TP : Clustering de stations Velib

1 Objectif

L'objectif de ce TP est de regrouper en classes homogènes les stations du réseau Velib parisien et d'interpréter les classes obtenues. **La principale caractéristique utilisée pour cela est le taux de vélos disponibles dans chaque station (taux de charge), disponible à une fréquence horaire**. On se focalisera sur la semaine du lundi 1er septembre 2014 au dimanche 7 septembre 2014. Plus spécifiquement, nous utiliserons une version lissée des données du package `funFEM` [1].

2 Données

Le jeu de données utilisé, dénommé `Velib` (téléchargement via `http://allousame.free.fr/m2mlds`), est caractérisé par 1189 stations décrites par les 171 variables suivantes (1 variable indiquant le nom des stations, 2 variables indiquant leur localisation géographique et 168 variables indiquant le taux de vélos disponibles à chaque heure de la semaine considérée :

- `station` : nom de la station,
- `longitude` : longitude,
- `latitude` : latitude,
- `lun00` : taux de vélos disponibles (ou taux de charge) le lundi 1er septembre 2014 à 0h : une valeur proche de 1 indique que la station est pleine, tandis qu'une valeur proche de 0 signifie que la station est vide,

  ⋮
- `dim23` : taux de vélos disponibles le dimanche 7 septembre 2014 à 23h.

La figure 1 montre la localisation géographique des stations ainsi que les taux de charge associés à cinq stations.

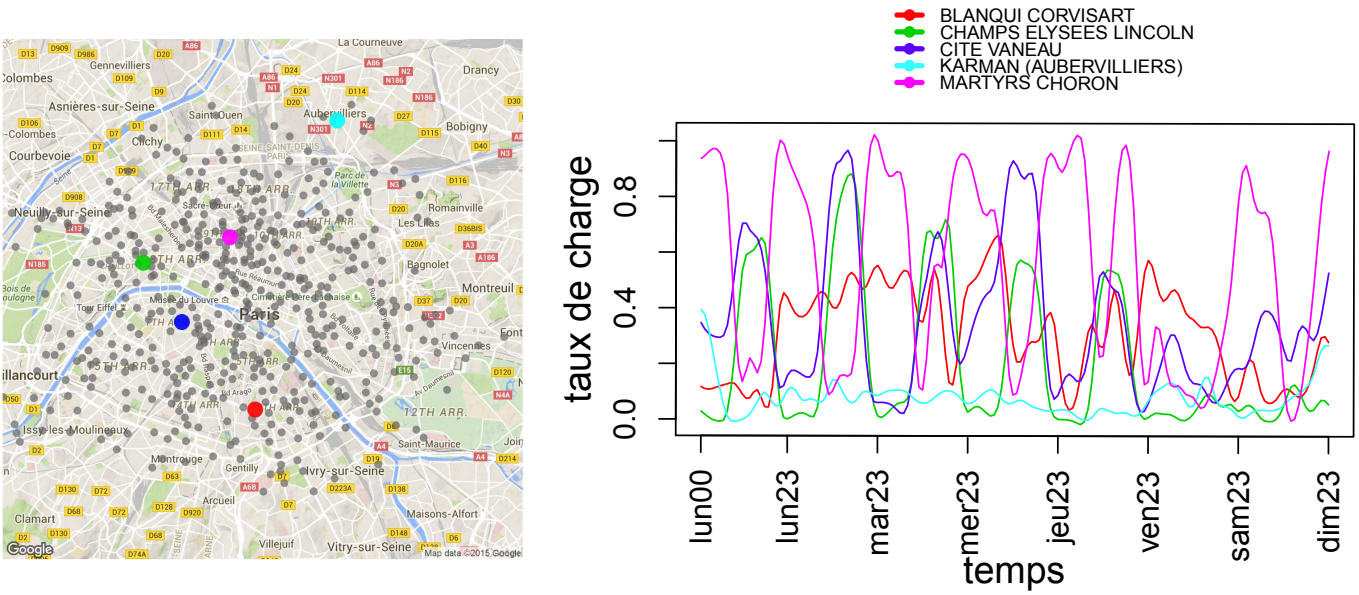

FIGURE 1 – (Gauche) Position des stations velib ; (Droite) Taux de charge de cinq stations observé au cours de la semaine (du lundi 1er septembre 2014 à 00h jusqu'au dimanche 7 septembre 2014 à 23h)

3 Travail à effectuer

1) Créer la matrice `data`, constituée uniquement des taux de charge. Il s'agit ici d'une matrice à 1189 lignes et 168 colonnes, sur laquelle les différents algorithmes de clustering seront appliqués.

2) Représenter graphiquement quelques courbes de charge. Observez-vous des différences sur celles-ci ?

3) Effectuez une ACP sur le tableau de données `data` et retenir les scores qui expliquent 95% de la variance des données. Cette étape doit notamment permettre de réduire le nombre de variables à manipuler, qui est initialement de 168.

4) A partir de ces nouvelles variables, effectuer un clustering en 6 classes des stations Vélib en utilisant les algorithmes CAH-Ward et k-means.

5) Pour chacune des méthodes, représenter graphiquement (sous la forme de courbes) les centres des clusters obtenus. Commenter les profils obtenus.

6) Représenter graphiquement les partitions obtenues en utilisant les coordonnées géographiques des stations. Proposer une interprétation des clusters en fonction de leur localisation spatiale.
Remarque : pour la représentation graphique sous forme de carte, vous pourrez télécharger le package `ggmap` *et l'utiliser de la manière suivante :*

```
position=c(longitude_min, latitude_min, longitude_max, latitude_max)
mymap=get_map(location=position, source="google", maptype="roadmap")
gm=ggmap(mymap)
gm+geom_point(aes(longitude,latitude),
data = data.frame(longitude=data_longitude,latitude=data_latitude),
alpha = .5, color="red", size = 4)
```

References

[1] C. Bouveyron and J. Jacques. *funFEM : an R package for functional data clustering*, March 2015.