Régression Laplacienne semi-supervisée pour la reconstitution des dates de pose des réseaux d'assainissement

Vivien Kraus *, Khalid Benabdeslem *,
Frederic Cherqui **

*Université Lyon 1 - 43 Bd du 11 Novembre 1918, 69622 Villeurbanne
**Univ Lyon, INSA-LYON, Université Claude Bernard Lyon 1,
DEEP, F-69621, F-69622, Villeurbanne

Résumé. La date de pose est souvent un facteur principal d'explication de la dégradation des conduites d'assainissement. Pour les gestionnaires de ces réseaux, connaître cette information permet ainsi (par l'utilisation de modèles de détérioration) de prédire l'état de santé actuel des conduites non encore inspectées. Cette connaissance est primordiale pour prendre des décisions dans un contexte de forte contrainte budgétaire. L'objectif est ainsi de reconstituer ces dates de pose à partir des caractéristiques du patrimoine et de son environnement. Les données à manipuler présentent plusieurs niveaux de complexité importants. Leurs sources sont hétérogènes, leur volume est important et les informations sur leur étiquetage (dates) sont limitées : seulement 24 % du linéaire est connu pour les réseaux d'assainissement de la métropole de Lyon. La base de données sous-jacente contient les caractéristiques connues des conduites (profil géométrique, matériau utilisé, etc.). Dans ce papier, nous proposons de mesurer l'effet et l'impact de quelques méthodes d'apprentissage statistique semi-supervisé, et de proposer ainsi une approche alternative adaptée à la reconstitution de ce type de données.

1 Introduction

Depuis la prolifération des bases de données partiellement étiquetées, l'apprentissage automatique a connu un développement important dans le mode semi-supervisé [Chapelle et al. (2006)]. Cette tendance est due à la difficulté de l'étiquetage des données d'une part et au coût de cet étiquetage quand il est possible, d'autre part. L'apprentissage semi-supervisé est un cas particulier de l'apprentissage à partir de données faiblement étiquetées [Li et al. (2013)], qui consiste en général à modéliser une fonction statistique à partir de données regroupant à la fois des exemples étiquetés et d'autres non-étiquetés. Pour aborder une telle problématique, deux grandes familles d'approches existent : celle basée sur la propagation de la supervision en vue de l'apprentissage supervisé [Zhu (2006)] et celle basée sur la transformation de la partie étiquetée en contraintes en vue de leur intégration dans un processus de clustering (non-supervisé) [Basu et al. (2008)]. Nous nous intéressons ici à la première famille d'approches avec une difficulté particulière. Il s'agit d'apprendre avec une partie supervisée relativement