

Régression semi-supervisée pour l'exploration de réseaux d'assainissement

Master financé par l'IMU et encadré par Khalid Benabdeslem¹ et Frédéric Cherqui², au sein de l'Université Claude Bernard Lyon 1, France

Contexte : le projet Hireau - <https://hireau.org/>

Les réseaux d'assainissement et d'eau potable ont été construits et étendus pour et par la ville. Aujourd'hui comme demain, ce patrimoine existant impact les pratiques de gestion : de nombreuses études ont montré l'importance primordiale de la connaissance de la date de pose ou du matériau des conduites pour estimer leur état actuel de détérioration et prédire leur dégradation. L'objectif du stage de master est de développer et tester de premières méthodes d'apprentissage semi-supervisé pour retrouver les dates de pose des conduites à Lyon (qui sont comprises en 1854 et aujourd'hui) à partir des caractéristiques des tronçons : hauteur, largeur, forme, longueur, matériau et importance dans le réseau, cf. Figure 1.

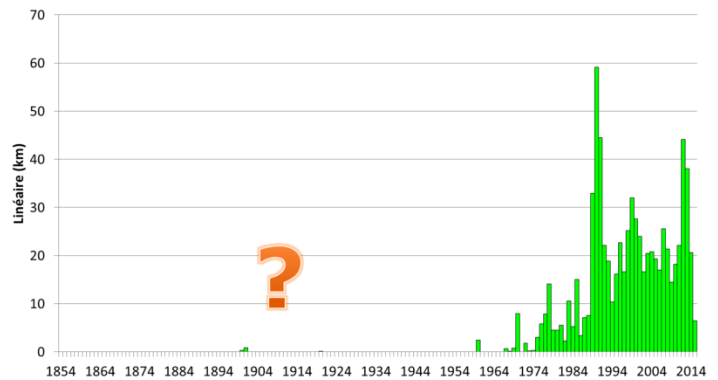


Figure 1 : 25% du linéaire des conduites d'assainissement possède une date de pose connue au démarrage du projet

Il s'agit principalement de :

- voir si les données fournies par la collectivité sont utilisables par les outils d'apprentissage semi-supervisé,
- tester plusieurs méthodes pour déterminer celles qui sont les plus adaptées au problème,
- identifier les améliorations nécessaires de la base de données pour améliorer la qualité de l'apprentissage.

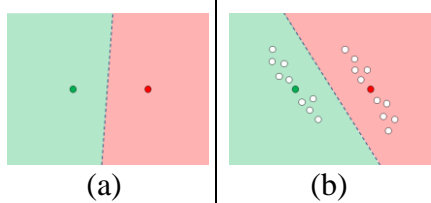


Figure 2 : apprentissage supervisé (a) et semi-supervisé (b) avec des points (données servant à l'apprentissage) et un trait pointillé qui représente le modèle obtenu. Les points blancs représentent les données pour lesquelles la cible n'est pas connue (master de Vivien Kraus)

Qu'est-ce l'apprentissage semi-supervisé ?

Pour comprendre le principe de l'apprentissage semi-supervisé, considérons une base de données partiellement étiquetée : pour certaines données, la cible (ici la date de pose) est connue et pour d'autres cette cible n'est pas connue. Dans le cas de l'apprentissage supervisé (fig. 2a) : seules les données avec une cible connue sont utilisées pour élaborer la fonction de prédiction. Pour l'apprentissage semi-supervisé (fig. 2b), les données non étiquetées sont également utilisées. Ces données sont prises en compte dans l'élaboration de la fonction de prédiction. Lorsque l'on rajoute l'information des données non étiquetées, l'apprentissage est différent, il peut être meilleur si certaines hypothèses sont respectées (régularité, partitionnement et variété).

Résultats

Trois méthodes d'apprentissage ont été testées : deux méthodes provenant de la littérature et une nouvelle méthode (combinant les deux précédentes) a été développée.

Les travaux ont permis de montrer que :

- les données de la collectivité peuvent être utilisables dans les outils permettant l'utilisation de méthodes d'apprentissage semi-supervisé,
- la nouvelle méthode semble prometteuse car elle réduit l'erreur sur la date prédite (erreurs quadratiques moyenne et relative),
- la base de données des conduites datées doit être complétée sur certaines périodes temporelles pour améliorer la qualité de la prédiction.

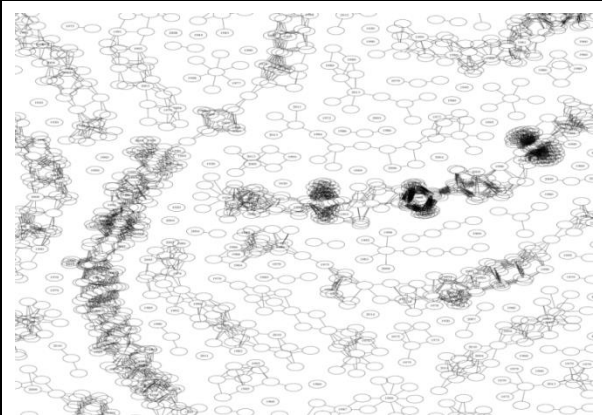


Figure 3 : graphe de canalisation obtenu (master de Vivien Kraus)

¹ Université Lyon 1 - 43 Bd du 11 Novembre 1918, 69622 Villeurbanne

² Université de Lyon, INSA-LYON, Université Claude Bernard Lyon 1, DEEP, F-69621, F-69622, Villeurbanne