Traitement de données historiques avec R

Arnaud Bringé

26 janvier 2017

Présentation du projet

Problématique

- Etude historique (Martigues XVIII^{ème} siècle)
- Projet piloté par I. Séguy (Ined, UR11)
- Etude de la mortalité (Epidémies)

Références :

- Séguy Isabelle, Bernigaud Nicolas, Bringé Arnaud, Signoli Michel, Tzortzis Stéfan. 2012. « A Geographic Information System for the Study of Past Epidemics: The 1705 Epidemic in Martigues (Bouches-du-Rhône, France) », Canadian Studies in Population, 39 (3-4), p. 107-122
- Séguy, Isabelle, Davide De Franco, Stephan Tzortzis, Arnaud Bringé, et Nicolas Bernigaud. 2016. « Measuring urban vulnerabilities in the early 18th century (Martigues, South of France) ». European Social Science History Conference (ESSHC).

Eléments géographiques

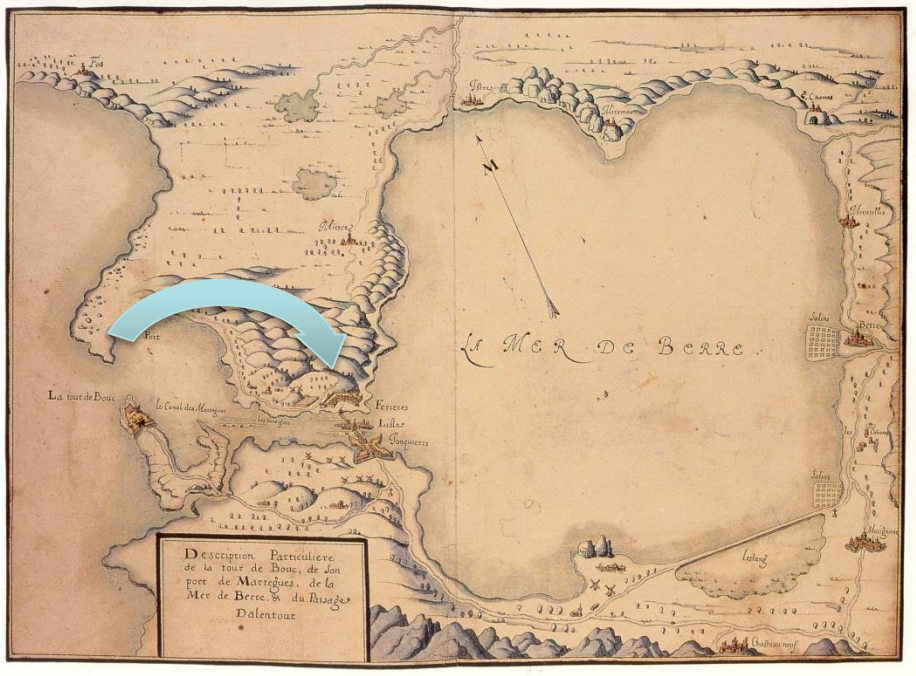
- Martigues, Venise provençale
- Communauté de 6000 habitants
- 3 paroisses
  - Ferrières, L'île, Jonquières

La ville de Martigues

Particularités historiques

- Sa topographie
- Son organisation urbaine
- La segmentation fortes en groupes sociologiques localisés.

Vue aérienne de Jonquières en 2009 (photo Georges Xuereb).



Croisement de sources

Analyse de sources historiques

- Recensements fiscaux (1701-1702-1716)
- Localisation des habitations (Matrices cadastrales 1716)
- Fichiers BMS (1702-1725)
  - Naissances : Infos sur Enfants + Parents + Parrains
  - Mariages : Infos sur Epoux + Parents + Témoins
  - Décès : Infos sur Ego + Conjoint éventuel + Parents

Présentation des données

Recensement fiscal (1702)

- Recensement exhaustif des membres du ménage
- Recensement par maison
- Identité (Nom-Prénom), Age, Profession
- Localisation par rue

Présentation des données

Fichiers BMS (1702-1725)

- Identifiants des époux, profession, âge au mariage
- Identifiant du baptisé, des parents, date de naissance
- Identifiant du défunt, profession, date de décès

Présentation des données

Fichier cadastral (1716)

- Identification du propriétaire
- Localisation relative des habitations (WNES)

Des problématiques multiples

- Cohérence des informations textuelles (lieux, noms, prénoms)
- Cohérence des dates mentionnées
- Traitement des valeurs manquantes
- Absences de codifications (professions, lieux …)

Cohérence des informations textuelles

Quelques exemples

- Orthographes différentes (Prénoms, noms, lieux)
- Homonymies
- Changements de noms (rues, quartiers)
- Professions peu informatives
- Problématiques en rapport avec la temporalité

Les chaînes de caractères sous R

Plusieurs solutions :

- grep
- Fonctions de base pour traitements de chaînes
- Le package stringi
- Le package stringr

Utilisation de grep

- Idée : Faire repérer diverses formes textuelles
- Utilisation :
  - Paramètre value=T pour affichage des valeurs, sinon indices
  - Paramètre ignore.case=T pour ne pas différencier la casse des caractères.

- Motifs à repérer : expressions régulières
- Syntaxe PERL :
  - Groupes de caractères (groupes, types de caractères …)
  - Quantificateurs (Nombre de caractères)

- Exemples de groupes :
  - \w : groupe de lettres
  - \d : groupe de chiffres
  - . : n'importe quel caractère
  - (k|\d{2}) : la lettre k, ou deux chiffres

Utilisation de grep

Les expressions régulières :

- ^ : Début de chaîne
- $: Fin de
- . : tout caractère
- | : disjonction (ou)
- [] : ensemble de caractères à égaler
- [^]: ensemble de caractères à ne pas égaler
- \\ : Recherche de caractères spéciaux (?, ^, $ …)
- * : Expression éventuellement présente
- + : Expression présente une ou plusieurs fois
- ? : Expression présente au plus une fois

Exemples

- grep(prenoms, "^A") : Prénoms qui commencent par A
- grep(prenoms, "[ai]ne$") : Prénoms qui se terminent par "ine" ou "ane"
- grep(prenoms, "au") : Prénoms qui contiennent la chaîne "au"

Autres fonctions

- gsub (Remplacement de chaine)

  Syntaxe : gsub(regex, remplacement, vecteur)

Quelques fonctions de base

- nchar
- paste
- substring(chaine, start, stop)
- strsplit

Le package stringr

Détection de sous-chaines

- str_count(vecteur,regex)
- str_locate(vecteur, regex)
- str_locate_all(vecteur, regex)
- str_detect(vecteur, regex)

Traitements des mots

- str_count(words, boundary("word"))
- str_split(words, " ")[[1]]
- str_split_fixed(words, " ")[[1]]

Le package stringr

Remplacement de chaînes

- str_replace(vecteur, regex, remplacement)
- str_replace_all(vecteur, regex, remplacement)
- str_extract(vecteur, regex)
- str_extract_all("The Cat in the Hat", "[a-z]+")
- str_match(vecteur, regex)
- str_match_all("The Cat in the Hat", "[a-z]+")

Conversions

- str_conv(x, "ISO-8859-1")
- str_to_upper(dog)
- str_to_lower(dog)
- str_to_title(dog)

Le package stringr

Extraction de sous-chaines

- str_subset(vecteur,regex)

Trier les éléments

- str_order(vecteur)
- str_sort(vecteur)

Utilitaires

- str_length(vecteur)
- str_trim(string, side = c("both", "left", "right"))
- str_c(str1,s tr2, sep) : Concaténation de chaines)

Le package stringr

Références

- Hadley Wickham, "Stringr: modern, consistent string processing". 2010. *R Journal*, vol. 2, no. 2, pp. 38–40.

Comparaison de chaînes

Exemple de problématique

- On dispose de fichiers nominatifs à deux dates (1701 et 1702), et on souhaite pouvoir raccorder les informations

Problèmes

- Orthographe
- Migrations

Comparaison de chaînes

Algorithme

- On va utiliser des techniques algorithmiques permettant de mesurer une distance entre 2 chaînes de caractères S1 et S2.

- La distance va être une mesure du nombre d'étapes minimal pour passer de S1 à S2, en utilisant uniquement les opérations élémentaires suivantes :
  - Insertion de caractère dans S1 ou S2
  - Suppression de caractère dans S1 ou S2
  - Substitution d'un caractère entre S1 et S2

Package stringdist

Fonctionnalités utilisées

Références

- Van der Loo, Mark PJ. 2014. « The stringdist package for approximate string matching ». *The R Journal*.
- Winkler, W.E.: String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. Proceedings of the Section on Survey Research Methods, American Statistical Association (1990)

Package stringdist

Algorithmes disponibles

Method name	Description
lv	Levenshtein distance. Compte le nombre d'insertions, suppressions et substitutions entre b et a
osa	Optimal String Aligment. Comme Lehvenshtein, mais permet la transposition des caractères adjacents.
Dl	Full Damerau-Levenshtein distance.
Hamming	Hamming distance. Compte le nombre de substitutions entre b et a (Chaînes de même longueur)
lcs	Longest common substring distance.
qgram	*q*-gram distance.
cosine	cosine distance between *q*-gram profiles
jaccard	Jaccard distance between *q*-gram profiles
jw	Jaro, or Jaro-Winker distance.
soundex	Distance based on soundex encoding

Package stringdist

Algorithmes disponibles

- Edits

Damerau-Levenshtein, Hamming, Levenshtein, optimal sting alignment

- qgrams

q- gram, cosine, jaccard distance

- heuristic metrics

Jaro, Jaro-Winkler

Application

Objectif :

- Trouver les propriétaires de 1716 dont le nom est "le plus proche" du propriétaire en 1702 (A Paroisse fixée)

- Il faut trouver un critère de comparaison pour les distances
- Il faut prendre en compte les multiples possibilités disponibles à une "distance minimale".

Package sqldf

Exemples de fonctionnalités

- Permet l'utilisation d'instructions SQL dans le traitement d'objets R
- Sélection d'observations, de variables (SELECT – FROM – WHERE)
- Fusion de données (INNER JOIN, OUTER JOIN)
- Agrégation de données (GROUP BY, HAVING)
- Sous-requêtes
- Résultat sous la forme d'objets R

Package plyr

Fonctionnalités utilisées

- cf exemples
- Analogie avec sqldf

Références

- Wickham, Hadley. 2011. « The split-apply-combine strategy for data analysis ». *Journal of Statistical Software* 40 (1). http://www.jstatsoft.org/v40/i01/paper

Package plyr

Opérations sur les datas frames

- arrange(df,cle1,cle2)
  - Trie un df selon les différentes clés
  - Utiliser l'option desc si tri selon ordre descendant
  - Attention l'ordre des lignes n'est pas respecté

Opérations sur les colonnes

- colwise : Application d'une fonction à toutes les colonnes
- numcolwise : Calcul sur toutes les colonnes

Package dplyr

- Package développé par Hadley Wickham (ggplot2)
  - Classe de données : tibbles
  - Définition d'une grammaire pour la gestion de données
  - Utilisation de pipes %>%

- Opérations élémentaires :
  - select() : Sélection de colonnes d'un jeu de données
  - filter() : Sélection de lignes en fonction de conditions logiques
  - summarise : Agrégation de données selon une clé
  - mutate : Construction de variables
  - arrange() : Tri des observations selon une ou plusieurs clés
  - group_by() : Opération sur des sous-groupes du jeu de données.

Package dplyr

Références

- Doc en ligne : https://cran.rstudio.com/web/packages/dplyr/
- Site R for Data Science : http://r4ds.had.co.nz/transform.html
- Feuille de présentation :
  - https://github.com/rstudio/cheatsheets/raw/master/source/pdfs/data-transformation-cheatsheet.pdf