

Divvy Bike Project

Rémy Garnier - Camille Palmier
28 février 2018

I- Introduction

But et intérêts de l'étude

Dans le cadre du projet, nous avons construit un package R.

```
library("DivvyBikeProject")
```

Nous présenterons les fonctions présentes dans notre package au fur et à mesure du rapport.

Remerciements

II - Présentation des données

Le jeu de données

Divvy est le système de partage de vélos de la ville de Chicago, avec 6 000 vélos disponibles à plus de 570 stations à travers Chicago et Evanston. Divvy, comme les autres systèmes de vélos de libre-service, se compose d'une flotte de vélos qui sont enfermés dans un réseau de stations d'accueil dans toute la région. Les vélos peuvent être déverrouillés à partir d'une station et renvoyés à n'importe quelle autre station du système. Divvy offre aux résidents et aux visiteurs une option de transport pratique pour se déplacer et explorer Chicago.

Divvy est un programme du Département des Transports de Chicago (CDOT), qui possède les vélos, les stations et les véhicules de la ville. Le financement initial du programme provenait de subventions fédérales pour des projets qui favorisent le redressement économique, réduisent la congestion du trafic et améliorent la qualité de l'air, ainsi que des fonds supplémentaires provenant du programme de financement des augmentations d'impôt de la Ville. En 2016, Divvy s'est étendu à la banlieue voisine d'Evanston grâce à une subvention de l'État de l'Illinois.

Sur le site de Divvy, les données sur les trajets et les stations sont disponibles au public. Tous les ans, deux jeux de données sont publiés : un premier allant de janvier à juillet et un second allant de août à décembre.

Chaque trajet à vélo est anonymisé et comprend :

- Jour et heure du début du voyage
- Fin du voyage jour et heure
- Station de départ
- Station d'arrivée
- Type d'utilisateur (membre, tour unique ou passe d'exploration)
- Si un membre voyage, le sexe et l'année de naissance du membre seront inclus

Exemple de 5 trajets réalisés en 2014 :

```
##   trip_id      starttime      stoptime bikeid tripduration
## 1 2355134 6/30/2014 23:57 7/1/2014 0:07 2006          604
## 2 2355133 6/30/2014 23:56 7/1/2014 0:00 2217          263
## 3 2355130 6/30/2014 23:33 6/30/2014 23:35 2798          126
## 4 2355129 6/30/2014 23:26 7/1/2014 0:24 173          3481
## 5 2355128 6/30/2014 23:16 6/30/2014 23:26 173          638
##   from_station_id      from_station_name to_station_id
## 1          131 Lincoln Ave & Belmont Ave          303
## 2          282 Halsted St & Maxwell St          22
## 3          327 Sheffield Ave & Webster Ave          225
## 4          134 Peoria St & Jackson Blvd          194
## 5          320 Loomis St & Lexington St          134
##   to_station_name usertype gender birthyear
## 1 Broadway & Cornelia Ave Subscriber Male 1988
## 2 May St & Taylor St Subscriber Male 1992
## 3 Halsted St & Dickens Ave Subscriber Male 1993
## 4 State St & Wacker Dr Subscriber Female 1988
## 5 Peoria St & Jackson Blvd Subscriber Female 1988
```

Les 5 premières lignes de la table des stations de 2014 :

```
##   id      name latitude longitude dpcapacity
## 1 43 Michigan Ave & Washington St 41.88389 -87.62465          43
## 2 44 State St & Randolph St 41.88473 -87.62773          27
## 3 33 State St & Van Buren St 41.87718 -87.62784          27
## 4 199 Wabash Ave & Grand Ave 41.89174 -87.62694          15
## 5 51 Clark St & Randolph St 41.88458 -87.63189          31
##   online date
## 1 2013-06-16
## 2 2013-06-16
## 3 2013-06-25
## 4 2013-08-10
## 5 2013-06-17
```

Analyses descriptives des données

Table 1: Nombre de trajets et stations par année.

Année	Stations	Trajets
2013	300	759 788
2014	300	2 454 634
2015	474	3 183 439
2016	581	3 595 383
2017	585	3 829 014
		Total : 13 822 258

Les vélos ont été mis en service en juin 2013 d'où un nombre de trajets plus faible. Ces chiffres nous montre une augmentation du nombre de trajets sur le réseau et une augmentation du réseau lieu même au cours du temps. C'est entre 2014 et 2015 que ces augmentations sont les plus importantes avec environ 30% de trajets supplémentaires et 174 nouvelles stations.

Visualisation géographique

Pour visualiser l'emplacement des stations à Chicago, nous les avons tracés sur un fond de carte de la ville.

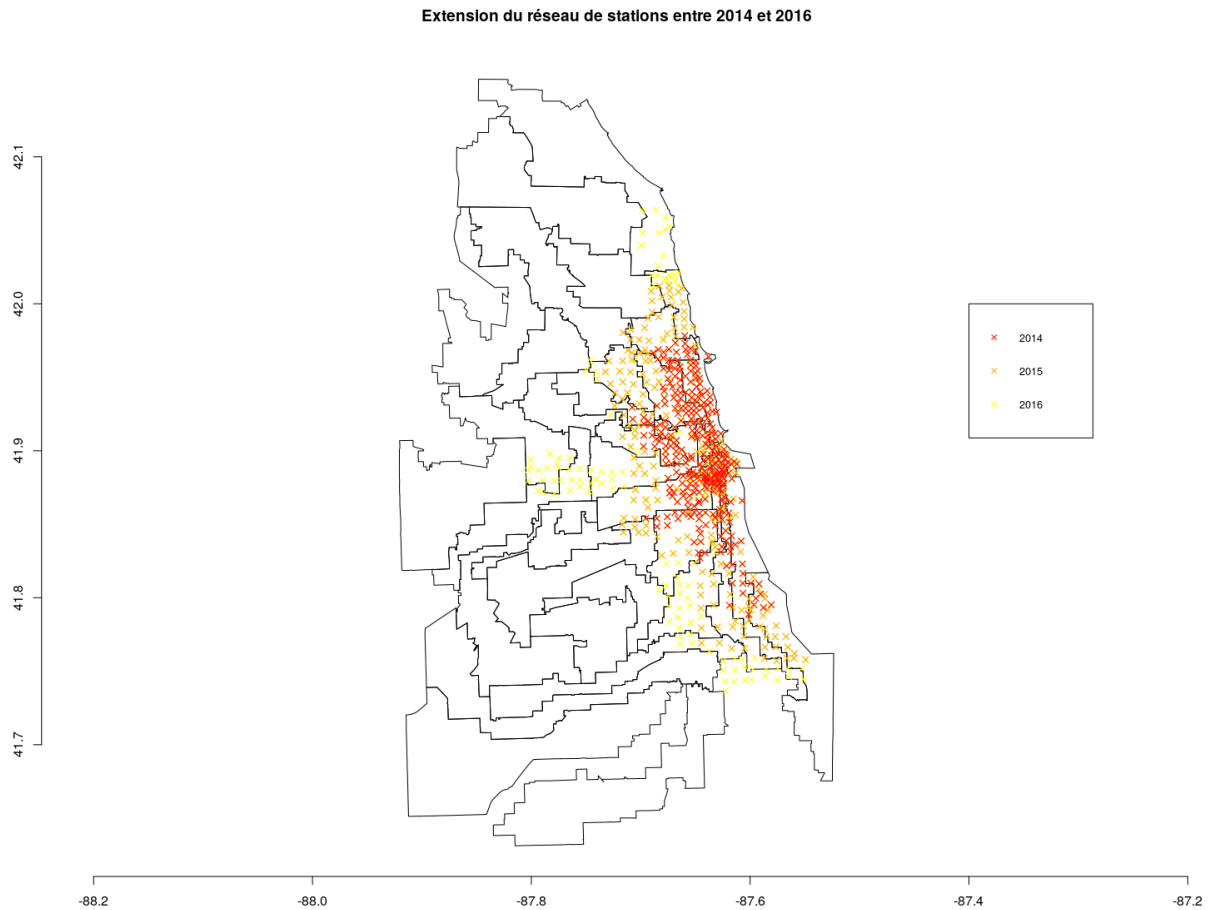


Figure 1:

Les stations vélos ont d'abord été réparti au centre de la ville avant de s'étendre vers la banlieue. Le réseau en 2017 est sensiblement le même que le réseau de 2016 (seulement 4 stations de plus).

Autres représentations

Si on regarde la fréquentation des stations en 2016, on remarque que les vélos les plus utilisés sont ceux du centre ville. Par ailleurs, il y a un fort écart entre les stations les moins utilisées (moins de 200 vélos empruntés par an) et les stations les plus utilisées (jusqu'à 92 000 vélos empruntés).

III - Transformation des données

Nous présenterons dans cette partie les manipulations que nous avons effectué sur les données. Cette partie est centrale puisqu'elle a nécessité beaucoup de travail et de temps.

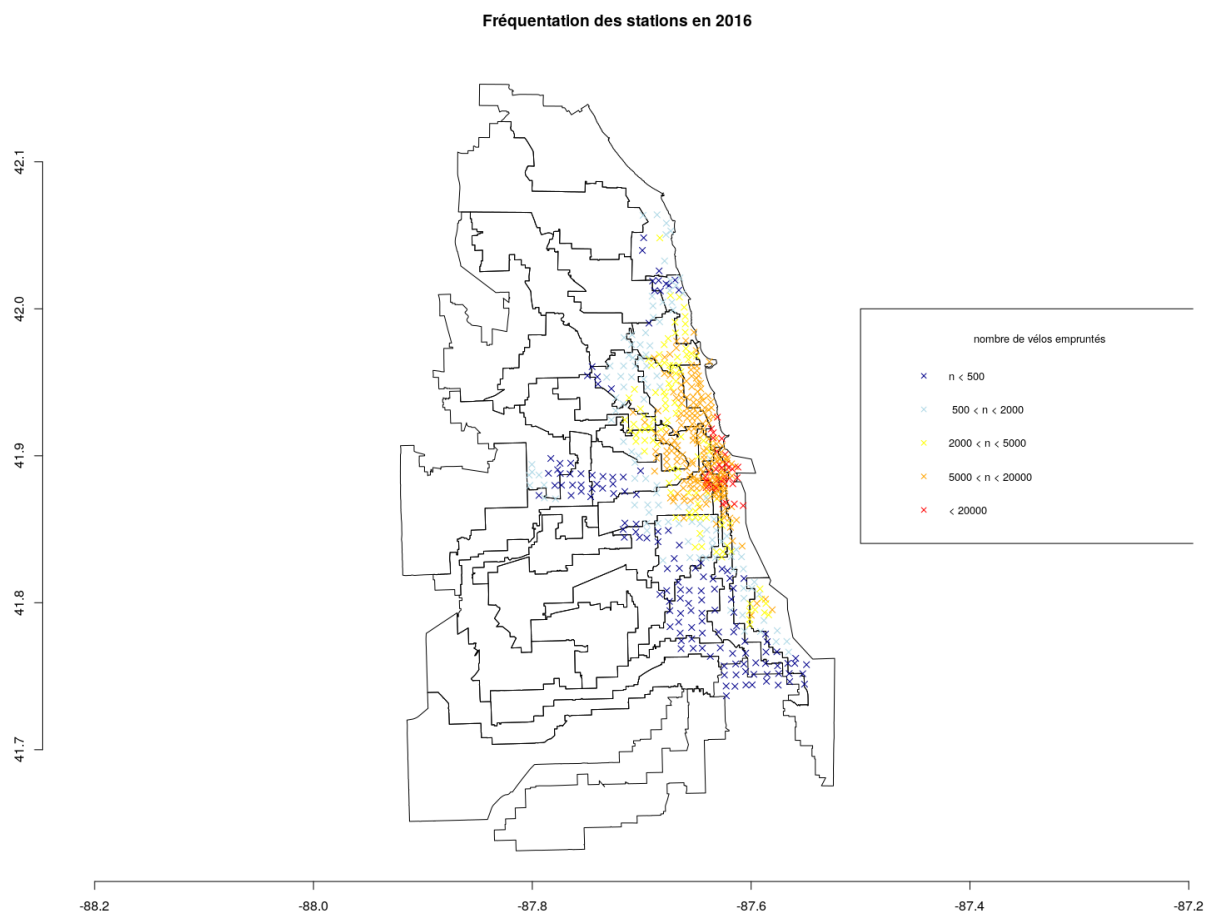


Figure 2: h

Les données que l'on trouve sur le site de Divvy sont divisées par semestre, ou par mois selon les années. Les données n'ont pas un format homogène. Les variables changent parfois de nom et le format des données n'est pas toujours le même. On a également été confronté à un problème au niveau des stations : elles peuvent changer de noms au cours du temps ou se déplacer de quelques mètres.

Notre jeu de données vit en grande dimension (plus de 13 millions de trajets et une dizaine de variables). Pour disposer de la capacité nécessaire pour faire tourner les modèles, nous avons réduit la dimension en réalisant plusieurs agrégations.

Agrégation temporelle

Nous avons compté, pour chaque station, le nombre de vélos entrant (nbE) et le nombre de vélos sortant (nbS) à chaque heure. Nous avons donc perdu l'information sur les trajets, puisque qu'on ne sait plus d'où vient chaque vélo. Par ailleurs, nous avons négligé les informations sur les utilisateurs.

L'agrégation temporelle (ainsi que la standardisation du format des données) se fait en utilisant la fonction **aggregateData** du package.

Agrégation spatiale

Le fait que de nouvelles stations s'ouvrent constamment, ainsi que le fait qu'elles sont parfois déplacées, nous a conduit à grouper les données géographiquement. Ainsi, nous supposons que des stations spatialement proche se comportent de la même façon.

Pour gérer les données géographiques, nous nous avons utilisé les packages R *rgdal* et *raster*. Etant donné un découpage spatial de la ville de Chicago contenue dans le Large SpatialPolygonDataFrame *area*, ainsi qu'un fichier *data* regroupant les données agrégées temporellement sur les *stations*, on obtient une agrégation spatiale via la fonction **groupByGeospatialData(data, area, stations, year, save_data = FALSE)**

Pour nos tests, nous avons choisi le découpage géographique par circonscriptions législatives pour la chambre basse de l'Illinois, obtenue sur le site du recensement américain. Ce découpage n'est objectivement pas très bon, notamment à cause de la pratique du gerrymandering. En effet, la forme des circonscriptions est généralement tortueuse (voir carte plus haut) et regroupe des populations différentes pour des considérations électorales. Cependant, il s'agissait du seul regroupement à la bonne échelle que nous avons pu trouver. Par ailleurs, comme on le verra, ce découpage est déjà assez informatif.

On peut ainsi regrouper les stations dans une vingtaine de regions géographiques différentes qui comportent chacune en moyenne 24 stations (avec un écart type de 18).

Ajouts de variables explicatives

Nous avons rajouté 3 variables explicatives à notre jeu de données :

- La pluviométrie (*pluvio*) et la température (*temp*) grâce au package *riem*.
- La variable "Day of Week" (*dow*) qui varie de 1 à 8 : lundi = 1, ..., vendredi = 7 et 8 pour les jours fériés.

Ces variables sont ajoutés respectivement par les fonctions **getMeteo** et **getSpecialDays**.

L'ensembles des manipulations décrites dans les trois derniers paragraphes ont été regroupées dans la fonction **BuildDataSet**.

Voici un extrait de nos données finales :

```
##           Time district      Day Hour nbE nbS      diff
## 1 2014-12-31 23:00:00      2 2015-01-01    0    2    1 0.03333333
```

```
## 2 2014-12-31 23:00:00      3 2015-01-01    0  0  0  0.00000000
## 3 2014-12-31 23:00:00      4 2015-01-01    0  0  0  0.00000000
## 4 2014-12-31 23:00:00      5 2015-01-01    0  0  3 -0.06521739
## 5 2014-12-31 23:00:00      6 2015-01-01    0  1  1  0.00000000
##  nb_stations dow  weekday    temp pluvio
## 1           30   8 NewYears 5.78629      0
## 2            1   8 NewYears 5.78629      0
## 3           24   8 NewYears 5.78629      0
## 4           46   8 NewYears 5.78629      0
## 5           44   8 NewYears 5.78629      0
```

Définition de notre série d'intérêt

Nous avons proposé 3 série d'interet différentes par région:

- **nbEstat** : le nombre moyen de vélos par station entrant dans la région à une heure donnée par
- **nbSstat** : le nombre moyen de vélos par station sortant de la région à chaque
- **diff** = nbEstat - nbStat : la différence entre le nombre de vélos entrant et sortant de chaque région par heure et par station

Nous estimons que cette dernière statistique est la plus utile, car elle peut servir à l'entreprise pour determiner comment elle doit déplacer les vélos entre les différentes régions afin de s'assurer que les stations ne soient jamais complètement vides ou complètement pleines.

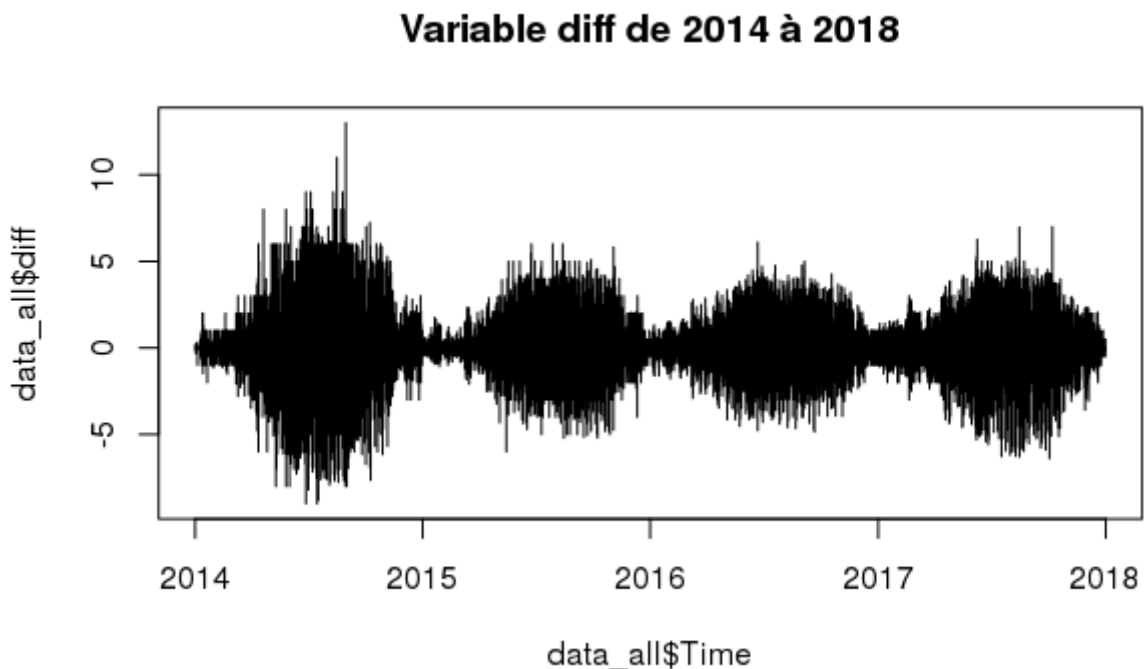


Figure 3:

Ce premier graphique permet de visualisation la variable représentant la différence entre le nombre de vélos entrant et le nombre de vélos sortant. La forme générale est la même : les vélos sont moins empruntés en

début d'année (période hivernale - janvier/février). C'est en été que la différence entrant/sortant est la plus importante.

Pour bien comprendre la forme de la série, intéressons nous à ces trois séries sur une période de temps restreinte : soit la deuxième semaine du mois de *juillet 2014*.

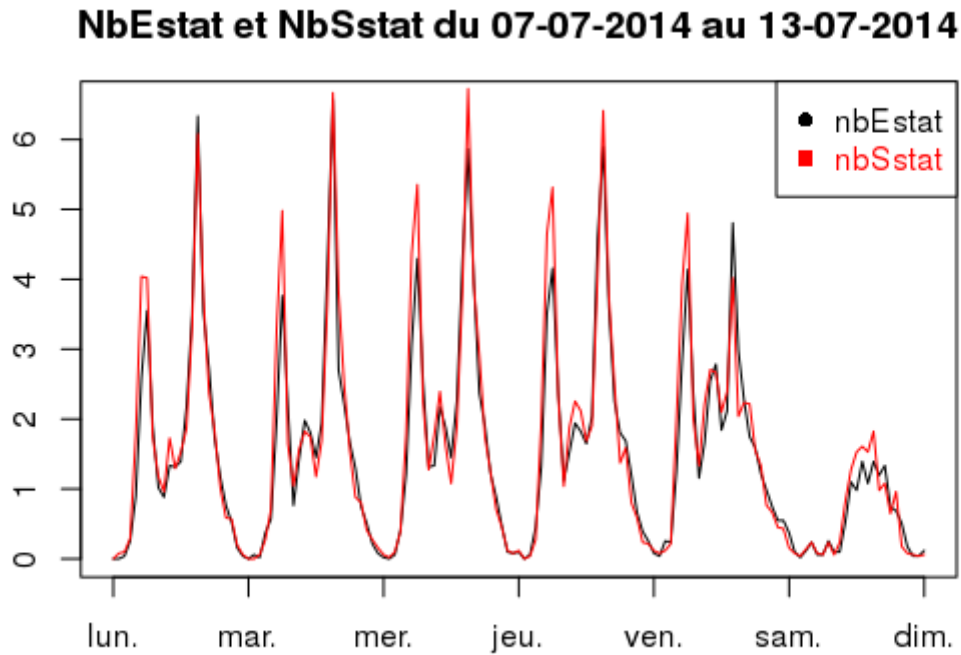


Figure 4:

Quelques visualisations

IV - Les modèles

Tendance / Saisonalité

Modèle GAM (GAMLSS) - significativité des variables

Random Forests (Adaboost)

Modèles auto-régressifs

Agrégation

V - Conclusion

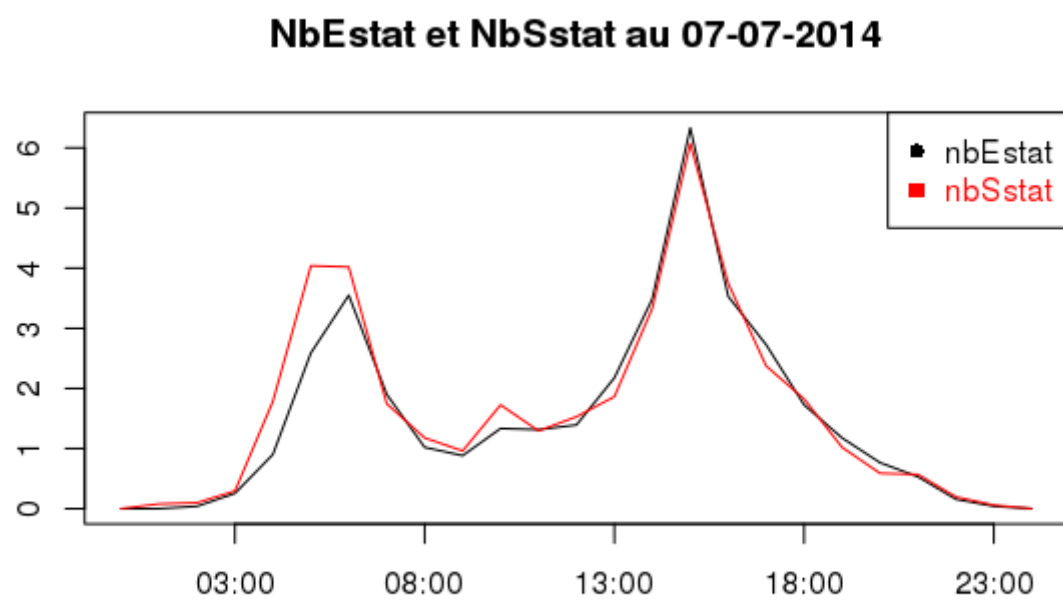


Figure 5: