# Estimation en temps réel des temps de trajet dans un réseau de bus à l'aide de données historiques

Nikita Marchant[1]

[1]Université Libre de Bruxelles, Département d'Informatique
nimarcha@ulb.ac.be

**Abstract**

## Introduction

Le but de ce projet est de prédire, en temps réel, l'heure d'arrivée aux arrêts d'un véhicule de transport en commun. Dans le cadre de ce travail, la prédiction sera effectuée pour des lignes de bus, tram et métro de la STIB[1].

Un réseau de transport en commun étant très difficile à modéliser à cause de sa complexité et par ce qu'il est très influencable par des événements stochastiques, l'approche qui sera utilisée ici sera non pas de modéliser le réseau pour prédire son état futur mais d'extrapoler les trajets des ses véhicules grâce à des données historiques récoltées au préalable.

## Métriques de performance

Pour pouvoir comparer plusieurs méthodes, il est important d'avoir des métriques bien définies mesurant la performance de celles-ci.

La littérature utilise souvent le RMSE[2]: l'erreur quadratique moyenne.

Cependant, cette métrique souffre d'un problème : qu'un bus soit annoncé une minute à l'avance ou une minute en retard compte de la même manière alors que dans un cas l'usager attend son bus une minute de plus et que dans l'autre il le rate.

*J'ajouterai ici une métrique qui résoud ce problème.*

## Algorithmes des prédiction

Plusieurs algorithmes seront implémentés et leur performances seront comparées.

### Méthodes naïves

Deux méthodes assez naïves seront implémentées pour servir de base pour comparer les méthodes proposées par la suite. La première, la plus simple, est de prédire tout le temps la même durée de trajet entre deux arrêts en utilisant simplement la durée spécifiée dans les horaires statiques.

La seconde, celle utilisée par la STIB, est de prédire le temps de trajet entre deux arrêts comme étant la moyenne des temps de trajets des trois derniers véhicules de la ligne étant passés sur ce tronçon.

### $k$ plus proches voisins

L'algorithme des $k$ plus proches voisins[3] est une méthode d'intelligence artificielle qui peut être utilisée aussi bien pour de la classification que pour de la régression. L'idée de cette méthode est de trouver les $k$ trajets les plus similaires à la cible et de les utiliser pour extrapoler le temps de trajet futur.

Pour cela, on projette chaque trajet dans un espace à $n$ dimensions avec $n+1$ étant le nombre d'arrêts déjà effectués par le véhicule dont on cherche à prédire le temps de trajet (ce véhicule sera appelé $\alpha$).

Les trajets sont donc représentés par le vecteur colonne $T_\alpha = \left(d_{1,\beta}, d_{2,\beta}, ..., d_{n,\beta}\right)^T$ avec $d_{i,\beta}$ étant la durée du trajet entre l'arrêt $i$ et $i+1$ pour le véhicule $\beta$. La similarité $s_{\alpha,\beta}$ entre deux trajets $\alpha$ et $\beta$ est définie comme la distance Euclidienne entre deux vecteurs :

$$s_{\alpha,\beta} = \sqrt{\sum_{j=0}^{n}(d_{j,\alpha} - d_{j,\beta})^2} \tag{1}$$

La prédiction du temps de trajet $\hat{d}_{n+1,\alpha}$ est donnée par la moyenne des temps de trajet pour ces arrêts des $k$ plus proches voisins de $\alpha$ dans l'espace.

$$\hat{d}_{n+1,\alpha} = \frac{1}{k}\sum_{j=0}^{k} d_{n+1,k} \tag{2}$$

[1]Société de transports en commun à Bruxelles (Belgique)

[2]Root Mean Squared Error

[3]$k$-nearest neighbors ou encore $k$-NN. (Hastie et al. (2009))

**Pondération :** Il serait peut-être appréciable que les performances récentes d'un véhicule influencent plus la prédiction que ses performances plus anciennes. Pour cela, on peut introduire la notion d'une distance pondérée dans laquelle certaines dimensions subiraient un homothétie pour en augmenter au diminuer le poids.

*J'ajouterai ici la définition de cette distance*

**Prise en compte d'autres critères que le temps de trajet :** Il pourrait être utile de prendre en compte d'autres critères que le temps de trajet pour déterminer la similitude entre deux trajets. Le jour de la semaine, la météo ou même l'heure sont des facteurs influençant le temps de trajet. L'heure étant une variable continue, nous pouvons simplement l'ajouter au vecteur $T_\alpha$ et utiliser ce vecteur de $n + 1$ éléments pour la distance euclidienne. Pour ce qui est du jour de la semaine ou de la météo, ce sont des variable discrètes pour lesquelles il faudra définir une nouvelle fonction de distance.

*Il faut encore que je fasse des recherches sur les k-NN avec variables discrètes si je m'aventure dans cette direction là.*

## Ensemble d'apprentissage

Pour constituer l'ensemble d'apprentissage, il a fallu enregistrer la position de l'ensemble des véhicules du réseau de la STIB toutes les 20 secondes pendant plusieurs mois ce qui resprésente approximativement un demi million de mesures par ligne et par sens.

Ces données ont une faible résolution temporelle (20 secondes) et spatiale (la position des véhicules est connue à un arrêt près) et sont parfois de mauvaise qualité (certains véhicules disparaissent pendant quelques minutes pour réapparaitre plus loin par exemple). De plus, il n'a pas été possible de récupérer un identifiant en plus de la position pour les véhicules ce qui rend l'identification de trajets non trivialle.

Il a donc fallu traiter ces données avant de pouvoir les utiliser pour effectuer des prédictions. *Explication rapide de la méthode*

Une fois ce traitement effectué, il reste un dernier détail à prendre en compte : les trajets extraits ne commencent ou ne finissent pas tous au terminus (soit parce que le véhicule n'as pas été jusqu'au bout ou à cause d'erreurs commises lors de l'extraction des trajets depuis les positions).

Les méthodes présentées ci-dessus ne fonctionnant pas avec des vecteurs contenant des éléments non définis, il y avait deux possibilités : soit ignorer les vecteurs incomplets soit extrapoler les valeurs manquantes.

La première méthode a été écartée car elle diminuait fortement la taille de l'ensemble d'apprentissage, la seconde a été privilégiée. Si le $i$ème élément d'un vecteur est manquant, il est remplacé par la moyenne des $i$èmes éléments des vecteurs dont la valeur est définie.

## Améliorations possibles

- Distance entre variables discrètes
- Meilleure imputation des valeurs manquantes

## References

Hastie, T. J., Tibshirani, R. J., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer.