12 JUIN 2015

PROJET MEDIAMOBILE

PREVISIONS DE TRAFIC

HUGO MORIN ET MOHAMED YASSINE LANDA

Tuteurs : Thierry Gruber
Nicolas Durrande
Xavier Bay

Sommaire

I – Contexte Général

Mediamobile est l'un des premiers opérateurs de services d'information trafic et de mobilité en temps réel en Europe. Mediamobile compte parmi ses clients plus de 20 constructeurs automobiles qui intègrent les services d'information trafic dans leurs systèmes de navigation embarqués, ainsi que des fabricants de systèmes de navigation autonomes (PND). Les services de Mediamobile sont également proposés par les opérateurs de téléphonie mobile et les médias.

Les enjeux liés à la mobilité sont aujourd'hui au cœur des préoccupations des entreprises et des particuliers. C'est sur ce créneau d'amélioration de la mobilité que Mediamobile, grâce à ses solutions V-Traffic, se positionne. V-Traffic permet d'emmener les utilisateurs à destination, de la manière la plus sûre, la plus rapide et la plus économique.

Les solutions V-Traffic s'occupe également de certains aspects sociétaux en permettant par exemple aux administrations de lutter efficacement contre l'impact environnemental et le gaspillage énergétique, aux entreprises d'être plus compétitives, et aux individus de retrouver leur liberté de mouvement, en circulant en toute sécurité.

Tuteurs : Thierry Gruber
Nicolas Durrande
Xavier Bay

II – Présentation du projet

La force de Mediamobile est de pouvoir fournir l'information la plus précise du marché, à partir de collectes d'information effectuées auprès de millions d'usagers sur la route. Pour être compétitif et minimiser le nombre d'erreurs, ils ont besoin de détecter automatiquement les fermetures de routes.

Or il s'agit d'un problème complexe car une absence de données n'est pas équivalente à une fermeture de routes (il peut s'agir d'une coupure, d'un problème FCD,…). Pour le résoudre l'utilisation de la loi de Poisson est envisagée, on définit la probabilité qu'aucun véhicule ne passe par: $e^{-\int q_i(t)dt}$.

Le but de notre projet réside dans le fait de trouver cette fonction qi(t), la fonction de débit de véhicule en conditions normales pour chaque arc i. Il s'agit donc de dégager des informations pertinentes du dataset, de faire du clustering de courbes pour ainsi être capable de mapper un nouveau (arc,jour) à un cluster et donc d'en prédire sa courbe de débit.

III – Formalisation du problème

1- Données disponibles et traitement :

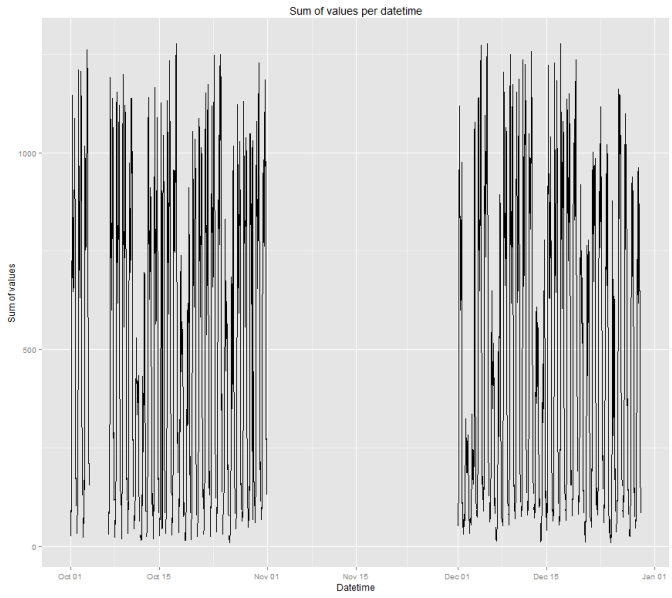
Les données fournies brutes sur lesquelles il fallait travailler étaient en format csv. Elles représentaient le nombre de voitures sur la route sur des plages horaires de 15 minutes pour chaque segment de la route et cela pour chaque jour des mois d'Octobre, Novembre et Décembre.

A partir de ces données le but est de repérer des tendances particulières ou des changements significatifs qui pourront expliquer/prédire certain phénomènes de trafic.

Pour faire cela, une première étape de nettoyage de données était nécessaire. Suppression de toutes les données du mois de Novembre ainsi que le 30 et 31 du mois de Décembre ainsi que des deux premières heures pour tous les couples arc-jour pour cause de leurs défaillances (absence de données…).

Une augmentation des données était aussi nécessaire pour faciliter leurs classifications. Principalement des colonnes « Jour de la semaine» ainsi que les jours de vacance.

Tuteurs : Thierry Gruber
Nicolas Durrande
Xavier Bay



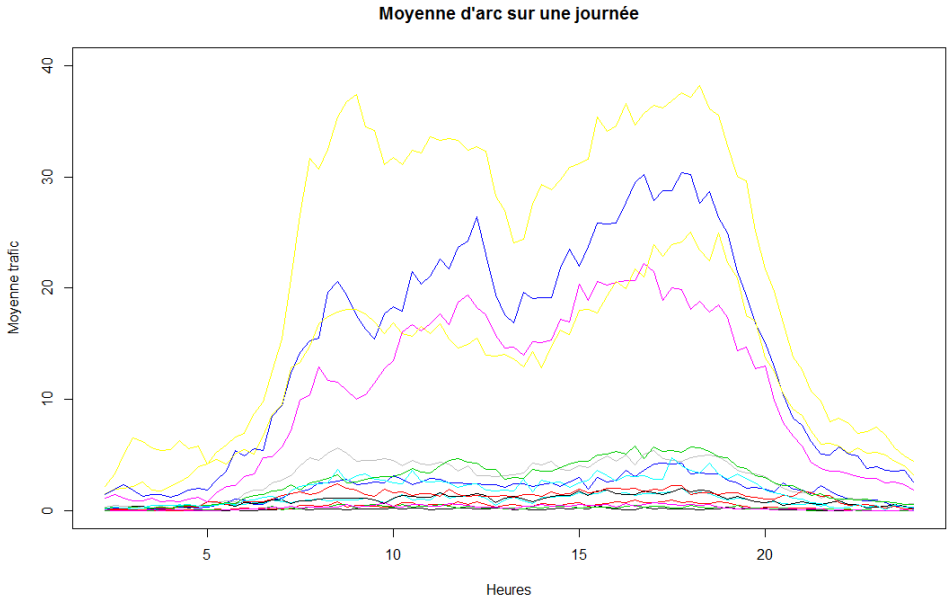
Les valeurs après traitement

2- Méthodologie :

On a divisé l'ensemble de données en deux sous-ensembles : un ensemble d'apprentissage et un ensemble de test. Les méthodes statistiques simples et le 'clustering' seront fait sur l'ensemble d'apprentissage et l'ensemble test servira à confirmer les résultats.

L'ensemble test fera 80% de l'ensemble de données en mains. La construction de cet ensemble se fait de façons aléatoire.

Une première vue sur l'ensemble d'apprentissage (en dessinant les graphiques de moyenne par jour pour les premiers arcs de cet ensemble) nous montre qu'il y a une grande disparité entre les arcs et les volumes de voitures qui y circulent.



Tuteurs : Thierry Gruber
Nicolas Durrande
Xavier Bay

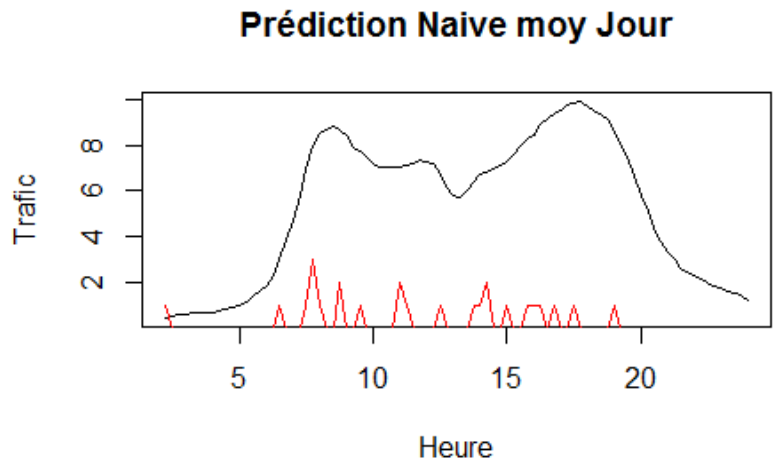
IV - Méthodes de prédiction naïves :

Nous allons essayer 3 méthodes de prédictions naïves et allons les tester à chaque fois avec une même ligne de notre ensemble test (Le trafic réel en rouge)

.

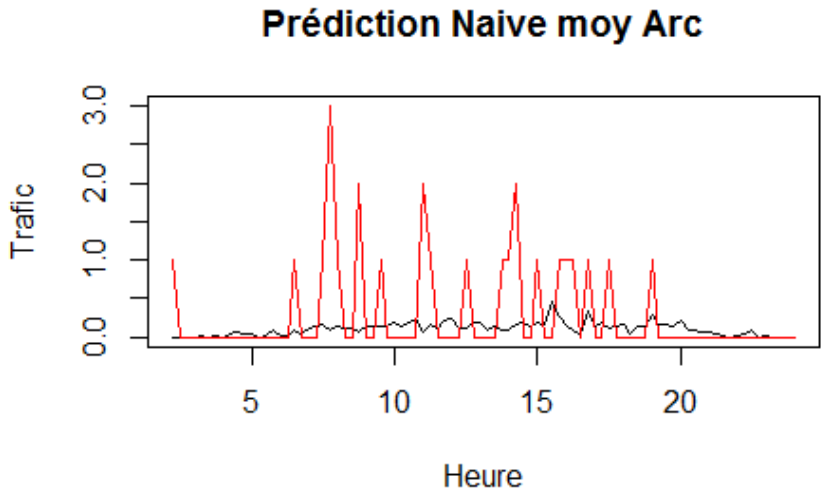
1- Prédiction par jour de la semaine :

En se basant sur le calcul de la moyenne sur tous les arcs de l'échantillon apprentissage pour le jour de l'arc testé :



2- Prédiction par moyenne de l'arc :

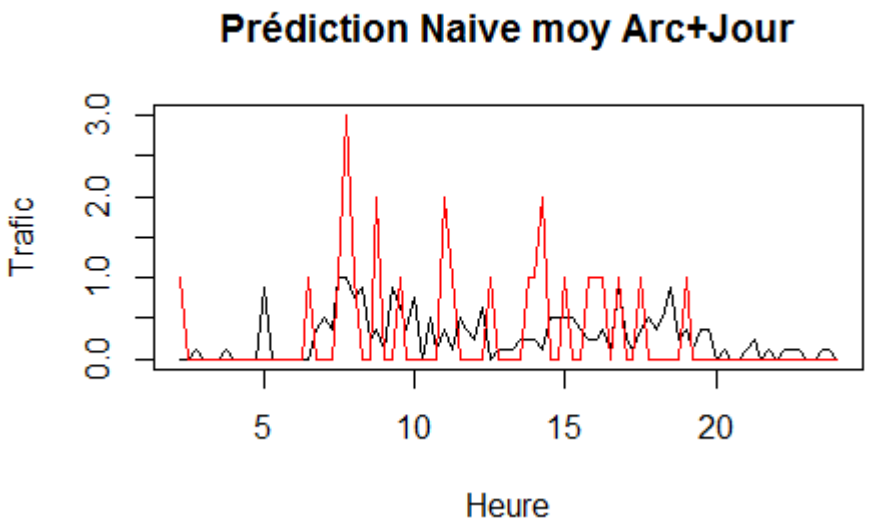
En se basant sur la moyenne de l'arc sur tous les jours :



Tuteurs : Thierry Gruber
Nicolas Durrande
Xavier Bay

3- Prédiction par moyenne d'arc-jour :

En se basant sur la moyenne de l'arc mais seulement en prenant considérant les jours de l'arc à prédire :



L'erreur (distance euclidienne) faite pour chacune des méthodes :

Méthode	Moy Jour	Moy Arc	Moy Arc + Jour
Erreur	5.888014	0.6042104	0.5665887

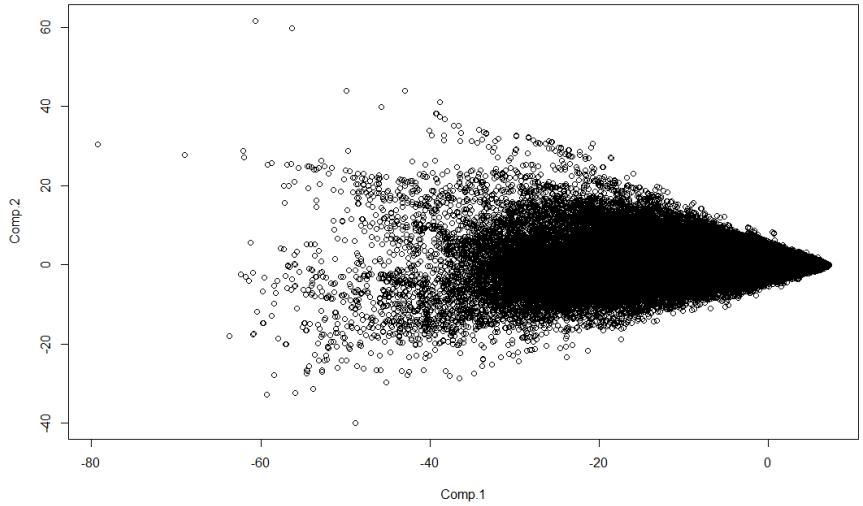
On voit bien que les méthodes qui se basent sur l'arc ont plus proche de la réalité que la première méthode.

Tuteurs : Thierry Gruber
Nicolas Durrande
Xavier Bay

V - Méthodes de prédiction avancées :

Pour cette partie nous allons utiliser des méthodes de classifications non supervisées pour diviser notre ensemble en différents clusters. Mais pour définir le nombre de cluster, puisqu'il doit être donné à au préalable à la fonction de classification, nous avons procéder à une PCA de nos données d'apprentissage pour voir leurs répartition.

1- PCA sur l'ensemble d'apprentissage :



Vu cette répartition, nous avons décidé de subdiviser le groupe en 4 clusters. Mais avant cela l'ensemble d'apprentissage devait être réduit considérablement (vers les 7500 observations) pour que les calculs puissent se faire sur la mémoire de la machine.

2- Classification Hierarchical Clustering :

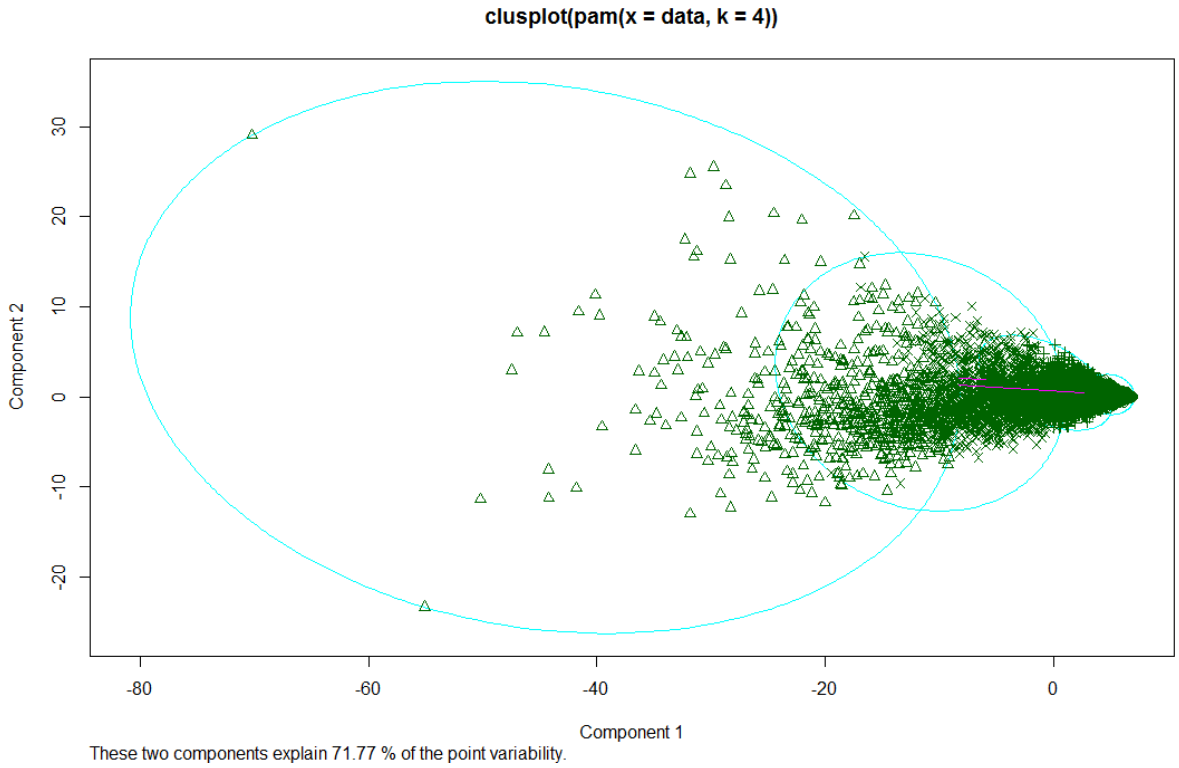
Avec l'utilisation de la fonction 'hclust' avec 'ave' (moyenne) et après la subdivation en 4 groupes on obtient:

Groupes	1	2	3	4
Nombre d'arc-jour	6795	202	3	1

Tuteurs : Thierry Gruber
Nicolas Durrande
Xavier Bay

3- Classification Partition Around Menoids :

Sur cette image là on peut observer les 4 clusters obtenus, et on voit que sur cette base on explique plus de70% des données.



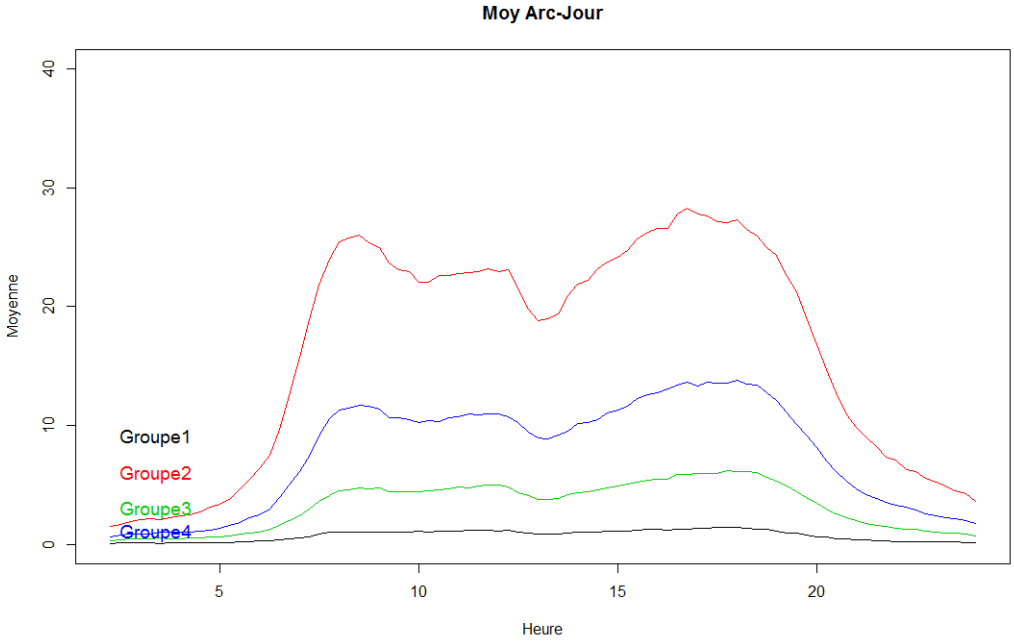
Une comparaison entre les groupes générés par les deux méthodes s'impose :

\Pam Hclust\	Groupe 1	Groupe 2	Groupe 3	Groupe 4
Groupe 1	2569	647	2044	1724
Groupe 2	0	13	0	0
Groupe 3	0	2	0	0
Groupe 4	0	1	0	0

On voit bien qu'avec la méthode PAM on obtient une meilleure répartition. C'est avec ces clusters là que nous allons travailler.

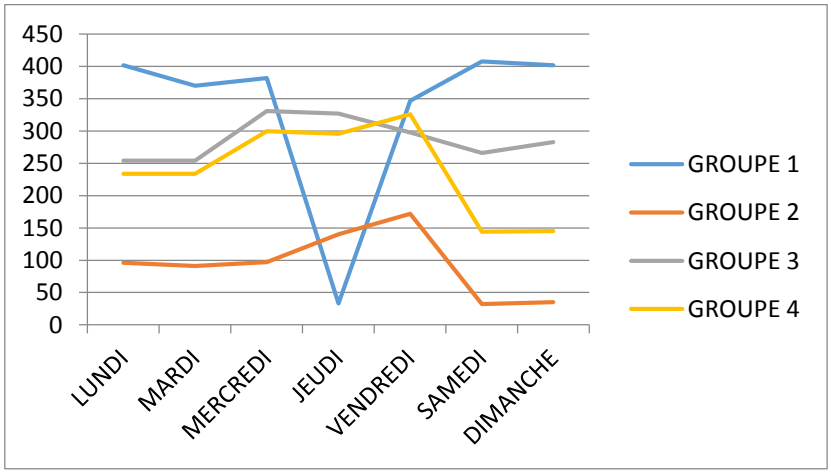
Tuteurs : Thierry Gruber
Nicolas Durrande
Xavier Bay

Pour analyser ces groupes, nous nous sommes penchés en premiers sur leurs moyennes:



On remarque que les 4 groupes ont presque la même forme courbe (même si elle plus « creuse » pour le groupe 2) mais que la différence entre les densités de trafic est très prononcée.

Une analyse par nombre d'arc par jour de semaine dans les différents groupes. En y le nombre d'arc-jour, donc l'interprétation d'une courbe doit porter plus sur sa forme que sur son positionnement par rapport aux autres courbes:



L'analyse pour voir l'incidence des jours de vacance sur les groupes ne donnant rien (même répartition sur les 4 groupes) laissent à croire que les groupes obtenus représentent une classification physique et non temporelle de certaines catégories d'arcs.

Une fonction de prévisions permet ensuite d'affecter chaque arc-jour de l'ensemble test à un cluster en calculant sa distance avec les 4 Medoids des clusters et qui décide de l'affecter ou pas.

Tuteurs : Thierry Gruber
Nicolas Durrande
Xavier Bay

VI – Conclusion

Pour résoudre ce problème et réussir à prévoir la densité de véhicules sur chaque arc à chaque jour de la semaine nous avons commencé par des méthodes naïves qui déjà nous donnent des résultats assez satisfaisants (prévision par moyenne d'arc-jour) mais encore insuffisants. Nous avons ensuite tenté d'améliorer ces résultats, grâce à une technique de clustering. La création de nos groupes semble satisfaisante, nous n'avons malheureusement pas été en mesure de calculer précisément l'efficacité de cette technique.

Même si nous n'avons pas pu apporter une réponse idéale à Médiamobile avec notre technique avancée, nous espérons leur avoir offert des pistes de recherches intéressantes. Ce projet fut pour nous fort intéressant, dans l'amélioration de nos compétences en R et la découverte de nouvelle technique de prédiction.

Annexe :

Une analyse par nombre d'arc par jours dans les différents groupes :

GROUPES	1	2	3	4
LUNDI	402	96	254	234
MARDI	370	91	254	234
MERCREDI	382	97	331	300
JEUDI	33	140	327	296
VENDREDI	347	172	298	326
SAMEDI	408	32	266	144
DIMANCHE	402	35	283	145
TOTAL	2344	663	2013	1679

Code fonction prévision pam :

```
prevision_pam <- function(df) {
  nbre.test <- nrow(df)
  previsions <- vector(mode="numeric", length=nbre.test)
  for( j in 1:nbre.test){
    cat('boucle row ',j,'¥n')
    df.data <- get_data(df)
    distance<- c(0,0,0,0)
    for( k in 1:4) {
      medoid <- dff.train.pam$medoids[k,]
      for( i in 1:88 ) {
        somme <- (df.data[j,i]-medoid[i])*(df.data[j,i]-medoid[i])

        distance[k]<- distance[k] + somme
      }
      distance[k]<- sqrt(distance[k])
    }
    distance.min <- min(distance, na.rm = FALSE)
    if( distance.min < 50 ) {
      for( k in 1:4) {
        if (distance.min == distance[k] ) {
          previsions[j] <- k
          cat('previsions[',j,'] <- ',k,' ¥n')
          break
        }else{
          previsions[j] <- 0
        }
      }
    }
  }
```


Tuteurs : Thierry Gruber
Nicolas Durrande
Xavier Bay

MINES
Saint-Étienne