IGR204 - M1 Project Description

Groupe M - TélécomParisTech, Mastère BGD 05/11/2018

1 Names of all project Members

Group M : Marine Ferrary - François Lecerf - Jérome Divac - Amaël Chaigneau - Christophe Thibault

2 Topic to be addressed, the datasets to use, where the data come from, and their formats

Evolution of first names in France between 1900 and 2016. We got two different databases : the first one is based on first names in France, and the second on first names for each department in France. We use the databases (état civil) from this page : https://www.insee.fr/fr/statistiques/2540004#consulter, and the data are composed by different variables : SEXE, PREUSUEL, ANNAIS, NUMBER.
Format [1] of the variables is :

- SEXE : character - 1 for male, and 2 for female
- PREUSUEL : character - first name (max:25 characters)
- ANNAIS : character - birth year (length : 4 characters)
- NUMBER: numeric - frequency (max : 8)

The first database contains 605 463 observations and the second one 3 520 641 with one new variable : DPT (department in France - three characters).

3 Description of data

- txt file, Fichiers France (métropole et départements d'outre-mer): 2Mo
- txt file, Fichiers par départements de naissance: 12 Mo

4 Description of the problem

Representation and observation of first names in function of year, and department over more than 100 years in France. Rank and original views.

[1] https://www.insee.fr/fr/statistiques/2540004#consulter