INF552 – DATA VIZUALISATION

JEREMIE DENTAN, PAUL THERON

Lien vers le répertoire GitHub du code : https://github.com/PaulRaid/INF552.git
Lien vers une page de visualisation du résultat : https://jdentan.com/departements/
Remarque : selon la rapidité de la connexion internet, le chargement des données des graphiques peut être lent.

INTRODUCTION

Dans un ouvrage nommé *Paris et le désert français* publié en 1947, Jean-François Gravier fut le premier à construire une explication à la fois globale et détaillée de ce que les géographes appellent la macrocéphalie parisienne (gonflement disproportionné de la ville capitale) et à lui apporter des propositions concrètes en prenant tout en compte : territoire, équipement, population, formation, activité professionnelle, mode de vie, transport (omniprésent dans l'ouvrage), logement (et son corollaire migratoire, éventuellement nécessaire), fiscalité...

Après avoir pris connaissance de ce constat fort très vrai à l'époque de parution de l'ouvrage, nous avons cherché à le mettre à l'épreuve du temps. La France a en effet mis en place dans les années 1980 avec les lois Defferre une politique connue sous le nom de l'acte 1 de la décentralisation. Quel a donc été l'effet de ces actions ? La position dominante de la métropole parisienne tend elle à disparaitre, ou à se renforcer ?

Nous avons, à l'aide de différents jeux de données fournis par l'INSEE, tenté de répondre à ces problématiques. Pour ce faire nous avons utilisé `d3` et diverses techniques de modélisation et d'animation.

I. DATASETS

1. PRESENTATION DES DATASETS

Nos *datasets* sont issus d'études effectués par l'INSEE, ils nous permettent d'extraire 5 variables démographiques par région et département, à savoir :

- L'espérance de vie, par département entre 1975 et 2020 ;
- La population, par département entre 1931 et 2018 ;
- Le PIB total et par habitant, par région entre 1990 et 2015 ;
- L'âge moyen des femmes à l'accouchement, par département entre 1975 et 2020 ;
- Le taux de natalité, par département entre 1975 et 2020.

Notons que chacune de ces variables fait référence à une définition plus ou moins complexe de l'INSEE (en particulier pour le calcul de l'âge moyen des femmes à l'accouchement et du PIB) ; pour plus de détail, le lecteur est invité à se rendre directement sur le site de l'INSEE.

Pour chacune de ces variables, nous avons recherché les séries temporelles de l'INSEE les plus précises possibles, que ce soit du point de vue temporel ou géographique. Ainsi :

- Nous avons cherché les séries avec un pas de temps le plus fin possible. Celui-ci était d'un an pour l'espérance de vie, l'âge moyen des femmes à l'accouchement, et le taux de natalité. Cependant, pour le PIB et la population, nous avons dû effectuer des régressions linéaires pour les dates manquantes.
- Nous nous sommes donnés pour objectif d'avoir des données à l'échelle du département (et non du canton ou de la région), car les départements constituent un bon équilibre entre des zones géographiques trop grandes (région) ou trop petites (canton, ce qui résulte en des frontières géographiques très arbitraires et pour lesquelles nous n'avons pas assez de données).

Notons cependant que pour le PIB par habitant, nous ne disposions que de données régionales, ce qui explique que tous les départements d'une région aient le même PIB.

2. PRE-PROCESSING

La phase de *pre-processing* a été une phase longe et très importante pour le succès de notre projet. En effet, bien que les bases de données de l'INSEE soient bien construites, certains éléments centraux étaient manquants :

- Les départements étaient désignés par leurs noms en toutes lettres et non leurs numéros, ce qui est problématique pour faire correspondre les lignes de différents *datasets* relatifs à un même département, surtout lorsque le nom présente quelques variations au niveau des majuscules, des espaces et des tirets. Nous avons donc utilisé une table de correspondance, ainsi que quelques ajustements manuels.
- Pour la population et le PIB, nous avons dû effectuer une régression linéaire pour les années manquantes.
- Pour l'espérance de vie, nous avons dû effectuer une moyenne entre les hommes et les femmes.
- Pour le PIB nous avons dû associer chaque département à une région pour en déterminer le PIB.
- Enfin, nous avons enregistrés toutes les données dans un format unique et standardisé afin de faciliter la lecture en vue de la représentation.

Cette phase de *pre-processing* est réalisée par le programme `cleaning.py`, qui a pour effet de sauvegarder des *datasets* nettoyés dans le dossier `donnees_clean`. À noter cependant que l'exécution a déjà été effectuée, et qu'il n'est pas nécessaire de l'exécuter à nouveau.

II. Visualisation

1. Objectifs

Nous nous sommes fixé plusieurs objectifs intermédiaires afin de mettre en évidence les phénomènes sous-jacents à l'évolution de la centralisation française. Il s'agit de mettre en évidence :

- La distribution géographique des variable démographiques ;
- Les corrélations entre ces variables ;
- L'évolution dans le temps de ces variables.

Pour répondre à ces objectifs, nous avons réalisé trois *plots* différents, utilisant chacune diverses méthodes de visualisation afin de répondre à ces trois objectifs.

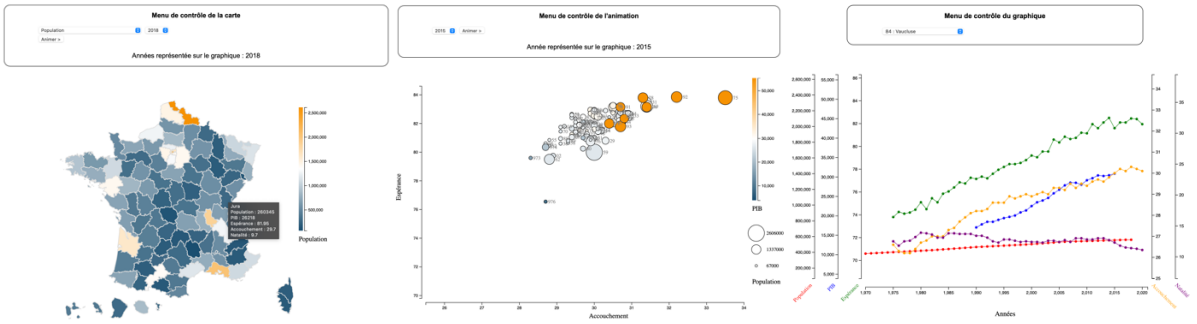

Figure 1 : Les trois plots de notre outil de visualisation

Les diverses techniques utilisées seront détaillées sous chaque item de visualisation.

2. Modelisation geographique des variables

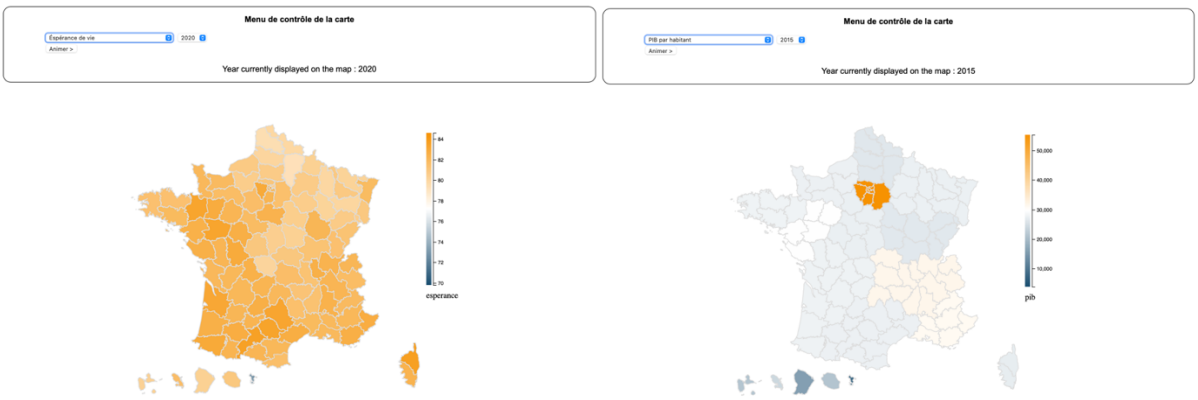

Figure 2 : Deux représentations différentes issues de notre premier plot

Nous avons dans un premier temps réalisé une carte de la France, avec les variables agrégées par département. Il est possible de visualiser l'évolution temporelle par département avec le bouton « Animer ».

Plusieurs choix graphiques ont été réalisés ici :

- L'utilisation de deux couleurs (bleu et orange) distinctes et fortement contrastées l'une par rapport à l'autre permet de capter l'attention et de distinguer clairement les différences de niveau.
- L'utilisation d'une *colorbar* à deux niveaux permet de mettre en évidence les extrêmes, que ce soit les maximaux ou les minimaux. En effet, comme décrit en introduction, nous nous intéressons aux spécificités de la région parisienne, et une *colorbar* à deux niveaux permet bien de mettre en lumière la spécificité de cette région comme on le voit à droite de la figure 2 dans le cas du PIB.

3. ÉVOLUTION TEMPORELLE PAR DEPARTEMENT

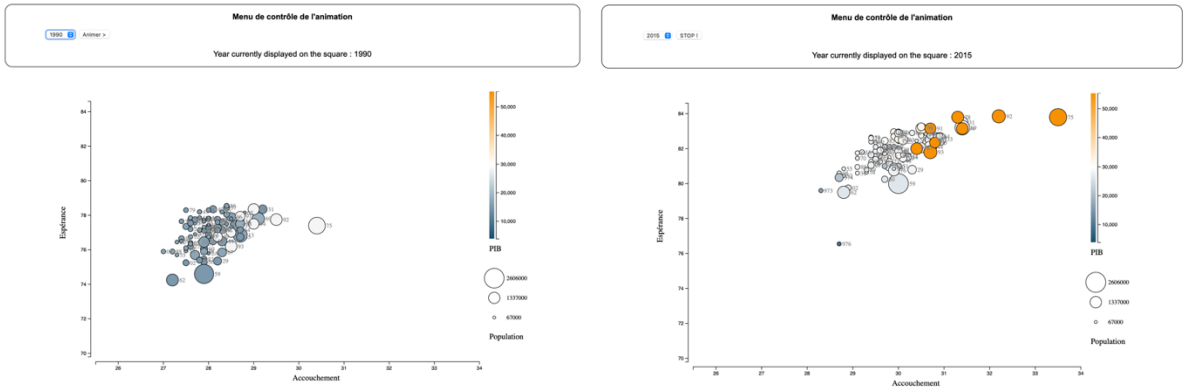

Figure 3 : Deux représentations différentes issues de notre deuxième plot

Notre deuxième *plot* présente un intérêt certain pour répondre à notre deuxième objectif, à savoir la visualisation de la corrélation entre les différentes variables. En effet, nous y avons représenté quatre de nos cinq variables (la cinquième, à savoir le taux de natalité, ne fournissait pas de résultats intéressants sur ce graphique), afin d'en visualiser l'évolution en parallèle, et donc la corrélation. Chaque cercle représente un département, dont la taille évolue avec la population, la couleur avec le PIB par habitant, et la position avec l'espérance de vie en fonction de l'accouchement.

Ces indicateurs correspondent mettent en lumière un certain modèle d'évolution des départements : avec le temps, les départements ont tendance à se déplacer vers le "nord-est" de notre graphique avec une couleur de plus en plus orange, ce qui correspond à un style de vie de plus en plus moderne et urbain (une espérance de vie élevée, des enfants qui arrivent de plus en plus tard, pour un département de plus en plus riche). Cette évolution est particulièrement visible et rapide pour les DROM, qui arrivent tardivement dans nos jeux de données.

Enfin, ce mode de représentation permet de mettre en lumière une différentiation très claire entre différentes régions (notamment les régions parisiennes et lyonnaises), certaines se retrouvant bien plus clairement isolées en 2015 qu'en 1990.

4. COMPARATIF DES EVOLUTIONS

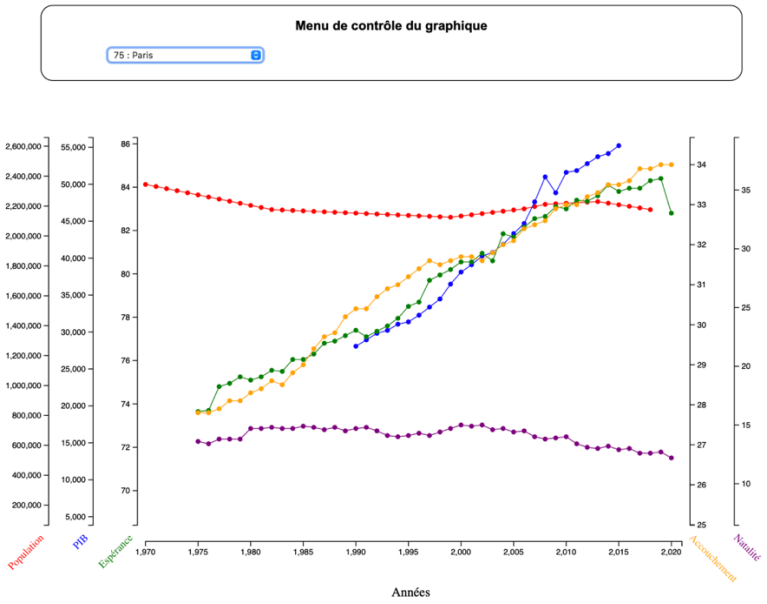

Figure 4 : Une représentation issue de notre dernier plot

Enfin, afin de répondre à notre dernier objectif, à savoir une visualisation précise de l'évolution temporelle de nos variables démographiques dans un département, nous avons réalisé un dernier *plot*, représenté sur la figure 4.

Ce dernier *plot* utilise le même mode de représentation pour nos cinq variables, à savoir la position verticale sur un axe. Ce mode de représentation étant celui avec lequel l'œil humain perçoit de façon la plus précise les variations et les différences, ce dernier *plot* permet de visualiser très précisément les corrélations entre les évolutions de nos variables au sein d'un département (ce qui est plus compliqué par exemple sur notre deuxième *plot*, car une comparaison entre une évolution de la taille d'un cercle et de sa couleur n'est pas aisée).

Pour ce dernier plot, les cinq axes verticaux sont superposés, ce qui complique la lecture d'une valeur en particulier. Cependant, l'objectif premier de ce *plot* est la comparaison des évolutions et non la lecture d'une valeur donnée ; par ailleurs la lecture d'une valeur en particulier peut se faire aisément en passant la souris sur un point, grâce à un *tooltip*.

III. Zoom sur nos animations

L'une des principales difficultés techniques de ce projet a été la réalisation d'animations qui s'exécutent en continu. En effet, les outils `d3` présentés en cours ne permettent que de réaliser des animations ponctuelles d'un état à un autre, mais pas des animations continues.

Pour faire face à cette difficulté, nous avons utilisé, pour chaque animation, une variable globale (par exemple `animationMap`) permet de retenir l'état de l'animation et les prochaines étapes à suivre :

- Cette variable globale mémorise les données à afficher, ainsi qu'un booléen indiquant si l'animation est en cours ou non ;

- Puis, une fonction récursive (par exemple `nextStepAnimationMap`) est munie d'un `setTimeout` natif en `JavaScript`. Cette dernière s'appelle récursivement pour itérer sur la variable globale et en mettre à jour l'état.

Cette implémentation, bien qu'elle fonctionne, n'est pas tout à fait satisfaisante. En effet, l'usage d'un `setTimeout` n'est pas optimal, car le temps de calcul nécessaire lors de chacun des appels récursifs entre en concurrence avec la fluidité de l'animation. Par exemple, nous avons initialement tenté une animation pour laquelle certains appels récursifs effectuaient l'animation sur plus d'une année. Nous avons alors constaté que l'animation était saccadée, s'arrêtant plus de temps sur les années pour lesquelles la fonction s'appelait récursivement.

Bien qu'une solution plus efficace existe sans doute, la solution que nous avons choisie nous semble cependant convenir pour les objectifs que nous nous étions fixés.

IV. Interprétation des résultats

Nous nous étions demandé si l'hégémonie parisienne avait réduit à a suite des politiques de décentralisations mises en place à partir des années 1980. Or, les visualisations que nous avons produites mettent en évidence le contraire. En effet, la région parisienne est en tête de file sur de nombreux indicateurs clefs, tels que la richesse des habitants et l'espérance de vie. De même, la région évolue clairement vers un modèle démographique plus moderne et se démarque des autres, avec des enfants qui arrivent de plus en plus tard dans la vie des habitants et une espérance de vie qui ne cesse de croître.

Toutefois, nos différentes courbes et animations ont pu mettre en évidence dans les région lyonnaise ou marseillaise présentent des motifs d'évolution similaires, avec cependant un « retard » sur la capitale. À l'inverse, certaines régions connues sous le dénominatif de « diagonale du vide », notamment en Lorraine et Ardenne, sont clairement identifiées en queue de file pour ces variables démographiques.

Ainsi, bien que nos représentations ne puissent en aucun cas se substituer à une étude sociologique approfondie pour capter la subtilité démographique de nos territoires, notre travail tend à montrer que l'écart entre Paris et le « désert français » ne s'est pas rétrécit, bien au contraire, mais s'interpréterait désormais plus comme un écart entre « Nos villes et le désert rural ».

Conclusion

Ce projet aura été pour nous très instructifs dans le cadre de notre formation en sciences des données. En effet, il nous a premièrement confronté à une étape importante d'un projet de data science, à savoir le nettoyage des données. Puis, il a nous a donné l'occasion de nous interroger en profondeur sur les questions auxquelles nous voulions répondre à partir de nos *datasets*, et de réfléchir à une stratégie pertinente pour les représenter au mieux. Enfin, ce projet nous a permis de progresser sur le plan technique, que ce soit au niveau de la maîtrise de `JavaScript` et `d3` ou lors du *pre-processing* des données.