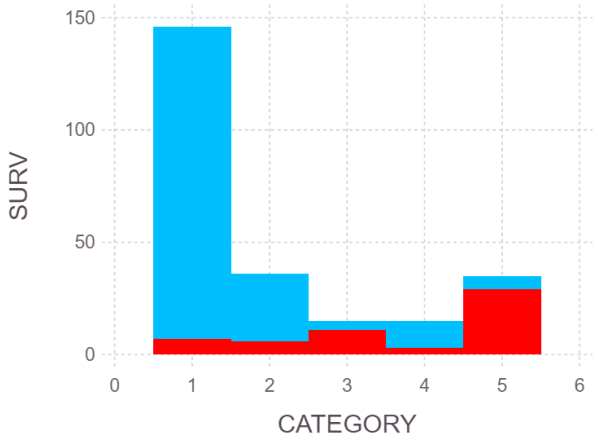

Figure 9: Catégorisation des EP avec les variables SUM et MAX_2H et des seuils fixes

comparerons l'efficacité d'une telle catégorisation par rapport à la catégorisation fixe.

4.2 Arbres de décision pour la définition de seuil

Nous définissons les seuils pour chaque catégorie de façon à minimiser le mélange d'EP générant une surverse ou non. Nous aimerions avoir les catégories 1 et 3 les plus "pures" possibles afin d'être sûrs qu'un EP rentrant dans une de ces catégories génère ou non une surverse. La catégorie 2 rassemble des EP dont nous ne pouvons pas facilement dire si elle génère une surverse. Dans un projet futur, il faudrait peut être analyser plus finement les EP de cette catégorie.

Les arbres de décision peuvent être utilisés pour prédire l'issu d'un EP, mais aussi de le catégoriser. En effet, en regarder les choix qui sont faits à chaque noeud de l'arbre, nous pouvons retrouver les règles de catégorisation de chaque feuille.

4.3 Construction d'un arbre et fonction d'entropie

Voici quelques explications sur les arbres de décision. Un exemple d'arbre peut être trouvé à la figure 10. A chaque étape de la construction de l'arbre, en partant de la racine (en haut) pour former des feuilles (en bas), nous séparons les EP en deux groupes (cette séparation se fait au niveau d'un noeud). Chacun de ces deux groupes est à nouveau séparé en deux, et ainsi de suite jusqu'à obtenir des feuilles. En-dessous de chaque feuille nous voyons le nombre d'EP groupés dans la feuille ainsi que le nombre de surverses générées (syntaxe : **Srv nb_surverses /**

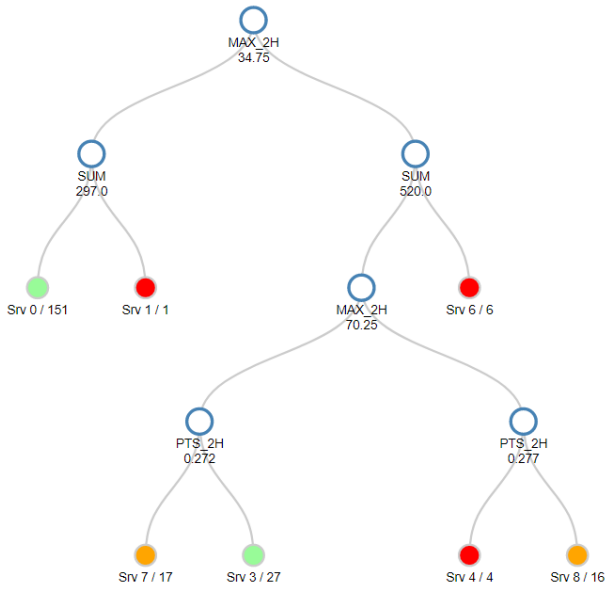

Figure 10: Exemple d'un arbre de décision créé avec l'ouvrage 4350-01D

nb_EP donc Srv 7/17 signifie qu'il y a 7 surverses pour 17 EP).

Les feuilles de cet arbre sont colorés pour visualiser facilement la proportion de surverses parmi tous les EP (vert = moins de 10% des EP génèrent une surverse, rouge = plus de 80% des EP génèrent une surverse, orange = mélange). Au niveau des noeuds nous voyons quel couple variables explicative / seuil est utilisé pour séparer les EP en deux groupes. La branche de gauche regroupe les EP qui ont cette caractéristique inférieure au seuil, et la branche de droite les EP supérieurs au seuil. Le choix du couple variable explicative / seuil est fait à chaque noeud en minimisant une fonction d'entropie, c'est à dire en minimisant le désordre au sein de chaque groupe.

Voici le détail de la fonction d'entropie. Pour tout couple (variable explicative / seuil) les EP sont divisés en deux groupes, gauche et droite. Dans chacun de ces deux groupes nous avons un certain nombre de surverses. Tout d'abord nous calculons l'entropie du groupe de gauche.

$$\begin{aligned} p_g &= \textit{proportion de surv. gauche} \\ &= \frac{\textit{nb surv. gauche}}{\textit{nb EP gauche}} \end{aligned} \quad (1)$$

$$Ent_g = -(p_g log(p_g) \ + \ (1-p_g) * log(1-p_g)) \quad (2)$$

De même nous calculons l'entropie de la branche de droite Ent_d. Finalement, nous calculons l'entropie moyenne résultante des deux branches :

$$Ent = \frac{nb\ EP\ g}{nb\ EP\ tot} * Ent_g \ + \ \frac{nb\ EP\ d}{nb\ EP\ tot} * Ent_d \quad (3)$$

Pour toutes les combinaisons possibles de couple (variable explicative / seuil) l'entropie est calculée. Le couple ayant la plus petite entropie est sélectionnée pour le noeud.

Nous décidons d'arrêter l'agrandissement de l'arbre après avoir atteint un certain niveau de profondeur. En traçant la valeur de la précision en fonction de la profondeur de l'arbre nous optons pour une profondeur qui donne la meilleure précision (ici 2). (La précision est obtenue avec un arbre entraîné sur 80% des données et une précision calculée sur les 20% restants).

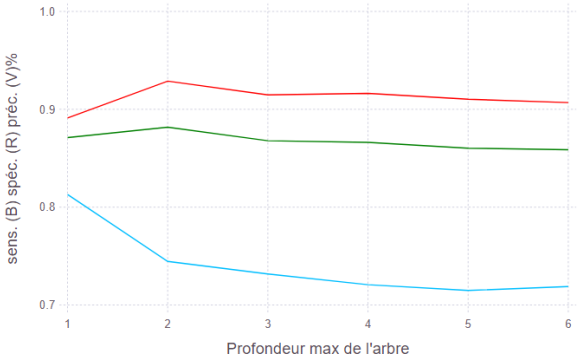

Figure 11: Recherche de la profondeur optimale de l'arbre de décision

4.4 Dans la pratique : recherche de seuils pour l'ouvrage 4340-03D

L'arbre de décision obtenu pour cet ouvrage est présenté dans la figure 12 et la catégorisation associée dans la figues 13

Les 3 variables explicatives utilisées sont MAX_2H, MEAN et MAX_1H.

La précision passe de 89% à 92% et la sensibilité de 71% à 77%, avec cette nouvelle catégorisation, par rapport à une catégorisation fixe.

Afin d'avoir une idée de la performance des arbres de décision sur les 56 ouvrages considérés, nous construisons un arbre avec une profondeur de 2 (fig. 14) et de 3 (fig. 15) puis nous tirons la précision, sensibilité et spécificité. Pour chaque catégorie (quelle que soit la définition considérée), nous labélisons l'EP avec le label majoritaire présent dans cette catégorie. Par exemple avec la figure 13 les EP des catégories 1

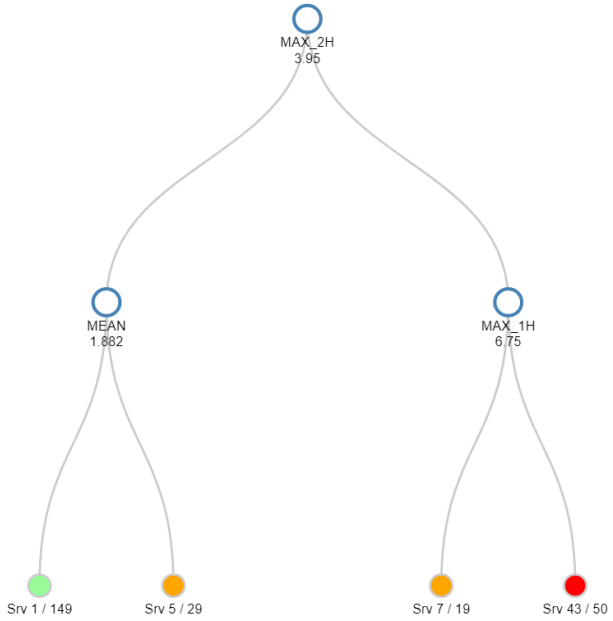

Figure 12: Arbre de décision obtenu pour l'ouvrage 4340-03D

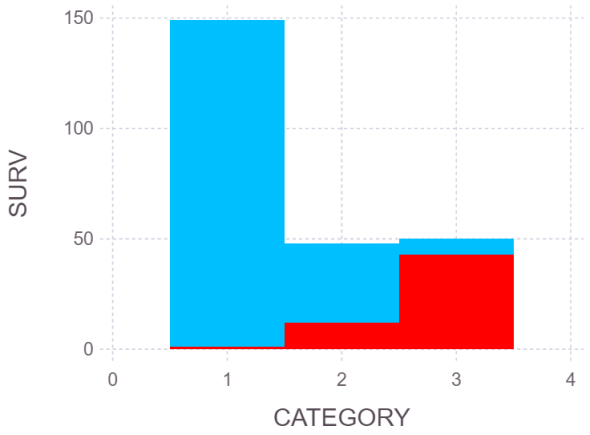

Figure 13: Catégorisation tirée de l'arbre de décision

et 2 sont marquées comme ne générant pas de surverse, puisque la majorité des EP qui la compose n'en génèrent pas, et la catégorie 3 labélisée comme générant tout le temps une surverse.

En fonction de la profondeur que l'on choisit, une catégorisation d'EP par ouvrage peut être meilleure que la catégorisation fixe. Il est risqué d'utiliser des arbres avec une profondeur plus grande que 3 car les règles risquent d'être ajustées trop spécifiquement à notre jeu de données, et la précision risque d'être mauvaise avec l'ajout d'une nouvelle année (2018 par exemple).

A noter également que nous comparons un modèle

```
----------COMPARAISON CATEGORISATION----------
Moyenne précision catégorisation arbre : 86.59
Moyenne précision catégorisation fixe : 86.7
Moyenne sensibilité catégorisation arbre : 36.93
Moyenne sensibilité catégorisation fixe : 46.15
Moyenne spécificité catégorisation arbre : 97.37
Moyenne spécificité catégorisation fixe : 94.14
```

Figure 14: Comparaison profondeur arbre profondeur 2

```
----------COMPARAISON CATEGORISATION----------
Moyenne précision catégorisation arbre : 88.3
Moyenne précision catégorisation fixe : 86.7
Moyenne sensibilité catégorisation arbre : 48.52
Moyenne sensibilité catégorisation fixe : 46.15
Moyenne spécificité catégorisation arbre : 97.78
Moyenne spécificité catégorisation fixe : 94.14
```

Figure 15: Comparaison profondeur arbre profondeur 3

à 5 catégories à un autre à 3 catégories. Ce dernier sera donc naturellement moins précis.

5 Limites

Les modèles n'ont été testés dans un premier temps que sur 56 des 170 ouvrages. Il faudrait voir comment évoluent la précision sur des ouvrages avec un très faible nombre de surverses (1 par an) et sur les ouvrages qui ont des données manquantes (83 ouvrages concernés).

Il n'y a que 3 stations météo qui sont utilisées pour représenter les précipitations au niveau de l'ouvrage. En ajoutant les données de stations aux alentour de Montréal et combinant les données, ils serait possible d'avoir une estimation plus fiable des précipitations.

Nos modèles sont basés sur la présence ou absence de surverses et non la quantité d'eau déversée dans le fleuve. Une étude plus poussée doit donc être effectuée pour le dimensionnement de bassin de rétention d'eaux usées par exemple.

6 Conclusion

Prédire l'occurrence d'une surverse d'un ouvrage en fonction des données pluviométriques est important puisque cela permet de voir l'évolution du nombre de surverses dans les années à venir. Différentes mesures peuvent être prises pour éviter les débordements, comme la construction de bassin de rétention ou la végétalisation de notre milieu urbain. Mais ces mesures restent chères et augmenter la précision de

la prédiction est essentiel. En adaptant le choix de variables explicatives par ouvrage on peut augmenter la précision de 8% au niveau de certains ouvrages. En moyenne la précision est augmentée de 1,9%.

Un deuxième aspect du projet était de tester une nouvelle forme de catégorisation d'EP, basée sur des arbres de décision. Cela permet d'utiliser toutes les variables disponibles et d'adapter les seuils pour chaque ouvrage. 3 catégories peuvent ainsi être créées pour chaque ouvrage. Avec des arbres d'une profondeur de 2 cette nouvelle catégorisation n'est pas meilleure qu'une catégorisation fixe. Mais en générant des arbres d'une profondeur de 3, nous augmentons 1,6% la précision.

Il serait intéressant de tester avec les données de l'année 2018 les deux aspects de ce projet. Une demande des données de surverses est en cours.

References

[1] B. Lavallée A. Mailhot, G. Talbot. Relationships between rainfall and combined sewer overflow (cso) occurrences. 2015.

[2] S. Bolduc A. Mailhot, G. Talbot. Évolution des régimes de précipitation en climat futur pour la région de montréal. 2019.

[3] A. Mailhot C. Fortier. Climate change impact on combined sewer overflows. 2014.