Dossier cartographique

Statistiques et cartographie avec R

2020-2021

Consignes

Ce devoir à la maison reprend l'ensemble des notions et méthodes de statistiques et de cartographie abordées en cours.

Il s'agit de produire un dossier individuel cartographique mis en page et structuré. Une attention particulière sera portée à la qualité des cartes et graphiques produits (couleur, habillage), ainsi qu'à leurs commentaires.

Le format de rendu est libre, mais il est recommandé de travailler avec un fichier Rmarkdown.

La date de rendu est le 08 février 2021. Merci d'envoyer un mail qui contienne :

- le rapport cartographique : un fichier html, un fichier pdf ou un fichier texte (word ou libre office), qui comprend texte, cartes et commentaires. Vous traiterez les questions dans l'ordre.
- Votre script, au format NOM_Prénom.R, qui reprendra dans sa structure les numéros de chaque question, et comportera les lignes de code utilisées pour produire les cartes, graphiques et résultats. Si vous travaillez dans un fichier Rmarkdown, vous pouvez directement inclure les lignes de code dans le rendu final.

Présentation des données

Vous disposez d'un jeu de données sur les élections présidentielles aux Etats-Unis, au format *GeoPackage*. Le fichier comporte, pour chaque comté ou *county*, les variables suivantes :

- fips5 : identifiant du comté
- state_abbv : Abréviation de l'Etat fédéré d'appartenance.
- county_name : Nom du comté
- state_name : Nom de lc'Etat fédéré
- votes_dem_2016 : nombre de votes pour le parti démocrate en 2016
- votes_gop_2016 : nombre de votes pour le parti républicain en 2016
- votes_dem_2020 : nombre de votes pour le parti démocrate en 2020
- votes_gop_2020 : nombre de votes pour le parti républicain en 2020
- total_votes2016 : nombre total de votes en 2016
- total_votes2020 : nombre total de votes en 2020
- TotalPopulation : population du comté en 2010
- White : population blanche en 2010
- Hispanic : population hispanique en 2010
- Black : population noire en 2010

- Asian : population asiatique en 2010
- MedianIncome : revenu médian par comté en 2019
- CBSA : nom de l'aire urbaine d'appartenance
- Code2013 : classification géographique du comté (voir documents joints)

L'Alaska a été exclu du jeu de données. Les données sont déjà projetées dans le système de coordonné approprié.

Vous avez également une carte et un schéma à disposition sur Slack et sur le site git du cours, qui détaillent la classification géographique de chaque comté selon leur appartenance urbaine ou rurale (plus d'informations, si vous le souhaitez, ici).

Exercice 1 (*5 points*)

1. Ouvrez le jeu de données *us_elections.gkpg* en utilisant le package 'sf'. Combien de variables et d'unités spatiales composent ce jeu de données ? Quelle fonction pouvez vous utiliser pour répondre à cette question ? Quel est le système de projection utilisée ?
2. Proposez une ligne de code qui recensent le nombre de comtés pour chaque Etat fédéré. Combien l'Etat de Caroline du Nord comporte-t-il de *counties* ?
3. En utilisant les fonctions du tidyverse, créez une nouvelle variable qui recode les modalités de la variable *Code2013* en utilisant les termes géographiques proposés par les documents complémentaires. Proposez une carte qui montre la classification des comtés, à l'image de celle fournie en Annexe, avec une palette de couleur appropriée.
4. En agrégeant les données à l'échelle des 917 CBSA (ou aires urbaines), calculez le pourcentage de la population étasunienne qui vit en dehors d'une aire urbaine en 2010.
5. Proposez une carte en cercles proportionnels qui montre la population des 10 plus grandes aires urbaines du pays. Indice : pour obtenir les centroïdes d'un polygone, il suffit d'utiliser la fonction st_centroid() du package sf, qui renvoie un objet sf avec des points.

```
points <- objetSFpolygone %>% st_centroid()
```

Aide pour les prochaines étapes.

La fonction *pivot_longer* de la suite tidyverse permet de ré-agencer un tableau pour calculer et visualiser plus facilement les données.

Les ligne de code ci-dessous calculent 1) l'effectif total de chaque groupe racial à l'échelle du pays, via les fonctions *group_by* et *summarize* puis 2)

- avec le paramètre "names_to", les variables, ou colonnes, "White", "Asian", "Black" et "Hispanic" deviennent les modalités d'une nouvelle variable qualitative de 4 modalités, intitulée "Race" ;
- avec le paramètre "values_to", les effectifs de chaque groupe racial deviennent les modalités d'une nouvelle variable quantitative, intitulée "Population".

On obtient ainsi un tableau avec 3 variables :

- l'état d'appartenance ;
- Une variable qualitative "Race" : la catégorie raciale, au format *character* ;
- Une variable quantitative "Population" : l'effectif de chaque catégorie raciale, au format *numeric*.

Les donnée sont ainsi dites *tidy*, c'est à dire "rangées". Chaque ligne du tableau correspond à 1 unité spatiale, 1 modalité de la variable qualitative et l'effectif de cette modalité. Autrement dit, une ligne = une unité spatiale = un groupe racial = une population.

On peut donc créer très facilement, avec la fonction *mutate*, une 4e colonne, intitulée “PCT”, qui stocke donc, pour chaque ligne et chaque unité spatiale, le pourcentage que représente chaque groupe racial.

La 3e étape du code permet de visualiser le nouveau tableau créé, par un diagramme en bâtons, avec le package ggplot2.

```
#################################################################################

library(tidyverse) #si ce n'est pas déjà fait

## Etape 1 : calcul des effectifs agrégés

nouvelObjet <- us_elections %>% # le jeu de données original
    st_set_geometry(NULL) %>% #j'enlève les données géométriques
    group_by(state_name) %>% #je regroupe éventuellement les entités selon un identifiant commun
    summarise(White = sum(... ), # Je calcule les totaux pour chaque groupe
              Black = sum(...), # à vous de remplacer les points de suspension !
              Hispanic = sum(...),
              Asian = sum(...))

nouvelObjet #regardez l'objet

# Etape 2 : Réagencement du tableau
nouvelObjet2 <- nouvelObjet %>%
  select(state_name, White, Black, Asian, Hispanic) %>% # sélection des variables à réagencer
  pivot_longer(-state_name, # identifiant que je souhaite conserver
               names_to = "Race", #le nom de la variable qualitative
               values_to = "Population") %>% #le nom de la variable quanti
  group_by(state_name) %>% # pour regrouper les individus selon leur état
  mutate(PCT = Population/sum(Population)*100) #je crée une colonne de pourcentage

nouvelObjet2 #regardez l'objet

# Etape 3 : Visualisation
ggplot(nouvelObjet2, aes(Race, Population,
                         fill =  Race)) +  #remplissage selon les modalités de la variable Race
  geom_col() #diagramme en bâton

#################################################################################
```

6. En vous inspirant de ces lignes de codes, que vous pouvez compléter et copier-coller, calculez la population totale de chaque groupe racial aux Etats-Unis, exprimée en valeur absolue et en pourcentage. Proposez ensuite un diagramme en bâton qui permette de visualiser le pourcentage respectif de chaque groupe racial aux Etats-Unis.
7. En utilisant les outils de la statistique descriptive, commentez la distribution de chaque groupe racial en pourcentage, à l’échelle des comtés, avec le vocabulaire et les visualisations graphiques appropriés.
8. Proposez une carte qui montre le pourcentage de la population afro-américaine (‘Black’) par comté. Expliquez votre choix de discrétisation et commentez les structures spatiales.

Exercice 2 : analyse des résultats électoraux (*7 points*)

1. Cartographier le vainqueur de chaque état fédéré en 2016 et en 2020, puis comparez et commentez les deux cartes.
2. Calculez, en pourcentage, les résultats des candidats Trump et Biden aux élections de 2016 et 2020 au

sein de chaque chaque comté. Précisez la formule utilisée pour calculer ces quatre nouvelles variables.

3. Avec les variables obtenues, comparez les résultats de Trump et de Biden en 2020, en cartographiant leurs pourcentages obtenus à l'échelle des comtés. Commentez la distribution statistique des résultats de chaque candidat, justifiez la discrétisation choisie, réalisez les deux cartes et commentez les structures spatiales. Vous pouvez vous référer au manuel de Lambert et Zanin, disponible sur Slack.

4. Afin de représenter l'intensité d'un vote à l'échelle locale, il est intéressant de mesurer et cartographier l'écart entre deux candidats. Explication : pour une élection opposant un candidat A à un candidat B, l'écart s'obtient en calculant le rapport entre la valeur absolue de la différence de vote entre A et B avec le nombre total de votes. L'écart se calcule ainsi de la manière suivante :

$$Ecart = \frac{|Vote_A - Vote_B|}{VoteTotal} * 100$$

Vous devrez montrer *sur une même carte* l'écart en faveur de Biden et l'écart en faveur de Trump. Décrivez les étapes du traitement des données, justifiez le choix de votre discrétisation, et commentez avec précision les structures spatiales révélées. Indice : une option peut être de diviser le jeu de données en deux : un objet pour les comtés où Trump est vainqueur ; un autre pour ceux où Biden est vainqueur. Ensuite, vous pouvez cartographier l'écart, en juxtaposant, sur une même carte, ces deux objets, mais avec un gradient de couleur différent selon l'identité du candidat.

5. Les élections sénatoriales en Georgie ont constitué une étape déterminante pour les élections de 2020. Proposez une carte commentée de l'écart de vote pour cet Etat, en regardant également la composition raciale de cet Etat à l'échelle des comtés. Quelles sont les caractéristiques géographiques, économiques et raciales des comtés remportés par Biden ?

6. La dissociation entre ruralité et urbanité a bonne presse pour expliquer les résultats électoraux aux Etats-Unis. A l'aide de boites à moustache (*boxplot*), représentez les résultats, en pourcentage, des candidats Trump et Biden selon la classification géographique des comtés (*voir séance 6, slide 36*). Quelles différences sont visibles entre le parti démocrate et le parti républicain lorsque l'on prend en compte la classification géographique des comtés ?

7. Proposez une nouvelle variable catégorielle qui classe les 917 aires urbaines selon leur population, d'après les seuils suivants : 20 000, 50 000, 100 000, 250 000, 500 000, 1 million, 2 millions, 5 millions. En utilisant cette nouvelle variable ordinale, proposez ensuite un graphique ou un tableau qui montre le nombre et le pourcentage d'aires urbaines remportées par Donald Trump et Joe Biden. Votre tableau ou graphique doit permettre, par exemple, de répondre aux questions suivantes : quel pourcentage d'aires urbaines entre 2 et 5 millions a remporté Donald Trump ? Combien d'aires urbianes de moins de 50 000 habitants a remporté Joe Biben ? Indice : il s'agit donc de proposer une analyse de fréquence.

Exercice 3 : analyse et modélisation des résultats (*8 points*)

1. Commentez, en proposant un nuage de points (*scatter plot*) et avec le vocabulaire approprié, la relation entre les résultats, en pourcentage par comté, de Trump aux élections de 2016 et ses résultats obtenus en 2020. Faites de même pour le parti démocrate. Calculez et interprétez les coefficient de corrélation.

2. Proposez un modèle de régression linéaire pour prédire les résultats, en pourcentage par comté, de Trump en 2020 en fonction des résultats de 2016. Faites de même pour Biden. Pour chaque modèle, écrivez la formule utilisée sous R, proposez une visualisation graphique, commentez les résultats du modèle et écrivez l'équation obtenue. Quel est le modèle le plus performant ?

3. D'après le modèle, quel serait en 2020 le résultat de Trump dans un comté où le parti républicain aurait obtenu 18% en 2016 ? Quel serait en 2020 le résultat de Biden dans un comté où le parti démocrate a obtenu 48% en 2016 ?

4. Cartographiez les résidus de vos deux modèles, en veillant à mettre en valeur leur caractère positif ou négatif, et commentez les structures spatiales des écarts au modèle. Pour rappel, les unités spatiales

dont les résidus sont peu éloignés du modèle doivent être inclues dans une classe centrale, afin de mettre en valeur graphiquement les unités spatiales qui s'en éloignent positivement ou négativement.

5. Calculez la densité par habitant par comté en 2010, puis visualisez et caractérisez la relation entre la densité et le résultat de Biden en 2020.

6. A l'échelle des comtés, y a t-il un lien entre le revenu médian et le résultat de chaque candidat ?

7. A l'aide d'un test du khi-deux, analysez le lien entre les comtés où Biden a été donné vainqueur 2020 et la classification géographique du comté. Dans un tableau, vous préciserez les effectifs réels et effectifs théoriques. Vous commenterez les résultats du test, en incluant un graphique et un commentaire des résidus. Quelle peut être la limite de ce test réalisé avec cette classification de 6 modalités ?

8. En utilisant les fichiers fournis, recoder la classification géographique des comtés en deux modalités : métropolitain, et non-métropolitain. En effet, d'après le Département de l'Agriculture, les "*Nonmetro Counties are commonly used to depict rural and small-town trends*". Effectuez de nouveau un test du khi-deux avec cette nouvelle variable et comparez les résultats avec le précédent modèle.