phacochr: un géocodeur pour les géocoder tous

Package R pour réaliser le géocodage d'adresses en Belgique

Joël Girès 1* Hugo Périlleux 2†

Résumé (max 300 mots)

Nous présentons `phacochr`: un géocodeur pour la Belgique sous forme de package R. Son principe est de produire de manière simple et rapide, à partir d'une liste d'adresses, une série d'informations nécessaires pour l'analyse spatiale: les coordonnées X-Y mais également d'autres informations utiles comme le secteur statistique (la plus petite unité géographique belge). Le programme fonctionne sur base de données publiques d'adresses quasi-exhaustives pour la Belgique compilées à partir des listes régionales (Région de Bruxelles-Capitale, Région flamande et Région wallonne). Par ailleurs, le géocodage est réalisé entièrement en local, permettant une confidentialité maximale dans le cas de données d'adresse sensibles. `phacochr` réalise également la cartographie des adresses localisées sur base de shapefiles intégrés. Nous désirons présenter ce package afin d'avoir des retours avisés de la communauté R à propos des solutions techniques choisies et des améliorations possibles.
Mots-clefs : Package – Géocodage – SIG

Logique de `phacochr`

Le géocodeur `phacochr` a été développé pour constituer une alternative face aux solutions existantes - notamment GoogleMap - tout en reposant entièrement sur des données publiques, des procédures libres, et un fonctionnement local[1].

La logique de `phacochr` est de réaliser une jointure inexacte entre la liste d'adresses à géocoder et les données publiques d'adresses installées en local (contenant les coordonnées X-Y). Lors de cette opération, `phacochr` dispose de plusieurs options: il peut notamment réaliser des corrections orthographiques (en français et néérlandais) préalables à la détection des rues ou procéder au géocodage au numéro le plus proche - de préférence du même côté de la rue - si les coordonnées du numéro indiqué sont inconnues (par exemple si l'adresse n'existe plus). En cas de non disponibilité du numéro de la rue, le programme indique les coordonnées du numéro médian de la rue. `phacochr` est compatible avec les 3 langues nationales: il géocode des adresses écrites en français, néérlandais ou allemand - option indispensable, la Belgique étant un pays trilingue. L'utilisateur/trice peut indiquer l'erreur acceptable maximale lors de la jointure inexacte. Augmenter ce paramètre accroît le pourcentage de rues trouvées, mais aussi d'erreurs réalisées. Dans le cas où la ou les langue(s) dans laquelle les adresses sont inscrites sont connues, l'utilisateur/trice peut les renseigner via un argument, ce qui augmente la vitesse et la fiabilité du processus en limitant le matching à la langue définie.

Données séparées

Le package `phacochr` a la particularité de ne pas contenir directement les données nécessaires au géocodage. Le fait que les données ne soient pas intégrées a l'avantage de donner à l'utilisateur/trice la possibilité de mettre à jour lui/elle-même les données.

Notre package dispose en effet d'une fonction qui télécharge les dernières données publiques d'adresses sur le site du service public fédéral les hébergeant: cette fonction les transforme et les enrichit pour les

*Observatoire de la Santé et du Social de Bruxelles-Capitale, jgires@ccc.brussels

†Université Libre de Bruxelles - Institut de Gestion de l'Environnement et d'Aménagement du Territoire, Hugo.Perilleux@ulb.be

[1]La documentation du package est disponible en ligne: https://phacochr.github.io/phacochr/index.html.

rendre compatibles avec le fonctionnement de `phacochr` et sauvegarde le résultat dans le répertoire de travail du package pour être utilisées lors du géocodage. Cette manière de faire a l'avantage majeur que l'utilisateur/trice peut réaliser lui/elle-même la mise à jour des données. Les données publiques d'adresses sont en effet mises à jour de manière hebdomadaire, et il aurait été beaucoup trop contraignant pour nous de mettre le package à jour toutes les semaines pour suivre ce rythme. Ce fonctionnement en local a par ailleurs l'utilité de pouvoir géocoder des données sensibles qui ne peuvent pas être envoyées sur un serveur via une API de géocodage.

Performance

Nous avons tenté d'optimiser le fonctionnement de `phacochr` en termes de performances et de qualité des résultats - travail qui est toujours en cours.

Le matching inexact sur lequel repose le `phacochr` est notamment parallélisé sur les n-1 cores du CPU afin d'augmenter la vitesse du calcul. Concernant la jointure des adresses trouvées avec les coordonnées géographiques contenues dans les données publiques d'adresses, seuls les arrondissements belges dans lesquels sont présents les codes postaux des données à géocoder sont chargés, pour augmenter la vitesse du traitement et soulager l'ordinateur. Néanmoins, la vitesse d'exécution par adresse suit une fonction inverse (1/x): `phacochr` est bien meilleur avec un nombre conséquent d'adresses. Ceci provient du fait que le "coût" marginal en temps de chargement des données est d'autant plus faible que les données sont nombreuses à géocoder.

Concernant la qualité du géocodage, nous avons réalisé des tests sur 18 bases de données réelles fournies par des collègues. `phacochr` possède une bonne capacité à trouver les adresses, puisque la médiane du pourcentage d'adresses trouvées est de 97%. Pour mesurer la fiabilité du package, nous avons mesuré la distance (euclidienne, en mètres) entre la géolocalisation opérée par `phacochr` avec ses réglages par défaut et les coordonnées spatiales déjà présentes dans deux bases de données. 97,6% des adresses géocodées sont localisées à moins de 100m de leurs coordonnées "réelles", montrant un degré de fiabilité tout à fait satisfaisant.

Développements futurs

`phacochr` est encore en phase de développement. La version disponible sur Github[2] est néanmoins pleinement fonctionnelle et a passé l'épreuve de multiples tests, au cours desquels des solutions ont été apportées aux problèmes posés par des structures de données diverses. La mise à disposition publique du package et sa présentation dans divers cénacles nous permettront de bénéficier de retours plus larges concernant des problèmes que nous n'aurions pas anticipés.

La présentation de `phacochr` lors des *Rencontres R* nous permettrait de soumettre la logique de notre code à des utilisateurs et développeurs avisés de R, dans un but d'amélioration de notre package. Divers éléments sont l'objet d'interrogations:

- Notre implémentation des données pour qu'elles puissent être mises à jour séparément du package est-elle optimale? Quelles solutions ont été apportées par d'autres développeurs face à cette question?
- La logique de chargement des données publique d'adresse est-elle optimale? Cette partie du code - déjà modifiée plusieurs fois - est l'un des goulots d'étranglement du traitement, et nous restons attentifs à des solutions alternatives plus performantes. La logique suivie implique notamment de mauvaises performances pour le géocodage d'un petit nombre d'adresses.
- Devons-nous diminuer les dépendances du packages? Par facilité de codage et pour une meilleure lisibilité du code, nous avons notamment utilisé quelques packages du `tidyverse`. Quelle implication dans la durée de ce type de dépendance?

[2] https://github.com/phacochr/