RAPPORT PROJET LONG

Réalisé par : Brenda TONLEU

Sous la direction de : M. Ricardo GUILHERME

Année universitaire 2023-2024

Table des matières

1 Introduction

Dans un monde en constante évolution, les dynamiques géodémographiques des centres urbains sont au cœur des préoccupations en matière de mobilité et de planification urbaine. La compréhension des déplacements des individus au sein des villes revêt une importance capitale pour anticiper les besoins en infrastructures, optimiser les ressources disponibles et améliorer la qualité de vie urbaine. Dans cette optique, l'analyse des données cellulaires émerge comme un outil précieux pour appréhender les schémas de mobilité urbaine et caractériser les modèles géospatiaux des zones urbaines.

Le présent projet se propose d'explorer ces aspects fondamentaux en utilisant des données cellulaires anonymisées fournies par Shanghai Telecom. Ces données riches et diversifiées nous offrent une opportunité unique d'analyser les comportements de déplacement des habitants de Shanghai, l'une des mégapoles les plus dynamiques au monde. En nous appuyant sur ces données, nous chercherons à analyser les schémas de mobilité urbaine, à exploiter les données cellulaires pour obtenir des insights précieux et à développer des modèles prédictifs pour soutenir la planification urbaine. Nous débuterons par une présentation détaillée du jeu de données fourni par Shanghai Telecom, avant d'entrer dans le vif de notre analyse. En combinant expertise en géospatialité, analyse de données et techniques de machine learning, nous visons à fournir des perspectives nouvelles et éclairantes pour répondre aux défis actuels de la mobilité urbaine et de la planification des villes.

2 Contexte

Ce document reporte notre étude effectuée sur l'analyse des Données et Apprentissage Collaboratif pour la Mobilité Urbaine et la Caractérisation Géospatiale dans le cadre du projet long.

3 Présentation du sujet

Dans ce projet, nous utilisons l'analyse des données cellulaires pour obtenir des informations sur les modèles de mobilité urbaine et la caractérisation géospatiale.

En analysant des données cellulaires anonymisées du dataset fourni par Shanghai Telecom, nous pouvons extraire des informations sur les déplacements des personnes, identifier les points chauds de la circulation et créer des modèles géospatiaux qui aident à la planification urbaine et à l'optimisation.

Objectifs de l'Étude :

- **Comprendre les schémas de mobilité urbaine :** Analyser les modèles de déplacement des populations urbaines, en identifiant les tendances, les points chauds de la circulation (Les zones ayant des attroupements et les zones moins occupées)et les schémas de déplacement.
- **Utiliser les données cellulaires pour obtenir des insights :** Extraire des informations précieuses sur les habitudes de déplacement (zone géographique : résidence, lieu de travail, etc.)) des utilisateurs dans les zones urbaines en exploitant des données anonymisées provenant des réseaux cellulaires.
- **Développer des modèles de machine learning prédictifs :** Construire des modèles de machine learning pour prédire les futurs déplacements des utilisateurs en fonction de leur historique de déplacements.

4 Présentation du dataset

Le jeu de données fourni par Shanghai Telecom contient plus de 7,2 millions d'enregistrements d'accès à Internet provenant de 3 233 stations de base différentes, recueillis sur une période de six mois.

Ces enregistrements proviennent de 9 481 téléphones mobiles utilisés dans la ville de Shanghai, en Chine.

Chaque enregistrement comprend des informations telles que :

- **Mois (Month) :** Indique le mois au cours duquel l'enregistrement a été effectué.
- **Date (Date) :** Spécifie la date de l'enregistrement.
- **Heure de début (Start Time) :** Indique l'heure de début de l'enregistrement.
- **Heure de fin (End Time) :** Indique l'heure de fin de l'enregistrement.
- **Localisation de la station de base (Base Station Location) :** Les coordonnées de longitude et de latitude de la station de base à partir de laquelle l'accès à Internet a été effectué.
- **Identifiant de l'utilisateur (User ID) :** L'identifiant unique du téléphone mobile utilisé pour l'accès à Internet.

Ce dataset offre une opportunité d'analyser les habitudes de déplacement des utilisateurs dans la ville de Shanghai, explorer les schémas de mobilité urbaine des habitants et de développer des modèles prédictifs pour anticiper les futurs déplacements.

5 Article relatifs étudiés

5.1 Understanding Mobile Traffic Patterns of Large Scale Cellular Towers in Urban Environment

Cet article vise à étudier et modéliser les schémas de trafic (traffic patterns) des grandes tours cellulaires (cellular towers) déployées dans des environnements urbains à grande échelle. L'article explore les données de trafic de 9 600 tours cellulaires, identifie cinq schémas de trafic distincts et établissent une corrélation entre ces schémas et les caractéristiques géographiques, telles que les zones résidentielles, les bureaux, les transports, les divertissements et les zones complètes.

Il démontre également comment ces schémas de trafic peuvent être utilisés pour comprendre les comportements humains, notamment en termes de mobilité urbaine.

Enfin, l'article propose une méthodologie de traitement du trafic qui intègre des informations temporelles, spatiales et spectrales pour extraire et modéliser les schémas de trafic des tours cellulaires.

5.1.1 Description du dataset

Le jeu de données utilisé dans l'étude est une trace cellulaire anonymisée collectée par un fournisseur de services Internet à Shanghai, en Chine, entre le 1er et le 31 août 2014.

Il contient des informations détaillées sur l'utilisation des données mobiles de 150 000 utilisateurs, telles que l'identifiant des appareils, les horaires de connexion, les ID et adresses des stations de base, ainsi que la quantité de données 3G ou LTE utilisées dans chaque connexion.

Le jeu de données contient au total 1,96 milliard de tuples, contribués par environ 9 600 stations de base à travers Shanghai.

5.1.2 Preprocessing et visualisation (fig 1 et fig 2)

Le prétraitement effectué consiste à :

— Elimination des entrées redondantes et conflictuelles telles que les journaux de trafic identiques causés par des problèmes techniques.
— Elimination des informations incomplètes sur les emplacements des stations de base
— Calcul de la densité de trafic pour chaque zone de la ville ; pour comprendre la distribution spatiale du trafic cellulaire à travers Shanghai.

La visualisation du jeu de données consiste à observer :

— Des schémas temporels fondamentaux du trafic, avec des pics de trafic pendant la journée et des périodes de faible trafic la nuit, correspondant aux habitudes de sommeil des humains.
— une concentration de trafic dans les zones centrales de la ville pendant les heures de travail et des niveaux de trafic plus faibles pendant la nuit.

5.1.3 Identification des patterns de traffic des tours cellulaires (fig 3)

L'article présente une méthodologie basée sur le vectoriseur de trafic, l'identificateur de modèle, et le réglage des métriques pour identifier les schémas de trafic parmi des milliers de tours cellulaires.

Ils parviennent à identifier cinq schémas de traffic distincts qui varient en termes de temps de pic, d'amplitude de trafic, et de régularité.

5.1.4 Compréhension des schémas de traffic

L'article explore les caractéristiques temporelles des schémas de trafic identifiés, fournissant des insights sur les comportements de trafic dans différentes zones urbaines fonctionnelles.

— **Caractéristiques Temporelles** : Des différences significatives entre les régions urbaines en termes de volumes de trafic, de moments de pointe et de creux, et de ratios entre les jours de semaine et les week-ends sont observées.

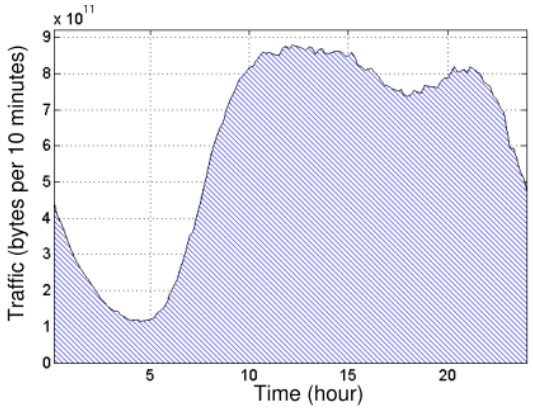

FIGURE 1 – Variation des connections sur une journée

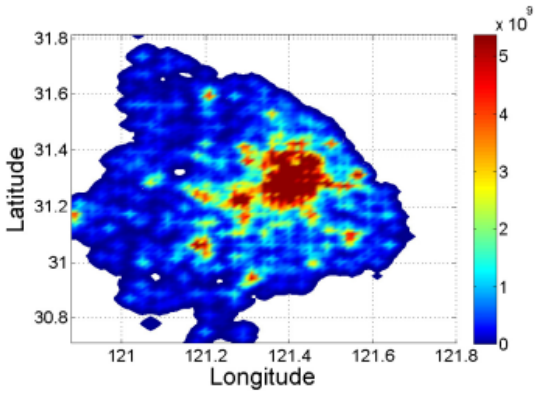

FIGURE 2 – Traffic

PERCENTAGE OF CELL TOWERS CLASSIFIED IN EACH CLUSTER

Cluster Index	Functional Regions	Percentage
1	Resident	17.55%
2	Transport	2.58%
3	Office	45.72%
4	Entertainment	9.35%
5	Comprehensive	24.81%

FIGURE 3 – 05 Schémas de traffic

— **Interrelations entre les schémas de trafic** : Des corrélations élevées sont trouvées entre les schémas de trafic résidentiels, de transport et de bureau, suggérant une routine quotidienne des populations actives. De plus, une similarité marquée entre le schéma de trafic de la zone comprehensive et la moyenne de tous les schémas, indiquant que la **zone comprehensive est un mélange des autres zones fonctionnelles**.

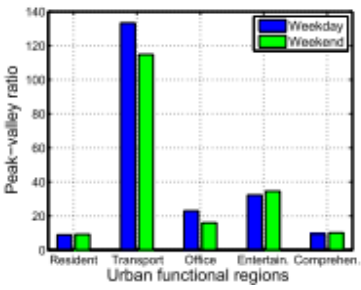

FIGURE 4 – Variation du traffic par groupe le week-end et en semaine

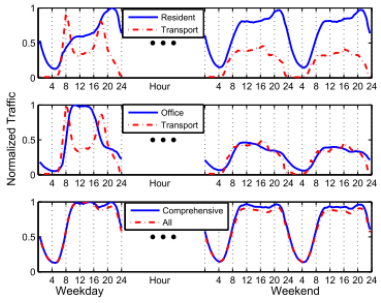

FIGURE 5 – Comparaison du traffic par groupe le week-end et en semaine

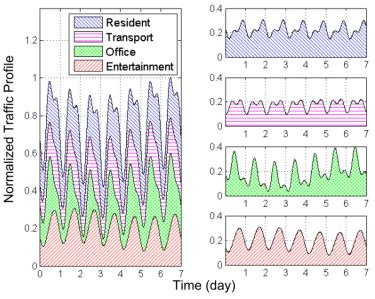

FIGURE 6 – Schémas des traffics

5.1.5 Conclusion

L'article étudié utilise des données de trafic cellulaire pour étudier les comportements humains, les performances réseau et les schémas de trafic des tours cellulaires à grande échelle. Elle se base sur un

modèle simple mais profond capable de caractériser les caractéristiques géographiques de la ville et la régularité de la communication humaine.

5.2 What Machine Learning Predictor Performs Best for Mobility Prediction in Cellular Networks ?

Cet article examine l'efficacité de quatre modèles de prédiction (DNN, le SVM, le XGBoost et le semi-Markov) de mobilité humaine dans les réseaux cellulaires, essentiels pour gérer le trafic mobile dans la mobilité dans les futurs réseaux cellulaires.

5.2.1 Dataset (fig 7)

- Le jeu de données utilisé dans cette étude a été créé de manière synthétique en utilisant un simulateur LTE conforme à la norme 3GPP dans MATLAB.
- La topologie du réseau comprenait 7 cellules macro, chacune avec trois secteurs, soit un total de 21 cellules (On peut le considérer comme 7 villes ayant chacune 03 stations de base ; donc 21 stations)
- Les modèles de mobilité ont été générés pour une semaine avec une granularité d'une minute en utilisant le modèle réaliste SLAW (Self-similar Least Action Walk).
- Chaque utilisateur mobile disposait de 10 080 observations au total.
- Environ 85 % des données ont été utilisées pour l'ensemble d'entraînement, tandis que les 15 % restants ont été utilisés pour tester l'exactitude de la prédiction.
- L'ensemble de données d'entraînement a également été divisé en deux pour l'entraînement (75 % des 85 %) et la validation (25 % des 85 %) en utilisant une validation croisée à quatre volets.
- Les caractéristiques d'entrée comprenaient **l'emplacement actuel de l'utilisateur mobile** et le **temps de séjour dans chaque cellule**, et pour **le DNN**, **les trois emplacements précédents** ont également été considérés comme des caractéristiques supplémentaires.

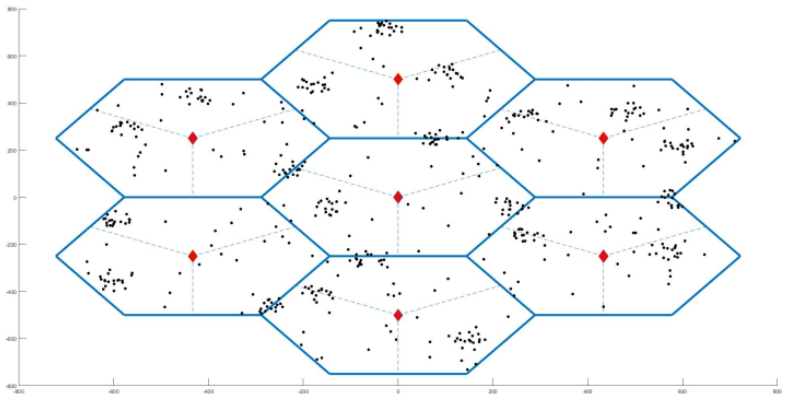

FIGURE 7 – Topologie du réseau artificiel créé

5.2.2 Modèles utilisés

Deep neural Network

L'approche utilisant un réseau de neurones profonds (DNN) pour la prédiction de mobilité repose sur l'utilisation d'un modèle à propagation avant avec plusieurs couches cachées. Ce modèle comprend

une couche d'entrée, une ou plusieurs couches cachées et une couche de sortie. Chaque couche est entièrement connectée aux couches adjacentes. L'activation des neurones dans les couches d'entrée et cachées est déterminée par la fonction d'activation ReLU (Rectified Linear Unit), qui introduit une non-linéarité dans le modèle. La couche de sortie est activée par une fonction softmax pour obtenir une distribution de probabilité sur les différentes classes de destination.

Le modèle est entraîné en minimisant la perte de cross-entropy, calculée en comparant la sortie prédite du modèle avec les étiquettes de classe réelles dans l'ensemble d'entraînement.

$$L = -y.log(y^)$$

La précision du modèle est évaluée en calculant le taux de classification correcte sur l'ensemble de données de validation.

$$Precision = \frac{TP+TN}{TP+TN+FP+FN}$$

Avec TP, FP, TN, FN respectivement le vrai positif (sorties correctement identifiées), le faux positif (FP) (sorties incorrectement identifiées), le vrai négatif (rejets corrects) et le faux négatif (rejets incorrects).

Pour éviter le surapprentissage, le modèle utilise une technique de **cross validation k-fold répétée**, où les données sont divisées en k partitions pour l'entraînement et la validation, répétées n fois.

Les hyperparamètres du modèle, tels que le nombre de couches cachées, le nombre de neurones par couche, la taille du lot et le nombre d'époques, sont sélectionnés en recherchant les hyperparamètres qui permettent d'avoir un meilleur score.

Le modèle final sélectionné comprend **06 hidden layers, chacune avec soixante neurones, et est entraîné avec 10 batchs et 50 époques**.

Les performances du modèle sont surveillées à chaque époque pour éviter le surapprentissage. La méthode se révèle efficace pour prédire les modèles de mobilité avec un taux de précision élevé, tout en évitant le surapprentissage.

XGBoost

XGBoost repose sur l'utilisation d'un modèle de Gradient Boosting Machine (GBM) amélioré et scalable. Il utilise un ensemble d'arbres de décision, appelés CART (Classification and Regression Trees), qui sont construits séquentiellement pour réduire le taux d'erreur de classification. Il est assez facilité d'utilisation et a une capacité de parallélisation de précision prédictive remarquable.

Le modèle XGBoost est entraîné en optimisant une fonction objective régularisée, qui combine une perte différenciable mesurant la différence entre les valeurs prédites et les valeurs réelles, avec des termes de régularisation pour éviter le surapprentissage et réduire la complexité du modèle. La fonction objective est simplifiée à chaque étape d'entraînement à l'aide d'une expansion de Taylor du deuxième ordre.

$$L(\theta)(t) = \sum l(y_i, y^(t-1) + f_t(x_i)) + \Omega(f_t)$$

Les paramètres (le poids des feuilles, les coefficients de régularisation et les paramètres spécifiques aux arbres, tels que la profondeur maximale de l'arbre et le taux d'apprentissage) ont été optimisés en utilisant une approche basée sur la validation croisée et la recherche de grille (grid search).

Les meilleures performances ont été obtenues pour **le nombre minimal de sous arbres, la profondeur maximale, l'échantillonnage de colonnes, la taille de pas et le taux de réduction des surajustements η qui contrôle le taux d'apprentissage et le surajustement fixés à 5, 3, 0.8, 50 et 0.01 respectivement**.

SVM

Les SVM (Support Vector Machine) sont des classifieurs à marge large qui cherchent à trouver une frontière de décision entre les classes en maximisant la marge entre les exemples de formation et la frontière de décision.

Le SVM utilisé dans ce cas était un SVM non linéaire avec un noyau de fonction de base radiale (RBF) pour modéliser les trajectoires de mobilité.

Le noyau RBF permet de cartographier les données dans un espace de caractéristiques de dimension supérieure, permettant ainsi de modéliser des relations non linéaires entre les données.

Les paramètres clés de SVM, (paramètre de régularisation C, paramètre du noyau ι), ont été optimisés à l'aide d'une grid search sur un sous-ensemble de l'ensemble de données d'entraînement.

Les meilleures performances ont été obtenues en utilisant le **noyau RBF avec les paramètres C et ι réglés sur leurs valeurs par défaut**.

Résultats obtenus (fig 8, fig 9)

	Xgboost	SVM	DNN	Semi-Markov
training time	20	2	73	0.01
Prediction time	0.81	1.4	0.48	0.001
Mean training accuracy	90.47	90.58	89.44	82.1
Mean testing accuracy	90.22	88.07	83.89	81.46

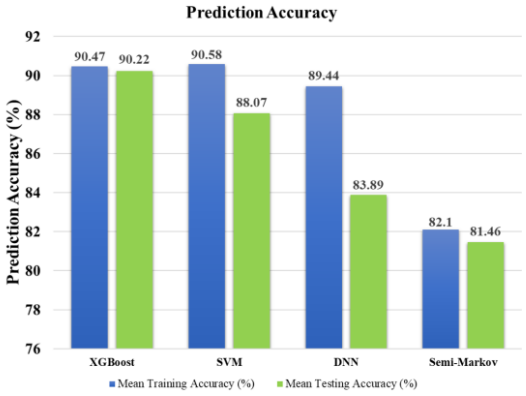

FIGURE 8 – Précision des prédicteurs

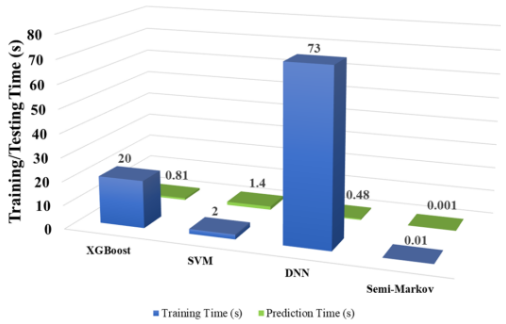

FIGURE 9 – Temps d'exécution

L'approche XGBoost a obtenu les meilleures performances en termes de précision de prédiction de mobilité, avec une précision moyenne de 90,22%.

En comparaison, l'approche SVM a également donné des résultats satisfaisants, bien que légèrement inférieurs à ceux de XGBoost.

D'autre part, l'approche DNN a montré des résultats prometteurs avec une précision supérieure à 80%, mais avec un temps d'entraînement plus élevé que les autres modèles.

Enfin, l'approche Semi-Markov a présenté une précision relativement inférieure, mais avec un temps d'exécution minimal, ce qui peut être crucial pour répondre aux exigences de qualité de service (QoS) strictes des réseaux cellulaires de 5G et au-delà.

6 Traitement du dataset

6.1 Preprocessing

Dans le dataset fourni par Shanghai Telecom, les informations indiquées sont le mois, la date, l'heure de début, l'heure de fin (en date time), Les coordonnées de longitude et de latitude de la station de base (base station) à partir de laquelle l'accès à Internet a été effectué et l'identifiant de l'utilisateur.

Au fil de l'évolution de notre projet, nous avons réalisé trois versions de preprocessing.

Première version

Nous avons commencé par :

- **Première version de gestion des valeurs manquantes** : Nous remplacons les valeurs manquantes des positions par les valeurs précédentes existantes.
- **Normalisation des coordonnées** : Les coordonnées de latitude et de longitude sont extraites, normalisées et centrées autour de zéro.
- **Encodage des positions (pos_id)** : Les positions géographiques sont encodées sous forme d'identifiants numériques (integer) pour pouvoir être utilisées comme caractéristiques d'entrée dans les modèles.
- **Création de séquences d'utilisateurs** : Les données sont organisées en séquences d'actions d'utilisateurs, avec des informations sur la position actuelle et la position suivante, ainsi que les caractéristiques d'entrée et les cibles de temps.

  User_1 : [[pos_1_1, time_1_1, additional_insights_1_1], [pos_1_2, time_1_2, additional_insights_1_2] ... [pos_1_n, time_1_n, additional_insights_1_n]]
- **Création d'insights supplémentaires** : Des caractéristiques supplémentaires telles que le mois, le jour, l'heure, la minute et la seconde correspondant à la date de début de connexion sont extraites des données temporelles pour fournir plus d'informations au modèle. De plus, de nouveau insights correspondant au temps de connection (end_time - start_time) sur une station et au temps de connection entre deux stations (next_start_time - end_time) pour un utilisateur sont rajoutés.
- **Encodage du temps** : Le temps est encodé en utilisant des fonctions trigonométriques (sinus et cosinus) pour capturer les aspects cycliques du temps, tels que les heures du jour, les mois de l'année, etc.
- **Création des cibles (pos_id_target et time_target)** : Les cibles sont préparées pour l'apprentissage supervisé.pos_id_target représente l'identifiant de la position suivante de l'utilisateur, et time_target représente les temps restants jusqu'à la fin de la connexion et jusqu'à la prochaine connexion.

Problème : Forte présence de données manquantes (fig 10)

Nous nous sommes rendus compte qu'il y avait énormément de données manquantes et que remplacer les positions manquantes par les précédentes pouvait avoir un impact très significatif sur nos modèles.

Où la dernière station rajoutée aux connexions ayant des données manquantes. On observe un **pic significatif** dans ce cas.

Deuxième version

Nous obtenons la deuxième version du preprocessing en supprimant les séquences d'utilisateurs contenant une ou plus d'une position manquante. (fig 11)

Nous observons donc ainsi **une balance de nombre de connexion par station meilleure que la précédente**.

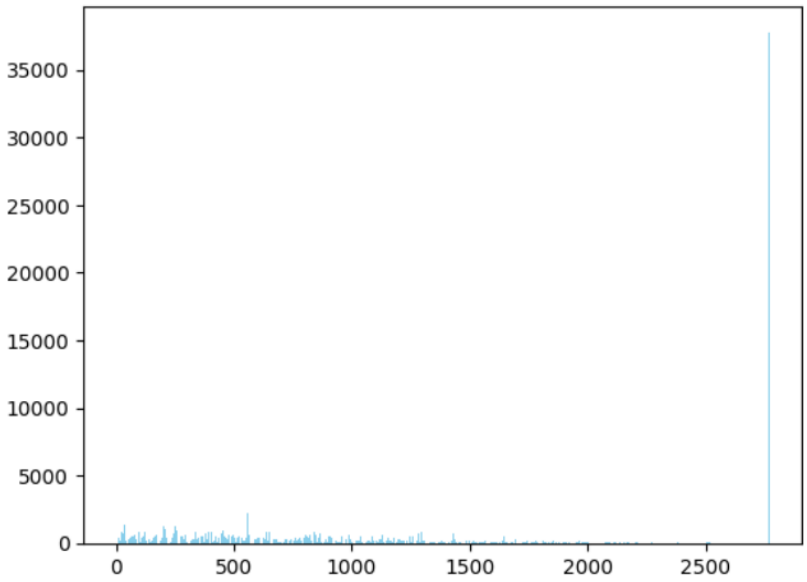

FIGURE 10 – Nombre de connexion pour chaque station v1

Problème : Forte présence de connexion successive à une même station (fig 12)

Par la suite, avec l'étude du comportement des utilisateurs nous nous sommes rendus compte que dans 72% des cas, la connexion suivante et la connexion précédente était effectuée sur la même station. Donc étant donné un utilisateur et sa position actuelle, on avait 72% de chance que la prochaine position dans le dataset soit la même (pos_id == pos_id_target)

Où le **modèle naïf correspondait au modèle qui prédit la même position étant donné un utilisateur à une position**.

Version finale

Les connexions répétées successives sont combinées pour former une seule, évitant ainsi d'avoir pour un utilisateur des successions de connexion à la même station. (fig 13)

Le prétraitement final de données effectué sur ce dataset comprend donc :

— **Gestion des valeurs manquantes (version 2)**
— **Normalisation des coordonnées**
— **Encodage des positions (pos_id)**
— **Création de séquences d'utilisateurs**
— **Gestion des connexions répétées**
— **Création d'insights supplémentaires**
— **Encodage du temps**
— **Création des targets** (pos_id_target et time_target)

6.2 Analyse et compréhension des patterns

Période de la journée

Nous avons analysé l'activité de connexion des utilisateurs tout au long de la journée et nous avons obtenus les résultats suivants : (fig 14, fig 15)

Compréhension

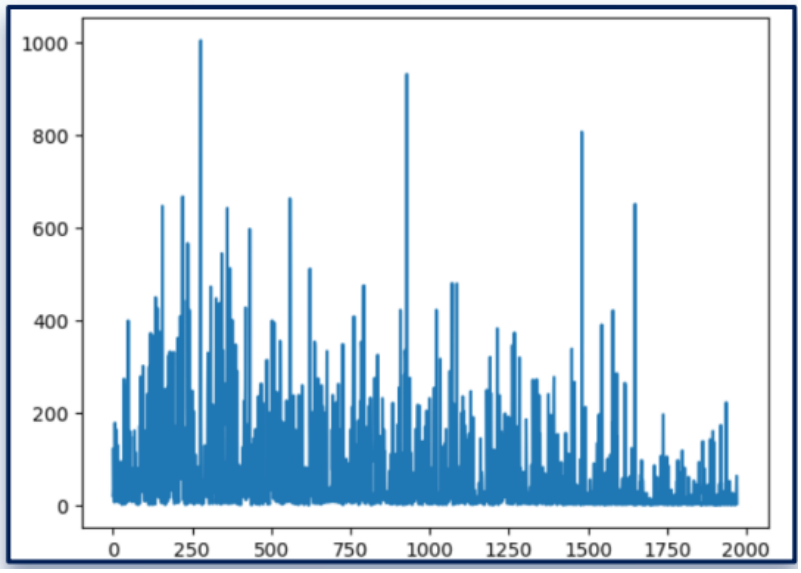

FIGURE 11 – Nombre de connexion pour chaque station v2

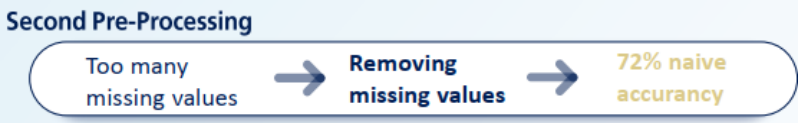

FIGURE 12 – Modèle naïf

Nous pouvons tirer des conclusions similaires à celles présentées dans l'article 1 par rapport à l'activité humaine à savoir :

- **Rythme quotidien typique** : L'analyse des connexions montre un schéma cohérent avec les habitudes quotidiennes des individus. Le pic d'activité le matin correspond probablement aux heures où les gens commencent leur journée de travail ou d'études, et ont donc besoin d'accéder aux services mobiles pour diverses tâches professionnelles ou personnelles.
- **Variations en fonction des heures de la journée** : Les variations dans le nombre de connexions tout au long de la journée reflètent les différentes activités humaines et les moments de repos. L'activité soutenue l'après-midi peut correspondre aux heures de travail prolongées, tandis que la baisse d'activité le soir et la nuit suggère que les gens se détendent ou dorment.

Jours de semaine et weekend

Tout comme l'article [1] le suggère, nous avons observés la variation de la tendande de connexion aucours de la semaine et du week-end et nous avons obtenus les résultats suivants : (fig 16, fig 17)

Compréhension

Nous constatons que les utilisateurs ont un ratio de nombre de connexion en semaine assez stable et uniforme contrairement au week-end où la variation n'est pas vraiment stable.

Cette variation peut s'expliquer par la **routine quotidienne en semaine** (tendance fixe) et **les loisirs et activités sociales le week-end** (pas de tendance particulière).

Regroupement en zone

Algorithme 1 : Threshold (fig 18)

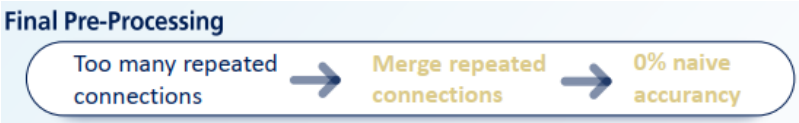

FIGURE 13 – Version finale

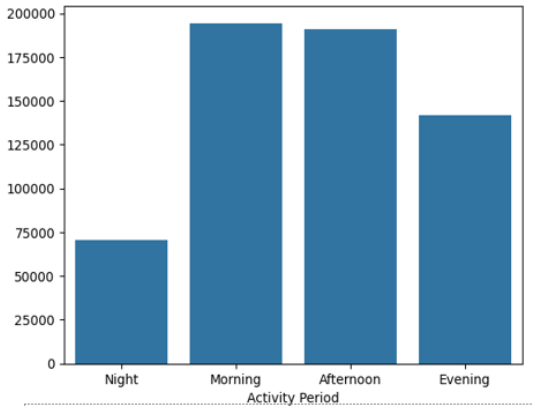

FIGURE 14 – Nombre de connexion par période

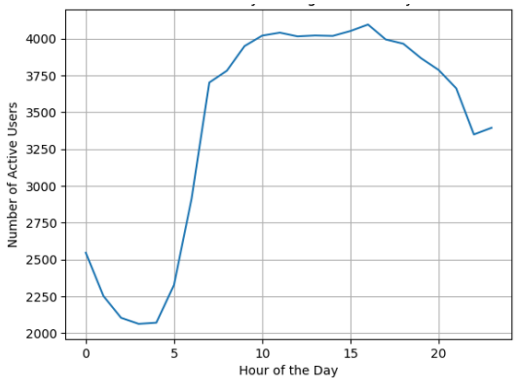

FIGURE 15 – Variation du nombre de connexion sur un jour

— Si la proportion le matin et l'après-midi est $\leq 40\%$, et la proportion le soir et la nuit est $\geq 30\%$, et la proportion maximale pendant les week-ends est $\geq 30\%$, alors cet endroit = "Zone de travail".
— Si la proportion le soir et la nuit est $\leq 40\%$, et la proportion du matin et de l'après-midi est $\leq 50\%$, et la proportion du matin et de l'après-midi pendant les week-ends est plus élevée que la proportion pendant la semaine, l'endroit est = "Zone résidentielle".
— Si les deux conditions ci-dessus ne sont pas remplies, l'endroit est = "Zone mixte".

Compréhension

En utilisant l'algorithme 1, on obtient :

Les résultats obtenus à partir de l'application de l'algorithme 1 sont répartis comme suit :
"Zone de travail" : 387 occurrences. "Zone résidentielle" : 252 occurrences. "Zone mixte" : 1879 occurrences.

Cela indique que parmi les données analysées, la plupart des occurrences ont été classées comme "Zone mixte", suivies par "Zone de travail" et "Zone résidentielle".

Ces résultats suggèrent que la majorité des observations correspondent à des zones mixtes, ce qui signifie qu'elles présentent des caractéristiques à la fois résidentielles et professionnelles.

Cependant, il y a également **un nombre significatif de zones de travail et de zones résidentielles identifiables**.

Algorithme 2 : DBSCAN (fig 19)

parameters : DBSCAN(eps=0.05, min_samples=35),
maximum distance of clustering :0.05, and the minimum number of samples : 35.
fit_predict() : cluster the standardized location data

On obtient :

Frequency of activity for each location during different time periods (weekends):

period	Night	Morning	Afternoon	Evening
pos_id				
0	37.837838	27.027027	16.216216	18.918919
1	1.562500	29.687500	26.562500	42.187500
2	21.052632	47.368421	26.315789	5.263158
3	18.000000	24.000000	32.000000	26.000000
4	6.521739	56.521739	26.086957	10.869565

FIGURE 16 – Ratio de connexion les jours de semaine

Frequency of activity for each location during different time periods (weekends):

period	Night	Morning	Afternoon	Evening
pos_id				
0	37.837838	27.027027	16.216216	18.918919
1	1.562500	29.687500	26.562500	42.187500
2	21.052632	47.368421	26.315789	5.263158
3	18.000000	24.000000	32.000000	26.000000
4	6.521739	56.521739	26.086957	10.869565

FIGURE 17 – Ratio de connexion le week-end

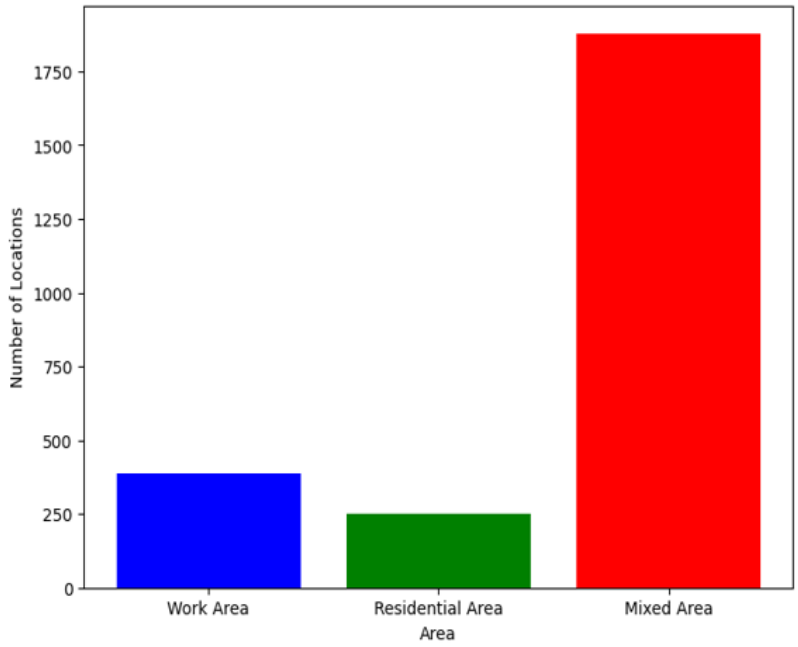

FIGURE 18 – Type de zone

Compréhension

On pourait tirer les même conclusions que précédemment.
Pour les perspectives possibles, on pourrait se concentrée sur la zone mixte pour tirer des conclusions utiles pour l'analyse.

6.3 Modèles de prédiction de where next

6.3.1 Neural Network models : LSTM et Transformers

Résultats correspondant à la deuxième version du preprocessing

Avant la fusions de connexions successives répetées, nous obtenons des scores de :
— LSTM : **70.75%**
— Transformers : **69.94%**

N.B : Ils se **rapprochent tous les deux de la prédiction naïve** qui était à 72.20% ;
Dans les cas où la station actuelle est différente de la prochaine station ; il y a à peu près **90% de chance que ces modèles ont tord** sur la prédiction.

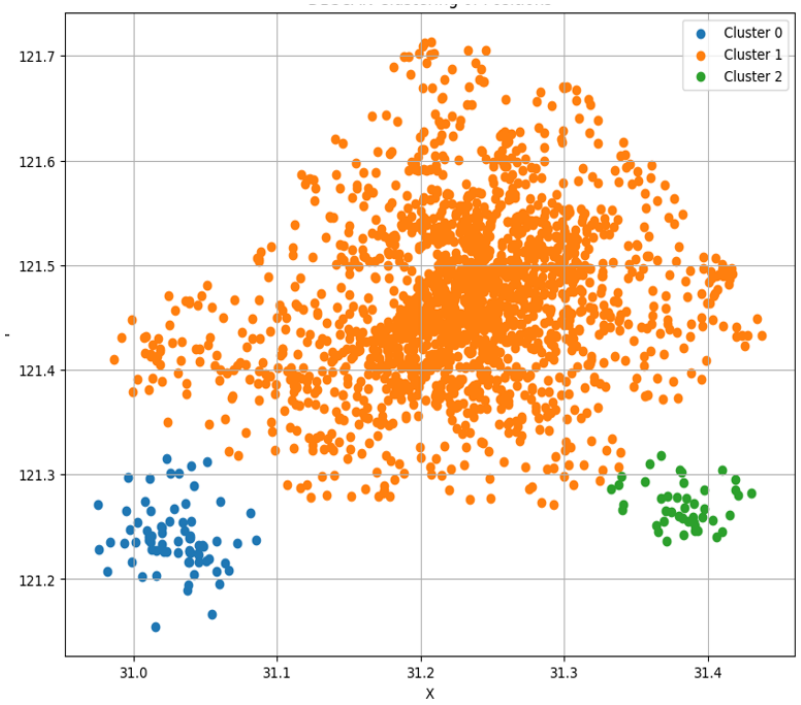

FIGURE 19 – Type de zone

Résultats correspondant à la dernière version du preprocessing

La recherche de meilleurs hyperparamètres est faite en utlisant la méthode d'optimisation Hyperopt basé sur les estimateurs de l'arbre Parzen.
En **mixant LSTM et Transformers**, nous obtenons un score de **44.62%**.

6.3.2 Base-tree models : Decision tree et Random forest

Particularités des inputs et outputs des arbres

- **Suppression des données avec moins de connexions** : Les données qui ont moins de connexions que le seuil spécifié nb_min sont supprimées pour garantir que les exemples de données utilisés pour l'apprentissage sont suffisamment informatifs.
- **Groupement des données par identifiant d'utilisateur** : Les données sont regroupées par identifiant d'utilisateur pour former des groupes homogènes. Cela permet de capter une certaine chronologie lors de l'analyse et l'entraînement de modèles de classification.

Le jeu de données final obtenu est donc de la forme :

- **Entrées** :
  - Les identifiants de position encodés (pos_id).
  - La latitude et la longitude (x et y) normalisées
  - Les identifiants d'utilisateur encodés (user_id)
  - Les variables de date et d'heure encodées si l'option date_time
  - Les temps jusqu'à la fin de la connexion et jusqu'à la prochaine connexion (time_to_end_and_target)
- **Sorties** : Identifiants de position cibles (pos_id_target) à prédire.

Résultats

Nous recherchons les meilleurs paramètres pour nos modèles (DecisionTreeClassifier et RandomForest) en utilisant une Grid Search CV à l'aide de la validation croisée.

Résultats correspondant à la deuxième version du preprocessing

Avant la fusions de connexions successives répetées, nous obtenons des scores de :

— Decision tree : **69.43%**

— Random forest : **70.81%**

N.B : Dans les cas où la station actuelle est différente de la prochaine station ; il y a à peu près 93% de chance que ces modèles ont tord sur la prédiction.

Résultats correspondant à la dernière version du preprocessing

Les meilleurs paramètres correspondant aux scores les plus élevés obtenus sont :

	Criterion	Max_depth	Min_samples_split	Splitter
Decision tree	gini	25	15	random

	n_estimators	min_samples_leaf	max_features
Random forest	50	5	0.7

Correspondant à des scores de :

— Decision tree : **37.29%**

— Random forest : **40.47%**

6.3.3 Commentaires

— Les **modèles obtenus en utilisant le deuxième preprocessing** ont plus de 90% tord lorsque la station actuelle est différente de la prochaine station (**une valid accuracy de moins de 10% dans ces cas là**) contrairement aux **modèles obtenus en utilisant le preprocessing final** qui ont **une valid accuracy comprise entre 37% et 44% dans ces cas là**.

Ce qui montre que **l'impact significatif du changement de preprocessing**.

— Bien que le modèle LSTM et transformers ait affiché les performances les plus élevées, il existe encore des opportunités d'amélioration et d'exploration dans plusieurs domaines.

D'abord, **une analyse plus approfondie des caractéristiques temporelles et spatiales** pourrait être entreprise pour découvrir des modèles plus complexes dans les données.

Ensuite, une **analyse plus poussée concernant la zone mixte** pourrait apporter des informations significatives dans le dataset et permettre une exploration plus approfondie des données séquentielles et spatiales, offrant ainsi des **insights informatifs**.

Enfin, avoir un **plus grand dataset ou un dataset plus informatif avec moins de données manquantes** pourrait contribuer à avoir de meilleurs performances. Dans notre cas, nous avons constaté que **50% des utilisateurs sont connectés une seule fois et 2/3 des utilisateurs ont moins de 7 connexions en 06 mois**. Ce qui montre que **beaucoup de données manquent à notre dataset qui déjà contient beaucoup de données manquantes**.

7 Comparaison avec les articles

Dans l'article [2] étudié, le meilleur meilleur modèle sur le jeu de données synthétiques était XGBoost. Dans notre cas, nous avons testé et le modèle performait très mal sur le jeu de données de la ville de Shanghai (moins de 10% sur le dataset final) ; raison pour laquelle nous n'avons pas présenté les résultats liés au modèle XGBoost.

En utilisant DNN comme modèle, avec l'article [2] on réussit à avoir un score de 83.89%, contrairement à notre cas où nous obtenons 44.62%.

Ces différences pourraient être dûes au fait que :

— **Différence significative de la taille du jeu de données** : Le jeu de données a été généré sur une semaine, et comporte 21 stations et 54 utilisateurs. Contrairement à notre cas où les données correspondent à une période de 06 mois ; avec 9481 différents utilisateurs, 3233 stations.

— **Données synthétiques =! Données réelles** : Le jeu de données utilisé dans l'article [2] est un jeu de données synthétique qui suit une certaine logique avec une granularité d'une minute et qui n'est pas bruité comme le notre.

Dans l'article [1], on réussit à **diviser les stations en 05 zones** (resident, transport, office, entertainment et comprehensive). Contrairement à notre cas où l'on a réussi à **diviser nos 3233 stations en 03 zones** (work area, residential area et mixed area).

Cette différence pourrait s'expliquer par le fait que **la zone mixed area qui contient un nombre significatif de stations contient les stations appartenant aux autres zones**. Cependant, il est assez difficile de différencier les zones. Une analyse plus poussée de cette zones pourrait améliorer la compréhension de notre jeu de données.

8 Conclusion

En conclusion, l'analyse des données de localisation et la modélisation prédictive du comportement des utilisateurs offrent des informations précieuses sur les modèles de mobilité urbaine. Bien que chaque modèle présente des niveaux de précision variables, le LSTM avec transformateurs apparaît comme le plus efficace, indiquant l'importance des architectures d'apprentissage en profondeur dans la capture de dépendances temporelles complexes. Les recherches futures pourraient approfondir l'analyse des zones mixtes et les caractéristiques spatio-temporelles, en affinant les modèles existants et en en explorant de nouveaux, y compris des modèles textuels pré-entraînés. On pourra ainsi répondre à de nombreux problèmes liés à la mobilité urbaine, la précision prédictive et à l'optimisation de l'allocation des ressources dans des environnements urbains dynamiques. En tirant parti des techniques avancées d'apprentissage automatique, nous pouvons mieux comprendre et répondre aux besoins changeants des populations urbaines, améliorant ainsi la planification urbaine et les services géolocalisés. Cependant, nous devons noter qu'il reste toujours assez difficile de prédire le déplacement des humains.

Reférences

[1]Understanding Mobile Traffic Patterns of Large Scale Cellular Towers in Urban Environment, Fengli Xu, Yong Li, Senior Member, IEEE, Huandong Wang, Pengyu Zhang, and Depeng Jin, Member, IEEE

[2] What Machine Learning Predictor Performs Best for Mobility Prediction in Cellular Networks ?, Hana Gebrie, Hasan Farooq and Ali Imran

[3]Yuanzhe Li, Ao Zhou, Xiao Ma, Shangguang Wang, Profit-aware Edge Server Placement, IEEE Internet of Things Journal, 2022, vol.9, no.1 ,pp.55-67 [PDF] [Sourcecode]

[4] Y. Guo, S. Wang, A. Zhou, J. Xu, J. Yuan, C. Hsu. User Allocation-aware Edge Cloud Placement in Mobile Edge Computing, Software : Practice and Experience, vol. 50, no. 5, pp. 489-502, 2020.[PDF] [Sourcecode]

[5] S. Wang, Y. Guo, N. Zhang, P. Yang, A. Zhou, X. Shen. Delay-aware Microservice Coordination in Mobile Edge Computing : A Reinforcement Learning Approach, IEEE Transactions on Mobile Computing, vol. 20, no.3, pp.939-953, 2021. [PDF]