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

Master financé par l'IMU et encadré par Khalid Benabdeslem[1] et Frédéric Cherqui[2], au sein de l'Université Claude Bernard Lyon 1, France

Contexte du 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'enjeu est donc pour la Métropole de Lyon de reconstituer les dates de pose des réseaux d'assainissement (24% du linéaire connu), et de fiabiliser les dates pour l'eau potable (97% du linéaire renseigné dont 27% supposé). Mais plus largement, l'enjeu est de démontrer que l'on peut reconstituer ces dates pour ainsi permettre à d'autres collectivités de répondre aux contraintes réglementaires et mettre en œuvre une réelle gestion patrimoniale. De plus, les ambitions du projet dépassent le cadre des réseaux d'assainissement et d'eau potable. HIREAU vise également à renseigner sur l'histoire de la ville et de son développement. C'est également un cas d'application pertinent de l'apprentissage semi-supervisé qui consiste à modéliser des fonctions de décision à partir de bases de données statistiques partiellement étiquetées (connues). Ce projet de 36 mois qui a émergé de la communauté IMU vise donc à faire progresser les connaissances via des interactions de compétences en Génie Civil, Géographie et Informatique, ainsi qu'avec les praticiens partenaires (Eau du Grand Lyon, la Métropole de Lyon et Veolia).

Objectif du travail de master

L'objectif du Master est donc d'identifier et étudier les difficultés de l'exploitation des données pour leur prétraitement optimal, afin de construire la base de travail, à partir de données issues du système d'information géographique (SIG) de la Métropole de Lyon. Cette étude doit également permettre de mesurer l'effet et l'impact des méthodes d'apprentissage statistique semi-supervisé, un champ disciplinaire qui consiste à modéliser des fonctions de décision à partir de base de données statistiques partiellement étiquetées. En effet, les bases sont construites à partir de différentes sources hétérogènes et après l'expertise qui ne pourra déterminer qu'une partie de la cible : les dates de poses à caractère continu. Concernant les données prises en compte, nous nous intéressons principalement à la vue des caractéristiques des tronçons (diamètre, matériau, forme, etc.) au détriment des aspects géographiques, car les caractéristiques ont une cohérence (certains matériaux sont surtout employés à une époque, par exemple), alors que des remplacements ponctuels pourraient remettre en question la continuité spatiale de la date de pose.

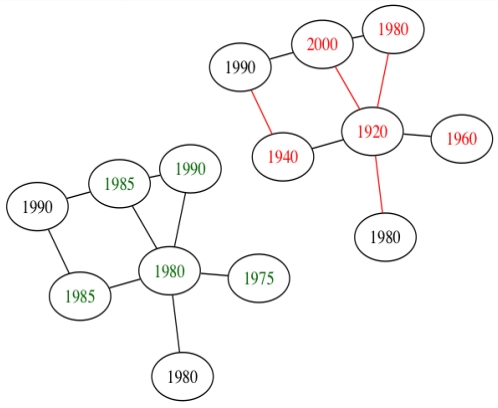

Figure 1 : l'objectif du travail de master est la prédiction des dates de pose des conduites d'assainissement à partir de méthodes d'apprentissage semi-supervisé

[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

Méthodes d'apprentissage semi-supervisé

Nous cherchons à prédire une information continue, la date de pose, en fonction des caractéristiques des tronçons. C'est une tâche de régression. Nous ne connaissons les dates de pose de quelques conduites (25 %), mais nous souhaitons utiliser les caractéristiques des conduites non datées. C'est un cadre dit semi-supervisé. L'apprentissage semi-supervisé est un paradigme de l'apprentissage automatique particulièrement intéressant, du fait qu'il tient compte de données labellisées ainsi que de données non labellisées[1].

Pour comprendre le principe de l'apprentissage semi-supervisé, considérons une base de données partiellement étiquetée : pour certaines données, la cible est connue et pour d'autres cette cible n'est pas connue. La Figure 2.a présente le cas de l'apprentissage supervisé : seules les données avec une cible connue sont utilisées pour élaborer la fonction de prédiction. Dans la Figure 2.b, les données non étiquetées sont également utilisées lors de l'élaboration de la fonction de prédiction. Dans le cas de l'apprentissage semi-supervisée, les données sont partitionnées et classées dans des groupes homogènes, et ces groupes sont pris en compte dans l'élaboration de la fonction de prédiction.

(a) (b)

Figure 2 : apprentissage supervisé (a) et semi-supervisé (b). Les points représentent les données servant à l'apprentissage : un point vert correspond à une cible connue (de valeur "vert"), un point rouge correspond à une cible connue (de valeur "rouge") et un point blanc correspond à une cible inconnue. Le trait bleu discontinu représente le modèle obtenu. La figure a) présente le cas de l'apprentissage supervisé et la figure b) présente le cas de l'apprentissage semi-supervisé.

Lorsque l'on rajoute l'information des données non étiquetées, l'apprentissage est différent. L'ajout de données non labellisées ne se traduit cependant pas nécessairement en une amélioration de l'apprentissage. Pour cela, il faut respecter les trois principales hypothèses de l'apprentissage semi-supervisé :

1. l'hypothèse de régularité (*smoothness* en anglais) : si des points sont proches les uns des autres, la cible de ces points doit être proche ;
2. l'hypothèse de partitionnement (*cluster* en anglais) : les zones à haute densité de points ont une valeur de la cible constante ;
3. l'hypothèse de variété (*manifold* en anglais) : les données se situent en réalité dans un espace de très faible dimension.

Plusieurs approches sont considérées dans le cas des dates de pose des réseaux d'assainissement de la Métropole :

Self-training et co-training

Le self-training consiste à apprendre un modèle à partir des canalisations datées, utiliser ce modèle pour prédire certaines dates manquantes, et recommencer l'apprentissage avec ces nouvelles conduites. Il s'agit de bien choisir les conduites à dater.

Le co-training consiste à séparer les données en vues indépendantes et apprendre un modèle pour chaque vue. Chaque modèle date certaines dates, qui deviendront des éléments d'apprentissage pour l'autre vue.

Adaptation de méthodes existantes

Des techniques existent pour adapter des méthodes supervisées en méthodes non-supervisées.

Par exemple, les SVM[2] (machines à vecteur support) permettent d'obtenir une fonction de décision qui maximise la marge entre les différentes classes. Par exemple, sur la Figure 2, la décision est une ligne qui sépare les deux classes, et la marge est l'écart entre la séparation et le point le plus proche. Le SVM permet de trouver une séparation qui soit le plus loin possible de toutes les canalisations, même celles qui ne sont pas datées.

Régularisations sur un graphe

On représente les conduites sur un graphe, où les conduites-nœuds sont reliées si elles se ressemblent suffisamment. Ensuite, parmi toutes les solutions de datation possibles, on choisit celles qui donnent des écarts de dates faibles entre conduites ressemblantes. Cette régularisation est implémentée dans beaucoup de méthodes de régression supervisée, comme la régression simple, les processus gaussiens, ou les SVM.

Méthodes retenues

Nous nous intéressons particulièrement à deux méthodes de régression semi-supervisées différentes. La première, *LapRLS*[3] pour *Laplacian-Regularized Least Squares*, est une méthode de propagation de labels très utilisée qui s'applique facilement à la régression en donnant. La seconde, *SSSL*[4] pour *Simple Semi-Supervised Learning*, utilise une étape non supervisée puis une étape supervisée, et permet de garantir un meilleur résultat que n'importe quelle méthode supervisée, si les données respectent certaines hypothèses.

Nous développons également une troisième méthode, nommée *LapS3L*, en utilisant l'algorithme *SSSL*, auquel nous rajoutons une régularisation laplacienne. L'algorithme procède en deux étapes :

1. Changement d'espace : on cherche une nouvelle représentation des conduites ;
2. Régression simple : on effectue une régression simple dans ce nouvel espace.

Dans l'algorithme original, le but de ces deux étapes est de traiter le problème semi-supervisé en deux sous-problèmes (non supervisé puis supervisé). En régularisant la seconde étape, nous rendons l'apprentissage meilleur.

Pré-traitement de la base

La base de conduites est très déséquilibrée : elle contient beaucoup de conduites récentes et relativement peu de conduites anciennes. Or, on peut savoir la répartition des dates de pose grâce aux bilans de pose par année. Ceci nous permet de faire l'hypothèse qu'il y a eu à peu près autant de conduites posées chaque année.

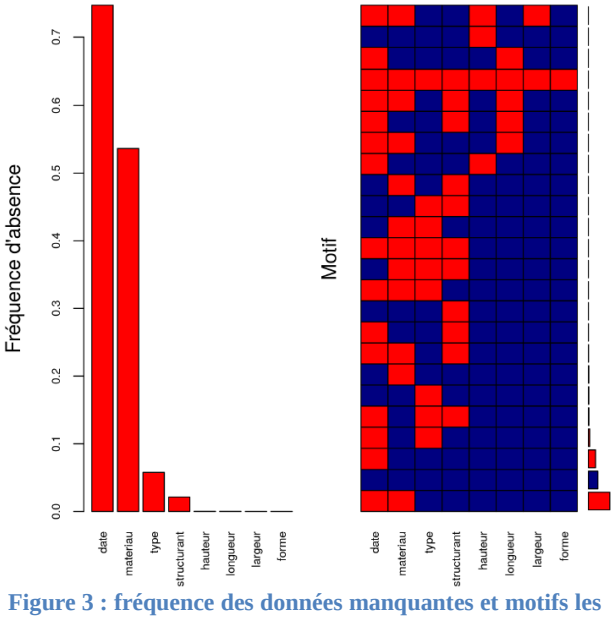

Figure 3 : fréquence des données manquantes et motifs les plus fréquents de données manquantes

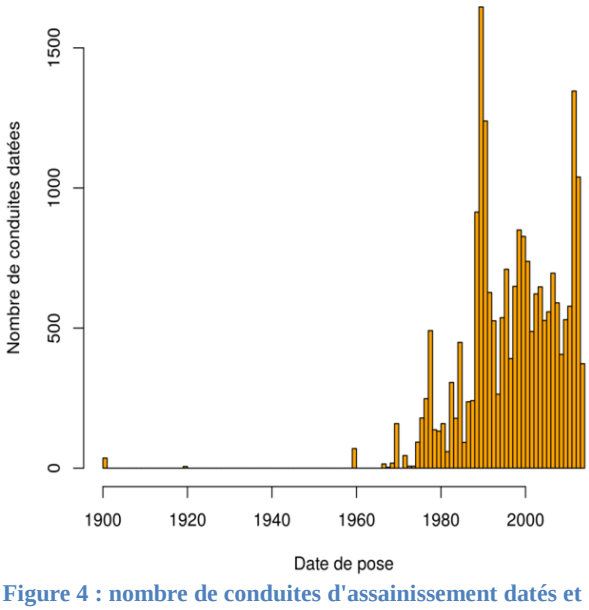

Figure 4 : nombre de conduites d'assainissement datés et date de pose (données Métropole de Lyon)

Pour rééquilibrer la base, on choisit donc des conduites uniformément selon leur date pour la partie supervisée, et un sous-ensemble quelconque pour la partie non supervisée.

Résultats

Le graphe des canalisations est représenté ci-dessous. On rappelle que le graphe se fonde uniquement sur les ressemblances entre conduites, et non pas sur les aspects géographiques.

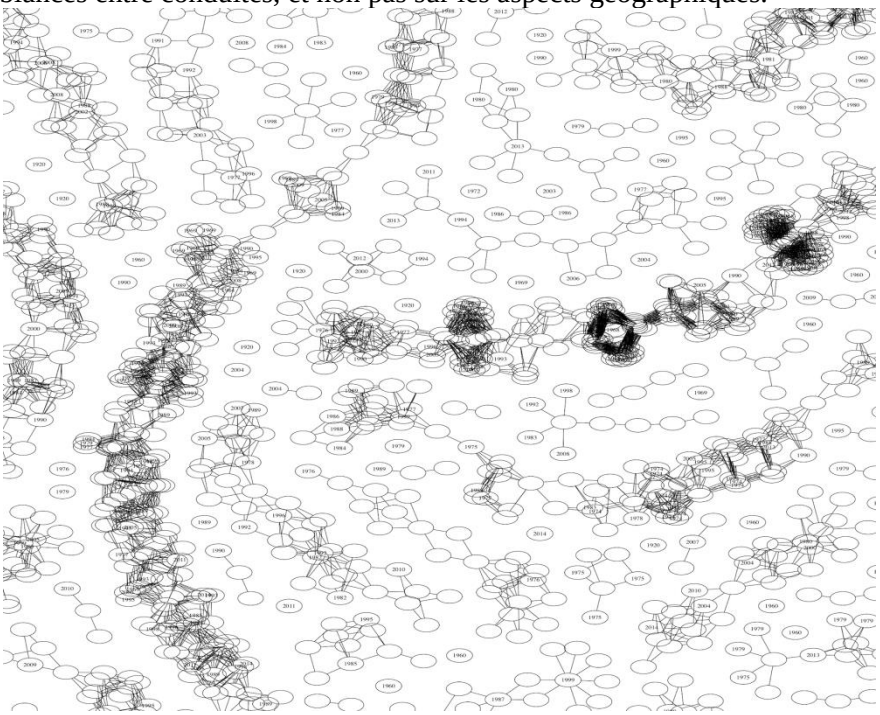

Figure 5 : graphe de canalisation obtenu

La régularisation introduite dans l'algorithme SSSL permet de considérer les conduites dans un plus grand espace (dont la dimension est le paramètre *s*). Nous obtenons une erreur de régression selon les quatre évaluations classiques : en valeur absolue MAE (erreur moyenne absolue) et RMSE (erreur quadratique moyenne), et en valeur relative RAE (erreur relative absolue) et RRSE (erreur quadratique relative). La Figure 6 présente ces résultats.

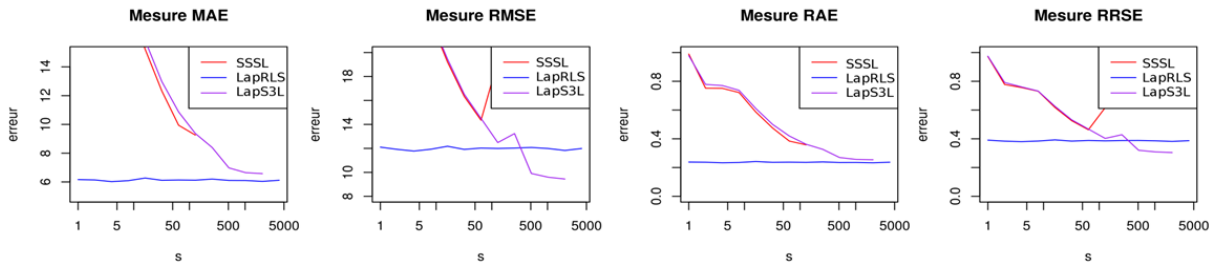

Figure 6 : Évolution de l'erreur en fonction du paramètre s, pour les méthodes *SSSL*, *LapRLS* (erreur constante, car s n'intervient pas) et la méthode proposée (*LapS3L*). Après un minimum d'erreur, la méthode SSSL produit une erreur de plus en plus grande qui se traduit par une erreur de calcul sur les flottants.

Références

[1] Chapelle, O., Scholkopf, B., Zien, A.: Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. IEEE Transactions on Neural Networks 20(3), 542–542 (2009).

[2] Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. pp. 144–152. COLT '92, ACM, New York, NY, USA (1992), http://doi.acm.org/10.1145/130385.130401.

[3] Belkin, M., Niyogi, P., Sindhwani, V.: On manifold regularization. In: AISTATS. p. 1 (2005), http://www.mit.edu/~9.520/Papers/Belkin-AISTATS-05.pdf

[4] Ji, M., Yang, T., Lin, B., Jin, R., Han, J.: A Simple Algorithm for Semi-supervised Learning with Improved Generalization Error Bound. arXiv:1206.6412 [cs, stat] (Jun 2012), http://arxiv.org/abs/1206.6412, arXiv: 1206.6412.