Projet de Stage : Détection d'Anomalies

Wensi DING

August 25, 2016

Outline

Introduction

Données :

- données brutes : imsi, event, cgi, timestamp
- données stop-segment : imsi, segmentType, tsStart, tsEnd, cuidStart, cuidEnd, speed
- données des cellules : cgi, cuid, longitude, latitude, rayon

Objective : identifier les **anomalies**

- comportements inhumains?
- comportements particuliers?
- imprécision ou erreurs dans les données ?
- etc...

Présentation des Modèles

Modèle Gaussien :

- chaque variable suit la loi gaussienne
- toutes les variables sont indépendantes
- anomalies sont les points avec les plus petites probabilités

Modèle LOF :

- lof : rapport entre la moyenne des densités des points voisins et la densité locale
- anomalies sont des points qui ont une densité relativement petite par rapport leurs voisins

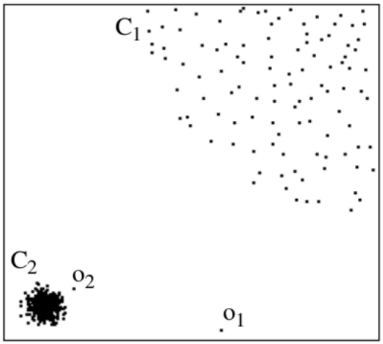
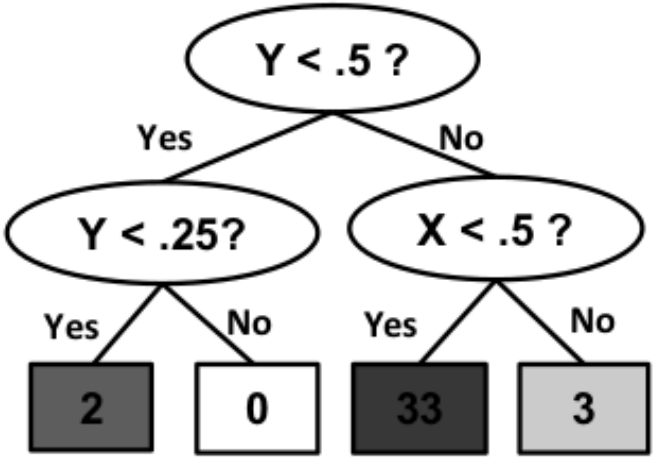


Figure 1: 2-*d* dataset DS1

Présentation des Modèles

Modèle Half Space Tree :

- Arbre de décision
- Chaque nœud représente les intervalles des toutes les dimensions
- anomalies sont des points qui entrent dans les feuilles avec les plus petits nombres de points



Variables Possibles

Avec les **données brutes**:

- **v_max** : le maximum des vitesses entre deux événements consécutifs
- **v_mean** : la moyenne sur l'ensemble des valeurs moyennes des vitesses entre deux événements consécutifs chaque jour
- **n_events_mean** : la moyenne du nombre d'événements chaque jour
- **n_events_sd** : la variance du nombre d'événements chaque jour
- **duree_stop_mean** : la moyenne des sommes de durée d'être immobile chaque jour
- **duree_stop_sd** : la variance des sommes de durée d'être immobile chaque jour

Variables Possibles

Avec les **données stop-segment**:

- **v_max_ss**, **v_mean_ss**, **duree_stop_mean_ss**, **duree_stop_sd_ss**
- **n_stop_mean** : la moyenne du nombre d'événements de type "STOP" chaque jour
- **n_stop_sd** : la variance du nombre d'événements de type "STOP" chaque jour
- **dis_ss** : la médiane des distances entre les endroits où la personne a fait le dernier stop chaque jour

Test et Conclusion

sur les variables :

- v_max ✓: capable de sortir deux types d'anomalies: cellules mal positionnées, oscillations entre deux cellules avec une distance grande
- v_max_ss : capable d'enlever les événements anormaux à cause de cellules très mal placées; pour les cellules pas trop mal positionnés ou les cellules avec rayons grands, le résultat peut varier
- v_mean : écrasé par les v_max très grands, pas très performant
- v_mean_ss ✓: capable de représenter une sorte de mobilité d'une personne en moyenne

Test et Conclusion

sur les variables :

- n_events_mean ✓: capable de sortir des imsi avec le nombre d'oscillations entre plusieurs cellules très grand
- n_events_sd : écrasé par les n_events_mean très grands, pas très performant
- n_stop_mean ✓: capable de représenter une sorte de mobilité d'une personne, exemple: un voyageur peut faire plein de stop pendant une journée
- n_stop_sd ✓: capable de représenter une sorte de mobilité d'une personne, exemple: un voyageur qui visite plusieurs villes peut faire plein de stop dans une journée de visite, très peu de stop dans une journée de déplacement

Test et Conclusion

sur les variables :

- duree_stop_mean : pas très précis pour représenter la vraie durée de stop
- duree_stop_sd : pas très précis à cause d'imprécision de duree_stop_mean
- duree_stop_mean_ss : capable de représenter une sorte de mobilité d'une personne, mais cette variable a une relation linéaire avec n_stop_mean
- duree_stop_sd_ss : pas de typique comportement sorti
- dis_ss : capable de représenter une sorte de mobilité dans le sens global, exemple: les personnes qui bougent tout le temps avec une grande distance chaque jour

Test et Conclusion

variables choisies :

- E1 : v_max, v_mean, n_events_mean, n_events_sd, duree_stop_mean, duree_stop_sd
- E2 : v_max, v_mean_ss, n_events_mean, n_stop_mean, n_stop_sd

sur les modèles :

1. Courbe ROC et AUC Score(référence : modèle gaussien)
   - LOF : 0.78 (E1) ; 0.82 (E2)
   - HS_Tree : 0.94 (E1) ; 0.93 (E2)

Test et Conclusion

2. Matrice de Confusion : 679 anomalies sur 157,107 imsi pour tous les trois modèles

Modèle Référent :	Gaussien	Gaussien	LOF
Modèle Testé :	LOF	HS_Tree	HS_Tree
Accuracy Score :	0.99	0.99	0.99
F1 Score :	0.23	0.22	0.11
Matrice de Confusion :	TN : 155906; FP, FN : 522; TP : 157	TN : 155897; FP, FN : 531; TP : 148	TN : 155825; FP, FN : 603; TP : 76

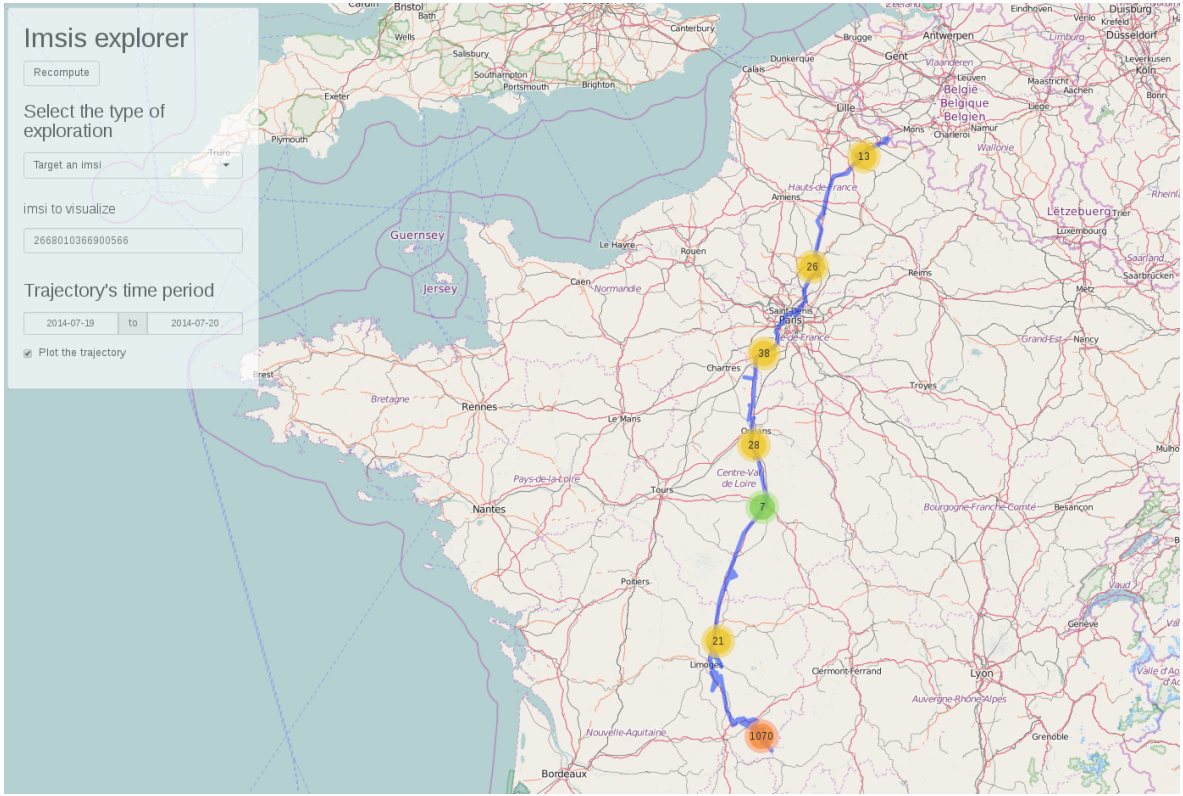
Table 1: Résultat matrice de confusion

Test et Conclusion

- Pour comparer ou évaluer les résultats des trois modèles, il vaut mieux utiliser F1 Score ou Similarité de Jaccard.
- Pour expliquer les grandes différences entre les trois modèles, une hypothèse possible : la transformation à la distribution gaussienne sur les variables dans le modèle gaussien a un effet important.

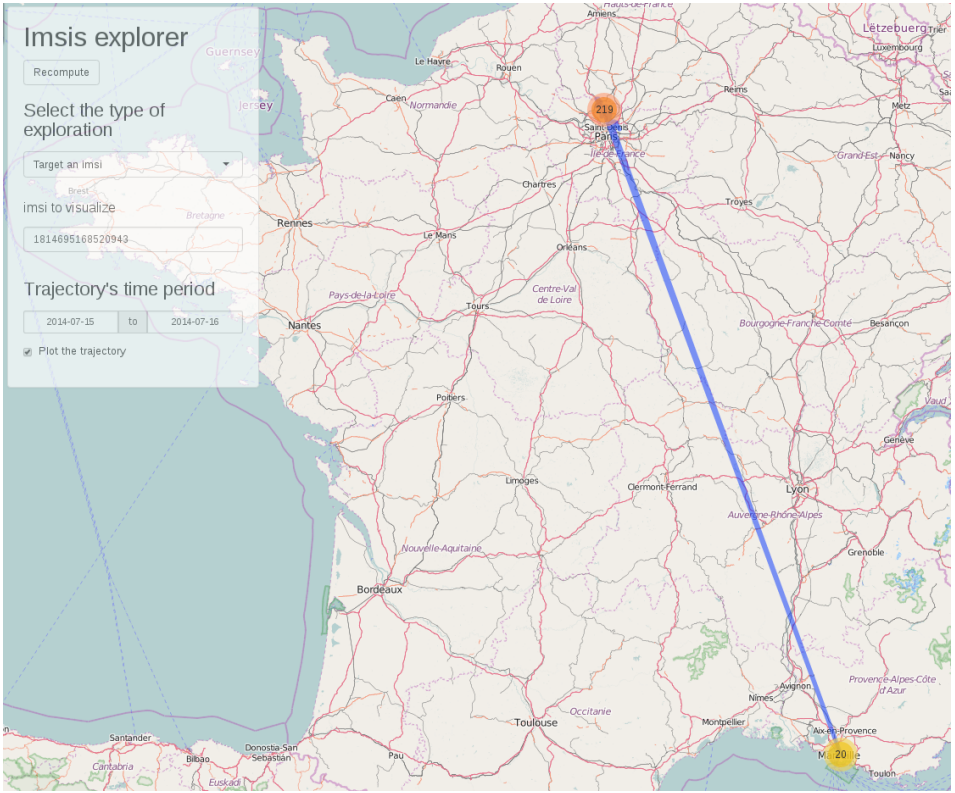
Quelques types d'anomalies

1. profile spécial



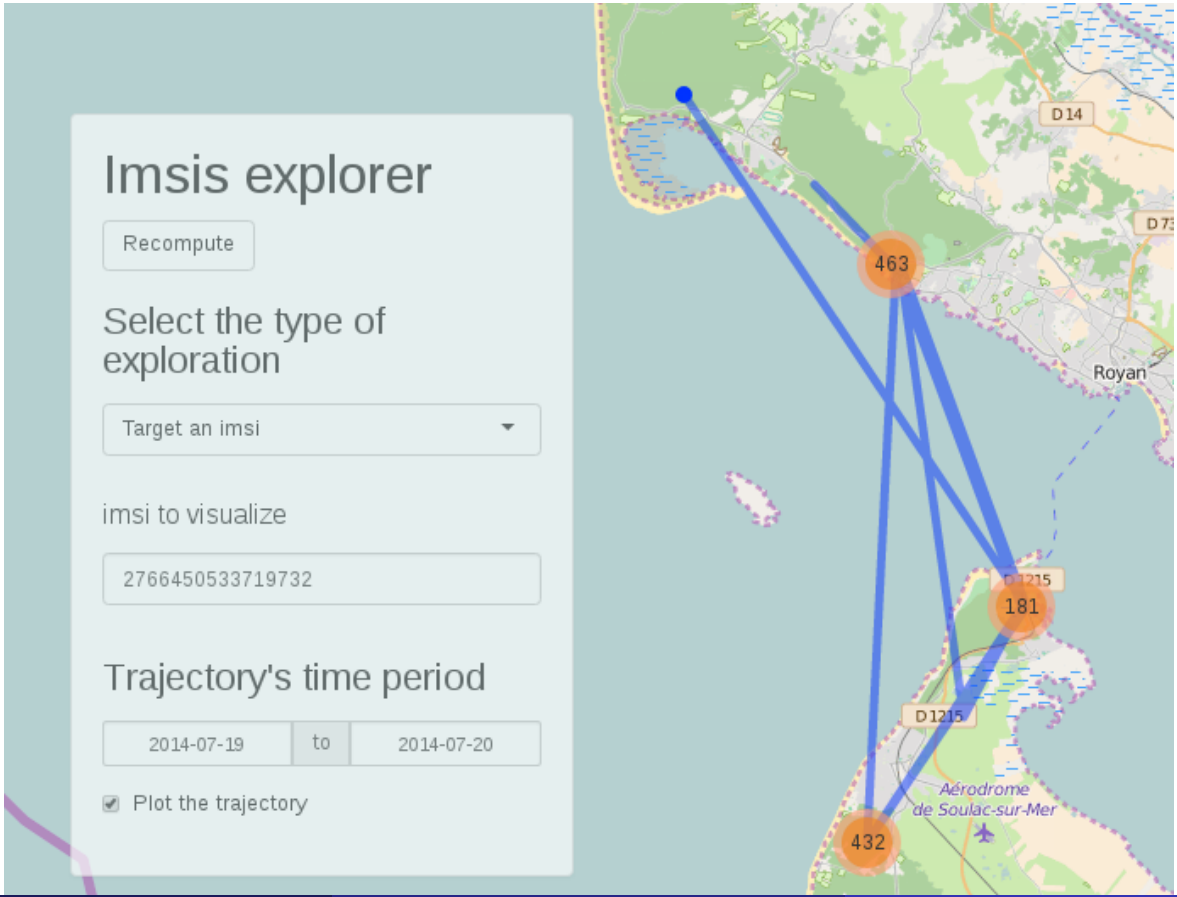
Quelques types d'anomalies

2. cellule très mal placée



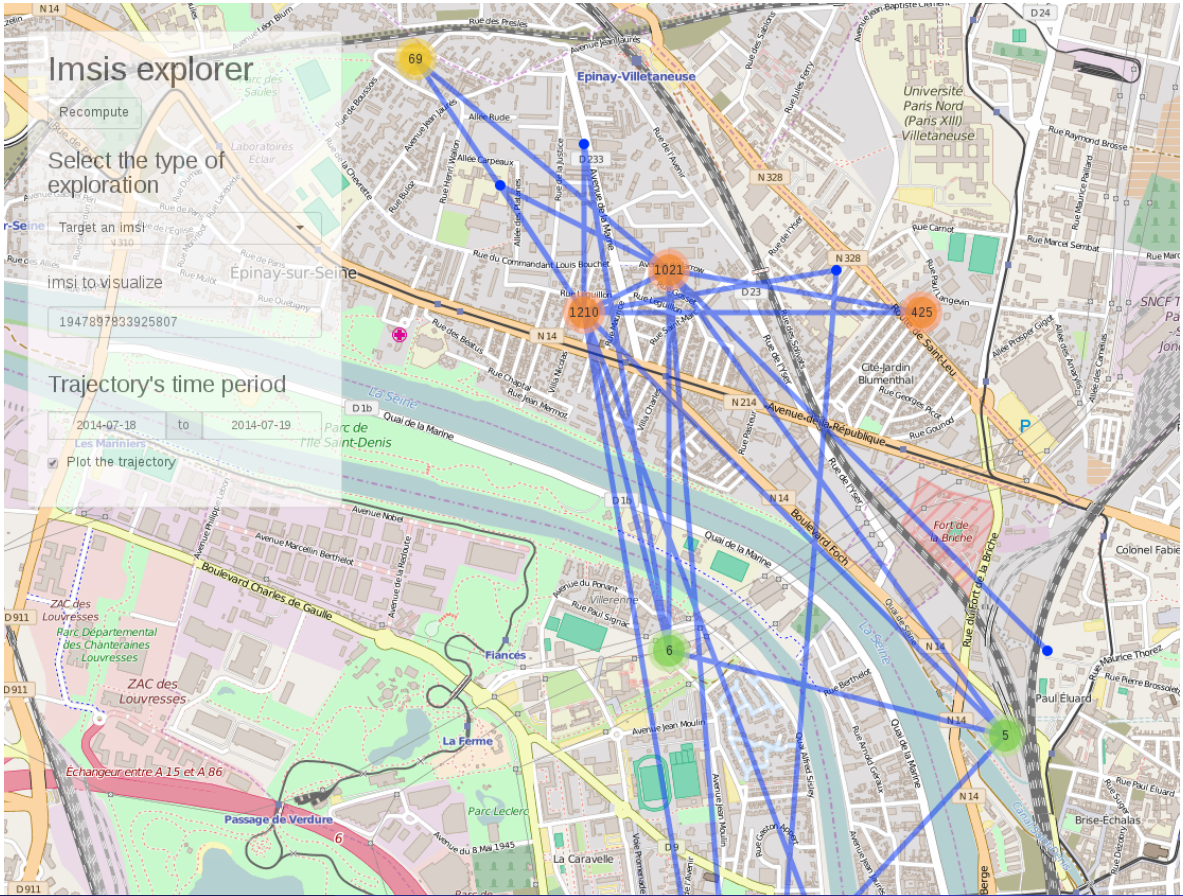
Quelques types d'anomalies

3. nombre d'événement explose



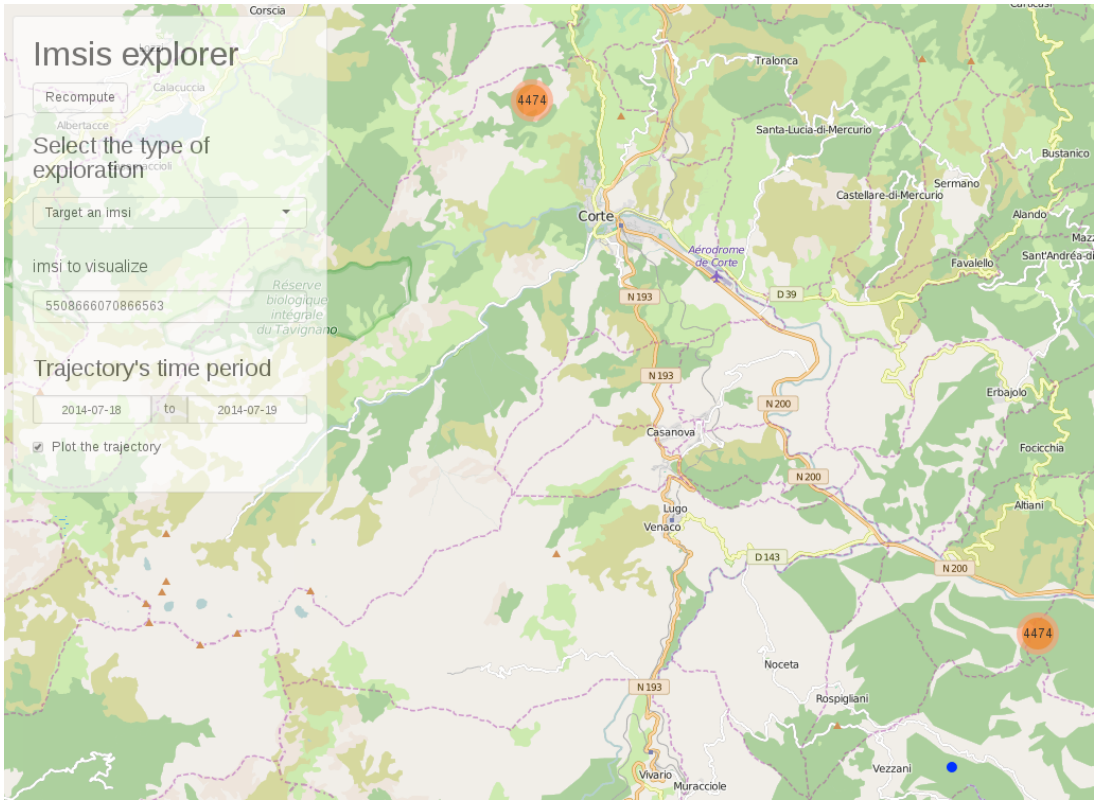
Quelques types d'anomalies

3. nombre d'événement explose



Quelques types d'anomalies

3. nombre d'événement explose



Proposition

- modèle : comprendre les différences des modèles
- variables : créer et examiner des nouvelles variables (par exemple: différencier les jours dans la semaine et les weekends; utiliser les données avec d'autre type d'événement; etc...)

Référence

- Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, Jörg Sander (2000). LOF: Identifying Density-Based Local Outliers. ACM, 1-58113-218-2/00/05
- Swee Chuan Tan, Kai Ming Ting, Tony Fei Liu. Fast Anomaly Detection for Streaming Data. Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence