

L'objectif de ce séminaire est d'échanger autour du logiciel de statistique libre, gratuit et multiplateforme *R* (<http://www.r-project.org/>). Il s'adresse aux praticiens impliqués dans le traitement quantitatif des données en sciences humaines et sociales (utilisateurs de données, chercheurs, ingénieurs ou étudiants) qu'ils aient ou non déjà utilisé le logiciel *R*.

Chaque séance est organisée autour de la présentation d'une expérience de traitement de données avec le logiciel (fonction spécifique et/ou packages). Le cadre de ces réunions est informel et les participants doivent se sentir libres d'intervenir afin de confronter leurs expériences.

Élisabeth Morand et Bénédicte Garnier (Ined),
Timothée Giraud (CNRS UMS Riata), Pascal Cristofoli (EHESS)
<https://enseignements-2016.ehess.fr/2016/ue/1010//>

Séance 3 : jeudi 23 Mars 2017 de 9h30 à 12h

Les arbres qui cachent les forêts ?

Arbres de régression et forêts aléatoires comme alternatives aux modèles de régressions standards en sciences sociales

Nicolas Robette
(CREST-LSQ ENSAE, Université Paris Saclay)

Parmi les innombrables méthodes d'apprentissage automatique, les arbres de décision se sont imposés depuis les années 1980 parmi les principaux outils pour résoudre les problèmes de classification et de régression. Ils ont depuis été perfectionnés et dépassés, avec notamment les algorithmes ensemblistes (bagging, forêts aléatoires, etc.). Dans le contexte des sciences sociales, cette boîte à outils semble à même de fournir une alternative crédible aux modèles de régression standards : ces algorithmes ne reposent pas sur des hypothèses contraignantes concernant les données (normalité, absence de multicollinéarité, etc.) et prend d'emblée et simplement en compte les interactions entre variables candidates à l'explication. De plus, leur mise en œuvre est maintenant largement facilitée par l'existence de packages *R* spécifiques

Prochaine séance le 18/5/2017