Méthodologie de la normalisation des adresses du REU

Juin 2023

La mise à disposition de la correspondance des adresses des électeurs avec les bureaux de vote du Répertoire électoral unique (REU) a nécessité des traitements visant à garantir la confidentialité des données à caractère personnel et à normaliser les informations géographiques pour en faciliter l'usage. Ce document présente ces traitements, réalisés par l'Insee avec l'aide d'Etalab.
Il reprend principalement les éléments présentés lors de l'atelier du 12 décembre 2022 au Lieu de la Transformation Publique.

1. Initialisation

Analyse exploratoire des données & preprocessing

A. Les champs à traiter

L'extraction du REU dont proviennent les données a été réalisée en septembre 2022. Les lignes du fichier brut correspond aux couples (adresse de rattachement, bureau de vote) distincts dans le Répertoire Électoral Unique, ce dernier étant mis à jour en permanence. Une ligne ne correspond donc pas nécessairement à un électeur distinct. Au sein de chaque ligne sont renseignées les adresses de rattachement et de contact (parfois différentes) ainsi que les informations liées au bureau de vote correspondant. Dans la normalisation des adresses, certains champs sont à prendre en considération plus que d'autres, notamment :

- **Champs pouvant comporter des données personnelles :**
  - Numéro de voie : *num_voie*
  - Nom de la voie : *voie*
  - Compléments : *complément_1* et *complément2*
  - Lieu-dit : *lieu_dit*
- **Autres champs :**
  - Code postal : *cp*
  - Nom de la commune (tel que rentré manuellement) : *commune*
  - Pays : *pays*

B. Premières opérations

- **Reconstruction de variables :**
  - Reconstruire les départements à partir des codes postaux (00 : inconnu, 99 : étranger)
  - Reconstituer les codes commune des adresses à partir des libellés des communes pour comparer au code commune du bureau de vote associé
- **Traitements initiaux :**
  - Passer tous les champs en minuscules
  - Retirer toute ponctuation (hors tirets) et les doubles espaces

2. Anonymisation

Retirer les informations personnelles des données

A. Procédure d'anonymisation

- Établir une liste des variations du mot “chez” et autres mots-clefs (hébergé, bàl, …)
- Pour chaque champ concerné, repérer les adresses pour lesquelles l’une des variations est présente
    - Certaines variations ne sont recherchées qu’en début de champ (ex: Monsieur, Mme)
- Pour les adresses concernées, lancer la méthode d’anonymisation

B. Méthode d’anonymisation

- Repérer le premier mot sensible identifié dans le champ (ex : chez, hébergé, …)
- Si celui-ci est précédé ou suivi de certains mots-clefs identifiés, laisser passer
    - Ex: le mot est précédé d’un déterminant ou suivi d’une préposition bien choisie
    - “Avenue chez les cygognes” ou “Rue de chez St-Jérôme”
- Sinon, regarder ce qui se trouve après le mot sensible :
    - Si un seul mot suit, laisser passer (généralement un nom de commerce, ex: chez Bertrand)
    - Si ” - ” apparaît plus loin, conserver uniquement ce qui se situe après
        - Ex: “3 rue des hêtres chez Mme Martin - résidence des fleurs”
    - Si un numéro ou type de voie (ex : rue) apparaît plus loin, conserver uniquement ce qui se situe après
        - Ex: “chez M. Robert 4 rue des marrons”
    - Sinon, tout supprimer après le mot sensible

C. Utilisation des adresses de référentiels

Les opérations susmentionnées ne garantissent pas à 100% le retrait de toutes les informations personnelles. L’ultime opération d’anonymisation est donc la normalisation des adresses par les référentiels nationaux, comme détaillé en section 4. Afin de garantir une anonymisation parfaite des données diffusées, ce sont donc bien les adresses de ces référentiels, eux-mêmes considérés parfaitement anonymisés, qui sont diffusées.

3. Nettoyage

Retirer les spécifications d’adresses nocives pour la normalisation et géolocalisation

A. Procédure de nettoyage

- Établir une liste de toutes les spécifications supplémentaires d’adresse (ex : étage, apt, …)
- Pour chaque champ concerné, distinguer l’adresse si l’une des spécifications est présente
    - On cherche également les codes postaux répétés dans le mauvais champ pour les y supprimer
- Pour les adresses concernées, lancer la méthode de nettoyage

B. Méthode de nettoyage

- Repérer la 1ère spécification sensible présente (ex: apt)
- Si celle-ci est précédée ou suivie de certains mots-clefs identifiés, laisser passer
    - Ex: On retirera “entrée” dans “5 rue des marrons entrée 3” mais pas dans “rue de l’entrée dorée”
- Sinon, regarder ce qui se trouve après le mot sensible :
    - Tout supprimer jusqu’à retomber sur un mot d’au moins 3 lettres
    - On enlève ainsi les “étage 3”, “apt B”, “lgt A 25”, …
- Recommencer jusqu’à avoir retiré toutes les spécifications inutiles

C. Exemple 1

- **Adresse initiale :** “6 rue de Sévigné / chez M. Jean Bernard chemin du Bourg, bât 3 apt 27”
- **Traitement initial :** “6 rue de sévigné chez m jean bernard chemin du bourg bât 3 apt 27”
- **Après anonymisation :** “6 rue de sévigné chemin du bourg bât 3 apt 27”
- **Après nettoyage :** “6 rue de sévigné chemin du bourg”

D. Exemple 2

- **Adresse initiale :** “4 Rue de chez Les Lièvres, étage n°42”
- **Traitement initial :** “4 rue de chez les lièvres étage n 42”
- **Après anonymisation :** “4 rue de chez les lièvres étage n 42”
- **Après nettoyage :** “4 rue de chez les lièvres”

E. Traitements spéciaux

- **Nettoyer *num_voie* :** On ne garde que les nombres éventuellement suivis de :
  - Une seule lettre
  - bis/ter
- **Générer l’adresse complète :** On concatène les différents champs de l’adresse
  - On cherchera notamment à éliminer les éventuels dupliqués dans 2 champs successifs

4. Géolocalisation

Passer des adresses rentrées manuellement à un référentiel officiel

A. Dédoublonner les adresses

- Le fichier initial comprend 26M adresses de rattachement
- Après anonymisation et nettoyage, certaines adresses sont devenues identiques. Il est donc possible de partiellement dédoublonner la table en fonction des adresses nettoyées, notamment pour limiter les entrées à l’API de la BAN et de la BANO
- En pratique, la table dédoublonnée a un peu moins de 20M d’entrées

B. Normaliser et géolocaliser les adresses

- On n’envoie ainsi pour géolocalisation que des champs traités (anonymisés et nettoyés)

- Un référentiel choisi pour normaliser et géolocaliser les adresses : la **Base Adresse Nationale** (BAN) et la **Base Adresse Nationale Ouverte** (BANO)

  - Via des requêtes à une API
  - Coordonnées GPS (WGS84, EPSG 4326)

- Un script repris de Christian QUEST, initialement écrit pour géocoder la base Sirene

C. Les résultats de la normalisation

- Les adresses issues des référentiels sont par définition :
  - Nettoyées et anonymisées
  - Géolocalisées
- Toutes les adresses données en entrée aux référentiels ne trouvent pas une adresse correspondante normalisée.
  - Certaines sont donc seulement identifiées au niveau de la voie, de la commune, voire pas du

tout.
    - Les variables de qualité renvoyées par les deux outils de géolocalisation sont présentes dans la table finale.

D. Principe des requêtes aux référentiels d'adresses

- On cherche à normaliser chaque adresse au sein de la commune déclarée.
    - Si on ne trouve pas au numéro de voie (ou que le score de fiabilité est trop bas), on cherche au niveau de la voie
    - Puis même chose à l'échelle de la commune
- Plusieurs bases différentes (BAN, BANO) sont requêtées les unes après les autres si l'on ne trouve pas de correspondance dans les précédentes.
- Tous les résultats renvoyés sont donc des normalisations jugées suffisamment fiables. De plus, deux filtres supplémentaires sont ajoutés :
    - Ne sont gardées que les adresses situées dans la même commune que le bureau de vote associé
    - Ne sont gardées que les adresses géolocalisées plus finement qu'à l'échelle de la commune, ou bien situées dans une commune avec un seul bureau de vote

5. Finalisation

Mise en forme des résultats issus de la BAN et de la BANO

A. Mise en forme des résultats

Une fois les adresses normalisées via les référentiels, un vrai dédoublonnage sur le couple (adresse, bureau de vote associé) à partir des lignes initiales est réalisé. Plusieurs colonnes sont alors ajoutées.

- A l'échelle des adresses :
    - *nb_adresses* : nombre de lignes dans le REU correspondant au couple (adresse normalisée, bv associé) considéré
        - Attention, cela ne correspond pas au nombre d'électeurs à l'adresse considérée, le fichier initial des adresses du REU étant déjà partiellement dédoublonné
- A l'échelle des bureaux de vote :
    - *nb_adresses_initial* : nombre d'adresses initialement présentes dans le fichier originel du REU correspondant au bureau de vote
    - *nb_adresses_final* : nombre d'adresses présentes dans le fichier diffusé des adresses du REU correspondant au bureau de vote
        - Les adresses manquantes, en très faible proportion, correspondent à celles n'ayant pas pu être normalisées et géolocalisées avec une certitude suffisante par la BAN ou la BANO

B. Table de correspondance expérimentale et ponctuelle des référentiels de bureaux de vote

Dans la grande majorité des cas, les identifiants des bureaux de vote sont identiques dans le REU et dans le système d'information centralisant les résultats électoraux du ministère de l'Intérieur. Il subsiste toutefois quelques cas (3 000 bureaux de vote sur 69 000), où ces identifiants sont différents. Afin de permettre plus facilement le rapprochement de ces données, nous proposons une table de correspondance

des identifiants des bureaux de vote avec les référentiels du site de l'Insee et du ministère de l'Intérieur. Ces correspondances ne revêtent aucun caractère officiel et ne sont qu'un travail expérimental conjoint entre l'Insee et Etalab afin d'offrir une possibilité de mise en cohérence facilitée des données. Celle-ci n'est donc pas fiable à 100%. En particulier, certains bureaux de vote n'ont pas pu être associés d'un référentiel à l'autre avec les informations à disposition. Certains identifiants (hors identifiant REU), bien que très rares, peuvent donc être manquants. De plus, le travail effectué n'est pertinent que pour les données considérées, les différents référentiels pouvant changer avec le temps.

Pour plus de détails sur la méthodologie employée, les étapes sont les suivantes :

- Constitution d'un indice de bureau de vote comme {code_commune}_{code_du_bv_dans_la_commune} des 2 côtés, suivi d'un nettoyage standard sur les codes des bureaux (notamment retirer les caractères spéciaux et les *leading zeros*). Pour les collectivités d'outre-mer (hors Saint-Martin et Saint-Barthélémy), conversion des codes communes selon les indications du ministère de l'Intérieur.
- Jointure classique sur les indices construits, ce qui met en concordance la majorité des bureaux de vote (~66k sur un total de 69k de bureaux dans le REU)
- Application des règles de conversion obtenues auprès du ministère pour les indices des bureaux de vote des villes à arrondissement (Paris, Lyon, Marseille), qui a permis de mettre en concordance la quasi-totalité des bureaux dans ces communes (~1.6k bv)
- Association systématique des bureaux de vote seuls dans leur commune dans les 2 référentiels (~400 bv)
- Association manuelle des bureaux de vote des 2 fichiers lorsque le rapprochement est « évident », par exemple des codes [1, 2, 3, …] d'un côté et [101, 102, 103, …] de l'autre, mais qu'on ne peut pas rapprocher automatiquement sans vérification à l'œil nu en raison d'exceptions existantes (~300 bv rattrapés)

A l'issue de ce traitement, environ 130 bureaux de vote côté REU n'ont pas trouvé leur équivalent dans l'autre fichier, et près de 250 côté SIE1. Une grande partie de ces rebuts sont localisés dans 3 communes : Belfort, Troyes et Dieppe, pour lesquelles les indices des bureaux de vote de chaque côté ne permettent pas de faire d'association avec certitude. Pour les autres, il s'agit peut-être de bureaux de vote divisés en deux d'un côté et pas de l'autre ou bien de changements de communes.

C. Publication

Il est prévu que les adresses normalisées du REU soient publiées en juin 2023 sous la forme de 2 fichiers :

- La table des adresses du REU
- La table des bureaux de vote du REU

Plusieurs propositions d'utilisation des fichiers dans le cadre de la génération de contours devraient suivre cette publication.