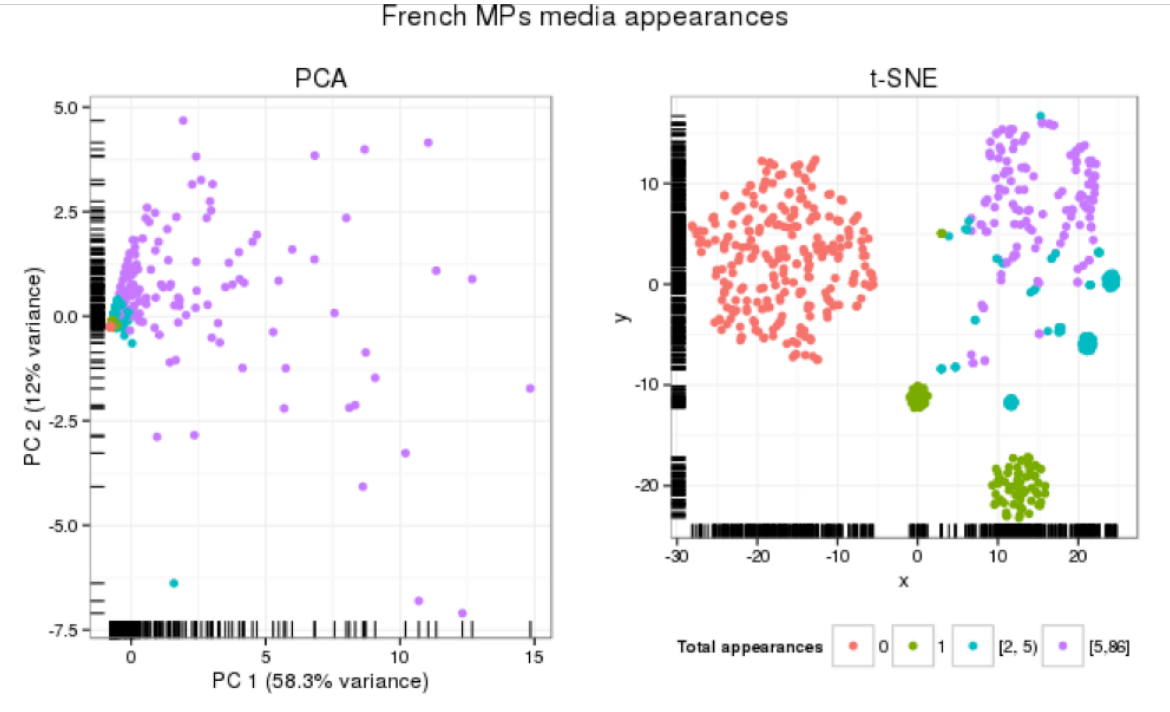
APPRENTISSAGE STATISTIQUE POUR LES SCIENCES SOCIALES

Julien Boelaert (julien.boelaert@gmail.com), Étienne Ollion (ollion@unistra.fr)

Formation SAGE 2016



Comparaison d'une analyse de correspondances et *machine learning* sur des données d'invitations médias des députés français[1]

Nées à la frontière des sciences informatiques, des statistiques et de l'ingénierie, les méthodes d'apprentissage statistique ont connu des développements rapides depuis le début des années 1980. Sous les dénominations alternatives ou connexes de *machine learning*, intelligence artificielle, ou *data mining*, elles désignent un ensemble de méthodes de traitement de données numériques. Leurs applications sont nombreuses, dans les domaines techniques, commerciales et scientifiques.

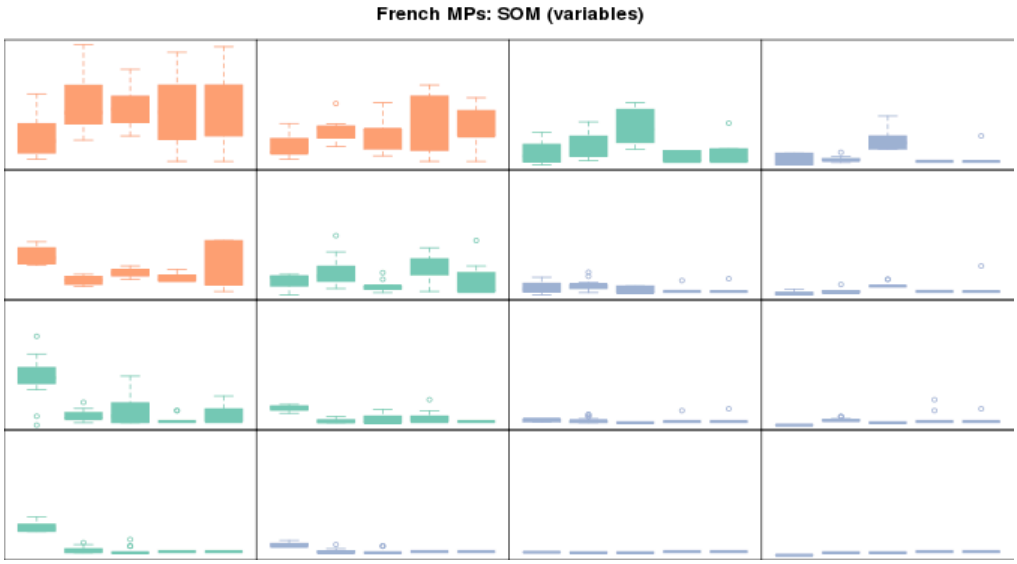
Que ce soit pour la description ou pour l'inférence, elles offrent en particulier de puissantes alternatives aux méthodes statistiques plus classiques, en particulier parce qu'elles proposent des modèles flexibles et parcimonieux en hypothèses. Si elles ont connu un grand succès en ingénierie et dans les sciences naturelles, elles sont encore peu exploitées en sciences sociales, alors même que leur flexibilité semble les y prédisposer.

L'objectif de cette formation est de proposer une introduction au fonctionnement et à l'utilisation de quelques unes de ces méthodes, en prenant le parti de la comparaison systématique avec les outils statistiques couramment utilisés en sciences sociales (analyse géométrique de données, régression linéaire, classifications). L'accent sera mis sur les grandes lignes théoriques, l'utilisation concrète des algorithmes, et leur articulation à des questionnements de sciences sociales. Chaque séance constituera donc une réflexion en action de l'intérêt et des usages possibles de ces méthodes en sciences sociales.

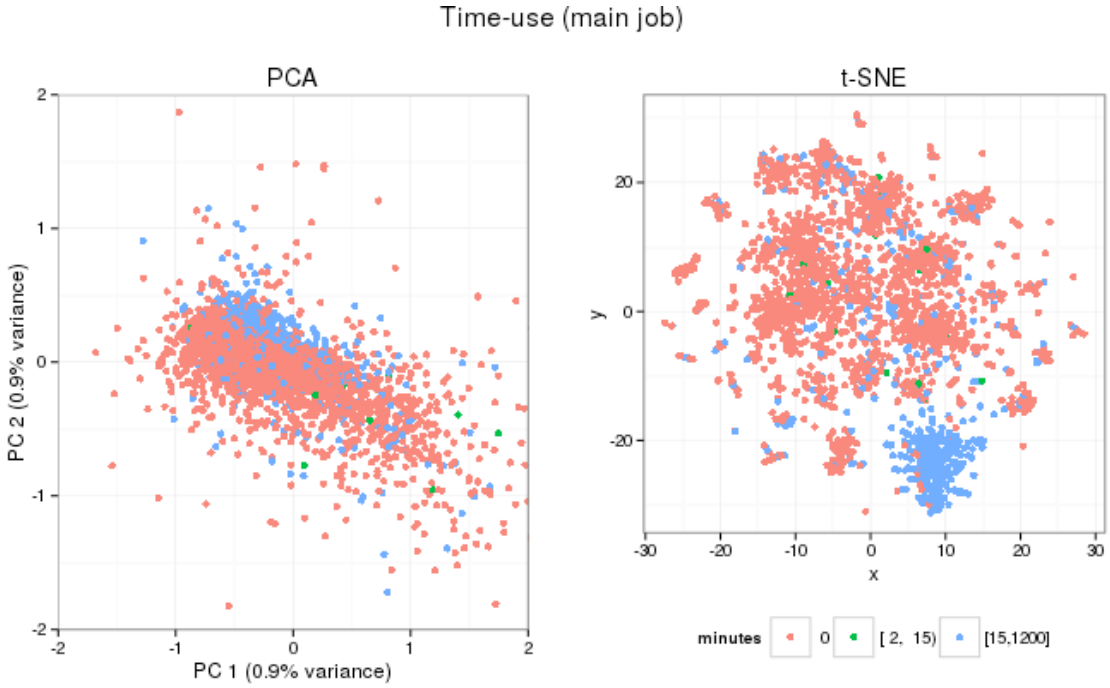
Pré-requis : une connaissance minimale en statistique (notions relatives à l'analyse géométrique de données, à la régression). Toutes les manipulations seront faites sous R, dont une courte séance introductive rappellera les rudiments.

[1] Données : (Delpierre, 2015).

Date	Programme
Vendredi 5.02 9h30-17h (Océanie)	*Clustering* : classification hiérarchique, k-means, cartes auto-organisatrices (SOM)
Vendredi 26.02 9h30-17h (Asie)	*Réduction de dimensionnalité et visualisation* : ACP/ACM, t-SNE, cartes auto-organisatrices
Vendredi 29.04 9h30-17h (Asie)	*Apprentissage supervisé* : Introduction générale (+ introduction à R), régression linéaire et extensions, arbres de régression, forêts aléatoires.



Cartes auto-organisatrices
Données: invitations médias des députés français



Comparaison PCA / t-SNE
Données: *American Time Use Survey* (2014)