Table des matières

1 Exploration des données

A partir de statistiques descriptives simples, expliquez votre approche initiale pour aborder ce jeu de données. Illustrez votre réponse avec quelques graphiques et/ou cartes.

Réponse :

Nous disposons de trois caractéristiques dans le jeu de données : un identifiant utilisateur (user_id), des données temporelles (timestamp_client) et des données spatiales (lat, lon). Ces données forment des séries-temporelles de localisation. L'objectif de ce projet est de déterminer le lieu d'habitation d'un utilisateur. Nous aborderons ce jeu de données en analysant l'aspect spatial et temporel des données.

1.1 Description générale

Nous vérifions d'abord la présence de doublons et valeurs manquantes dans le jeu de données.

Le jeu de données peut être résumé comme :

Statistique	Valeur
Taille de l'échantillon	19985
Nombre d'utilisateurs	12
Période de tracking	698 jours et 15 h 10 min 21 s

Table 1: Statistiques Générales

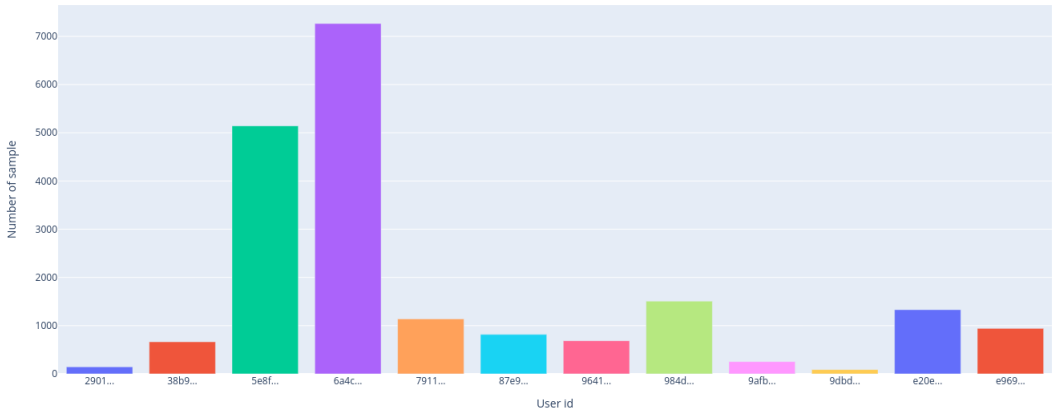

FIGURE 1: Distribution des mesures par utilisateur

Statistique	Valeur
Moyenne	1665.4
Mediane	879.5
Ecart-type	2212.9

TABLE 2: Statistiques de la distribution des mesures par utilisateur

D'aprés la Figure 1 on remarque que les données ne sont pas réparties équitablement entre les utilisateurs. Il y a une forte dispersion des valeurs autour de la moyenne ($\sigma = 2219.9$). Pour certains utilisateurs il peut être difficile de prédire le lieu d'habitation de manière précise.

1.2 Description temporelle

Dans cette partie nous explorons la caractéristique temporelle du jeu de données.

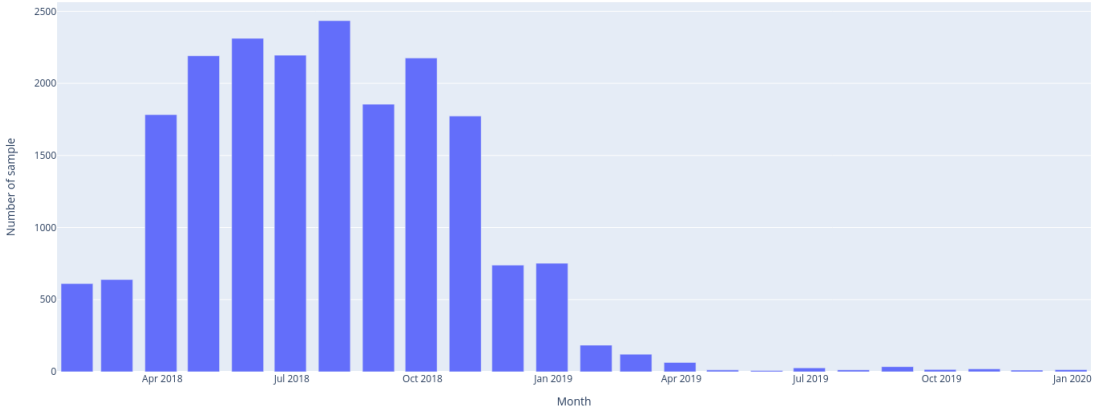

FIGURE 2: Distribution des mesures par mois

Les mesures sont principalement concentrées entre Février 2018 et Avril 2019. Aprés Avril 2019 la fréquence des mesures diminue.

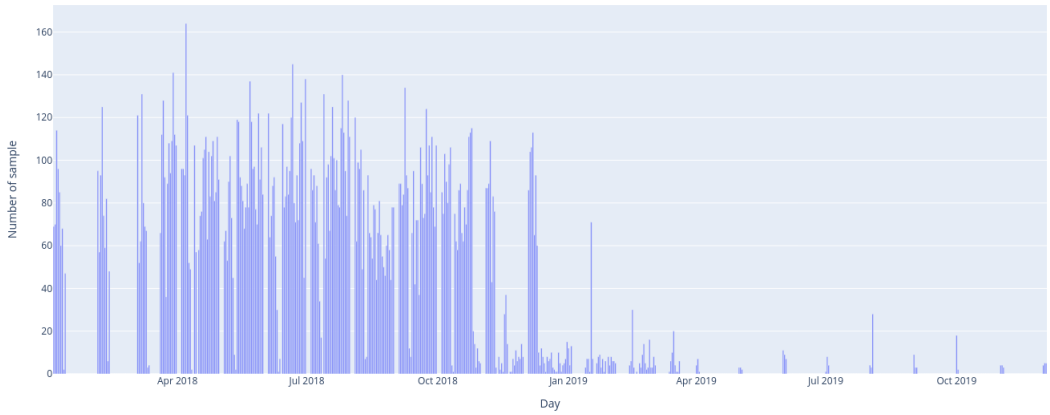

FIGURE 3: Distribution des mesures par jour

Les mesures sont distribuées de manière irrégulières sur les jours de la période.

FIGURE 4: Distribution des mesures par jour de la semaine par utilisateur

Pour certains utilisateurs les mesures sont distribuées de manière (presque) équiprobable (Figure 4.1, 4.2), pour d'autres la répartition est irrégulière (Figure 4.8).

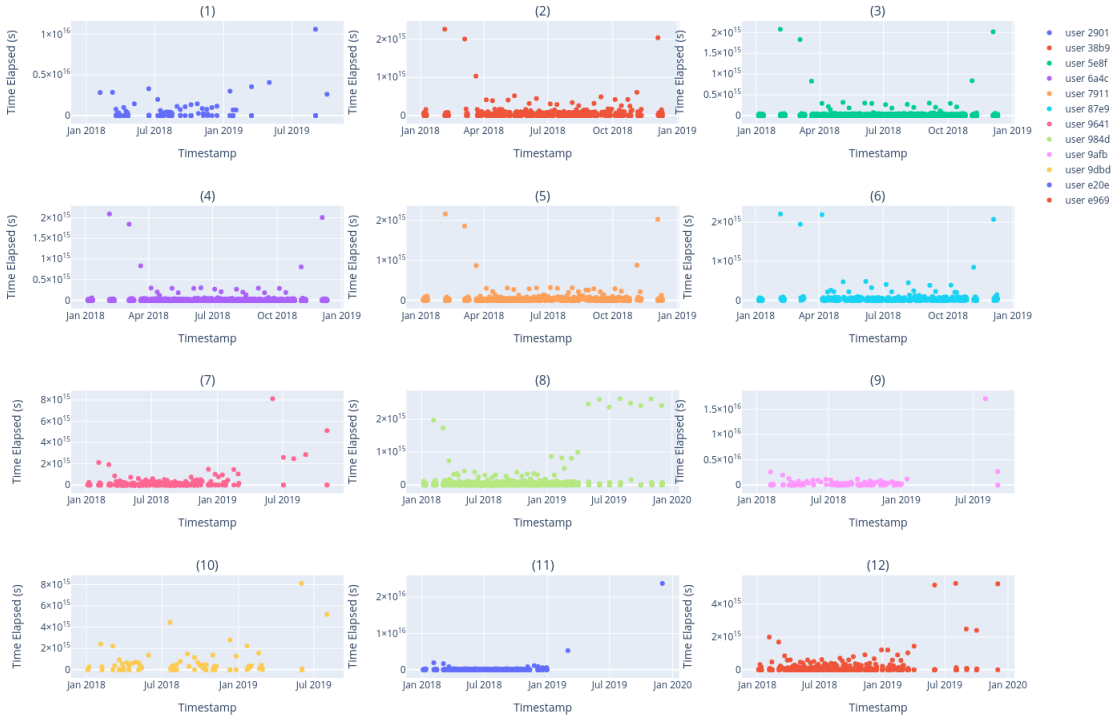

FIGURE 5: Distribution du temps écoulé entre deux mesures

Nous remarquons que pour chaque utilisateur il existe des points pour lesquels le temps écoulé est grand. Cet écart peut être dû à des pertes de connections GPS ou des erreurs de localisation. Ces points engendrent du bruit dans le jeu de données.

1.3 Description Spatiale

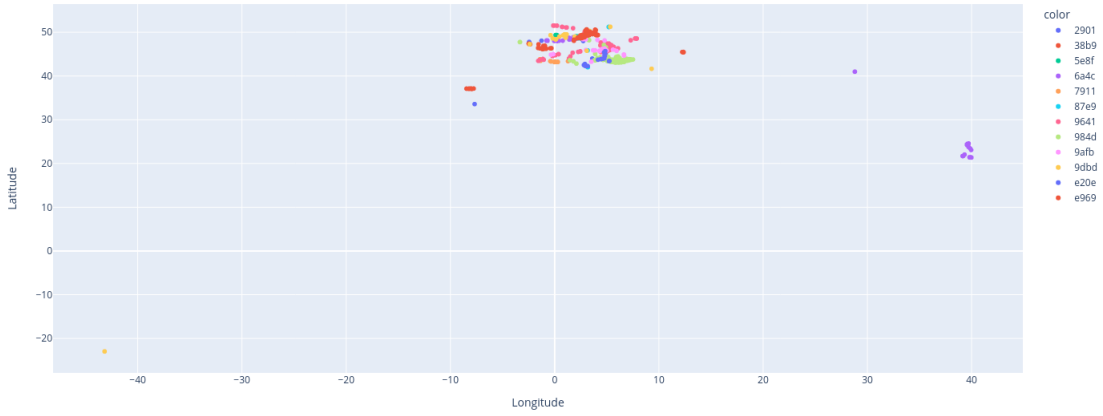

FIGURE 6: Visualisation des données GPS

On remarque que le nuage de points est concentré autour d'une même zone géographique (pays par exemple). Les utilisateurs proviennent tous d'une zone géographique proche.

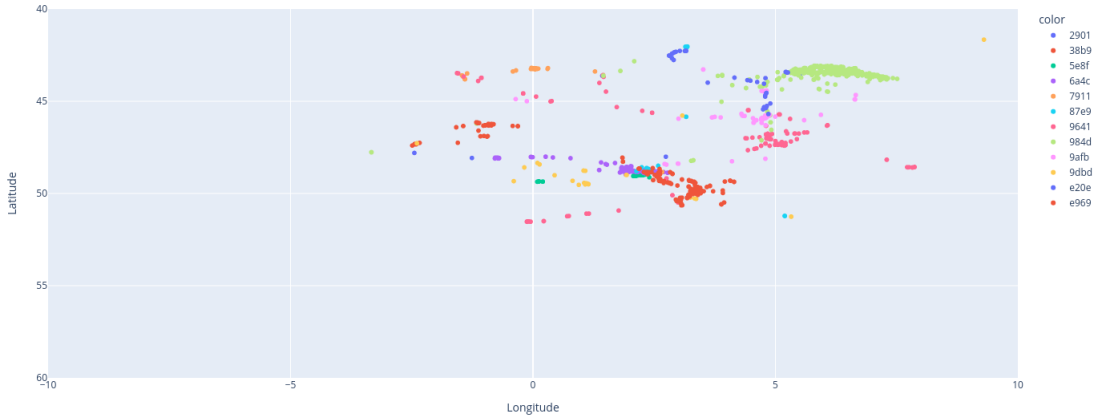

FIGURE 7: Visualisation des données GPS (zoom)

On remarque que si pour certains utilisateurs les mesures GPS sont distantes, pour d'autres les mesures sont confondues. Ces utilisateurs sont proches les uns des autres. Il peut être difficile de séparer ces zones par clustering.

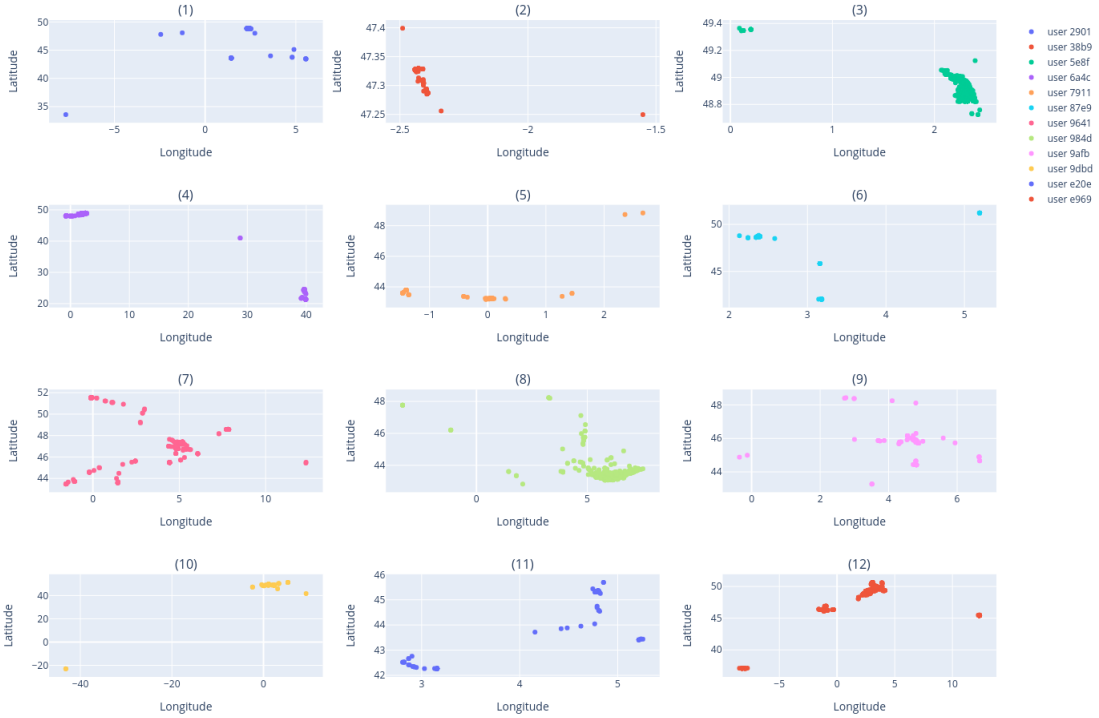

FIGURE 8: Visualisation des données GPS par utilisateur

Pour chaque utilisateur, les nuages de points sont différents. Pour certains utilisateurs (Figure 8.1, 8.5, 8.9) on distingue des zones denses, tandis que pour d'autres (Figure 8.6, 8.11, 8.12) les points sont dispersés.

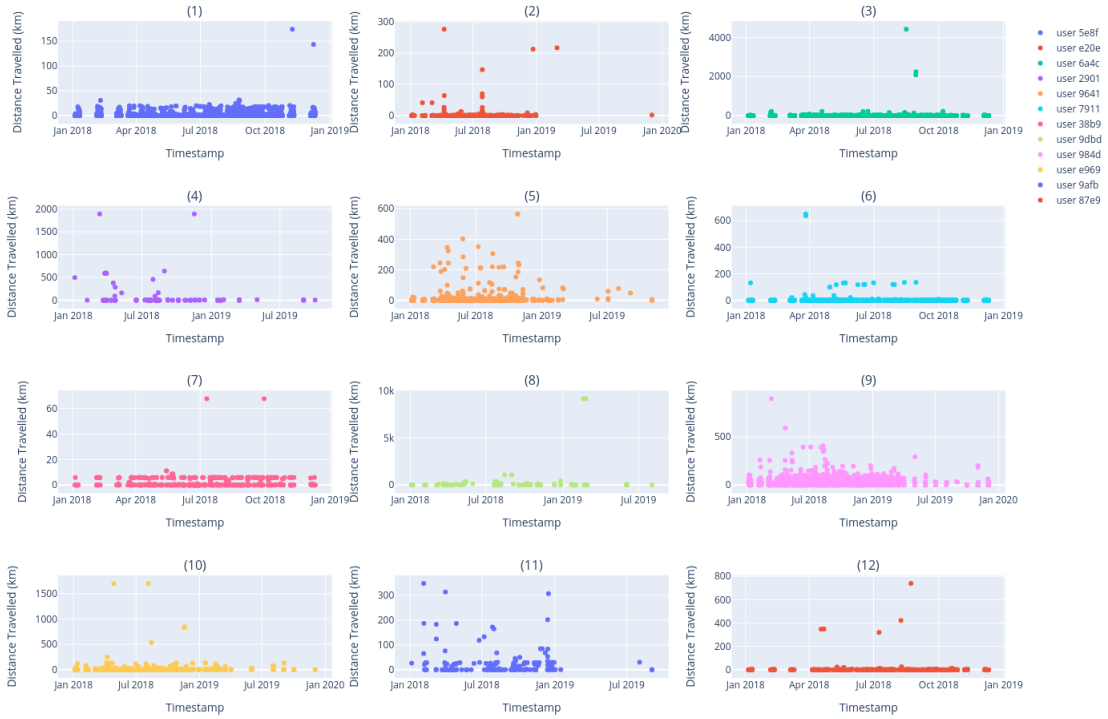

FIGURE 9: Distribution de la distance (en km) entre deux mesures consécutives par utilisateur

De même que pour le temps, on remarque de grands écarts de distances entre les mesures. Ces écarts peuvent êtres dû également à des pertes de connections ou des déplacements en transport. Ces points bruitent également le jeu de données car nous sommes plus interessés ici par le temps passé dans un endroit que le trajet pour y arriver.

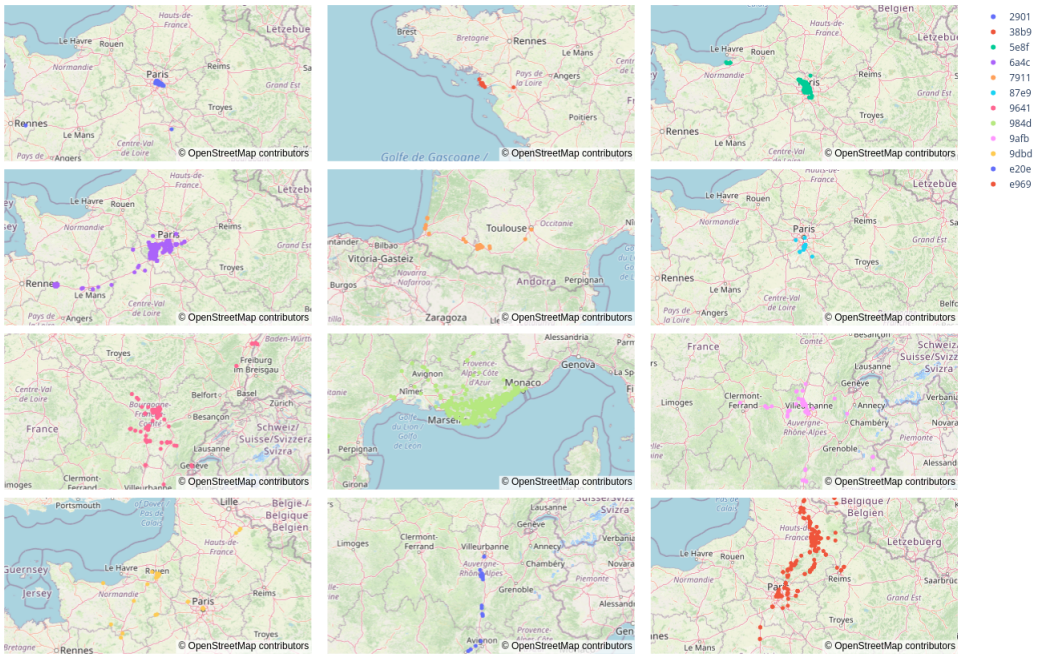

FIGURE 10: Vue cartographique des mesures GPS par utilisateur

Nous pouvons également visualiser la position et les déplacements des utilisateurs sur une carte. Nous nous servirons de ces cartes pour vérifier visuellement les performances de notre approche.

2 Modélisation

2.1 De quel type de problème d'apprentissage s'agit-il ?

Réponse :

Dans ce projet nous devons prédire le lieu d'habitation des utilisateurs, comme nous ne disposons pas des lieux d'habitation des utilisateurs il s'agit d'un problème d'apprentissage non-supervisé.

2.2 Concevez un modèle simple permettant d'estimer le lieu d'habitation de l'utilisateur (latitude, longitude) en expliquant vos choix algorithmiques : hypothèses, feature engineering, choix de l'algorithme, hyperparamètres.

Réponse :

Nous disposons dans ce jeu de données de mesure de la localisation des utilisateurs dans le temps. Pour prédire le lieu d'habitation de l'utilisateur nous souhaitons regrouper les différents emplacements où l'utilisateur a passé le plus de temps. Nous utiliserons donc des méthodes de clustering afin de d'effectuer ce regroupement.

Pendant l'exploration de données nous avons remarqué que :

- La distribution des distances parcourues entre les points présente des valeurs extrêmes (grande distance entre les points), pouvant correspondre par exemple à des trajets en transport (avion, voiture)
- Certains enregistrements sur une période peuvent correspondre à des déplacements de l'utilisateur
- Les mesures ne sont pas régulières dans le temps
- Les localisations dépendent de la précision de l'appareil de mesure. Si un utilisateur reste au même endroit pendant un certain temps les données de latitude et longitude ne seront pas les mêmes (dispersion des valeurs)

Ces observations constituent un bruit dans le jeu de données. Afin d'améliorer la précision de notre prédiction nous filtrerons ces valeurs.

Nous emploierons deux stratégies de filtrage, En notant $\mathbf{L}_i^j$ la i^{ème} mesure du j^{ème} utilisateur :

- Filtrer les points ayant une distance parcourue supérieure à un certain seuil

$$\text{Si } \mathbf{L}_{i-1}^j - \mathbf{L}_i^j > \xi, \text{ alors nous supprimons } \mathbf{L}_i^j \tag{1}$$

  Dans le but de réduire le bruit généré par la mesure de l'appareil, le seuil ξ devra être proche de l'incertitude de la mesure.

- Filtrer les points ayant une vitesse supérieure à la vitesse de marche moyenne.

$$\text{Si } \frac{\mathbf{L}_{i-1}^j - \mathbf{L}_i^j}{\Delta t_{i-1,i}} > \gamma, \text{ alors nous supprimons } \mathbf{L}_i^j \tag{2}$$

  Dans le but de réduire les points correspondant à des déplacements de l'utilisateur, γ devra être proche de la vitesse de marche moyenne.

Pour déterminer l'incertitude liée aux mesures GPS, nous isolons un ensemble de mesures proches correspondant à l'emplacement d'un utilisateur (visualisé en Figure 10) grâce à un algorithme de clustering. Nous calculons la distance moyenne entre les points de ce cluster pour obtenir notre estimation.

Pour la vitesse de marche moyenne, nous savons qu'elle est estimée entre 4 et 5 km/h. Nous choisirons 4.5km/h soit 1.25 m/s.

Même en supprimant l'ensemble de ces points, le jeu de données reste bruité. Afin de limiter l'impact du bruit et puisque nous ne connaissons pas, a priori, le nombre de clusters, nous utiliserons DBSCAN (Density-Based Spatial Clustering of Applications with Noise). DBSCAN présente l'avantage de clusteriser par zone de densité et de créer un cluster 'Noise' pour détecter les valeurs aberrantes (outliers).

Trois hyperparamètres sont importants pour DBSCAN :

- ε, qui représente la distance maximum entre deux points pour qu'ils soient considérés comme voisins. Pour choisir nous utiliserons la méthode visuelle 'Elbow curve' puis nous validerons ces valeurs en nous aidant des maps (Figure 10).

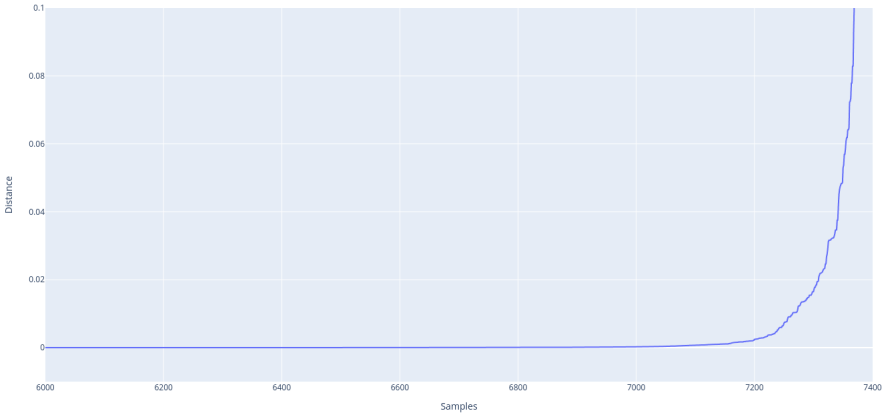

FIGURE 11: K-distance Elbow Curve

On trouve grâce à cette méthode $\varepsilon = 0.040$, cependant expériementalement cette valeur donne un mauvais résultat de clustering. On trouve empiriquement $\varepsilon = 5.10^{-5}$.

- min_samples, qui représente le nombre de points dans un voisinage pour qu'un point soit considéré comme principal. Comme certains utilisateurs ont peu de points, min_samples sera proche du nombre de mesures pour ces utilisateurs. Nous choisissons donc min_samples = 3
- La mesure utilisée pour calculer la distance entre deux points. Nous utilisons la distance d'Haversine qui sert à calculer la distance entre deux coordonnées GPS.

Afin de prédire le lieu d'habitation des utilisateurs nous allons donc :

(1) Nettoyer les données avec les stratégies évoquées précédement (Figure 12)
(2) Effectuer un clustering avec l'algorithme DBSCAN sur tous les utilisateurs (Figure 13, Figure 14)
(3) Récupérer pour chaque utilisateur le cluster le plus dense
(4) Estimer le centre du cluster (lieu d'habitation), en prenant le point du cluster qui a le plus de voisins proches
(5) Vérifier visuellement les performances de notre approche (Figure 15)

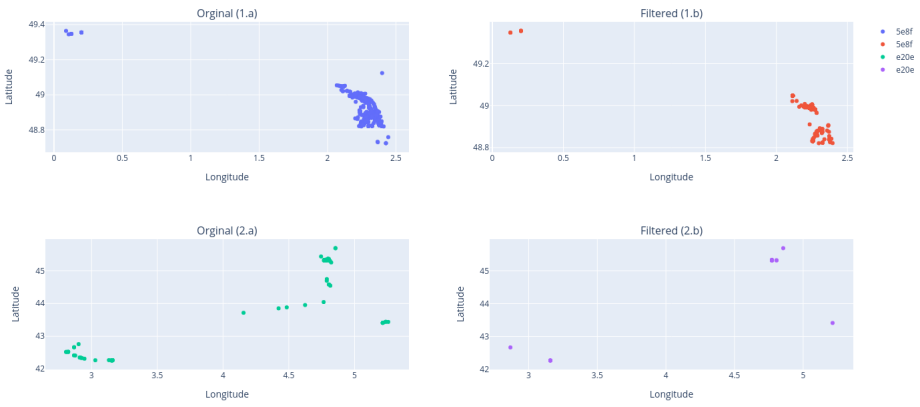

FIGURE 12: Visualisation des données GPS avant (colonne de gauche) et aprés (colonne de droite) filtrage pour 4 utilisateurs

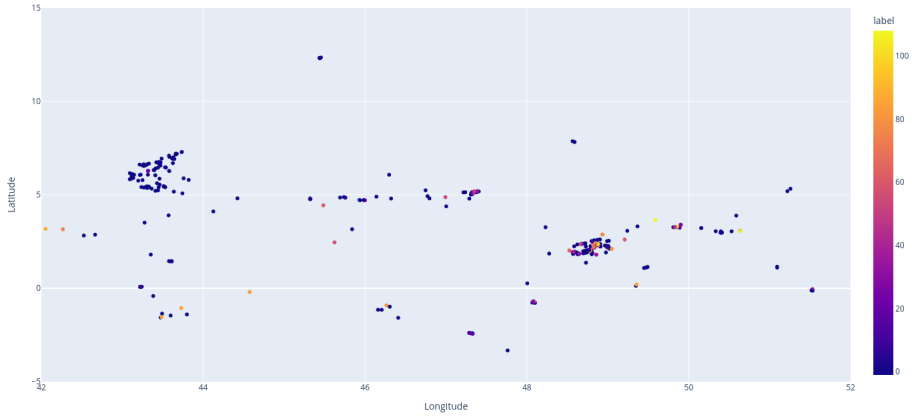

FIGURE 13: Visualisation du clustering avec DBSCAN

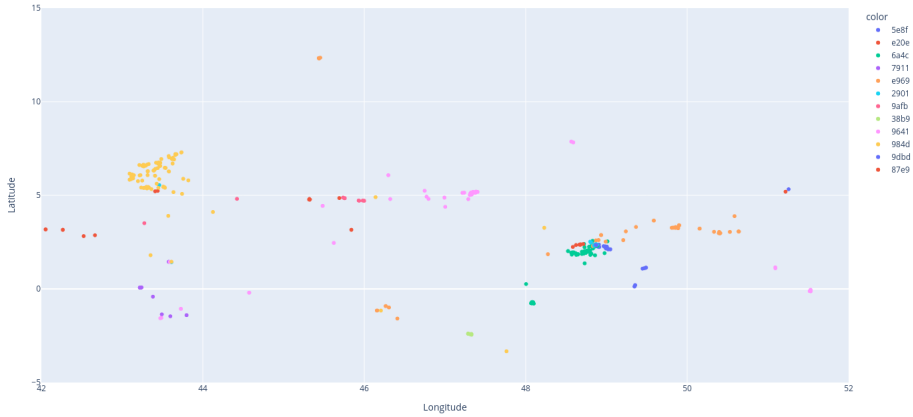

FIGURE 14: Visualisation du ground-truth clustering

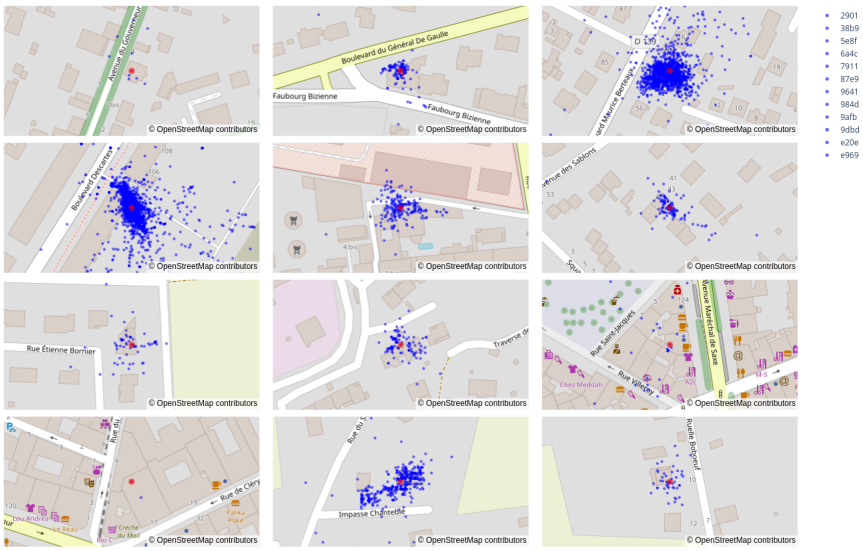

FIGURE 15: Vue cartographique des lieux d'habitation prédits pour chaque utilisateur

2.3 Vous avez estimé le lieu d'habitation d'un utilisateur. Quelles informations complémentaires pourrait-on extraire de ce jeu de données ?

Réponse :

Nous pourrions également extraire de ce jeu de données les sous-locations les plus fréquentées par un utilisateur (son lieu de travail, son magasin favori). Aprés avoir identifié toutes les locations où l'utilisateur a l'habitude d'aller, nous pourrions nous servir des séries temporelles de trajectoire pour prédire la localisation d'un utilisateur en fonction de l'heure et du jour de la semaine (GMM, Markov Models).

2.4 Comment pourrait-on évaluer la performance du modèle (en faisant l'hypothèse que vous pouvez collecter des données supplémentaires) ?

Réponse :

En faisant l'hypothèse que nous pouvons récolter le lieu d'habitation des utilisateurs et la précision/incertitude des appareils de localisation. Nous pourrions labéliser les points du jeu de données qui seraient à une distance (incertitude de l'appareil) inférieure ou égale au lieu d'habitation.

Nous pourrions ainsi mesurer pour chaque utilisateur le nombre de points effectivement bien clusterisés et ainsi utiliser des mesures de performances de clustering (Homogeneity, Completeness ou Mutual Information based scores). Nous pourrions également mesurer l'écart entre la localisation GPS prédite et la localisation GPS réelle pour obtenir une mesure de la performance sur les coordonnées prédites (Mean Square Error)

3 Bibliographie

(1) Location Knowledge Base

(2) Simple GPS data visualization using Python and Open Street Maps
(3) Using GPS to Learn Significant Locations and Predict Movement Across Multiple Users
(4) A geographical location prediction method based on continuous time series Markov model
(5) GPS Data Analysis