Programme des séances 2016-2017

Séance 1 : **24 novembre 2016** – Adrien Comminges *"Flux de travail intégré et conception d'outils pour la recherche et l'enseignement avec R et shiny"*

L'utilisation du langage-logiciel R pour la modélisation de l'information géographique présente plusieurs avantages majeurs vis-à-vis des outils historiques de l'analyse de données (SAS) et des SIG (ArcGis). Cette présentation fera le point de ces avantages pour construire des chaînes de traitement robustes et pour fluidifier le travail en équipe. Concernant le premier point, R permet de construire une même chaîne de traitements impliquant l'acquisition, le nettoyage, la mise en forme, la modélisation et la présentation des résultats, et cela en manipulant des objets de différents types (tableaux, objets spatiaux, graphes). Concernant le second point, R facilite le travail d'équipe grâce à deux intégrations logicielles : R + langages à balises (markdown, latex) pour faire du dynamic reporting ; R + HTML/CSV pour concevoir des plateformes de calcul.

Séance 2 : **26 janvier 2017** – Arnaud Bringé (Ined) *"Traitement de données historiques avec R"*

La présentation sera effectuée à partir de la juxtaposition de plusieurs sources de données historiques du 18ème siècle. Elle a pour cadre la ville de Martigues, victime de la dernière épidémie de peste en France (1720). Les données proviennent de listes nominatives issues de recensements fiscaux et de registres paroissiaux (Baptêmes-Mariages-Sépultures).

Ce type de sources est notamment caractérisé par la présence de nombreuses données textuelles, qui permettent notamment d'identifier les individus et la construction de généalogies. Ces données textuelles existent aussi très fréquemment pour caractériser des lieux (naissance, mariage, décès, origine) ou des professions. En préalable à tout traitement ou à tout regroupement, elles nécessitent d'être harmonisées. Nous montrerons dans un premier temps, quelles fonctions R utiliser afin d'homogénéiser au maximum ces données textuelles. Nous décrirons dans cette première partie l'utilisation des packages stringr pour le traitement des chaines de caractères et stringdist pour le calcul de distances entre chaînes.

La juxtaposition de plusieurs sources nécessite une homogénéisation des informations, tant au niveau des variables que des observations considérées. Nous décrirons dans cette deuxième partie l'utilisation du package sqldf. Enfin, l'analyse de ces sources nominatives a conduit au calcul de statistiques à un niveau agrégé (famille, maison). Nous décrirons dans cette dernière partie l'utilisation des packages plyr et dplyr.

Séance 3 : **23 mars 2017** Nicolas Robette (CREST-LSQ ENSAE, Université Paris Saclay)

"Les arbres qui cachent les forêts ? Arbres de régression et forêts aléatoires comme alternatives aux modèles de régressions standards en sciences sociales"

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 multicolinéarité, 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

Séance 4 : **18 mai** 2017 François Briatte (Université Catholique de Lille,École européenne des sciences politiques et sociales (ESPOL))

"Web Scraping et APIs avec R"

Cette présentation fournit un aperçu des différentes méthodes de collecte de données numériques via R, soit au moyen des techniques de Web scraping (collecte de données à partir de pages HTML), soit au moyen d'APIs (interfaces de programmation applicative) avec lesquels R est capable de communiquer.