Séminaire RUSS

R à l'usage des sciences sociales

à l'**EHESS** bâtiment Le France
190-198 avenue de France, 75013 Paris
salle du conseil B, R-1

http://russ.site.ined.fr/fr/

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 Riate), Pascal Cristofoli (EHESS)

https://enseignements-2016.ehess.fr/2016/ue/1010//

A l'Usage Des Sciences Sociales

Séance 2 : **jeudi 26 janvier 2017 de 9h30 à 12h**

Traitement de données historiques avec R

Arnaud Bringé (Ined – service méthodes statistiques)

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**.

Nombre de places limité à 30 - **inscription indispensable à**
http://www.ined.fr/fr/actualites/rencontres-scientifiques/seminaires-colloques-ined/russ-janvier-2017/

Prochaines séances les 23/3 et 18/5