Visualization of massive data

Loïc PIREZ, Maxime BETEMPS

Problématique : Comment la situation démographique en Russie évolue-t-elle dans le temps ?

1 - Dataset

Description du dataset :

Ce dataset contient les informations démographiques de régions de la Russie tel que le taux de natalité, le taux de mortalité, etc.

Les différentes colonnes du dataset sont :

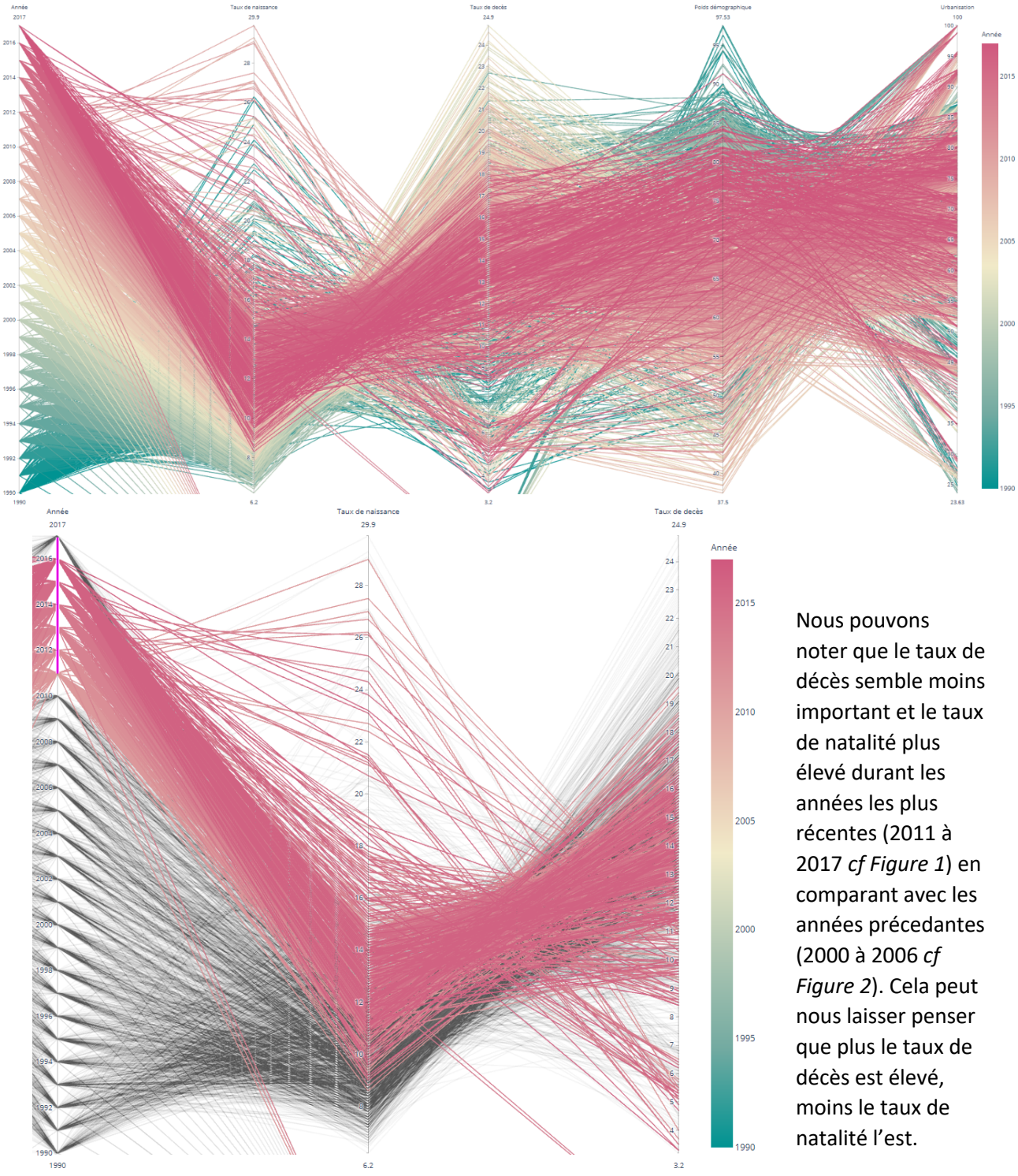
- ID : Identifiant
- Year : Année à laquelle la donnée a été
- Region : Nom de la région concernée
- Npg : Variation naturelle
- Birth_rate : Nombre de naissances sur 1000 personnes
- Death_rate : Nombre de décès sur 1000 personnes
- Gdw : « General Demographic Weight », nombre de personnes ne pouvant pas travailler pour 100 personnes pouvant travailler
- Urbanization : Pourcentage de la population vivant en milieu urbain

Le dataset original utilisé est disponible à l'adresse suivante :
https://www.kaggle.com/dwdkills/russian-demography

2 - Visualisations

A - Parallel coordinates

Nous avons décidé de représenter les données avec la méthode *Parallel coordinates* afin d'identifier des corrélations potentielles entre les données.



Nous pouvons noter que le taux de décès semble moins important et le taux de natalité plus élevé durant les années les plus récentes (2011 à 2017 *cf Figure 1*) en comparant avec les années précédantes (2000 à 2006 *cf Figure 2*). Cela peut nous laisser penser que plus le taux de décès est élevé, moins le taux de natalité l'est.

Figure 1

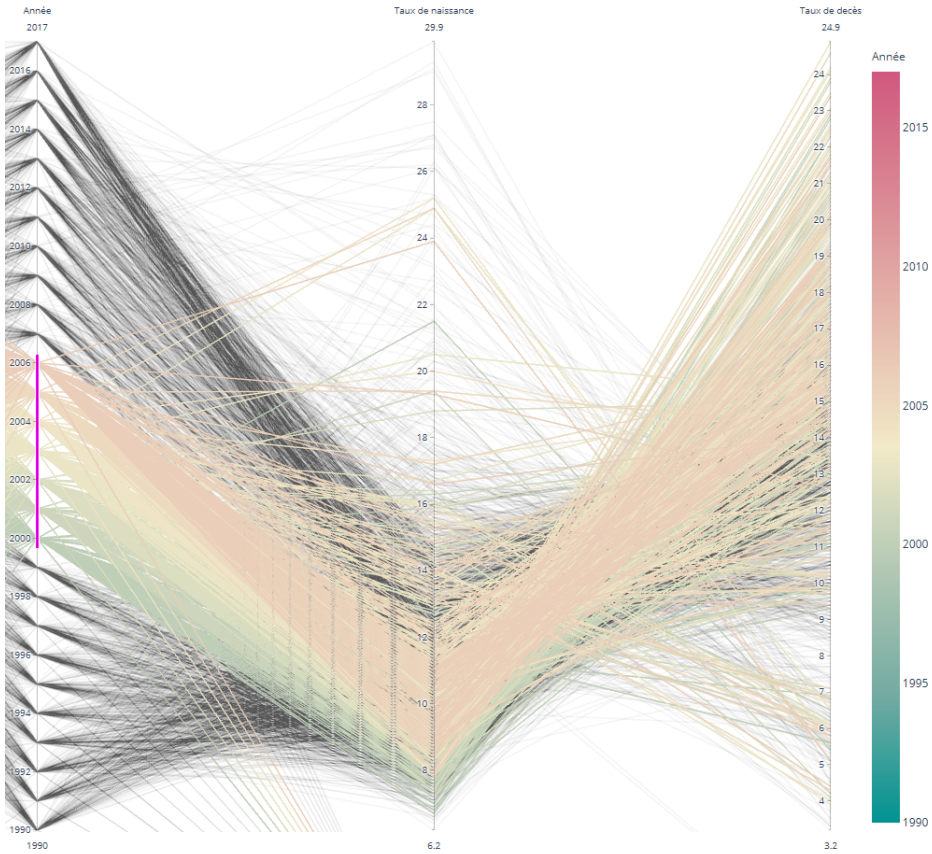

Année
2017
2015
2014
2012
2010
2008
2006
2004
2002
2000
1998
1996
1994
1992
1990
1990
Taux de naissance
29.9
28
26
24
22
20
18
16
14
12
10
8
6.2
Taux de decès
24.9
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3.2
Année
2015
2010
2005
2000
1995
1990

Figure 2

B - Categorical scatterplots

Sur la seconde visualisation, nous avons utilisé la méthode *Categorical scatterplots*.

Nous avons souhaité observer la variation naturelle dans les différentes régions de Russie. Le graphique ci-dessous représente celle-ci pour toutes les années (1990 à 2017) (cf *Figure 3*).

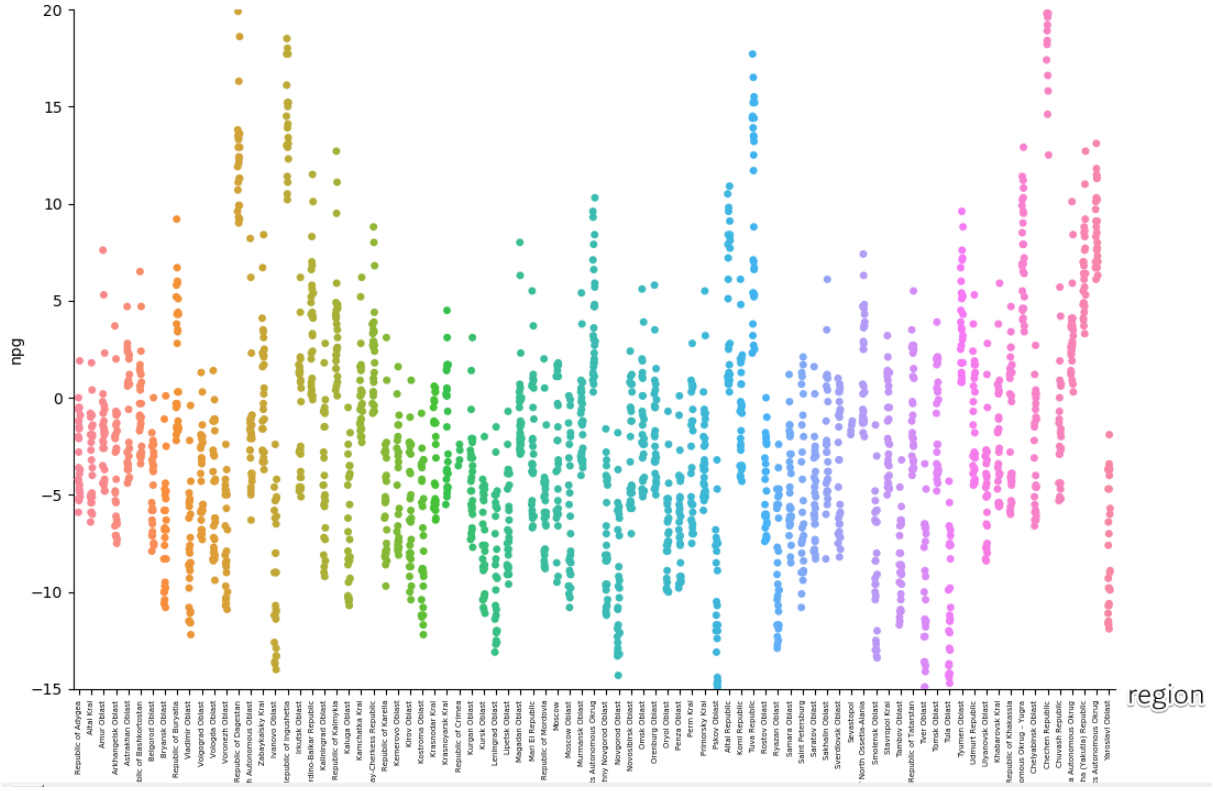


Figure 3

Nous souhaitons identifier l'évolution de la variation naturelle au cours des années, donc nous avons généré un graphique pour chaque année séparemment (voir dans le dossier render du projet pour observer chaque année).

Figure 4

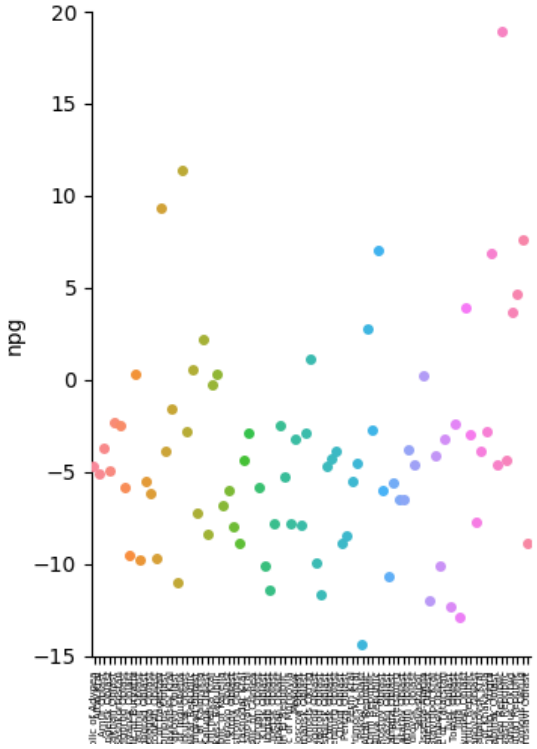
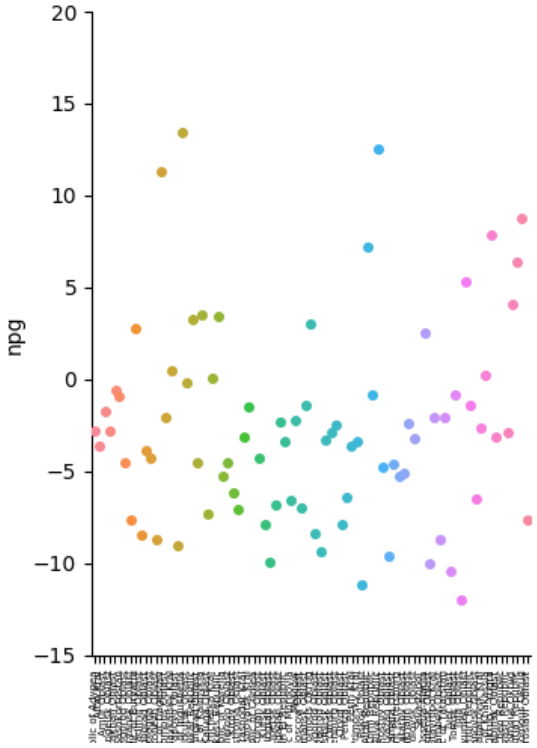


Figure 5



En 2006 (*cf Figure 4)* nous pouvons constater une grande augmention du taux de naissance en Russie sur l'ensemble des régions comparé à l'année 2007. (*cf Figure 5)*. En effet, en 2007, selon un article de The Guardian (https://www.theguardian.com/world/2007/sep/12/russia.matthewweaver), le gouverment Russe à mit en place des mesures pour augmenter le taux de natalité « qui est en déclin depuis 1991 ».

3 - Analyse quantitative

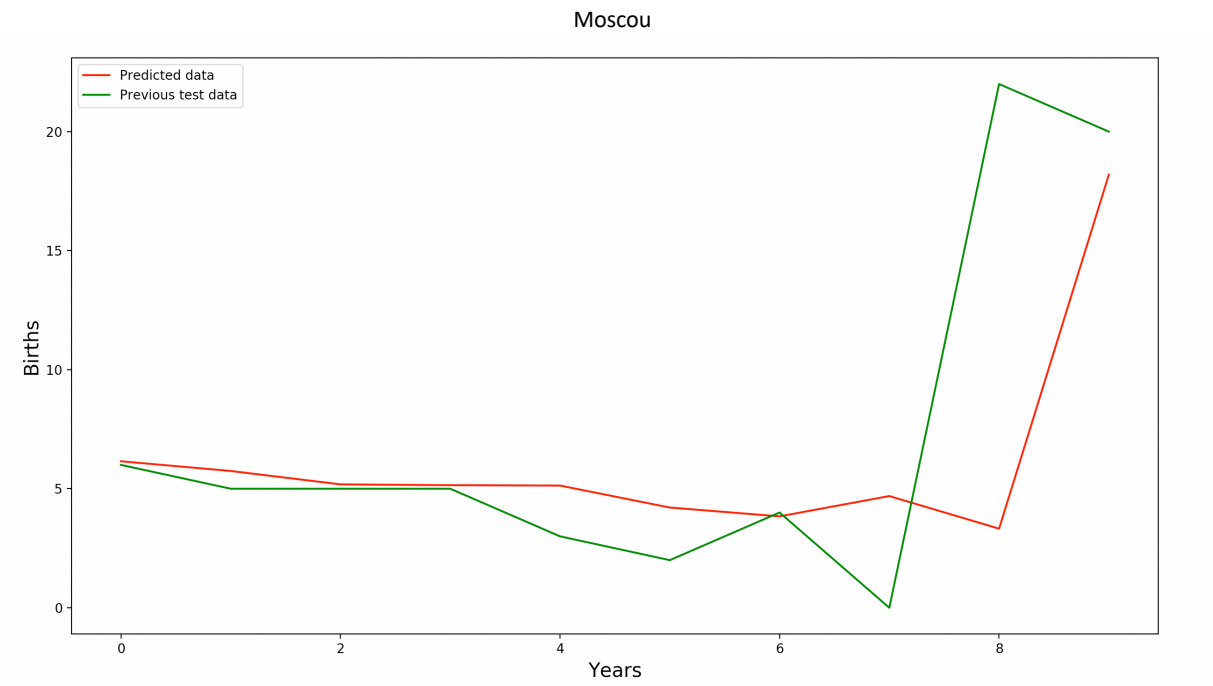


Figure 6

Le nombre de données mise à dispositions nous permettent d'effectuer une prédiction de la croissance du taux de natalité pour une région sur un nombre d'années défini (*cf Figure 6*).

Afin d'y parvenir, nous pouvons découper le processus de prédiction en 4 parties :

- 1er traitement,
- Encodage,
- Reframing,
- Splitting.

1er Traitement

Le 1er traitement consiste en une extraction de la data et une suppression des données manquantes/mal formatées/inutiles.

Encodage

Le stade d'encodage qui consiste à réassembler les données selon un protocole.

Reframing

Le reframing est la troisième étape du traitement des données où nous allons copier le dataframe et dans cette copie nous allons nous séparer des colonnes que nous ne voulons pas prédire.

Splitting

Dernière étape du traitement des données : le « Splitting ».
Ce dernier va consister à séparer notre data entre la data qui nous servira à l'entrainement et la data qui nous servira aux tests.

Le modèle

Nous allons maintenant, après le traitement de la data, aborder le modèle. Nous utilisons la régression logistique qui est un modèle consistant à modéliser au mieux un modèle mathématique simple à des observations réelles nombreuses. (qui sont ici les relevés du taux de naissance pour une période donnée.)