Christopher Belinguier
Manon Charvet
Yrieix Leprince
Rémi Sanchez
Thai-An Than Trong

**TP Géolocalisation**
INF 730
Date de rendu : 22/12/17

## Répartition des données en train-val-test

Le fichier des variables explicatives (mess_train_list.csv) a été réparti en une partie *train* et une partie *val* afin de valider les modèles testés. La partie *val* est composée de message émis par des objets différents de ceux présent dans le *train*. De plus, la distribution des variables $rssi$, $Nseq$ et $Bsid$ dans *val* est similaire à celle de *test* afin d'évaluer le plus précisement possible les performances de notre modele sur le test. On obtient donc 3 échantillons différents :

— $X_{train}|y_{train}$ : pour le training
— $X_{val}|y_{val}$ : pour la validation
— $X_{test}$ : pour l'évaluation finale

## Création de la matrice des features

La matrice initiale a tout d'abord été reprise pour effectuer des tests. Cette matrice consistait à affecter, pour chaque message, désigné par un index, la valeur 1 aux colonnes des stations ayant reçu un signal, et 0 sinon.

Elle a ensuite été enrichie avec d'autres variables explicatives telles que le RSSI, la latitude et la longitude. Dans le cadre du projet, plusieurs modèles de matrices ont été explorés.

| index | $bsid_1$ | $bsid_2$ | ... | $bsid_p$ |
|---|---|---|---|---|
| 0 | 0 | 1 | ... | 0 |
| 1 | 1 | 0 | ... | 0 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| n | 1 | 0 | ... | 1 |

TABLE 1 – Matrice initiale

Pour la matrice ci-dessous, les données de puissance du signal ont été rajoutées en multipliant les variables catégorielle par la donnée de RSSI.

| index | $bsid_1$ | $bsid_2$ | ... | $bsid_p$ |
|---|---|---|---|---|
| 0 | 0 | $rssi_{0,0}$ | ... | 0 |
| 1 | $rssi_{1,1}$ | 0 | ... | 0 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| n | $rssi_{n,1}$ | 0 | ... | $rssi_{n,x}$ |

TABLE 2 – La matrice prenant en compte le rssi

Pour la matrice suivante, les données de latitude et de longitude des stations ont été introduite. Les termes suivants ont été définis :

$$\forall (i,j) \in [0,n] \times [1,p], \begin{cases} \mu_{lat} = \frac{1}{n}(\sum_{k=1}^{p} lat_k) * min_{rssi} \\ \mu_{lng} = \frac{1}{n}(\sum_{l=1}^{p} lng_l) * min_{rssi} \\ lat_{i,j} = rssi_{i,x} * lat_j, \forall x \in [1,p] \end{cases}$$

A chaque message, désigné par un index, on affecte d'une part la latitude aux colonnes des stations ayant reçu un signal, et d'autre part la longitude. La matrice a de ce fait le double de la taille de la matrice précédente, à la place des valeurs non renseignées on met un terme correspondant à la moyenne de la latitude ou de la longitude multipliée par le rssi minimum.

| index | $bsid_1lat$ | $bsid_2lat$ | ... | $bsid_plat$ | $bsid_1lng$ | $bsid_2lng$ | ... | $bsid_plng$ |
|---|---|---|---|---|---|---|---|---|
| 0 | $\mu_{lat}$ | $lat_{0,2}$ | ... | $\mu_{lat}$ | $\mu_{lng}$ | $lng_{0,2}$ | ... | $\mu_{lng}$ |
| 1 | $lat_{1,1}$ | $\mu_{lat}$ | ... | $\mu_{lat}$ | $lng_{1,1}$ | $\mu_{lng}$ | ... | $\mu_{lng}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| n | $lat_{n,1}$ | $\mu_{lat}$ | ... | $lat_{n,p}$ | $lng_{n,1}$ | $\mu_{lng}$ | ... | $lng_{n,p}$ |

TABLE 3 – Amélioration de la matrice

# Machine Learning

## Détermination du meilleur modèle

Sur les échantillons *train* et *val*, le meilleur modèle ainsi que les paramètres associés ont été estimés avec une gridsearch. La validation croisée choisie est une stratégie de *leave one device out*.

Différents algorithmes ont été testés, dont des Ramdom Forest, des Support Vecteurs Machines et XGBoost.

## Régression et prédictions

Le meilleur modèle calculé à l'étape précédente est entrainé sur les données $X_{train}$ et $y_{train}$ et ensuite appliqué aux variables explicatives d'évaluation $X_{val}$. Cette fonction renvoie les valeurs des latitudes et longitudes prédites.

# Résultats

Le meilleur score est obtenu avec un Random Forest avec 100 arbres pour prédire la latitude et 10 arbres pour prédire la longitude. A 80%, l'erreur commise est de 3864km.