Projet 4IF MID-FD 2013/2014

« Fouille de données INSEE sur les communes du Rhône»

Jean-François Boulicaut – Mehdi Kaytoue

Objectifs de l'étude. Les données contenues dans le fichier « jeu69.txt » représentent une sélection dans une collection de données publiées par l'INSEE concernant les communes françaises. Cette sélection concerne des données statistiques pour les communes du département du Rhône collectées auprès de différentes sources : état civil, recensement, marché de l'emploi, DGI, DARES, CLAP, etc. Les variables renseignées portent sur cinq thèmes principaux : la population, le logement, les revenus, l'emploi, et les établissements.

Vous allez répondre à un appel d'offre du Grand Lyon qui souhaite sélectionner une équipe de spécialistes (disons un binôme) pouvant travailler à l'analyse et à la valorisation de telles données pour les besoins de ses collectivités territoriales. Pour concourir, il s'agit de mener à bien une première étude dans des délais très courts (2 semaines). Vous allez remettre un rapport qui s'appuiera sur des processus de fouille implémentés sous KNIME.

Ce rapport doit vous permettre d'être convaincant sur votre maîtrise méthodologique et technique de l'analyse et de la fouille de données. Vous partirez des données fournies mais vous pourrez aussi les compléter (avec, par exemple, la géolocalisation des communes, les pyramides des ages ou encore des données de recensement actualisées). Le Grand Lyon ne pose pas de questions précises mais s'intéresse à la découverte d'éléments de typologie des communes du Rhône. Il se peut que le choix d'un point de vue particulier, par exemple le positionnement au service d'une commune particulière, aide à dégager des questionnements pertinents : sur tel espace de paramètres socio-économiques, quels sont les communes qui ressemblent à une ou plusieurs communes choisies ?

Votre étude s'appuiera sur l'utilisation de méthodes et des outils d'exploration et de fouille de données que vous avez découvert ces dernières semaines lors des séances encadrées de Travaux Dirigés.

- Exploration des distributions, traitement des données exceptionnelles et des valeurs manquantes ;
- Sélection et dérivation d'attributs (descripteurs), normalisations, discrétisations et autres transformations ;
- Comparaison/choix des techniques de « clustering » à utiliser, choix du nombre de groupes à calculer, évaluation de la qualité des groupements (stabilité, qualité objective, qualité subjective) ;
- Interprétations des groupements, notamment au moyen de la construction de modèles interprétables comme des arbres de décision ;
- Prédiction au moyens des méthodes de classification supervisées (arbres de décision, règles, modèles de régression, etc).

Les données à traiter sont des données réelles et elles ne contiennent pas de groupes créés artificiellement. L'espace des communes (au regard d'une sélection d'un ensemble d'attributs) est sans doute assez « continu » et il est peu probable qu'il existe des groupes ayant des frontières très marquées. Pour autant, il y a peut-être des tendances dans la distribution des communes, ainsi que des communes pouvant être considérées comme « typiques » d'une zone de cet espace (ou d'une zone d'un sous-espace). Sur ce jeu de données, le même comportement se rencontrera sans doute en classification supervisée, où des classes ayant des contours flous ne pourront pas permettre d'obtenir de très faibles taux d'erreur

(dans ce cas 60 ou 70% de classement correct, dans une visée plutôt descriptive, peut alors rester raisonnable pour souligner des tendances). Noter que vous travaillez dans un contexte de fouille de données très exploratoire et qu'il est possible que même une utilisation experte des techniques ne fasse pas toujours apparaître de groupes. De telles conclusions sont en elles mêmes des résultats intéressants (homogénéité des distributions).

Compléments sur les données fournies. Le jeu de données est disponible dans l'espace pédagogique sur la baie. Il fournit, pour chaque commune du Rhône, la valeur d'un certain nombre de variables. La description des variables est dans le fichier « RSTA08_variables.pdf ». Des définitions et terminologies liées à ce jeu de données sont présentées dans « RSTA08_definitions.pdf ». L'information géographique est contenue dans les premières colonnes de la table. Il s'agit des champs : CODGEO, REG, DEP, ARR, CV, ZE1990 et EPCI. Dans le fichier « jeu69.txt », les champs sont séparés par une tabulation (TAB). Les champs vides donnent deux tabulations de suite. Dans KNIME, pour la lecture, choisir le séparateur TAB, cocher « read column header », ne pas cocher « read row IDs » (sinon la première des variables est utilisée comme identifiant des objets et non plus comme attribut) et ne pas cocher non plus « ignore spaces and tabs ». Vérifier que les types des variables ont bien été reconnus (« I » pour entier, « S » pour une chaîne, « D » pour un nombre en représentation flottante). Vérifier que les champs vides ont été identifiés comme des valeurs manquantes (valeur « ? »).

Compte rendu. Le travail s'effectue par binôme. Le compte-rendu du projet doit contenir les indications permettant de reproduire vos manipulations (hormis les valeurs par défaut des paramètres lorsque celles-ci sont utilisées), une vue des principaux « workflows KNIME » utilisés, et les éléments chiffrés et graphiques (exportations des graphiques et copies d'écrans, éventuellement agrémentées de toutes les informations « manuelles » qui peuvent faciliter l'interprétation). Ce rapport n'est pas destiné au représentant du Grand Lyon mais à vos enseignants qui devront bien comprendre votre questionnement (i.e., le ou les objectifs des traitements réalisés), les expérimentations effectuées et vos interprétations pour apprécier votre maitrise des méthodes et techniques de fouille de données (variété et pertinence des méthodes, validité des choix et de leur mises en œuvre, clarté des analyses et du compte rendu).

Signification des champs géographiques

CODGEO : Code Département (69) suivi d'un code commune
REG : Code région 82 Rhône-Alpes
DEP : Département 69 Rhône
ARR : Arrondissement 69 suivi d'un code d'arrondissement,
CV : Code Canton-Ville
EPCI : numéro d'établissement public de coopération intercommunale

Ce sont des regroupements de communes ayant pour objet l'élaboration de « projets communs de développement au sein de périmètres de solidarité ». Le numéro EPCI semble être égal à ZZZZZZZZZ dans la base lorsqu'il n'y a pas d'EPCI à cet endroit (mais apparaîtra dans KNIME comme une valeur manquante si on laisse à l'outil le soin de reconnaître seul le type de la colonne)

ZE1990 : Numéro de zone d'emploi (définition de 1990 ?).

Il représente l'espace géographique à l'intérieur duquel la plupart des actifs résident et travaillent. Par exemple pour Meyzieu on a : CODGEO=69282 ; REG=82 ; DEP=69 ; ARR=691 ; CV=6937 ; ZE1990=8211 ; EPCI=246900245.