Auteur : Didier LEPICAUT - Salarié Groupe EDF
Objet : Learning par étude de cas
Période : Centre de formation
Jury Académique : Serge GRATTON - Co-Directeur Ms Valdom
Intervenant : Paul PIDOU - Salarié SOPRA STERIA
Date : 08/12/18

Problème : Anonymiser des données de géolocalisation

CADRAGE	
date	11 décembre 2018
version	Pré-configuration
source des données	décembre 2018, France, fichier.csv
nombre de variables	2 variables quantitatives, lattitude et longitude
nombre d'enregistrements	1083
périmètre	NA
problème	Construire une typologie de points de coordonnées géographiques
type d'analyse	Méthode d'apprentissage statistique non supervisée
langage de programmation	Langage R
Note destinée au lecteur	Présentations des résultats clefs

- SOMMAIRE -

Résumé

(1) Motivation entreprise : un problème d'anonymiser les coordonnées GPS des points adresses des salariés observés [?],

(2) Nature scientifique du problème : dans la famille des méthodes d'apprentissage statistique non supervisées, identifier le Classifieur qui sait bien gérer les données quantitatives coordonnées GPS latitudes et longitudes,

(3) Résultats :

. **Apprentissage statistique** : pas de classifieur miracle mais les travaux empiriques des géomaticiens flèchent l'utilisation du classifieur DBSCAN,

. **Retour d'expérience** :

- Un problème « qui semble facile de prime abord » mais qui nécessitera des procédures itératives d'audit « a posteriori » pour vérifier que les regroupements obtenus de points sont bien homogènes à l'intérieur et biens hétérogènes entre eux,
- Mise en garde en revanche : avec du temps et des moyens, des experts en traitement des données arrivent à remonter jusqu'à l'individu observé.

(i.e. conclusions des travaux « Quantification de l'anonymat dans les bases de données » Ecole polytechnique de Louvain, Université catholique de Louvain, 2017)

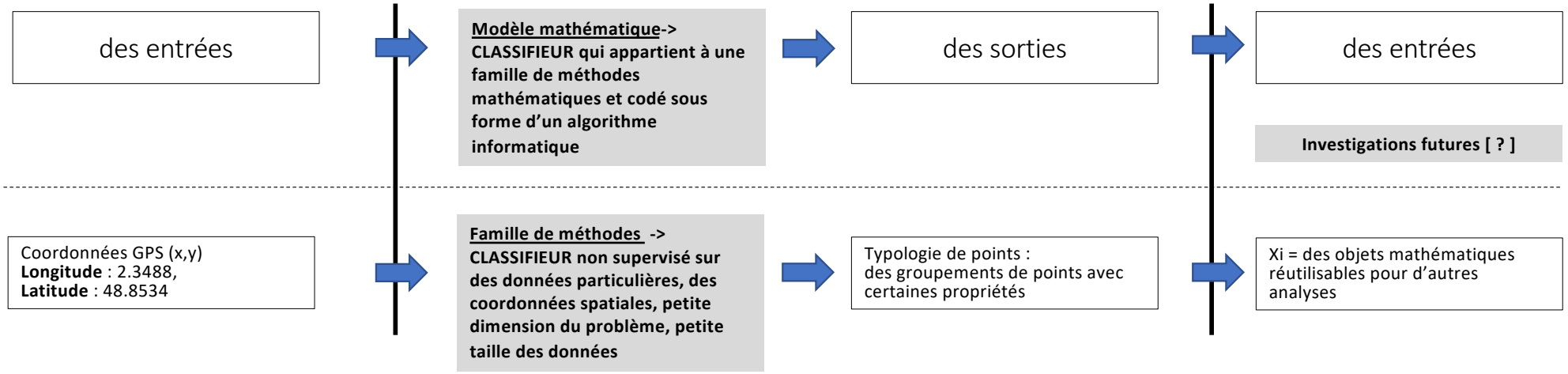
Le problème

Demande: *« … anonymiser des données de géolocalisation afin de respecter la vie privée tout en gardant l'information la plus granulaire possible. Ceci dans le but de ne pas gêner les investigations futures qui utiliseront ces données … »*

« vocabulaire entreprise »

analyse de la question

« vocabulaire scientifique »



Le modèle

Source ~ Philippe BESSE - Insa 31 – UE Apprentissage Statistique - Insa 2018>19

1. Famille de modèle d'analyse en apprentissage statistique

- **Famille** : méthode de classification non supervisée (ou clustering),
- **Données** : observations de p variables quantitatives sur ces n individus et se ramener au tableau des distances deux à deux entre les individus,
- **Objectifs** : répartir les individus en classes homogènes. Ceci est fait en optimisant un critère visant à regrouper les individus dans des classes, chacune la plus homogène possible et, entre elles, les plus distinctes possible.
- **Critère de mesure d'éloignement** : dissemblance, dissimilarité ou distance entre individus quantitatifs.

2. Modèle statistique retenu

Classifieur DBSCAN = Density-based spatial clustering of applications with noise

- **Classifieur récent** basé sur une estimation locale de la densité comme son acronyme le désigne (Ester et al. 1996) .
- **2 paramètres à régler :** nombre minimum de points et rayon d'une boule, il regroupe itérativement les points par paquet sur la base de leur voisinage (nombre minimum d'individus) à l'intérieur d'une boule de rayon ε.
- **Principe :** repose sur la notion de ε-voisinage d'un individu ou point défini comme l'ensemble des points appartenant à la boule de rayon ε centrée sur ce point. En plus du rayon ε, un autre paramètre est considéré : MinPts qui précise un nombre minimum de points à prendre en compte dans cette boule.

3. L'implémentation algorithmique

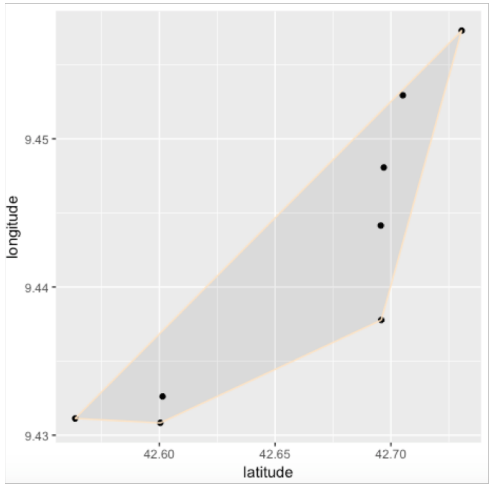
- **Initialisation** : algorithme démarre d'un point arbitrairement fixé qui n'a pas encore été visité. Si le ε-voisinage de ce point contient suffisamment de points, la construction d'une classe démarre. Sinon ce point est, au moins temporairement, étiqueté comme atypique mais peut être reconsidéré s'il apparaît par la suite dans le ε-voisinage du cœur d'une autre classe.
- **Parcours** : Si un point est un cœur, tous les points de son ε-voisinage sont affectés à sa classe puis chacun est considéré par l'algorithme. Si l'un de ces points est un cœur son ε-voisinage vient compléter la classe et le processus continue jusqu'à la complétion de la classe des points atteignables.
- **Arrêt** : l'algorithme considère un autre point pas encore visité pour renouveler le procédé jusqu'à ce que tous les points soient étiquetés.

3. Optimisation et critère de la qualité de la prédiction

- **Réglage des valeurs des paramètres ε et minPts** : opérer de façon experte à partir de la compréhension des données et de leur environnement.
  . Valeur minPts : apparaît comme une borne inférieure pour la taille des classes.
  . Valeur ε : il est conseillé (dbscan de R) de tracer le graphe des distances des k plus proches voisins (kNNdistplot) avec minPts pour valeur de k.
  La recherche d'un genou dans ce graphe est une indication pour le choix de ε.
- **Critère de la qualité de la prédiction** : pas de mesure quantitative car on est en « non supervise ».
  . cet algo est a utiliser quand tous les autres classifieurs, méthodes de classification non supervisées ne fonctionnent pas sur le jeu de data ?
  . demander aux métiers de vérifier les classes de points formées ?
- **Avantages :**
  . pas nécessaire de définir a priori le nombre de classes, celui-ci est une conséquence du choix des paramètres ε et minPts.
  . robuste aux observations atypiques et même propose de les détecter. Il ne classe pas les points isolés.
- **Inconvénients :** difficile de fixer les valeurs minPts et ε quand les données présentent un mélange de classes.

Les résultats

ième cluster de points
+ son POLYGONE simple

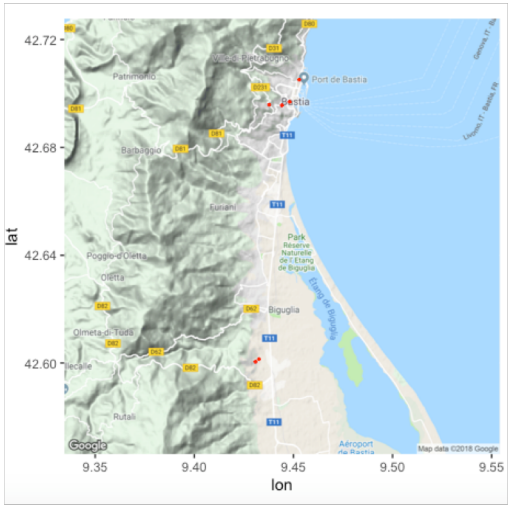


Objet statistique

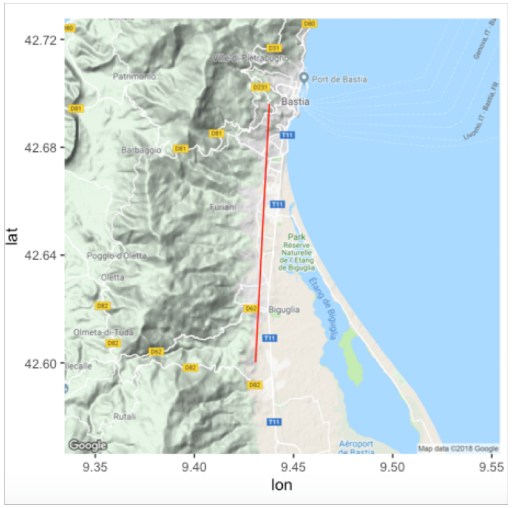
```
> summary(dvi)
    latitude        longitude
 Min.   :42.56   Min.   :9.431
 1st Qu.:42.60   1st Qu.:9.432
 Median :42.70   Median :9.441
 Mean   :42.66   Mean   :9.442
 3rd Qu.:42.70   3rd Qu.:9.449
 Max.   :42.73   Max.   :9.457
```

get_googlemap () accède à l'API Google Static Maps version 2

association cluster points,
fonds de carte

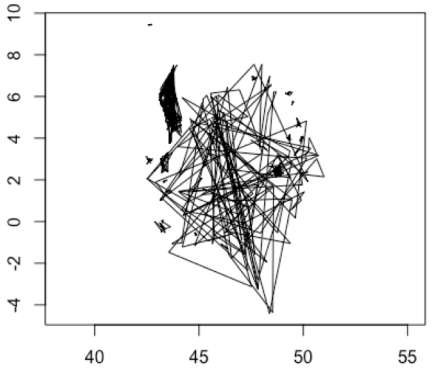


association cluster points,
polygone, fonds de carte

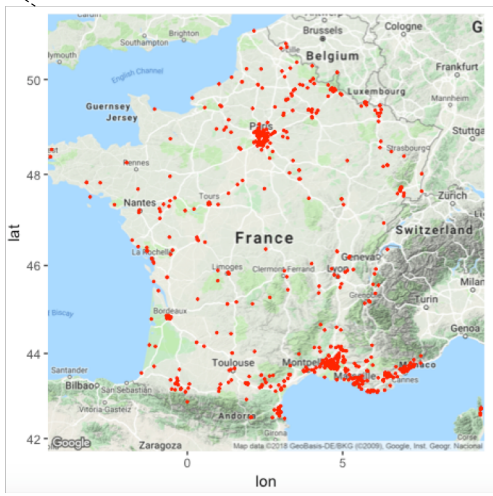


Echelon Géographie Locale

graphe du META-CLUSTER (cluster de clusters)
+ graphe du META-POLYGONE (poly de polys)

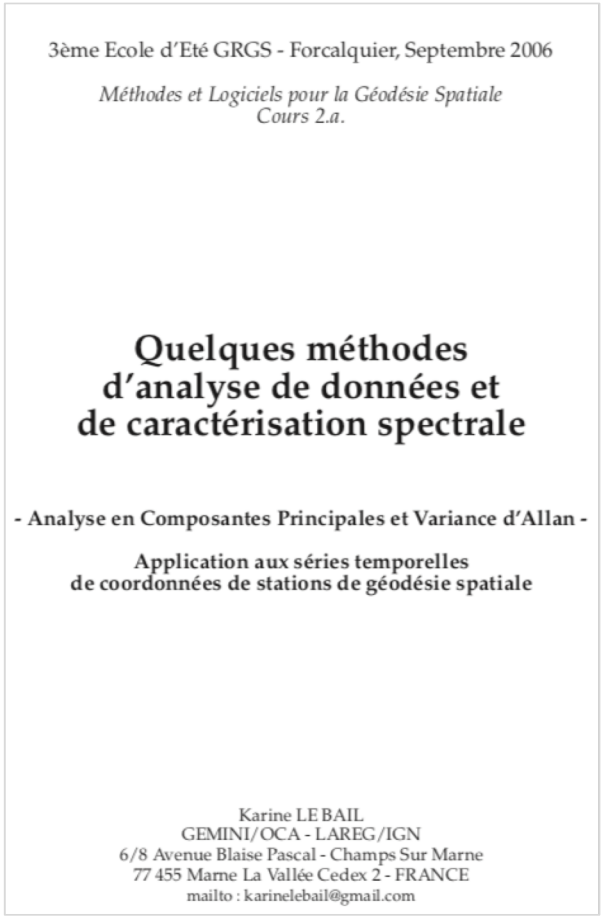


association des clusters de points,
fonds de carte (1083 enregistrements)



Echelon Géographie Nationale

Bibliographie



3ème Ecole d'Eté GRGS - Forcalquier, Septembre 2006

Méthodes et Logiciels pour la Géodésie Spatiale
Cours 2.a.

Quelques méthodes d'analyse de données et de caractérisation spectrale

- Analyse en Composantes Principales et Variance d'Allan -

Application aux séries temporelles de coordonnées de stations de géodésie spatiale

Karine LE BAIL
GEMINI/OCA - LAREG/IGN
6/8 Avenue Blaise Pascal - Champs Sur Marne
77 455 Marne La Vallée Cedex 2 - FRANCE
mailto : karinelebail@gmail.com

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

SEMIN-R

Utiliser des fichiers spatiaux dans des applications R

BAUDOIN, Raymond

Muséum national d'Histoire naturelle
Inventaire et suivi de la biodiversité

Annexes : Algorithme DBSCAN … pseudo code

https://fr.wikipedia.org/wiki/DBSCAN

```
DBSCAN(D, eps, MinPts)
   C = 0
   pour chaque point P non visité des données D
      marquer P comme visité
      PtsVoisins = epsilonVoisinage(D, P, eps)
      si tailleDe(PtsVoisins) < MinPts
         marquer P comme BRUIT
      sinon
         C++
         etendreCluster(D, P, PtsVoisins, C, eps,
MinPts)

etendreCluster(D, P, PtsVoisins, C, eps, MinPts)
   ajouter P au cluster C
   pour chaque point P' de PtsVoisins
      si P' n'a pas été visité
         marquer P' comme visité
         PtsVoisins' = epsilonVoisinage(D, P', eps)
         si tailleDe(PtsVoisins') >= MinPts
            PtsVoisins = PtsVoisins U PtsVoisins'
      si P' n'est membre d'aucun cluster
         ajouter P' au cluster C

epsilonVoisinage(D, P, eps)
   retourner tous les points de D qui sont à une
distance inférieure à epsilon de P
```

https://cran.r-project.org/web/packages/dbscan/dbscan.pdf

Usage

```
dbscan(x, eps, minPts = 5, weights = NULL, borderPoints = TRUE, ...)

## S3 method for class 'dbscan_fast'
predict(object, newdata = NULL, data, ...)
```

Référence

didier le picaut

Data Scientist

Diplômé :

. Ms Data Sciences Insa de Rouen

. Msc International HR Sorbonne Business School

. Ecole Nationale des Ponts et Chaussées

. Dess Marketing, Iae d'Orléans

. Mst économétrie, UFR S.Eco d'Orléans