Rapport projet R

Jakob Maier, Guillaume Lambert

25/12/2021

Contents

0.1 Le jeu de données

0.1.1 Présentation des données

On présente maintenant les données que l'on a collectées et traitées afin de les utiliser dans la tâche de prédiction.

Données de comptage routier

Les données de comptage routier sont disponibles sur le site **Paris Data** ("Paris Data," n.d.), regroupant les jeux de données de la Ville de Paris, en cliquant ici. Elles sont collectées grâce à des boucles électromagnétiques implantées dans la chaussée sur plus de 3000 tronçons de voies. L'historique des données s'étend de 2014 à ???? au pas horaire, selon la variable **t_1h**. Deux types de données sont fournies : le **taux d'occupation** qui correspond au temps de présence de véhicules sur la boucle en pourcentage d'une heure et le **débit** qui est le nombre de véhicules ayant passé le point de comptage pendant une heure.

Pour une année, il y a environ 29 millions observations réparties sur plus de 3000 points. On a donc procédé à une transformation des données afin de réduire leur nombre et les truscturer. La première étape est l'agrégation des points d'observations selon les libellés (nom de rue) associés dans le jeu de données. Ensuite, on a déterminé les principaux axes de Paris en moyennant le nombre de voitures par heure des points d'observations partageant un même libellé puis en choisissant les 200 premières valeurs (hors périphérique). On obtient le graphique 1 et on en déduit le graphe simplifié 2 de Paris composé de 69 arêtes, correspondant à 69 jeux de données.

Figure 1: Représentation en rouge des points d'observation associés aux 200 libellés les plus fréquentés (hors périphérique)

On obtient finalement 2 variables au pas horaire : **nbCar** (nom des variables) le nombre de voitures et **rateCar** le taux d'occupation. A partir de ces 2 variables, on en construit des autres en les retardant, d'une semaine (**nbCarLaggedWeek**, **rateCarLaggedWeek**), d'un jour (**nbCarLaggedDay**, **rateCarLaggedDay**) et d'une heure (**nbCarLaggedHour**, **rateCarLaggedHour**). En effet, il semble pertinent

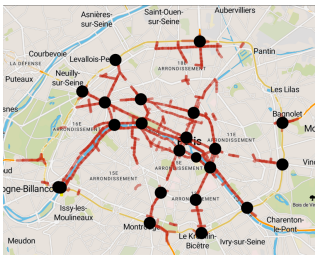

Figure 2: Graphe simplifié de Paris

d'observer l'état du traffic routier à des temps antérieurs : historique à très court terme de l'eure précédente et historique au même instant du cycle journalier et hebdomadaire précédent. Cela est confirmé dans la partie ????.

Enfin, il est très important de noter que ces données sont incomplètes. Une étude approfondie de cette incomplétude et de la complétion est faite dans la partie ????.

0.1.1.1 Variables temporelles

Le traffic routier étant relié à l'activité humaine, nous avons ajouté de nombreuses variables temporelles (notamment grâce à la librairie *lubridate*).

- **year**, **month**, **day**, **hour** en décomposant la variable **t_1h**
- **time** le numéro de l'observatio
- **toy** de 0 à 1 selon la position de l'observation dans l'année en cours
- **weekdays** le jour de la semaine et **weekendsIndicator** l'indicatrice si le jour est un jour du weekend
- **winterHolidaysIndicator** et **summerHolidaysIndicator** les indicatrices des vacances d'hiver et d'été définies à partir de ("Les Archives Du Calendrier Scolaire," n.d.)
- **bankHolidaysIndicator**

0.1.1.2 Index de la situation sanitaire en rapport avec le covid

0.1.1.3 Météo

0.1.2 Complétion du jeux de données

text

"Les Archives Du Calendrier Scolaire." n.d. https://www.education.gouv.fr/les-archives-du-calendrier-scolaire-12449.

"Paris Data." n.d. https://parisdata.opendatasoft.com/page/home/.