Nouvelles méthodes d'apprentissage automatique pour la planification des réseaux mobiles

Danny Qiu

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2023IPPAS010

Thèse de doctorat

Nouvelles méthodes d'apprentissage automatique pour la planification des réseaux mobiles

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom SudParis

École doctorale n°626 Institut Polytechnique de Paris (EDIPP)
Spécialité de doctorat : Informatique

Thèse présentée et soutenue à Palaiseau, le 28/11/2023, par

DANNY QIU

Composition du Jury :

Laurent TOUTAIN Professeur, IMT Atlantique Campus de Rennes	Président, Rapporteur
Sidi Mohammed SENOUCI Full professor, Université de Bourgogne	Rapporteur
Samia BOUZEFRANE Professeure, Cnam Paris	Examinatrice
Fabrice CLÉROT Ingénieur de recherche, Orange Innovation	Examinateur
Hossam AFIFI Professeur, Télécom SudParis (SAMOVAR)	Directeur de thèse
Alassane SAMBA Ingénieur de recherche, Orange Innovation	Examinateur, Co-encadrant de thèse
Yvon GOURHANT Ingénieur de recherche, Orange Innovation	Invité, Co-encadrant de thèse
Hassine MOUNGLA Professeur, Université Paris Cité	Invité

Remerciements

Je tiens tout d'abord à remercier mes encadrants de thèse : Yvon Gourhant, pour avoir guidé mes travaux, et je suis reconnaissante d'avoir bénéficié de son expérience et de son savoir ; Hossam Afifi, mon directeur de thèse, sans qui nombre de réalisations n'auraient pu voir le jour ; et enfin Alassane Samba, pour sa confiance et sa sagesse.

Je remercie également mes rapporteurs, Laurent Toutain et Sidi Mohammed Senouci pour avoir accepté d'évaluer le contenu de ces trois années de recherche, ainsi que Fabrice Clérot, Samia Bouzefrane et Hassine Moungla pour avoir pris connaissance de mes travaux et assisté à ma soutenance.

J'adresse de chaleureux remerciements aux personnes d'Orange avec lesquels j'ai partagé de nombreux moments conviviaux, qu'ils soient ingénieurs, chercheurs, stagiaires, apprentis ou doctorants. Merci en particulier à toute l'équipe OSONS, qui m'a accueillie durant cette thèse et dont je suis ravie d'avoir fait partie ; et à l'équipe ARC, par laquelle tout a commencé.

Enfin, je remercie mes amis et ma famille. Merci pour vos encouragements et votre écoute en toutes circonstances. À mes parents et mon conjoint : merci pour votre présence, votre affection et votre soutien indéfectible.

Table des matières

1 Introduction **3**
- 1.1 La connectivité : un bien essentiel au développement . . . 3
- 1.2 Un réseau mobile de plus en plus sollicité . . . 3
- 1.3 Renforcer le réseau mobile : à chaque contexte sa stratégie de déploiement . . . 4
- 1.4 Apprendre des déploiements passés pour améliorer les décisions futures . . . 4

2 État de l'art **7**
- 2.1 Le réseau d'accès radio . . . 7
- 2.2 Apprentissage automatique . . . 11
- 2.3 Les données géo-spatiales . . . 22
- 2.4 Discussion . . . 27

3 Segmentation des sites en fonction de leur affluence **31**
- 3.1 Introduction . . . 31
- 3.2 Quantifier et qualifier les relations entre les données mobiles et le tissu urbain . . . 32
- 3.3 Méthode d'apprentissage pour la classification . . . 33
- 3.4 Résultats et interprétation . . . 41
- 3.5 Conclusion . . . 45

4 Réduction de dimension **47**
- 4.1 Introduction . . . 47
- 4.2 Données utilisées . . . 47
- 4.3 Relation entre l'AE linéaire et l'ACP . . . 49
- 4.4 Auto-encodeur convolutif . . . 51
- 4.5 Auto-encodeur convolutif variationnel . . . 54
- 4.6 Auto-encodeur convolutif variationnel discret . . . 57
- 4.7 Conclusion . . . 58

5 Prédiction du traffic hebdomadaire médian **61**
- 5.1 Introduction . . . 61
- 5.2 Travaux relatifs à la régression multi-cible . . . 61
- 5.3 Méthodes d'apprentissage pour la régression multi-cibles . . . 63
- 5.4 Résultats et interprétation . . . 65
- 5.5 Conclusion . . . 68

6 Modélisation de la couverture de service des cellules **69**
- 6.1 Introduction . . . 69
- 6.2 Travaux existants . . . 70
- 6.3 Problématiques de l'étude . . . 73
- 6.4 Mesures utilisées . . . 74
- 6.5 Formes géométriques utilisées . . . 76
- 6.6 Stratégies de dimensionnement . . . 79
- 6.7 Agrandissement de régions de Voronoi sur des données agrégées au niveau cellulaire . . . 85
- 6.8 Résultats . . . 86
- 6.9 Interprétation et discussion . . . 89
- 6.10 Conclusion . . . 90

7 Prédiction de l'impact du déploiement de nouvelles cellules **93**
7.1 Introduction . . . 93
7.2 Méthode d'apprentissage pour la prédiction de l'impact de la mise à jour d'un secteur . . 93
7.3 Résultats et interprétation . . . 101
7.4 Conclusion . . . 105

8 Priorisation des déploiements 5G à partir des prédictions de trafic et de débit moyen **107**
8.1 Introduction . . . 107
8.2 Méthode d'apprentissage pour ordonner les déploiements de la 5G . . . 108
8.3 Résultats et interprétation . . . 117
8.4 Conclusion . . . 124

9 Conclusion et perspectives **125**
9.1 Bilan . . . 125
9.2 Perspectives . . . 126

A Interface graphique : Déploiement de réseau mobile assisté par l'IA **129**

Acronymes

ACP Analyse en Composantes Principales. 49, 51, 53
AE Auto-Encodeur. 49
AECV Auto-Encodeur Convolutif Variationnel. 54–58
AECVD Auto-Encodeur Convolutif Variationnel Discret. 47, 57, 59
AP Average Precision. 100, 105
Arcep Autorité de régulation des communications électroniques. 8, 10, 126

CatBoost Categorical Boosting. 17, 31, 35, 41–43, 63, 65–67, 98, 114, 125
CDR Call Detail Record. 9
CNN Convolutional Neural Network. 15

D4D Data for Development. 9, 31, 33
DL Downlink. 10
DTW Dynamic Time Warping. 22, 66–68

EVS Explained Variance Score. 21, 66, 67

GBDT Gradient Boosted Decision Trees. 16, 17, 62–65, 67, 68, 103–105, 112, 114, 115, 117, 121, 122
GloVe Global Vectors for Word Representation. 37, 38

INSEE Institut National de la Statistique et des Études Économiques. 27, 82, 109, 119

KNN K-nearest neighbors. 18, 65, 66, 98, 103, 105

lasso Least Absolute Shrinkage and Selection Operator. 17, 112, 114, 117, 119, 121, 122
LightGBM Light Gradient Boosting Machine. 17, 63, 65, 98, 103–105, 114, 117, 119, 120, 125

MAE Mean Absolute Error. 21, 88, 114, 117, 121
MAPE Mean Absolute Percentage Error. 21
MAPL Maximum Allowable Path Loss. 82
MaxE Maximum Error (erreur maximale). 22, 66, 67
MedAE Median Absolute Error. 21
MLP Multilayer Perceptron. 65
MSE Mean Squared Error. 20, 22, 55, 66, 67

NDCG Normalized Discounted Cumulative Gain. 100, 101, 105, 108, 114, 117, 120, 121
nMAE Normalized Mean Absolute Error. 21, 66, 67

OSM OpenStreetMap. 26, 27, 36–39, 65, 109, 110, 119

PRB Physical Resource Block. 10, 80, 82, 87, 89, 93, 95, 103–105

RAN Radio Access Network. 7, 10

RF Random Forest (forêt aléatoire). 35, 42, 65, 66, 98
RMSE Root Mean Squared Error. 20, 21, 66, 67, 99, 103–105
RR Reciprocal Rank. 100, 101, 105

SHAP SHapley Additive exPlanations. vii, 42–45, 117
SMAPE Symmetric Mean Absolute Percentage Error. 21, 99, 104, 105
SMOTE Synthetic Minority Over-sampling Technique. 112, 117, 124
SMS Short Message Service. 9, 10
ST Single-Target Regression. 63, 65, 67, 68, 98
SVC Support Vector Classifier. 35, 42, 43
SVM Support Vector Machine. 18, 65, 98
SVR Support Vector Regressor. 98

UE User Equipment. 10
UL Uplink. 10

XGBoost eXtreme Gradient Boosting. 17, 62, 63, 65, 98, 103, 105, 114, 115, 121, 122, 125

Table des figures

2.1 Schéma des éléments d'un réseau mobile. 8
2.2 Illustration du surapprentissage 13
2.3 Processus d'apprentissage itératif appliqué à la classification 14
2.4 Représentation d'un arbre de décision à 8 feuilles, de profondeur 3 16
2.5 Illustration des premières étapes de l'algorithme des k-moyennes 19
2.6 Exemple de donnée représentée par un point 24
2.7 Exemple de donnée représentée par une ligne brisée 24
2.8 Exemple de données représentées par des polygones 24
2.9 Illustration de la fonction découpage de PostGIS 25
2.10 Illustrations de formes géométriques englobantes 26
2.11 Illustration de la fonction d'intersection de PostGIS 26
2.12 Architecture globale d'un outil de planification guidé par la donnée 29

3.1 Série temporelle construite à partir des statistiques d'appel et de SMS 34
3.2 Construction d'une signature hebdomadaire médiane 35
3.3 Exemple de données : Variables et cibles de deux stations de base 37
3.4 Dendrogramme des groupes d'attributs OpenStreetMap 40
3.5 Centres de classe des groupes de k-moyennes pour les stations de base de Dakar et de Thiès 41
3.6 Diagramme de Voronoi des positions des stations de base 42
3.7 Graphe beeswarm des valeurs SHAP des sites de Dakar 44
3.8 Graphes de décision SHAP de deux stations de base mal classifiées. 45

4.1 Échantillon de 9 observations, représentées en 1D et en 2D 48
4.2 Distribution des données mobiles dans l'espace réduit de l'ACP 50
4.3 Illustration générique d'un auto-encodeur 50
4.4 Architecture d'un auto-encodeur linéaire 51
4.5 Comparaison des distributions des observ. réduites par l'ACP et par l'auto-encodeur linéaire 52
4.6 Architecture d'un auto-encodeur convolutif 53
4.7 Comparaison des distributions des observations réduites par l'ACP et par l'AEC 54
4.8 Illustration de données générées par un auto-encodeur convolutif 54
4.9 Architecture de l'auto-encodeur convolutif variationnel 55
4.10 Comparaison de l'espace latent de l'AECV avec un échantillon de loi normale multivariée 56
4.11 Illustration de données générées par un AECV 56
4.12 Illustration d'un espace latent non informatif 57
4.13 Architecture d'un AEVD encodant les données réduites avec $m = 1$ vecteurs de taille 2 . . 58
4.14 Comparaison des distributions des observations réduites par l'ACP et par l'AEVD 59
4.15 Architecture d'un AEVD encodant les données réduites avec $m = 4$ vecteurs de taille 2 . . 59

5.1 Illustration de centres de classes obtenus avec les k-moyennes sur les signatures 62
5.2 Couverture des stations de base en Île-de-France modélisées par le diagramme de Voronoi 62
5.3 Architectures et flux de données des méthodes C2C et D2D. 64
5.4 Exemple d'une signature prédite par CatBoost nativement multi-cible et CatBoost-ST . . 68

6.1 Carte thermique des positions des utilisateurs collectées en Île-de-France. 71
6.2 Enveloppes concaves des positions d'utilisateurs rattachés à trois sites tri-sectorisés 72
6.3 Évaluation du modèle de couverture d'une cellule 74
6.4 Formes géométriques utilisées pour modéliser la couverture des cellules 76

6.5 Exemple d'un polygone de Voronoi divisé en trois régions sectorielles. 77
6.6 Influence de la longueur de normalisation sur la concavité des χ-shapes. 80
6.7 Redimensionnement des formes géométriques selon différentes stratégies 81
6.8 Précision et rappel des modèles obtenus sur les positions des utilisateurs 86
6.9 Courbes de la MAE en fonction du rapport d'agrandissement et de la fréquence 88
6.10 Précision et rappel du diagramme de Voronoi sectorisé à différentes échelles 89

7.1 « Effet vacances » en région parisienne vu à travers le trafic des stations de base 94
7.2 Chronologie des données collectées 94
7.3 De la création des données d'entraînement jusqu'à l'évaluation des modèles 96
7.4 Modélisation de la couverture des secteurs 97
7.5 Recherche du rapport d'agrandissement optimal 97

8.1 Recherche du rapport d'agrandissement paramétrant le modèle de couverture 109
8.2 Frontières administratives et cellules 4G/5G du sud-ouest de la France 110
8.3 Nuages de points des prédictions de LightGBM en fonction des cibles 118
8.4 Scores NDCG du classement des secteurs établi en fonction du trafic 5G attendu 119
8.5 Interprétation d'un LightGBM à travers des prédictions du trafic 5G avec SHAP 119
8.6 Nuage de point des prédictions du modèle XGBoost par rapport aux cibles 121
8.7 Scores NDCG du classement des secteurs établi en fonction du débit moyen 4G/5G attendu 122
8.8 Interprétation d'un XGBoost à travers des prédictions du débit moyen 4G/5G avec SHAP 123

9.1 Architecture d'un système de planification guidé par la donnée 127

A.1 Captures d'écran de l'interface graphique 130
A.2 Capture d'écran d'un polygone dessiné avec l'outil Mapbox Draw. 131
A.3 Capture d'écran de graphes affichés avec plotly.js. 131
A.4 Exemple de valeurs prises par les variables passées aux modèles prédictifs 132

Liste des tableaux

3.1 Méthodes utilisées dans la littérature pour étudier l'activité des réseaux mobiles 32
3.2 Hyperparamètres utilisés pour chaque modèle d'apprentissage 36
3.3 Clé et valeurs utilisées pour le calcul des variables 37
3.4 Liste des valeurs des attributs OSM 39
3.5 Proportion des effectifs de classe obtenus avec les k-moyennes 41
3.6 Moyenne et écart-type des mesures de classification des stations de base de Dakar 42
3.7 Moyenne et écart-type des mesures de classification des stations de base de Thiès 45

5.1 Mesures des erreurs de prédictions des modèles C2C 66
5.2 Mesures des erreurs de prédictions des modèles D2D 66

6.1 Nombre de positions utilisateurs collectées 70
6.2 Résumé des méthodes appliquées et des problèmes abordés 73
6.3 Bilan de liaison dans le sens montant 83
6.4 Rapport optimal et MAE minimaux en mètres 88
6.5 Paramètres de dimensionnement des couvertures des cellules 90
6.6 Précision/rappel du diagramme de Voronoi sectorisé redim. avec un rapport d'agrand. optimisé 91

7.1 Entrées et cibles des données d'entraînement 98
7.2 Hyperparamètres des SVMs 99
7.3 Hyperparamètres des KNNs 99
7.4 Hyperparamètres des modèles ensemblistes 99
7.5 Exemple de priorisation du refarming du spectre 2100 MHz pour des secteurs LTE 100
7.6 Médiane des taux de croissance des indicateurs par configuration existante et ajoutée . . . 102
7.7 Valeurs-p du test de Mann-Whitney U sur la distrib. des taux de croiss. des indic. de perf. 102
7.8 RMSE des prédictions des composantes 104
7.9 SMAPE des prédictions des composantes 104
7.10 Évaluation des classements obtenus à partir des prédictions des taux de croissance 105

8.1 Résumé des variables utilisées pour l'entraînement 111
8.2 Hyperparamètres des modèles entraînés pour prédire le trafic 5G 115
8.3 Hyperparamètres des modèles entraînés pour prédire le débit moyen 4G/5G 116
8.4 MAE des prédictions de trafic 5G sur les données test 117
8.5 Coefficients non nuls du lasso entraîné à prédire le trafic 5G 119
8.6 Liste des variables importantes sélectionnées avec LightGBM-Ideal 120
8.7 MAE des prédictions de débit moyen 4G/5G sur les données test 121
8.8 Vingt premiers coefficients non nuls du lasso entraîné à prédire le débit moyen 4G/5G . . 123

Chapitre 1

Introduction

1.1 La connectivité : un bien essentiel au développement

Prise pour acquis, discrète mais omniprésente dans les pays développés, le rôle essentiel de la connectivité s'est brutalement rappelé aux Français le 17 mars 2020, date du premier confinement visant à freiner la pandémie mondiale de la Covid-19. Grâce à Internet, les possibilités de télétravail, d'études à distance, de courses à domicile et de visioconférence ont permis de mitiger l'impact économique, éducatif et de préserver une forme de lien social.

Pour autant, l'absence de connectivité a également exacerbé les inégalités. En 2022, l'Union internationale des télécommunications (UIT) rapportait qu'un tiers de la population mondiale n'avait pas accès à Internet [1]. Lors des restrictions sanitaires, ce sont des milliards d'individus dans l'impossibilité de télétravailler et d'écoliers dont la scolarité s'est brutalement arrêtée [2]. La connectivité est un moteur important de la croissance économique, c'est pourquoi l'Organisation des Nations unies (ONU) a établi une feuille de route pour parvenir à une connectivité universelle d'ici 2030 [3]. L'ONU définit deux axes de développement : l'universalité et l'efficacité. L'universalité vise à permettre l'accès à tous d'Internet, tandis que l'efficacité vise à fournir une connexion fiable, sécurisée et rapide. Conjointement avec le réseau fixe, le développement du réseau mobile permettra d'atteindre ces deux critères.

Grâce aux réseaux mobiles 4G/5G et à la démocratisation du smartphone, les utilisateurs peuvent profiter des mêmes loisirs que sur un ordinateur fixe, depuis n'importe quel endroit : streaming de vidéos et de musique, jeux en ligne, messagerie instantanée, réseaux sociaux... Mais surtout, le téléphone connecté est un vecteur d'information qui facilite grandement les mobilités : services de cartographie, information de trafic en temps réel, réservation de co-voiturage, autant de possibilités qui permettent planifier les trajets et de trouver des alternatives en cas d'incident pendant le déplacement. Enfin, les modes de paiement dématérialisés et les services de banque en ligne accessibles sur le smartphone améliorent la sécurité des transactions, facilitent les achats et la consommation.

1.2 Un réseau mobile de plus en plus sollicité

Le réseau mobile est une architecture permettant d'acheminer une information d'un point à un autre, et dont la connexion finale avec l'utilisateur se fait via les ondes radio. Grâce aux progrès techniques, il a connu de nombreuses générations technologiques qui ont accompagné les besoins des utilisateurs. Déployée à partir de 2009, la majorité des connexions haut-débit actuelles sont réalisées avec la 4G [4].

Tout réseau possède une capacité finie, c'est-à-dire qu'il prend en charge un nombre simultané d'utilisateurs et transmet une quantité d'information limités durant une période donnée. Plus il y a d'utilisateurs et plus le besoin est important, plus la qualité de service est dégradée. Elle peut parfois même être simplement insuffisante pour certains usages. Une qualité de service dégradée se ressentira des temps de chargement de pages plus longs, des vidéos saccadées, des difficultés à passer des appels, etc.

Selon les projections actuelles, l'usage du smartphone continuera de croître partout dans le monde. On estime que 70% du trafic mondial est attribuable au visionnage de vidéos [5], en grande partie dominée par le streaming. Ce chiffre devrait passer à 80% d'ici 2028. L'usage des réseaux sociaux (hors vidéos) constitue quant à lui 9% du trafic global. Pour répondre à ces attentes, les solutions qui s'offrent aux opérateurs sont les suivantes : le renforcement de la couverture 4G, le déploiement d'une nouvelle génération de réseau possédant une plus grande capacité, ou une combinaison des deux. Commercialisée depuis 2019, la 5G est la nouvelle génération des réseaux mobiles qui devra améliorer la qualité de service existante.

La version *standalone* (SA) exploitant les fréquences 26 GHz permettra de multiplier par dix le débit maximal perçu par l'utilisateur et de diviser la latence par dix [6].

1.3 Renforcer le réseau mobile : à chaque contexte sa stratégie de déploiement

Le renforcement ou le déploiement d'une couverture réseau est réalisé grâce à l'installation de nouveaux équipements. Dans le cadre de la thèse, on s'intéresse aux déploiements des stations de base et des antennes, qui composent le réseau d'accès radio. Il s'agit de la partie du réseau qui reçoit les communications des terminaux des utilisateurs et les transmet au reste du réseau (et inversement). Le déploiement du réseau d'accès est très coûteux ; aussi bien au niveau matériel (déploiement de nouveaux supports d'antennes, raccordement au reste du réseau) qu'au niveau de l'acquisition des fréquences. En France par exemple, l'attribution des fréquences est réalisée par l'Autorité de régulation des communications électroniques (Arcep) à l'issue d'enchères [7]. L'autorité administrative rapporte que l'investissement des opérateurs dans les licences mobiles s'est élevé à 2,8 milliards d'euros en 2020, et qu'en moyenne, 1,8 milliards d'euros ont été investis par an sur la période 2016-2020 dans les boucles locales 4G/5G [8]. Par opérateur mobile, les dépenses d'investissement (CAPEX) et d'exploitation (OPEX) du réseau d'accès 4G/5G représentent 45 à 50% du coût total de possession [9]. Pour ces raisons, une grande attention est accordée à la planification afin d'identifier les sites les plus rentables, c'est-à-dire présentant une demande importante de connectivité. Dans le même temps, pour être compétitifs, les opérateurs mobiles se doivent de fournir la meilleure qualité de service possible.

De plus, derrière les chiffres globaux sur la croissance des usages, la situation du réseau mobile et les projections attendues sont différentes selon les régions du monde. Pour cette raison, chaque contribution de la thèse portera sur un mode de déploiement différent, qui répondra à des besoins et des situations particulières. On prendra pour zones d'études des régions du Sénégal et de la France.

L'Afrique subsaharienne, l'Asie-Pacifique et l'Amérique latine seront les régions qui verront la plus forte croissance du taux d'adoption de smartphones, ces derniers devenant plus abordables [10]. Les nouveaux utilisateurs viendront d'une population jeune, qui a grandi avec Internet. En 2022, l'Union internationale des télécommunications (UIT) rapporte qu'en Afrique, 55% des 15-24 ans utilisent internet contre 36% pour le reste de la population [1]. Ce sont les usages de ces « enfants du numériques » qui conduiront l'extension de la couverture 4G/5G pour développer l'accès au haut débit. Suite à la crise de la Covid-19, la 5G a connu un amorçage précoce en Afrique du Sud. Au Sénégal, une partie du fixe haut-débit transite par le réseau 4G, et pourrait être renforcé dans un futur proche, par la 5G *Fixed Wireless Access*. Il est en effet projeté que la 5G sera déployée courant 2023 au Sénégal [11].

En Europe, les smartphones et ses usages sont déjà bien implantés. En France par exemple, selon le baromètre du numérique édition 2022 [12], 87% des Français de plus de 12 ans sont équipés d'un smartphone. C'est l'équipement majoritairement utilisé pour se connecter à Internet (47%, contre 39% pour l'ordinateur). Cependant, le rapport souligne aussi la différence de satisfaction des abonnés sur la qualité des appels, messages et navigation Internet (80% de satisfaction) en comparaison avec les appels vidéos et les jeux en ligne (60% de satisfaction). Ainsi, les problématiques des opérateurs mobiles seront tournées vers la densification de la couverture 4G et le déploiement de la 5G pour augmenter la capacité du réseau. En 2018, au moment de la ré-attribution des fréquences 900MHz, 1800MHz et 2100MHz [13], l'Arcep met en place le *New Deal*. Il s'agit d'un accord passé en 2018 entre les opérateurs mobiles, l'Arcep et le gouvernement, dans lequel les opérateurs s'engagent à accélérer le déploiement à minima de la 4G dans 20 000 zones prioritaires.

1.4 Apprendre des déploiements passés pour améliorer les décisions futures

L'objectif de la thèse est d'adapter les méthodes d'apprentissage automatique aux problématiques de déploiement du réseau mobile. On distingue deux formes de déploiement : l'extension de couverture et l'augmentation de la capacité. Pour chacune, il faut réfléchir à la manière dont les modèles pourraient apprendre des déploiements passés pour aider à anticiper les stratégies de déploiements futurs.

L'apprentissage automatique (de l'anglais *machine learning*) est un vaste domaine d'étude de par la grande variété des méthodes d'apprentissage, des modèles et des applications. Les progrès techniques ont permis de doter les ordinateurs d'une grande puissance de calcul, pour résoudre des tâches telles que la

reconnaissance d'objets dans le traitement de l'image, le traitement du langage naturel (domaine dont font partie les agents conversationnels), la détection d'anomalies ou la génération de données synthétiques. La force des modèles d'apprentissage réside dans leur faculté à traiter un volume conséquent de données afin d'extraire et de modéliser des relations entre des phénomènes, sans connaissance à priori des règles qui les régissent. Le fonctionnement du réseau mobile génère un volume abondant de données sur le trafic, l'affluence, la charge ou encore le débit observable au niveau des stations et des cellules, et chaque nouveau déploiement apporte une quantité de données supplémentaires. En sauvegardant ces informations, on peut constituer un historique des performances du réseau au fur et à mesure des déploiements successifs. Grâce à ces données et à l'apprentissage automatique, on espère pouvoir anticiper plus dynamiquement l'évolution de la demande d'un territoire suite à la création de nouveaux aménagements. Par exemple, on pourrait mieux anticiper le dimensionnement capacitif nécessaire suite à la construction d'un nouveau centre commercial, voire d'un pôle urbain.

La première forme de déploiement étudiée, l'extension de couverture, consiste à déployer des stations de base et des antennes sur de nouveaux emplacements. Pour améliorer les décisions de déploiement, le premier volet de la thèse est dédié à l'entraînement des modèles d'apprentissage automatique pour estimer l'activité des futures stations de base. La difficulté du travail est d'estimer cette affluence sur la base de données qui devront être disponibles avant le déploiement ; dans les cas d'étude, on suppose qu'on ne possède aucune connaissance sur les performances du réseau. La solution est de s'appuyer sur des données externes pour décrire les activités humaines et inférer les usages réseaux aux endroits non couverts. Dans le Chapitre 3, on s'attache à prédire la classe d'activité des futures stations de base. Par la suite, on étend ces travaux en cherchant à prédire des données plus précises. Les temps d'entraînement parfois longs de certains modèles ont conduit à l'étude et à la mise au point d'un modèle de réduction de données (auto-encodeur) dans le Chapitre 4. Ce modèle est ensuite utilisé dans le Chapitre 5, où l'objectif est d'estimer le trafic hebdomadaire médian des futurs stations de base. Ces travaux ont débouché sur l'implémentation d'une interface graphique pour démontrer comment ces modèles peuvent être intégrés comme des outils d'aide au déploiement (anticipation de la demande et du dimensionnement), en prenant comme cas d'étude la région de Dakar au Sénégal.

La deuxième forme de déploiement étudiée est la densification du réseau. Contrairement à l'extension de couverture, la densification consiste à déployer de nouvelles cellules sur des sites existants pour augmenter la capacité de ces derniers. Pour réaliser les dernières études de la thèse, il a été nécessaire d'employer un modèle de couverture plus précis, développé et évalué dans le Chapitre 6. Dans le Chapitre 7, on commence par étudier l'efficacité des déploiements ré-utilisant des bandes de fréquences 2G/3G pour améliorer les performances des cellules 4G existantes. Dans le Chapitre 8, on étudie ensuite comment prioriser le déploiement des cellules 5G sur des sites 4G, grâce à la prédiction du volume de trafic 5G et du débit moyen utilisateur attendus après l'augmentation de capacité.

Publications

Conférences

- Qiu, D., Samba, A., Afifi, H., & Gourhant, Y. (2021). Classifying urban fabrics into mobile call activity with supervised machine learning. In 2021 International Wireless Communications and Mobile Computing (IWCMC) (pp. 1948–1953). doi :10.1109/IWCMC51323.2021.9498606
- Qiu, D., Samba, A., Afifi, H., & Gourhant, Y. (2022). Transforming Urban Fabric into Mobile Call Traffic signatures. In 2022 IEEE International Conference on Communications (ICC) (pp. 377–382). doi :10.1109/ICC45855.2022.9838654
- Qiu, D., Lavergne, M., Samba, A., Afifi, H., & Gourhant, Y. (2023). Estimating the impact of adding new cells on the traffic of neighboring cells. In GLOBECOM 2023 - 2023 IEEE Global Communications Conference.

Journal

- D. Qiu, A. Samba, H. Afifi and Y. Gourhant, "A Study on Simple Geometries for Modelling User Equipment Geospatial Attachment to Mobile Cells," in IEEE Access. doi : 10.1109/ACCESS.2023.3315129.

Brevets

— Qiu, D., Samba, A., & Gourhant, Y. « Procédé de gestion de ressources d'au moins un équipement d'accès d'un réseau de télécommunications, dispositif, équipement d'accès, équipement de contrôle, système et programmes d'ordinateur correspondants. ». Référence n°FR3128348, publié le 21/04/2023.

— Qiu, D., Samba, A., & Gourhant, Y. « Procédé d'évaluation d'un déploiement d'une configuration candidate d'un équipement d'accès radio d'un réseau de télécommunications, dispositif, système de gestion et programme d'ordinateur correspondants ». Déposé le 28/10/2022.

Chapitre 2

État de l'art

À l'ère du numérique, la collecte et l'analyse des données deviennent incontournables pour observer et étudier les événements, développer de nouveaux modèles pour améliorer notre compréhension et anticiper des phénomènes variés, qu'ils soient de nature biologique, météorologique, financière, sociologique au autre encore. Le domaine des télécommunications n'y échappe non plus ; lorsqu'elles sont analysées, les données produites par le réseau mobile permettent de se présenter une idée globale de leur fonctionnement grâce aux outils statistiques. Ces chiffres permettent de rendre compte et d'orienter des décisions stratégiques, à la fois en matière de marketing et de planification.

Dans la Section 2.1, on commencera par donner une définition des équipements du réseau d'accès radio concernés par les travaux de la thèse, ainsi qu'une description de la nature des données collectées pour l'apprentissage automatique. La Section 2.2 introduit les concepts d'apprentissage automatique qui positionnent la thèse et/ou qui seront utilisés par la suite, ainsi que les usages qui sont faits de ces modèles appliqués aux données de télécommunication. Cependant, l'utilisation seule des données du réseau ne permet pas d'entraîner des modèles à estimer les performances futures de sites qui ne sont pas encore déployés. Pour pallier cette difficulté, l'idée générale est d'utiliser des données externes corrélées aux activités humaines, qui sont disponibles aux endroits où l'on souhaite réaliser des déploiements, et de faire apprendre aux modèles le lien entre ces données externes et les performances du réseau. La Section 2.3 présente quelques notions associées aux données géographiques ainsi que les sources de données utilisées. Pour autant que l'on sache, la littérature sur la planification des réseaux mobiles entièrement guidée par l'apprentissage automatique et la donnée (en opposition à l'utilisation de simulateurs) est assez peu fournie. La Section 2.4 termine l'état de l'art en positionnant cette thèse par rapport aux quelques travaux existants, en présentant le paradigme de planification dans lequel ces travaux, ainsi que celui de la thèse, s'inscrit.

2.1 Le réseau d'accès radio

Pour connaître l'état de santé du réseau mobile, les performances ou encore les taux d'utilisation et d'affluence, des sondes sont installées au niveau du réseau d'accès radio. Dans cette section, on présente d'abord le vocabulaire employé pour désigner les équipements, puis la nature des données utilisées par les travaux de la thèse et enfin, les raisons justifiant l'intérêt porté par le secteur des télécommunications pour l'apprentissage automatique.

2.1.1 Vocabulaire du réseau mobile

Réseau mobile Une représentation simple de l'architecture du réseau mobile est de le distinguer en deux parties : le **réseau d'accès radio** (RAN : *Radio Access Network*) et le **cœur de réseau** auquel il est raccordé (Figure 2.1). Le RAN correspond à l'interface sans fil du réseau qui reçoit les communications provenant des équipements utilisateurs (UE : *User Equipment*) et les transmet au cœur de réseau, et inversement. L'UE peut être par exemple un smartphone, tablette ou un ordinateur. Le cœur de réseau connecte le RAN au reste du réseau mobile et aux réseaux externes comme Internet. L'architecture précise des réseaux mobiles, les équipements et leurs fonctions dépendent de la technologie considérée. Pour plus de détails sur les réseaux 2G, 3G, 4G et 5G, nous référons le lecteur aux cours du professeur Frédéric Launay [14-17]. Par la suite, on s'intéresse uniquement au RAN, et plus précisément à la station de base et aux antennes qui sont les deux équipements concernés par les données collectées.

FIGURE 2.1 – Schéma des éléments d'un réseau mobile.

L'antenne est l'émetteur-récepteur qui transmet et reçoit les signaux électromagnétiques entre l'équipement utilisateur et le reste du réseau mobile. La zone de service, ou couverture d'une antenne, correspond à la région géographique dans laquelle un utilisateur peut être rattaché à l'antenne. Celle-ci est raccordée à une station de base.

La station de base Dans la thèse, on considère simplement la station de base comme une infrastructure constituée d'un mât sur lequel sont déployées des antennes, et qui possède des fonctions pour le traitement des signaux radios. Il est courant de déployer les antennes par groupe de trois (on dit qu'elles sont trisectorisées). Chaque antenne est orientée une direction différente, appelée également azimut. On utilise le mot **site** pour désigner l'emplacement géographique de la station de base et des antennes associées.

Cellule On emploie le terme de cellule pour désigner la couverture d'une antenne transmettant à une fréquence et dans une direction donnée. Il s'agit d'une des échelles les plus fines à laquelle on dispose des données mobiles ; on parlera par exemple du nombre d'utilisateurs rattachés à chaque cellule, le volume de données y transitant, le débit fourni... En outre, le croisement des données mobiles avec des données externes se fera au moyen d'un modèle de couverture de service des cellules.

Secteur On définit le secteur comme étant l'ensemble des antennes transmettant dans la même direction.

Les bandes de fréquence auxquelles transmettent les antennes diffèrent selon la technologie considérée. L'usage de plusieurs fréquences permet d'augmenter la capacité du réseau mobile. Cela a pour conséquence d'augmenter le nombre d'utilisateurs pris en charge simultanément, ainsi que la quantité de ressources allouées à chacun d'entre eux. Les fréquences de bande attribuées aux opérateurs ainsi que les largeurs de bande dépendent de la génération de la technologie, et varient selon la zone géographique. Dans les Chapitres 7 et 8, on travaille sur des données cellulaires d'un opérateur français, distinguées par bande de fréquence. Les fréquences concernées par les travaux sont les suivantes :

- la bande des 700 MHz, attribuée en 2015 à la 4G et la 5G.
- la bande des 1800 MHz. Historiquement attribuée à la 2G, les opérateurs mobiles ont obtenu progressivement l'autorisation depuis 2013 d'utiliser cette bande pour la 4G.
- la bande des 2100 MHz est historiquement attribuée à la 3G mais depuis 2017, les opérateurs mobiles sont autorisés à utiliser cette bande pour la 4G et la 5G.
- les bandes des 800 MHz et 2600 MHz sont attribuées à la 4G dès son introduction.
- la bande des 3500 MHz est attribuée à la 5G dès son introduction.

Pour les fréquences où l'usage peut être mixte, le choix de déployer une ou plusieurs technologies peut varier selon les régions, en fonction des stratégies de couverture adoptées par les opérateurs.

Largeur de bande Pour chaque bande de fréquence, chaque opérateur dispose d'une largeur de bande variable résultant des enchères organisées par l'Autorité de régulation des communications électroniques (Arcep). Du 21 août 2021 au 8 février 2025 [18], l'Arcep alloue une largeur de bande par opérateur :

- de 10 MHz à 20 MHz selon l'opérateur pour la bande 700 MHz.
- 20 MHz pour la bande 800 MHz.
- de 29.6 MHz à 40 MHz pour les bandes 1800 MHz, 2100 MHz et 2600 MHz.
- de 70 MHz à 90 MHz pour la bande 3500 MHz.

La largeur de bande est un critère important de la qualité de service : plus elle est large, plus le débit utilisateur maximal est important. En effet, le théorème de Shannon-Hartley montre qu'il est possible de transmettre une information de manière fiable dans un canal de transmission bruité si ce débit ne

dépasse pas une borne maximale, qu'on appelle la capacité [19]. En notant P la puissance du signal et N la puissance du bruit, la relation entre la largeur de bande W et la capacité C est :

$$C = W \log_2 \left(1 + \frac{P}{N}\right)$$

2.1.2 Données du réseau mobile étudiées

Les données mobiles à notre disposition sont collectées à deux échelles : la cellule et l'équipement utilisateur. On qualifie ces données de temporelles, ou se présentant sous forme de séries temporelles, car elles sont collectées de manière périodique.

Données publiques

De nombreux travaux académiques sur les mobilités humaines et l'étude de l'activité des réseaux ont pu voir le jour grâce à la publication de données de la part des opérateurs mobiles. Ces données concernent principalement les statistiques d'appel.

Les statistiques d'appel (CDR : *Call Detail Record*) sont des données utilisateur collectées au niveau des stations de base. Lorsqu'un abonné passe un appel, les données sur le début de l'appel, sa durée, la station de base auquel il est rattaché, et divers autres indicateurs sont collectés. En agrégeant les statistiques d'appel de tous les utilisateurs par station de base ou par cellule, on obtient une série temporelle sur le nombre d'utilisateurs connectés d'une granularité très fine. Les statistiques d'appel ne permettent pas de connaître l'emplacement exact d'un utilisateur. En revanche, il est commun d'approximer sa position en supposant qu'il est situé dans la zone couverte par la station de base, que l'on modélise avec le diagramme de Voronoi. Cela permet de créer des cartes du trafic du réseau mobile à la granularité spatiale de la couverture des sites [20]. Dans le Chapitre 6, on étudie la précision de ce diagramme par rapport à un échantillon de positions réelles d'utilisateurs, et on propose des paramètres pour améliorer le modèle.

Exemples célèbres de données publiques

- *D4D-Challenge* [21] : En 2012, l'opérateur mobile Orange lance le défi « *Data for Development* » (D4D). En publiant des statistiques d'appels et de SMS anonymisés collectées sur une durée de 5 mois en Côte d'Ivoire, l'objectif est d'acquérir une meilleure compréhension des mobilités humaines, des dynamiques socio-économiques et de l'activité des réseaux mobiles.
- *D4D-Senegal* [22] : En 2014, le défi D4D est réitéré avec des données mobiles du Sénégal. Comme pour le précédent challenge, les données publiées sont des statistiques d'appels et de SMS anonymisés. Cette fois-ci, la période de collecte est d'un an.
- *Milan dataset* [20] : Dans le souci de proposer un jeu de données public, exhaustif et servant de référence pour la reproductibilité des travaux de recherche, un ensemble d'universités et d'industriels italiens a publié ce que l'on surnomme le « *Milan dataset* ». Ce jeu de données carroyées se présente sous forme de couches de données provenant de sources diverses : météo, télécommunications, réseaux sociaux, réseau électrique, actualités. Les deux régions décrites par ces données sont la ville de Milan et la province autonome de Trente. Les données de mobiles disponibles correspondent à deux mois de statistiques d'appels.

Dans le domaine des télécommunications, ces données publiques ont abouti à un grand nombre de travaux portés sur la prédiction dynamique du trafic grâce à l'apprentissage automatique [23-26]. Les données des défis D4D ont également débouché sur plusieurs études de mobilité des populations [27, 28].

Données privées

Le nombre d'utilisateurs permet de rendre compte de la demande de connectivité et de l'affluence au niveau d'un site. Dans le contexte du déploiement, il est également important de considérer d'autres indicateurs de performance pour dimensionner le réseau. Pour cette raison, on s'intéresse également aux indicateurs suivants :

- le taux d'usage des blocs de ressources qui reflète la charge des cellules
- le débit moyen utilisateur qui est un indicateur de qualité de service
- le volume de trafic des données dans le sens descendant qui permet de rendre compte du besoin de connectivité des utilisateurs

La liaison montante/descendante (UL/DL : *uplink/downlink*) caractérise le sens dans lequel se déroule la transmission radio entre le RAN et l'UE :

- On parle de liaison descendante lorsque l'émetteur est la station de base et le récepteur est l'équipement utilisateur. Il s'agit du sens sollicité par l'utilisateur pour la navigation Internet, le visionnage de vidéos, la musique en streaming ou le téléchargement de fichiers.
- On parle de liaison montante lorsque l'émetteur est l'équipement utilisateur et le récepteur est la station de base. Ce sens est sollicité par les applications de visio-conférence, la diffusion de vidéos en temps réel, le téléversement de photos et vidéos.

Les indicateurs de performance de charge, débit et trafic sont tous différenciés selon le sens de la liaison. Par la suite, on s'intéresse uniquement aux indicateurs de performance pour la liaison descendante. Cependant, les travaux réalisés peuvent être adaptés aisément pour des études sur la liaison montante.

Ressources d'une cellule Lorsque deux équipements sont en communication radio, le récepteur reçoit toujours le message de l'émetteur avec du bruit causé par le milieu de propagation (canal de transmission). Pour maximiser la quantité d'information pouvant être transmise tout en garantissant que le message initial puisse être retrouvé, le concept de blocs de ressource est introduit pour définir la quantité d'information pouvant être transmise, sur quels intervalles de temps et de fréquence.

- Pour la 4G, le bloc de ressource physique (PRB : *Physical Resource Block*) est la plus petite unité allouée à un utilisateur [29]. Un PRB est défini dans les domaines fréquentiel et temporel par une largeur de bande de 180 kHz et une durée de 0.5ms (appelé un *slot*). Le procédé de codage utilisé pour transporter l'information, appelé OFDM (*orthogonal frequency-division multiplexing*), permet de transporter l'information sous forme de symboles sur ces blocs. Le nombre de symboles est de 84 pour un PRB.
- Pour la 5G, étant donné une quantité de symboles fixée, l'intervalle de fréquence et de temps sur lesquels cette quantité est transportée varie en fonction d'un indice de numérologie μ [30, 31]. Cet indice a été introduit pour adapter le découpage des ressources en fonction de l'ordre de grandeur des fréquences, car celui-ci modifie les caractéristiques du canal de transmission. Le bloc de ressource est défini uniquement dans le domaine fréquentiel, sur une largeur de bande variable. Pour donner un exemple, si $\mu = 0$, alors on transporte 168 symboles sur un bloc de ressource de taille 180 kHz et un intervalle de 1 ms.

Le nombre de blocs de ressources (et donc la quantité d'information transmise) est limité par la largeur de bande. Pour la 4G par exemple, la plus petite largeur possible est de 1.4 MHz (6 blocs) et la plus grande de 20 MHz (100 blocs). Lorsque la proportion de blocs alloués par une station de base dépasse un certain seuil, le réseau tendra à congestionner en ce point.

Qualité de service La qualité de service caractérise la capacité d'un opérateur à assurer aux utilisateurs l'accès et l'usage du réseau mobile selon des critères chiffrés. L'Arcep, réalise régulièrement des campagnes de mesures de la qualité de service en se basant sur les critères ci-dessous [1] :

> — *le taux de communications de qualité vocale parfaite,*
> — *le taux de SMS reçus en moins de 10 secondes,*
> — *le débit moyen pour le téléchargement et l'envoi de fichiers,*
> — *le taux de pages web chargées en moins de 10 secondes,*
> — *le taux de vidéos d'une durée de 2 minutes visualisées avec une qualité parfaite.*

Comme les données collectées sur lesquelles on travaille sont collectées au niveau des stations de base, le seul indicateur exploitable qui s'approche d'un des critères cités précédemment est le débit moyen des utilisateurs.

Le débit décrit le volume de données transmis par unité de temps. Pour la 4G et la 5G, l'ordre de grandeur des débits s'expriment en mégabits ou en gigabits par seconde. Le débit maximal théorique dépend directement du nombre de PRBs disponibles, de la numérologie et de la modulation utilisée pour transporter les bits de données sur les symboles (QPSK, 16QAM, 64QAM, 256QAM). Un grand nombre de PRBs disponibles, un indice de numérologie élevé (pour la 5G uniquement) et une modulation performante (i.e. 256QAM) entraîneront un débit théorique maximal plus important. Les fréquences basses portent plus loin que les fréquences hautes, qui en revanche offrent généralement de meilleurs

1. https ://www.arcep.fr/nos-sujets/la-qualite-de-service-mobile.html

débits compte-tenu d'une plus grande largeur de spectre attribuée aux opérateurs. On réfère le lecteur à la spécification technique 3GPP TS38.306 [32] pour les éléments de calcul du débit théorique de la 5G.

Trafic On désigne par trafic le volume de données mobiles qui transite par une cellule ou une station de base au cours d'une période donnée (jour, semaine, mois, année...). Le trafic est généré par les utilisateurs lorsqu'ils utilisent des services nécessitant d'accéder à Internet. Cet indicateur peut être utilisé pour mesurer la rentabilité d'un site : plus le trafic d'une station de base est élevé, plus le retour sur investissement est élevé sur le territoire couvert.

2.2 Apprentissage automatique

2.2.1 Pourquoi utiliser l'apprentissage automatique ?

Le fonctionnement du réseau mobile produit un volume de données massif constitués d'indicateurs de performance très variés. L'utilisation d'algorithmes de l'apprentissage automatique (*Machine learning*) vient naturellement à l'esprit car ces modèles sont capables de traiter un jeu de données complexe et bruité pour en extraire des relations entre des variables que l'on espère pertinentes et plus complètes que les approches classiques. Dans les télécommunications, l'intelligence artificielle couvre un champ de recherche très large allant de l'étude des données mobiles et des déplacements des utilisateurs à des problématiques liées à la sécurité des réseaux, à la signalisation ou aux capteurs de réseaux sans fil [33]. Au niveau du réseau d'accès, l'étude de l'état de l'art s'est focalisée sur deux aspects relatifs à la gestion et à la planification du réseau : l'analyse de l'activité des réseaux mobiles, et la modélisation de la qualité du signal.

Analyse de l'activité des réseaux mobiles L'apprentissage automatique a ouvert la possibilité de connaître le trafic futur du réseau mobile à une granularité spatio-temporelle très fine : à l'échelle de la cellule, et pour un futur proche pouvant aller de quelques jours à quelques secondes. Les modèles entraînés peuvent être utilisés pour améliorer la gestion du réseau mobile. Par exemple, la prédiction de la charge permettrait de rendre l'allocation des ressources plus dynamiques, en allumant ou éteignant des sites en fonction de l'affluence anticipée [24, 34]. Par ailleurs, en analysant les données passées, certains algorithmes sont capables d'extraire des informations sur leur périodicité, leur tendance, les valeurs extrêmes, et ainsi détecter les anomalies pouvant être liées à des problèmes de sécurité [35, 36].

Prédiction de la qualité du signal La modélisation de la qualité du signal est une étape fondamentale de la planification des réseaux, puisque de celle-ci va dépendre le positionnement des nouveaux sites et la configuration des antennes. L'approche traditionnelle est d'utiliser des modèles d'affaiblissement de propagation du signal pour dimensionner la couverture du réseau. Les modèles sont dits « semi-empiriques » car il s'agit de formules prenant en compte des variables comme la distance de l'utilisateur, la hauteur de l'émetteur et du récepteur, mais aussi des constantes dont les valeurs ont été estimées par des mesures de terrain [37, 38]. Ces dernières années, de très nombreuses études ont montré la supériorité des modèles d'apprentissage pour estimer la propagation du signal [39-43]). Cela s'explique par le fait que les modèles guidés par la donnée modélisent mieux les phénomènes réels que les modèles semi-empiriques, dont les relations entre les paramètres sont parfois construites sur des hypothèses très simples. Depuis plus récemment, il existe des modèles de propagation beaucoup plus précis qui se basent sur les équations électro-magnétiques et la technique de ray tracing [44]. Cependant, ils sont également plus complexes et coûteux en temps de calcul, ces modèles devant être exécutés pour chaque site de déploiement. Là encore, l'apprentissage automatique est utile car il permet de réduire les temps d'exécution, en entraînant des modèles comme des fonctions d'approximation de ces modèles de propagation [45, 46].

Prédiction de l'activité du réseau pour la planification L'objectif poursuivi par la thèse est proche de ces deux aspects, à des différences près :

- comme dans le premier cas, on cherche à prédire des indicateurs d'activité des réseaux mobiles. Cependant, on s'écarte de la prédiction dynamique, car on cherche à faire cette prédiction pour des sites ou des cellules qui n'existent pas encore, des variations de performance des sites existants suite à de nouveaux déploiements.
- comme dans le deuxième cas, les modèles entraînés ont pour finalité d'assister la recherche d'emplacements optimaux pour le déploiement. Cependant, les sites choisis dépendront de variables

dont le volume et le profil temporel dépend de l'activité et de la mobilité des utilisateurs, et non de la configuration du terrain. Cette approche complémente les travaux de planification radio en estimant le profil de la demande, ce qui permet d'affiner le dimensionnement du réseau, et en identifiant les sites les plus intéressants pour les études de marketing.

Cette section commence par introduire les notions d'apprentissage automatique utilisées au cours de la thèse. Ensuite, on présentera les principes de l'apprentissage supervisé et non supervisé ainsi que les modèles associés à chacun. Enfin, on détaillera les mesures d'erreur couramment utilisées pour l'entraînement et l'évaluation des modèles.

2.2.2 Définitions et vocabulaire général

L'apprentissage automatique (en anglais *machine learning*) est un domaine à la croisée de l'informatique, de la statistique et des mathématiques appliquées qui rassemble un grand nombre d'algorithmes (ou modèles). Un modèle peut être simplement vu comme une boîte noire, prenant des données en entrée, et produisant d'autres données permettant d'accomplir une tâche (génération de données, débruitage, classification, prédiction de valeurs...). On parle d'apprentissage car avant d'être exploitables, les modèles sont composés de nombreux paramètres qu'il faut ajuster par un processus souvent itératif, réalisé sur un ensemble de données. On parle d'automatique, car cet ajustement est fait sans intervention humaine ; au cours de l'apprentissage, on fixe un critère qui dépend du type d'apprentissage, que l'ajustement des paramètres devra chercher à remplir. Une fois entraîné, on utilise les capacités de généralisation des modèles en leur passant des données qui ne faisaient pas partie de leur entraînement pour exploiter leurs sorties.

Variables Les variables sont les données qui sont passées en entrée au modèle pour réaliser les prédictions. Elles sont choisies de sorte à contenir des informations pouvant décrire, expliquer ou être corrélées à une tâche que l'on souhaite accomplir ou au phénomène que l'on veut prédire. Elles doivent être disponibles au moment de la phase de prédiction.

Observations On utilise le terme observation pour désigner les données rattachées à l'objet concerné par l'apprentissage. Ces données sont constituées des variables d'entrées, et lorsqu'il s'agit d'un apprentissage supervisé, des cibles à prédire. Par exemple si l'on cherche à prédire le trafic des stations de base, une observation sera un ensemble de caractéristiques de la station de base. Si l'on fait de la classification d'images, une observation sera une image.

Données d'entraînement, de validation, de test On définit souvent deux à trois jeux de données en apprentissage automatique :

- les données d'entraînement correspondent aux données sur lesquels les paramètres des modèles sont ajustés
- les données de validation correspondent à des données utilisées pendant la phase d'entraînement mais qui ne sont pas vues par les modèles. Elles correspondent souvent à une fraction des données d'entraînement initiales, et sont utilisées pour vérifier que le modèle ne fait pas de surapprentissage. Le surapprentissage correspond à un choix de paramètres qui rend les modèles très performants sur les données d'entraînement, mais beaucoup moins bons sur des nouvelles données (Figure 2.2).
- les données de test sont des données qui ne sont pas vues pendant l'entraînement du modèle. Elles peuvent correspondre à une dernière évaluation simulant l'arrivée de nouvelles données, ou correspondre directement aux nouvelles données.

Selon la tâche à accomplir, les apprentissages se distinguent en trois classes principales : l'apprentissage supervisé, l'apprentissage non-supervisé, et l'apprentissage par renforcement. Les problématiques rencontrées dans la thèse sont traités par des modèles supervisés et non-supervisés. Le détail de ces deux apprentissages et des modèles associés sont détaillés ci-après.

2.2.3 Apprentissage supervisé

L'apprentissage supervisé est particulièrement adapté aux tâches de classification (prédiction d'une catégorie) et de régression (prédiction d'une valeur continue). Le principe de cet entraînement est d'ajuster les paramètres des modèles de manière à minimiser l'écart entre des valeurs « cibles » et les sorties produites.

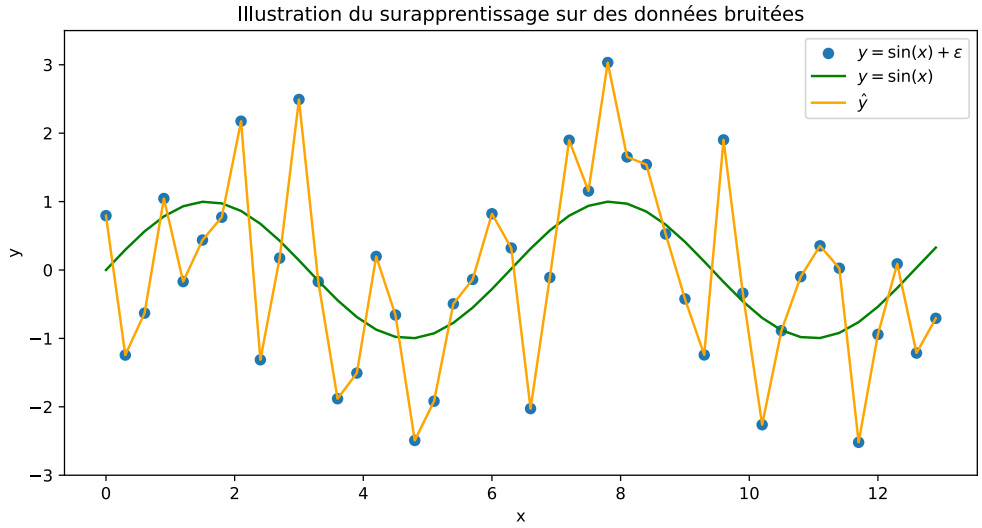


FIGURE 2.2 – Illustration du surapprentissage sur des données (points bleus) générées à partir de la fonction sinus (courbe verte) auquel on a rajouté un bruit gaussien. La courbe jaune est une illustration d'un modèle faisant du surapprentissage : les points bleus sont parfaitement interpolés, mais les points d'abscisse x ne faisant pas partie de l'entraînement seront estimés avec le bruit appris par le modèle.

Formalisation Pour des données tabulaires, on formalise les données et l'apprentissage de la manière suivante. Soit $\{(X_i, y_i)\}_{i=0}^{n-1}$ un jeu de données composé de $n \in \mathbb{N}$ observations, où $X_i \in \mathbb{R}^p$ est le vecteur constitué des $p > 0$ variables de la i-ème observation et $y_i \in \mathbb{R}^q$ $(q > 0)$ est le vecteur des q valeurs cibles associées. Pour une observation d'indice i, le modèle f prend en entrée les variables X_i et produit un vecteur $\hat{y}_i \in \mathbb{R}^n$ qui correspond à la prédiction du modèle :

$$f(X_i) = \hat{y}_i$$

Notations matricielles En pratique, les variables des observations et des cibles sont passées par blocs aux modèles. Pour certaines explications, on adoptera les notations suivantes : on note $X = (X_0, ..., X_{n-1}) \in \mathcal{M}_{n,p}$ la matrice des variables de toutes les observations, et $y = (y_0, ..., y_{n-1}) \in \mathcal{M}_{n,q}$ la matrice des cibles associées.

Classification Pour la classification, les données du jeu d'entraînement sont taggées à l'avance (parfois manuellement) dans des classes. Ces tags constituent les données cibles $y_i \in \mathbb{N}$ que l'on cherchera à prédire pour les nouvelles données. Le vecteur y_i se réduit donc à une seule valeur représentant le tag $(n = 1)$. Dans les réseaux mobiles, les travaux de classification portent principalement sur la classification de trafic chiffré pour reconstituer les usages du réseau, ou pour la détection d'anomalies et de communications malveillantes [33].

Régression Pour la régression simple, le vecteur $y_i \in \mathbb{R}$ est de taille 1 comme pour la classification, mais les valeurs peuvent être continues $(n = 1)$. Dans la littérature, les travaux de régression sur les données mobiles visent principalement à prédire le trafic. Par exemple, il est possible de prédire divers indicateurs de charge du réseau dans un futur proche à partir du trafic dans le passé proche [24], ou bien de prédire leur charge en croisant d'autres indicateurs de performance [34]. Les travaux mentionnés précédemment sur la prédiction de la propagation du signal correspondent également des problèmes de régression [39-43]. Pour certaines contributions de la thèse (Chapitres 5 et 7), on réalise des régressions multi-cibles. Dans ce cas, on cherche à prédire plusieurs valeurs et le vecteur y_i est de taille supérieure à 1. Le traitement de la régression multi-cibles sera abordé dans le Chapitre 5.

Entraînement itératif La plupart des algorithmes populaires (comme les réseaux de neurones ou les arbres de décision) sont entraînés suivant un processus itératif. Au cours d'une itération, on calcule l'erreur de prédiction entre $\hat{y}_i$ et y_i. Les différentes fonctions d'erreur utilisables sont présentées plus loin dans la section. Cette erreur est utilisée pour ajuster les paramètres de manière à ce qu'elle soit plus faible à l'étape suivante. L'entraînement s'arrête lorsque l'erreur ne peut plus être minimisée. La Figure 2.3

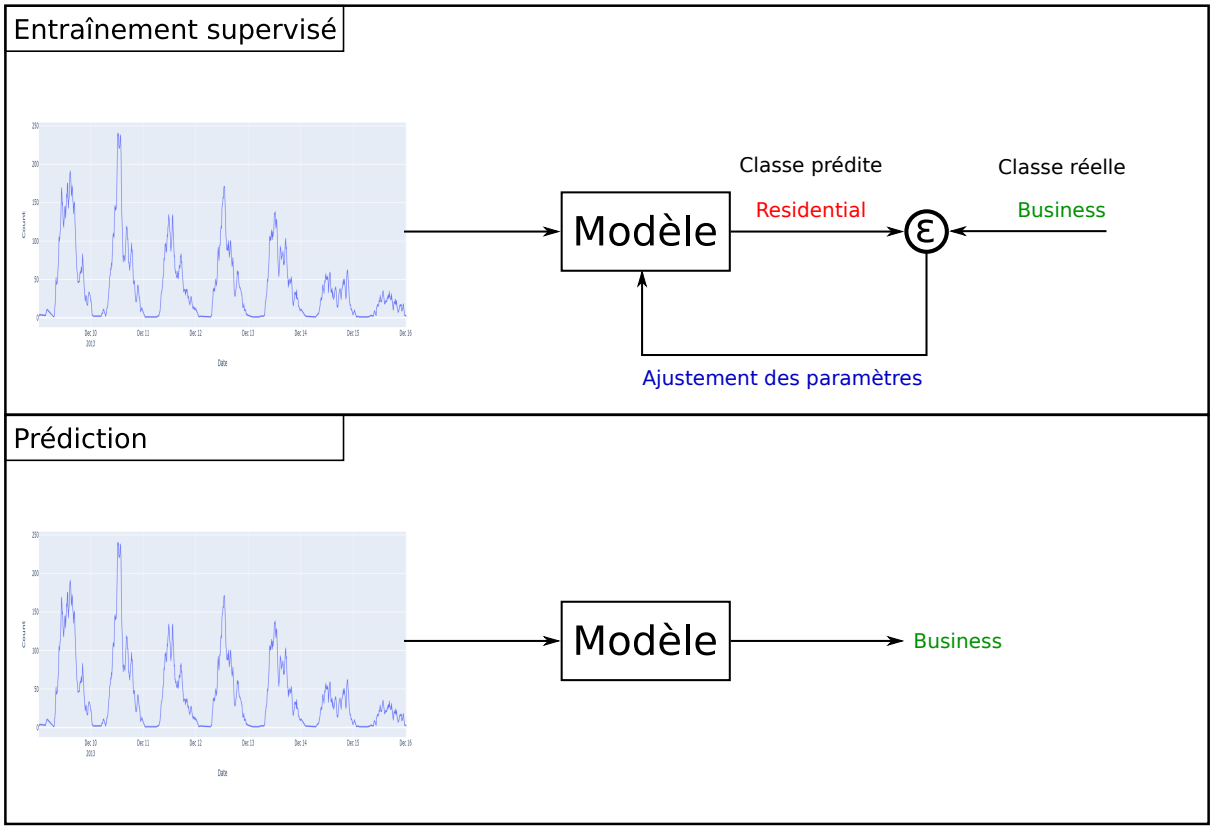


FIGURE 2.3 – Processus d'apprentissage itératif appliqué à un modèle pour classifier le profil du trafic des cellules. Le schéma du haut présente l'entraînement itératif sur une observation annotée à l'avance. Le schéma du bas montre la phase de prédiction sur les données une fois que le modèle est bien entraîné. Pour l'illustration, la donnée est identique à celle de l'entraînement, mais en pratique, on classifie des nouvelles données non annotées qui n'ont pas été vues par le modèle.

schématise la phase d'entraînement supervisé pour la classification du trafic hebdomadaire d'une cellule, puis la phase de prédiction lorsque les paramètres sont ajustés.

Difficultés couramment rencontrées durant l'apprentissage La recherche du modèle minimisant l'erreur d'entraînement rencontre souvent deux difficultés : le surapprentissage et le non déterminisme de certains entraînements.

Le phénomène de surapprentissage est observable lorsque les performances de prédiction du modèle sont très dégradées sur des nouvelles données alors qu'elles sont faibles sur les données d'entraînement. Cela arrive lorsque le modèle s'est trop spécialisé sur les données d'entraînement et perd en faculté de généralisation (Figure 2.2). En fonction du modèle utilisé, le surapprentissage peut être évité de plusieurs manières : utilisation de données de validation pour arrêter les itérations au bon moment, apprentissage ensembliste, réduction du nombre de variables...

Le non déterminisme des entraînements implique qu'en fonction du modèle et du type d'entraînement, deux modèles avec des hyperparamètres identiques et suivant le même processus d'apprentissage peuvent avoir des performances différentes. En effet, l'entraînement peut introduire des mécanismes aléatoires, comme par exemple la sélection au hasard d'un sous-ensemble d'observations à chaque itération. Cette technique permet de résoudre diverses difficultés, comme la recherche de paramètres satisfaisants dans un temps raisonnable pour des modèles complexes (ex : descente de gradient stochastique pour les réseaux de neurones), ou bien pour limiter le surapprentissage. Il est alors préférable d'évaluer un même modèle sur plusieurs entraînements et sur des fractions de données différentes afin d'obtenir des statistiques sur les erreurs de prédiction. La validation croisée permet d'effectuer ce genre d'évaluation.

Validation croisée à k blocs Le principe est de partitionner les données disponibles en k sous-ensembles, et de réaliser k entraînements pour un modèle donné. Pour chaque entraînement, les données

d'apprentissage sont constituées de $k-1$ blocs, le dernier étant réservé pour l'évaluation. La validation croisée permet de pouvoir calculer une erreur moyenne et un écart-type sur les k blocs de test et d'avoir des résultats statistiques représentatifs de la variabilité de l'entraînement du modèle. Une variante de cette méthode consiste à répéter plusieurs fois la validation croisée. Cela permet de conserver la même taille de blocs, mais d'avoir un plus grand nombre d'évaluations, grâce à un partitionnement différent à chaque répétition.

2.2.4 Modèles utilisant l'apprentissage supervisé

Dans cette section, on présentera brièvement le principe de fonctionnement des modèles utilisés dans la thèse : les réseaux de neurones, les arbres de décisions et les modèles ensemblistes dérivés qui ont fait l'objet d'études approfondies, et d'autres modèles couramment répandus dans la littérature.

Les réseaux de neurones artificiels

Les réseaux de neurones artificiels (*Artificial Neural Network*) sont des architectures composées d'un empilement de couches elles-mêmes constituées de neurones. On modélise des connexions entre les neurones qui peuvent échanger des informations entre elles. L'architecture la plus répandue est celle du réseau de neurones à propagation en avant (*Feedforward Neural Network*), où les connexions sont établies entre les neurones de couches différentes, dans un sens unidirectionnel. Dans certains modèles, les connexions également être bidirectionnelles et/ou posséder des connexions à l'intérieur des couches. C'est le cas des réseaux de neurones récurrents (RNN : *Recurrent Neural Network*) [47].

Les paramètres du modèle sont les poids associées aux connexions, ces poids permettant de filtrer ou d'amplifier l'information transmise d'un neurone à l'autre. Au moment de l'entraînement, les erreurs propagées à travers le réseau permettent donc d'ajuster la valeur de ces poids. Dans le Chapitre 4, on présente une formalisation des paramètres des réseaux de neurones et de l'entraînement.

Déclinaisons L'architecture la plus basique des réseaux de neurones est celui à propagation en avant. Il permet de résoudre des problèmes simples de régression et de classification. Toutefois, les réseaux de neurones se sont grandement diversifiés et complexifiés selon les domaines de recherche, pour s'adapter au mieux à la nature des variables d'entrées. Sans être exhaustif, on peut toutefois citer les modèles suivants et leurs usages traditionnels :

- les réseaux de neurones convolutifs (CNN : *Convolutional Neural Network*) [48] pour le traitement de l'image. Certaines données du réseau mobile peuvent être représentées sous le même format que des images. Les CNNs permettent alors de réaliser de la classification de trafic [49] grâce à l'extraction des relations spatiales, ou de la prédiction de trafic en combinant le modèle avec un réseau récurrent [24],
- les LSTMs (*Long Short-Term Memory*) [50] et réseaux de neurones récurrents pour la prédiction de données temporelles, à partir de données temporelles passées. Ce type de modèle peut être utilisé pour prédire le trafic du réseau mobile dans un futur proche [24, 51],
- les auto-encodeurs [52] pour la compression, le débruitage et la génération de données. Une fois le modèle entraîné, il peut être utilisé pour réduire la dimension des données mobiles afin de faciliter le regroupement et l'analyse des données mobiles [53].
- les transformeurs [54] pour le traitement du langage naturel.

Il est à noter que ces dernières années, de plus en plus de ponts entre des domaines de recherche différents sont établis et les modèles ne se cantonnent plus toujours à leurs usages traditionnels. Par exemple, la première version du générateur d'images d'OpenAI (DALL·E) utilisait un transformeur pour traduire des représentations de texte en représentations compressées d'images. Le décodeur d'un auto-encodeur était ensuite utilisé pour décompresser et reconstituer ces images. Pour pouvoir étendre les travaux du Chapitre 3 et prédire l'activité hebdomadaire horaire des stations de base, on explore la possibilité d'entraîner une architecture de transformeur similaire sur les données tabulaires. Pour cela, on étudie dans le Chapitre 4 les techniques de réduction de dimension et de reconstitution des données pour réduire les temps d'entraînement, et la discrétisation de l'espace latent pour être compatible avec l'architecture du transformeur considéré. L'auto-encodeur variationnel discret mis au point dans cette étude est par la suite utilisé dans le Chapitre 5.

Les arbres de décision et l'apprentissage ensembliste

L'arbre de décision est un modèle qui partitionne l'espace en sous-régions appelées des « feuilles ». À chaque feuille est associée une valeur qui correspond à la sortie du modèle pour toute entrée qui est associée dans cette région. Pour la classification, il s'agira du label de la classe et pour la régression, d'une valeur continue. Certaines implémentations supportent les problèmes multi-cibles où chaque partition est associée à un vecteur de valeurs [55].

Traitement des entrées Les variables d'entrées sont associées à une feuille de la manière suivante. Soit i l'indice d'une observation, X_i les variables associées ainsi qu'un arbre de décision entraîné pour une tâche quelconque. Les variables sont traitées étapes par étapes au niveau des « nœuds ». Un nœud est un embranchement au niveau duquel la valeur prise par une variable est comparée à une valeur seuil. Le résultat de la comparaison amène les variables à un autre nœud et à une autre comparaison sur une autre variable, et ainsi de suite jusqu'aux feuilles. Le nombre de comparaisons maximales possibles est appelé la profondeur de l'arbre. La Figure 2.4 schématise le traitement d'une observation par un arbre à 8 feuilles et 6 nœuds qui classifie le trafic d'une station de base en fonction du tissu urbain qu'elle couvre. Les variables d'entrées sont la religion, l'industrialisation et la population d'un territoire.

Lors de l'entraînement, l'arbre construit des nœuds tant que les partitions formées sont « impures », c'est-à-dire qu'elles contiennent des cibles de nature hétérogènes, mélangeant plusieurs classes, ou présentant une forte variance. Pour la classification, on utilise couramment l'indice de Gini pour mesurer l'impureté (*impurity*) du nœud. En régression, on peut par exemple utiliser l'erreur quadratique moyenne ou l'erreur absolue moyenne.

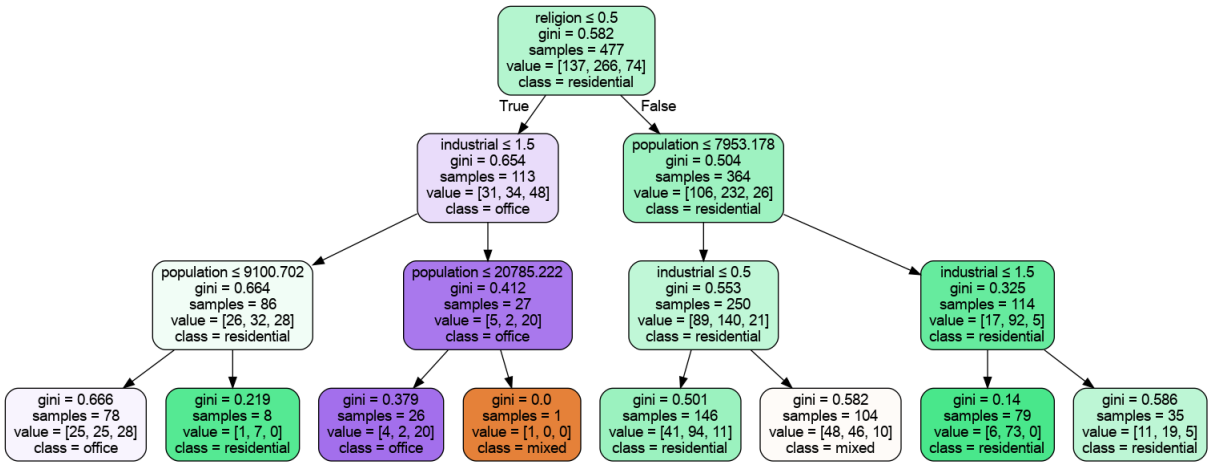


FIGURE 2.4 – Représentation d'un arbre de décision à 8 feuilles et de profondeur 3. La tâche de classification comporte trois labels : *office*, *residential* et *mixed*

Avantages et limites Les arbres de décisions ont l'avantage d'être facilement interprétables. Il existe ainsi plusieurs algorithmes permettant de savoir sur quelles variables s'appuie un arbre pour être précis. Ces modèles sont aussi appréciés pour leur faculté à traiter des variables hétérogènes (entiers, valeurs continues, catégories) avec assez peu de pré-traitement. Un arbre de décision seul peut en revanche souffrir de surapprentissage dès qu'il possède trop de feuilles ou de profondeur, ou de sous-apprentissage dans le cas contraire. On préférera les associer à des techniques d'apprentissage ensembliste pour compenser ces limitations.

Apprentissage ensembliste L'apprentissage ensembliste consiste à utiliser un groupe de modèles pour calculer une prédiction finale en agrégeant les prédictions de chaque modèle. Les estimateurs de groupe sont qualifiés de faibles (*weak learners*), car seuls, leur précision est plutôt faible, voire à peine meilleure que le hasard [56]. Il existe deux structures ensemblistes répandues utilisant les arbres de décision : les forêts aléatoires et les *Gradient Boosted Decision Trees* (GBDT) :

— Dans une **forêt aléatoire** de n arbres, ces derniers sont entraînés en parallèle, chacun sur un sous-ensemble des données d'entraînement [57]. Ce sous-ensemble peut être distinct ou non [58].

La prédiction résultant d'une forêt aléatoire est typiquement la moyenne des prédictions de tous les arbres de régression, ou la moyenne des probabilité de classe des arbres de classification [59]. Les arbres de décision sont des modèles classiques de la prédiction du trafic des cellules du réseau mobile [60].

— Dans un **GBDT**, les arbres de décision sont entraînés suivant la méthode du *gradient boosting* (amplification du gradient). Le principe est d'entraîner itérativement des estimateurs les uns à la suite des autres, chaque estimateur étant entraîné sur l'erreur résiduelle de l'estimateur à l'itération précédente. L'estimateur le plus utilisé pour ses performances consistantes est l'arbre de décision. Parmi les modèles les plus populaires, on peut citer CatBoost [61], LightGBM [62] et XGBoost [63]. Ces modèles sont en général plus performants que les forêts aléatoires, mais comme les réseaux de neurones, ils ont une tendance au surapprentissage. Pour cette raison, les librairies implémentant des GBDTs proposent d'utiliser des données de validation pendant l'entraînement pour trouver l'itération d'arrêt. On choisira d'arrêter l'entraînement lorsque l'erreur de prédiction sur les données de validation ne s'améliore plus, ou juste avant qu'elle n'augmente. Dans la littérature, de nombreux travaux de recherche employant des GBDTs sont portés sur la prédiction de la qualité du signal des réseaux mobiles en intérieur [64] et en extérieur [65] à partir de variables sur la configuration des antennes et de la distance des utilisateurs.

Régression linéaire multiple et régularisation

La régression linéaire multiple est une méthode de régression réalisée sur des observations $\{(X_i, y_i)\}_{i=0}^{n-1}$ et paramétrée par un ensemble de coefficients associés à chaque variable. Pour rappel, les vecteurs $X_i \in \mathbb{R}^p$ sont les p variables associées à une observation i, et le vecteur $y_i \in \mathbb{R}$ correspond à la valeur cible. On note $X = (X_0, ..., X_{n-1}) \in \mathcal{M}_{n,p}$ la matrice des variables de toutes les observations, et $y = (y_0, ..., y_{n-1}) \in \mathbb{R}^n$ le vecteur contenant toutes les cibles. La relation entre X et y modélisée par la régression linéaire multiple est :

$$y = X\beta + \epsilon$$

où $\beta \in \mathbb{R}^p$ est le vecteur des coefficients à déterminer et $\epsilon \in \mathbb{R}^p$ est l'erreur de régression.

La méthode des moindres carrés est une méthode pour déterminer les coefficients β. L'objectif est de trouver le vecteur β^* qui minimise l'erreur quadratique, c'est-à-dire :

$$\beta^* = \underset{\beta}{\operatorname{argmin}} ||y - X\beta||_2^2$$

où $||.||_2$ désigne la norme euclidienne.

Le problème fréquemment rencontré par la régression linéaire multiple est la colinéarité entre les variables. La colinéarité dégrade la fiabilité de l'interprétation des coefficients β, à la fois en valeur et en importance. L'utilisation de termes de régularisation permet de limiter ce problème. Dans certaines contributions de la thèse, on utilise les modèles *ridge* et *lasso*.

Ridge regression En apprentissage automatique, la régression *ridge* est une méthode de régression qui introduit un terme de régularisation L2 ainsi qu'un paramètre $\lambda \geq 0$ contrôlant son importance [66].

Le coefficient β_r recherché minimise la relation :

$$\beta_r^* = \underset{\beta_r}{\operatorname{argmin}} \frac{1}{n} ||y - X\beta_r||_2^2 + \lambda ||\beta_r||_2^2$$

Régression lasso Le lasso (*Least Absolute Shrinkage and Selection Operator*), est une autre variante qui régularise la méthode des moindres carrés en ajoutant un terme de régularisation L1. Dans ce cas, on cherche le coefficient β_l^* qui minimise :

$$\beta_l^* = \underset{\beta_l}{\operatorname{argmin}} \frac{1}{n} ||y - X\beta_l||_2^2 + \lambda ||\beta_l||_1^2$$

où $||.||_1$ désigne la norme 1 (somme des valeurs absolues des coefficients d'un vecteur).

Autres modèles supervisés

La méthode des k plus proches voisins (KNN : *K-nearest neighbors* [67, 68]) est utilisée aussi bien pour la régression que pour la classification. Soit i l'indice d'une observation dont on souhaite obtenir une prédiction. Celle-ci est calculée en fonction de k valeurs issues du jeu d'entraînement : il s'agit des k cibles dont les variables sont les plus proches des variables X_i passées en entrée (d'où le terme de voisins). Dans la forme la plus simple, en classification, on attribue à l'observation i la classe la plus représentée chez ses voisins les plus proches. En régression, il est courant d'attribuer la moyenne des valeurs cibles des k voisins. Une autre variante consiste à attribuer un poids à la valeur ou à la classe des voisins en fonction de leur distance à X_i, puis à effectuer une moyenne pondérée.

La machine à vecteurs de support (SVM : *Support Vector Machine*) est un algorithme utilisé pour la régression ou la classification :

— en classification (*Support Vector Classifier*), le principe est de trouver l'hyperplan dans un espace telle que la distance des classes des observations à cet hyperplan soit maximisée. La technique définit une marge autour de l'hyperplan dans lequel on minimise le nombre d'observations s'y trouvant. Cette condition permet une bonne capacité de généralisation du modèle.

— en régression (*Support Vector Regressor*) [69, 70], on cherchera l'hyperplan tel qu'il est « le plus proche possible » de tous les points, car c'est ce sont les points de cet hyperplan qui constitueront les prédictions. La marge de l'hyperplan est utilisée pour ignorer les points qui sont à l'intérieur (car le modèle estime bien ces points) pour se concentrer sur les points les plus éloignés de l'hyperplan. Ce modèle est utilisé dans des études comparatives sur les algorithmes pour la prédiction du trafic des cellules [34, 60].

2.2.5 Apprentissage non supervisé

L'apprentissage non-supervisé est utilisé pour des tâches de regroupement d'observations en fonction des similarités qu'il peut exister entre les variables. Contrairement à l'entraînement supervisé, il n'y a pas de valeur(s) cibles car les données ne sont pas annotées. Néanmoins, une connaissance des données traitées peut permettre d'attribuer une explication et une classe à chaque groupe obtenu a posteriori.

Les techniques non-supervisées débouchent sur diverses applications :

— Une fois entraînés, les modèles non-supervisés peuvent classer de nouvelles observations en les positionnant par rapport aux observations apprises.

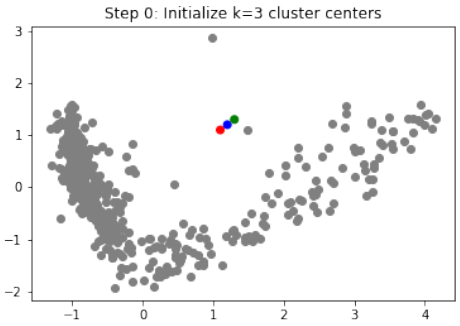
— Elles permettent d'annoter automatiquement les données, qui peuvent être employées par la suite pour de l'apprentissage supervisé. La différence avec l'application précédente est que les variables utilisées pour l'apprentissage supervisé peuvent être différentes des variables utilisées pour le regroupement. Cette méthode est utilisée dans le Chapitre 3 pour la prédiction de la classe d'activité du trafic d'une station de base.

— Combinées avec des techniques de réduction de dimension, elles peuvent servir pour de l'analyse exploratoire, pour améliorer la compréhension de la structure des données, la distribution des observations, etc.

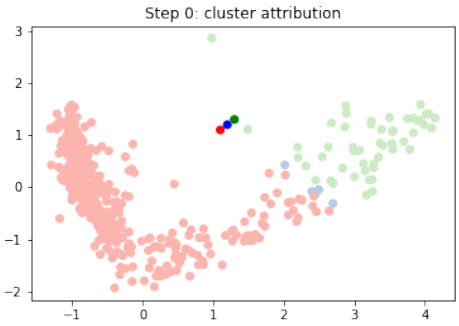
Il existe de nombreux algorithmes de regroupement, les plus classiques étant disponibles dans les librairies d'apprentissage automatique [59, 71]. Dans la thèse, on utilisera l'algorithme des k-moyennes ou le regroupement hiérarchique pour les tâches de regroupement.

L'algorithme des k-moyennes permet de partitionner les observations en k groupes, le paramètre k étant choisi à l'avance. Il s'agit d'un algorithme itératif :

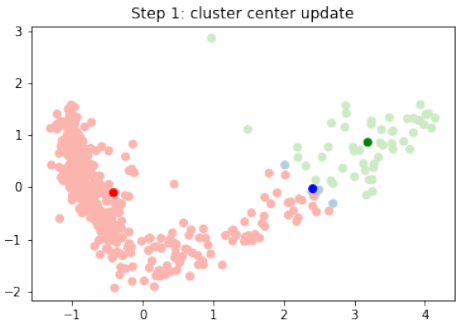
1. On initialise k vecteurs de la même taille que le nombre de variables, qu'on appelle des centroïdes (Figure 2.5(a)).
2. À une itération donnée, pour chaque observation i, on calcule la distance de X_i à tous les centroïdes, et on lui associe le centroïde dont il est le plus proche (Figure 2.5(b)). On obtient donc k groupes, composés pour chacune des observations les plus proches d'un centroïde.
3. Les centroïdes sont « mis à jour » en leur attribuant comme nouvelle valeur le barycentre de leur groupe (Figure 2.5(c)).
4. L'attribution et la mise à jour des centroïdes sont réitérées (Figure 2.5(d)) jusqu'à ce que l'algorithme converge. Les groupes obtenus minimisent la variance intra-classe et maximisent la variance inter-classes.



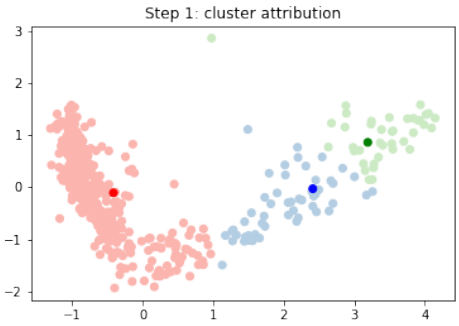
(a) Étape 0 : Initialisation des centres de classe



(b) Étape 0 : Attribution des classes à chaque observation



(c) Étape 1 : Mise à jour des centres de classe



(d) Étape 1 : Attribution des classes à chaque observation

FIGURE 2.5 – Illustration des premières étapes de l'algorithme des k-moyennes ($k = 3$) sur le trafic de stations de base (réduites à deux dimensions suite à l'application de l'analyse en composantes principales).

La méthode des k-moyennes est une méthode populaire pour grouper le trafic des cellules par profil d'activité. L'analyse de ces classes permet d'étudier l'utilisation des sols et les activités humaines dans les régions urbaines à une échelle proche de celle du quartier [72-74].

Les algorithmes de regroupement hiérarchique partitionnent les données de suivant deux méthodes possibles :

- la méthode descendante : en partant d'un unique groupe contenant toutes les observations, le jeu de données est divisé itérativement en partitions de plus en plus petites.
- la méthode ascendante : au début de l'exécution, chaque observation appartient à un groupe dont il est le seul membre. Les observations sont itérativement rassemblées dans des groupes de plus en plus grands jusqu'à englober toutes les observations. À chaque étape, le choix des groupes à fusionner est déterminé en calculant une mesure de dissimilarité entre tous les groupes. Les deux groupes possédant la plus petite dissimilarité sont fusionnés.

La méthode du regroupement hiérarchique ascendant est détaillée et utilisée dans la Section 3.3.4.

2.2.6 Mesures des erreurs de prédiction

Le paramétrage des modèles durant l'entraînement nécessite d'utiliser des mesures pour quantifier les erreurs de prédiction. Ces mesures sont également utiles pour évaluer la généralisation des modèles sur des données tests, et pour comparer les modèles entre eux. Cette section fait l'inventaire des mesures utilisées dans les contributions de la thèse. On distingue les mesures utilisées pour la classification de celles pour la régression.

Classification

Soit un jeu de données constitué de N observations à ranger dans une classe parmi $K > 1$ possibles. Pour une observation quelconque i, on note $y_i \in \{1, ..., K\}$ sa classe (cible) et $\hat{y}_i \in \{1, ..., K\}$ la classe prédite par un modèle.

La précision globale est le ratio entre le nombre de bonnes classifications et le nombre total d'observations :

$$\text{précision globale} = \frac{\sum_{i=1}^{N} \mathbb{1}_{\{\hat{y}_i = y_i\}}}{N}$$

Bien que cette mesure est facilement interprétable, elle n'est pas toujours satisfaisante en raison du déséquilibre des classes qu'il peut exister dans un jeu de données.

L'étude de la précision et du rappel pour chaque classe $k \in \{1, ..., K\}$ permet de mieux se représenter la précision des modèles dans chaque cas. Pour calculer ces deux mesures, on commence par distinguer quatre sous-ensembles :

- l'ensemble tp des vrais positifs, correspond aux observations correctement classées dans k.
- l'ensemble fp des faux positifs, correspond aux observations incorrectement classées dans k.
- l'ensemble tn des vrais négatifs, correspond aux observations correctement classées comme n'appartenant pas à k.
- l'ensemble fn des faux négatifs, correspond aux observations incorrectement classées comme n'appartenant pas à k.

Précision Pour une classe k donnée, on définit la précision comme le ratio entre le nombre de classification correctement réalisées et le nombre d'observations classées par le modèle dans k :

$$\text{précision} = \frac{tp}{tp + fp}$$

Rappel Pour une classe k donnée, on définit le rappel comme le ratio entre le nombre de classifications correctement réalisées et le nombre d'observations appartenant réellement à la classe k :

$$\text{rappel} = \frac{tp}{tp + fn}$$

F-score Pour une classe k donnée, on définit le F-score comme la moyenne harmonique entre la précision et le rappel :

$$2 \cdot \frac{\text{précision} \cdot \text{rappel}}{\text{précision} + \text{rappel}}$$

Le F-score se rapproche de 1 lorsque la précision et le rappel sont proches de 1. Il suffit donc qu'une seule de ses composantes soit faible pour que la mesure soit pénalisante.

Régression

Soit un jeu de données constitué de N observations, et un problème de régression consistant à prédire $q > 0$ valeurs continues à partir de celles-ci. Pour une observation i, on note $y_i \in \mathbb{R}^q$ les valeurs cibles associées et $\hat{y}_i \in \mathbb{R}^q$ les valeurs prédites par un modèle.

L'erreur quadratique moyenne (MSE : *Mean Squared Error*) est l'une des mesures les plus utilisées dans l'entraînement des modèles de régression. Il s'agit de la moyenne de la somme des différences au carré entre chaque prédiction et la valeur réelle correspondante :

$$\text{MSE}(y, \hat{y}) = \frac{1}{N} \sum_{i=0}^{N-1} (y_i - \hat{y}_i)^2$$

Plus la prédiction est proche de la réalité, plus l'erreur sera faible.

La racine carrée de l'erreur moyenne (RMSE : *Root Mean Squared Error*) est simplement dérivée de la MSE :

$$\text{RMSE}(y, \hat{y}) = \sqrt{\text{MSE}(y, \hat{y})}$$

Cette mesure a l'avantage d'être exprimée dans la même unité que les données cibles.

L'erreur absolue moyenne (MAE : *Mean Absolute Error*) est la moyenne de la somme des différences absolues entre chaque prédiction et la valeur réelle correspondante.

$$\text{MAE}(y, \hat{y}) = \frac{1}{N} \sum_{i=0}^{N-1} |y_i - \hat{y}_i|$$

L'unité de cette mesure est donc la même que celle de la valeur que l'on cherche à prédire. Le choix de la RMSE ou de la MAE dépend de la distribution des erreurs, qui dépend elle-même de la distribution des variables [75].

L'erreur absolue moyenne standard (nMAE : *Normalized Mean Absolute Error*) est la normalisation de la MAE par la moyenne des valeurs absolues des cibles. Contrairement aux précédentes mesures dont les valeurs sont sensibles à l'échelle des données, la nMAE calcule un écart relatif qui permet de comparer des erreurs de prédictions d'échelles différentes.

$$\text{nMAE}(y, \hat{y}) = \frac{\sum_{i=0}^{N-1} |y_i - \hat{y}_i|}{\sum_{i=0}^{N-1} |y_i|}$$

L'erreur absolue médiane (MedAE : *Median Absolute Error*) est une mesure qui prend la valeur médiane des différences absolues entre chaque prédiction et la valeur réelle correspondante :

$$\text{MedAE}(y, \hat{y}) = \text{médiane}(\mid y_1 - \hat{y}_1 \mid, \ldots, \mid y_n - \hat{y}_n \mid)$$

L'erreur relative absolue moyenne (MAPE : *Mean Absolute Percentage Error*) exprime l'erreur de prédiction en proportion par rapport à la valeur cible.

$$\text{MAPE}(y, \hat{y}) = \frac{1}{N} \sum_{i=0}^{N-1} \frac{|y_i - \hat{y}_i|}{max(\epsilon, |y_i|)}$$

Pour éviter que le dénominateur d'un élément de la somme soit nul, l'implémentation de `scikit-learn` [59] prend la plus grande valeur entre la cible et une très petite valeur ϵ. Cette mesure peut s'exprimer en pourcentage ; par exemple une MAPE de 0.13 correspond à une erreur de 13%.

L'erreur relative absolue symétrique moyenne (SMAPE : *Symmetric Mean Absolute Percentage Error*) est formulée de la manière suivante :

$$\text{SMAPE}_f = \frac{1}{n} \sum_{s=0}^{N-1} \frac{|\hat{y}_s^f - y_s^f|}{|y_s^f| + |\hat{y}_s^f|}$$

La SMAPE est une erreur relative bornée entre 0 et 1 (ou 0% et 100%).

Le score de variance expliquée (EVS : *Explained Variance Score*) mesure à quel point un modèle explique la variance des observations :

$$\text{EVS}(y, \hat{y}) = 1 - \frac{Var(y - \hat{y})}{Var(y)}$$

Un EVS de 1 équivaut à une variance des prédictions égale à celle des cibles. Il est insensible au décalage qu'il peut exister dans les prédictions. Par exemple, pour une observation i donnée, sa cible $y_i = (10, 11, 12)$ et sa prédiction $\hat{y}_i = (20, 21, 22)$ associées, on obtient $\text{Var}(y_i - \hat{y}_i) = 0$ et donc $\text{EVS} = 0$. Une comparaison entre un EVS faible et un score R2 élevé peut permettre d'identifier s'il existe un tel biais.

Le coefficient de détermination linéaire de Pearson (R^2) est exprimé de la manière suivante :

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{N-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{N-1}(y_i - \bar{y})^2}$$

où $\bar{y} = \frac{1}{n}\sum_{i=0}^{N-1} y_i$.

Le numérateur correspond à la variance entre la cible et la prédiction et le dénominateur à la variance entre la cible et la moyenne de toutes les cibles. Lorsque les prédictions du modèle sont très proches des valeurs réelles, $R^2 \to 1$. Lorsque $R^2 = 0$, le modèle n'est pas meilleur qu'une ligne de base qui renverrait la moyenne des cibles. Lorsque $R^2 < 0$, il est pire que la ligne de base.

L'erreur maximale (MaxE) est égale à la plus grande différence absolue des prédictions :

$$\text{MaxE}(y, \hat{y}) = max(|y_i - \hat{y}_i|)$$

La déformation temporelle dynamique (DTW : *Dynamic Time Warping*) consiste à mesurer la similarité de deux séries après une opération de ré-alignement. Utilisée au départ pour des travaux de reconnaissance du langage [76], ses propriétés sont aussi intéressantes pour la comparaison de séquences. Pour comprendre intuitivement son utilité, on peut prendre l'exemple des fonctions sinus et cosinus définies sur le domaine $[0, 2\pi]$. La valeurs de ces deux fonctions sont décalées d'une phase de $\frac{\pi}{2}$. Là où une mesure comme la MSE présenterait une valeur élevée, la valeur de la DTW serait égale à 0 après le ré-alignement des valeurs des deux fonctions.

L'implémentation utilisée dans la thèse pour calculer la DTW est celle du module Python `tslearn` [77].

2.2.7 Conclusion

Cette section a présenté diverses techniques d'apprentissage utilisées couramment pour la classification, la prédiction de performances et le groupement de données issues du réseau mobile.

Cependant, l'utilisation seule des données de performance du réseau ne suffit pas à entraîner les modèles automatiques pour accomplir les objectifs envisagés dans la thèse. En particulier, comment estimer le trafic futur d'équipements dans des zones où l'on ne dispose d'aucune donnée historique ? La solution proposée est de combiner les données réseaux avec des sources externes permettant de modéliser la demande des utilisateurs, afin de rendre l'estimation du trafic possible et d'améliorer la précision des prédictions dans les scénarios de densification.

Les données externes considérées proviennent des sources utilisées par les équipes de géo-marketing pour identifier les zones de déploiement rentables. Elles sont de nature géographique, démographique et socio-économique, et décrivent l'environnement urbain. Bien que l'on sait qu'elles sont corrélées d'une certaine façon à l'activité du réseau mobile, l'identification précise des variables importantes est difficile à réaliser « à la main ». Les éléments qui composent ces jeux de données sont très divers, hétérogènes et de distributions différentes. En utilisant l'apprentissage automatique, on espère pouvoir extraire les variables pertinentes et identifier des relations, comme des groupes de sites de trafics similaires corrélés à des occupations de sols spécifiques. L'interprétation des modèles permettrait, en partie, de valider ou d'infirmer nos intuitions initiales, et de possiblement apprendre de nouvelles relations entre l'activité des réseaux et le tissu urbain.

2.3 Les données géo-spatiales

2.3.1 Introduction

Les données exogènes (ou informations externes) d'intérêt sont des données provenant de sources qui ne concernent pas les télécommunications, mais qui peuvent expliquer des phénomènes liés à l'activité humaine. On peut citer par exemple la distribution des points d'intérêts et de l'utilisation des sols, qui permettent de prédire la demande en trafic d'un territoire [78, 79] ainsi que la force du signal reçu [80].

Dans cette section, on introduit les notions techniques liées aux systèmes d'information géographique (SIG) qui permettent de traiter les données externes et de les mettre en cohérence avec les données mobiles. On présentera ensuite les sources de données fournissant les informations liées à la cartographie, la distribution de population et la démographie.

2.3.2 Représentation des données dans l'espace : les système de coordonnées

Les données géo-spatiales sont des données sur des objets, des infrastructures ou des valeurs auxquelles on associe une géométrie et une localisation dans l'espace.

Le système de coordonnées géographique est un système de coordonnées (comme le système de coordonnées sphériques ou cartésiennes) utilisé pour localiser un objet :

- À la surface de la Terre, grâce à des coordonnées horizontales. Ces coordonnées sont des mesures angulaires de latitude et de longitude.
- À une certaine distance de la surface, auquel cas la mesure de l'altitude est rajoutée aux deux premières coordonnées.

Pour pouvoir utiliser un système de coordonnées géographiques, il faut se munir d'un système géodésique. Celui-ci définit entre autres le modèle utilisé pour représenter la Terre, ainsi que le point d'origine. Dans la thèse, les coordonnées géographiques sont exprimées à la surface de la planète dans le système *World Geodetic System 1984* (WGS 84) [81]. La longitude 0 se situe au méridien de référence de l'IERS (*International Reference Meridian*) et la latitude 0 à l'équateur.

Le système de coordonnées projetées est un système de coordonnées cartésiennes défini sur une surface plane. Pour représenter la Terre sur une surface plane, on utilise une projection cartographique (ou transformation plane), qui définit la distorsion appliquée au globe pour l'« aplatir ». Il existe une multitude de projections car il n'est pas possible d'aplatir un objet sphérique sans que certaines parties soient fortement déformées. En pratique, il faut donc choisir la projection qui déforme le moins possible la région étudiée. Les coordonnées projetées sont particulièrement utiles pour le calcul des distances, des aires ou des opérations géométriques.

La projection de Mercator (EPSG :3857) est une transformation qui préserve les angles mais pas les aires. Elle est très utilisée pour représenter les cartes du monde et dans les services de cartographie (ex : OpenStreetMap). Les distorsions sont de plus en plus importantes à mesure qu'on s'éloigne de l'équateur.

Dans la thèse, les projections locales utilisées sont la projection de **Lambert-93** (EPSG :2154) pour les études réalisées en France, et la projection de **Yoff** (EPSG :31028) pour celles réalisées au Sénégal.

2.3.3 Représentation sous forme de données informatiques

En informatique, les données géo-spatiales sont principalement stockées sous deux formes : le format *raster* et le format vectoriel.

Le format raster est un format matriciel : l'espace est partitionné en une grille d'unités spatiales, qui correspondent chacune à un pixel [82]. La couleur du pixel dépend de la valeur associée à l'emplacement (l'altitude du terrain par exemple). Plus la résolution est élevée, plus la donnée est précise. L'acquisition se fait principalement par photographie aérienne ou satellite. Les données raster sont couramment utilisées pour afficher des fonds de carte, des cartes d'utilisation des sols ou d'élévation de terrain. Le stockage et l'affichage des données est moins complexe que les données vectorielles. C'est cependant ce deuxième format que l'on utilisera car il permet de manipuler les données pour étudier les relations spatiales entre les objets.

Le format vectoriel Les données vectorielles décrivent la forme et la localisation d'un objet à l'aide de primitives géométriques. L'objet représenté peut être de nature variée :

- une statistique démographique, la somme des niveaux de vie des habitants dans un carré géographique (des données carroyées par exemple).
- une surface caractérisée par une utilisation du sol.
- une infrastructure comme un bâtiment, une route, un aménagement urbain.
- des éléments naturels : rivières, forêts, espaces verts.
- des frontières administratives.

Les données manipulées dans la thèse utilisent les formats standard *Well-Known Binary* (WKB) pour la représentation machine et *Well-Known Text* [83] pour la représentation humaine. Le WKT définit trois primitives géométriques :

— le type `Point`, est utilisé pour positionner les objets ponctuels, ou que l'on peut rapporter à un point. Il est caractérisé par deux coordonnées géographiques ou cartographiques.
Exemple de syntaxe : `POINT(x y)`.

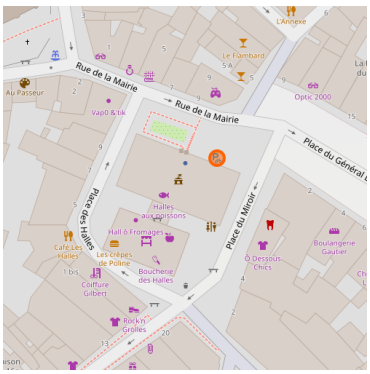
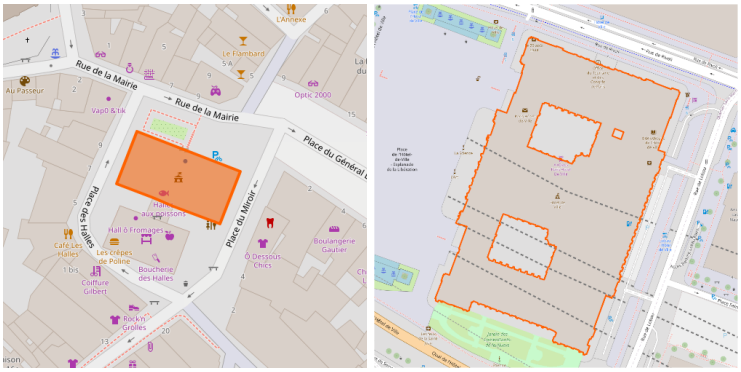


FIGURE 2.6 – Un parking à vélos représenté par un point. Source : OpenStreetMap [84]

— le type `Linestring`, constitué d'une suite de points, il donne le tracé d'une ligne brisée en positionnant les extrémités des segments dans l'ordre donné.
Exemple de syntaxe pour une ligne à 4 points : `LINESTRING(x0 y0, x1 y1, x2 y2, x3 y3)`

FIGURE 2.7 – Une rue représentée par une ligne brisée. Source : OpenStreetMap

— le type `Polygon` formé d'une suite de points dont le dernier est confondu en position avec le premier. Il décrit les formes des surfaces, qui sont représentées fermées, avec ou sans trous à l'intérieur.
Exemple de syntaxe pour un triangle : `POLYGON((x0 y0, x1 y1, x2 y2, x0 y0))`
Exemple de syntaxe pour un triangle avec un trou : `POLYGON((x0 y0, x1 y1, x2 y2, x0 y0), (x4 y4, x5 y5, x6 y6, x4 y4))`



(a) Hôtel de ville de Lannion (b) Hôtel de ville de Paris

FIGURE 2.8 – Deux hôtels de ville représentés par un polygone plein (gauche) et creux (droite). Source : OpenStreetMap

Les géométries peuvent être regroupées en groupes de même type (`MultiPoint`, `MultiLineString`, `MultiPolygon`) ou non (`GeometryCollection`). Les informations de localisation peuvent être représentées avec des coordonnées géographiques ou projetées.

2.3.4 Manipulation des données : les logiciels et bibliothèques SIG

Les données sont stockées et traitées à l'intérieur d'un système d'information géographique (SIG). Un SIG est un système remplissant de nombreuses fonctions, dont le stockage, le traitement et la visualisation des données cartographiques. Dans la thèse, on utilisera principalement les outils suivants :

— QGIS [85] : pour la visualisation des données vectorielles et raster. Parmi ses nombreuses fonctionnalités, il permet le chargement de multiples couches de données et l'utilisation de fond de carte provenant de services tiers comme par exemple Google Earth, MapBox ou OpenStreetMap.

— PostgreSQL [86]/PostGIS [87] : PostgreSQL est un système de gestion de base de données relationnelle. Les données sont récupérées en effectuant des requêtes SQL. L'extension PostGIS rajoute des structures pour optimiser le stockage et l'indexation des données géospatiales, des fonctions pour effectuer des traitements sur les données géométriques (dont font partie les données vectorielles), et faciliter la conversion entre systèmes de coordonnées.

— GeoPandas [88], shapely [89] : GeoPandas est un module Python qui est basé sur le module de traitement de données tabulaires Pandas. Il rajoute un type pour faciliter l'analyse des données spatiales. Il propose également des fonctionnalités d'affichage permettant l'exploration des données, et fait appel à la librairie shapely pour effectuer les opérations géométriques.

Les implémentations de PostGIS, GeoPandas et shapely veillent à se rapprocher des standards définis par l'*Open Geospatial Consortium* (OGC), qui est également à l'origine des formats WKB/WKT.

Les opérations géo-spatiales disponibles dépendent de la bibliothèque utilisée. Sans toutes les énumérer, la liste des fonctions géométriques utiles au cours de la thèse comporte entre autres :

— l'obtention de formes résultant de l'intersection ou de l'union de deux formes initiales.

— la division de polygones (Figure 2.9).

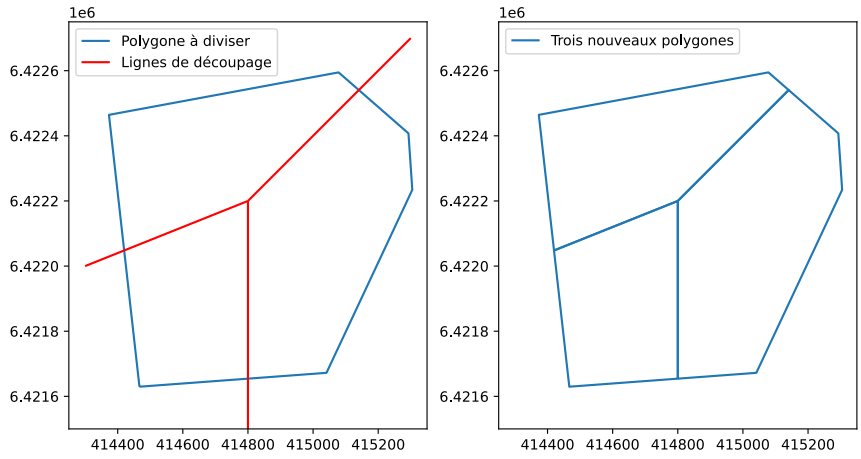


FIGURE 2.9 – Division d'un polygone suivant un ensemble de lignes grâce à la fonction découpage de PostGIS

— l'obtention d'enveloppes convexes, concaves ou du cercle minimum englobant un objet (Figure 2.10).

— la possibilité de réaliser des jointures spatiales entre deux tableaux sur la condition que les géométries des deux tableaux présentent une intersection non nulle. La Figure 2.11 illustre la fonction qui permet de récupérer les bâtiments intersectant avec un polygone quelconque. Les données de l'image de gauche peuvent être stockées sous forme de deux tableaux : un tableau A contenant les formes des bâtiments, et un tableau B ne contenant qu'un polygone. Les données de l'image de droite sont stockées dans un tableau C composé des lignes de la table B qui intersectent spatialement avec A.

— le calcul d'aires, de longueurs, la création de tessellations (ex : Voronoi), etc.

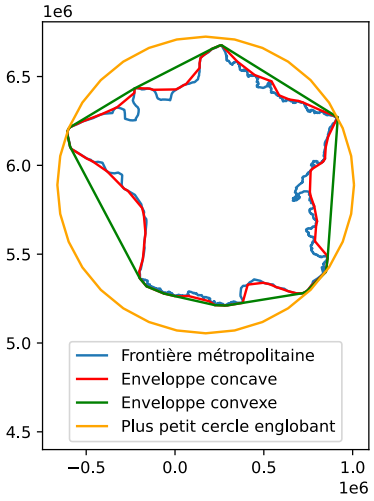


FIGURE 2.10 – Illustration des enveloppes convexes et concaves et du plus petit cercle englobant les frontières de France métropolitaine

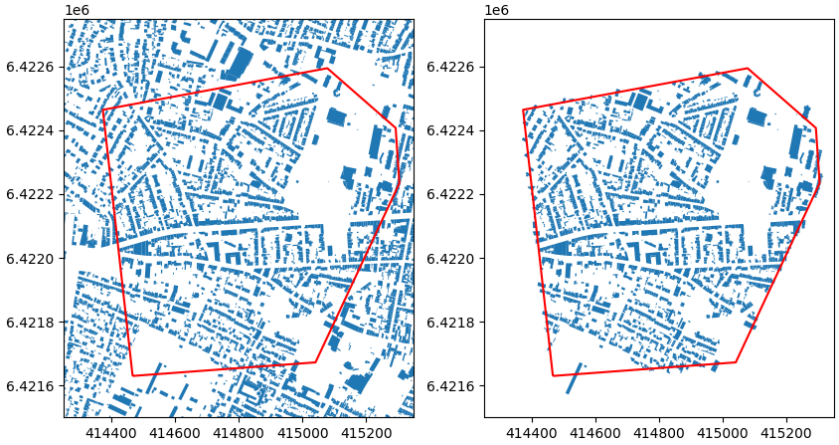


FIGURE 2.11 – Filtrage des bâtiments intersectant le polygone grâce à la fonction intersection de PostGIS

2.3.5 Sources de données utilisées

Les données externes utilisées dans les contributions proviennent essentiellement de deux sources : OpenStreetMap et la plateforme Humanitarian Data Exchange. Le choix d'utiliser des sources publiques et mondiales permet de favoriser la reproductibilité des méthodes proposées dans la thèse, pour les rendre facilement applicables à d'autres régions voire pays. Les Chapitres 7 et 8 prenant la France comme zone d'étude, on utilise en plus les données démographiques de l'Institut national de la statistique et des études économiques.

La mise en cohérence de toutes ces sources nous permet d'avoir des informations sur le tissu urbain d'une région. On définit par ce terme l'ensemble des infrastructures, des bâtiments et équipements composant un espace urbain.

La base de données OpenStreetMap (OSM) est utilisée pour obtenir les descriptions géographiques de l'utilisation des sols et des points d'intérêts [84]. On rappelle ici la structure des données OSM qui est documentée dans le wiki du projet [2]. Les géométries primitives sont le nœud (*node*) et le chemin (*way*). Leurs usages sont principalement les suivants :

- le nœud, assimilable au type `Point` du format WKT, permet de positionner les points d'intérêts.
- Le chemin ouvert (*open way*), composé d'une suite de nœuds, est assimilable à la description de la `LineString` du format WKT. De manière générale, les axes de transports, les cours d'eaux, les lignes électriques,... ont pour géométrie le chemin ouvert.

2. https ://wiki.openstreetmap.org/

— Le chemin fermé est un chemin dont le dernier nœud est confondu avec le premier. Il est assimilable au type Polygon du format WKT. Les formes de certaines infrastructures comme les bâtiments et les zones d'utilisation du sol sont décrits par des chemins fermés.

Ces éléments, ou objets, peuvent être regroupés en collections qu'on appelle les relations.

Pour un objet de type nœud, chemin ou relation, on associe un ou plusieurs attributs (ou balise, en anglais : *tags*) pour apporter des précisions sur sa nature. Un attribut est composé d'une paire clé/valeur, exprimée sous la forme « `clé=valeur` ». Par exemple, l'un des attributs les plus répandus sur les polygones est « building=yes » pour signifier qu'il s'agit de bâtiments. Le projet OSM est collaboratif et permet aux contributeurs d'ajouter librement les attributs qu'ils souhaitent. Cependant, à des fins d'unification pour l'affichage et le regroupement des données, une liste de balises recommandées est fournie par la communauté. L'avantage de ce système et qu'il permet d'avoir une cartographie très détaillée, et évoluant au fil du temps. Le désavantage rencontré pour l'apprentissage automatique est la présence d'un grand nombre varié d'attributs non exploitables car non informatifs (noms de marques, numéros de bâtiment par exemple) ou très peu représentés. Dans le Chapitre 3.3.4, une étude est conduite pour évaluer la qualité du regroupement des attributs par sémantique grâce à des outils du langage naturel.

The Humanitarian Data Exchange est une plateforme rassemblant des données publiques humanitaires publiées par des tiers. Dans le cadre de nos travaux, on utilise les cartes de distribution des populations françaises et sénégalaises publiées par Meta [90]. Ces cartes possèdent une granularité spatiale de 30×30 mètres.

L'Institut national de la statistique et des études économiques (INSEE) publique environ tous les cinq ans, les données socio-économiques de près de 30 millions de ménages distribués sur le territoire français. Ces données se présentent sous forme de carreaux de tailles variable, pouvant aller de 32 km à 200 m de côté. La dernière mise à jour de 2022 fournit les données du dispositif sur les revenus localisés sociaux et fiscaux de 2017 [91]. Chaque carreau renferme des informations sur sa taille, sa localisation ainsi qu'un ensemble de variables dont par exemple : le nombre des individus par tranche d'âge, la taille des ménages, les niveaux de vie, l'âge des logements,...

2.4 Discussion

2.4.1 En résumé

La Section 2.1 débute le chapitre en donnant une description de la partie du réseau mobile concernée par les travaux de thèse, ainsi que les données collectées. On a présenté dans la Section 2.2 le principe de l'apprentissage automatique et leur application aux données mobiles. On remarque cependant que la plupart des solutions existantes ne sont pas adaptées à la prédiction du trafic de futures infrastructures. La difficulté réside dans l'identification des données à passer en entrée aux modèles, à défaut de posséder des données historiques sur le trafic de la zone à couvrir. On propose d'utiliser des sources de données externes qui permettent de modéliser la demande utilisateur à travers le tissu urbain dans la Section 2.3.

Pour clôturer ce chapitre, on discutera de la place des contributions de la thèse par rapport aux méthodes de déploiement classique. On présente d'abord les grandes lignes du processus de planification. Ensuite on présente le paradigme dans lequel s'inscrivent les solutions de la thèse, à savoir la planification guidée par la donnée. Enfin, on discutera des avantages et des limites que peut apporter cette nouvelle approche dans le processus actuel.

2.4.2 Les grandes lignes de la planification d'un réseau

On distingue deux étapes dans le processus de planification :

1. des études géo-marketing sont réalisées pour estimer la demande des utilisateurs et définir les emplacements d'intérêt pour un nouveau déploiement.
2. à partir de la demande attendue et des critères de qualité de service, on recherche le dimensionnement et de la configuration radio répondant à ces critères.

Études géo-marketing

Les territoires concernés par un nouveau déploiement sont identifiés et étudiés en prenant en compte un grand nombre de données :

— les images satellites et la distribution de population pour identifier les populations non couvertes par une technologie (notamment en zone rurale)
— les infrastructures de transport (comme les chemins de fer) pour optimiser le raccordement des nouveaux sites au réseau existant.
— le contexte socio-économique pour comprendre le profil des clients ou des futurs clients ainsi leur usage du réseau mobile.

Ces études sont en général réalisées sur le long terme. Par exemple, des travaux ont été réalisés pour estimer la demande régionale annuelle en trafic sur plusieurs années, en combinant les statistiques de population, les plans d'aménagement du territoire et les parts de marchés des opérateurs [92].

Études de dimensionnement

Le travail de dimensionnement prend en compte trois éléments : la demande, les performances et la configuration du réseau. En supposant deux des trois éléments connus, l'objectif est d'optimiser le troisième de manière à répondre aux deux objectifs fixés. Pour cela, des formules mathématiques analytiques ou semi-empiriques sont utilisées pour modéliser les relations entre tous les éléments. Par exemple, en supposant la demande connue (grâce aux études précédentes) et pour une infrastructure réseau donnée, il est possible de proposer des modèles mathématiques pour estimer le débit moyen utilisateur [93].

2.4.3 Approche automatique

Grâce aux données mobiles massives, l'apprentissage automatique permet d'aborder le problème de la planification sous une perspective différente, qui ne correspond pas exactement aux deux étapes décrites précédemment. Qu'il s'agisse de compensation de panne de cellules [94], de configuration d'antennes pour améliorer la qualité de service [95] ou de positionnement de futurs sites pour maximiser la couverture [80], les travaux de recherche récents s'accordent sur une architecture commune plaçant les modèles d'apprentissage au cœur des simulations du comportement du réseau mobile.

Cette architecture peut être qualifiée de « guidée par la donnée » (*data-driven*) car les décisions reposent sur les prédictions d'un modèle d'apprentissage paramétré grâce à des données historiques. Elle est fondée sur deux étapes :

1. L'entraînement d'un modèle d'apprentissage pour la prédiction des performances du réseau mobile, en utilisant des données géographiques, socio-économiques (comme celles utilisées dans les études de géo-marketing) et éventuellement sur la configuration du réseau mobile (Figure 2.12, Étape I). Implicitement, le modèle apprend à lier les caractéristiques d'un territoire et les équipements déployés à la demande des utilisateurs. On espère ainsi qu'il aura appris les bons exemples de déploiement. L'étape suivante permettra de trouver les bonnes configurations pour les déploiements futurs.
2. Tester différentes configurations de déploiement en fournissant ces configuration au modèle entraîné, et garder celle maximisant les objectifs fixés à l'avance : couverture, trafic, débit, etc. (Figure 2.12, Étape II). À cette étape, le modèle d'apprentissage fait office de simulateur des performances du réseau.

Les contributions de la thèse développent l'Étape I en fournissant des simulateurs de performance du réseau mobile entraîné sur des données passées. La recherche du déploiement optimal à l'Étape II dépasse le cadre de cette thèse et fait l'objet de travaux futurs.

Avantages et limites

Avantages L'Étape I de l'approche automatique combine les deux étapes de la planification traditionnelle. En effet, les modèles apprennent conjointement la demande d'un territoire (études géo-marketing) et la réponse apportée par l'opérateur en termes d'infrastructures déployées (études de dimensionnement). Ces deux éléments sont automatiquement appris et combinés pour estimer les performances du réseau. Les avantages que l'on peut espérer tirer de cette architecture sont :

— un gain de temps du traitement des données de sources différentes et de leur mise en relation, en filtrant les données non pertinentes.
— une augmentation du réalisme des simulations grâce à une meilleure estimation de la demande attendue, cette dernière étant directement représentée par des éléments géo-spatiaux précis : points d'intérêts, aménagements du territoire, espaces peuplés ou non...

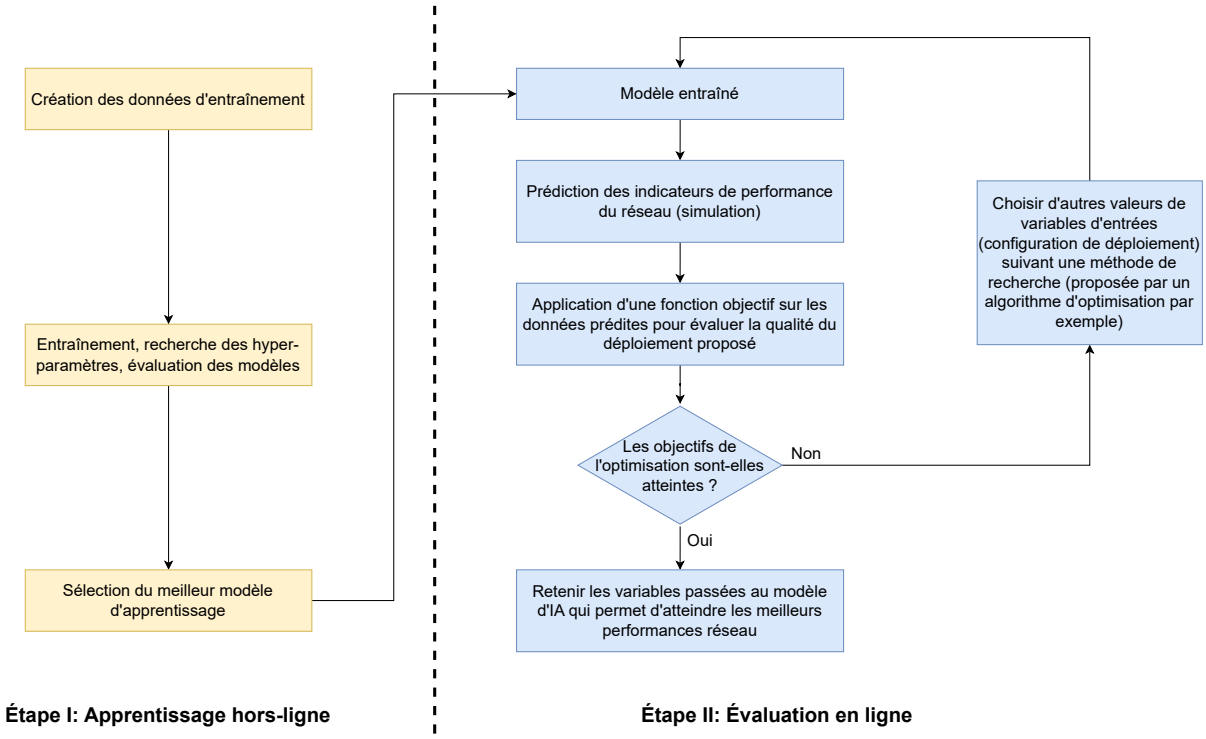


FIGURE 2.12 – Architecture globale d'un outil de planification utilisant comme données de simulation les prédictions d'un modèle d'apprentissage.

— quantification de la fiabilité des simulations grâce aux erreurs de prédiction sur des données de test à l'Étape I.

Limites Comme les modèles analytiques classiques, les prédictions des modèles d'apprentissage reposent sur des hypothèses sur le comportement des usagers du réseau et sur la maturité des technologies. Cependant, ces hypothèses ne sont pas explicitées au moment de l'entraînement. Il est cependant important de les connaître pour identifier les limites des simulations. La plus grande supposition est que le territoire sur lequel on effectue des déploiements possède les mêmes propriétés que des territoires vus pendant l'entraînement. En particulier, on peut citer les aspects suivants :

— état du déploiement du réseau : si l'entraînement est réalisé sur une technologie dont le déploiement est mature (i.e. pénétration de la 4G/5G), alors les performances prédites sont supposées celles d'un réseau de même maturité.
— part de marché : les prédictions supposent que le territoire couvert par les déploiements présentera des parts de marché similaires à celles des territoires vus pendant l'entraînement, au même stade de maturité de la technologie.
— offre/demande : pour des prédictions fiables sur un nouveau territoire, il faut que les dynamiques d'utilisation, d'offre et de demande soient similaires aux territoires vus pendant l'entraînement, au même stade de maturité technologique.

Chapitre 3

Segmentation des sites en fonction de leur affluence

3.1 Introduction

En 2021, l'Union internationale des télécommunications estime que 95% de la population mondiale est couverte par la 3G ou une technologie supérieure [96]. Pour autant, les problématiques d'extension de couverture sont toujours d'actualité car la couverture apportée par la 3G n'est pas suffisante pour accéder à une connectivité efficace et de haut débit. Dans le rapport de l'ONU sur la connectivité des Pays les Moins Avancés (PMAs), l'organisation estime qu'en 2025, 26 PMAs seront toujours loin d'atteindre l'objectif 9.c du développement durable concernant la connectivité universellement accessible. Le déploiement de la couverture 4G et 5G un accès à Internet haut-débit essentiel au développement économique en permettant une meilleure inclusion financière, un accès aux soins et à l'éducation facilité par la téléconsultation et la visioconférence, et une modernisation de l'agriculture.

Le déploiement de nouvelles stations de base est coûteux, c'est pourquoi les opérateurs ont tout intérêt à identifier les sites rentables apportant la connectivité au plus grand nombre de personnes. Dans ce contexte, que peut apporter l'apprentissage automatique pour faciliter les déploiements ? En partant de données publiques sur les statistiques d'appels et de données exogènes, on présente dans ce chapitre une méthode d'apprentissage automatique permettant de prédire une information simple sur le profil d'activité d'un futur site. Cette méthode se décompose en deux parties :

1. On regroupe d'abord les profils d'affluence des sites (volume d'appels au cours du temps) en utilisant l'apprentissage non supervisé, puis on interprète les classes obtenues en fonction du profil de périodicité et des heures de pointe.

2. Une fois les classes définies et les données mobiles annotées, on propose d'utiliser l'apprentissage supervisé pour prédire la classe d'un futur site à partir des caractéristiques géographiques qu'il couvrira. Les variables liées à la géographie et la démographie proviennent de sources publiques. Les modèles d'apprentissage utilisés dans cette étude sont CatBoost, les forêts aléatoires et les machines à vecteurs de support.

Dans cette étude, nous utilisons des données portant sur la région de Dakar et la ville de Thiès. Les données de distribution de population proviennent des cartes publiées par Meta sur la plateforme Humanitarian Data Exchange [90]. Les données des points d'intérêt et d'usage des sols proviennent de la base de données OpenStreetMap [84], et les statistiques d'appel et de SMS du jeu de données D4D-Senegal [22].

Dans la Section 3.2, on commence par résumer les connaissances de la littérature sur l'analyse conjointe des données mobiles avec l'usage du sol. On présente ensuite dans la Section 3.3 les méthodes utilisées pour regrouper et classifier les données mobiles. La Section 3.4 traite des résultats de l'apprentissage des modèles. Enfin, la Section 3.5 conclut ce chapitre.

3.2 Quantifier et qualifier les relations entre les données mobiles et le tissu urbain

Les études réalisées dans un grand nombre de pays différents ont montré qu'il existe une corrélation entre le volume d'appels des stations de base, l'utilisation des sols et les points d'intérêt qu'ils couvrent. L'utilisation du sol d'une région géographique correspond à son exploitation humaine pour une activité spécifique : industrielle, commerciale, agricole, etc. À l'échelle du bâtiment, les points d'intérêt (PI) décrivent les sites remarquables concentrant des activités humaines précises. Les PIs peuvent être par exemple des centres commerciaux, des restaurants ou des écoles.

Utilisation du volume d'appels pour caractériser l'utilisation des sols

La téléphonie mobile accompagnant le quotidien des hommes, les données générées par l'usage des réseaux fournissent des informations précieuses sur l'utilisation des sols. Plusieurs études (résumées dans le Tableau 3.1) ont été menés pour comparer les similitudes géographiques entre les volumes d'appel et les cartes d'utilisation des sols.

L'analyse du volume d'appel des stations de base est réalisée en regroupant les données grâce à des algorithmes de groupement par similarité ou à une source externe pour annoter les sites. Cela permet de résumer la distribution des données mobiles en un nombre réduit de classes, et de faciliter leur compréhension et leur comparaison avec les cartes.

Le nombre de classes de volume d'appel obtenu varie entre trois et cinq selon les études citées précédemment. On relève deux manières de comparer les similitudes géographiques entre la distribution de l'activité des sites et les cartes d'utilisation des sols :

- Un modèle est entraîné de manière supervisée pour prédire la classe d'utilisation du sol à partir des données mobiles. Plus la corrélation entre l'activité des réseaux et l'utilisation des sols est grande, plus la précision globale du modèle de classification est élevée.
- Dans les cas non supervisés, l'association des classes aux positions des cellules permet d'obtenir une carte découpée en zones d'activité du trafic. En utilisant la mesure de précision globale, cette carte peut ensuite être comparée à une carte d'utilisation des sols, cette dernière servant de vérité de terrain.

Référence	Méthode de classification	Classes d'activités	Degré de similarité
[97]	Annotation des données avec une carte d'urbanisation	— résidentielle — commerciale — industrielle — parcs — divers	Le modèle de classification montre une précision globale de 54% en comparaison avec la carte d'urbanisme de l'aire métropolitaine de Boston.
[98]	Détection de communauté	— résidentiel — quartiers d'affaires — industriel — nocturne	La comparaison avec le cadastre de Madrid et de Barcelone donnent des précisions respectives de 65% et de 60%.
[99]	Partitionnement diffus	— résidentielle — industrielle — commerciale — espaces ouverts — autres	La comparaison avec un plan d'urbanisme de Singapour donne une précision globale de 58.03%
[72]	K-moyennes	— résidentielle — industrielle — mixte	Comparaison qualitative uniquement dans les régions de Madrid et Barcelone

TABLE 3.1 – Résumé des méthodes utilisées dans la littérature pour étudier l'activité des réseaux mobiles conjointement avec l'utilisation des sols

Ces études ont mis en évidence un lien entre la nature de l'utilisation du sol et l'activité des réseaux mobiles. Les heures de pointe du réseau ainsi que leur intensité dépendent des activités humaines, rythmées

par des routines de travail, loisirs et déplacements pendant lesquels le téléphone est très souvent utilisé. Deux classes sont communes à toutes les études : la classe résidentielle et la classe industrielle. Dans les quartiers résidentiels, le réseau mobile rencontre un pic d'activité le soir, après que les personnes soient rentrées du travail. Dans les quartiers industriels, le pic d'activité survient en pleine journée, uniquement pendant les jours de semaine travaillés.

Plusieurs auteurs constatent cependant que les cartes d'urbanisme sont parfois inexactes (plan prévisionnel, ou avec granularité géographique insuffisante), ce qui peut expliquer la précision limitée des comparaisons. Il est donc possible que les données mobiles reflètent plus exactement et fidèlement la manière dont un territoire est réellement exploité.

Exploitation des points d'intérêt pour la prédiction de l'activité mobile

Les points d'intérêt, plus ponctuels géographiquement, permettent de décrire plus précisément l'activité humaine. Chaque classe d'activité du réseau mobile identifiée peut être caractérisé par des proportions spécifiques de types de points d'intérêt [53]. Cette spécificité peut aussi être validée en mesurant l'homogénéité des proportions des points d'intérêts par classe [100].

Il est possible d'aller plus loin en utilisant les points d'intérêt comme variables d'entrées pour affiner les prédictions de trafic des modèles automatiques [25, 78]. Dans ces travaux, les prédictions sont dynamiques : elles reposent non seulement sur les données exogènes, mais aussi sur les données du trafic passé. Cependant, dans une situation d'extension de couverture, on ne connaît pas le trafic attendu du réseau mobile dans la zone étudiée. Dans la section suivante, la méthode que l'on développe s'appuie entièrement sur des données cartographiques fines pour prédire le type d'activité du réseau mobile.

3.3 Méthode d'apprentissage pour la classification

3.3.1 Sources des données

Les données utilisées dans cette étude proviennent de trois sources :

— D4D-Senegal Challenge pour les statistiques d'appel et de SMS provenant du réseau mobile d'un opérateur privé au Sénégal (voir Section 2.1.2).
— OpenStreetMap pour les points d'intérêt et l'utilisation des sols (voir Section 2.3.5).
— Humanitarian Data Exchange pour la distribution des populations (voir Section 2.3.5)

Concernant les statistiques d'appel et de SMS, on manipule des données agrégées par station de base. On obtient une série temporelle qui permet d'estimer le nombre d'utilisateurs connectés à chaque station au cours du temps. Un exemple de série est donnée sur la Figure 3.1.

3.3.2 Énoncé du problème

Dans cette section, on présente la méthode utilisée pour prédire le profil d'affluence d'une station de base. Ces profils n'étant pas définis dans le jeu de données initial, on commence par regrouper les séries temporelles de manière non supervisée pour créer des classes et annoter les stations de base.

Création des classes (cibles d'apprentissage)

Le regroupement en classes est réalisé sur les signatures hebdomadaires médianes (*Median Week Signature* [100]) des stations de base. La signature hebdomadaire médiane du nombre d'utilisateurs d'une station de base peut être considéré comme la compression de la série temporelle pour n'en garder que les caractéristiques d'une semaine typique (heure de pointe, périodicité). Ainsi, une classe est un ensemble de stations de base partageant des heures de pointe et des heures creuses similaires.

Signature hebdomadaire médiane La signature hebdomadaire médiane est initialement formalisée dans les travaux de Furno et al. [100]. On en propose ici une ré-écriture en adoptant une notation vectorielle.

Soit $\mathcal{B} = \{0, 1, 2, ..., N-1\}$ les identifiants de N stations de base. On note v^i l'indicateur de performance mesuré périodiquement pour la station de base d'identifiant $i \in \mathcal{B}$. Dans cette étude, v^i correspond au nombre d'utilisateurs.

Soient $\mathcal{D}$ l'ensemble des jours (au format `YYYY-MM-DD`), $\mathcal{T}$ l'ensemble des heures (au format `HH:MM:SS`) et $H = \{d \in \mathcal{D}, t \in \mathcal{T} | h = (d, t)\}$ l'ensemble des couples (date, heure) pour lesquels les indicateurs sont

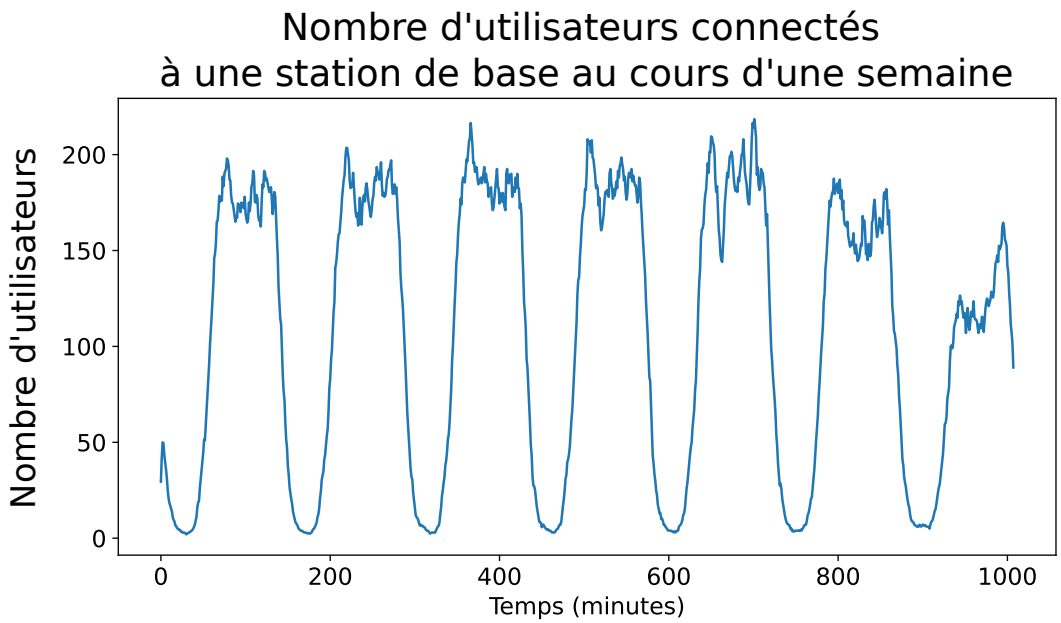


FIGURE 3.1 – Nombre d'utilisateurs d'une station de base sur une semaine construite à partir des statistiques d'appel et de SMS

collectés. La série temporelle v^i est donc de taille $|H|$. On note $v^i(h)$ la valeur de l'indicateur à la date $h = (d, t)$ pour la station i. Les séries temporelles de toutes les stations de base forment l'ensemble $\mathcal{V} = \{v^i(h) | h \in H, 0 \leq i \leq N-1\} \subset \mathbb{R}$.

On définit la fonction dow qui donne le jour de la semaine d'une date $d \in \mathcal{D}$:

$$\text{dow(d)} \in \{\text{MON, TUE, WED, THU, FRI, SAT, SUN}\}$$

La signature hebdomadaire médiane est le résultat de la fonction s :

$$s : \mathcal{V}^{|H|} \longrightarrow \mathbb{R}^u$$

$$(v^i(h_1), ..., v^i(h_{|H|})) \longmapsto s(v^i(h_1), ..., v^i(h_{|H|})) = \begin{bmatrix} \mu_{\frac{1}{2}}(\{v^i(d,t) | \text{dow}(d) = \text{MON}, t = t_1\}) \\ \mu_{\frac{1}{2}}(\{v^i(d,t) | \text{dow}(d) = \text{MON}, t = t_2\}) \\ \vdots \\ \mu_{\frac{1}{2}}(\{v^i(d,t) | \text{dow}(d) = \text{SUN}, t = t_{|\mathcal{T}|}\}) \end{bmatrix}$$

où u est la longueur de la signature et $\mu_{1/2}(\cdot)$ est l'opérateur médian.

Si l'on considère une granularité horaire, la signature est un vecteur de taille :

$$u = 7 \text{ (days)} \times 24 \text{ (hours)} = 168$$

Normalisation L'étude ne portant que sur l'allure des signatures et non leur trafic, la signature s_i d'une station est ensuite normalisée suivant le score standard :

$$z_i = \frac{s_i - \overline{s_i}}{\sigma(s_i)},$$

Regroupement La méthode des k-moyennes est utilisée pour regrouper les signatures normalisées. Elle est paramétrée par le nombre k de groupes à former. La méthode Elbow permet de trouver la valeur k idéale minimisant la variance intra-classe et maximisant celle inter-classe. En l'appliquant aux données de Dakar et Thiès, elle suggère de partitionner les signatures en $k = 4$ ou $k = 5$ groupes. Cependant, pour $k > 3$, plusieurs groupes partagent la même heure de pointe, avec des amplitudes de volume différentes. On préférera utiliser le paramètre $k = 3$ pour n'avoir que des groupes d'heures de pointe différentes, ce qui rejoint le choix de travaux précédents sur ce jeu de données [101].

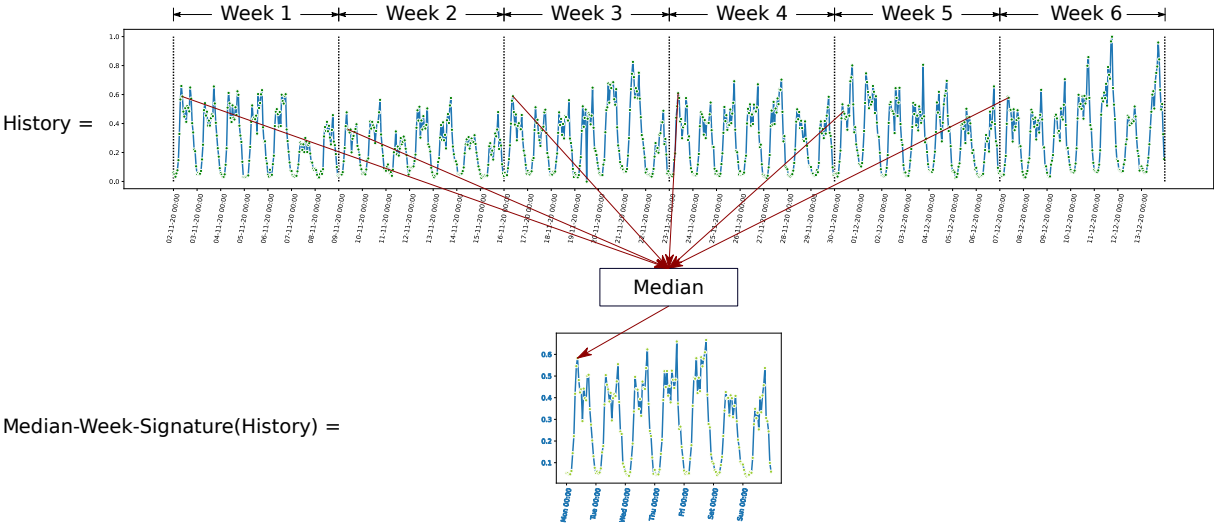


FIGURE 3.2 – Construction d'une signature hebdomadaire médiane

L'application des k-moyennes permet d'attribuer une classe à chaque station de base en fonction du profil de sa signature. C'est cette classe qu'on cherche à prédire pour les futures stations de base à déployer. Les signatures annotées, couplées aux deux sources externes, constituent le jeu d'entraînement.

3.3.3 Classification

La tâche de classification consiste à attribuer une classe de trafic à une station de base en se basant sur le tissu urbain couvert par celle-ci.

Formalisation Soit $K \in \mathbb{N}$ le nombre de classes d'activités. Soit N le nombre de stations de bases du jeu d'entraînement. Soit $i \in \mathbb{N}$ l'identifiant d'une station de base et (X_i, y_i) l'observation associée :

— $X_i \in \mathbb{R}^p$ est un vecteur de $p \in \mathbb{N}$ variables, décrivant le tissu urbain de la région couverte par la station de base i. Le j-ème élément du vecteur X_i correspond soit au nombre de points d'intérêt, soit au nombre de fois où une utilisation du sol est recensée, ou soit à la population comptabilisée dans la couverture de la station. Le détail des variables est donné dans le Tableau 3.3.

— $y_i \in \{1, ..., K\}$ correspond à la classe à laquelle appartient la station de base i. Les stations de base appartenant à la même classe possèdent des propriétés d'heure creuse et heure de pointe similaires.

Un modèle d'apprentissage de classification peut être vu comme une fonction f tel que pour toute station de base i :

$$\begin{aligned} f : \mathbb{R}^p &\longrightarrow \{1, ..., K\} \\ X_i &\longmapsto f(X_i) = \hat{y}_i \end{aligned}$$

où $\hat{y}_i$ est la prédiction du modèle. Une prédiction est correcte si elle est égale à la vraie classe, autrement dit, si $\hat{y}_i = y_i$. La performance du modèle est évaluée à l'aide de quatre mesures : la précision globale (toutes classes comprises), la précision, le rappel et le F-score (pour chaque classe). Le calcul de ces mesures est donné dans la Section 2.2.6.

Modèles d'apprentissage pour la classification

Afin de savoir si le jeu de données constitué contient des variables informatives, les modèles suivants ont été entraînés et évalués : CatBoost [61, 102], les forêts aléatoires (RF) et les machines à vecteurs de support (SVC) [59]. Les hyperparamètres retenus ont été obtenus en utilisant une recherche par quadrillage (*grid search*). Ils sont rapportés dans le Tableau 3.2.

Lignes de base Pour valider l'efficacité de la méthode proposée, on définit également deux modèles simplistes produisant des performances de référence qui devront être dépassés par les modèles précédents.

— Le modèle B1 catégorise toutes les stations de base dans la classe la plus fréquente statistiquement.

— Le modèle B2 est entraîné uniquement sur les variables liées à la distribution des populations, pour évaluer l'importance des variables géographiques.

Modèle	Hyperparamètres
CatBoost	`depth=3, l2_leaf_reg=8, iterations=200, learning_rate=0.15`
CatBoost B2	`depth=6, l2_leaf_reg=4, iterations=250, learning_rate=0.60`
RF	`max_depth=9, min_samples_leaf=1, min_samples_split=4, n_estimators=70`
RF B2	`max_depth=7, min_samples_leaf=2, min_samples_split=3, n_estimators=30`
SVC	`kernel='linear', C=8.3`
SVC B2	`kernel='rbf', C=4.0`

Table 3.2 – Hyperparamètres utilisés pour chaque modèle d'apprentissage

Variables d'entrée du modèle d'apprentissage

Afin de créer les variables du jeu d'entraînement, les données d'OpenStreetMap et de Humanitarian Data Exchange sont croisées aux aires de service des stations de base, modélisées par le diagramme de Voronoi.

Formalisation Soit $\mathbf{a} = \{a_i\}_{i=0}^{N-1}$ l'ensemble des régions de Voronoi modélisant la couverture de N stations de base. Pour une station de base i de couverture a_i, on note $X_i = (x_i^{(1)}, ..., x_i^{(p)}) \in \mathbb{R}^p$ le vecteur des p variables géographiques et de population couvertes par a_i.

Variables OpenStreetMap Les données géographiques stockées dans la base de données OSM sont décrites par des attributs. Un attribut est une paire `clé=valeur`.

Pour $j \in \{1, ..., p-3\}$, soit $x_i^{(j)}$ une variable issue d'OSM. Celle-ci est obtenue à partir de tous les objets géographiques couverts par la station i, de clé k_j et dont les valeurs appartiennent à un ensemble V_j. Les clés et les ensembles de valeurs ont été choisies empiriquement, en testant les groupements de valeurs améliorant la classification. La Table 3.3 résume les variables OSM utilisées, ainsi que la clé et les valeurs dont elles dépendent.

On définit l'ensemble des données géographiques associées à la variable j (toutes observations confondues) de la manière suivante :

$$F_j = \{\text{OBJ}(k_j, v) | v \in V_j\}$$

où $\text{OBJ}(k, v)$ est un objet géographique de clé k et de valeur v. Dans cette étude, on considère uniquement les objets dont la forme géométrique est un point ou un polygone.

On note $F_{j|a_i} \in F_j$ l'ensemble des objets possédant une intersection géographique non vide avec la couverture a_i.

La relation entre $F_{j|a_i}$ et $x_i^{(j)}$ est :

$$x_i^{(j)} = |F_{j|a_i}|$$

où $|\cdot|$ désigne la cardinalité de l'ensemble.

Autres variables Les trois dernières variables sont définies comme suit :
- $x_i^{(p-2)}$ est la surface de a_i.
- $x_i^{(p-1)}$ est la population à l'intérieur de a_i.
- $x_i^{(p)} = \frac{x_i^{(p-2)}}{x_i^{(p-1)}}$ est la densité de population.

L'implémentation a été réalisée avec les modules Python `overpy` [103, 104] pour la récupération des données OSM et `geopandas` [88] pour le traitement des données.

Exemple de données La Figure 3.3 montre deux exemples de stations de base avec leurs variables et leur classe (obtenue précédemment avec les k-moyennes).

Nom de variable (j)	**Clé** (k_j)	**Valeurs** (V_j)
industrial	landuse	construction, brownfield, industrial, military, commercial, retail
rural	landuse	farmland, farmyard, orchard
university	amenity	university, college
school	amenity	school
swimming_pool	leisure	swimming_pool
sports	leisure	pitch, sports_centre, stadium
shop	shop	kiosk, convenience, supermarket
marketplace	amenity	marketplace
medical	amenity	doctors, clinic, dentist
food	amenity	restaurant, fast_food
bank	amenity	bank
pharmacy	amenity	pharmacy
religion	amenity	place_of_worship
hotel	tourism	hotel, guest_house, apartment
government	office	government
office	office	ngo, igo, diplomatic, foreign_national_agency, administrative
public_transport	public_transport	Toutes les valeurs disponibles
highway	highway	motorway, trunk, primary, secondary, tertiary

TABLE 3.3 – Clé et valeurs utilisées pour le calcul des variables.

site_id	surface	industrial	rural	university	...	office	public_transport	highway	population	density	label
1	3.968813	13.0	0.0	0.0	...	0.0	0.0	3.0	4033.30185	1016.248824	2
2	0.892388	4.0	0.0	0.0	...	0.0	0.0	0.0	901.56159	1010.279911	1

FIGURE 3.3 – Variables et classes associées à deux stations de base. Les variables utilisées pour l'apprentissage sont les colonnes de `surface` à `density`, et la classe cible est la colonne `label`.

3.3.4 Analyse sémantique du tissu urbain avec GloVe

Avant de choisir arbitrairement les variables géographiques, une piste de recherche consistait à regrouper sémantiquement les valeurs des attributs pour former les variables. Cette idée n'a pas abouti car elle nécessitait des approfondissements en traitement du language naturel qui dépassent le cadre de la thèse. Néanmoins, on documente l'analyse réalisée pour les personnes souhaitant reprendre cette idée.

Principe OpenStreetMap est un projet collaboratif dans lequel les contributeurs peuvent décrire les objets géographiques avec un ensemble d'attributs recommandés, mais ils peuvent également proposer les

leurs. Cela résulte en un grande nombre d'attributs, dont certains pouvant décrire la même chose. Pour grouper les attributs similaires, l'idée est de vectoriser les valeurs des attributs avec la représentation GloVe et d'appliquer un regroupement hiérarchique. Pour rappel, un attribut est décrit par une clé et une valeur, sous la forme « `clé=valeur` »

Plongement lexical et vectorisation Le plongement lexical (*word embedding*) désigne une technique du traitement du language naturel qui vise à représenter des mots sous une forme interprétable par une machine. La représentation courante est l'association pour chaque mot d'un vecteur qui encode sa sémantique (d'où le terme vectorisation). Le but est de construire une distribution de vecteurs telle que des mots sémantiquement proches soient également proches en terme de distance vectorielle. La distance employée varie selon les méthodes.

Global Vectors for Word Representation (GloVe) est une représentation vectorielle des mots de la langue anglaise. Elle est publiée par l'université de Stanford sous Licence Apache 2.0 [105]. Les représentations vectorielles existent en plusieurs versions en fonction du corpus sur lequel elles sont pré-entraînées : Wikipedia 2014, Gigaword 5, Common Crawl ou Twitter.

La version de GloVe que l'on utilise est la version 6B, entraînée sur le corpus Wikipedia 2014 et Gigaword 5. Elle contient 400 000 mots vectorisés, pour lesquels plusieurs tailles de vecteurs sont disponibles : 50, 100, 200 et 300. On choisit les vecteurs de taille 50, car le regroupement devrait être plus efficace que dans les dimensions supérieures.

Traitement des attributs On retient les valeurs des 254 attributs les plus courants dans la base de données OSM. Un traitement de certaines valeurs a été nécessaire pour qu'elles puissent correspondre à une représentation vectorielle de GloVe :

- traduction en anglais des valeurs des attributs s'ils sont en français.
- retrait du préfixe « `TYPE` » s'il est présent.
- remplacement du tiret bas « `_` » par un trait d'union « - ».
- si une valeur composée d'un trait d'union n'a pas de représentation vectorielle dans GloVe, on ne garde que la première partie du mot.

Certaines valeurs sont confondues après ce traitement, c'est pourquoi il n'en reste plus que 221. Leur liste est donnée dans la Table 3.4. On peut remarquer que certains mots sont très proches, comme par exemple « *administration* » et « *administrative* », ou « *telecom* » et « *telecommunication* ». Ces paires de valeurs pourraient être rassemblées automatiquement dans un même groupe.

Analyse par regroupement hiérarchique ascendant En partant de groupes isolés composés d'un seul vecteur, ceux-ci sont fusionnés itérativement jusqu'à obtenir un seul groupe.

Soient c_1 et c_2 deux groupes de valeurs vectorisées. En notant $||.||_2$ la norme euclidienne d'un vecteur $u \in \mathbb{R}^{50}$, la mesure de dissimilarité entre c_1 et c_2 utilisée est le saut maximum :

$$d(c_1, c_2) = \max_{x \in c_1, y \in c_2} (||x - y||_2)$$

où les vecteurs $x, y \in \mathbb{R}^{50}$ sont les représentations vectorielles des valeurs de deux attributs.

Implémentation L'algorithme de regroupement hiérarchique utilisé est celui implémenté dans la librairie Scipy [106]. En fixant à $0.8 d_{max}$ la distance maximale entre les groupes, on obtient 12 groupes et une valeur isolée. Les groupes et les relations de proximité entre les valeurs sont présentées dans le dendrogramme de la Figure 3.4 (zoomer sur support numérique pour lire les valeurs).

Interprétation En considérant les thématiques dominantes de chaque groupe, une interprétation possible est la suivante :

- Enseignement, littérature
- Secteur de la santé
- Administration, finance
- Santé, sécurité
- Équipement urbain extérieur, informatique, commerces spécialisé
- Énergie
- Technologie, éléments divers

administration	casino	electrical	hospital	nutrition	stationery
administrative	charging	electronics	household	office	studio
advertising	cheese	embassy	houseware	optician	supermarket
agrarian	chemist	employment	ice-cream	outdoor	surveillance
alcohol	childcare	energy	igo	paint	swimming-pool
animal	cinema	nsfw	incubator	parking	tableware
antiques	clinic	estate	ingo	pasta	tailor
appliance	clothes	fabric	institute	pastry	taxi
architect	cobbler	farm	insurance	payment	telecom
art	coffee	fashion	interior	perfumery	telecommunication
arts	college	fast-food	internet	pet	telephone
association	community	ferry	jewelry	pharmacy	theatre
assurance	company	financial	kindergarten	photo	ticket
atm	compressed-air	fire	kiosk	physician	tobacco
baby	computer	food	kitchen	place	toilets
bag	confectionery	foreign	language	police	townhall
bakery	consulting	forestry	laundry	political	toys
bank	convenience	foundation	lawyer	post	trade
bar	copy	fountain	library	post-office	training
beauty	cosmetics	fuel	lighting	printing	travel
bed	courthouse	furniture	lottery	prison	tyres
bench	couture	gambling	mall	pub	university
beverages	coworking	games	marketplace	public	upholsterer
bicycle	curtain	garden	massage	ranger	vacant
boat	customs	gas	medical	recycling	vacuum
books	dairy	general	mobile-phone	religion	variety
boutique	deli	gift	monastery	research	vending
brothel	dentist	glass	money	restaurant	veterinary
bureau	department	government	mortuary	school	video
bus	diplomatic	grave	motorcycle	scuba-diving	waste
business	diy	greengrocer	musical	seafood	water
butcher	doctors	grocery	newsagent	sewing	watering
cafe	dojo	haberdashery	newspaper	shelter	weapons
camera	drinking-water	hairdresser	ngo	shoes	wedding
car	driving	hardware	nightclub	shop	wholesale
car-sharing	dry-cleaning	herbalist	nursery	social	wine
carpet	educational	hifi	nursing	sports	

Table 3.4 – Liste des valeurs des attributs OSM

— Activités nocturnes
— Éléments divers
— Services liés au bien-être personnel, à la maison
— Magasins de nourriture, boisson, divers, restaurants
— Transport

On observe que plusieurs groupes rassemblent des mots de thématiques proches (par exemple le premier et le dernier sur le dendrogramme, en partant du haut). Les paires de mots (*administration*,*administrative*) et (*telecom*, *telecommunication*) se retrouvent dans les mêmes groupes.

Cependant, la méthode n'est pas parfaite : certains groupes contiennent plusieurs thématiques (5e et 10e groupes), tandis qu'un même thème se retrouve parfois dans plusieurs groupes (2e et 4e groupes). Surtout, le principal problème est le caractère non spécialisé de la représentation vectorielle. Dans un contexte d'étude sur l'activité des réseaux mobiles, les regroupements de certains termes doivent être pensés différemment. Par exemple, ce sont plutôt les valeurs associées à des points d'intérêts concentrant des routines humaines similaires qui devraient se retrouver ensemble. Suivant ce raisonnement, les valeurs « *school* » et « *university* » qui désignent respectivement les écoles primaires et les universités ne devraient pas se retrouver dans le même groupe au risque de dégrader l'apprentissage des modèles. C'est sur la base de ces constats qu'on privilégie le regroupement arbitraire des attributs OSM.

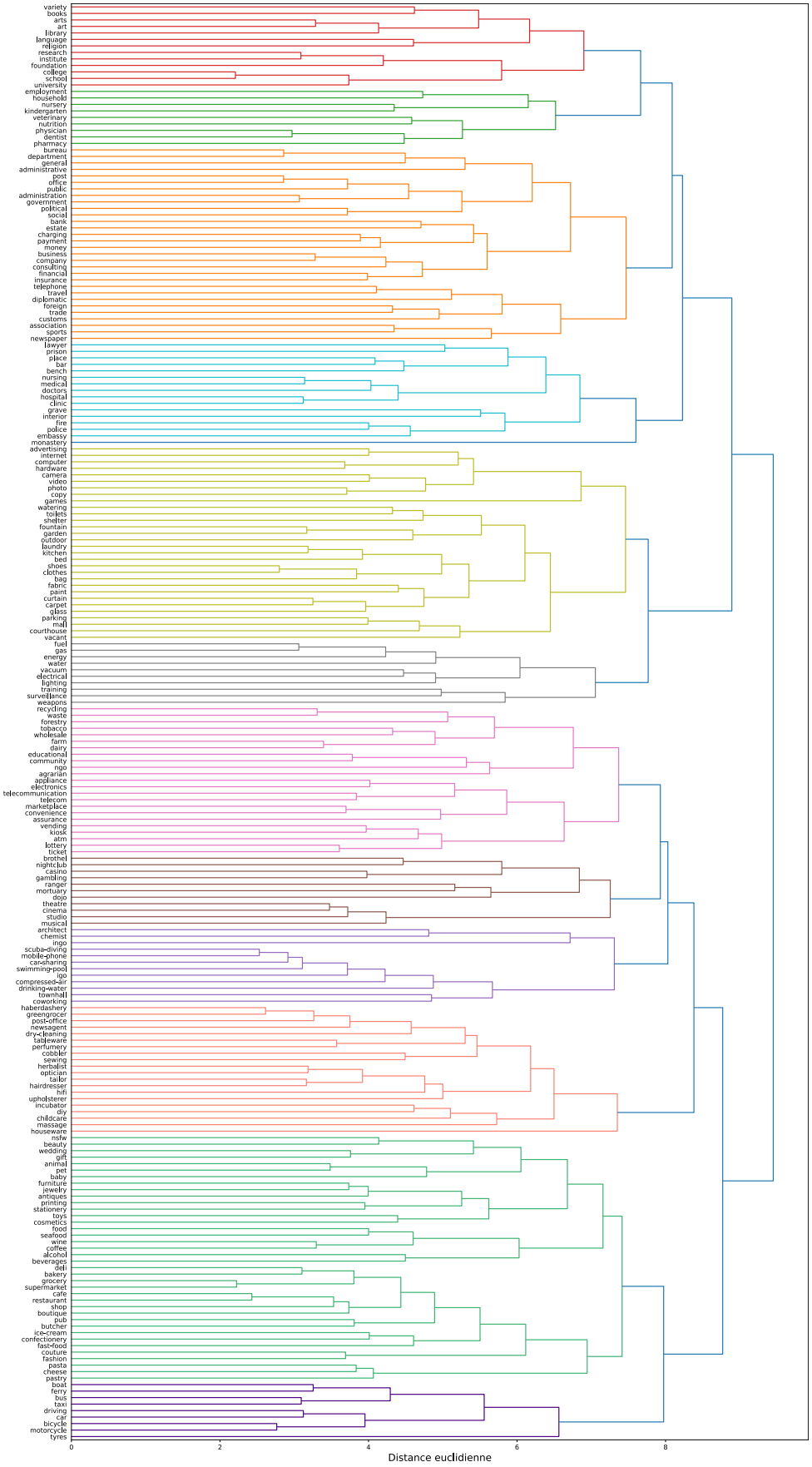


FIGURE 3.4 – Dendrogramme illustrant les résultats du regroupement hiérarchique des attributs d'OpenStreetMap.

3.4 Résultats et interprétation

Cette section présente du regroupement non supervisé et de la classification des stations de base. Dans un premier temps, on caractérise les classes obtenues par rapport aux heures de pointe. On présente également la similarité des classes d'activités entre la région de Dakar et la ville de Thiès. Dans un deuxième temps, les modèles sont entraînés sur un sous-ensemble de stations de bases de Dakar, et évalués sur le reste des stations de la région, et sur celles de Thiès. L'évaluation des modèles sur Thiès permet d'analyser les capacités de généralisation du modèle sur un territoire géographiquement éloigné mais d'activité mobile similaire.

3.4.1 Comparaison du regroupement des signatures de Dakar et de Thiès

La Figure 3.5 illustre les centres des classes obtenus en appliquant les k-moyennes sur les signatures des stations de base de la région de Dakar et de la ville de Thiès. On peut observer que les deux régions ont des centres de classe relativement similaires. La Table 3.5 montre également que la distribution des proportions de classe est proche. En fonction des heures de pointe, on caractérise les classes de la manière suivante :

— classe MP (*Morning Peak*) : pic d'utilisateurs survenant le matin
— classe NP (*Night Peak*) : pic survenant le soir
— classe T-P (*Two-Peaks*) : pic survenant le matin et le soir.

Heure de pointe Lieu	Matin (MP)	Soir (NP)	Mixte (T-P)
Région de Dakar	272 (57%)	78 (16%)	127 (27%)
Thiès	19 (53%)	5 (14%)	12 (33%)

TABLE 3.5 – Proportion des effectifs de classe obtenus avec les k-moyennes

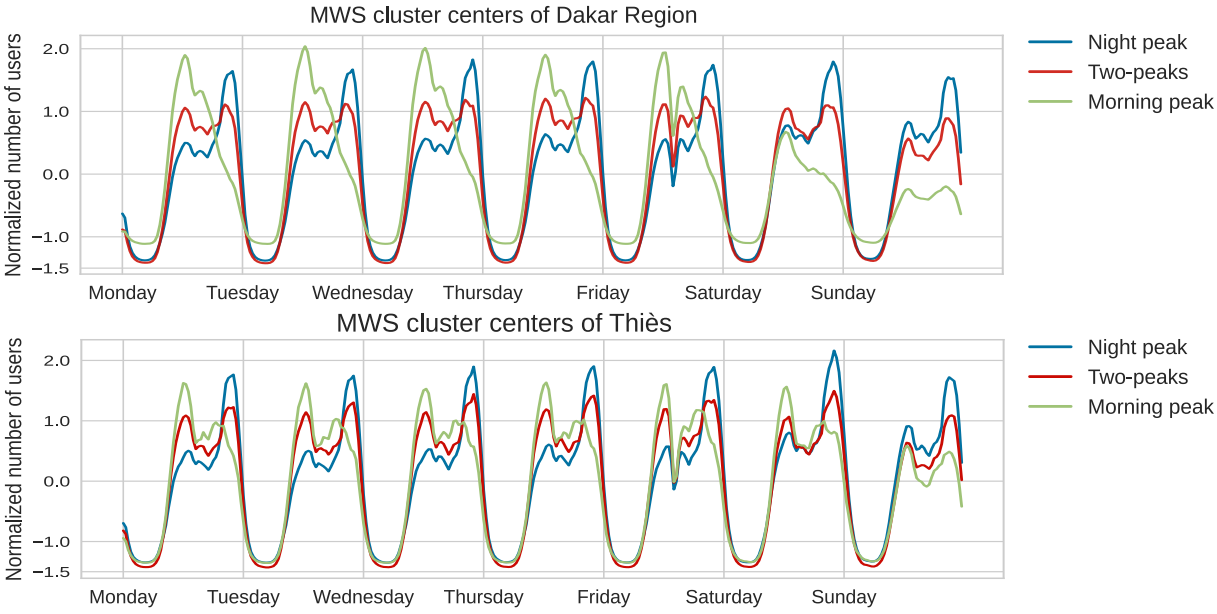
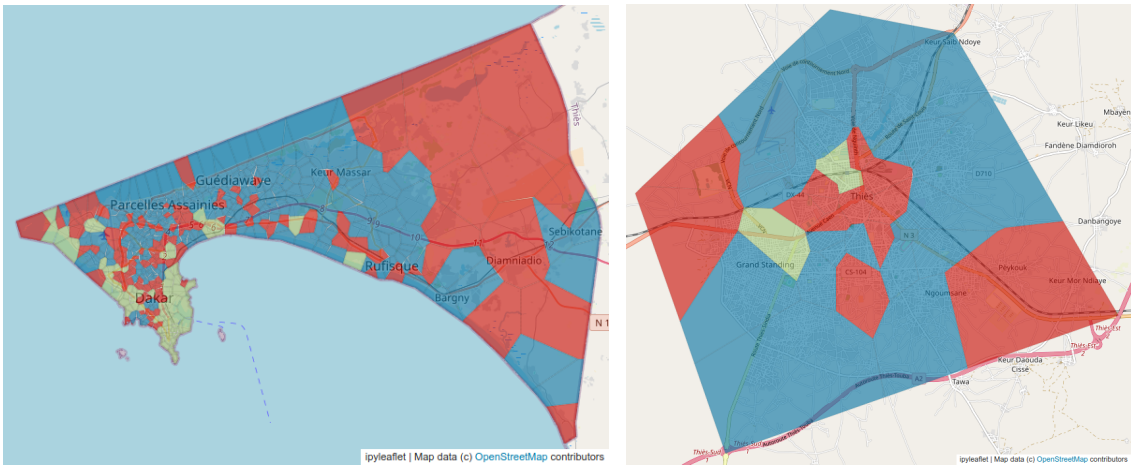


FIGURE 3.5 – Centres de classe obtenus avec les k-moyennes pour les stations de base de Dakar (haut) et de Thiès (bas). Chaque courbe correspond au nombre d'utilisateurs médianisé et normalisé en fonction du jour de semaine et de l'heure (le pas de temps du graphique est de 24h).

3.4.2 Classification sur la région de Dakar

En utilisant les résultats obtenus précédemment, plusieurs modèles sont entraînés à ranger les stations de base dans les classes MP, NP et T-P. Les performances d'un CatBoost, d'une forêt aléatoire et d'un



(a) Dakar Region (b) Thiès

FIGURE 3.6 – Diagramme de Voronoi des positions des stations de base colorisé en fonction de l'appartenance à l'une des trois classes. En bleu : NP, Vert : MP, Rouge : T-P.

SVC ont été évaluées en mesurant la précision, le rappel, le F-score des classes et la précision globale des prédictions. Pour chaque modèle, les entraînements ont été réalisés en répétant 100 fois la validation croisée à k blocs, pour $k = 10$. Par la suite, nous interprétons le meilleur modèle avec les valeurs SHAP pour donner une description du tissu urbain à chaque classe.

Performances

Les résultats renseignés dans le Tableau 3.6 montre que les trois modèles CatBoost, RF et SVC surpassent les lignes de base B1 et B2 de plus de 10% en précision globale. L'utilisation des variables géographiques contribue grandement à l'apprentissage des modèles. Le meilleur modèle est CatBoost, suivi de près par le SVC.

Classe	Mesure	B1	CatBoost B2	RF B2	SVC B2	CatBoost	RF	SVC
NP	Précision	0.57	0.58 ± 0.01	0.59 ± 0.01	0.57 ± 0.00	$\mathbf{0.73 \pm 0.01}$	0.71 ± 0.01	0.72 ± 0.01
	Rappel	**1.00**	0.91 ± 0.03	0.85 ± 0.01	0.98 ± 0.00	0.91 ± 0.01	0.92 ± 0.01	0.90 ± 0.01
	F-score	0.72	0.70 ± 0.01	0.69 ± 0.01	0.72 ± 0.00	$\mathbf{0.80 \pm 0.01}$	$\mathbf{0.80 \pm 0.01}$	$\mathbf{0.80 \pm 0.01}$
MP	Précision	0.00	0.13 ± 0.08	0.33 ± 0.08	0.00 ± 0.00	$\mathbf{0.73 \pm 0.04}$	0.72 ± 0.04	0.67 ± 0.04
	Rappel	0.00	0.04 ± 0.04	0.11 ± 0.03	0.00 ± 0.00	$\mathbf{0.56 \pm 0.03}$	0.50 ± 0.04	0.55 ± 0.03
	F-score	0.00	0.05 ± 0.05	0.15 ± 0.03	0.00 ± 0.00	$\mathbf{0.61 \pm 0.03}$	0.57 ± 0.03	0.59 ± 0.03
T-P	Précision	0.00	0.25 ± 0.09	0.35 ± 0.05	0.03 ± 0.03	0.46 ± 0.04	0.42 ± 0.05	$\mathbf{0.47 \pm 0.04}$
	Rappel	0.00	0.10 ± 0.04	0.17 ± 0.02	0.01 ± 0.01	$\mathbf{0.28 \pm 0.03}$	0.23 ± 0.03	0.26 ± 0.02
	F-score	0.00	0.13 ± 0.05	0.22 ± 0.03	0.01 ± 0.01	$\mathbf{0.33 \pm 0.03}$	0.28 ± 0.03	0.32 ± 0.03
Toutes classes	Précision globale	0.57	0.55 ± 0.01	0.55 ± 0.01	0.56 ± 0.00	$\mathbf{0.69 \pm 0.01}$	0.67 ± 0.01	0.68 ± 0.01

TABLE 3.6 – Moyenne et écart-type des mesures de classification des stations de base de la région de Dakar. Le F-score est moyenné sur la répétition de la validation croisée à k blocs.

Interprétation du modèle à l'aide des valeurs SHAP

Les valeurs SHAP [107, 108] sont utilisées pour interpréter CatBoost, le meilleur modèle de notre étude. On utilise le graphe *beeswarm* (Figure 3.7) pour analyser les valeurs importantes.

Sur un graphe de type beeswarm (littéralement : essaim d'abeille), les variables sont ordonnées par importance décroissante. On dit d'une variable qu'elle est importante si elle contribue grandement à la précision des prédictions du modèle. Pour chaque variable, un point correspond à la valeur de cette variable pour une station de base dans le jeu d'entraînement. La couleur du point illustre la valeur de la variable (élevée ou faible). L'axe horizontal donne la valeur SHAP d'une variable. Si la variable d'une observation

possède une valeur SHAP positive (respectivement négative), cela signifie qu'elle contribue positivement (respectivement négativement) à influencer la prédiction pour qu'elle soit de la classe considérée. Le graphe est composé des données de toutes les stations de base. Cela permet d'extraire les tendances globales des variables par rapport aux valeurs SHAP et de savoir si ces variables sont corrélées positivement ou négativement avec la prédiction d'une classe.

Exemple Sur la Figure 3.7 a, la variable la plus importante contribuant à prédire la classe nocturne NP est le nombre de sites religieux (*religion*) ; on observe une corrélation positive entre les valeurs élevées de cette variable et les valeurs SHAP. Cela signifie que les stations de base appartenant à la classe NP tendent à couvrir les régions contenant plus de lieux de religieux que d'autres classes.

Interprétation des pics d'activité On observe des corrélations opposées entre les classes NP et MP pour les trois premières variables les plus importantes. Les variables importantes pour prédire la classe T-P sont différentes.

- La classe MP est corrélée positivement avec les infrastructures en lien avec le travail dans les quartiers d'affaires : usage du sol industriel, équipements gouvernementaux, bureaux, universités, banques. Spécifique à la région de Dakar, on trouve un grand nombre de piscines dans les zones touristiques. L'activité touristique est suffisamment significative pour qu'elle soit fortement corrélée et exploitée par CatBoost.
- La classe NP est positivement corrélée avec la densité de population, les magasins, les écoles primaires et les sites religieux. Il s'agit de données fréquemment situées à l'intérieur ou à proximité des zones résidentielles.
- La classe T-P a le plus grand nombre de corrélations positives. Elle est décrite à la fois par les corrélations positives de la classe MP et de la classe NP, autrement dit, par des variables de travail et résidentielles : bureaux gouvernementaux, universités, écoles primaires, sites religieux. Les stations de base de cette classe couvrent des zones résidentielles moins densément peuplées que les autres, faiblement industrialisées, mais équipées de nombreux services tertiaires.

Par conséquent, le jeu de données contient des informations sur le tissu urbain apprenables par les modèles pour classer l'activité du réseau mobile. Les outils d'interprétation sont utiles pour comprendre quelles variables sont décisives pour la bonne prédiction du modèle étudié, et indirectement, quels éléments du tissu urbain sont corrélés avec les heures de pointe du réseau.

3.4.3 Classification sur la ville de Thiès

Thiès est une ville dont les classes d'activité mobile sont similaires à celles de Dakar, en proportions et en pics d'activité. Cette observation nous a mené à nous questionner sur les performances de généralisation des modèles. Est-il possible d'entraîner un modèle sur les données d'une région et de classifier correctement l'activité des stations de base d'une autre ville, sur la base de son tissu urbain ?

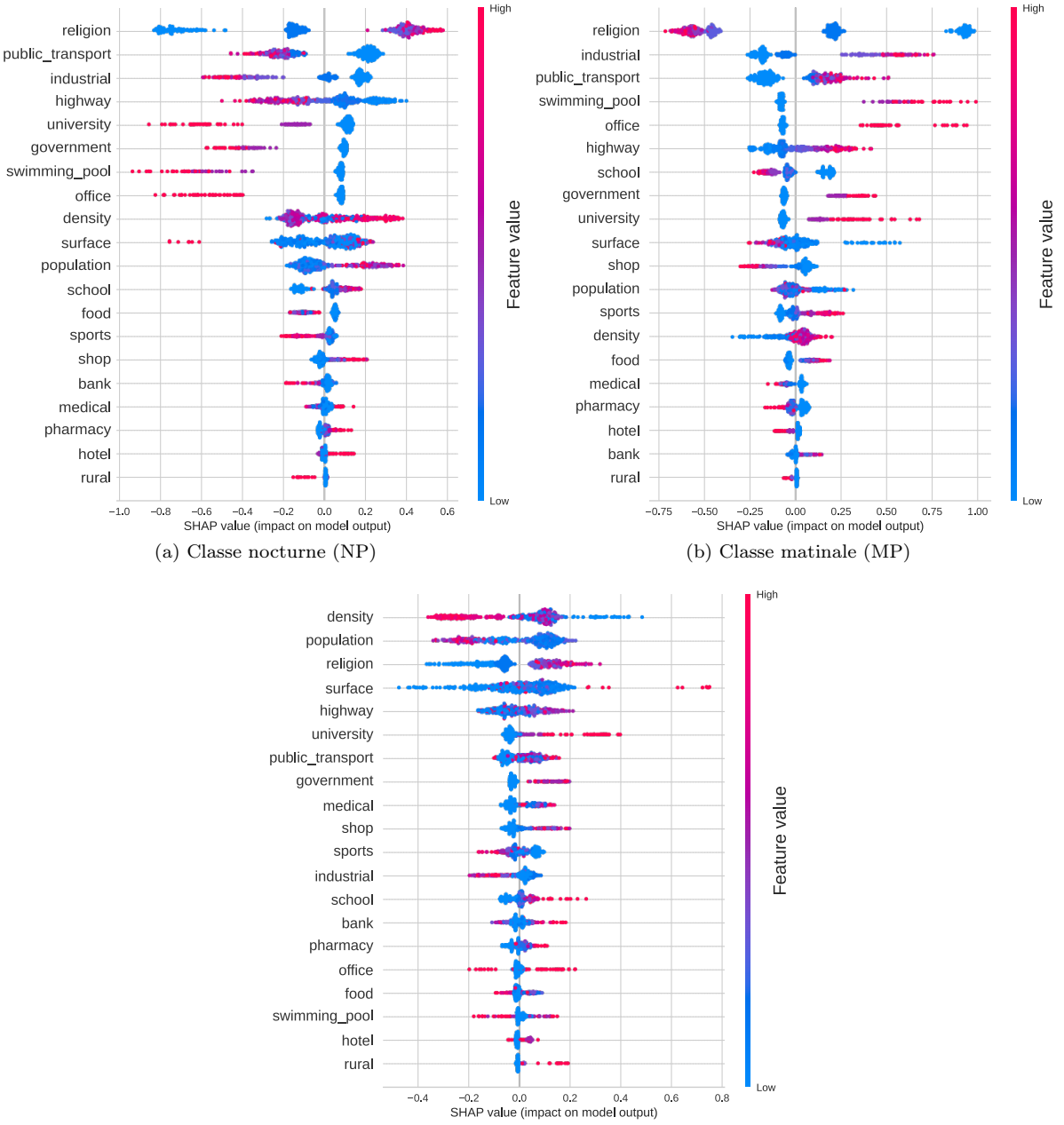
Performances

Bien que la précision globale des modèles d'apprentissage ne soit pas plus élevée que les lignes de base, ils achèvent tout de même de meilleures précision, rappel et F-score que le modèle B1 pour les classes MP et T-P (Tableau 3.7). Ces modèles restent un minimum informatifs et spécialisés, moyennant des usages similaires du réseau mobile.

Dans ce contexte, nous observons que le meilleur modèle est le SVC. Il est possible que le modèle CatBoost ait surappris sur les données de Dakar.

Expliquer les erreurs de classification avec SHAP

Pour comprendre pourquoi la précision des modèles s'est dégradée sur les données de Thiès, on analyse deux exemples d'erreur de classification de CatBoost. La Figure 3.8 montre les valeurs SHAP de deux stations de base de la classe MP qui ont été rangées dans la classe NP par le modèle. Sur l'image, « Class 0 » correspond à la classe T-P, « Class 1 » à la classe MP et « Class 2 » à la classe NP. Chaque graphe décrit de manière synthétique comment les variables d'une station de base sont traitées par le modèle, et si elles contribuent positivement ou négativement au fait que la prédiction soit d'une certaine classe ou non. Chaque graphe possède trois courbes, car au-delà de deux classes, les modèles d'apprentissage calculent un niveau de certitude (ou de probabilité) que la station de base appartienne à chacune des classes. La classe sélectionnée est celle pour laquelle la probabilité est la plus élevée. Sur les images, il



(a) Classe nocturne (NP)

(b) Classe matinale (MP)

(c) Classe mixte (T-P)

FIGURE 3.7 – Graphe beeswarm des valeurs SHAP des sites de Dakar

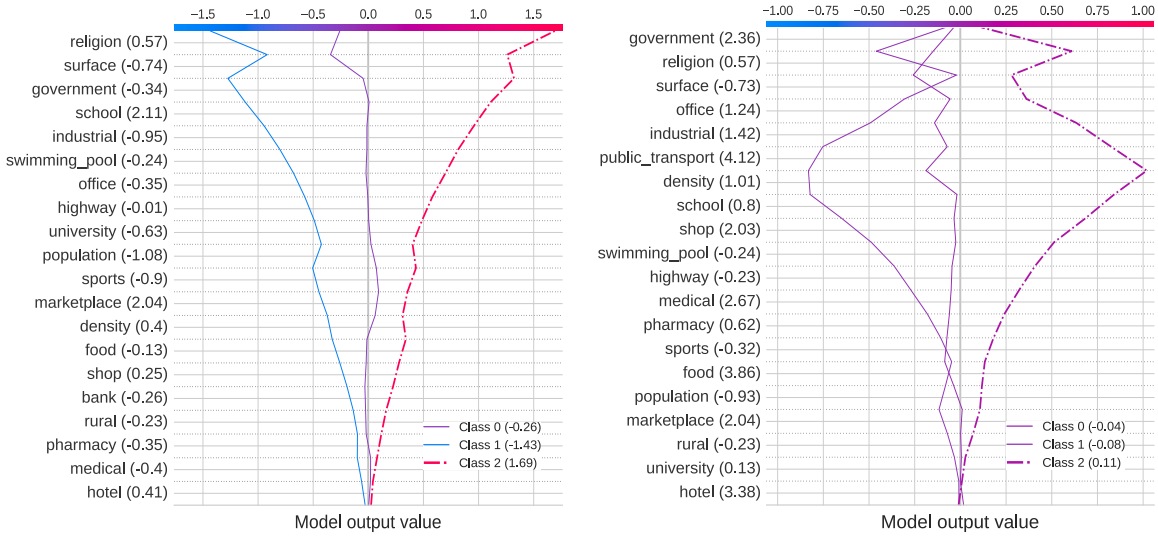
s'agit de celle dont la ligne est en pointillés. L'axe horizontal est la sortie du modèle avant normalisation (*softmax*), c'est pourquoi les valeurs ne se situent pas entre 0 et 1 comme on l'attendrait d'une probabilité. Le graphe se lit de bas en haut, par ordre décroissant d'importance des variables ; la valeur du bas où se rejoignent les trois classes est la probabilité attendue si les modèles ne se servent pas des variables pour faire la prédiction. À titre informatif, les valeurs des variables normalisées sont affichées à côté de leur nom.

BS 556 (Figure 3.8 a)

Dans cet exemple, la classe NP est prédite avec une probabilité nettement supérieure à celle de la classe MP. Le modèle est certain de sa mauvaise classification. Si l'on regarde la courbe de décision de MP (en bleu), on voit que seules les variables de l'aire de service de la station (*surface*) et de la population augmentent la probabilité de prédire la bonne classe, mais toutes les autres variables la diminuent. On peut supposer certaines variables caractéristiques d'une classe à Dakar ne le sont pas autant pour la classe

Classe	Mesure	B1	CatBoost	RF	SVC
NP	Précision	0.53	0.62 ± 0.01	0.59 ± 0.01	$\mathbf{0.66 \pm 0.01}$
	Rappel	**1.00**	0.81 ± 0.02	0.89 ± 0.01	0.79 ± 0.02
	F-score	0.69	0.70 ± 0.01	0.71 ± 0.01	$\mathbf{0.72 \pm 0.01}$
MP	Précision	0.00	$\mathbf{0.29 \pm 0.03}$	0.27 ± 0.03	0.16 ± 0.01
	Rappel	0.00	$\mathbf{0.42 \pm 0.04}$	0.22 ± 0.03	0.20 ± 0.00
	F-score	0.00	$\mathbf{0.34 \pm 0.03}$	0.24 ± 0.03	0.17 ± 0.00
T-P	Précision	0.00	0.42 ± 0.06	0.52 ± 0.06	$\mathbf{0.58 \pm 0.02}$
	Rappel	0.00	0.14 ± 0.02	0.16 ± 0.03	$\mathbf{0.33 \pm 0.01}$
	F-score	0.00	0.21 ± 0.03	0.24 ± 0.04	$\mathbf{0.42 \pm 0.01}$
All	Glob. Acc.	0.53	0.54 ± 0.01	$\mathbf{0.55 \pm 0.01}$	$\mathbf{0.55 \pm 0.01}$

TABLE 3.7 – Moyenne et écart-type des mesures de classification des stations de base de Thiès.



(a) BS 556 (b) BS 557

FIGURE 3.8 – Graphe de décision pour les valeurs SHAP de deux stations de base mal classifiées. Class 0 : T-P, Class 1 : MP, Class 2 : NP

correspondante à Thiès.

BS 557 (Figure 3.8 b)

Dans ce cas de figure, la mauvaise décision du modèle est moins certaine : si l'on regarde les graduations en haut du graphe, la différence de probabilité entre la classe MP et NP est de 0.19 seulement (contre 3.12 dans l'exemple précédent).

Pour les classes MP et NP, les variables ne contribuent pas toutes dans le même sens de prédiction. Dans le cas précédent, les courbes de MP et de NP étaient presque strictement croissantes ou décroissantes. Ici, les courbes de décision oscillent fortement. L'abondance de magasins, la forte densité de population, le nombre d'écoles et de bâtiments liés au secteur santé poussent le modèle à prédire la classe NP car elles sont des caractéristiques du tissu résidentiel de Dakar.

On peut conclure que malgré des profils similaires d'affluence du réseau mobile, la structure urbaine des deux zones est assez différentes. Comme Thiès est une zone urbaine plus petite, on peut supposer que différents services se retrouvent groupés dans le même quartier, là où ils seraient répartis dans plusieurs quartiers de fonctions différentes à Dakar.

3.5 Conclusion

Dans cette étude, le regroupement non supervisé des données d'affluence des stations de base a mis en évidence trois classes d'activités. On a montré qu'il était possible pour les modèles de prédire la classe

d'activité d'une station de base uniquement en fonction des données géographique et de distribution de population. Ces modèles réalisent de meilleures performances que les lignes de base. Par ailleurs, l'interprétation des modèles permet de qualifier les classes en fonction des activités humaines, et de recouvrer les classes résidentielles, industrielles et mixtes rencontrées dans la littérature. Ces résultats sont également intéressants quant à l'usage de données géographiques publiques, malgré des imprécisions possibles dans les régions manquant de contributeurs. Les campagnes massives de cartographie réalisées régulièrement pour le projet OpenStreetMap sont précieuses pour renforcer la fiabilité de cette source.

Ces premiers résultats concluants laissent penser qu'il est possible de prédire des informations plus précises sur l'activité d'une station de base future. Les deux chapitres suivants seront consacrés à la prédiction directe de la signature hebdomadaire médiane. Cette signature étant de grande dimension (168 valeurs), les temps d'entraînement des modèles peuvent être longs. Une solution à ce problème est l'utilisation d'un auto-encodeur pour entraîner les modèles à prédire des représentations discrètes des données mobiles, qui seront ensuite reconstituées par l'auto-encodeur. Dans le Chapitre 4, on étudiera différentes techniques de réduction de dimension pour aboutir à l'entraînement d'un auto-encodeur variationnel utilisé pour la prédiction des signatures du Chapitre 5.

Chapitre 4

Réduction de dimension

4.1 Introduction

Les travaux du Chapitre 3 aboutissent à des méthodes permettant de prédire une information simple sur le profil de l'affluence d'une future station de base. Pour pouvoir prédire directement la signature hebdomadaire médiane, une des méthodes considérées repose sur l'utilisation d'un auto-encodeur convolutif variationnel discret (AECVD) pour réduire la dimension des données mobiles. Deux raisons ont motivé cette méthode : la première est d'adapter le transformeur DALL·E 1 d'OpenAI utilisé pour la génération d'images à partir de texte [109] à la génération de signatures à partir de données tabulaires. Ce transformeur est entraîné sur les données réduites dans l'espace latent d'un auto-encodeur de ce genre. La deuxième motivation vient de l'observation que la taille des signatures (168 valeurs) peut induire un temps d'entraînement assez long chez certains modèles d'apprentissage. Pour accélérer l'entraînement, une solution est de le réaliser dans une dimension plus faible.

L'AECVD est un modèle dont l'architecture et les valeurs d'hyperparamètres recommandées par la littérature sont adaptées à la génération d'images en haute résolution. Les données mobiles sur lesquelles on travaille sont de dimension moindre et nécessitent d'adapter le modèle pour que la compression et la reconstitution des données soient adéquates. Plus précisément, il faudra veiller à ce que les caractéristiques de l'activité des réseaux mobiles soient conservées. Par exemple, les pics d'activités et la diversité des périodicités (industrielle, résidentielle, transports...) doivent pouvoir êtres générés par le décodeur. Or le choix des hyperparamètres et la conception de l'architecture d'un réseau de neurones ne sont pas des tâches triviales. Pour comprendre empiriquement l'influence de chaque hyperparamètre sur la qualité de la génération de données, il est important d'analyser visuellement la distribution des données réduites dans l'espace latent. Par ailleurs, l'étude du fonctionnement des générations précédentes d'auto-encodeurs peut faciliter cette compréhension, car ils introduisent progressivement les concepts sur lesquels reposent l'AECVD.

Ce chapitre est structuré comme suit :

- La Section 4.2 présente les données mobiles utilisées pour la réduction et la génération de données.
- La Section 4.3 étudie les similarités de l'espace réduit de l'analyse en composantes principales avec l'espace latent du réseau de neurones le plus simple pour ce travail : l'auto-encodeur linéaire.
- La Section 4.4 présente l'influence de la non-linéarité apportée par l'auto-encodeur convolutif sur la réduction de données.
- La Section 4.5 illustre ensuite les apports du modèle variationnel pour améliorer la génération des données.
- La Section 4.6 présente enfin les variations architecturales permettant de représenter les données dans un espace discret, adapté à l'utilisation du transformeur.
- La Section 4.7 conclut le travail.

4.2 Données utilisées

Les données mobiles utilisées ont été collectées sur 3 445 stations de bases 4G situées en Île-de-France, sur une période allant du 2 novembre 2020 au 30 août 2021. L'indicateur de performance étudié est le nombre d'utilisateurs. Il est représenté par une série temporelle pour chaque station de base. Pour créer les observations des données d'entraînement, chaque série est découpée sous-ensembles décrivant le nombre d'utilisateurs par semaine.

Nombre d'observations La période de collecte des données s'étendant sur 43 semaines, le nombre maximal d'observations utilisables est :

$$N_{\max} = \text{nombre de stations} \times \text{nombre de semaines} = 3\ 445 \times 43 = 148\ 135$$

En pratique, toutes les données ne sont pas exploitables car elles peuvent être indisponibles en raison de pannes du réseau, ou incomplètes pour un site déployé pendant la période de collecte. Après traitement des données, il reste $N = 143\ 805$ observations. Les valeurs du nombre d'utilisateurs sont normalisées dans l'intervalle $[0, 1]$.

Représentation vectorielle Par la suite, on note $\{X_i\}_{i=0}^{N}$ l'ensemble des observations de taille $N = 143\ 805$. Pour une observation i, le vecteur de variables $X_i \in \mathbb{R}^p$ contient p variables, chacune représentant le nombre d'utilisateurs connectés à la même cellule et à une date donnée. Les données sont collectées toutes les heures pendant une semaine, donc :

$$p = \text{nombre de jours de la semaine} \times \text{nombres d'heures dans une journée} = 7 \times 24 = 168$$

Représentation matricielle Pour les modèles employant des couches convolutives, on transforme le vecteur X_i en une matrice de taille $(7, 24, 1)$. Chaque ligne représente l'affluence d'une cellule sur une journée, et chaque colonne représente l'affluence d'une même heure sur tous les jours d'une semaine. La troisième dimension de taille 1 permet d'arranger les variables de manière à passer une image en niveau de gris aux modèles.

Exemples La Figure 4.1 donne la représentation vectorielle (graphes) et matricielle (images) d'un échantillon de neuf observations :

- Sur la représentation vectorielle, les pics d'activité correspondent aux valeurs proche de 1. Les variables d'heures successives sont fortement dépendantes.
- Sur la représentation matricielle, les pics d'activité correspondent aux zones en jaune. La dépendance horaire est toujours apparente suivant l'axe horizontal de l'image. Cette représentation en deux dimensions permet de rapprocher les périodicités journalières le long de l'axe vertical.

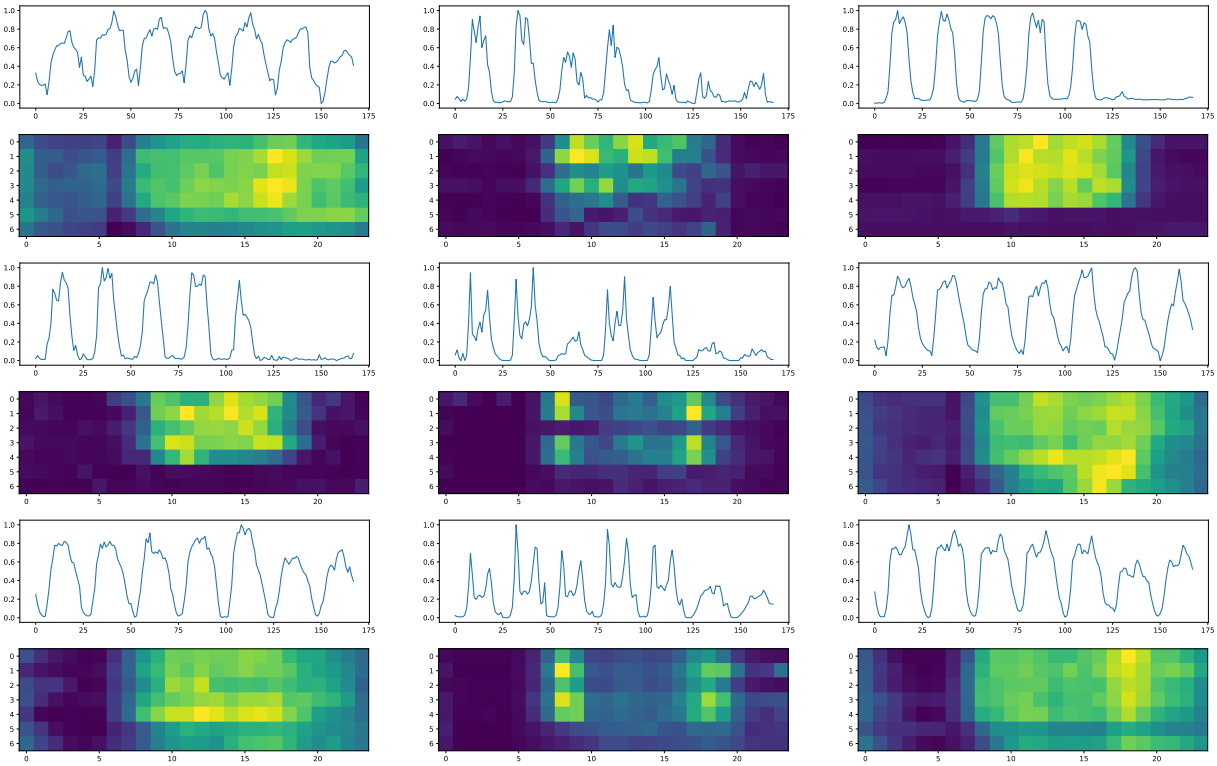


FIGURE 4.1 – Échantillon de 9 observations, représentées en 1D (graphes) et en 2D (images colorisées avec la palette verdigris)

4.3 Relation entre l'AE linéaire et l'ACP

Analyse en composantes principales

L'analyse en composantes principales (ACP) est une technique de réduction de dimension [110]. Elle transforme des variables potentiellement corrélées en variables orthogonales grâce à un changement de la base dans laquelle elles sont projetées. Les composantes principales sont les vecteurs de cette nouvelle base. Elles sont ordonnées de manière à expliquer la variance des données de façon décroissante. Après la réduction de dimension, le nombre de variables est égal au nombre de composantes principales que l'on choisit de garder. Un critère de détermination peut être la proportion de variance expliquée que l'on souhaite conserver. La réduction en deux dimensions est une pratique courante car elle permet de visualiser la distribution des données réduites avec un nuage de points.

Visualisation des données dans l'espace de l'ACP en dimension 2 Pour étudier comment les observations sont distribuées dans la dimension réduite de l'ACP, les données initiales sont préalablement étiquetées dans 6 classes avec la méthode des k-moyennes. Les centres de classe sont illustrés sur la Figure 4.2. Chaque point correspond à une observation (le trafic d'une cellule du réseau mobile) réduite à deux variables. Les points sont colorisés par appartenance à une classe : les membres d'un même groupe partagent un profil d'affluence similaire au représentant de classe. On peut constater une certaine continuité de la distribution des points. En effet, les centres de classe similaires sont proches dans l'espace, et l'évolution d'une classe à une autre semble plutôt régulière. On observe par exemple que les observations de la classe A et B, de représentants identiques mais d'amplitudes légèrement différentes, sont contigues et représentent le trafic des cellules résidentielles. Les classes C, E et F sont également proches : les pics d'activités du lundi au vendredi sont similaires, mais diffèrent le week-end. La classe C correspond à des cellules plutôt implantées dans des centres commerciaux, la E dans des bureaux et la classe F est d'activité mixte. La classe D est située entre la C et la E ; les deux pics journaliers sont plus exacerbés que les deux classes, mais l'allure de la courbe le week-end est plus proche de celui de la classe E. Elle représente donc les cellules desservant des utilisateurs dans des zones de mobilités (transports en communs, grands axes routiers par exemple).

Auto-encodeurs

L'ACP est une technique classique de la réduction de dimension pour l'analyse des données. Pour le traitement de données massives et complexes, la littérature tend à privilégier l'usage des auto-encodeurs, réputés très performants mais surtout, fournissant de nombreuses applications pour les données réduites.

Les auto-encodeurs AE sont des réseaux de neurones possédant une structure dite « en sablier » car les couches proches de l'extérieur sont plus grandes que celles à l'intérieur (Figure 4.3). Le milieu constitue un goulot d'étranglement où on récupère les représentations réduites (ou latentes) des vecteurs d'entrée. Celles-ci sont représentées dans ce qu'on appelle « l'espace latent » du modèle. Les auto-encodeurs sont entraînés pour approximer la fonction identité, c'est-à-dire :

$$f(x) = x,\ x \in \mathbb{R}^n, n \in \mathbb{N}$$

Encodeur/décodeur L'auto-encodeur se décompose en deux ensembles qui se rejoignent au niveau de la couche du vecteur latent : l'encodeur et le décodeur. L'encodeur est chargé de compresser les données pour ne conserver que l'information essentielle, tandis que le décodeur est chargé de décompresser la donnée latente pour reconstituer l'entrée initiale.

Applications Grâce à leurs propriétés, les auto-encodeurs sont utilisés pour des tâches de compression, de débruitage et de génération de données. C'est ce dernier aspect que nous étudierons à travers l'analyse des espaces latents des modèles.

Auto-encodeur linéaire

La version la plus simple de l'auto-encodeur est l'auto-encodeur linéaire. Ce modèle est linéaire car l'information passée en entrée du modèle est propagée d'une couche à une autre sans être modifiée par une fonction d'activation non linéaire.

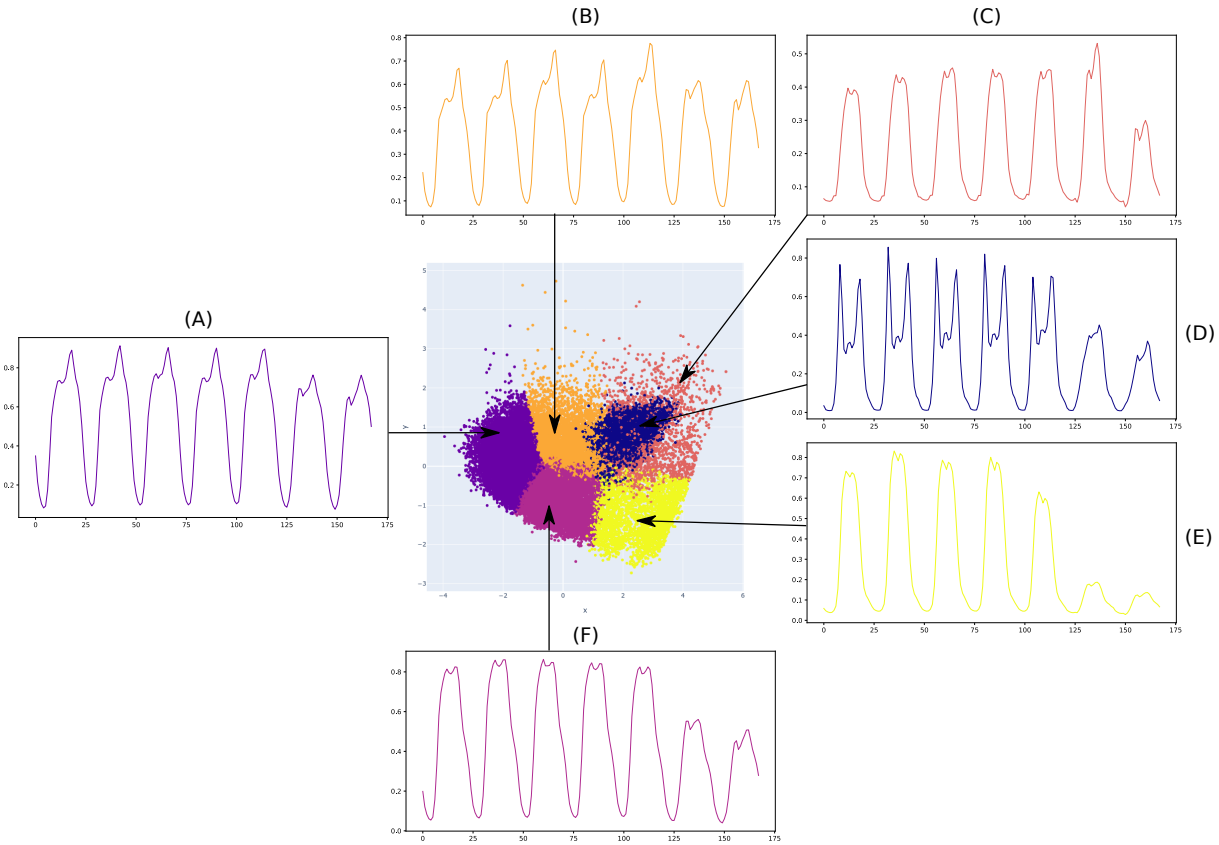


FIGURE 4.2 – Distribution des données mobiles dans l'espace réduit de l'ACP à deux composantes principales. Les points sont colorisés par groupes d'activité similaire. Les graphes correspondent aux centres de classes dans l'espace initial des données hebdomadaires. L'abscisse correspond à la date en heure et l'ordonnée au nombre normalisé d'utilisateurs.

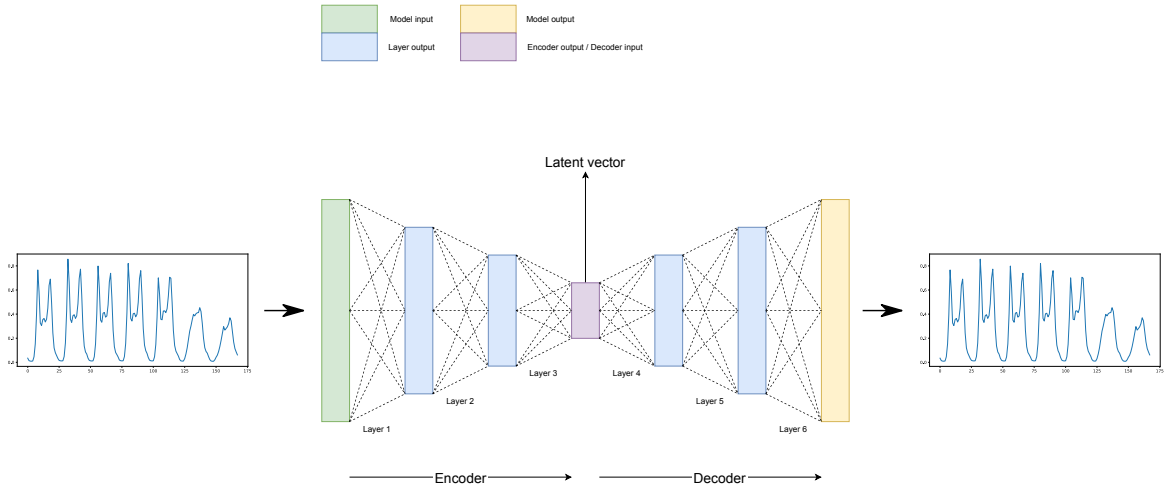


FIGURE 4.3 – Illustration générique d'un auto-encodeur

Réseaux de neurones à propagation en avant On décrit ici formellement l'architecture d'un réseau de neurones à propagation en avant, qui est le mode de propagation de tous les réseaux de neurones considérés dans cette étude.

Soit $X = (X_0, ..., X_{n-1}) \in \mathcal{M}_{n,p}$ un jeu de données stocké dans une matrice composée de n observations et de p variables. Pour $0 \leq i \leq n-1$, on note $X_i \in \mathbb{R}^p$ le vecteur des variables de l'observation d'indice i. Pour un entraînement supervisé, les cibles associées aux observations sont rassemblées dans la matrice $y = (y_0, ..., y_n) \in \mathcal{M}_{n,q}$, où $y_i \in \mathbb{R}^q$.

Soit $f : \mathbb{R}^p \longrightarrow \mathbb{R}^q$ la fonction décrivant un réseau de neurones artificiels composé de $L > 0$ couches.

On modélise une couche l de m_l neurones par un vecteur $X_i^{(l)} \in \mathbb{R}^{m_l}$. En suivant cette notation, la première couche correspond aux vecteurs d'entrée ($X_i^{(0)} = X_i$) et la dernière couche au vecteur de sortie ($X_i^{(L-1)} = \hat{y}_i = f(X_i)$).

Entre deux couches $X_i^{(l)}$ et $X_i^{(l+1)}$ composés de m_l et m_{l+1} neurones respectivement, les connexions entre les neurones sont représentées par une matrice de poids $W_i \in \mathcal{M}_{m_{l+1},m_l}$, et optionnellement par un vecteur de biais et une fonction d'activation. La matrice W_i connecte chaque neurone de $X_i^{(l)}$ à tous les neurones de $X_i^{(l+1)}$. On parle de couches complètement connectées (*fully connected*) car chaque neurone suivant est calculé à partir de tous les neurones précédents.

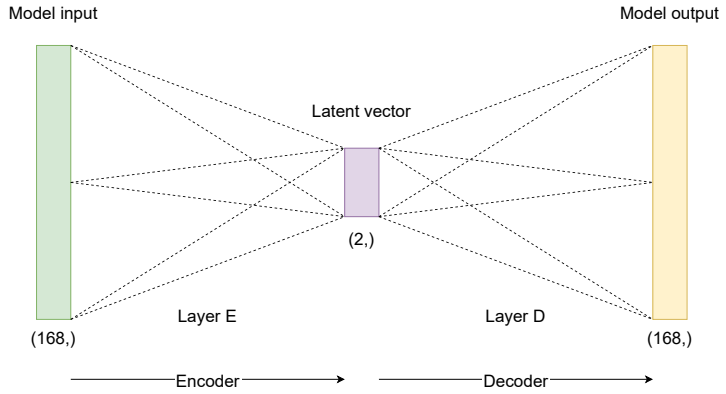


FIGURE 4.4 – Architecture d'un auto-encodeur linéaire

Architecture L'auto-encodeur linéaire (Figure 4.4) est composé d'une seule couche cachée $x_l \in \mathbb{R}^r$. Elle résulte de la multiplication de l'entrée par une matrice d'encodage $E \in \mathcal{M}_{r,p}(\mathbb{R})$:

$$x_l = Ex$$

La couche de sortie y résulte de la multiplication de la couche cachée par une matrice de décodage, notée $D \in \mathcal{M}_{p,r}(\mathbb{R})$:

$$y = Dx_l$$

Les auto-encodeurs sont entraînés à minimiser l'erreur de reconstruction. La mesure utilisée est l'erreur quadratique moyenne :

$$\text{Loss}(x, X_{\text{rec}}) = \text{MSE}(x, x_{\text{rec}}) = \frac{1}{n}\sum_i (x_i - x_{i_{rec}})^2$$

Pour un modèle parfaitement entraîné, la sortie attendue est égale à l'entrée, c'est-à-dire :

$$y = Dx_l = DEx$$

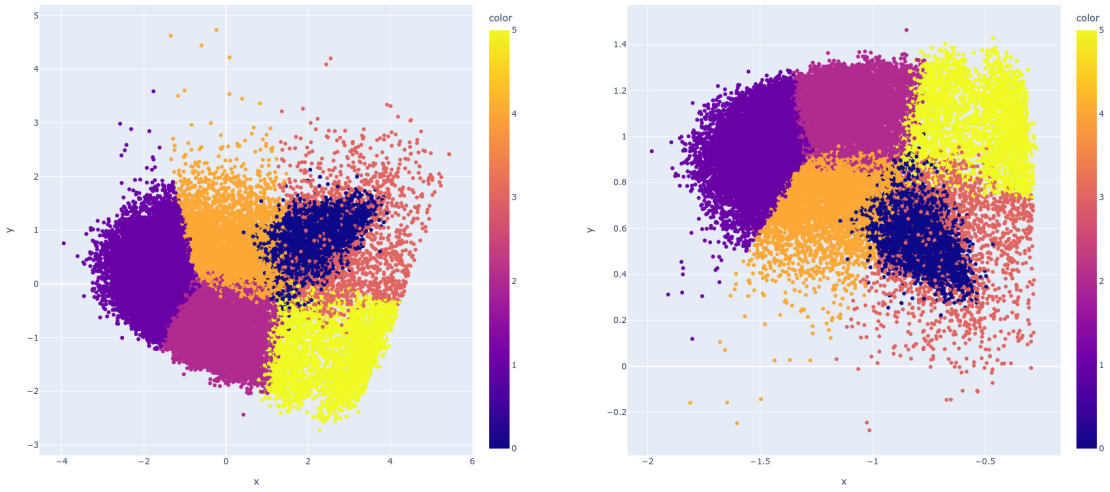
Par identification, le produit DE est égal à la matrice identité de taille $p = 168$ (taille d'une observation).

Analyse de l'espace latent

La Figure 4.3 compare la distribution des données mobiles dans l'espace de l'ACP avec celle dans l'espace latent de l'auto-encodeur linéaire. Les groupes à l'intérieur des nuages de points sont distribués de la même manière, à une rotation et une échelle près. Cette observation illustre le fait que l'espace latent d'un auto-encodeur linéaire est semblable à l'espace généré par l'ACP [111].

4.4 Auto-encodeur convolutif

En pratique, l'efficacité des auto-encodeurs réside dans leur non linéarité grâce aux fonctions d'activation. En reprenant les notations précédentes, la relation entre deux couches $X_i^{(l)}$ et $X_i^{(l+1)}$ entièrement connectées par une matrice W_i et une fonction d'activation a_i est :



(a) ACP

(b) Auto-encodeur linéaire

FIGURE 4.5 – Comparaison des distributions des observations réduites par l'ACP (à gauche) et par l'auto-encodeur linéaire (à droite). L'espace latent de l'auto-encodeur est de dimension $r = 2$.

$$X_i^{(l+1)} = a_i(W_i X_i^{(l)})$$

Parmi les fonctions d'activation couramment utilisées, on peut citer la sigmoïde, la tangente hyperbolique (tanh), l'unité linéaire rectifiée (ReLU : *Rectified Linear Unit* [112]) et la LeakyReLU [113].

L'auto-encodeur convolutif est un modèle non linéaire qui comporte des couches entièrement connectées, mais aussi des couches connectées par des filtres convolutifs [48].

Les filtres convolutifs font intervenir de nombreux paramètres qui sont difficilement interprétables. On suppose cependant que leur efficacité provient du traitement de l'information couche par couche. Les éléments complexes d'une image sont décomposés en des formes simples dans les premières couches (des morceaux de courbes, des traits, des contrastes...) qui sont rassemblées pour former des représentations plus complexes (des contours, des fragments d'images) à mesure que l'on s'enfonce dans les couches.

Une connection convolutive est composée de plusieurs matrices de convolutions (aussi appelées des filtres ou des *kernels*). Ces matrices sont généralement carrées, de dimensions impaires. Soit $k \in \mathcal{M}_{2n+1,2n+1}(\mathbb{R})$ une matrice de convolution. Pour l'expression de la convolution qui suit, on indexera les éléments de la matrice entre $-n$ et n. Ainsi :

— l'élément en première ligne, première colonne est indexé $(-n, -n)$,
— l'élément au centre de la matrice est indexé $(0, 0)$
— l'élément en dernière ligne, dernière colonne est indexé (n, n).

Formule de convolution Pour l'entraînement de ces modèles, on utilise la représentation matricielle des données X. On note $X_i^{(a,b)}$ un élément de la matrice X_i situé aux coordonnées (a, b). L'élément $X_{i+1}^{(a,b)}$, situé en même position à la couche suivante est calculé à partir d'un filtre convolutif k suivant la formule suivante [114] :

$$X_{i+1}^{(a,b)} = (X_i * k)^{(a,b)} = \sum_{l=-n}^{n} \sum_{m=-n}^{n} k^{(l,m)} X_i^{(a-l,b-m)}$$

Contrairement aux couches entièrement connectées, seuls les éléments localement proches de la position (a, b) à la couche précédente rentrent dans le calcul de la valeur de l'élément de même position à la couche suivante. De nombreuses variantes (*padding*, *stride*...) peuvent être appliquées aux filtres de convolution pour modifier la taille des données à travers les couches.

Architecture La Figure 4.6 illustre les dimensions des couches, colorisés en fonction du type de connection entre celles-ci. Les dimensions et paramètres d'utilisation des filtres sont choisis de manière à réduire la hauteur et la largeur des couches, tout en augmentant leur profondeur. La fonction d'activation choisie pour les couches est de type LeakyReLU :

$$a(x) = \begin{cases} 0.01x & x < 0 \\ x & x \geq 0 \end{cases}$$

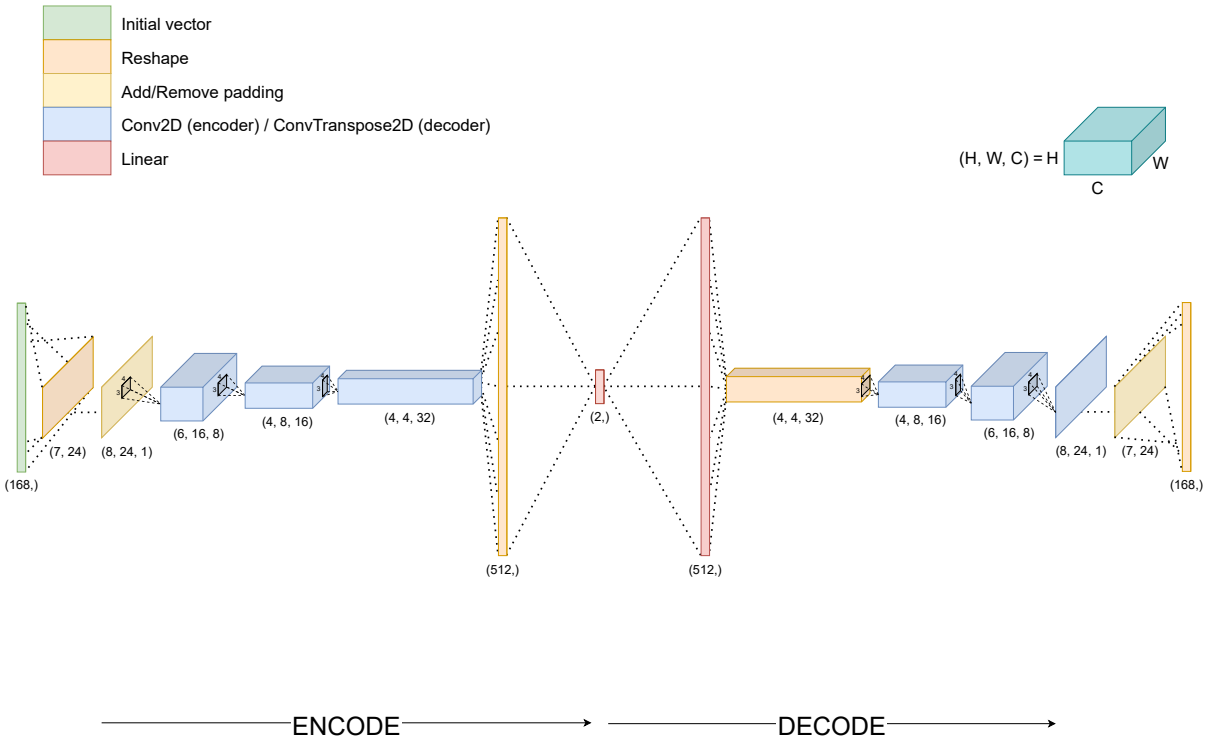


FIGURE 4.6 – Architecture de l'auto-encodeur convolutif utilisé pour l'étude

Analyse de l'espace latent

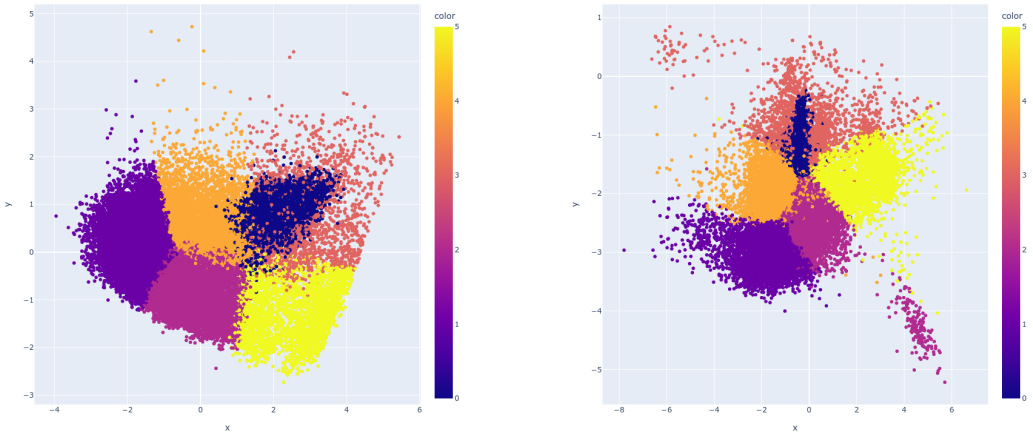
La Figure 4.7 compare la distribution des observations dans l'espace de l'ACP avec l'espace latent de l'auto-encodeur convolutif. On peut observer que les relations de proximité entre les groupes sont conservés. Cependant, on voit l'effet de la non linéarité sur l'espace latent : certains points sont plus écartés de la zone de forte densité. Ils correspondant probablement à des observations atypiques (aberrantes, bruitées, peu représentées).

Génération de données

La génération des données consiste à échantillonner aléatoirement des points dans l'espace latent. La Figure 4.8 présente la génération de 49 trafics de cellule à partir de points pris uniformément dans l'espace. On observe que les points de la moitié haute de la grille d'échantillonnage génèrent des profils très bruités et peu réalistes, contrairement aux générations du bas qui sont de meilleure qualité. En comparant avec la Figure 4.7, on remarque qu'il y a très peu de données d'entraînement au dessus de l'ordonnée y.

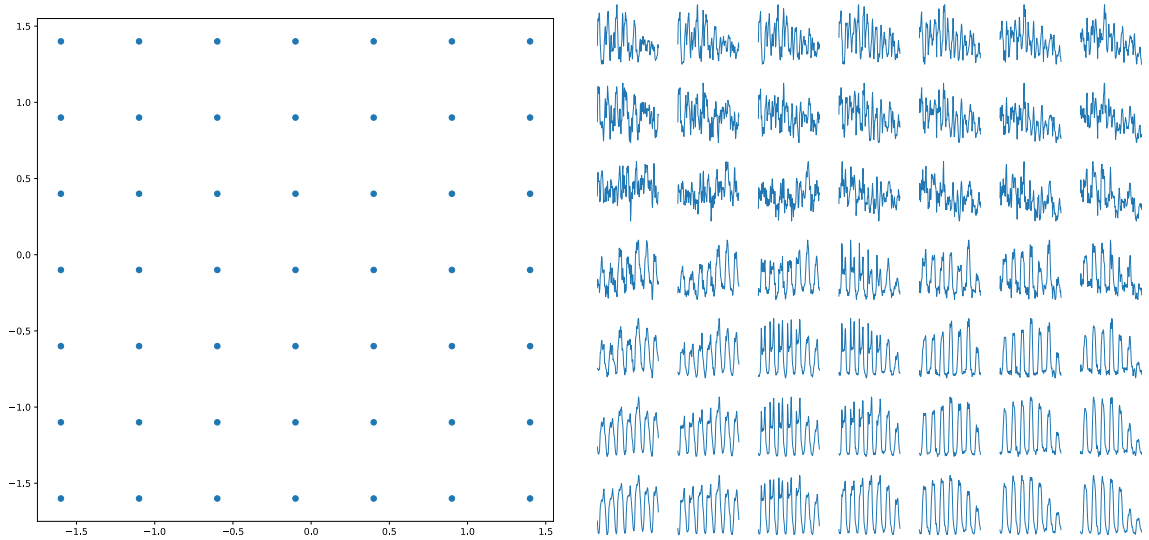
Par conséquent, il est préférable de prendre des points dans les endroits de l'espace où la densité des données d'entraînement est élevée. En effet, il n'est pas certain que les endroits de faible densité soient bien « connus » du modèle, ce qui pourrait générer des données synthétiques peu réalistes.

L'inconvénient d'un auto-encodeur convolutif tel que celui-ci que l'entraînement n'est pas déterministe (en partie dû à la descente de gradient stochastique). La forme de l'espace latent peut donc varier d'un entraînement à un autre. Cela rend le choix de l'intervalle d'échantillonnage difficile à déterminer.



(a) ACP (b) Auto-encodeur convolutif

FIGURE 4.7 – Comparaison des distributions des observations réduites par l'ACP (à gauche) et par l'auto-encodeur convolutif (à droite).



(a) Grille d'échantillonage z dans l'espace latent (b) Données $d(z)$ générées en sortie du décodeur

FIGURE 4.8 – Génération de données par l'auto-encodeur convolutif à partir d'une grille d'échantillonnage sur l'espace latent

4.5 Auto-encodeur convolutif variationnel

Pour rendre l'espace latent plus régulier et limiter les zones inconnues, une solution est contraindre les observations latentes à rester dans un cadre pré-déterminé à l'avance.

L'auto-encodeur variationnel utilise l'approche bayésienne pour améliorer la qualité de génération de données synthétiques [115]. Cette approche a pour conséquence que le modèle en lui-même n'est plus déterministe, c'est à dire qu'une même entrée passée à l'auto-encodeur peut générer plusieurs sorties différentes, bien que restant semblables à l'entrée. L'entraînement utilise l'astuce de reparamétrisation (*reparameterization trick*) pour que la rétropropagation du gradient (et donc l'entraînement du modèle) reste possible.

Architecture La Figure 4.9 illustre l'architecture de l'auto-encodeur convolutif variationnel (AECV) utilisé. Comme le modèle précédent, il possède des couches convolutives pour traiter les images. La différence se trouve au niveau des couches latentes : au lieu de prédire un seul vecteur latent, deux vecteurs de taille $r = 2$ sont renvoyés par l'auto-encodeur. On note ces vecteurs μ_X et σ_X.

L'astuce de reparamétrisation consiste à donner en entrée au décodeur est un vecteur aléatoire z suivant la loi normale multivariée $\mathcal{N}(\mu_X, \sigma_X)$. Pour cela, on échantillonne d'abord une variable aléatoire ϵ suivant la loi normale multivariée $\mathcal{N}(0, I_2)$. Ensuite, le vecteur z est obtenu suivant la relation :

$$z = \mu_x z + \sigma_x$$

De cette manière, le modèle reste probabiliste grâce à la variable ϵ qui ne dépend pas de l'apprentissage, mais peut être entraîné en ajustant les paramètres dont dépendent μ_X et σ_X.

Erreur de reconstruction Pour régulariser l'espace latent, la pratique commune est de le contraindre à se rapprocher de la distribution gaussienne (Figure 4.10 (a)). Pour cela, on ajoute un terme supplémentaire à l'erreur quadratique moyenne. Ce terme est la divergence de Kullback-Leibler (KL) :

$$\text{Loss}(X, X_{\text{rec}}) = \text{MSE}(X, X_{\text{rec}}) + \beta \text{KL}(N(\mu_X, \sigma_X), N(0, I_2))$$

où X correspond à l'entrée et X_{rec} à sa reconstitution en sortie du modèle. La divergence est d'autant plus petite que μ_X se rapproche de 0, et σ_X de I_2.

Le terme $\beta > 0$ est un hyperparamètre permettant d'équilibrer la force des deux termes pour trouver un compromis. En effet, une bonne reconstitution mettra l'accent sur la minimisation de la MSE, mais conduira à un espace moins régulier. Au contraire, un espace plus régulier peut rendre la reconstitution moins bonne car on force des observations très différentes à être plus proches dans l'espace latent. Cela peut être plus difficile pour le modèle de les distinguer au moment du décodage.

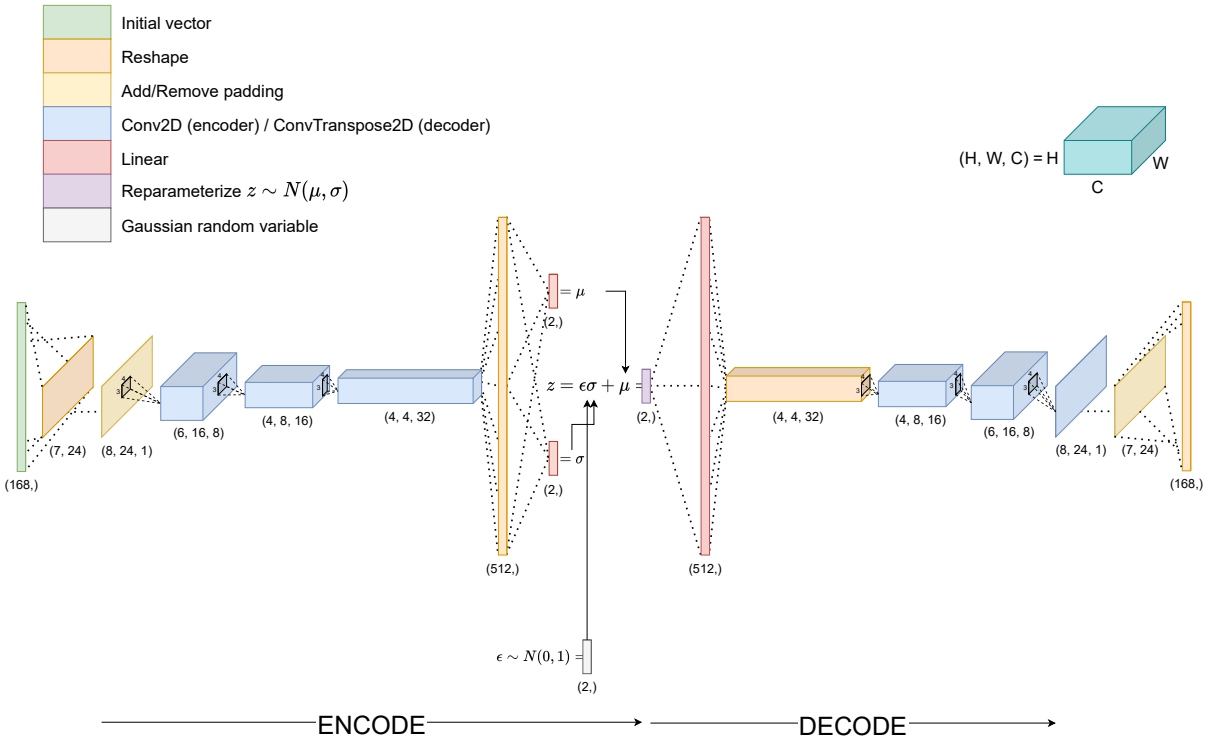
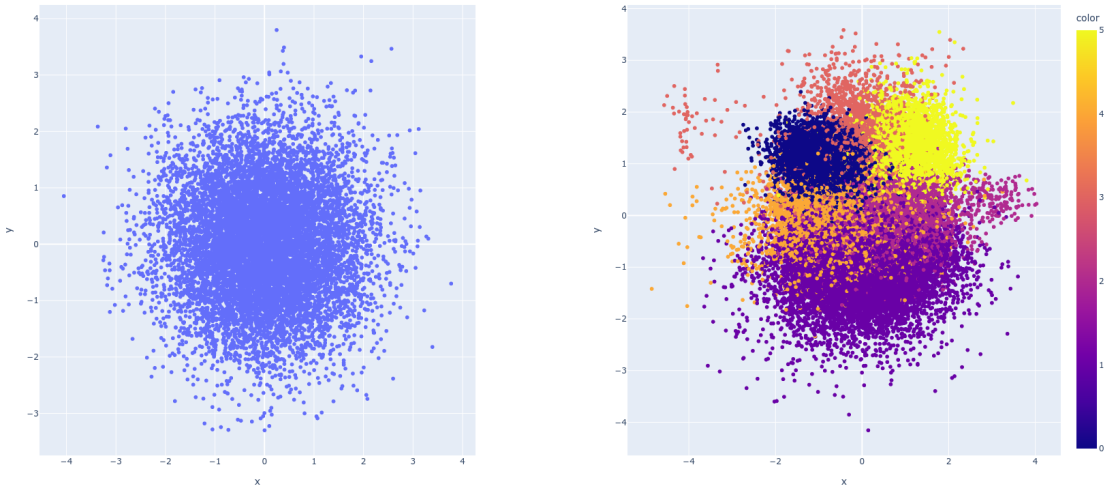


FIGURE 4.9 – Architecture de l'auto-encodeur convolutif variationnel utilisé pour l'étude

Analyse de l'espace latent

La Figure 4.10 compare un échantillon d'observations issus loi normale multivariée $\mathcal{N}(0, I_2)$ (a) avec la distribution des observations latentes de l'AECV (b). Par rapport aux modèles non variationnels, on constate que l'espace latent est plus régulier. Le nuage de point est similaire à la distribution $\mathcal{N}(0, I_2)$, tout en gardant les observations de même groupe ensemble, et le même agencement entre les groupes. Bien que certains points s'éloignent de la distribution normale, ceux-ci restent tout de même proches du centre de masse.



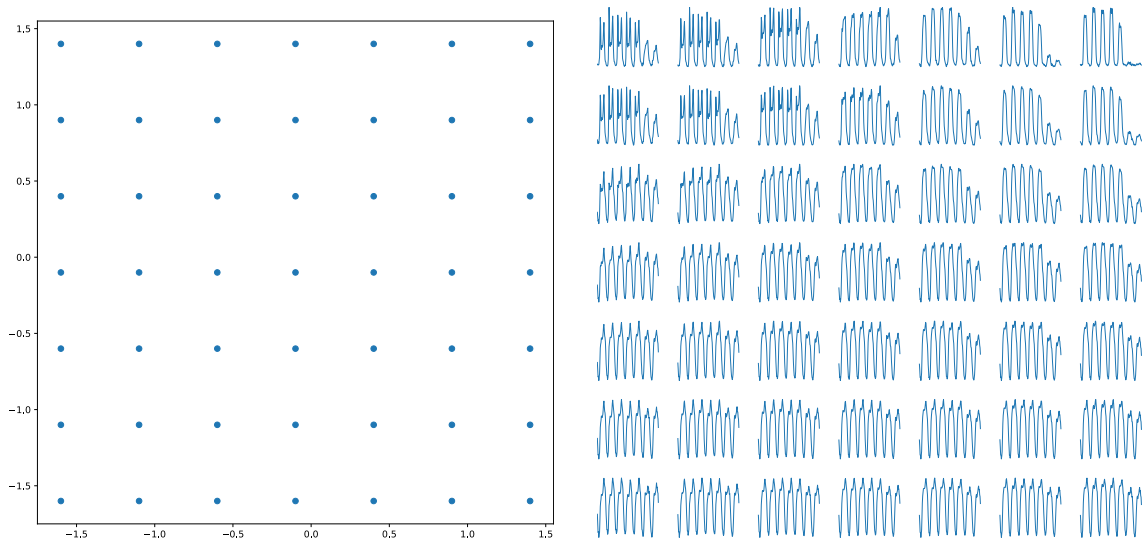
(a) Échantillons issus de $\mathcal{N}(0, I_2)$

(b) Sortie produite par l'encodeur de l'AECV dans l'espace latent

FIGURE 4.10 – Comparaison de l'espace latent de l'AECV avec un échantillon de loi normale multivariée $\mathcal{N}(0, 1)$

Génération de données

La Figure 4.11 montre la génération des données suivant la même grille que la Figure 4.8. On peut voir que les données synthétiques sont réalistes, sans motifs bruités ou aberrants. Pour cause, les points échantillonnés se situent tous dans la zone de distribution des données latentes. La contrainte de régularisation permet de garder cette grille d'échantillonnage fixe pour tous les AECV entraînés de cette façon.



(a) Grille d'échantillonnage de z

(b) Données $d(z)$ générées en sortie du décodeur

FIGURE 4.11 – Génération de données par l'auto-encodeur convolutif variationnel suivant une grille d'échantillonnage dans l'espace de forte densité de $\mathcal{N}(0, I_2)$

Posterior collapse

Durant l'entraînement de l'AECV, on a pu observer empiriquement qu'une régularisation trop forte produisait le phénomène de *posterior collapse*. Cela se traduit par la dégénération de toutes les reconstitutions en une sortie unique proche de la valeur moyenne de toutes les observations du jeu d'entraînement.

La Figure 4.12 illustre ce phénomène observé depuis l'espace latent. Bien que les points se distribuent selon une gaussienne, tous les groupes sont mélangés. Il n'y a donc pas de conservation d'une information permettant de distinguer les différents profils d'affluence contenus dans les données. Le décodeur ne pouvant pas se baser sur les coordonnées latentes d'un point pour reconstruire correctement l'entrée, la meilleure solution pour minimiser l'erreur de reconstruction est de renvoyer une constante qui minimise la distance à toutes les entrées.

En pratique, pour éviter l'effondrement, une heuristique courante est de réduire le coefficient β, de la faire varier au cours de l'entraînement (recuit simulé), ou de renforcer le poids du terme de reconstitution en sommant les erreurs de reconstitution au lieu de les moyenner.

Les vraies causes de cet effondrement ne font cependant pas l'unanimité. Des études récentes [116, 117] tendent à dire que le déséquilibre entre l'erreur de reconstitution et de régularisation n'est pas la cause principale du *posterior collapse*, mais qu'une variance σ_x trop élevée peut en être à l'origine.

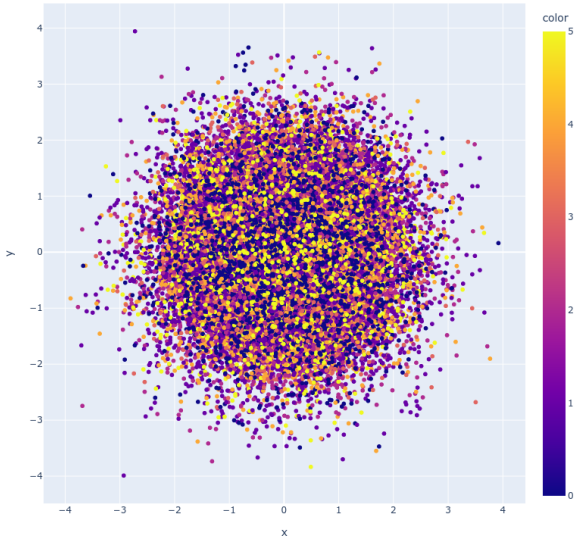


FIGURE 4.12 – Illustration d'un espace latent non informatif touché par le posterior collapse

4.6 Auto-encodeur convolutif variationnel discret

L'auto-encodeur convolutif variationnel discret (AECVD) est un auto-encodeur convolutif variationnel dont les vecteurs de l'espace latent prennent un nombre fini et discrets de valeurs possibles. Ce modèle est utilisé par la première version de DALL·E [109] pour la génération en haute résolution d'images réalistes.

L'espace latent étant discret, le décodeur peut être interfacé à des modèles produisant des sorties discrètes. Dans le cas de DALL·E, un transformer réalise la traduction du langage naturel vers une représentation réduite et discrétisée d'une image dans l'espace latent du décodeur, qui reconstitue cette image en haute résolution.

Architecture La Figure 4.13 illustre l'architecture de l'AECVD employé dans l'étude. Elle est similaire à celle de l'AECV utilisé précédemment. La différence se trouve une fois encore à l'interface entre l'encodeur et le décodeur. Les explications qui suivent sont en basées sur le billet de blog de Charlie Snell [118] et sur une implémentation open source d'AECVD [119].

Dans ce modèle, on souhaite que z soit un vecteur parmi n_T vecteurs possibles, ce qui rend l'espace latent discret et incompatible avec la rétropropagation du gradient qui nécessite une continuité de z à l'entraînement. La solution est d'utiliser la distribution de Gumbel-Softmax [120], de paramètres $x_e = (x_e^{(0)}, ..., x_e^{n_T}) \in \mathbb{R}^{n_T}$ et $\tau \in \mathbb{R}_+^*$. x_e est le vecteur des probabilités associés aux n_T vecteurs, et τ est un paramètre appelé la température. Soit y un vecteur échantillonné de la distribution de Gumbel-Softmax(x_e, τ). Les éléments y_i de y sont exprimés de la façon suivante :

$$y_i = \frac{\exp\left((\log(x_i) + \epsilon_i)/\tau\right)}{\sum_{j=1}^{n_T} \exp\left((\log(x_j) + \epsilon_j)/\tau\right)}$$

Lorsque la température τ tend vers 0, y_i tend vers un vecteur catégoriel (*one-hot*) dont un seul élément vaut 1 et tous les autres sont nuls. L'élément d'indice non nul est utilisé pour choisir un vecteur parmi les n_T possibles. Dans la suite, on emploiera le mot *Codebook* pour désigner la liste des n_T vecteurs de l'espace latent et Codebook$[i]$ pour désigner le vecteur d'indice i.

L'astuce de reparamétrisation fonctionne de la même façon que pour l'AECV. Pour rappel, dans la section précédente, on construisait le vecteur aléatoire z avec un vecteur $\epsilon \sim \mathcal{N}(0, I_2)$ et les sorties μ_x et σ_x de l'encodeur. Ici, z est construit à partir d'un vecteur aléatoire $\epsilon \in \mathbb{R}^{n_T}$ et les sorties $x_e = (x_e^{(0)}, ..., x_e^{n_T})$ de l'encodeur (Figure 4.13). Les éléments du vecteur ϵ sont des variables indépendantes et identiquement distribuées qui suivent la loi standard de Gumbel. Le vecteur z est obtenu par la formule suivante :

$$z = \sum_{i=0}^{n_T} y_i \times \text{Codebook}[i]$$

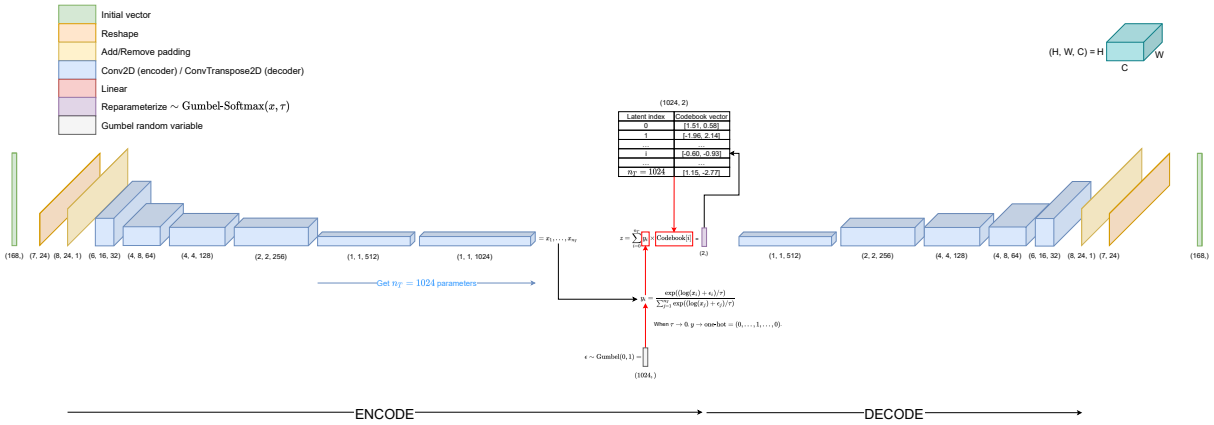


FIGURE 4.13 – Architecture de l'auto-encodeur convolutif variationnel discret utilisé pour l'étude. Les données réduites sont encodées avec $m = 1$ vecteurs de taille 2

À partir de cette expression, on peut remarquer que lorsque τ est proche de 0, z prend la valeur d'un vecteur latent du *Codebook*.

Pendant l'entraînement, le vecteur latent que le décodeur prend en entrée est une combinaison linéaire des vecteurs du *Codebook*. Si l'on veut récupérer la représentation latente d'une observation cependant, on prendra le vecteur du *Codebook* tel que l'élément de y de même index est de valeur maximale. C'est de cette manière que l'espace latent présenté dans la Figure 4.14(b) est obtenu.

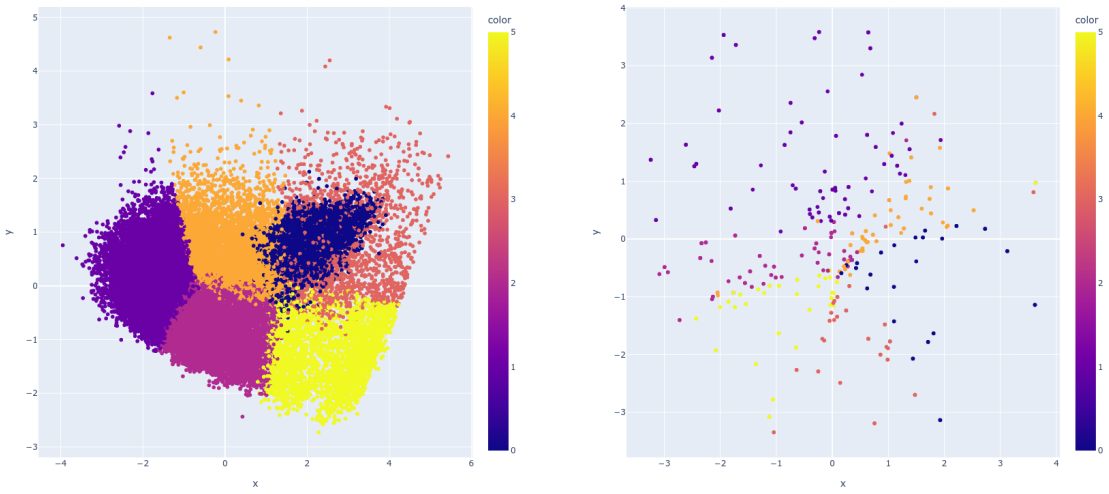
Analyse de l'espace latent

On peut observer que la densité des points est beaucoup plus faible que dans les autres modèles ; beaucoup d'observations sont superposées car elles sont rattachées au même vecteur latent. Bien qu'on ait choisi un *Codebook* de taille $n_T = 1024$, toutes les possibilités ne sont pas exploitées. On peut aussi voir que les agencements entre les observations intra-groupes et inter-groupes sont conservés.

Ouverture Bien que suffisant pour la taille de notre jeu de données, un espace latent discrétisé de 1024 possibilités est assez limité pour la génération d'images en haute résolution. En réalité, on peut recueillir un groupe de m vecteurs $x_{e_1}, ..., x_{e_m}$ en sortie de l'encodeur, réaliser pour chacun la reparamétrisation, et avoir ainsi une combinaison de m vecteurs latents pour représenter une entrée (Figure 4.15). Si l'on prend $m = 4$, cela nous fait $n_T^4 = 1024^4 \sim 10^{12}$ encodages possibles dans l'espace latent !

4.7 Conclusion

L'exploration des données du trafic cellulaire nous a permis d'étudier un ensemble de techniques de réduction de dimension et de génération de données. L'idée est donner une intuition du fonctionnement



(a) ACP (b) AECVD

FIGURE 4.14 – Comparaison des distributions des observations réduites par l'ACP (à gauche) et par l'auto-encodeur convolutif variationnel discret (à droite).

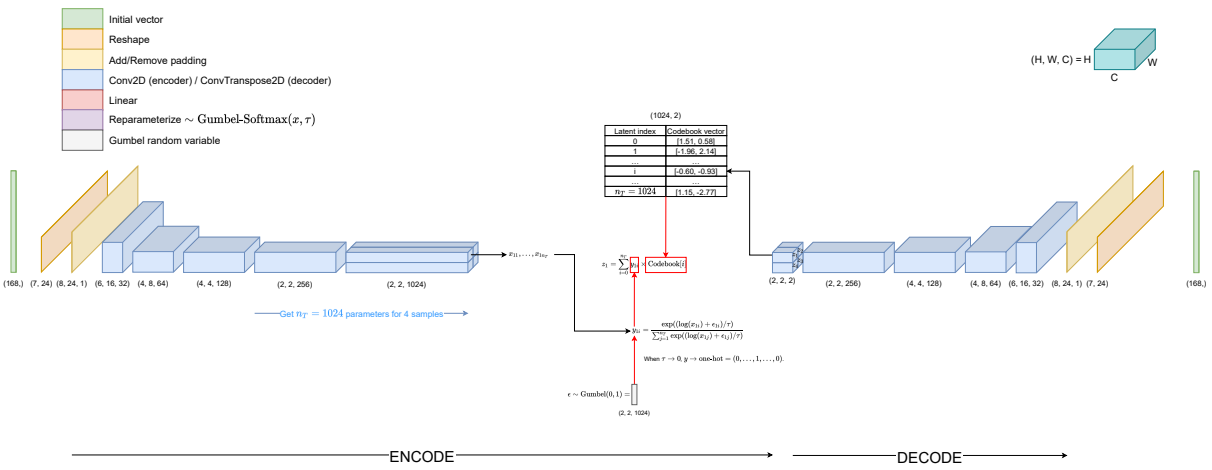


FIGURE 4.15 – Architecture de l'auto-encodeur convolutif variationnel discret codant les données dans l'espace latent avec une combinaison de $m = 4$ vecteurs de taille 2 du *Codebook*.

de ces modèles à travers l'observation de la distribution des données compressées dans l'espace latent. Bien que la structure et la distribution des données dans ces espaces diffèrent selon les modèles employés, l'agencement des groupes d'observations les uns par rapport aux autres est similaire : les groupes d'activités de nature résidentielle sont restés proches dans tous les espaces latents et s'opposent aux groupes dont le profil est marqué par les activités professionnelles humaines avec un faible trafic le week-end. L'étude de l'auto-encodeur convolutif variationnel a montré l'importance de régulariser l'espace latent pour éviter la génération de données synthétiques aberrantes. On a ainsi pu observer la variété de profils d'activités générés, qui sont semblables aux centres de classes des données réelles.

Les études de ce chapitre ont permis de mettre au point l'auto-encodeur convolutif variationnel discret qui sera utilisé dans le Chapitre 5. Il permet d'entraîner les modèles à prédire la signature hebdomadaire médiane de l'affluence des utilisateurs dans un temps plus court. Au lieu de prédire les 168 valeurs d'une signatures, les modèles apprennent à prédire 4 valeurs discrètes dans l'espace latent de l'AECVD, puis ces représentations sont reconstituées par le décodeur en signatures.

Chapitre 5

Prédiction du traffic hebdomadaire médian

5.1 Introduction

Les études réalisées dans le Chapitre 3 avaient pour objectif de mieux comprendre les interactions entre le réseau mobile, le paysage urbain et les activités humaines. Les données mobiles, par leur abondance, sont souvent complémentaires et plus fiables que les recensements traditionnels [121] pour développer les aménagements et les transports urbains. Inversement, on a pu montrer qu'il était possible d'utiliser des données cartographiques précises (points d'intérêts et peuplement) pour prédire une information concise sur l'affluence d'une station de base. L'objectif était de savoir si une station de base à déployer dans un futur proche présenterait un usage de type résidentiel, industriel ou mixte. Ces trois classes ont été obtenues automatiquement en utilisant l'apprentissage supervisé pour regrouper les signatures hebdomadaires médianes des stations de base. Pour rappel, ces signatures décrivent les indicateurs de performance des stations en situation de routine (hors événements et périodes atypiques). Pour plus de détails, on réfère le lecteur au Chapitre 3, Section 3.3.2.

Une limite de la méthode précédente est le choix subjectif (et parfois ambigu) du nombre de classes. Pour illustrer ce propos, la Figure 5.1 présente les centres de classe résultant d'un regroupement des signatures en six classes, et la Figure 5.2 montre leur distribution spatiale. Les centres de classe 5.1c et 5.1d représentent les groupes de station de base d'affluence mixte. Dans le premier cas, les pics d'activité semblent dominés par l'usage résidentiel. Dans le deuxième cas, ils semblent dominés par un usage industriel. La classification des stations dans une classe ou l'autre peut être difficile car elles sont très proches. D'un autre côté, on peut craindre de perdre des informations précieuses en construisant un groupe de moins.

Ce chapitre a pour objectif d'étudier la possibilité de prédire directement les signatures. La tâche à apprendre n'est plus une classification, mais une régression multi-cibles. Une partie du travail sera consacrée à étudier les performances d'un ensemble de modèles nativement multi-cibles ou non, à la fois en termes de qualité des prédictions et de temps de calcul.

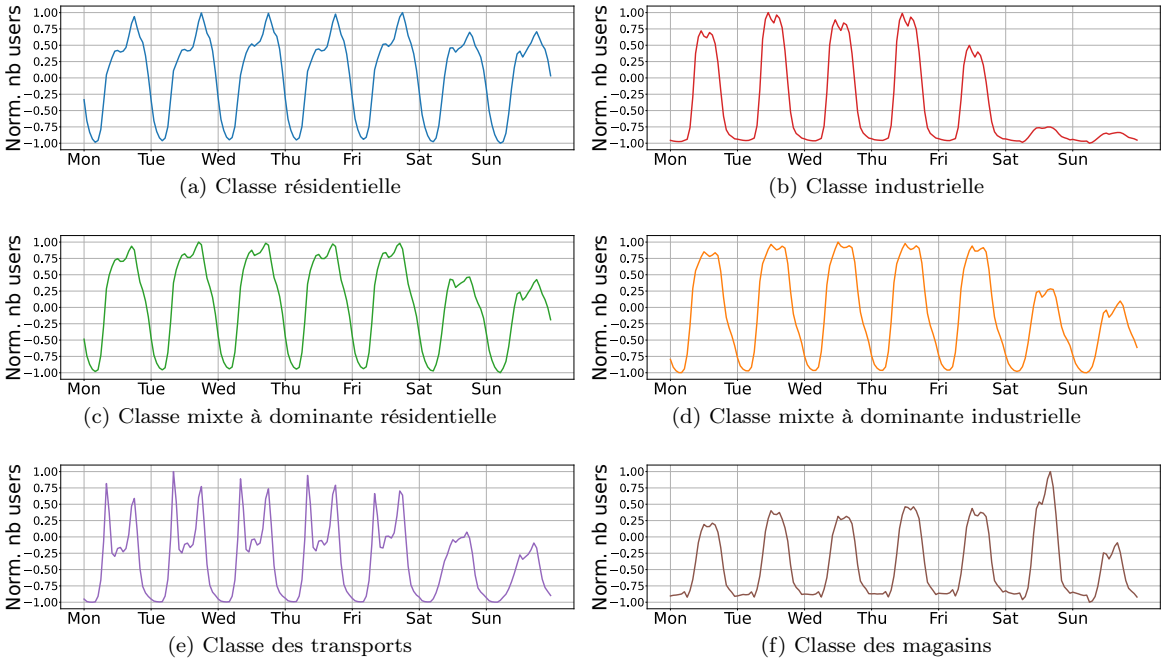
La suite du chapitre est structurée comme suit : la Section 5.2 présente les travaux relatifs à la régression multi-cible, la Section 5.3 formalise le problème de régression multi-cible et décrit les méthodes employées, la Section 5.4 évalue les performances des modèles entraînés sur des données de la région Île-de-France et la Section 5.5 conclut l'étude.

5.2 Travaux relatifs à la régression multi-cible

Réseaux de neurones et de l'apprentissage profond

La régression multi-cible (*multi-output regression*) est une variante de la régression simple. Au lieu de prédire une seule valeur continue, les modèles sont entraînés à prédire plusieurs valeurs. Dans les travaux de recherche liés aux données mobiles, la régression multi-cible concerne principalement la prédiction dynamique du trafic en s'appuyant sur des réseaux de neurones récurrents [25, 78].

Dans le cas présent, on cherche à prédire des données statiques à partir d'entrées tabulaires statiques. Le traitement des données tabulaires par les réseaux de neurones profonds n'est pas très documenté et les résultats sont très variables.



(a) Classe résidentielle

(b) Classe industrielle

(c) Classe mixte à dominante résidentielle

(d) Classe mixte à dominante industrielle

(e) Classe des transports

(f) Classe des magasins

FIGURE 5.1 – 6 centres de classes de signatures hebdomadaires médianes obtenus avec la méthode des k-moyennes

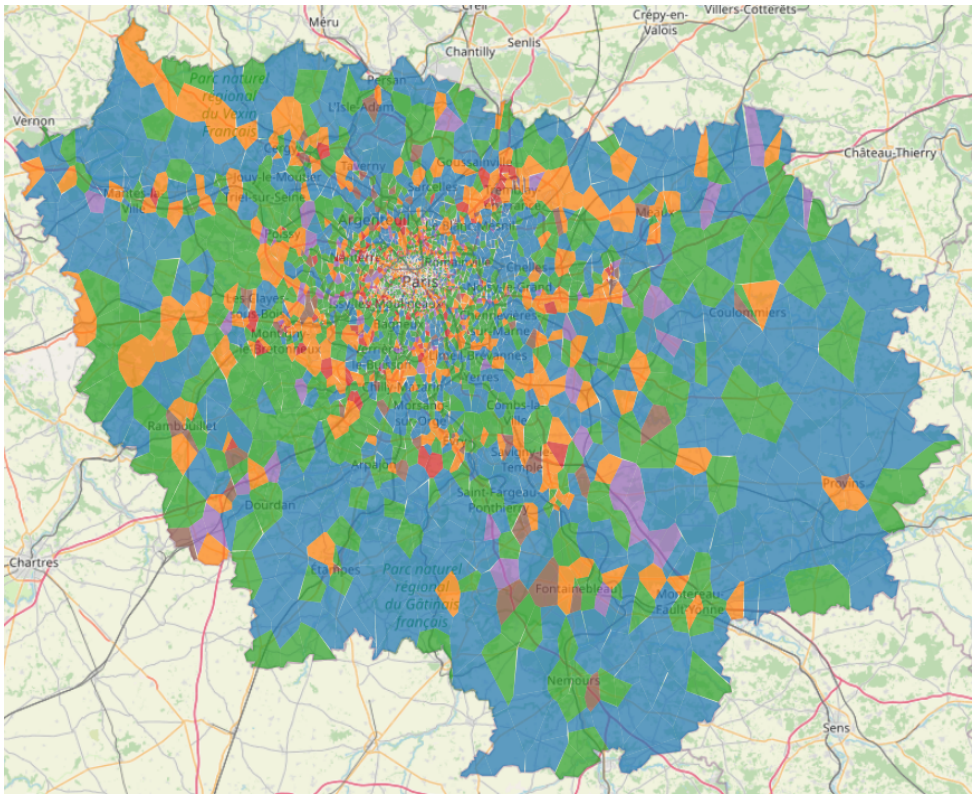


FIGURE 5.2 – Couverture des stations de base en Île-de-France modélisées par le diagramme de Voronoi. Les sites sont colorisés en fonction de leurs classes d'appartenance suivant le code couleur utilisé dans la Figure 5.1

D'une part, une étude a montré que le modèle XGBoost présentait des performances supérieures aux modèles profonds sur un ensemble de jeux d'entraînements différents [122]. Une autre étude a rapporté que les GBDTs sont actuellement les modèles à l'état de l'art pour la classification des données tabulaires [123].

Les résultats seraient d'autant plus tranchés que l'apprentissage est réalisé sur des petits jeux de données, c'est-à-dire comportant moins de 1 million d'observations [124].

D'autre part, les transformeurs [54] semblent être les modèles les plus prometteurs parmi les réseaux de neurones pour l'apprentissage des données tabulaires. Des chercheurs sont parvenus à égaler les performances des GBDTs sur quinze tâches de classification [123]. Cette architecture est aussi utilisée pour la classification de transactions bancaires, la prédiction d'indicateurs de pollution de l'air et la génération de données de transaction synthétiques [125].

A notre connaissance, aucune étude sur les performances des transformeurs n'a été réalisée sur des données tabulaires pour la régression multi-cible de données mobiles.

Autres méthodes multi-cibles

À part les réseaux de neurones, d'autres modèles d'apprentissages sont nativement compatibles avec la régression multi-cible. On emploie le terme « natif » dans le sens où un seul modèle est entraîné pour la prédiction de valeurs multiples. Les arbres de décision et les forêts aléatoires font partie de cette catégorie. En pratique, tout modèle de régression simple peut être adapté à la régression multi-cible grâce à la méthode dite de « *Single-Target* » (ST) [126]. La méthode ST consiste simplement à entraîner autant de modèles qu'il y a de valeurs à prédire.

Choix des modèles

À partir des travaux existants, une liste de modèles d'apprentissage à évaluer a été constituée. Elle comporte les GBDTs les plus populaires (CatBoost [61], LightGBM [62], XGBoost [63]), la machines à vecteurs de support, la méthode des k plus proches voisins, la régression *ridge* et le transformeur.

L'architecture du transformeur qui a été adoptée découle d'une réflexion sur le rapprochement de notre problème avec la génération de données [127], et plus particulièrement de la génération d'images. Les sorties que l'on cherche à prédire sont des signatures qui sont certes statiques, mais dont les valeurs successives dépendent les unes des autres, car elles sont construites à partir de séries temporelles périodiques. En comparaison, les images sont des données statiques, mais dont les pixels sont localement dépendants pour représenter des formes, des contours et des couleurs. Une idée est donc d'essayer d'adapter un modèle prenant des données semblables à des données tabulaires pour générer des images. La première version de DALL·E est un transformeur capable de générer des images à partir du language naturel [109]. On cherchera à adapter cette architecture en voyant les signatures comme des images et les données tabulaires comme un langage où chaque ligne est une phrase, chaque variable est un mot et sa valeur est la signification de ce mot.

5.3 Méthodes d'apprentissage pour la régression multi-cibles

5.3.1 Sources des données

Variables Comme dans le Chapitre 3, les variables passées en entrée aux modèles proviennent de la base de données OpenStreetMap [84] et du Humanitarian Data Exchange [90].

Cibles Les données mobiles utilisées pour produire les signatures des stations de base sont fournies par un opérateur mobile français. Elles sont collectées entre le 2 novembre 2020 et le 30 août 2021. Heure par heure, elles décrivent le nombre d'utilisateurs connectés aux 3 445 stations de base 4G de l'Île-de-France.

5.3.2 Régression multi-cible

Formalisation Soit $\{(X_i, y_i)\}_{i=1}^{N}$ des données d'entraînement associées à $N \in \mathbb{N}$ stations de base telles que :

- $X_i \in \mathbb{R}^p$ est le vecteur des p variables décrivant le tissu urbain couvert par la station de base d'indice i. Une variable peut être : le nombre de points d'intérêts d'un type donné, le nombre de zones d'un type d'utilisation de sol donné, la surface du point d'intérêt ou de la zone, le nombre d'habitants couverts, ou la surface de la couverture.
- $y_i \in \mathbb{R}^u$ est un vecteur représentant la signature hebdomadaire médiane [100], c'est-à-dire le nombre médian d'utilisateurs attachés à la station de base i au cours d'une semaine. Dans cette étude, la signature d'une station de base possède une granularité horaire, c'est à dire que $u = 7 \times 24 = 168$.

La solution recherchée est modélisée par une fonction $f : \mathbb{R}^p \longrightarrow \mathbb{R}^u$ telle que pour toute observation $i \in \{1, ..., N\}$:

$$f(X_i) = y_i + \epsilon_i$$

où $\epsilon_i \in \mathbb{R}^u$ désigne l'erreur de prédiction.

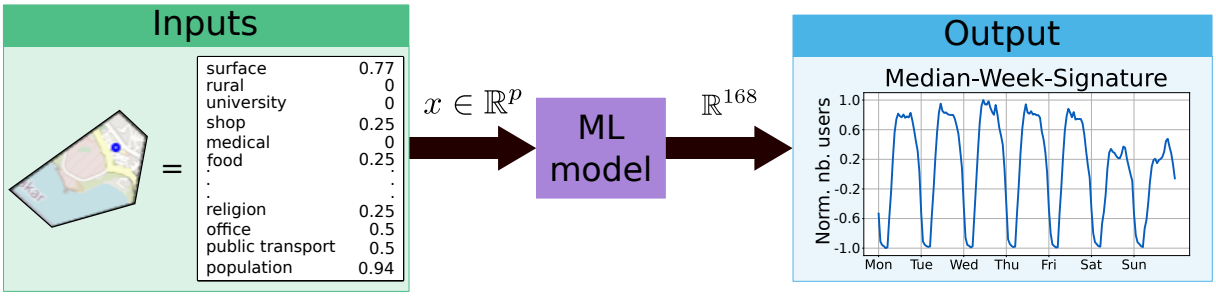
On propose deux approches, surnommées « C2C » et « D2D » pour minimiser les erreurs de prédictions. Elles sont illustrées sur la Figure 5.3.

Méthode « C2C » (*continuous to continuous*) Dans cette méthode, on considère simplement que la fonction f est un modèle d'apprentissage. Ce modèle peut être un GBDT, une machine à vecteurs de support, la méthode des k plus proches voisins, ou un régresseur ridge. Il prend en entrée des variables continues et produit des sorties continues (Figure 5.3a).

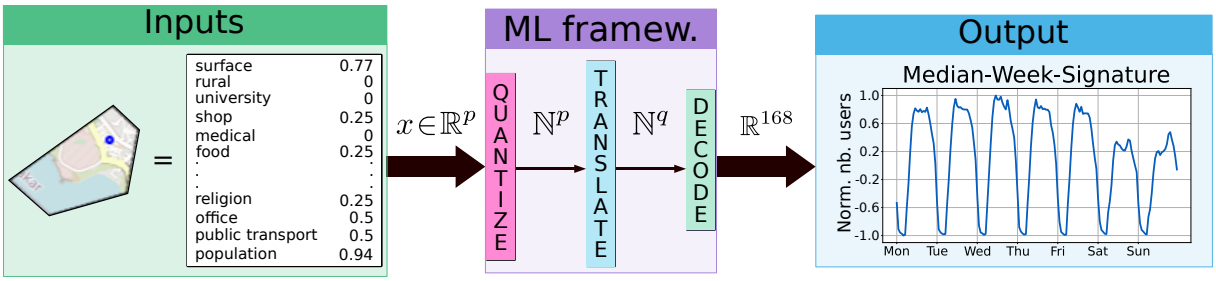
Méthode « D2D » Avec certains modèles, le temps d'apprentissage et/ou de prédiction peut être assez long en utilisant la méthode précédente, car il faut prédire 168 valeurs. Pour réduire le temps de calcul, on s'inspire de l'architecture de DALL·E [109] pour entraîner les modèles à prédire une représentation compressée des données. On note $f = d \circ f_{\text{D2D}} \circ q$ l'architecture suivante en trois composantes :

1. Une fonction de quantification $q : \mathbb{R}^p \longrightarrow \mathbb{N}^p$, qui transforme le vecteur de variables continues en vecteur de variables discrètes. Cette étape est nécessaire lorsqu'on utilise le transformeur.
2. Un modèle d'apprentissage $f_{\text{D2D}} : \mathbb{N}^p \longrightarrow \mathbb{N}^r$, qui « traduit » le tissu urbain en une représentation réduite de la signature constituée de $r < u$ variables.
3. Une fonction $d : \mathbb{N}^r \longrightarrow \mathbb{R}^u$, qui correspond au décodeur d'un auto-encodeur variationnel à valeurs discrètes entraîné sur les signatures.

Le modèle f_{D2D} prend entrée des variables discrètes et produit des sorties discrètes (Figure 5.3b). Notons que f_{D2D} est un modèle de classification (et non de régression) multi-cible.



(a) C2C



(b) D2D

FIGURE 5.3 – Architectures et flux de données des méthodes C2C et D2D.

5.3.3 Cibles

Pour chaque station de base, les signatures hebdomadaires médianes sont calculées à partir des données mobiles suivant la procédure décrite dans la Section 3.3.2.

5.3.4 Variables d'entrées

Les variables décrivant le tissu urbain couvert par chaque station de base sont obtenues par intersection spatiale des données de sources externes (points d'intérêts, utilisation des sols, populations) avec le diagramme de Voronoi modélisant la couverture des stations de base. Pour des explications plus approfondies, nous référons le lecteur à la Section 3.3.3 du Chapitre 3. Par rapport au travail précédent, on inclut la surface des objets (et pas seulement leur compte) parmi les variables. La surface des bâtiments pourrait être corrélée à l'intensité des activités humaines, ce qui à son tour, pourrait influencer les profils d'activité des stations de base. À l'issue du traitement des données, on obtient $p = 409$ variables.

Stockage des données Les données OSM sont stockées dans une base de données PostgreSQL/PostGIS pour tirer parti de l'indexation spatiale et optimiser les requêtes d'intersection, d'agrégation et de jointure spatiales.

5.3.5 Implémentation

On détaille ici l'implémentation des étapes de quantification (Quantize), traduction (Translate) et décodage (Decode) illustrées sur la Figure 5.3.

Quantification (D2D uniquement) Les variables sont quantifiées avec la méthode des k-moyennes, de paramètre $k = 20$.

Traduction (C2C et D2D) En l'absence du support multi-cible dans l'implémentation de certains modèles d'apprentissage, ceux-ci sont entraînés suivant la méthode *single target* (autant de modèles entraînés que de sorties à prédire). Le cas échéant, ils sont suffixés du terme « -ST » dans les tableaux des résultats.

Les régresseurs évalués sont les suivants :

— des GBDTs, a priori très performants sur les données tabulaires : CatBoost, LightGBM et XGBoost. Les trois modèles sont entraînés suivant la méthode ST. Au moment où nous réalisons cette étude, CatBoost est le seul à supporter nativement la régression multi-cible. On comparera la variante CatBoost-ST avec sa version nativement multi-cible.

— des modèles de régression plus simples, provenant de la bibliothèque Python `scikit-learn` [59] : la forêt aléatoire (RF, RF-ST), la méthode des k plus proches voisins (KNN), le régresseur *ridge* paramétré automatiquement par la validation croisée (RCV-ST), la machine à vecteurs de support (SVM-ST) et le perceptron multicouche (MLP).

— Le transformeur utilisé est adapté d'une implémentation open source [128] de DALL·E [109]. Si les performances du modèle s'avèrent intéressantes, on pourrait exploiter les avantages des réseaux de neurones comme le transfert d'apprentissage (d'un cas d'étude d'un pays à un autre par exemple). La taille des transformeurs étant proportionnelle à la taille des variables d'entrée, des limitations de mémoire ont été rencontrées avec 409 variables. Une analyse en composantes principales a été réalisée pour ne garder que $p_{red} = 42$ variables, qui expliquent 55% de la variance initiale.

Décodage (D2D uniquement) Avant d'entraîner l'auto-encodeur, les signatures, initialement des vecteurs de taille 168, sont redimensionnées en matrices de taille $(7, 24, 1)$. On pense que cette représentation en 2D pourrait faciliter l'extraction des périodicités journalières (suivant l'axe vertical) et des dépendances horaires (suivant l'axe horizontal) par les filtres de convolution.

L'auto-encodeur est entraîné pour compresser les signatures de taille $p = 168$ en une combinaison de $r = 4$ vecteurs pris dans un *Codebook* de taille 64. Le *Codebook* est une liste de taille finie indexant un ensemble de vecteurs. Son fonctionnement est décrit dans le Chapitre 4, Section 4.6.

L'encodeur est composé de couches convolutives 2D tandis que le décodeur est composée de couches convolutives 2D transposées. L'architecture complète est semblable à celle illustrée sur la Figure 4.15.

En sortie du décodeur, les images sont redimensionnées en vecteurs de taille 168.

5.4 Résultats et interprétation

Avant de présenter les résultats, on présente d'abord les lignes de base, les mesures d'erreur de prédiction et la configuration matérielle utilisées.

	MSE		RMSE		nMAE		DTW		EVS		MaxE		Time	
Model	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Train	Predict
CatBoost	**0.085**	0.005	**0.259**	0.007	0.320	0.018	1.979	0.046	**0.812**	0.011	**0.615**	0.015	13 :30 :02	00 :00 :03
CatBoost-ST	0.086	0.005	0.261	0.006	**0.311**	0.017	**1.946**	0.045	0.810	0.011	0.644	0.014	13 :40 :32	00 :03 :29
RF	0.088	0.005	0.263	0.007	0.323	0.019	2.004	0.047	0.807	0.011	0.622	0.015	00 :03 :25	00 :00 :04
RF-ST	0.089	0.005	0.266	0.007	0.317	0.019	1.975	0.044	0.804	0.012	0.656	0.014	02 :23 :39	00 :07 :44
LightGBM-ST	0.089	0.005	0.266	0.007	**0.311**	0.017	1.948	0.044	0.803	0.011	0.678	0.015	00 :16 :31	00 :00 :35
XGBoost-ST	0.097	0.005	0.282	0.007	0.321	0.017	2.030	0.041	0.780	0.012	0.776	0.016	00 :49 :17	00 :00 :16
RCV-ST	0.095	0.004	0.279	0.006	0.342	0.020	2.102	0.046	0.799	0.011	0.649	0.133	00 :00 :12	00 :00 :01
SVM-ST	0.097	0.005	0.281	0.007	0.333	0.019	2.093	0.048	0.793	0.012	0.663	0.015	00 :16 :12	00 :04 :21
B2 CatBoost	0.100	0.006	0.285	0.008	0.349	0.020	2.108	0.053	0.792	0.014	0.651	0.016	00 :55 :46	00 :00 :02
B2 RF	0.104	0.007	0.287	0.009	0.326	0.020	2.071	0.060	0.786	0.016	0.663	0.019	00 :04 :28	00 :00 :02
B2 KNN	0.106	0.007	0.290	0.009	0.326	0.018	2.091	0.055	0.782	0.016	0.674	0.018	00 :00 :01	00 :00 :01
KNN	0.124	0.005	0.321	0.007	0.361	0.019	2.343	0.047	0.753	0.013	0.752	0.014	00 :00 :01	00 :00 :01
B1 AVG	0.129	0.006	0.331	0.009	0.411	0.023	2.382	0.041	0.749	0.012	0.729	0.012	00 :00 :01	00 :00 :01
MLP	0.138	0.008	0.332	0.009	0.357	0.018	2.429	0.067	0.730	0.013	0.764	0.018	00 :04 :59	00 :00 :01

TABLE 5.1 – Mesures des erreurs de prédictions des modèles C2C

	MSE		RMSE		nMAE		DTW		EVS		MaxE		Time	
Model	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Train	Predict
CatBoost-ST	**0.113**	0.007	**0.292**	0.009	**0.319**	0.017	**2.032**	0.059	**0.744**	0.016	**0.714**	0.020	03 :10 :31	00 :00 :15
RF-ST	0.114	0.006	0.298	0.008	0.326	0.016	2.074	0.052	0.740	0.013	0.727	0.017	00 :02 :03	00 :00 :43
XGBoost-ST	0.115	0.006	0.297	0.008	0.326	0.017	2.067	0.048	0.742	0.014	0.724	0.018	00 :15 :04	00 :00 :06
RF	0.116	0.006	0.301	0.008	0.327	0.015	2.086	0.051	0.736	0.014	0.733	0.018	00 :01 :17	00 :00 :23
SVM-ST	0.117	0.006	0.305	0.007	0.337	0.018	2.100	0.050	0.734	0.015	0.736	0.017	00 :03 :43	00 :02 :49
LightGBM-ST	0.121	0.014	0.310	0.022	0.346	0.030	2.150	0.150	0.733	0.030	0.743	0.046	00 :10 :23	00 :00 :50
RCV-ST	0.125	0.006	0.313	0.007	0.343	0.017	2.147	0.046	0.728	0.015	0.750	0.017	00 :00 :25	00 :01 :29
KNN	0.162	0.010	0.360	0.011	0.504	0.015	2.462	0.011	0.664	0.021	0.853	0.025	00 :00 :10	00 :00 :04
Transformeur	0.230	0.012	0.431	0.011	0.467	0.024	2.890	0.079	0.577	0.021	0.958	0.022	03 :32 :47	00 :00 :45

TABLE 5.2 – Mesures des erreurs de prédictions des modèles D2D (après décodage)

Lignes de base

Deux lignes de bases ont été établies pour estimer l'efficacité des méthodes proposées :
— Le modèle B1 retourne systématiquement la moyenne des signatures composant le jeu d'entraînement. Il représente le biais du jeu de données.
— Les prédictions des modèles de type B2 dépendent des voisins géographiquement proches de la station dont on veut connaître la signature. Ils prennent deux variables en entrée : la latitude et la longitude du site. Les modèles d'apprentissage entraînés sur les positions des stations de base sont : la méthode des k plus proches voisins (B2 KNN), CatBoost (B2 CatBoost) et la forêt aléatoire (B2 RF).

Mesures

Les mesures suivantes sont utilisées pour évaluer les erreurs de prédiction des modèles : l'erreur quadratique moyenne (MSE), sa racine carrée (RMSE), l'erreur absolue moyenne standard (nMAE), le score de variance expliquée (EVS), l'erreur maximale (MaxE) et la déformation temporelle dynamique (DTW).

Ces mesures sont calculées sur des observations individuelles, puis moyennées sur l'ensemble des données de test. Un bon modèle doit avoir des mesures faibles de MSE, RMSE, nMAE, DTW et MaxE, et un EVS élevé.

Implémentations utilisées Les métriques MSE, MaxE et EVS sont calculées avec les méthodes de la bibliothèque `scikit-learn`. La DTW est calculée en utilisant la bibliothèque `tslearn` [77]. Pour l'expression de ces mesures, on réfère le lecteur à la Section 2.2.6 de l'état de l'art.

Configuration matérielle

Les entraînements ont été réalisés avec le matériel suivant :
— 1× Intel(R) Core(TM) i9-10980XE CPU 3.00GHz

— 8× Crucial DIMM DDR4 Synchronous 2666 MHz
— 1× GeForce RTX 2080 Ti GPU

Les modèles de l'étape « Traduction » ont été entraînés sur les CPUs, et le décodeur sur le GPU.

Les performances rapportées sont les moyennes des mesures d'erreur sur 5 répétitions de validation croisée à 10 blocs, pour un total de 50 entraînements par modèle.

5.4.1 Évaluation des performances

Résultats des modèles C2C

À partir des résultats rapportés dans le Tableau 5.1, on analyse les performances et les temps d'exécution des modèles. En particulier, on compare les différences de temps de calcul entre les modèles nativement multi-cibles et ceux entraînés avec la méthode ST.

Performances de prédiction Les GBDTs et la forêt aléatoire sont les modèles les plus performants, ce qui est cohérent avec la littérature sur les données tabulaires. Les modèles des k plus proches voisins et du perceptron multicouche souffrent très probablement de la malédiction de la dimension à cause du nombre élevé de variables d'entrée. Mis à part ces deux modèles, les autres dépassent les performances des lignes de base B1 et B2. Les données du tissu urbain contiennent donc des informations latentes sur la demande des utilisateurs qui permettent de prédire des signatures plus précises que le biais (B1) et que le trafic localisé (B2).

Les temps d'exécution relevés correspondent au temps total des 50 entraînements. Malgré le multi-threading, les modèles ST prennent plus de temps à être entraînés que leur équivalent multi-cible natif (lorsqu'il existe). On observe de grands écarts de temps de calcul entre les modèles. Par exemple, CatBoost et CatBoost-ST réalisent les meilleures prédictions, mais ont aussi les temps de calculs les plus longs. Cette différence peut s'expliquer par le fait que Les GBDTs n'ont pas tous les mêmes manières d'optimiser les paramètres. Le temps d'entraînement de la forêt aléatoire est plus faible que les GBDTs, car les entraînements des arbres qui la compose sont parallélisés. En comparaison, les arbres des GBDTs sont entraînés séquentiellement. Les modèles nativement multi-cibles ont le plus faible temps de prédiction, ce qui est logique puisque toutes les valeurs d'une signature sont prédites par un seul modèle.

Comparaison de la méthode ST avec les modèles nativement multi-cibles En comparant CatBoost et la forêt aléatoire avec leurs variantes ST respectives, on observe que les modèles ST ont des erreur nMAE et DTW plus faibles, tandis que les versions nativement multi-cibles ont des MSE, RMSE, EVS et MaxE plus faibles. Pour interpréter cette séparation entre les mesures, on a effectué une inspection visuelle des prédictions et de leur cible correspondante. On suppose qu'il existe un compromis entre une bonne prédiction de l'allure de la signature et un bon alignement de la courbe prédite avec la cible. La Figure 5.4 illustre cette hypothèse, avec un profil correctement prédit pour les jours de travaillés, mais une heure de pointe décalée de quelques heures. La MSE pénalise fortement ce décalage, mais pas la DTW car la distance entre la cible et la prédiction sont calculées après le ré-alignement des courbes. Les modèles multi-cibles sont entraînés avec des mesures qui pénalisent un mauvais alignement des prédictions, quitte peut-être à « émousser » l'intensité des pics de d'activité. Les modèles ST, qui traitent les sorties indépendemment, s'adaptent peut-être mieux à la distribution de leur sortie unique, mais vont présenter des erreurs plus importantes avec les métriques sans ré-alignement.

En conclusion, malgré les meilleures performances de CatBoost, la forêt aléatoire semble le meilleur compromis performance/temps d'entraînement. En fonction des besoins et des critères de planification, on pourra soit privilégier un modèle nativement multi-cible constitué d'arbres de décision (préférence pour des heures de pointes correctes), ou un modèle ST (préférence pour l'exactitude de l'intensité des pics d'activités).

Résultats D2D Le Tableau 5.2 rapporte les performances des modèles suivant la méthode D2D.

Comparaison avec la méthode C2C On observe une dégradation des performances : seul un modèle (D2D CatBoost-ST) possède des erreurs nMAE et DTW plus faibles que celles des lignes de base B1 et B2. Le modèle D2D CatBoost-ST est comparable au modèle C2C XGBoost-ST sur ces deux mesures. Comme les lignes de base ont une MSE et RMSE plus faible que le modèle D2D CatBoost-ST, mais une

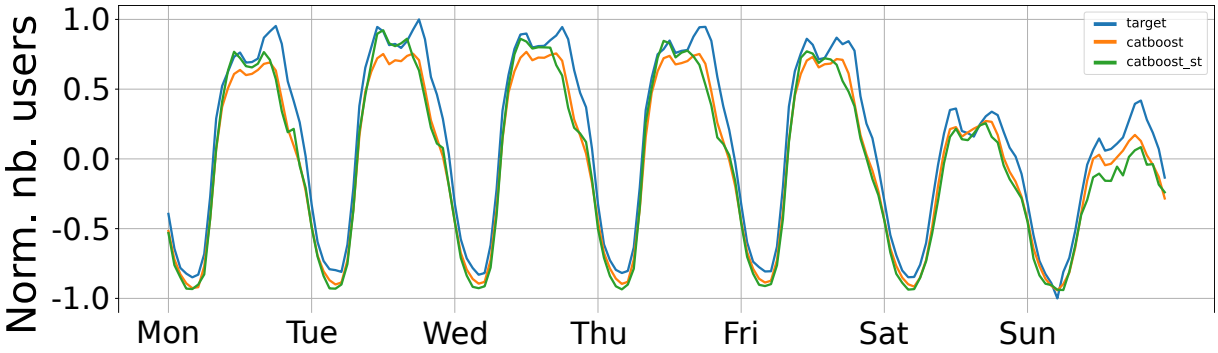


FIGURE 5.4 – Exemple d'une signature prédite par un CatBoost nativement multi-cible, un CatBoost-ST et la signature réelle. La prédiction de CatBoost-ST est légèrement plus décalée que l'autre prédiction, mais les pics d'activités des jours de la semaine sont mieux prédits.

DTW plus élevée, on peut supposer que le décodage reconstitue les signatures avec des pics d'activité décalés.

En revanche, les performances du transformeur sont loin derrière. Deux raisons peuvent expliquer ces mauvaises performances :

— premièrement, la taille du jeu d'entraînement est insuffisante. À titre comparatif, les travaux de [125] ont été réalisés sur des jeux d'entraînement d'environ 40000 échantillons.

— deuxièmement, nous avons dû réduire la taille d'un transformeur initialement pensé pour la génération d'images de haute résolution. D'autres adaptations sont probablement nécessaires concernant le nombre de couches, une simplification plus poussée de l'architecture, l'erreur d'entraînement à utiliser, etc.

Temps d'exécution Globalement, les temps d'exécutions des modèles D2D sont plus faibles que les modèles C2C correspondants. Le coût d'entraînement et de prédiction supplémentaire lié à l'auto-encodeur est négligeable. On relève un temps d'entraînement d'une dizaine de minutes pour 1000 itérations, et un temps de prédiction inférieur à la seconde. La plupart des modèles D2D ST ont des temps de prédiction inférieurs à leurs versions C2C ST, et un temps d'entraînement bien plus faible.

En conclusion, la méthode D2D produit des performances dégradées, mais possède des temps d'entraînement et de prédiction intéressants. Des travaux supplémentaires seraient nécessaires pour améliorer les performances des modèles D2D et l'architecture des réseaux de neurones.

5.5 Conclusion

Dans ce chapitre, on étend la méthode du Chapitre 3 pour prédire des informations plus détaillées sur le réseau mobile. De nombreux modèles d'apprentissage ont été évalués sur leur capacité à « traduire » des données du tissu urbain en signatures d'activité des stations de base. Les GBDTs et la forêt aléatoire sont les plus adaptés pour cette tâche de régression multi-cibles.

On propose également une méthode pour réduire le temps d'entraînement en apprenant aux modèles à prédire des représentations compressées des signatures. Cette approche a mené à une dégradation de la qualité des prédictions, bien que le meilleur modèle reste comparable aux performances des lignes de base. L'une des raisons pour laquelle une architecture si complexe a été utilisée était la volonté d'étudier les transformeurs sur ce type de problème. Elle était motivée par leurs performances à l'état de l'art dans les domaines de la vision par ordinateur et le traitement du langage, ainsi que la possibilité de transférer l'apprentissage. Cependant, des travaux plus approfondis (sur l'architecture, la représentation des données tabulaires passées en entrée, l'erreur de prédiction utilisée pour l'apprentissage...) devront être réalisés pour obtenir des résultats comparables aux modèles basés sur les arbres de décision.

Les modèles d'apprentissage entraînés avec cette méthode pourraient trouver leur application dans des problèmes d'extension de couverture, pour connaître à l'avance la demande attendue sur les territoires non couverts. Cet aspect est développé dans l'Annexe A, où les meilleurs modèles de l'étude (CatBoost et la forêt aléatoire) sont pré-entraînés sur des données du Sénégal, et intégrés à une interface graphique de démonstration.

Chapitre 6

Modélisation de la couverture de service des cellules

6.1 Introduction

Jusqu'à présent, les études exploitaient des données mobiles à l'échelle de la station de base, toutes fréquences confondues. Pour descendre à une granularité plus fine, comme celle du secteur, il est nécessaire d'avoir un modèle de couverture adapté à cette échelle.

La qualité de l'information qui peut être apprise depuis les sources externes dépend fortement du traitement des données et de la manière de les croiser. Par exemple, quelles variables choisit-on de garder ? Comment croiser les données externes avec celles du réseau mobile de façon à ce que le résultat soit pertinent ?

Les travaux de ce chapitre visent à répondre à la deuxième question. Pour maximiser les corrélations entre les performances des cellules, l'activité humaine et le tissu urbain, l'idéal serait de pouvoir extraire les données externes situées uniquement dans leur zone de service. En théorie, il est possible d'obtenir la couverture exacte d'une cellule à partir des formules de propagation électromagnétique [44]. En pratique, l'utilisation de ces modèles n'est pas à la portée de tous, car ils nécessitent une connaissance parfaite de la zone de déploiement (ex : élévation du terrain, modèles 3D des bâtiments), font intervenir des paramètres complexes et sont coûteux en temps de calcul. Une solution alternative est d'utiliser un modèle semi-empirique dont les paramètres sont ajustés sur un échantillon de valeurs du signal mesuré depuis des positions d'utilisateurs. Il est cependant difficile d'obtenir un échantillon exhaustif à travers les campagnes de mesure, et le traitement de ce type de données pose des problèmes de confidentialité. Par conséquent, la plupart des études nécessitant un modèle de couverture mobile ont recours à des géométries simples qui requièrent peu (voire pas) de paramètres de configuration radio.

La manière la plus courante de partitionner une région géographique pour obtenir une carte de couverture est le diagramme de Voronoi. Cet algorithme prend en entrée les coordonnées géographiques des stations de base et renvoie les régions d'attachement des équipements utilisateurs. L'hypothèse sur laquelle se fonde ce modèle est de considérer qu'un utilisateur est attaché à la station de base dont il est le plus proche. À notre connaissance, très peu d'études ont été conduites pour analyser la validité du diagramme de Voronoi. De plus, la couverture modélisée est celle d'une station de base, alors que les infrastructures actuelles sont composées d'antennes tri-sectorielles. Il peut être pertinent d'étudier les performances du réseau mobile à cette plus petite échelle. En effet, les activités humaines ne sont pas homogènes dans l'espace, donc chaque secteur peut recevoir du trafic très différent. Malheureusement, la littérature à ce sujet ne semble pas très fournie. C'est pourtant une question très intéressante pour l'apprentissage automatique, car la disposition d'un modèle fiable de couverture sectorielle permettrait d'avoir des jeux de données plus précis et plus fournis. Les données mobiles à l'échelle sectorielle sont au minimum trois fois plus importantes que si elles étaient agrégées par station de base, et même plus si l'on différencie les cellules par bande de fréquence.

Ce chapitre présente deux études pour paramétrer des modèles de couverture sectorielle et évaluer leur précision.

La première étude a pour objectif de proposer et de comparer plusieurs modèles de couverture sectorielle. Ces derniers sont dérivés de modèles simples de couverture de station de base, qui sont subdivisés en sous-régions. Les deux approches considérées pour estimer la couverture d'un site et de ses secteurs sont :

— Chaque polygone de Voronoi modélise la couverture d'une station de base, et chaque sous-région un secteur du site.
— La couverture de chaque station de base est modélisée par un cercle, découpé en secteurs circulaires, un par secteur d'antenne.

La taille des polygones et des cercles est modifiée au moyen de plusieurs stratégies de dimensionnement, certaines dépendant de paramètres de configuration radio et de formules de propagation semi-empiriques (Hata, UMa, RMa). Pour évaluer la précision des modèles, les couvertures sont superposées à des données de terrain. Ces données sont composées des emplacements des utilisateurs et de leur cellule de rattachement au début de la communication. L'objectif est de quantifier à quel point les formes des modèles géométriques se superposent bien aux positions réelles des utilisateurs.

Les données de terrain échantillonnées dans la région Île-de-France sont constituées de 526 000 positions d'utilisateurs, rattachés à 3 255 cellules. Elles peuvent être obtenues via le mécanisme de *Minimization of Drive Tests* (MDT) [129]. Les données ont été collectées sur une période allant du 1er octobre 2022 au 3 mars 2023, et ne concerne que les utilisateurs 4G. Au moment de l'étude, la 4G est la technologie dominante : c'est donc celle pour laquelle on peut avoir une distribution d'utilisateurs la plus représentative de la réalité. La Figure 6.1 présente la distribution spatiale des donées collectées, colorisées selon le contexte géographique. Le détail du nombre d'observations par contexte est donné dans le Tableau 6.1.

Les résultats de l'étude montrent que le modèle de couverture approchant au plus près la distribution spatiale des utilisateurs est le modèle cellulaire dérivé du diagramme de Voronoi. La précision du modèle est d'autant meilleure que les polygones sont uniformément agrandis d'un rapport dépendant de la bande de fréquence de la cellule.

La seconde étude a pour objectif d'évaluer l'efficacité d'utiliser des données spatiales agrégées au niveau des stations de base pour trouver le rapport optimal d'agrandissement des polygones de Voronoi. L'intérêt est de proposer une alternative basée sur des données certes moins détaillées, mais aussi moins sensibles et donc potentiellement plus accessibles. Les données décrivent la distribution des distances des utilisateurs à leur cellule de rattachement (calculées à partir du *Timing Avance*). Le rayon limite empirique est défini comme la distance pour laquelle 95% des distances des utilisateurs connectés se situent en dessous de cette valeur. Bien que moins précise, on montre que les rapports d'agrandissement résultants de cette méthode améliorent la modélisation de la couverture de cellule en comparaison avec les dimensions par défaut du diagramme de Voronoi.

Contexte	Nombre de positions utilisateurs collectées	Nombre de cellules
Rural	82 901	418
Périurbain	366 014	2 180
Urbain	77 090	657
Total	526 005	3 255

Table 6.1 – Nombre de positions utilisateurs collectées

6.2 Travaux existants

Les problèmes d'apprentissage automatique faisant appel à des sources externes pour prédire les performances du réseaux mobile sont étudiés sous deux angles principalement : à l'échelle de l'utilisateur ou à l'échelle du site. En fonction de la granularité spatiale, la méthode employée pour croiser les données exogènes est différente.

Dans le premier cas, la zone géographique est discrétisée en une grille de petites unités spatiales (le plus souvent des carreaux). Pour chaque unité, on y associe la performance du réseau mobile perçue par les utilisateurs dans la zone. Les données exogènes sont associées à chaque unité en réalisant des opérations d'intersection spatiale. Cette représentation géographique permet d'utiliser la répartition des points d'intérêt et des utilisations des sols pour prédire divers indicateurs comme le trafic [78] ou la force du signal reçu [80]. L'avantage de cette représentation est le degré élevé de précision spatiale, mais elle requiert en contrepartie d'importantes ressources pour stocker, traiter et anonymiser les données.

La seconde approche consiste à prédire la performance des réseaux mobiles au niveau des stations de base [79]. Bien que spatialement moins précises, les données d'entraînement agrégées sont moins sensibles et moins complexes à stocker. En revanche, c'est dans cette situation qu'il devient nécessaire de disposer d'un modèle de couverture, car c'est avec la carte des couvertures des sites (ou des cellules) que l'on partitionne la zone géographique et que l'on croise les données externes avec les données mobiles.

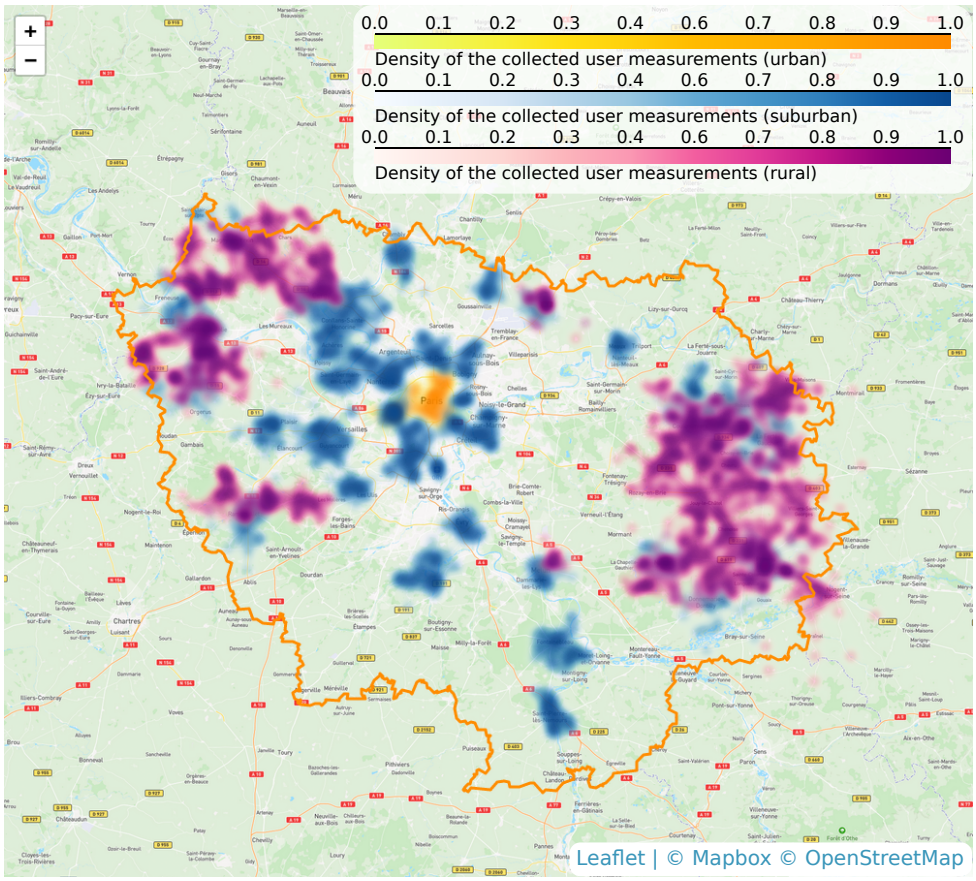


FIGURE 6.1 – Carte thermique des positions des utilisateurs collectées en Île-de-France. On distingue par couleur les utilisateurs attachés à des cellules urbaines (jaune), périurbaines (bleu) et rurales (violet). Le tracé orange correspond aux limites administratives de la région.

Le diagramme de Voronoi est une représentation couramment utilisée pour modéliser la couverture des stations de base. Son usage suppose que l'on considère que chaque utilisateur est connecté à la station de base la plus proche de lui. La définition de proximité dépend de la fonction distance utilisée. Par défaut, il s'agit de la distance euclidienne, mais des études proposent également d'utiliser la puissance de transmission des stations de base ou des secteurs pour rattacher les utilisateurs à l'équipement où la puissance reçue est la plus grande [130]. En dehors du problème étudié, les diagrammes de Voronoi sont couramment utilisés dans les travaux de géométrie stochastique [131] et dans les études sur les activités humaines. Par exemple, le trafic des réseaux mobiles est croisé avec le tissu urbain pour estimer la distribution en temps réel des populations par fonction d'unité urbaine [132], ou pour analyser la corrélation entre le niveau d'urbanisation et l'utilisation des réseaux mobiles [133]. Quelques travaux existent également sur la modélisation de la couverture des cellules à l'échelle cellulaire. Ces modèles divisent les régions de Voronoi en sous-régions en fonction de l'azimut [134], du tilt et de la hauteur de l'antenne [135]. Ces études considèrent que toutes les cellules situées sur le même secteur, indépendemment de la bande de fréquence, ont la même couverture.

Comme le modèle de Voronoi ne se base sur aucune configuration radio ou propriété de propagation, on peut s'attendre à ce que la précision du modèle soit limitée. Cependant, aucune étude pour comparer l'adéquation du diagramme avec la réalité n'a été publiée jusqu'à très récemment. A notre connaissance, un seul travail à ce sujet a montré qu'en moyenne, le diagramme de Voronoi ne couvre pas totalement la distribution des utilisateurs attachés à une station de base. Les auteurs rapportent une correspondance d'environ 50% entre le polygone et la distribution des utilisateurs produite par un logiciel de simulation [136]. L'étude présente concorde avec la conclusion de cette étude sur la nécessité de redimensionner les polygones de Voronoi si l'on veut mieux refléter la réalité. Cependant, on choisit de représenter la réalité du terrain avec un échantillon de mesures réelles, et une approche différente pour évaluer la précision des géométries. Pour que la couverture échantillonnée soit représentative de la couverture réelle, la distribution des positions collectée doit être uniforme. D'après nos observations, les positions collectées sont toujours situées au niveau de grands axes routiers, de routes résidentielles et

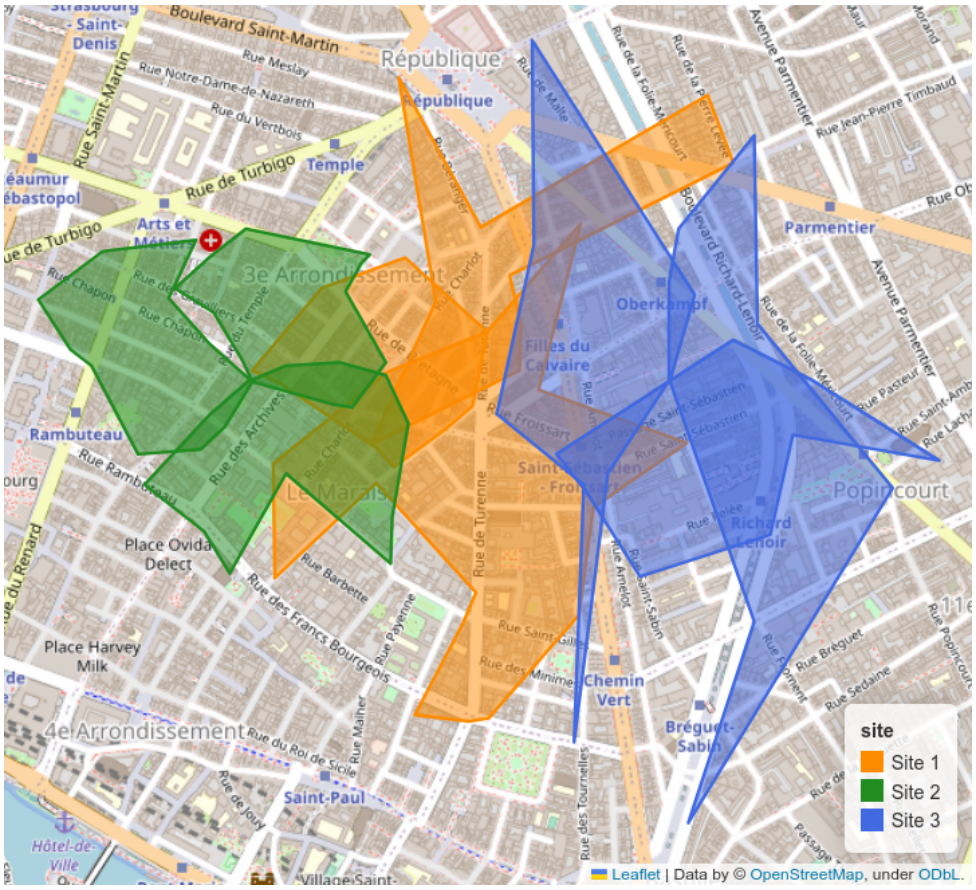


FIGURE 6.2 – Enveloppes concaves (χ-shapes, longueur de normalisation $\lambda_P = 0.67$) des positions d'utilisateurs rattachés à trois cellules tri-sectorisées (neuf cellules).

de bâtiments, ces infrastructures formant une structure très dense et continue. Le degré d'urbanisation est suffisamment élevé pour que l'on suppose que les données collectées sont uniformément distribuées dans l'espace. La Figure 6.2 montre les formes approximatives (enveloppes concaves) des distributions des mesures collectées pour trois sites tri-sectorisés. Ces mesures sont suffisamment abondantes pour couvrir les espaces entre les sites, en laissant peu de trous de couverture.

Par ailleurs, on s'attache à proposer des mesures pour comparer plusieurs modèles géométriques avec la réalité, mais aussi entre eux. Pour cela, on se sert des enveloppes convexes et concaves des localisations des utilisateurs attachés à la même cellule. Les enveloppes obtenues permettent d'avoir des bornes inférieures et supérieures pour les mesures utilisées. L'enveloppe convexe d'un ensemble est définie comme l'unique plus petit ensemble convexe contenant cet ensemble [137]. On utilise l'implémentation de PostGIS [87] qui utilise l'algorithme Graham scan [138, 139]. Contrairement à l'enveloppe convexe, une enveloppe concave n'est pas unique. Plusieurs algorithmes existent pour créer des enveloppes concaves selon des critères définis. On peut citer par exemple les α-*shapes* [140] et les χ-*shapes* [141]. Dans ce papier, on utilisera les χ-shapes pour leur facilité de paramétrage et d'interprétation.

Le modèle circulaire est une autre couverture géométrique classique, utilisée surtout dans les débuts de la planification [142]. L'idée est simplifier le problème en ramenant le diagramme de rayonnement de l'antenne à celui d'une antenne isotrope. Ensuite, on utilise les équations de propagation du signal pour déterminer le rayon de couverture limite de la station. Il existe différentes sortes de modèles de propagation, des modèles semi-empiriques utilisés depuis les déploiements du GSM jusqu'aux modèles analytiques développés récemment [143]. Ces derniers sont très complexes à intégrer dans des travaux de recherche en raison de la quantité des connaissances du terrain et de la radio nécessaires. Pour cette raison, on choisira d'utiliser les modèles semi-empiriques les plus courants : le modèle Hata [37], connu pour être fiable dans les zones urbaines en France [144], et les modèles UMa et RMa [38], adaptés pour des fréquences appartenant à l'intervalle $0.5 - 100$ GHz. Le rayon limite de la couverture est défini comme la distance à laquelle la puissance reçue par un équipement est égale à son seuil de sensibilité. La méthode, détaillée dans la Section 6.6.3 est basée sur les travaux de Marceau Coupechoux [145] pour l'établissement

de bilans de liaison.

Concernant la seconde étude, le rapport d'agrandissement des polygones de Voronoi est ajusté sur des rayons de couverture estimés à partir du *Timing Advance* (TA) [29, 146]. Le timing advance est une mesure discrétisée du temps mis par un signal pour voyager entre la station de base et l'utilisateur. Elle permet de synchroniser les transmissions entre deux équipements.

6.3 Problématiques de l'étude

Label	**Équipement**	**Géométrie**	**Division sectorielle ?**	**Stratégie de dimensionnement**	**Hyperparamètre**	**Questions abordées**
voronoi-site	Station de base	Voronoi	Non	Redimensionnement uniforme du diagramme	Rapport d'agrandissement dans $[0.3, 3]$	2.a, 5
voronoi-cell-nosplit	Cellule sectorisée	Voronoi	Non	Redimensionnement uniforme du diagramme	Rapport d'agrandissement dans $[0.3, 3]$	2.b, 5
voronoi-cell	Cellule sectorisée	Voronoi	Oui	Redimensionnement uniforme du diagramme	Rapport d'agrandissement dans $[0.3, 3]$	1.a, 2, 4.a, 5
uma/rma+circle	Cellule sectorisée	Circulaire	Oui	Rapport d'agrandissement dérivé de UMa/RMa	n_{PRB}	1.b, 4.b.ii, 5
uma/rma+voronoi	Cellule sectorisée	Voronoi	Oui	Rapport d'agrandissement dérivé de UMa/RMa	n_{PRB}	1a, 4.b.ii, 5
hata+circle	Cellule sectorisée	Circulaire	Oui	Rapport d'agrandissement dérivé de Hata	n_{PRB}	1b, 4.b.i, 5
hata+voronoi	Cellule sectorisée	Voronoi	Oui	Rapport d'agrandissement dérivé de Hata	n_{PRB}	1a, 4.b.i, 5
voronoi+circle	Cellule sectorisée	Circulaire	Oui	Redimensionnement du cercle aux dimensions de Voronoi	Rapport d'agrandissement dans $[0.3, 3]$	1.b, 4.c, 5
convexhull	Cellule sectorisée	Enveloppe convexe	Oui	-	-	3, 5
concavehull	Cellule sectorisée	Enveloppe concave	Oui	-	-	3, 5
voronoi-cell	Cellule sectorisée	Voronoi	Oui	Rapport d'agrandissement ajusté avec des données cellulaires agrégées	Rapport d'agrandissement $s^* \in \mathbb{R}$	6

TABLE 6.2 – Résumé des méthodes appliquées et des problèmes abordés

L'un des buts du chapitre est d'évaluer quelle géométrie (parmi celles présentées dans la Section 6.5) est la plus proche de la couverture réelle d'une cellule sectorisée. Contrairement à une carte des *best servers* qui partitionne la zone en régions telle que l'utilisateur soit rattaché à la cellule assurant la meilleure qualité de service, les travaux présents cherchent à déterminer si un utilisateur peut être connecté à une cellule indépendemment de l'état de la connexion. Par conséquent, il est fortement probable que la carte de couverture obtenue présente des superpositions entre plusieurs cellules, du fait de mécanismes variés comme le *handover*. On propose de redimensionner les formes géométriques à l'aide de différentes stratégies détaillées dans la Section 6.6 pour évaluer comment l'agrandissement ou la réduction des formes peut impacter la précision des modèles.

Afin de comparer les modèles géométriques, on utilise les mesures de précision et de rappel (Section 6.4) utilisées en classification binaire. On définit une borne inférieure et supérieure de la précision admissible en utilisant la précision des enveloppes concaves et convexes des positions des utilisateurs par cellule d'attachement (Section 6.5.4).

Dans la Section 6.7, on détaille la méthode utilisée pour trouver le rapport d'agrandissement tel que les cellules de Voronoi soient à la même échelle que le rayon limite d'une cellule estimé avec le timing advance.

Pour résumer les problématiques abordées, ce chapitre se structure autour des questions suivantes :

1. Quelle est la meilleure forme géométrique modélisant la couverture cellulaire : (a) les polygones issus du diagramme de Voronoi, (b) les secteurs circulaires ?
2. (a) Comment la précision et le rappel varient-ils entre les modélisations de la couverture de la station de base et celles de la couverture cellulaire ? (b) À l'échelle de la cellule, que gagne-t-on à diviser la couverture de la station de base ?
3. Comment construire des mesures et définir des valeurs de référence pour comparer les modèles ?
4. Est-il préférable de redimensionner les géométries : (a) pour les polygones de Voronoi, avec un rapport d'agrandissement unique pour toutes les cellules, (b) pour la géométrie circulaire, avec un rapport d'agrandissement le ramenant aux dimensions du polygone de Voronoi correspondant, (c) avec un rapport d'agrandissement différent pour chaque cellule, dérivé d'une formule de propagation comme (i) Hata, (ii) UMa/RMa ?
5. Est-ce que les résultats varient significativement à travers les bandes de fréquence ?
6. Quelles performances obtient-on en ajustant la taille des polygones de Voronoi grâce aux rayons limites de couverture estimées avec le timing advance ?

Pour répondre à ces questions, plusieurs combinaisons de géométries et de stratégies de redimensionnement ont été testées. Elles sont renseignées dans la Table 6.2.

Note : La troisième et la dernière ligne de la table partagent le même label car on utilise le même modèle géométrique, mais pour des études différentes. Les conclusions sont données en interprétant les mesures moyennées sur l'ensemble des cellules échantillonnées.

6.4 Mesures utilisées

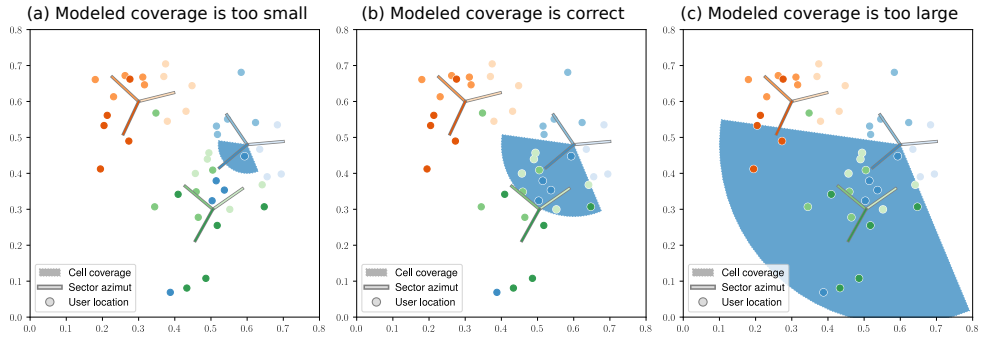


FIGURE 6.3 – Évaluation du modèle de couverture d'une cellule (en bleu) à plusieurs échelles. (a) précision=1, rappel=0.2 ; (b) : précision=0.5, rappel=0.8 ; (c) : précision=0.23, rappel=1

6.4.1 Notation

Pour une combinaison de géométrie et de stratégie de redimensionnement, le modèle de couverture est évalué en mesurant la précision et le rappel moyen obtenus sur les données du terrain. On calcule la précision et le rappel des modèles pour chaque bande de fréquence (les données terrains sont aussi partitionnées par fréquence).

Soit $\mathcal{C} \subset \mathbb{N}$ l'ensemble des identifiants des cellules venant des données du terrain, $\mathcal{M} \subset \mathbb{R}^2$ l'ensemble des positions d'utilisateurs rattachés aux cellules appartenant à $\mathcal{C}$, et $\mathcal{F} \subset \mathbb{N}$ l'ensemble des fréquences pouvant être déployées sur une station de base.

Pour $f \in \mathcal{F}$, on définit l'ensemble $\mathcal{C}_f \subset \mathcal{C}$ qui contient toutes les cellules transmettant à la fréquence f, et $\mathcal{M}_f \subset \mathcal{M}$ le sous-ensemble des positions utilisateurs rattachés aux cellules appartenant à $\mathcal{C}_f$. Pour une cellule $c \in \mathcal{C}_f$, on définit $M_c \subset \mathcal{M}_f$ comme l'ensemble des positions utilisateurs rattachés à c. Enfin, on désigne par $\mathrm{Cov}_{\mathrm{cell}}(c) \subset \mathbb{R}^2$ la région couverte par une cellule telle que modélisée par un modèle géométrique.

6.4.2 Précision et Rappel

Pour évaluer l'adéquation entre un modèle de couverture et la distribution des données collectées, on prend les moyennes de la précision et du rappel mesurés sur chaque cellule. La précision et le rappel sont des mesures classiques de la classification binaire.

Soient $f \in \mathcal{F}, c \in \mathcal{C}_f$. La précision est définie comme le ratio entre le nombre de vrais positifs et la somme des vrais et faux positifs. Les vrais positifs sont les emplacements de M_c géographiquement contenus dans $\mathrm{Cov}_{\mathrm{cell}}(c)$. Les faux positifs sont les positions contenues dans $\mathrm{Cov}_{\mathrm{cell}}(c)$ mais qui n'appartiennent pas à M_c. En termes mathématiques :

$$\text{précision}(c) = \frac{|\{Q \in \mathcal{M}_f | Q \in M_c \cap \mathrm{Cov}_{\mathrm{cell}}(c)\}|}{|\{Q \in \mathcal{M}_f | Q \in \mathrm{Cov}_{\mathrm{cell}}(c)\}|}$$

La précision moyennée sur toutes les cellules de $\mathcal{C}_f$ est :

$$\overline{P} = \frac{1}{|\mathcal{C}_f|} \sum_{c \in \mathcal{C}_f} \text{précision}(c)$$

Le rappel est défini comme le ratio entre le nombre de vrais positifs et la somme des vrais positifs et faux négatifs. Les faux négatifs correspondent aux emplacements de M_c qui ne sont pas contenus dans $\mathrm{Cov}_{\mathrm{cell}}(c)$.

$$\text{rappel}(c) = \frac{|\{Q \in \mathcal{M}_f | Q \in M_c \cap \mathrm{Cov}_{\mathrm{cell}}(c)\}|}{|\{Q \in \mathcal{M}_f | Q \in M_c\}|}$$

Le rappel moyenné sur toutes les cellules de $\mathcal{C}_f$ est :

$$\overline{R} = \frac{1}{|\mathcal{C}_f|} \sum_{c \in \mathcal{C}_f} \text{rappel}(c)$$

Comme il est souhaitable d'avoir des modèles dont les géométries se superposent au mieux avec les ensembles de positions M_c, le rappel est la mesure la plus importante à maximiser (Figure 6.3 (b)). Cependant, l'utilisation seule du rappel peut mener à choisir un modèle surestimant de loin la couverture réelle. En effet, plus la forme est grande, plus elle est susceptible de couvrir toutes des positions de M_c et de n'avoir que des vrais positifs (Figure 6.3 (c)). Une précision trop faible par rapport à un seuil de référence pourrait aider à détecter ce problème.

La précision permet aussi de détecter les formes sous-dimensionnées car cette valeur est plus élevée pour des géométries plus petites (Figure 6.3 (a)). Les modèles géométriques qui ont une précision trop élevée par rapport à un seuil limite peuvent aussi être écartés.

Pour fixer une borne de précision minimale et maximale raisonnable, on propose d'utiliser les précisions des enveloppes convexes et concaves de M_c (Section 6.5.4).

Référence voronoi-site

On envisage qu'une perte de précision et de rappel puisse survenir en passant de l'échelle de la station de base à celle de la cellule. Pour cette raison, en guise de référence, on calcule la précision et le rappel obtenus par le diagramme de Voronoi. Ces mesures s'expriment avec des ensembles de vrais positifs, faux positifs et faux négatifs différents. Soit $\mathcal{B}$ l'ensemble des identifiants des stations de base. Pour une station de base d'identifiant $b \in \mathcal{B}$, une fréquence $f \in \mathcal{F}$, $M'_{f,b}$ est l'ensemble des positions des utilisateurs rattachés à une cellule de fréquence f appartenant à b. On y associe $\mathrm{Cov}_{\mathrm{BS}}(b)$ la région modélisant la couverture de la station de base (ici, un polygone de Voronoi).

La précision et le rappel d'une station de base sont donnés par :

$$\text{précision}'(b) = \frac{|\{Q \in \mathcal{M}_f | Q \in M'_{f,b} \cap \mathrm{Cov}_{\mathrm{BS}}(b)\}|}{|\{Q \in \mathcal{M}_f | Q \in \mathrm{Cov}_{\mathrm{BS}}(b)\}|}$$

$$\text{rappel}'(s) = \frac{|\{Q \in \mathbb{R}^2 | Q \in M'_{f,b} \cap \text{Cov}_{\text{BS}}(b)\}|}{|\{Q \in \mathbb{R}^2 | Q \in M'_{f,b}\}|}$$

On moyenne ensuite la précision et le rappel sur l'ensemble $\mathcal{B}$ restreinte à la fréquence f.

Référence voronoi-cell-nosplit

On mesure également le gain supposé de diviser la couverture des stations de base en sous-régions. Pour cela, on définit un modèle de référence qui considère que la couverture cellulaire est égale à la couverture de la station de base.

Soit $f \in \mathcal{F}, c \in \mathcal{C}_f$ et $b \in \mathcal{B}$ la station de base où c est déployée. Les expressions de précision et de rappel sont celles de précision(c) et rappel(c), à la différence que $\text{Cov}_{\text{cell}}(c) = \text{Cov}_{\text{BS}}(b)$, puisqu'on ne divise pas la couverture de b.

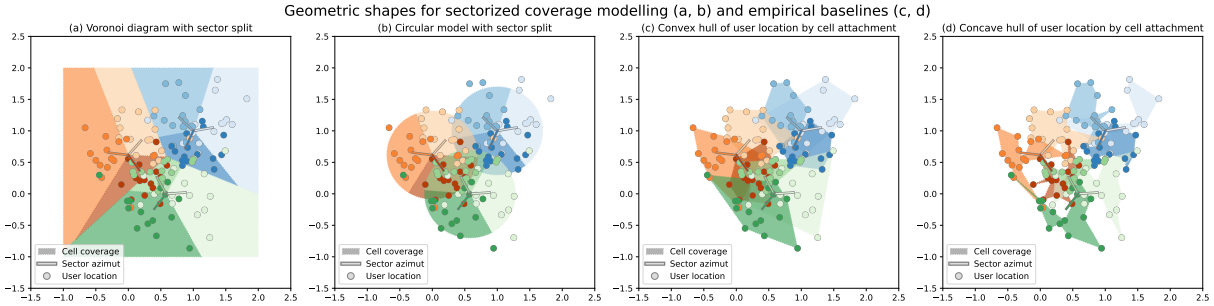


FIGURE 6.4 – Les Figures (a) and (b) illustrent les formes géométriques qui seront évalués à différentes échelles dans la suite du chapitre. Les Figures (c) and (d) illustrent les enveloppes convexe et concave des positions des utilisateurs par cellule de rattachement. À titre illustratif, la configuration topographique des sites et les positions des utilisateurs sont générées aléatoirement, et peuvent ne pas correspondre exactement à la distribution réelle des sites et des utilisateurs.

6.5 Formes géométriques utilisées

Pour obtenir les modèles de couverture cellulaire, on commence par modéliser la couverture de leur station de base. Dans l'étude, celle-ci est soit un polygone du diagramme de Voronoi, soit un cercle. Ensuite, on divise la couverture du site en sous-régions, une par secteur en fonction de l'orientation azimutale des antennes. La Figure 6.4(a) illustre le modèle cellulaire basé sur Voronoi et la Figure 6.4(b) le modèle basé sur le cercle.

Les Figures 6.4(c) et 6.4(d) illustrent respectivement les enveloppes convexes et concaves des positions utilisateurs par cellule d'attachement. Leur utilisation est développée dans la Section 6.5.4.

6.5.1 Modèle de Voronoi pour les stations de base

Soit $n \in \mathbb{N}$ le nombre de stations de base, $S = \{P_1, ..., P_n\} \subset \mathbb{R}^2$ l'ensemble des points correspondant à leurs coordonnées de localisation.

La région de Voronoi d'une station de base située au point P est :

$$\text{Vor}(P) = \{Q \in \mathbb{R}^2 | \forall P' \in S \;\; ||\text{QP}|| \leq ||\text{QP}'||\}$$

où $||\text{QP}||$ est la distance euclidienne entre Q et P.

6.5.2 Modèle de cercle pour les stations de base

Soit P le point d'emplacement d'une station de base. Le modèle de cercle, de rayon R_l, définit la région de couverture :

$$D(P) = \{Q \in \mathbb{R}^2 | ||\text{QP}|| \leq R_l\}$$

Le calcul du rayon limite R_l est expliqué dans la Section 6.6.3.

6.5.3 Division de la couverture des stations de base en secteurs

En réutilisant les notations adoptées précédemment, on divise la couverture des stations de base en autant de sous-régions qu'il y a de secteurs déployés sur le site. Pour illustrer les explications qui suivent, la Figure 6.5 reprend les notations utilisées pour diviser un polygone de Voronoi en couverture cellulaire.

La carte de la couverture cellulaire peut varier d'une fréquence à une autre, car toutes les fréquences ne sont pas présentes sur tous les sites, voire pour un site donné, sur tous les secteurs. Pour cette raison, la division de la couverture des stations de base est réalisée fréquence par fréquence. Soit $f \in \mathcal{F}$ une fréquence déployée sur une station de base b située au point P, munie d'un modèle de couverture $\mathrm{Cov}_{\mathrm{BS}}(b)$. On adopte un repère orthonormal muni d'une base $(\vec{\mathbf{PX}}; \vec{\mathbf{PY}})$ (avec donc P pour point d'origine).

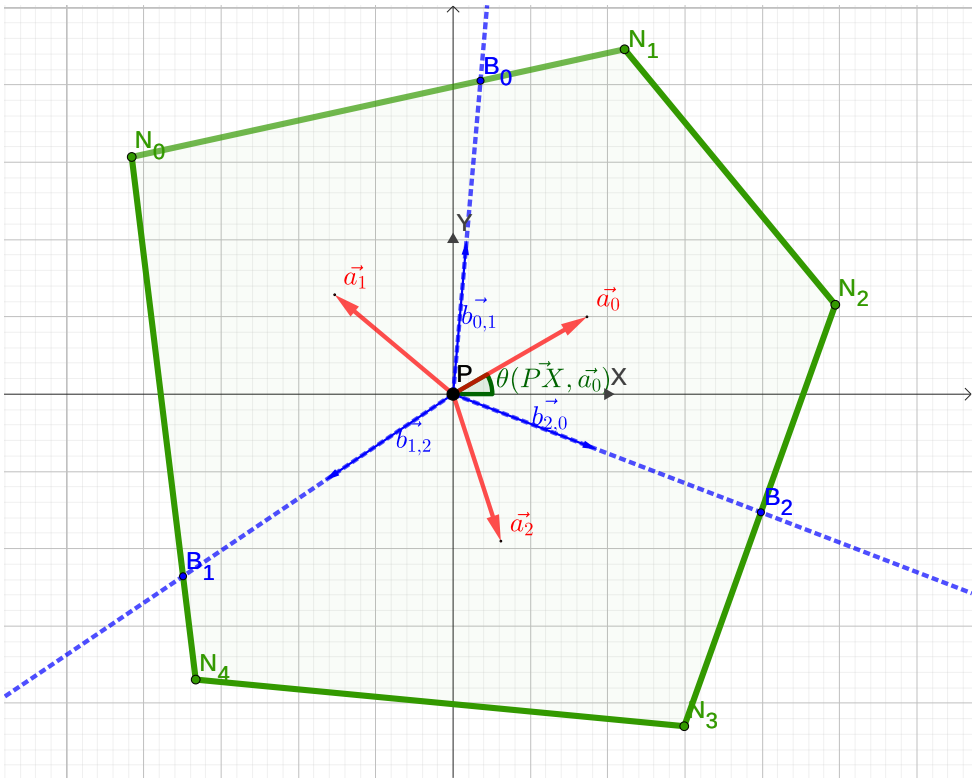


FIGURE 6.5 – Exemple d'un polygone de Voronoi divisé en trois régions sectorielles. $\mathrm{Cov}_{\mathrm{BS}}(b)$ est le polygone $N_0N_1N_2N_3N_4N_0$, et $\mathrm{Cov}_{\mathrm{cell}}(c_0)$ est le polygone $PB_0N_1N_2B_2P$.

Pour tous vecteurs $\vec{\mathbf{u}}, \vec{\mathbf{v}} \in \mathbb{R}^2$, on adopte la notation suivante :

— $\widehat{(\vec{\mathbf{u}}, \vec{\mathbf{v}})}$ est l'angle géométrique formé par deux vecteurs avec P comme sommet.

— $\theta(\vec{\mathbf{u}}, \vec{\mathbf{v}})$ est la mesure de l'angle $\widehat{(\vec{\mathbf{u}}, \vec{\mathbf{v}})}$ en radians. Les angles sont mesurés dans le sens trigonométrique et dans l'intervalle $[0, 2\pi[$.

— $(P, \vec{\mathbf{u}})$ est la demi-droite de direction $\vec{\mathbf{u}}$, avec pour origine P.

Soit $\mathcal{A}_{P,b} = \{\vec{\mathbf{a}}_0, ..., \vec{\mathbf{a}}_{m-1}\} \subset \mathbb{R}^2$ l'ensemble des vecteurs représentant les directions de transmission des cellules de la station P à la fréquence b. L'ensemble $\mathcal{A}_{P,b}$ est ordonné tel que :

$$\theta(\vec{\mathbf{PX}}, \vec{\mathbf{a}}_0) < ... < \theta(\vec{\mathbf{PX}}, \vec{\mathbf{a}}_{m-1})$$

La région $\mathrm{Cov}_{\mathrm{BS}}(b)$ est partitionnée en m sous-régions $\{\mathrm{Cov}_{\mathrm{cell}}(c_0), ..., \mathrm{Cov}_{\mathrm{cell}}(c_{m-1})\}$.

Soit $\vec{\mathbf{b}}_{k,k+1}$ le vecteur directeur tel que $(P, \vec{\mathbf{b}}_{k,k+1})$ est la bissectrice de $\widehat{(\vec{\mathbf{a}_k}, \vec{\mathbf{a}_{k+1}})}$.

Soit $B_{k,k+1}$ (respectivement $B_{k-1,k}$) l'intersection de $(P, \vec{\mathbf{b}}_{k,k+1})$ (respectivement $(P, \vec{\mathbf{b}}_{k-1,k})$) avec la frontière de $\mathrm{Cov}_{\mathrm{BS}}(b)$, que l'on note $\partial\mathrm{Cov}_{\mathrm{BS}}(b)$.

Soit c_k l'identifiant de la cellule déployée sur le k^e secteur, $\mathrm{Cov}_{\mathrm{cell}}(c_k) \subset \mathrm{Cov}_{\mathrm{BS}}(b)$ la région couverte par c_k de frontière :

$$\partial\mathrm{Cov}_{\mathrm{cell}}(c_k) = \mathrm{PB}_{k-1,k} \cup \mathrm{PB}_{k,k+1} \cup \{Q \in \partial\mathrm{Cov}_{\mathrm{BS}}(b) | \theta(\vec{\mathbf{PX}}, \vec{\mathbf{b}}_{k-1,k}) \leq \theta(\vec{\mathbf{PX}}, \vec{\mathbf{PQ}}) \leq \theta(\vec{\mathbf{PX}}, \vec{\mathbf{b}}_{k,k+1})\}$$

Par simplicité de notation, on a omis le terme modulo m, qui s'applique lorsque $k = 0$ (i.e. $k-1 \equiv m-1 \pmod m$) et $k = m-1$ (i.e. $k+1 \equiv 0 \pmod m$).

Remarque On peut donner quelques éléments pour montrer que la division en secteur opérée revient à considérer qu'un utilisateur est rattaché à la cellule « la plus proche ».

On propose de définir l'emplacement d'une cellule c_k en définissant un point $A_k \in (P\vec{\mathbf{a}}_k)$ qui appartient à la demie droite d'origine P et de direction $\vec{\mathbf{a}}_k$. De plus, on conditionne tous les points $\{A_i\}_{i=[\![0,m-1]\!]}$ associés aux m secteurs du site à être équidistants de P. C'est-à-dire qu'ils vérifient la propriété suivante :

$$||\mathrm{PA}_0|| = ... = ||\mathrm{PA}_k|| = ... = ||\mathrm{PA}_{m-1}||$$

Tant que ces deux contraintes sont vérifiées, la valeur numérique de $||\mathrm{PA}_k||$ peut être choisie arbitrairement (on peut prendre 50 m, 100 m...). Si on applique le diagramme de Voronoi sur $\{A_i\}_{i=[\![0,m-1]\!]}$ et qu'on le restreint à la couverture de la station de base $\mathrm{Cov_{BS}}(b)$, on obtient m sous-régions qu'on note $\{\mathrm{Vor}(A_0)_{|P}, ..., \mathrm{Vor}(A_m)_{|P}\}$. L'expression de $\mathrm{Vor}(A_k)_{|P}$ est :

$$\mathrm{Vor}(A_k)_{|P} = \{Q \in \mathrm{Cov_{BS}}(b) | \forall A' \in \{A_i\}_{i=[\![0,m-1]\!]} \ \ ||\mathrm{QA}|| \leq ||\mathrm{QA'}||\}$$

Cas particulier n°1 Lorsque $m = 1$, il n'y a qu'un seul secteur, donc on ne réalise pas de division sectorielle, et

$$\mathrm{Vor}(A_0)_{|P} = \mathrm{Cov_{BS}}(b)$$

Cas particulier n°2 Pour $m = 2$, on prolonge la division portée par la demi-droite $(P, \vec{\mathbf{b}}_{0,1})$ en une droite de même direction. On démontre ensuite que les deux sous-régions résultantes correspondent aux régions de Voronoi de A_0 et A_1 en suivant le même raisonnement que pour $m \geq 3$

Autres cas Pour $m \geq 3$ (sites tri-sectorisés ou plus), comme tous les points de $\{A_i\}_{i=[\![0,m-1]\!]}$ sont situés à égale distance de P, par définition P appartient à toutes les régions. On admet que cela a pour conséquence que pour tout $k \in [\![0, m-1]\!]$, la frontière partagée par A_k avec A_{k-1} (respectivement A_{k+1}) est un segment dont l'une des extrémités est P. Comme les points de l'ensemble sont ordonnés par mesure d'angles, A_k ne partage pas de frontière avec les points d'index non successifs.

En notant $L_k = \mathrm{PB}_{k-1,k} \cup \mathrm{PB}_{k,k+1}$ et

$$F_k = \{Q \in \partial\mathrm{Cov_{BS}}(b) | \theta(\vec{\mathbf{PX}}, \vec{\mathbf{b}}_{k-1,k}) \leq \theta(\vec{\mathbf{PX}}, \vec{\mathbf{PQ}}) \leq \theta(\vec{\mathbf{PX}}, \vec{\mathbf{b}}_{k,k+1})\}$$

on veut montrer que :

$$\partial\mathrm{Vor}(A_k)_{|P} = L_k \cup F_k = \partial\mathrm{Cov_{cell}}(c_k)$$

Pour cela, on décompose la frontière de $\mathrm{Vor}(A_k)_{|P}$ en deux ensembles λ_k et ϕ_k tels que :

$$\partial\mathrm{Vor}(A_k)_{|P} = \lambda_k \cup \phi_k$$

Où :
— λ_k est l'ensemble des frontières de $\mathrm{Vor}(A_k)$ partagées avec $\mathrm{Vor}(A_{k-1})$ et $\mathrm{Vor}(A_k)$,
— $\phi_k = \partial\mathrm{Cov_{BS}}(b) \cap \mathrm{Vor}(A_k)_{|P}$ est la partie de la frontière de $\mathrm{Cov_{BS}}(b)$ dans la région de A_k

En utilisant la définition du diagramme de Voronoi, l'ensemble λ_k est défini par :

$$\begin{aligned}\lambda_k &= \{Q \in \mathrm{Vor}(A_k)_{|P} \ | \ ||\mathrm{QA}_k|| = ||\mathrm{QA}_{k-1}||\} \cup \{Q \in \mathrm{Vor}(A_k)_{|P} \ | \ ||\mathrm{QA}_k|| = ||\mathrm{QA}_{k+1}||\} \\ &= [\mathrm{P}\beta_{k-1,k}] \cup [\mathrm{P}\beta_{k,k+1}]\end{aligned}$$

où $\beta_{k-1,k} \in \mathbb{R}^2$ (respectivement $\beta_{k,k+1}$) est le point d'intersection du segment séparant $\mathrm{Vor}(A_k)_{|P}$ de $\mathrm{Vor}(A_{k-1})_{|P}$ (respectivement $\mathrm{Vor}(A_{k+1})_{|P}$) avec la frontière de $\mathrm{Cov_{BS}}(b)$. Autrement dit :

$$\lambda_k \cap \partial\mathrm{Cov_{BS}}(b) = \{\beta_{k-1,k}, \beta_{k,k+1}\}$$

On peut également exprimer ϕ_k en utilisant les points $\beta_{k-1,k}$ et $\beta_{k,k+1}$:

$$\phi_k = \{Q \in \partial\mathrm{Cov}_{\mathrm{BS}}(b) | \theta(\vec{\mathbf{PX}}, \vec{\beta}_{k-1,k}) < \theta(\vec{\mathbf{PX}}, \vec{\mathbf{PQ}}) < \theta(\vec{\mathbf{PX}}, \vec{\beta}_{k,k+1})\}$$

Pour montrer que $\lambda_k = L_k$ et $\phi_k = F_k$, on doit montrer que $\beta_{k-1,k} = B_{k-1,k}$ et $\beta_{k,k+1} = B_{k,k+1}$.

Par définition des $\{A_i\}_{i=[\![0,m-1]\!]}$, $||\mathrm{PA}_k|| = ||\mathrm{PA}_{k-1}||$ donc $\mathrm{A}_{k-1}\mathrm{PA}_k$ est un triangle isocèle en P. Par conséquent, le segment bissecteur $[\mathrm{PB}_{k-1,k}]$ est aussi la médiatrice de $[\mathrm{A}_k\mathrm{A}_{k-1}]$. Donc, $\forall Q \in [\mathrm{PB}_{k-1,k}]$, $||\mathrm{QA}_k|| = ||\mathrm{QA}_{k-1}||$, et $[\mathrm{PB}_{k-1,k}] \subset [\mathrm{P}\beta_{k-1,k}] \subset \lambda_k$. Or, on a aussi $B_{k-1,k} \in \partial\mathrm{Cov}_{\mathrm{BS}}(b)$, donc $B_{k-1,k} \in \lambda_k \cap \partial\mathrm{Vor}_{\mathcal{S}}(P)$, et $B_{k-1,k} = \beta_{k-1,k}$

De la même façon, on montre avec le triangle $\mathrm{A}_k\mathrm{PA}_{k+1}$ que $B_{k,k+1} = \beta_{k,k+1}$.

Il en résulte que $L_k = \lambda_k$ and $F_k = \phi_k$.

Conclusion : $\partial\mathrm{Vor}(A_k)_{|P} = L_k \cup F_k = \partial\mathrm{Cov}_{\mathrm{cell}}(c_k)$

6.5.4 Enveloppes empiriques de référence

En supposant que les données collectées sont représentatives de la distribution réelle des utilisateurs par cellule, la forme de la couverture réelle d'une cellule se situe quelque part entre les enveloppes convexes et concaves empiriques de ces positions.

Soit $c \in \mathcal{C}$ une cellule, et $M_c = \{m_1, ..., m_N\}$ l'ensemble de N positions d'utilisateurs attachés à c. Les enveloppes convexes et concaves de M_c sont des polygones qui contiennent tous les points de l'ensemble, et dont les sommets sont des points appartenant à M_c. Le rappel associé à ces enveloppes est donc toujours égal à 1. Grâce aux contraintes utilisées pour construire ces polygones, leurs précisions peuvent servir de valeurs de référence. Ces valeurs sont des seuils minimum et maximum de valeurs admissibles qui indiquent si une géométrie surestime ou sous-estime la couverture réelle.

Enveloppe convexe L'enveloppe convexe de M_c est l'ensemble formant le plus petit polygone convexe qui contient tous les points de M_c (Figure 6.4 (c)). Dans la géométrie en deux dimensions, une région est dite convexe si, pour toute paire de points appartenant à cette région, le segment joignant les points est également contenu dans la région.

On se sert de la nature convexe et minimale de cette enveloppe pour considérer que la précision atteinte est un seuil minimal. En effet, il s'agit aussi de la plus grande surface possible englobant M_c tel que les sommets de l'enveloppe appartiennent tous à M_c. Toute géométrie ayant une précision globale $\overline{P}$ plus faible que le seuil est une indication qu'elle est surdimensionnée.

Enveloppe concave Comme le nom l'indique, une enveloppe concave de M_c n'a pas de contrainte de convexité (Figure 6.4 (d)). L'objectif recherché en utilisant une enveloppe concave est de capturer la forme du nuage de points formé par M_c plus fidèlement que l'enveloppe convexe. Les χ-shapes sont une famille d'enveloppes concaves issues d'un algorithme basé sur l'érosion des arêtes de la triangulation de Delaunay [147] appliquée à M_c [141]. Cet algorithme est paramétré par la longueur de normalisation $\lambda_P \in [0, 1]$, où une valeur de 0 produit une enveloppe de concavité maximale, tandis qu'une valeur de 1 produit une enveloppe convexe (Figure 6.6).

La précision d'une enveloppe de type χ-shape obtenue avec $\lambda_P = 0$ est l'une des plus élevées possibles compte tenu de la très forte concavité de la forme. L'enveloppe est formée de quasiment tous les points (si ce n'est tous) de M_c. Par conséquent, on utilisera cette définition d'enveloppe concave pour établir une borne supérieure de la précision. Un modèle géométrique dont la précision moyenne $\overline{P}$ est plus élevée que la précision de χ-shape sous-estime probablement la couverture réelle.

6.6 Stratégies de dimensionnement

Les modèles de couverture sont redimensionnés pour comprendre quels rapports d'agrandissement doivent être utilisés pour que les formes géométriques se superposent au mieux avec les positions réelles. On redimensionne d'abord la couverture de la station de base avant de la diviser en secteurs.

Trois stratégies de redimensionnement sont évaluées, chacune contrôlée par un paramètre :

1. Pour les polygones de Voronoi, on redimensionne la couverture de la station de base en utilisant un rapport d'agrandissement unique (Figure 6.7 (b) and (c)) qui est le paramètre de cette méthode.

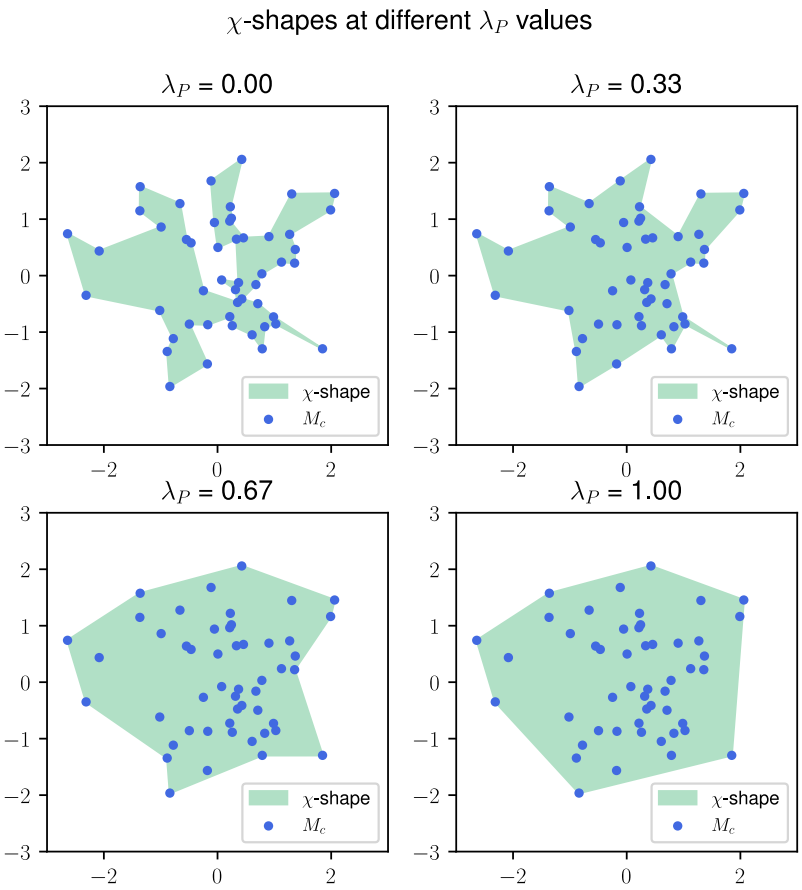


FIGURE 6.6 – Illustration de l'influence de la longueur de normalisation λ_P sur la concavité des χ-shapes.

2. Pour les cercles, on redimensionne chacun d'entre eux par un facteur différent, de telle sorte que le rayon final soit égal au rayon du plus petit cercle englobant le polygone de Voronoi correspondant. Pour évaluer différentes tailles de couverture, on redimensionne donc d'abord le diagramme de Voronoi avant de redimensionner les cercles correspondants. C'est pourquoi le paramètre de cette méthode est également le rapport d'agrandissement unique des polygones de Voronoi (Figure 6.7 (f)).
3. Pour les deux formes, on étudie aussi un redimensionnement non uniforme. Le facteur est choisi de façon à ce que le rayon du cercle englobant vérifie l'équation d'affaiblissement de propagation maximal. Dans cette méthode, le paramètre contrôlant l'affaiblissement de propagation maximal est le nombre de PRBs reçus par l'équipement utilisateur en bordure de cellule (6.7 (e)).

6.6.1 Redimensionnement uniforme de Voronoi

Soit $s \in \mathbb{R}$ un rapport d'agrandissement. Soit $b \in \mathcal{B}$ la station de base située à l'emplacement $P \in \mathbb{R}^2$ et de couverture modélisée $\mathrm{Cov}_{\mathrm{BS}}(b) \subset \mathbb{R}^2$. Tout point $Q \in \partial\mathrm{Cov}_{\mathrm{BS}}(b)$ situé à la frontière de la couverture est transformé en point Q' appartenant à la frontière de la couverture redimensionnée grâce à une fonction de redimensionnement f_s de la forme :

$$\begin{aligned} f_s : \mathbb{R}^2 &\longrightarrow \mathbb{R}^2 \\ Q &\longmapsto f_s(Q) = P + s(Q - P) = Q' \end{aligned}$$

La transformation est réalisée en prenant P comme origine. Tout rapport $s < 1$ réduit les formes (Figure 6.7 (b)), et tout $s > 1$ les agrandit (Figure 6.7 (c)). Dans nos expérimentations, l'intervalle de recherche du rapport optimal est $[0.3, 3]$

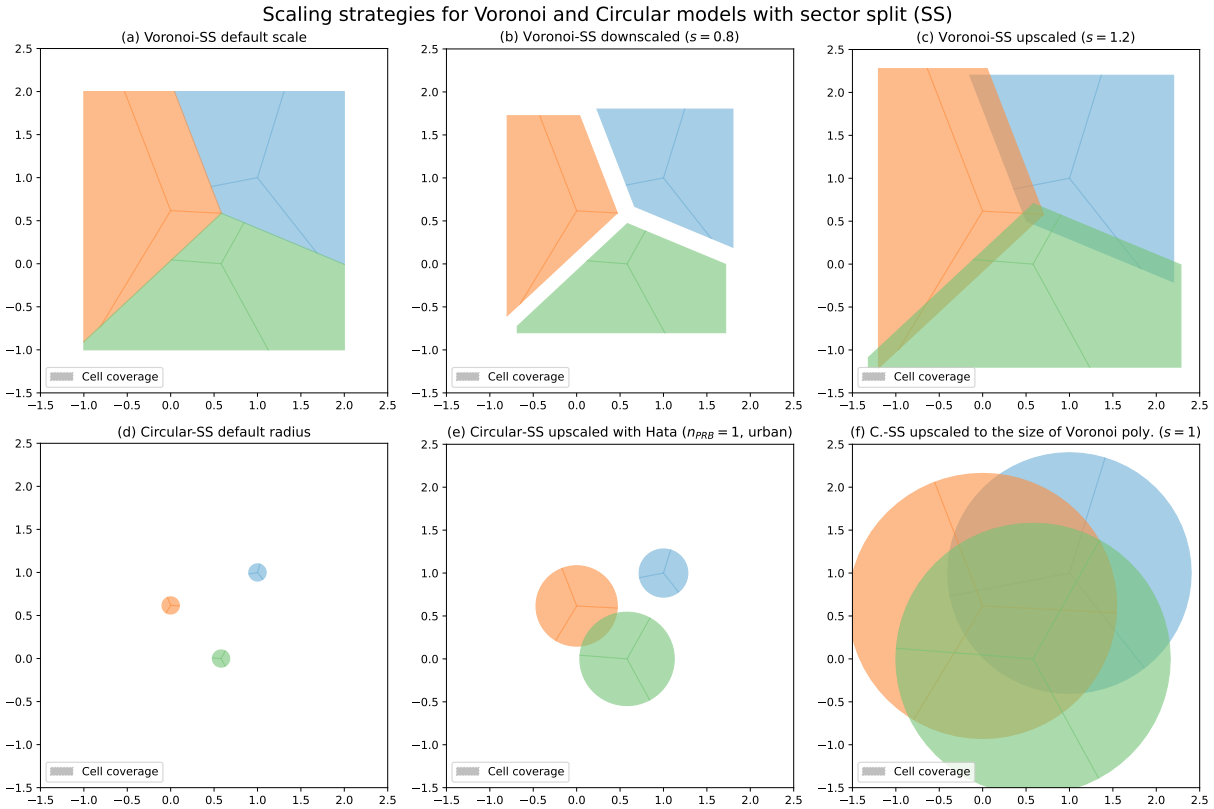


FIGURE 6.7 – Modèles géométriques redimensionnés selon les différentes stratégies (première ligne : Voronoi, deuxième ligne : cercle).

6.6.2 Redimensionnement de cercles dépendant du diagramme de Voronoi

Soit r_c le rayon par défaut du modèle circulaire. Pour une station de base d'identifiant b, soit r_v le rayon du plus petit cercle englobant le polygone de Voronoi associé à b (tel que tous les sommets du polygone soient à l'intérieur de celui-ci).

Pour évaluer la précision et le rappel à différentes échelles, on redimensionne d'abord les polygones de Voronoi uniformément en utilisant le rapport s et la fonction f_s. Le cercle est ensuite redimensionné de sorte que son rayon soit égal à sr_v.

Soit $s' \in \mathbb{R}$ le rapport d'agrandissement du cercle modélisant la couverture de b. L'expression de s' est :

$$s' = s\frac{r_v}{r_c}$$

Le rayon r'_c du cercle redimensionné obtenu par la relation :

$$r'_c = s'r_c = s\frac{r_v}{r_c}r_c = sr_v$$

ce qui est égal au rayon du plus petit cercle englobant le polygone de Voronoi redimensionné. La Figure 6.7 (f) donne un exemple de cercles redimensionnés à la taille des polygones du diagramme de Voronoi.

6.6.3 Rapport d'agrandissement dérivé des expressions de propagation du signal

Pour une station de base d'identifiant b avec un modèle de couverture $\mathrm{Cov}_{\mathrm{BS}}(b)$, soit r le rayon du plus petit cercle englobant $\mathrm{Cov}_{\mathrm{BS}}(b)$ (si le modèle est un cercle, il est confondu avec celui-ci). On note R_l le rayon limite tel que la puissance reçue par un équipement utilisateur est égale à son seuil de sensibilité.

Le rapport d'agrandissement s est donné par la relation :

$$s = \frac{R_l}{r}$$

On fait varier le seuil de sensibilité en changeant le nombre de PRBs n_{PRB} alloués à l'utilisateur en bordure de cellule. La Figure 6.7 (e) illustre le redimensionnement des cercles pour $n_{\mathrm{PRB}} = 1$. Dans la suite, on présente les expressions utilisées pour calculer R_l à partir de l'affaiblissement de propagation maximal. La plupart des notations et la méthode de calcul du bilan de liaison est emprunté aux cours de Marceau Coupechoux [145].

Affaiblissement de propagation maximal

Le bilan de liaison est calculé sur la contrainte que la puissance $P_r(R)$ reçue par un récepteur à une distance R doit être supérieure ou égale à sa sensibilité S. On suppose que la transmission est limitée par le sens montant, c'est-à-dire par le cas où l'équipement utilisateur est l'émetteur et la station de base est le récepteur. Pour un récepteur situé à la distance R de l'émetteur, la puissance reçue $P_r(R)$ est égale à la puissance d'émission P_t de l'antenne augmentée par son gain G_{UE} et soustraite de l'affaiblissement de propagation $\mathrm{PL}(R)$ et des marges M :

$$P_r(R) = P_t + G_{\mathrm{UE}} - \mathrm{PL}(R) - M \geq S$$

En réalité, le gain de l'antenne émettrice varie en fonction de la direction de transmission du signal. Dans cette étude, on ramène le diagramme de rayonnement d'une antenne à celui d'une antenne isotropique émettant à la puissance $\mathrm{EIRP} = P_t + G_{\mathrm{UE}}$ dans toutes les directions.

On réécrit l'équation précédente de la manière suivante :

$$P_r(R) = \mathrm{EIRP} - \mathrm{PL}(R) - M \geq S$$

L'affaiblissement de propagation maximal (MAPL : *maximum allowable path loss*) est la valeur telle que la puissance reçue par le récepteur est égale à la limite de sa sensibilité. Le rayon limite auquel le MAPL est atteint est noté R_l. Il vérifie les équations suivantes :

$$\begin{cases} \mathrm{PL}(R_l) = \mathrm{MAPL} \\ \mathrm{P}_r(R_l) = \mathrm{EIRP} - \mathrm{MAPL} - M = S \end{cases} \tag{6.1}$$

On peut réarranger les expressions précédentes en fonction de R_l, et prendre l'inverse de la fonction d'affaiblissement de propagation PL, qu'on note PL^{-1} :

$$\begin{cases} R_l = \mathrm{PL}^{-1}(\mathrm{MAPL}) \\ \mathrm{MAPL} = \mathrm{EIRP} - S - M \end{cases} \tag{6.2}$$

La Table 6.3 renseigne les valeurs d'EIRP, S, M ainsi que les modèles de propagation utilisés. Sauf mentionné explicitement, les unités de mesure des variables sont celles indiquées dans le tableau.

Attribution de contexte à une cellule

Les modèles de propagation se déclinent en plusieurs variantes en fonction du contexte de propagation, à savoir s'il est urbain, périurbain ou rural. On détermine ce contexte en se basant sur la situation générale de l'urbanisation de l'Île-de-France. On utilise par exemple les frontières administratives des départements pour séparer les zones urbaines des zones périurbaines, et la tranche d'unité urbaine [148] pour identifier les espaces ruraux.

La tranche d'unité urbaine est une donnée statistique fournie par l'INSEE. Elle se base sur le concept d'unité urbaine. L'unité urbaine est définie par l'INSEE comme « *une commune ou un ensemble de communes présentant une zone de bâti continu (pas de coupure de plus de 200 mètres entre deux constructions) qui compte au moins 2 000 habitants* » [149].

En fonction de leur population, les unités urbaines ainsi que les communes en faisant partie se voient attribuer une tranche. La tranche d'unité urbaine est une catégorie numérique comprise entre 0 et 8. Entre 0 et 7, plus le nombre est élevé, plus l'unité urbaine est peuplée. La tranche 8 correspond à l'unité urbaine de Paris. Les communes appartenant à la tranche 0 sont situées en dehors d'une unité urbaine. On considère ces zones comme étant rurales.

Les règles suivantes sont appliquées pour déterminer le contexte de propagation d'une cellule :

— si la cellule est localisée dans le département 75 (Paris), alors le contexte de propagation est urbain.
— si la cellule est déployée dans le département 77, 78, 91 ou 95, et si la tranche d'unité urbaine est 0, alors le contexte est rural.

— sinon, le contexte est de propagation est considéré comme périurbain.

	Unité	**Variable**	**Formule**	**Value**
Paramètres				
Fréquence	MHz	f	Paramètre	$\{700, 800, 1800, 2100, 2600\}$
Nombre de PRBs	-	n_{PRB}	Paramètre	[1 ;50]
Bande passante	MHz	W	$0.180 \times n_{\text{PRB}}$	[0.180 ;9]
Hauteur de la station de base	m	h_{BS}	Donnée de configuration réseau	[1.3, 98.3]
Hauteur équipement utilisateur	m	h_{UE}	Valeur typique	1.7
Température	K	T	Valeur typique	290
Constante de Boltzmann	$\text{J}\cdot\text{K}^{-1}$	k_B	Constante	1.38×10^{-23}
Densité du bruit	dBm/Hz	N_d	$10\log_{10}(k_B T \times 10^3)$	-174
Probabilité de couverture	-	P_{cov}	Valeur typique	0.95
Fonction quantile de la distribution normale	-	Q	$\sqrt{2}\text{erf}^{-1}(2p-1)$	-
Contexte	-	-	Value définie à partir des topographies réelles	{Urbain, Périurbain, Rural}
Écart-type effet de masque	dB	σ_{SF}	Valeurs typiques	8 si Contexte = Urbain 7 si Contexte = Périurbain 6 si Contexte = Rural
Transmission				
Puissance	dBm	P_t	Valeur typique	23
Gain d'antenne	dBi	G_{UE}	Valeur typique	0
EIRP	dBm	EIRP	$P_t + G_{\text{UE}}$	23
Réception				
SINR cible	dB	SINR	Valeur typique	-6
Puissance du bruit	dBm	N_{pow}	$N_d + 10\log_{10}(W \times 10^6)$	[-121.45 ; -104.46]
Facteur de bruit	dB	N_{fig}	Valeur typique	5
Gain d'antenne	dBi	G_{BS}	Valeurs typiques	16 si $f \in \{700, 800\}$ 18 si $f \in \{1800, 2100\}$ 19 si $f = 2600$
Pertes câble	dB	L_{cable}	Valeur typique	2
Sensibilité	dBm	S	SINR + N_{pow} + N_{fig} − $G_{\text{BS}} + L_{\text{cable}}$	[-139.45, -119.46]
Marges				
Interférence	dB	M_I	Valeur typique	3
Effet de masque	dB	M_{SF}	Formule de Jakes : $\sigma_{\text{SF}} Q(P_{\text{cov}})$	[9.84; 13.12]
Corps humain	dB	M_{body}	Valeur typique	1
Pénétration bâtiments	dB	M_{building}	Valeurs typiques	18 si Contexte = Urbain 15 si Contexte = Périurbain 12 si Contexte = Rural
Total marges	dB	M	M_I + M_{SF} + M_{body} + M_{building}	[25.84 ; 35.12]
Rayon de cellule				
Affaiblissement de propagation maximal	dB	MAPL	EIRP − S − M	[107.34 ; 136.61]
Modèle de propagation	-	$PL \in$ {H, UMa, RMa}	{Hata, UMa, RMa}	-
Rayon	m	R_l	PL^{-1}(MAPL)	[250; 13000] (Voir Section 6.8.)

TABLE 6.3 – Bilan de liaison dans le sens montant

Modèle Hata

Le modèle de propagation Hata se décline en trois variantes : urbain, périurbain et rural. En utilisant les notations de variable de la Table 6.3, on inverse les formules de Hata pour obtenir des expressions qui nous permettent de calculer R_l.

Urbain Comme l'Île-de-France est une région fortement urbanisée et que les fréquences étudiées sont supérieures à 400 MHz, le facteur de correction de la hauteur d'antenne $a(h_{\mathrm{UE}})$ est égal à :

$$a(h_{\mathrm{UE}}) = 3.2[\log_{10}(11.75h_{\mathrm{UE}})]^2 - 4.97$$

Avec le rayon limite R_l exprimé en kilomètres et le modèle de propagation Hata noté H_{U}, l'équation 6.1 devient :

$$\begin{aligned}\mathrm{MAPL} &= \mathrm{H_U}(R_l) \\ &= 69.55 + 26.16\log_{10}(f) - 13.82\log_{10}(h_{\mathrm{BS}}) - \mathrm{a}(h_{\mathrm{UE}}) + [44.9 - 6.55\log_{10}(h_{\mathrm{BS}})]\log_{10}(R_l)\end{aligned}$$

En inversant H_{U}, on obtient :

$$\begin{aligned}\mathrm{H_U^{-1}(MAPL)} &= R_l \\ &= 10^{\frac{\mathrm{MAPL} - 69.55 + 13.82\log_{10}(h_{\mathrm{BS}}) + \mathrm{a}(h_{\mathrm{UE}}) - 26.16\log_{10}(f)}{44.9 - 6.55\log_{10}(h_{\mathrm{BS}})}}\end{aligned} \tag{6.3}$$

Périurbain Le modèle périurbain de Hata H_{SU} est formulé de la façon suivante :

$$\mathrm{H_{SU}}(R_l) = \mathrm{H_U}(R_l) - 2\left[\log_{10}\left(\frac{f}{28}\right)\right]^2 - 5.4$$

En inversant H_{SU} :

$$\begin{aligned}\mathrm{H_{SU}^{-1}(MAPL)} &= R_l \\ &= \mathrm{H_U^{-1}(MAPL)} \times 10^{\frac{5.4 + 2\left[\log_{10}\left(\frac{f}{28}\right)\right]^2}{44.9 - 6.55\log_{10}(h_{\mathrm{BS}})}}\end{aligned} \tag{6.4}$$

Rural Le modèle rural quasi-open de Hata H_{RU} est formulé de la façon suivante :

$$\mathrm{H_{RU}}(R_l) = \mathrm{H_U}(R_l) - 4.78[\log_{10}(f)]^2 + 18.33\log_{10}(f) - 35.94$$

En inversant H_{RU} :

$$\begin{aligned}\mathrm{H_{RU}^{-1}(MAPL)} &= R_l \\ &= \mathrm{H_U^{-1}(MAPL)} \times 10^{\frac{35.94 - 18.33\log_{10}(f) + 4.78[\log_{10}(f)]^2}{44.9 - 6.55\log_{10}(h_{\mathrm{BS}})}}\end{aligned} \tag{6.5}$$

Modèles 0.5-100 GHz

Le standard 3GPP définit deux modèles de propagation pour les macrocellules : Urban Macro (UMa) et Rural Macro (RMa). On utilise le modèle UMa pour les cellules urbaines et périurbain et, et le modèle RMa pour les cellules rurales. Au lieu d'exprimer l'affaiblissement de propagation comme une fonction de la distance entre la station de base et l'utilisateur (R_l), les modèles UMa et RMa sont exprimés en fonction de la distance d_{3D} entre le sommet de la station et l'utilisateur.

En appliquant le théorème de Pythagore, la relation entre R_l, d_{3D} et h_{BS} est :

$$R_l = \sqrt{d_{\mathrm{3D}}^2 - h_{\mathrm{BS}}^2} \tag{6.6}$$

Urban Macro (UMa) On considère un scénario pessimiste où il n'est pas possible d'avoir une propagation du signal en ligne de vue à la distance du rayon limite. L'affaiblissement de propagation est calculé de la manière suivante :

$$\begin{aligned}\text{MAPL} &= \text{UMa}(d_{3\text{D}}) \\ &= 13.54 + 39.08\log_{10}(d_{3\text{D}}) \\ &\quad + 20\log_{10}(f) - 0.6(h_{\text{UE}} - 1.5)\end{aligned}$$

En réarrangeant l'expression précédente, la distance $d_{3\text{D}}$ est exprimée ainsi :

$$d_{3\text{D}} = 10^{\frac{\text{MAPL} - 13.54 + 0.6(h_{\text{UE}} - 1.5) - 20\log_{10}(f)}{39.08}}$$

Rural Macro (RMa) On suppose que le rayon limite est toujours au-delà de la distance de rupture (*breakpoint distance*) d_{bp} exprimée en mètres [38]. La formulation adoptée est donc :

$$\text{RMa}(d_{3\text{D}}) = a\log_{10}(d_{3\text{D}}) + bd_{3\text{D}} + c$$

où :

$$\begin{cases} a = 60 + 0.03h^{1.72} \\ b = 0.002\log_{10}(h) \\ c = 20\log_{10}\left(40\pi\frac{f}{3}\right) - 0.044h^{1.72} - 40\log_{10}(d_{\text{bp}}) \\ d_{\text{bp}} = \frac{2}{3}\pi f h_{\text{BS}} h_{\text{UE}} \times 10^{-2} \end{cases}$$

La fonction inverse RMa^{-1} est la branche principale (0) de la fonction W de Lambert, qu'on note W_0.

$$\begin{aligned} d_{3\text{D}} &= \text{RMa}^{-1}(\text{MAPL}) \\ &= \frac{a}{b\log(10)} W_0\left(\frac{10^{(\text{MAPL}-c)/a} b\log(10)}{a}\right) \end{aligned} \tag{6.7}$$

6.7 Agrandissement de régions de Voronoi sur des données agrégées au niveau cellulaire

Cette section détaille la seconde étude consistant à utiliser les données des distances des utilisateurs agrégées au niveau de la cellule pour ajuster le rapport d'agrandissement des polygones de Voronoi. L'efficacité de cette méthode est analysée en utilisant les mêmes mesures et données de terrain que la première étude.

Le jeu de données agrégées est composée des distributions cumulatives empiriques des distances des utilisateurs pour chaque cellule. Pour une cellule d'identifiant $c \in \mathcal{C}$, soit $R_{\text{TA}} : \Omega \to \mathbb{R}$ la variable aléatoire correspondant à la distance d'un utilisateur à c. L'espace des événements $\Omega \subset \mathbb{N}$ est l'ensemble des valeurs possibles prises par le timing advance qui a été mesuré entre l'utilisateur et c. Les mesures de temps discrétisées sont converties en distances en utilisant la formule :

$$d = 3 \times 10^8 \frac{N_{TA}}{2}$$

où $\frac{N_{TA}}{2}$ est le temps mis par le signal pour parcourir la distance entre l'émetteur et le récepteur. Ce temps est déterminé à partir de la valeur discrète $T_A \in \{0, ..., 1282\}$ et de la fréquence d'échantillonnage $T_S = \frac{1}{15000 \times 2048}$ [146] :

$$N_{TA} = 16 T_A T_S$$

Soient $(r_{c,1}, ..., r_{c,n}) \in \mathbb{R}$ les n réalisations de R_{TA}. Elles correspondent à n mesures de distances d'utilisateurs à la cellule c.

Pour $t \in \mathbb{R}$, $F_n(t)$ est la fonction de répartition empirique de R_{TA}. En utilisant la notation $\mathbb{1}_A$ pour désigner la fonction indicatrice d'un événement A, l'expression de $F_n(t)$ est :

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{r_{c,i} \leq t}$$

Le rayon empirique limite $\hat{R}_c \in \mathbb{R}$ est la plus petite valeur vérifiant que $F_n(\hat{R}_c) \geq 0.95$. On filtre ainsi 5% des mesures pour limiter les valeurs aberrantes.

Soient $(c_1, ..., c_m) \in \mathcal{C}$ les m cellules constituant le jeu de données agrégées, $\hat{\mathbf{R}} = (\hat{R}_{c_1}, ..., \hat{R}_{c_m})$ est le vecteur tel que le i^e élément correspond au rayon limite empirique de c_i, et $\mathbf{r} = (r_{c_1}, ..., r_{c_m})$ est le vecteur tel que le i^e élément est le rayon du plus petit cercle englobant la région de Voronoi de c_i. Lorsque le polygone est redimensionné d'un rapport s, le rayon du cercle englobant devient $\mathbf{r}' = s\mathbf{r}$. Le rapport optimal est la valeur qui minimise l'erreur absolue moyenne entre $\mathbf{r}'$ et $\hat{\mathbf{R}}$:

$$\begin{cases} \text{MAE}(\hat{\mathbf{R}}, \mathbf{r}') = \frac{1}{n}\sum_{i=0}^{n} |\hat{\mathbf{R}}_{c_i} - s\mathbf{r}_{c_i}| \\ s^* = \arg\min_{s \in \mathbb{R}} \text{MAE}(\hat{\mathbf{R}}, s\mathbf{r}) \end{cases} \tag{6.8}$$

6.8 Résultats

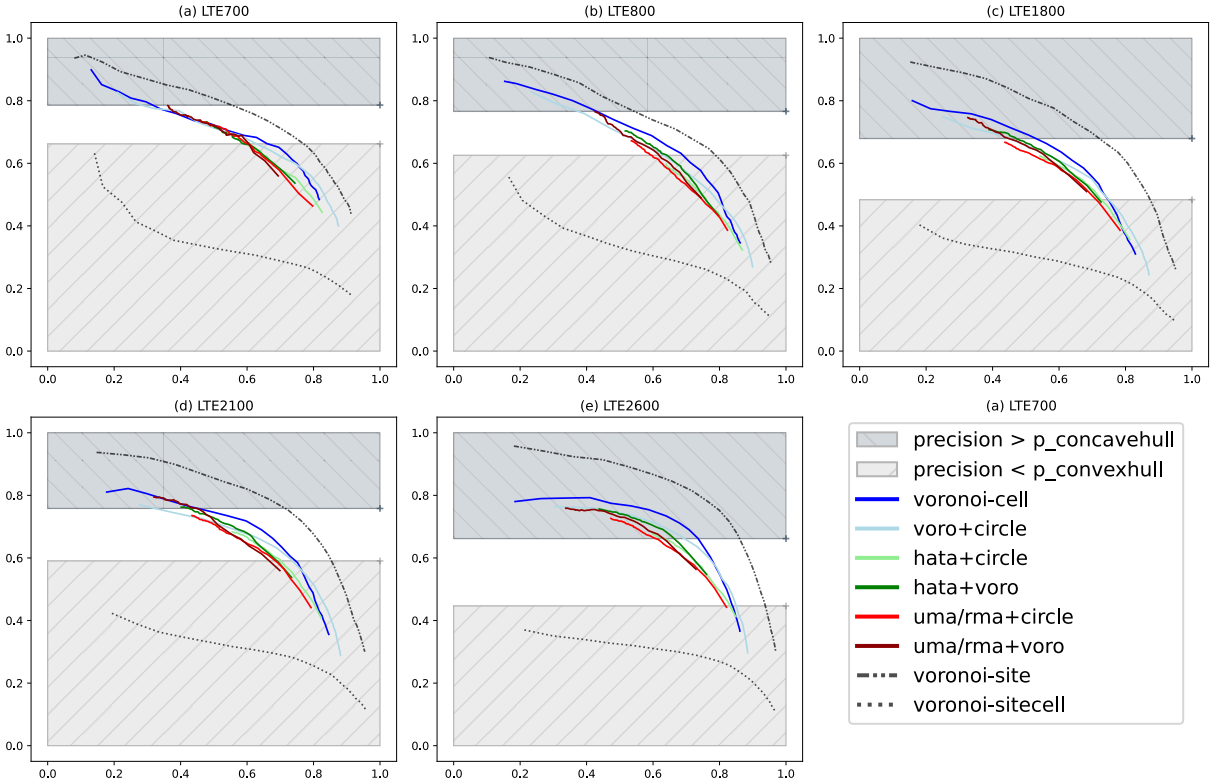


FIGURE 6.8 – Précision et rappel des modèles obtenus sur les données de terrain. Un sous-graphe correspond à l'évaluation de plusieurs modèles pour des cellules de même bande de fréquence, et une courbe d'un sous-graphe correspond à l'évaluation d'un modèle à différentes échelles.

On présente d'abord les résultats de la comparaison entre les modèles géométriques issus de Voronoi et ceux issus des cercles en combinaison avec différentes stratégies de redimensionnement. Ensuite, on présente la précision et le rappel obtenus par la couverture cellulaire de Voronoi redimensionnée en ajustant le rapport d'agrandissement sur les données agrégées. On comparera notamment ces mesures avec celles du diagramme par défaut.

6.8.1 Comparaison des modèles géométriques et stratégies de mise à l'échelle

La Figure 6.8 montre la précision moyenne $\overline{P}$ et le rappel moyen $\overline{R}$ des modèles géométriques par bande de fréquence. Une courbe correspond à une combinaison de géométrie avec une stratégie de di-

mensionnement. Pour une courbe donnée, un point sur celle-ci correspond en abscisse au rappel et en ordonnée à la précision du modèle redimensionné suivant la valeur du rapport d'agrandissement ou du nombre de PRBs. On se rapportera à la Table 6.2 pour une description des combinaisons géométrie/-dimensionnement évalués. Chaque combinaison est associée à un label qui correspond à celui renseigné dans la légende des graphes.

Bien que les graphes font penser aux courbes de précision/rappel des classificateurs binaires, ceux-ci présentent des différences majeures dans leur conception et leur interprétation. Par exemple, il est peu probable d'obtenir une précision ou un rappel de 1. D'une part, parce qu'on analyse les précisions et rappels moyennés sur toutes les cellules des données du terrain. D'autre part, les paramètres contrôlant les rapports d'agrandissement (nombre de PRBs, ou rapport lui-même) varient dans des intervalles dont les bornes ne garantissent pas qu'on obtiendra une géométrie couvrant tous les positifs (rappel = 1), ou un seul positif (précision = 1).

Le paramètre d'une stratégie de dimensionnement influence la précision et le rappel de la manière suivante. Pour une stratégie paramétrée par le rapport s (voronoi-cell, voro+circle, voronoi-site, voronoi-sitecell), plus celui-ci est grand, plus les formes géométriques sont grandes et couvrent plus d'utilisateurs attachés à la cellule, ce qui augmente le rappel moyen. Cela signifie aussi que les formes vont couvrir plus d'utilisateurs non rattachés à la cellule, ce qui diminue la précision moyenne. Par conséquent, lorsque l'on lit les graphes paramétrés par un rapport s, celui-ci augmente quand on lit la courbe de gauche à droite. Pour les stratégies paramétrées par le nombre de PRBs n_{PRB} reçu par un équipement utilisateur (hata+circle, hata+voro, uma/rma+circle, uma/rma+voro), le rayon limite R_l croît à mesure que n_{PRB} diminue. Donc n_{PRB} décroît lorsqu'on lit les graphes de la gauche vers la droite. L'efficacité de deux combinaisons (forme géométrique, stratégie de redimensionnement) A et B peuvent être comparées en observant la position de leur courbe l'une par rapport à l'autre. Si la courbe de A est au-dessus de celle de B, alors la combinaison A couvre mieux les positions des données terrains, et par extension, est un meilleur modèle de couverture.

Avec ces éléments d'interprétation, on répond à présent aux questions 1-5 posées dans la Section 6.3.

Quelle est la meilleure forme géométrique modélisant la couverture cellulaire : les sous-régions des polygones de Voronoi ou les secteurs circulaires ?

Le modèle géométrique qui suit le mieux la distribution spatiale des utilisateurs par cellule est le modèle voronoi-cell, qui divise les polygones de Voronoi en secteurs. On observe ce résultat dans toutes les bandes de fréquences, et l'écart est d'autant plus grand que la fréquence est élevée. Pour les plus basses fréquences (LTE700), les modèles sont quasiment équivalents. Pour les géométries redimensionnées en fonction des équations de propagation, les modèles de type Voronoi (hata+voro, uma/rma+voro) sont un peu plus précis que les secteurs circulaires (hata+circle, uma/rma+circle).

(a) Comment la précision et le rappel varient-ils entre la modélisation de la couverture de la station de base et la modélisation de la couverture cellulaire ? (b) À l'échelle de la cellule, que gagne-t-on à diviser la couverture de la station de base en secteurs ?

— (a) On constate un dégradation de la précision et du rappel quand on subdivise la couverture de la station de base en couverture sectorielle. Cela peut s'expliquer par la superposition intra-site des couvertures des cellules. Tout comme ils ne sont pas toujours rattachés au site le plus proche, les utilisateurs ne sont pas systématiquement attachés au secteur le plus proche.
— (b) Si on considère que la couverture de chaque cellule est confondue avec celle de sa station de base, alors la précision et le rappel chutent significativement. Cela montre que même si les couvertures intra-secteur sont superposées, la plupart des utilisateurs restent connectés au secteur le plus proche. Cela confirme donc l'intérêt de diviser la couverture des sites.

Quelles mesures et valeurs de référence utiliser pour comparer les modèles ?

Les enveloppes convexes et concaves des positions d'utilisateurs sont utilisées pour établir un intervalle de précisions admissibles. En fonction des bandes de fréquence, la borne supérieure p_concavehull varie entre 0.65 et 0.8, et la borne inférieure p_convexhull entre 0.45 et 0.65. Ces valeurs montrent qu'il existe un fort recouvrement inter-site. Elles montrent aussi le besoin d'agrandir les géométries pour augmenter le rappel et s'autoriser un certain degré de superposition de couverture.

Est-il préférable de redimensionner les géométries : (a) pour les polygones de Voronoi, avec un rapport d'agrandissement identique pour toutes les cellules, (b) pour la géométrie circulaire, avec un rapport le ramenant aux dimensions du polygone de Voronoi correspondant ou (c) avec un rapport d'agrandissement différent pour chaque cellule ? Pour le dernier cas, quelle formule de propagation est la plus adaptée : (i) Hata, (ii) UMa/RMa ?

La meilleure stratégie est de redimensionner uniformément les géométries avec un rapport d'agrandissement (voronoi-cell), suivie par le redimensionnement des cercles à l'échelle du diagramme de Voronoi (voro+circle), puis la mise à l'échelle avec un rapport dérivé d'équations de propagation. Dans ce dernier cas, les courbes sont assez proches, même si les géométries dimensionnées avec Hata (hata+voro, hata+circle) sont légèrement au-dessus d'UMa/RMa (uma/rma+voro, uma/rma+circle).

Est-ce que les résultats varient significativement à travers les bandes de fréquence ?

Bien que les valeurs de précision et de rappel varient à travers les fréquences, les allures des courbes et leurs positions les unes par rapport aux autres restent similaires.

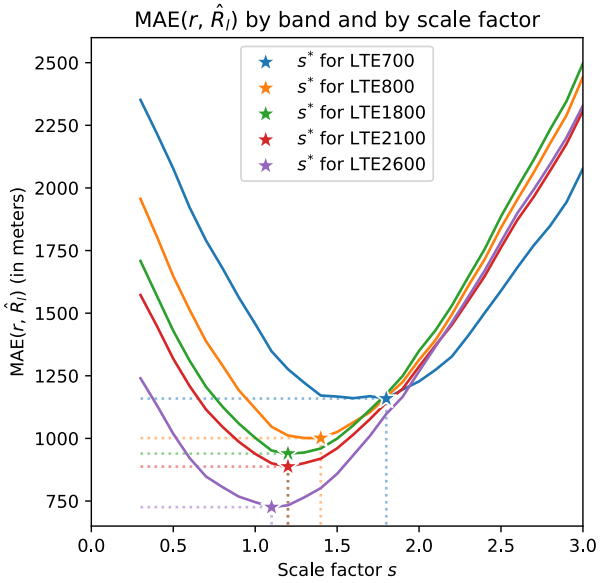


FIGURE 6.9 – Courbes de l'erreur MAE($\hat{\mathbf{R}}, \mathbf{r}'$) en fonction du rapport d'agrandissement et par bande de fréquence

Band	s^*	MAE($\hat{R}, s^*r$)
700	1.8	1159
800	1.4	1001
1800	1.2	939
2100	1.2	888
2600	1.1	726

TABLE 6.4 – Rapport optimal et MAE minimaux en mètres

Quelle efficacité obtient-on en ajustant les polygones de Voronoi grâce aux rayons de couverture limite estimées avec le timing advance ?

Les données utilisées pour ajuster le rapport d'agrandissement des polygones de Voronoi décrivent la fonction de répartition empirique des distances des utilisateurs cellule par cellule. La source fournit 12 valeurs possibles pour $\hat{R}$ en km : $I = \{0.35, 0.7, 1.1, 2.2, 3.6, 5.8, 8.0, 10.0, 15.0, 20.0, 25.0, 30.0\}$.

La Table 6.4 rapporte les valeurs de s^* et de MAE($\hat{R}, s^*r$) minimales des graphes sur la Figure 6.9. La MAE, exprimée en mètres, est relativement élevée, mais les résultats sont impactés par la granularité grossière des classes de distances qui composent les données agrégées. En effet, la distance moyenne entre deux valeurs consécutives de I est 2.7 km, 1.4 km si l'on ne considère que les classes de distances < 10 km. En moyenne donc, pour une cellule c, l'erreur absolue entre $\hat{R}_c$ et le rayon du plus petit cercle englobant le polygone de Voronoi redimensionné par s^* est de l'ordre de grandeur d'une classe de I.

Le tableau montre aussi que s^* est toujours supérieur à 1, ce qui veut dire que d'après la méthode d'ajustement employée, les polygones de Voronoi devraient toujours être agrandis. Plus la fréquence est basse, plus s^* est grand.

La Figure 6.10 présente la courbe de précision/rappel de voronoi-cell des rapports d'agrandissement dans $[0.3, 3]$, par pas de 0.1. Les performances de la géométrie par défaut sont repérées d'un triangle et celles de la géométrie redimensionnée par une étoile. Le rappel des polygones de Voronoi agrandis avec

s^* est toujours plus élevé. Pour les cellules LTE700 et LTE800, la précision est légèrement en dessous de la borne inférieure, et pour les cellules LTE2600, celle-ci est au-dessus de la borne supérieure.

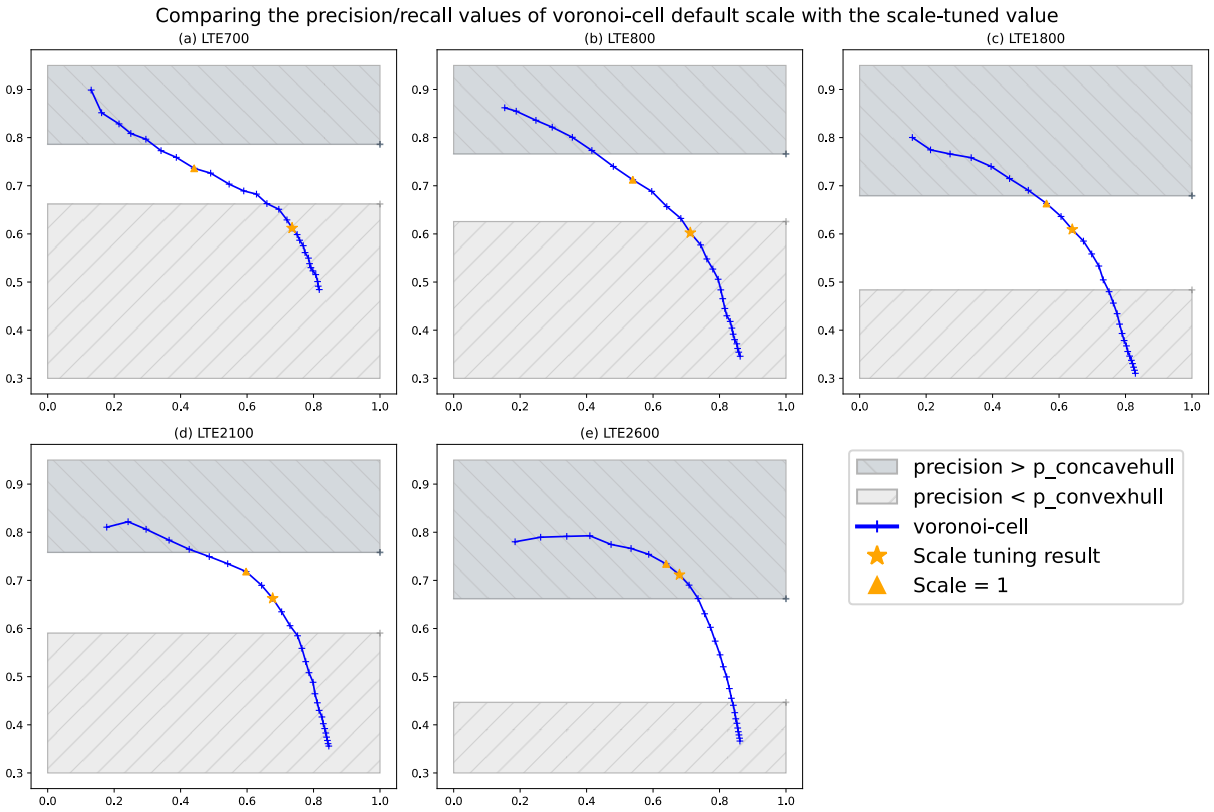


FIGURE 6.10 – Précision et rappel du diagramme de Voronoi sectorisé à différentes échelles. Les mesures mises en valeur sont celles l'échelle par défaut (marqueur triangle) et le diagramme redimensionné avec le rapport optimal (marqueur étoile). Les autres marqueurs correspondent aux autres rapports d'agrandissement utilisés pour créer la courbe.

6.9 Interprétation et discussion

On constate le diagramme de Voronoi, malgré sa simplicité, est le modèle le plus adapté pour représenter la couverture cellulaire. Les polygones doivent être redimensionnés d'un rapport d'agrandissement uniforme pour se rapprocher de la réalité. Bien que cette étude ne puisse donner d'explication certaine sur la raison pour laquelle ce modèle est le meilleur, on peut proposer des éléments d'interprétation.

Le réseau mobile est conçu pour qu'un équipement utilisateur soit rattaché à la cellule dont il perçoit le meilleur signal. Dans la zone géographique étudiée, il est probable que la densité des stations de base est si élevée que le rayon limite calculé en utilisant l'affaiblissement de propagation maximal surestime le rayon de service réel. De plus, cette forte densité du réseau pourrait aussi rendre les bordures réelles des cellules plus semblables aux frontières des polygones de Voronoi qu'à des cercles. Le redimensionnement suivant un rapport unique préserve l'information de proximité et de densité des cellules.

Pour les zones urbaines, les résultats recommandent d'agrandir les cellules de Voronoi d'un rapport appartenant à l'intervalle $]1.0, 2.0]$, pour prendre en compte les phénomènes à l'origine des superpositions de couverture, comme le mécanisme de handover. Une manière de choisir un rapport ou un nombre de PRBs adéquat est prendre la valeur telle que la précision du modèle soit équivalente à celle de l'enveloppe convexe. De cette manière, on obtient le plus grand rappel possible sans trop risquer de surestimer la couverture de service réelle. Le Tableau 6.5 montre les hyperparamètres retenus pour toutes les fréquences 4G.

Dans le cas où les positions des utilisateurs ne sont pas disponibles, on a montré qu'il était possible d'avoir un bon rapport d'agrandissement en l'ajustant sur des distances d'utilisateurs agrégées au niveau des cellules, même si les données sont moins précises. Pour éviter le surdimensionnement, un compromis pourrait être de prendre la moyenne (possiblement pondérée) entre le rapport par défaut (égal à 1) et s^*.

Bande	Modèle	Paramètre	Précision	Rappel
	convexhull	-	*0.66*	*1.0*
	voronoi-cell	$s = 1.5$	0.66	**0.66**
	voro+circle	$s = 1.0$	0.66	0.63
700MHz	hata+circle	$n_{\text{PRB}} = 38$	0.66	0.61
	uma/rma+voro	$n_{\text{PRB}} = 4$	0.66	0.61
	uma/rma+circle	$n_{\text{PRB}} = 14$	0.66	0.60
	hata+voro	$n_{\text{PRB}} = 10$	0.66	0.59
	convexhull	-	*0.63*	*1.0*
	voronoi-cell	$s = 1.3$	0.63	**0.68**
	hata+voro	$n_{\text{PRB}} = 15$	0.63	0.64
800MHz	hata+circle	$n_{\text{PRB}} = 50$	0.60	0.64
	voro+circle	$s = 0.8$	0.64	0.62
	uma/rma+voro	$n_{\text{PRB}} = 8$	0.63	0.61
	uma/rma+circle	$n_{\text{PRB}} = 29$	0.63	0.59
	convexhull	-	*0.48*	*1.0*
	voronoi-cell	$s = 1.7$	0.48	**0.75**
	voro+circle	$s = 1.1$	0.48	0.75
1800MHz	hata+circle	$n_{\text{PRB}} = 6$	0.48	0.73
	hata+voro	$n_{\text{PRB}} = 1$	0.48	0.73
	uma/rma+circle	$n_{\text{PRB}} = 3$	0.48	0.72
	uma/rma+voro	$n_{\text{PRB}} = 1$	0.51	0.68
	convexhull	-	*0.59*	*1.0*
	voronoi-cell	$s = 1.5$	0.58	**0.75**
	voro+circle	$s = 0.9$	0.60	0.73
2100MHz	hata+circle	$n_{\text{PRB}} = 11$	0.59	0.70
	uma/rma+circle	$n_{\text{PRB}} = 5$	0.58	0.69
	hata+voro	$n_{\text{PRB}} = 2$	0.59	0.69
	uma/rma+voro	$n_{\text{PRB}} = 2$	0.61	0.64
	convexhull	-	*0.45*	*1.0*
	voro+circle	$s = 1.5$	0.44	**0.85**
	voronoi-cell	$s = 2.2$	0.44	0.84
2600MHz	hata+circle	$n_{\text{PRB}} = 2$	0.46	0.82
	uma/rma+circle	$n_{\text{PRB}} = 1$	0.44	0.82
	hata+voro	$n_{\text{PRB}} = 1$	0.55	0.76
	uma/rma+voro	$n_{\text{PRB}} = 1$	0.56	0.73

TABLE 6.5 – Paramètres permettant de dimensionner les couvertures des cellules. Ils sont choisis de manière à ce que la précision des modèles soit égale à celle de l'enveloppe convexe.

De cette façon, les performances du modèle se situent sur la courbe précision/rappel de la Figure 6.10 entre les marqueurs triangle et étoile.

Dans les zones moins peuplées, les résultats peuvent être biaisés par la fiabilité des données du terrain. En milieu urbain, la densité de population et d'infrastructures est telle qu'on peut supposer une distribution uniforme des positions collectées. Cependant, les populations rurales sont principalement concentrées dans des petites villes et des villages au milieu d'espaces non peuplés. Les modèles de couverture ajustés sur des positions rurales pourraient sous-estimer la portée de service d'une cellule car aucun utilisateur n'a été suffisamment loin des habitations pour fournir une telle information. Cet aspect peut être limitant pour les études cherchant à analyser les performances du réseau en tout point de l'espace. Mais les cartes de couverture obtenues resteraient pertinentes pour des études conjointes de la couverture de service avec les zones d'activités humaines où le réseau est le plus sollicité.

6.10 Conclusion

À travers ce chapitre, nous avons développé une méthode pour comparer des modèles de couverture cellulaire par rapport à un échantillon de données du terrain constitué des positions de rattachement

Fréquence (MHz)	**Rapport**	**Rappel**	**Précision**
700	$s = 1$	0.44	0.74
	$\mathbf{s^* = 1.8}$	**0.74**	0.61
800	$s = 1$	0.54	0.71
	$\mathbf{s^* = 1.4}$	**0.71**	0.60
1800	$s = 1$	0.56	0.66
	$\mathbf{s^* = 1.2}$	**0.64**	0.61
2100	$s = 1$	0.60	0.72
	$\mathbf{s^* = 1.2}$	**0.68**	0.66
2600	$s = 1$	0.64	0.73
	$\mathbf{s^* = 1.1}$	**0.68**	0.71

TABLE 6.6 – Comparaison de la précision et du rappel du diagramme de Voronoi sectorisé à l'échelle initiale (s=1) avec le redimensionnement ajusté (s^*).

des utilisateurs aux réseau mobile. On a utilisé les notions d'enveloppe convexe et concave pour définir un intervalle de référence pour la précision, afin d'éviter que les modèles proposent des couvertures surestimées ou sous-estimées.

Les couvertures cellulaires étudiées sont basées sur deux formes géométriques : les polygones de Voronoi et les cercles. Trois stratégies de dimensionnement ont été testées : le redimensionnement uniforme des polygones de Voronoi, la mise à l'échelle des cercles à la dimension des polygones de Voronoi, et la mise à l'échelle dépendant de formules d'affaiblissement du signal. Les résultats montrent que la division sectorielle des polygones de Voronoi, combinée à un agrandissement uniforme est la solution qui se rapproche le plus de la vérité du terrain. Cette contribution supporte la validité d'utiliser le diagramme de Voronoi comme approximation simpliste de la couverture des réseaux mobiles.

Une solution alternative a aussi été proposée pour paramétrer le redimensionnement des polygones de Voronoi en l'absence de données précises sur les positions des utilisateurs. Celle-ci se base sur des données décrivant la distribution des distances utilisateurs cellule par cellule.

Pour la thèse, ces travaux nous permettent d'aborder les études des chapitres suivants en descendant à l'échelle sectorielle. Le gain direct est l'augmentation des données exploitables par les modèles d'apprentissage, puisqu'une ligne des données d'entraînement ne décrira plus un site, mais un secteur.

Chapitre 7

Prédiction de l'impact du déploiement de nouvelles cellules

7.1 Introduction

Dans les zones urbaines, la problématique des opérateurs mobiles n'est pas tant l'extension de couverture que la mise à jour des infrastructures existantes. Les usages numériques des clients évoluent, et leur consommation de données mobiles ne cesse de croître. L'enjeu est d'augmenter progressivement la capacité du réseau avant que celui-ci n'arrive à saturation. Comme il est difficile d'obtenir l'autorisation de déployer des nouveaux sites dans les villes, on privilégie l'ajout de nouvelles fréquences sur les secteurs des sites existants. Deux mises en œuvres sont envisageables :

— La largeur de bande étant une ressource limitée, les autorités de régulation autorisent petit à petit les opérateurs mobiles à ré-utiliser (*spectrum refarming*) les fréquences historiquement attribuées à une technologie ancienne (2G/3G) pour les technologies plus récentes (4G/5G). Dans ce chapitre, on étudiera le déploiement des cellules 4G, qui perdure du fait la réutilisation des fréquences 1800 MHz et 2100 MHz attribuées historiquement à la 2G et la 3G respectivement.

— Le déploiement de cellules d'une nouvelle technologie, pour lesquelles des nouvelles de bandes de fréquences ont été attribuées aux opérateurs (par exemple la bande 3490-3800 MHz pour la 5G). Ce cas de figure est étudié dans le Chapitre 8.

La ré-utilisation des fréquences n'est pas une décision systématique, car il faut prendre en compte le contexte géographique et de l'état du réseau. Afin de déterminer s'il est bénéfique d'ajouter une nouvelle cellule pour décongestionner le réseau, on cherche à prédire l'impact de ce nouveau déploiement sur les performances des cellules qui existaient sur le secteur mis à jour. Cela permettrait de cibler finement et efficacement les emplacements nécessitant un renforcement de capacité tout en limitant les dépenses de déploiement.

Les expériences ont été réalisées avec des données 4G recueillies sur des cellules situées en Île-de-France, sur la période du 4 janvier 2021 au 1er août 2022. Deux indicateurs de performance sont étudiés : la disponibilité des cellules qui dépend du pourcentage d'occupation des blocs de ressource physiques (PRBs), et le ratio entre le pourcentage d'utilisation des PRBs et le nombre d'utilisateurs. On utilise l'apprentissage automatique pour prédire les taux de croissance de ces indicateurs pour chaque secteur mis à jour.

Le reste du chapitre est structuré comme suit : la Section 7.2 présente les méthodes utilisées pour entraîner et évaluer les modèles d'apprentissage, la Section 7.3 présente les résultats et la Section 7.4 conclut les travaux.

7.2 Méthode d'apprentissage pour la prédiction de l'impact de la mise à jour d'un secteur

Pour expliquer les méthodes employées, on commence par donner la définition de l'évolution de l'activité des cellules considérée durant l'étude. Ensuite, on présentera les données d'entraînement, les modèles d'apprentissage et les manières de les évaluer.

Configuration sectorielle On utilisera régulièrement les termes de configuration (sectorielle) existante et configuration (sectorielle) ajoutée. La configuration existante est l'ensemble des fréquences couvrant déjà un secteur. Celles qui sont ajoutées au cours d'un déploiement forment la configuration ajoutée.

7.2.1 Établir les tendances du trafic

Saisonnalités des données mobiles Pour évaluer l'impact de l'ajout d'une nouvelle configuration, il faut prendre en compte certains facteurs qui peuvent biaiser l'étude. L'activité du réseau mobile varie fortement au cours du temps et est caractérisée par une forte périodicité journalière et hebdomadaire [100, 150]. Ces phénomènes rendent l'étude difficile à réaliser à une petite échelle temporelle. De plus, la saisonnalité annuelle est non négligeable et ne permet pas de comparer facilement l'évolution du réseau entre deux mois successifs. Cette saisonnalité est causée par la migration des clients. La Figure 7.1 est un exemple de saisonnalité estivale à Paris. La somme des utilisateurs quotidiens est carroyée dans une grille 500 $m \times 500$ m pour les mois de juin, d'août et de septembre 2021. Le nombre d'utilisateurs est au plus bas en août car les gens ont quitté la capitale pour partir en vacances, puis retrouve son niveau normal pour la rentrée. Sans connaître cette information, on pourrait faussement imputer aux nouveaux déploiements l'amélioration des indicateurs de performance entre juin et août, et la dégradation de ceux-ci entre août et septembre.

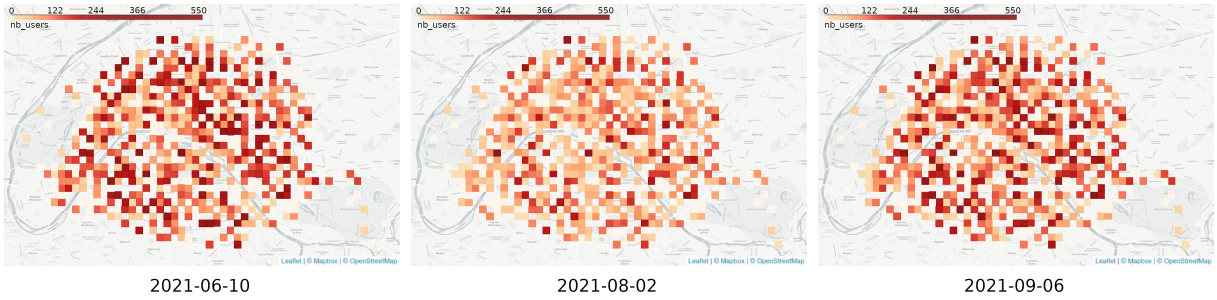


FIGURE 7.1 – « Effet vacances » en région parisienne vu à travers les données carroyées du trafic des stations de base. Un carré vide traduit une absence de station de base.

Évolution annuelle Pour limiter l'intervention des saisonnalités dans l'étude, on évalue l'évolution annuelle des indicateurs de performance. Soit $m \in \{\text{JAN}, ..., \text{DEC}\}$ le mois d'une année. Pour un secteur quelconque et un indicateur de performance donné, on suppose qu'un déploiement a eu lieu entre le mois m d'une année y et le même mois de l'année suivante (Figure 7.2). Pour étudier son impact, on calcule le taux de croissance de l'indicateur de performance des cellules à partir de ses valeurs moyennes sur les mois (m, y) et $(m, y+1)$.

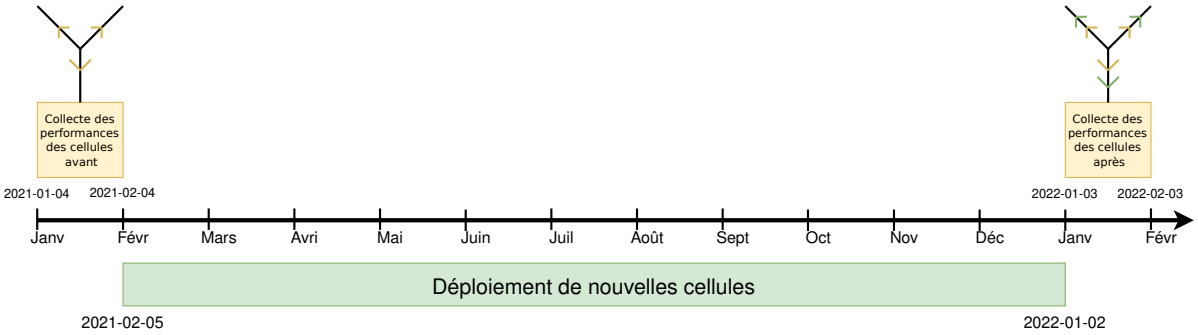


FIGURE 7.2 – Chronologie des données collectées pour comparer les performances des cellules avant et après déploiement

Indicateurs de performance utilisés Soit F_e l'ensemble des bandes de fréquence pouvant exister sur des sites avant le déploiement capacitif. Soit s l'identifiant d'un secteur. Les indicateurs de performance utilisés sont calculés à partir de deux composantes :

— La proportion de PRBs utilisés : on note $R_s^f(m, y)$ le pourcentage de PRBs utilisés par la cellule couvrant le secteur s à la fréquence $f \in F_e$. Cette valeur est moyennée sur le mois m de l'année y.
— Le nombre d'utilisateurs connectés : on note $U_s^f(m, y)$ le nombre d'utilisateurs connectés à la cellule couvrant le secteur s à la fréquence f. Cette valeur est aussi moyennée sur le mois m de l'année y.

Les indicateurs pour lesquels on calcule le taux de croissance sont dérivés de ces composantes. Il s'agit de la disponibilité des ressources et du pourcentage d'utilisation des PRBs rapporté au nombre d'utilisateurs.

La disponibilité d'une cellule est définie par la formule suivante :

$$A_s^f(m, y) = 100 - R_s^f(m, y)$$

Le taux de croissance de la disponibilité est donc :

$$\begin{aligned} \mathrm{AGR}_s^f(m, y) &= \frac{A_s^f(m, y+1) - A_s^f(m, y)}{A_s^f(m, y)} \\ &= \frac{R_s^f(m, y) - R_s^f(m, y+1)}{100 - R_s^f(m, y)} \end{aligned} \tag{7.1}$$

La variation de la disponibilité reflète l'évolution de la congestion du secteur mis à jour. Un taux positif signifie que la congestion de la cellule diminue. Pour deux cellules ayant la même réduction d'usage des PRBs, le taux de croissance de la disponibilité est plus important pour la cellule dont l'occupation initiale était la plus élevée.

Le pourcentage d'utilisation des PRBs des cellules divisé par le nombre d'utilisateurs, noté RU est défini par la relation :

$$\mathrm{RU}_s^f(m, y) = \frac{R_s^f(m, y)}{U_s^f(m, y)}$$

Le taux de croissance de RU est :

$$\mathrm{RUGR}_s^f(m, y) = \frac{\mathrm{RU}_s^f(m, y+1) - \mathrm{RU}_s^f(m, y)}{\mathrm{RU}_s^f(m, y)} \tag{7.2}$$

On suppose que ce ratio est un indicateur reflétant mieux la qualité de service que la disponibilité des cellules, car la quantité de ressources allouées par équipement affecte le débit maximal théorique qu'il peut espérer. Un taux de croissance positif devrait donc suggérer une amélioration de la qualité de service pour l'utilisateur.

7.2.2 Entraînement des modèles d'apprentissage

Le problème abordé est une régression multi-cible que l'on formalise de la manière suivante. Soit $\{(X_i, y_i)\}_{i=1}^N$ des données d'entraînement associées à $N \in \mathbb{N}$ secteurs telles que :

— $X_i \in \mathbb{R}^p$ est le vecteur des p variables décrivant la configuration existante, la configuration ajoutée, les valeurs des composantes des cellules avant le déploiement et le tissu urbain associés au secteur d'indice i.
— $y_i \in \mathbb{R}^u$ est un vecteur de u cibles composées des valeurs mensuelles des composantes R et U des cellules après le déploiement. On a observé qu'il était plus précis de prédire les composantes plutôt que les taux de croissance directement. Le nombre de sorties est égal à :

$$u = \text{nombre de composantes} \times |F_e| = 2 \times 4 = 8$$

Un modèle d'apprentissage est modélisé par une fonction $f : \mathbb{R}^p \longrightarrow \mathbb{R}^u$ telle que pour toute observation $i \in \{1, ..., N\}$:

$$f(X_i) = y_i + \epsilon_i$$

où $\epsilon_i \in \mathbb{R}^u$ désigne l'erreur de prédiction.

L'apprentissage aura pour but de paramétrer le modèle de manière à minimiser les erreurs de prédiction sur le jeu d'entraînement. Par la suite, on détaille les étapes pour créer ces données, le choix des modèles ainsi que les mesures utilisées pour les évaluer. La Figure 7.3 donne une vision d'ensemble de toute la procédure.

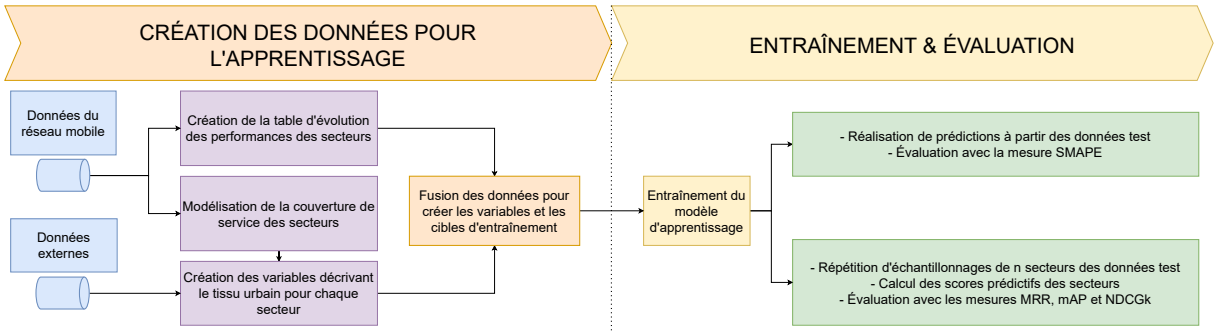


FIGURE 7.3 – Diagramme du processus de création des données d'entraînement jusqu'à l'évaluation des modèles

Création des données d'entraînement

Sources de données Les données mobiles ainsi que la topographie des cellules proviennent des bases de données de l'opérateur mobile. On reprend les mêmes sources externes que dans le chapitre précédent (OpenStreetMap [84], Human Data Exchange [90]), auxquelles on rajoute les données socio-économiques de l'Insee [91].

Évolution de l'activité des secteurs On utilise des variables binaires pour décrire les configurations, respectivement existante et ajoutée, d'un secteur. On note $F_n = \{700, 800, 1800, 2100, 2600\}$ l'ensemble des fréquences pouvant être ajoutées, et $F_e = \{800, 1800, 2100, 2600\}$ l'ensemble des fréquences pouvant exister sur un secteur. Les cellules 700 MHz ont été déployées durant la période de l'étude, et n'existaient pratiquement pas auparavant.

Pour chaque fréquence appartenant à F_e, on construit les variables et les cibles à partir des composantes R et U, moyennées par mois. Pour une observation dont les variables sont mesurées sur le mois m de l'année y, les cibles correspondantes sont les valeurs de R et U mesurées sur le mois m de l'année suivante. Les valeurs sont normées dans l'intervalle $[0, 1]$. Lorsque qu'une fréquence de F_e ne fait pas partie de la configuration existante, les valeurs de R et U ne sont pas disponibles et sont mises à 0. Les prédictions correspondantes sont filtrées et mises à 0 également.

Les indicateurs et les taux de croissance sont calculés à partir des formules (7.1) et (7.2) de la Section 7.2.1. La justesse des estimations des taux de croissance sera mesurée à l'aide de fonctions de score empruntées aux systèmes de recommandation.

Découpage des données Pour un historique de données allant du mois $m \in M$ de l'année y au mois $m' \in M$ de l'année $y + k + 1$, $(k \in \mathbb{N})$, les observations sont constituées en calculant l'évolution des secteurs sur des périodes glissantes : $(m, y, y + 1), (m_2, y, y + 1), ..., (m', y + k, y + k + 1)$ jusqu'à ce que toutes les données historiques soient utilisées.

Dans le cas où un même déploiement apparaît sur plusieurs périodes, les éléments R et U sont moyennés sur cette période, et on n'en garde qu'une observation. On garde également les secteurs qui n'ont pas été mis à jour (configuration ajoutée nulle), car on a observé qu'ils amélioraient les performances des modèles, probablement en rendant l'espace des observations plus régulier.

Modélisation de la couverture de service des secteurs On associe les données externes aux secteurs des stations de base en intersectant les positions des données externes avec les formes géographiques modélisant la couverture de service des secteurs. Précédemment, on se servait du diagramme de Voronoi pour décrire la couverture des sites et réaliser des prédictions à cette échelle. Pour modéliser la couverture plus finement, on adopte le modèle de Voronoi cellulaire décrit dans le Chapitre 6, à une variante près. Au lieu de redimensionner puis découper les régions du diagramme de Voronoi, ceux-ci sont d'abord divisés en sous-régions puis redimensionnés en prenant comme origine le centre des nouveaux polygones (Figure 7.4). Ce choix a pour effet de créer des superpositions inter-sites et intra-sites. Le rapport d'agrandissement utilisé est choisi pour minimiser l'erreur absolue moyenne entre le rayon du plus petit cercle englobant chaque polygone de Voronoi avec le rayon empirique du secteur correspondant (Figure 7.5. Ce rayon empirique est la moyenne des rayons limites de chaque bande de fréquence déployée sur le secteur, eux-mêmes étant calculés à partir de données de timing advance. Dans l'étude, la valeur du rapport

d'agrandissement utilisée est de 1.3. Les détails du modèle de couverture, ainsi que l'étude de sa fiabilité sont présentés dans le Chapitre 6.

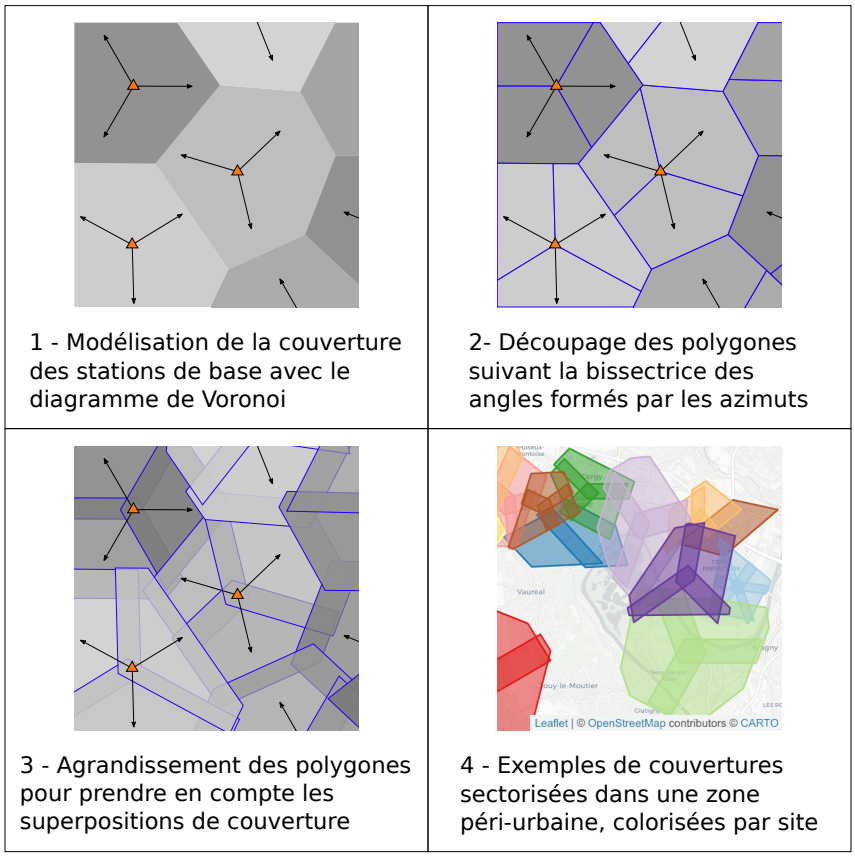


FIGURE 7.4 – Modélisation de la couverture des secteurs.

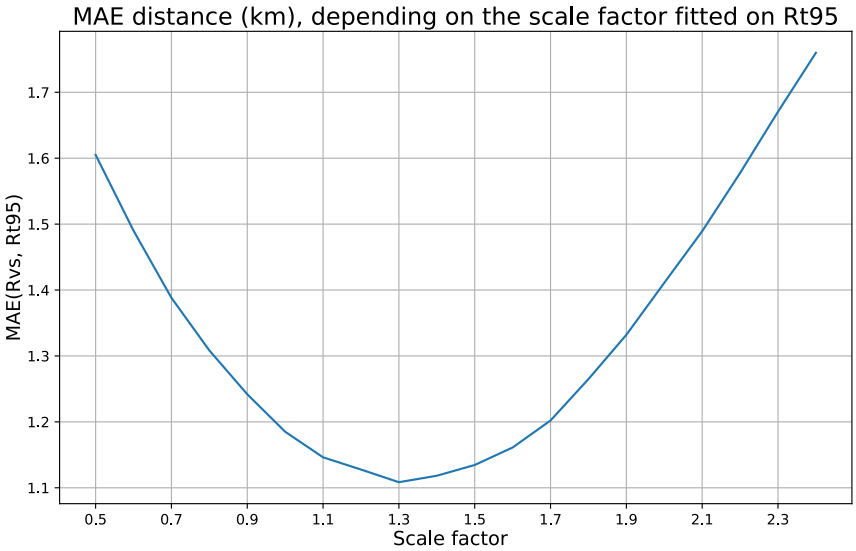


FIGURE 7.5 – Recherche du rapport d'agrandissement optimal minimisant en moyenne la différence absolue (MAE) entre le rayon limite empirique (Rt95) et le rayon du cercle minimal englobant le polygone de Voronoi (Rvs) de chaque site.

Création du tissu urbain associé aux secteurs La technique pour obtenir les variables à partir des sources externes est la même que dans le Chapitre 5, Section 5.3.4. On applique l'analyse en composantes principales pour réduire le nombre de variables tout en gardant 90% de variance expliquée. Les N facteurs

résultants sont notés $C_s^1(m,y),...,C_s^N(m,y)$.

Création des données d'entraînement Les variables du tissu urbain sont concaténées aux variables des données mobiles. Au total, on compte $p = 175$ variables d'entrée et $u = 8$ cibles, décrites dans le Tableau 7.1.

Type de donnée	Provenance des données	Description	Variables
Variables d'entrée	Réseau mobile	Fréquences ajoutées	$n_s^{700}...n_s^{2600}$, $n_f = 1$ si la bande f est ajoutée, sinon 0
Variables d'entrée	Réseau mobile	Fréquences existantes	$e_s^{800}...e_s^{2600}$, $e_f = 1$ si la bande f existait avant le déploiement, sinon 0
Variables d'entrée	Réseau mobile	Ressources avant déploiement	$R(m,y)_s^{800},...,R(m,y)_s^{2600}$
Variables d'entrée	Réseau mobile	Utilisateurs avant déploiement	$U(m,y)_s^{800},...,U(m,y)_s^{2600}$
Variables d'entrée	Sources externes diverses	Tissu urbain	$C_s^1(m,y),...,C_s^N(m,y)$
Cibles	Réseau mobile	Ressources après déploiement	$R(m,y+1)_s^{800},...,R(m,y+1)_s^{2600}$
Cibles	Réseau mobile	Utilisateurs après déploiement	$U(m,y+1)_s^{800},...,U(m,y+1)_s^{2600}$

TABLE 7.1 – Entrées et cibles des données d'entraînement

Entraînement

Choix des modèles Les travaux du Chapitre 5 ont montré que les réseaux de neurones profonds n'étaient pas adaptés à la petite quantité de données tabulaires qu'on dispose. C'est pourquoi on n'utilisera que des modèles classiques d'apprentissage machine : des modèles ensemblistes entraînés avec la méthode du gradient boosting (CatBoost [61], LightGBM [62] et XGBoost [63]), les forêts aléatoires (RF), l'algorithme des plus proches voisins (KNN) et la machine à vecteurs de support pour la régression (SVR) implémentés avec la librairie scikit-learn [59]. Pour les modèles LightGBM, XGBoost et SVR qui ne supportent pas nativement la régression multi-cible, on utilise la méthode ST comme dans le Chapitre 5.

Les estimateurs faibles entraînés avec la méthode de gradient boosting sont les arbres de décision et les modèles linéaires (lorsque l'option est disponible). Pour les librairies le supportant, les arbres ont été boostés avec et sans les techniques de *Dart dropout* [151] et de *Goss* [62] (*Gradient-Based One Side Sampling*).

Autres modèles comparatifs Ce problème est le premier dans la thèse pour lequel on dispose de données mobiles comme variables d'apprentissage. On peut donc se demander si les données externes sont toujours aussi informatives. Pour mesurer leur importance, les modèles KNN-NoFab et LightGBM-NoFab sont entraînés sur les données mobiles uniquement.

Un modèle de référence est aussi utilisé pour évaluer le gain apporté par l'apprentissage automatique par rapport à une connaissance naïve des données historiques. Ce modèle suppose que pour une combinaison donnée de fréquences existantes et ajoutées, l'impact du déploiement est constant et ne dépend pas des performances des cellules. Les valeurs de R et U prédites sont calculées sur les données d'entraînement, en moyennant les valeurs post-déploiement des composantes sur les observations de même configurations ajoutée et existante.

Hyperparamètres Les hyperparamètres des modèles ont été optimisés avec la bibliothèque Optuna [152]. Les valeurs sont données dans les Tableaux 7.2, 7.3, et 7.4 pour les SVMs, les KNNs et les modèles ensemblistes respectivement.

Partitionnement en données d'entraînement et test La méthode de validation croisée à k blocs répétée a été appliquée pour diviser les données initiales en données d'entraînement et données test. La séparation a été réalisée par site et non par secteur pour limiter au maximum les fuites de données.

Label	Noyau	Coefficient	Degré	Paramètre de régularisation
SVR-Sigm	Sigmoïde	0	-	20.5
SVR-RBF	Gaussien	-	-	9.6
SVR-Poly	Polynomial	9	2	1

TABLE 7.2 – Hyperparamètres des SVMs

Label	Paramètre de dist. de Minkowski	Poids des voisins	Nb. de voisins
KNN	1	Poids inversement proportionnel à la distance des voisins	13
KNN-NoFab	2	Poids inversement proportionnel à la distance des voisins	20

TABLE 7.3 – Hyperparamètres des KNNs

Label	Profondeur max.	Nb. d'estimateurs	Fraction d'observ.	Taille feuilles max.	Type de boosting	Frac. de variables	Taux d'apprentissage	Nb. feuilles max.
Random For.	11	400	1.0	4	-	-	-	-
CatBoost	10	10000	0.8	2	GBDT	0.8	0.03	-
LightGBM-DT	6	100	0.5	21	GBDT	0.8	0.06	20
LightGBM-NoFab	6	100	0.5	21	GBDT	0.8	0.06	20
LightGBM-Dart	7	100	0.7	16	Dart	1.0	0.14	18
LightGBM-Goss	6	100	0.1	28	Goss	1.0	0.05	20
XGBoost-DT	7	100	0.8	-	GBDT	1.0	0.8	32
XGBoost-Dart	7	100	0.8	-	Dart	0.8	0.8	32
XGBoost-Linear	-	100	1.0	-	GBLinear	1.0	0.5	-

TABLE 7.4 – Hyperparamètres des modèles ensemblistes. Certaines valeurs sont indisponibles parce qu'elles ne s'appliquent pas au modèle considéré.

Évaluation

Les modèles d'apprentissage sont évalués de deux manières différentes : d'une part sur la précision des prédictions des valeurs R et U par rapport aux cibles, d'autre part sur leur efficacité à prioriser les déploiements des secteurs en fonction de l'impact attendu.

Erreurs de prédiction (en sortie de modèle) La précision des prédictions des modèles est évaluée en utilisant les mesures de RMSE et SMAPE. Plus les erreurs sont faibles, meilleur est le modèle. L'unité de la RMSE est la même que celle de la variable étudiée, tandis que la SMAPE est une valeur normée entre 0 et 100.

Les erreurs de prédiction sont calculées par fréquence $f \in F_e$ et par composante. On note y_s^f la cible (R_s^f ou U_s^f) de la fréquence f déployée sur le secteur s et $\hat{y}_s^f$ la prédiction. Les mesures sont calculées de la manière suivante :

$$\text{RMSE}(y_s^f, \hat{y}_s^f) = \sqrt{\frac{1}{n}\sum_{s=1}^{n}(\hat{y}_s^f - y_s^f)^2}$$

$$\text{SMAPE}(y_s^f, \hat{y}_s^f) = \frac{1}{n}\sum_{s=1}^{n}\frac{|\hat{y}_s^f - y_s^f|}{|y_s^f| + |\hat{y}_s^f|}$$

Mesures de l'erreur de classement La valeur des erreurs de prédiction n'est pas toujours évidente à interpréter. On peut notamment se demander si elle est suffisamment petite de sorte que le modèle soit fiable pour de l'aide à la décision. Pour cette raison, on propose d'évaluer les modèles dans un scénario qui se servirait des prédictions pour prioriser les déploiements. Le scénario considéré est le suivant : on dispose d'un ensemble de secteurs et on ne peut déployer de nouvelles cellules que sur un sous-ensemble d'entre eux dans un premier temps. L'idée est de calculer un score de priorité pour chaque secteur, de manière à ce que l'ordre chronologique des déploiements ait le plus gros impact positif sur les performances du réseau. Le Tableau 7.5 illustre ce scénario avec un ensemble de cinq secteurs pour lesquels on été prédits l'évolution de la disponibilité des cellules suite à l'ajout de la fréquence 2100 MHz. Ces prédictions permettent de calculer la priorité de déploiement pour chaque secteur. Le classement obtenu peut être vu comme une recommandation d'ordre de déploiement pour réduire le plus efficacement la congestion du réseau.

Secteur	Bandes existantes	Ajouts	Évolution de disponibilité	Priorité
S_A 	800, 1800, 2600	2100	+1%, +2%, +5%	1
S_B 	800	2100	+10%	2
S_C 	1800	2100	-2%	5
S_D 	1800, 2600	2100	-7%, +6%	3
S_E 	800, 1800	2100	0%, -2%	4

TABLE 7.5 – Exemple de priorisation du refarming du spectre 2100 MHz pour des secteurs LTE, sur la base des prédictions du taux de croissance annuel de disponibilité des cellules.

Score d'un secteur Soit n_s le nombre de secteurs échantillonnés aléatoirement sans remise sur les données test. À partir des prédictions des composantes R et U, on estime les taux de croissance des indicateurs de performance grâce aux équations (7.1) et (7.2). Pour calculer le score d'un secteur, on calcule d'abord un score intermédiaire cs sur chaque cellule. En notant GR le taux de croissance de l'indicateur considéré, pour une cellule de fréquence f déployée sur un secteur s donné :

$$\mathrm{cs}(\mathrm{GR}_s^f) = \begin{cases} 0 & \text{si } \mathrm{GR}_s^f < 0 \\ 1 & \text{si } \mathrm{GR}_s^f = 0 \text{ ou f est non-existant} \\ \frac{2}{1+e^{-\mathrm{GR}_s^f}} & \text{si } \mathrm{GR}_s^f > 0 \end{cases}$$

La fonction score du secteur s est la moyenne des scores intermédiaires :

$$\mathrm{score}(\mathrm{GR})_s = \frac{1}{|F_e|} \sum_{f \in F_e} \mathrm{cs}(\mathrm{GR}_s^f)$$

Le score intermédiaire cs a été conçu en considérant les aspects suivants :
— Le score est toujours positif (cs > 0) pour être compatible avec les mesures utilisées pour évaluer la justesse des classements prédictifs.
— Une dégradation de la performance du réseau (un impact négatif) diminue le score.
— Une fréquence f qui est non-existante est traitée comme une cellule qui ne subit aucune évolution de performance suivant le déploiement.
— Le score est plafonné à 2 pour une question de régularité. Plus le score est élevé, plus l'impact du déploiement est positif sur une cellule. Si une cellule c_1 a un score de 1.2, et une cellule c_2 un score de 1.8, alors l'impact du déploiement sur le secteur de c_2 est plus positif que sur le secteur de c_1.

Évaluation du classement prédictif Pour chaque échantillon de secteurs à classer par un modèle d'apprentissage, on calcule les scores réels et prédictifs des secteurs. On obtient d'un côté le classement réel, de l'autre le classement prédictif, tous deux par ordre décroissant de score.

Les mesures utilisées sont le rang réciproque (RR : *Reciprocal Rank*), la précision moyenne (AP : *Average Precision*) et le gain cumulatif normalisé escompté (NDCG : *Normalized Discounted Cumulative Gain*). Ces mesures sont toutes normalisées entre 0 et 1. Plus la valeur est élevée, plus le classement prédictif est proche du classement réel.

Le rang réciproque (RR) est une mesure évaluant à quel point la recommandation la plus pertinente réellement est placée en tête du classement prédictif. Sa formule est la suivante :

$$\mathrm{RR(GR)} = \sum_{k=1}^{n_s} \frac{1}{k}[\mathrm{score(GR)}_k = \max_{1 \leq i \leq n_s} \mathrm{score(GR)}_i]$$

où $\mathrm{score(GR)}_k$ est le score d'un secteur calculé sur les taux de croissance réels, classé au rang k grâce au modèle prédictif. La somme est réduite au seul élément dont le score réel est maximal.

Précision moyenne Contrairement à la mesure précédente, la précision moyenne permet de prendre en compte si plusieurs recommandations pertinentes sont en tête du classement prédictif. Cette mesure est utilisée en classification binaire pour obtenir une moyenne des précisions $P(k)$ aux rangs k, pour $k \in [\![1, n_s]\!]$ (parfois aussi abrégé P@k). Pour un secteur s quelconque, on convertit son score en un indicateur binaire de performance avec la fonction rel :

$$\mathrm{rel}(\mathrm{score(GR)}_s) = \begin{cases} 1 & \text{si } \mathrm{score(GR)}_s > 1 \\ 0 & \text{sinon} \end{cases}$$

La précision moyenne est formulée de la manière suivante :

$$\mathrm{AP(GR)} = \frac{\sum_{k=1}^{n_s} P(k) \times \mathrm{rel}(\mathrm{score(GR)}_k)}{\text{nombre de taux de croissance positifs}}$$

où k est le secteur de rang prédictif k et $P(k)$ est la précision calculée sur les k premières valeur du classement réel.

Le gain cumulatif normalisé escompté (NDCG) permet d'évaluer si les recommandations les plus pertinentes sont placées en tête de classement. La différence avec la précision moyenne est que la pertinence n'est pas décrite par une valeur binaire, plusieurs recommandations pertinentes pouvant posséder des valeurs différentes. Les scores des secteurs sont directement utilisés pour calculer le NDCG. Cette mesure est normalisée en effectuant le ratio entre le gain cumulatif escompté DCG et le gain cumulatif escompté idéal IDCG obtenu avec le classement réel. Le calcul du DCG est le suivant :

$$\mathrm{DCG} = \sum_{k=1}^{n_s} \frac{\mathrm{score(GR)}_k}{\log_2(k+1)}$$

où $\mathrm{score(GR)}_k$ est le score calculé sur les taux de croissance réels de l'indicateur, classé au rang k grâce au modèle prédictif. Le NDCG est donc exprimé ainsi :

$$\mathrm{NDCG} = \frac{\mathrm{DCG}}{\mathrm{IDCG}}$$

7.3 Résultats et interprétation

7.3.1 Analyse statistique

Avant de présenter les résultats des modèles d'apprentissage, on présente une étude statistique visant à mesurer la différence d'évolution des secteurs mis à jour par rapport à des secteurs qui n'ont pas changé de configuration sur la période étudiée. Cette analyse préalable permet de vérifier qu'il y existe bien une différence d'évolution du trafic entre les secteurs mis à jour et les secteurs témoins, avant de se lancer dans l'entraînement des modèles d'apprentissage.

Secteur témoin Un secteur appartient au groupe témoin si :

- Il n'a reçu aucun déploiement pendant la période étudiée.
- Aucun site dans son voisinage proche n'a reçu de déploiement non plus. Cette contrainte permet d'éviter la sélection de secteurs dont les évolutions de performance ont été influencée par des déploiements proches. Le rayon de proximité est choisi arbitrairement, en fonction du contexte géographique. Cette étude exploitant des données d'Île-de-France, le rayon d'attachement des utilisateurs est en moyenne de quelques centaines de mètres. On choisit donc un rayon de voisinage de 1 km en supposant que deux sites éloignés de cette distance ne peuvent s'influencer.

Mode opératoire Les données mobiles collectées par cette étude s'étend de janvier 2021 à janvier 2022. Elles comportent 2670 secteurs témoins et 1060 secteurs mis à jour. Comme pour les données d'entraînement, on calcule les taux de croissance de la disponibilité des cellules (AGR) et de la proportion de ressources occupées par utilisateur (RUGR). Ces taux sont rassemblés par configurations existantes et ajoutées, et on calcule les taux médians des groupes formés.

Résultats de l'analyse

Le Tableau 7.6 rapporte les valeurs médianes des taux de croissance des secteurs par configuration. L'indication « Groupe témoin » dans la colonne des fréquences ajoutées signifie qu'aucune fréquence n'a été ajoutée sur les secteurs existants de ce groupe. Pour chaque sous-colonne des colonnes « AGR médian » et « RUGR médian », le taux de croissance le plus élevé parmi les différentes configurations ajoutées est en gras.

Test statistique Pour chaque taux médian, on réalise le test de Mann-Whithney U avec une valeur-p de 0.05. Le test repose sur l'hypothèse nulle que la distribution des taux de croissance des secteurs mis à jour est égale à la distribution des taux de croissance des sites témoins. Les valeurs-p inférieures à 0.05 sont soulignées dans le Tableau 7.7. On souligne dans le Tableau 7.6 les taux de croissance des déploiements significativement différents des taux des groupes témoins.

Fréquence existante	**Fréquence ajoutée**	**AGR médian (%)**				**RUGR médian (%)**			
		800	**1800**	**2100**	**2600**	**800**	**1800**	**2100**	**2600**
1800	800, 2100, 2600	-	**3.0**	-	-	-	**20.0**	-	-
	Groupe témoin	-	-0.6	-	-	-	-3.0	-	-
800, 1800	2100, 2600	**2.0**	**3.0**	-	-	-3.0	**40.0**	-	-
	Groupe témoin	-0.4	-0.7	-	-	**-0.7**	-7.0	-	-
800, 1800, 2100, 2600	700	**2.0**	-2.0	-2.0	**-0.6**	**6.0**	-10.0	**-10.0**	-10.0
	Groupe témoin	-0.2	**-1.0**	**-1.0**	**-0.6**	0.8	**-6.0**	**-10.0**	**-7.0**
800, 1800, 2600	2100	**3.0**	**0.3**	-	**0.6**	0.6	**10.0**	-	**8.0**
	700, 2100	**3.0**	-0.9	-	**0.6**	**1.0**	6.0	-	-1.0
	Groupe témoin	-0.5	-1.0	-	-1.0	**1.0**	-4.0	-	-9.0
800, 2600	1800, 2100	0.8	-	-	5.0	3.0	-	-	**40.0**
	700, 1800, 2100	**15.0**	-	-	**7.0**	**30.0**	-	-	20.0
	Groupe témoin	-2.0	-	-	-1.0	3.0	-	-	-10.0

TABLE 7.6 – Médiane des taux de croissance des indicateurs par configuration existante et ajoutée. AGR : taux de croissance de la disponibilité, RUGR : taux de croissance de la proportion de ressources occupées par utilisateur. Pour chaque sous-colonne des colonnes « AGR médian » et « RUGR médian », le taux de croissance le plus élevé parmi les différentes configurations ajoutées sont en gras. Les résultats soulignés sont statistiquement différents des valeurs du groupe témoin.

Fréquence existante	**Fréquence ajoutée**	**Valeur-p pour l'AGR**				**Valeur-p pour le RUGR**			
		800	**1800**	**2100**	**2600**	**800**	**1800**	**2100**	**2600**
1800	800, 2100, 2600	-	9.8e-06	-	-	-	2.8e-05	-	-
800, 1800	2100, 2600	3.8e-05	3.1e-16	-	-	8.2e-01	1.5e-14	-	-
800, 1800, 2100, 2600	700	1.6e-39	9.9e-01	5.6e-01	3.8e-04	4.2e-04	1.0e+00	5.0e-01	9.9e-01
800, 1800, 2600	2100	1.6e-04	5.8e-05	-	1.8e-06	3.3e-01	3.0e-06	-	1.5e-07
	700, 2100	2.8e-06	2.0e-01	-	9.6e-06	4.1e-01	3.4e-04	-	2.8e-02
800, 2600	1800, 2100	4.3e-01	-	-	9.5e-05	4.4e-01	-	-	9.3e-07
	700, 1800, 2100	5.5e-08	-	-	1.0e-08	5.5e-06	-	-	3.8e-07

TABLE 7.7 – Valeurs-p du test de Mann-Whitney U sur l'hypothèse nulle d'égalité de distribution des taux de croissance entre les ensembles des secteurs mis à jours et des secteurs témoins. Les résultats soulignés sont statistiquement significatifs pour une valeur-p égale à 0.05.

Un exemple de lecture des résultats La première ligne du Tableau 7.6 montre les taux de croissance médians des indicateurs pour les secteurs 1800 MHz, auxquels on a rajouté des cellules 800 MHz, 2100 MHz et 2600 MHz dans l'année. En comparaison, les médianes des secteurs témoins ne disposant que de la bande 1800 MHz sont aussi rapportées. Les secteurs mis à jour ont un taux de croissance positif et plus élevé que les secteurs témoins. Cette différence est statistiquement significative (valeur-p égale à 9.8e-06 d'après le Tableau 7.7). Leur congestion a diminué et la qualité de service s'est améliorée par rapport aux secteurs non mis à jour.

Au total, on observe que 14 taux de disponibilité sur 17 sont plus élevés que les niveaux des données témoins. Pour 12 de ces valeurs, la différence est statistiquement significative. Concernant la proportion de ressources occupées par utilisateur, 10 valeurs sur 17 montrent une amélioration significative de la qualité de service par rapport aux témoins. On remarque que pour toutes les configurations où les secteurs témoins présentent un taux de croissance plus élevé, cet écart n'est pas statistiquement significatif dans la majorité des cas.

Interprétation Globalement, les configurations existantes où les déploiements ont le plus amélioré les performances sont celles avec peu de fréquences (≤ 2), qui ont vu l'ajout de beaucoup de nouvelles bandes (≥ 2). Les ajouts de fréquences élevées (respectivement faibles) ont plus d'impact sur les cellules de fréquence élevées (respectivement faibles) existantes. Par exemple, les déploiements des cellules 2100 MHz semblent avoir un plus fort impact sur les cellules 1800 MHz et 2600 MHz, et l'ajout des cellules 700 MHz impacte plutôt les cellules 800 MHz existantes. Ces résultats pourraient s'expliquer par la manière dont la charge est répartie entre les bandes de fréquences ; pour les utilisateurs mobiles, leur charge aura tendance à se répartir entre les bandes 700 MHz et 800 MHz, et pour les utilisateurs fixes, entre les bandes 1800 MHz, 2100 MHz et 2600 MHz.

Conclusion Ces résultats montrent que l'impact des déploiements sur l'amélioration des performances du réseau est mesurable, et positif dans la plupart des cas. Par conséquent, on peut espérer pouvoir entraîner des modèles à prédire cet impact, à partir des données mobiles et du tissu urbain.

7.3.2 Machine learning

Cette section présente les résultats de l'évaluation des modèles. L'apprentissage a été réalisé sur un jeu de données de 5705 secteurs, avec 10 répétitions de validation croisée à $k = 5$ blocs. On analyse d'abord la précision des prédictions des composantes R et U, puis la qualité des recommandations pour réduire en priorité la congestion ou améliorer la qualité de service.

Évaluation de l'erreur de prédiction en sortie du modèle

Analyse de la RMSE Le Tableau 7.8 présente les erreurs de prédiction de chaque cible mesurées avec la RMSE. Chaque cible est la composante R ou U d'une fréquence. Les valeurs en gras indiquent la plus faible erreur obtenue, tous modèles confondus. Comme il n'y a pas un modèle unanimement meilleur sur toutes les cibles, on choisit d'ordonner les modèles par somme croissante des erreurs sur toutes les colonnes.

On peut analyser les résultats du tableau par blocs de performance similaires :

— les GBDTs sont en tête. LightGBM-Dart est le modèle avec la plus faible erreur pour la prédiction du nombre d'utilisateurs, et Catboost a la plus faible erreur pour la prédiction de l'utilisation des PRBs. Globalement, le gain d'utiliser une variante de gradient boosting (Goss ou Dart) n'est pas un hyperparamètre améliorant notablement les prédictions.

— Quelque soit le noyau utilisé (polynomial, sigmoïde ou gaussien), les SVRs ont des performances similaires et sont en bas du classement.

— La méthode des k plus proches voisins est très sensible à la taille de l'entrée. Le modèle KNN-NoFab est entraîné sur les données mobiles, qui ne constituent que 21 variables. Il souffre donc moins du fléau de la dimension que le KNN entraîné sur 175 variables.

— XGBoost-Linear est le seul modèle appliquant le gradient boosting sur des estimateurs qui ne sont pas des arbres de décision. On peut donc en déduire que l'estimateur utilisé est un hyperparamètre plus important que les variantes de gradient boosting adoptées.

— Dans les travaux précédents, la forêt aléatoire (Random Forest dans les tableaux) faisait partie des meilleurs modèles, en particulier dans le Chapitre 5. Dans cette étude, tous les GBDTs dépassent ce modèle.

— La référence choisie se trouve en fin de classement. On constate un écart minimal de 0.036 de point de pourcentage d'usage des PRBs prédit avec LightGBM-Dart, et de 0.024 points pour le nombre normé d'utilisateurs. L'utilisation des données mobiles pré-déploiement affine les prédictions par rapport à l'utilisation d'une constante apprise uniquement sur les combinaisons de fréquences ajoutées et déployées.

On note par ailleurs la faible contribution des variables du tissu urbain, le modèle LightGBM-NoFab n'étant pas spécialement moins bon que les autres GBDTs. La part d'information expliquant les valeurs des cibles qui aurait pu être apportée par les données externes est très probablement déjà contenue dans les données mobiles.

Analyse de la SMAPE Le Tableau 7.9 présente les erreurs de prédiction mesurées avec la SMAPE.

L'ordre des meilleurs modèles est globalement conservé par rapport au tableau précédent.

Bien que les RMSE associées à R et U soient d'ordre de grandeur similaire, on observe que l'erreur relative est plus élevée pour la cible U. Cette différence entre les deux composantes pourrait s'expliquer par le fait que le nombre d'utilisateurs n'est pas un indicateur borné de nature. Même normalisé, il peut être plus difficile à prédire que le pourcentage d'usage des PRBs. Cette mesure a l'intérêt de ramener les erreurs de prédiction à des grandeurs normalisées et comparables.

Par rapport à la référence, l'écart des erreurs de LightGBM-Dart est d'au minimum 8.5 points pour l'usage des PRBs et de 11.6 points pour le nombre d'utilisateurs.

Modèle évalué	**% d'utilisation des PRBs (R)**				**Nombre d'utilisateurs (U)**			
	800	**1800**	**2100**	**2600**	**800**	**1800**	**2100**	**2600**
LightGBM-Dart	**0.045±5e-4**	0.050±7e-4	0.039±5e-4	0.047±6e-4	**0.034±4e-4**	**0.039±5e-4**	**0.040±5e-4**	**0.056±6e-4**
LightGBM-DT	0.048±5e-4	0.049±7e-4	0.039±5e-4	**0.046±5e-4**	0.035±4e-4	**0.039±5e-4**	**0.040±5e-4**	**0.056±6e-4**
CatBoost	0.049±5e-4	**0.048±6e-4**	**0.038±5e-4**	**0.046±6e-4**	0.036±5e-4	0.040±6e-4	**0.040±5e-4**	**0.056±6e-4**
LightGBM-NoFab	0.048±5e-4	0.050±7e-4	0.040±5e-4	0.047±5e-4	0.035±4e-4	**0.039±5e-4**	0.041±5e-4	**0.056±6e-4**
LightGBM-Goss	0.049±5e-4	0.050±7e-4	0.039±5e-4	0.047±5e-4	0.036±4e-4	**0.039±5e-4**	**0.040±5e-4**	0.057±6e-4
XGBoost-DT	0.049±5e-4	0.050±7e-4	0.039±5e-4	0.048±6e-4	0.036±4e-4	0.040±5e-4	0.041±5e-4	0.058±6e-4
XGBoost-Dart	0.049±5e-4	0.050±7e-4	0.039±5e-4	0.048±6e-4	0.036±4e-4	**0.039±5e-4**	0.041±6e-4	0.059±7e-4
KNN-NoFab	0.051±5e-4	0.052±6e-4	0.040±5e-4	0.050±5e-4	0.037±5e-4	0.041±5e-4	0.041±6e-4	0.060±6e-4
XGBoost-Linear	0.054±5e-4	0.052±6e-4	0.042±6e-4	0.048±6e-4	0.040±4e-4	0.040±5e-4	0.042±6e-4	0.058±7e-4
Random For.	0.059±5e-4	0.051±7e-4	**0.038±5e-4**	0.048±5e-4	0.041±5e-4	0.041±6e-4	**0.040±6e-4**	0.059±6e-4
SVR-RBF	0.053±5e-4	0.058±6e-4	0.047±6e-4	0.055±6e-4	0.043±4e-4	0.049±5e-4	0.046±6e-4	0.063±7e-4
SVR-Poly	0.054±5e-4	0.058±7e-4	0.046±6e-4	0.055±6e-4	0.044±5e-4	0.048±5e-4	0.046±6e-4	0.063±7e-4
SVR-Sigm	0.055±5e-4	0.058±7e-4	0.046±6e-4	0.055±6e-4	0.044±5e-4	0.048±5e-4	0.047±6e-4	0.063±7e-4
KNN	0.055±8e-4	0.062±8e-4	0.046±6e-4	0.058±8e-4	0.040±6e-4	0.046±6e-4	0.045±6e-4	0.068±8e-4
Référence	0.082±9e-4	0.10±1e-3	0.075±9e-4	0.096±1e-3	0.059±8e-4	0.065±7e-4	0.064±9e-4	0.10±1e-3

TABLE 7.8 – RMSE ± erreur type de la moyenne des prédictions des composantes

Modèle évalué	**% d'utilisation des PRBs (R)**				**Nombre d'utilisateurs (U)**			
	800	**1800**	**2100**	**2600**	**800**	**1800**	**2100**	**2600**
LightGBM-Dart	**9.60±0.08**	9.13±0.09	9.59±0.09	9.37±0.1	18.5±0.1	17.6±0.1	20.1±0.2	**18.8±0.1**
LightGBM-DT	10.2±0.08	9.01±0.09	9.52±0.1	**9.27±0.09**	**19.0±0.1**	**17.5±0.2**	20.2±0.2	18.9±0.1
CatBoost	10.4±0.07	**8.80±0.09**	**9.33±0.09**	9.37±0.09	19.1±0.1	17.6±0.2	**19.8±0.2**	19.3±0.2
LightGBM-NoFab	10.3±0.08	9.18±0.09	9.84±0.1	9.42±0.09	19.2±0.1	17.6±0.2	20.4±0.2	19.0±0.1
LightGBM-Goss	10.5±0.08	9.10±0.1	9.61±0.09	9.42±0.09	19.4±0.1	17.8±0.2	20.5±0.2	19.1±0.2
XGBoost-Dart	10.4±0.08	9.12±0.1	9.71±0.09	9.51±0.09	19.2±0.1	17.8±0.2	20.6±0.2	19.5±0.2
XGBoost-DT	10.4±0.08	9.13±0.1	9.74±0.09	9.51±0.09	19.2±0.1	17.8±0.2	20.6±0.2	19.5±0.2
Random Forest	12.2±0.08	9.25±0.09	9.40±0.09	9.68±0.09	21.6±0.2	18.7±0.2	20.1±0.2	20.0±0.2
KNN-NoFab	10.9±0.08	9.61±0.07	9.91±0.09	10.1±0.08	20.7±0.2	18.8±0.1	20.8±0.2	20.9±0.2
XGBoost-Linear	11.6±0.09	9.68±0.1	10.7±0.1	9.97±0.09	26.7±0.3	19.6±0.2	22.1±0.2	21.0±0.2
KNN	11.8±0.1	11.5±0.09	11.5±0.1	11.9±0.1	22.3±0.2	21.3±0.1	23.4±0.2	23.5±0.1
SVR-RBF	11.7±0.08	11.6±0.1	12.4±0.1	12.0±0.1	28.2±0.2	23.3±0.2	24.9±0.2	23.6±0.2
SVR-Poly	12.0±0.09	11.4±0.1	12.2±0.1	11.9±0.1	31.0±0.3	23.2±0.2	25.0±0.2	23.6±0.2
SVR-Sigm	12.2±0.09	11.4±0.1	12.2±0.1	11.9±0.1	31.7±0.3	23.2±0.2	25.0±0.2	23.6±0.2
Référence	18.1±0.1	19.6±0.1	19.6±0.1	20.4±0.1	32.2±0.2	29.2±0.2	32.8±0.2	35.4±0.2

TABLE 7.9 – SMAPE ± erreur type de la moyenne des prédictions des composantes

Évaluation des classements prédictifs

Chaque entraînement est évalué sur 20 échantillons de 10 secteurs tirés aléatoirement sans remise. Pour chaque échantillon, on compare le classement calculé à partir des taux de croissance réels avec le classement calculé à partir des prédictions de U et R. Les résultats sont rapportés dans le Tableau 7.10.

Globalement sur toutes les cibles, les meilleurs modèles pour la recommandation de déploiements sont les SVRs, KNN-NoFab, XGBoost-Linear et LightGBM-Dart. Ce constat est tout de même nuancé en fonction de la mesure considérée :

— La méthode des k plus proches voisins obtient les valeurs RR les plus élevés. Le modèle KNN est le plus performant pour classer les secteurs par taux de croissance de la disponibilité, et le modèle KNN-NoFab pour classer les secteurs par taux de croissance du pourcentage d'utilisation des PRBs rapporté au nombre d'utilisateurs (0.11 points au dessus de la référence dans les deux cas). Le RR est une mesure exigeante, qui évalue les modèles sur le rang auquel ils recommandent le secteur le plus positivement impacté par un déploiement.

— La MAP et le NDCG sont des mesures qui notent favorablement les classements où les impacts les plus positifs sont globalement proposés en tête de classement. Les modèles ayant une valeur AP élevée sont les SVRs, une majorité des GBDTs et KNN-NoFab ($0.19 - 0.21$ points au dessus de la référence dans les meilleurs cas). Le modèle KNN et la forêt aléatoire ont les valeurs les plus faibles.

— Le gain NDCG est quasiment le même pour tous les modèles, à l'exception du KNN et de la référence ($0.12 - 0.13$ points au dessus de la référence dans le meilleur cas). Dans cette étude, ce gain ne permet pas de comparer efficacement les modèles entre eux.

Modèle évalué	**Classement prédictif par AGR**			**Classement préd. par %PRB/utilisateur**		
	RR	**AP**	**NDCG**	**RR**	**AP**	**NDCG**
SVR-RBF	0.40±0.01	0.47±0.01	**0.75±7e-3**	0.39±0.01	**0.47±0.02**	**0.71±9e-3**
KNN-NoFab	0.42±0.01	0.46±0.01	0.74±7e-3	**0.41±0.01**	0.46±0.02	0.70±0.01
XGBoost-Linear	0.43±0.02	**0.48±0.02**	**0.75±8e-3**	0.38±0.01	0.43±0.02	0.69±0.01
SVR-Sigm	0.39±0.01	0.47±0.01	0.74±8e-3	0.38±0.01	**0.47±0.02**	0.70±9e-3
SVR-Poly	0.39±0.01	0.47±0.01	0.74±8e-3	0.38±0.01	0.46±0.02	0.70±0.01
LightGBM-Dart	0.43±0.01	**0.48±0.01**	**0.75±6e-3**	0.35±0.01	0.43±0.02	0.68±0.01
CatBoost	0.43±0.02	0.45±0.02	**0.75±8e-3**	0.37±0.01	0.41±0.02	0.69±0.01
LightGBM-NoFab	0.34±0.01	0.44±0.02	0.73±7e-3	0.37±0.01	0.45±0.02	0.70±0.01
XGBoost-DT	0.36±0.02	0.42±0.02	0.73±8e-3	0.37±0.02	0.45±0.02	0.69±0.01
LightGBM-DT	0.34±0.02	0.45±0.01	0.73±7e-3	0.36±0.01	0.44±0.02	0.69±0.01
LightGBM-Goss	0.34±0.01	0.46±0.01	0.73±6e-3	0.35±0.01	0.43±0.02	0.69±0.01
XGBoost-Dart	0.34±0.02	0.43±0.02	0.73±8e-3	0.36±0.01	0.44±0.02	0.68±9e-3
KNN	**0.45±0.01**	0.39±0.01	0.71±6e-3	0.37±0.01	0.36±0.02	0.66±0.01
Random Forest	0.34±0.01	0.42±0.02	0.73±7e-3	0.34±0.01	0.37±0.01	0.67±0.01
Référence	0.34±0.01	0.29±0.01	0.62±7e-3	0.30±0.008	0.26±0.01	0.59±9e-3

TABLE 7.10 – Évaluation des classements obtenus à partir des prédictions des taux de croissance. Chaque valeur est exprimée comme la moyenne des mesures de classement ± erreur type de la moyenne

Au premier abord, les résultats sont étonnants : l'ordre des meilleurs modèles maximisant les mesures de recommandation est différent de celui minimisant les erreurs de prédiction. Une interprétation possible est que les mesures de recommandation évaluent les comportement des modèles à prédire correctement les valeurs extrêmes (puisqu'on priorise les taux de croissance les plus élevés), tandis que les valeurs RMSE et SMAPE sont plutôt représentatives des valeurs cibles majoritaires.

7.4 Conclusion

On a proposé dans ce chapitre une méthode d'apprentissage pour prédire comment le trafic des cellules existantes évoluerait après l'ajout d'une ou plusieurs nouvelles cellules. Les modèles ont d'abord été évalués avec des mesures de régression sur les prédictions des composantes servant à calculer les taux de croissance de la disponibilité et de la qualité de service des secteurs. Ensuite, ces estimations de taux de croissance ont servi à ordonner les cellules par ordre décroissant d'impact positif de déploiement. Les classements obtenus ont été évalués avec des mesures empruntées aux systèmes de recommandation. Les résultats ont montré que le choix du modèle pour l'aide au déploiement dépendait de la contrainte de classement souhaitée (prendre les k meilleurs estimations parmi n secteurs candidats, ou mettre à jour uniquement le secteur avec la meilleure évolution attendue).

Ces évaluations ont aussi permis de mesurer le gain d'utiliser l'apprentissage automatique par rapport à un modèle de référence supposant que l'impact d'un déploiement est constant pour des configurations de secteurs identiques. Dans la thèse, cette étude est la première à utiliser les données mobiles comme

variables d'entrée. Leur présence rend les variables du tissu urbain moins importantes car elles contiennent probablement des informations redondantes.

Les résultats ont aussi montré que les modèles présentant les erreurs de prédiction les plus faibles ne sont pas ceux qui maximisent les mesures de recommandation de déploiement. L'explication la plus plausible est que pour ce genre de tâche, il importe que les valeurs extrêmes soient bien prédites, au détriment d'une erreur de régression plus élevée.

Dans le Chapitre 8 qui suit, on approfondira les méthodes de ce chapitre. D'une part, les méthodes d'entraînement seront ajustées en accordant une attention à la qualité des prédictions des valeurs extrêmes élevées. D'autre part, l'apprentissage automatique réutilisera la même structure de variables d'entrée pour prédire la performance des nouveaux sites déployés.

Chapitre 8

Priorisation des déploiements 5G à partir des prédictions de trafic et de débit moyen

8.1 Introduction

En France, le déploiement commercial de la 5G a été lancé fin 2022, période à laquelle les fréquences dans la bande 3500 MHz furent attribuées aux opérateurs mobiles. Dans un premier temps, ce déploiement est « non *stand alone* » car il consiste à installer des cellules 5G sur les sites 4G et à les raccorder au cœur de réseau 4G. Les premières zones concernées sont principalement les grandes villes et les axes routiers. Elles présentent une forte concentration de sites 4G et de clients, facilitant les nouveaux déploiements et permettant d'augmenter efficacement la capacité et le débit du réseau à ces endroits là.

Le nouveau déploiement d'une technologie est toujours accompagnée d'un coût élevé d'investissements : chaque opérateur mobile a dépensé entre 600 millions et 850 millions d'euros pour l'acquisition des fréquences dans la bande 3,4 - 3,8 GHz [7], et les dépenses d'investissement et d'exploitation du réseau d'accès 5G représenteraient 45 à 50% du coût total de possession [9]. Pour être rentables, les opérateurs doivent être les plus attractifs possibles, en sachant qu'un futur client choisira son opérateur parmi un nombre de critères, dont la qualité de service. Ils peuvent pour cela être orientés grâce aux enquêtes de l'Arcep qui publie tous les ans un rapport d'enquête sur la qualité de service en métropole [153].

Pour aider les opérateurs mobiles à choisir les emplacements où la demande sera la plus importante, ou garantissant la meilleure amélioration de qualité de service, on propose d'utiliser l'apprentissage automatique pour prédire le trafic des cellules 5G et le débit moyen sur la base des indicateurs de performance des sites 4G et du tissu urbain. Avec plus de 22 000 sites 5G ouverts commercialement fin 2022 [154], le volume de données sur les performances du réseau 5G commence à devenir conséquent. Comme dans le Chapitre 7, on utilisera une mesure empruntée aux systèmes de recommandation pour évaluer la qualité de priorisation prédite des déploiements. On complète ainsi la contribution précédente, de manière à anticiper non seulement l'impact d'un déploiement sur l'infrastructure existante, mais aussi les performances des nouvelles cellules.

Dans l'étude de cas suivante, on s'intéressera à un jeu de données 4G/5G du sud-ouest de la France. Ces données sont partitionnées en données d'entraînement et données test. Les données test ont été choisies pour constituer un ensemble de sites localement déployés dans la même zone, cette zone devant être une ville de taille assez grande pour posséder un nombre de sites conséquents sur lesquels évaluer les modèles. On choisira arbitrairement les sites où la 5G est déployée sur les sites 4G de la métropole de Toulouse (totalisant ainsi 750 secteurs 4G/5G). Le reste des sites 4G/5G du sud-ouest forment les données d'entraînement. On créé ainsi un scénario où l'on cherche à déployer la 5G dans une grande ville en s'aidant de modèles d'apprentissages entraînés sur des données de déploiements passés. On pourra ensuite comparer les prédictions réalisées avec les données réelles. En fonction de l'indicateur de performance à prédire, on arrive à des conclusions différentes :

— pour le trafic 5G, les gains de l'apprentissage automatique en comparaison avec l'usage de la connaissance du volume du trafic 4G dans la zone test ne semblent pas évidents.

— pour l'évolution du débit moyen, les modèles prédictifs permettraient de mieux prédire les zones où l'évolution de la qualité de service serait la plus importante. Pour le classement des 100 premiers

sites, le score NDCG du meilleur modèle est de 0.76 contre 0.70 pour le modèle priorisant les déploiements par débit 4G croissant.

Ce chapitre est structuré comme suit : la Section 8.2 présente les données d'entraînement et les modèles utilisés, la Section 8.3 rapporte et interprète les performances des modèles prédictifs et la Section 8.4 conclut ce chapitre.

8.2 Méthode d'apprentissage pour ordonner les déploiements de la 5G

8.2.1 Données d'entraînement

Les variables utilisées sont similaires à celles du Chapitre 7, à savoir qu'on utilise les données mobiles de la 4G combinées au tissu urbain. Mais contrairement au chapitre précédent, on prédit les indicateurs de performance du nouveau site déployé et non l'évolution de l'activité des cellules existantes. Ces deux contributions sont complémentaires et peuvent être combinées pour avoir une vision d'ensemble de l'impact d'un déploiement.

Un résumé des variables d'entrée et des variables cibles est donné dans le Tableau 8.1. Elles sont les mêmes pour la prédiction du trafic tout comme pour la prédiction du débit moyen.

Données du réseau mobile

Granularité de l'équipement En général, la capacité d'un site est augmentée en déployant des cellules sur tous les secteurs. Avec le modèle de couverture cellulaire établi dans le Chapitre 6, on peut prédire les indicateurs de performance pour chaque cellule sectorisée. Par conséquent, on pourrait proposer des déploiements plus fins, ne concernant que certains secteurs. Cette particularité pourrait être utile lorsque seul un secteur est fortement sollicité, car orienté vers une zone de forte activité humaine par exemple. Dans le cas où la granularité recherchée est celle du site, on pourra arrêter la construction du modèle de couverture au redimensionnement du diagramme de Voronoi sans réaliser le découpage cellulaire. Au total, le jeu de données compte 3900 secteurs 4G sur lesquels sont déployés la 5G 3500 MHz.

Indicateurs de performance Contrairement aux études précédentes, les données mobiles utilisées proviennent de deux technologies. Puisque l'on s'intéresse au déploiement de la 5G sur les sites 4G existants, on utilise des indicateurs de performance de la 4G comme variables d'entrée pour prédire ceux des cellules 5G. Ces indicateurs sont le volume du trafic et le débit moyen des cellules de bande de fréquence 700 MHz, 800 MHz, 1800 MHz, 2100 MHz et 2600 MHz :

- Les prédictions du trafic visent à aider à prioriser les déploiements rentables. On suppose que plus le trafic est attendu est important, plus le nombre de clients 5G le sera aussi.
- Les prédictions du débit moyen visent à identifier les zones où l'amélioration de la qualité de service sera potentiellement la plus importante, en calculant le taux de croissance de celle-ci.

Ces prédictions font l'objet de deux entraînements séparés pour des études distinctes.

Taux de croissance du débit moyen Le débit moyen pondéré d'un secteur sur les fréquences d'un ensemble $\mathcal{F}$ est :

$$\text{débit\_moyen} = \frac{\sum_{b \in \mathcal{F}} \text{débit}_b \times \text{trafic}_b}{\sum_{b \in \mathcal{F}} \text{trafic}_b}$$

où trafic_b est le trafic écoulé par la cellule de fréquence $b \in \mathcal{F}$ à la vitesse débit_b.

Le débit moyen avant déploiement d'une nouvelle cellule est calculé sur l'ensemble des fréquences $\mathcal{F}_{avant}$:

$$\mathcal{F}_{avant} = \{\text{LTE700}, \text{LTE800}, \text{LTE1800}, \text{LTE2100}, \text{LTE2600}\}$$

Après déploiement d'une cellule 5G 3500 MHz, le nouveau débit moyen est calculé sur les fréquences $\mathcal{F}_{after}$:

$$\mathcal{F}_{after} = \{\text{LTE700}, \text{LTE800}, \text{LTE1800}, \text{LTE2100}, \text{LTE2600}, \text{NR3500}\}$$

Le taux de croissance du débit moyen d'un secteur est calculé à partir des débits moyens avant et après déploiement :

$$\frac{\text{débit\_après} - \text{débit\_avant}}{\text{débit\_avant}}$$

Traitement de la granularité temporelle Pour chaque indicateur et chaque cellule, les données mobiles sont initialement de granularité horaire sur la période allant du 1er novembre 2022 au 31 mai 2023. Cette granularité est plus fine que nécessaire pour le problème considéré, c'est pourquoi le trafic et le débit moyen sont agrégés à l'échelle du mois, puis on prend la valeur mensuelle maximale :

— Les données du trafic sont agrégées en sommant le trafic horaire.

— Les données débit moyen sont agrégées en moyennant les valeurs horaires.

Ce traitement permet de connaître le trafic maximal attendu et la meilleure qualité de service atteignable.

Données externes

Les données externes utilisées pour compléter les variables d'entrée proviennent des mêmes sources que dans le Chapitre 7 : OSM [84] (utilisations des sols, points d'intérêts), INSEE [91] (démographie, niveaux de vie carroyés à 200 m de côté) et Humanitarian Data Exchange (distribution de population) [90]. Ces données du tissu urbain sont croisées avec celles des cellules en adoptant le modèle de couverture cellulaire développé dans le Chapitre 6 qui dérive du diagramme de Voronoi.

Paramétrage de la couverture des secteurs 4G/5G Pour rappel, le modèle de couverture utilisé consiste d'abord à appliquer le diagramme de Voronoi sur les positions des stations de base. Ensuite, chaque région de Voronoi est découpée en sous-régions en fonction de l'orientation des antennes. Pour savoir comment dimensionner les polygones de Voronoi pour qu'ils soient à l'échelle de la couverture 5G réelle, on utilise des données empiriques sur la distribution des distances des utilisateurs calculées avec le timing advance. Ces données permettent de trouver le rapport d'agrandissement s tel que le rayon du cercle minimal englobant le polygone de Voronoi soit le plus proche possible du rayon limite empirique. Le rapport d'agrandissement obtenu est $s = 1.3$ (Figure 8.1).

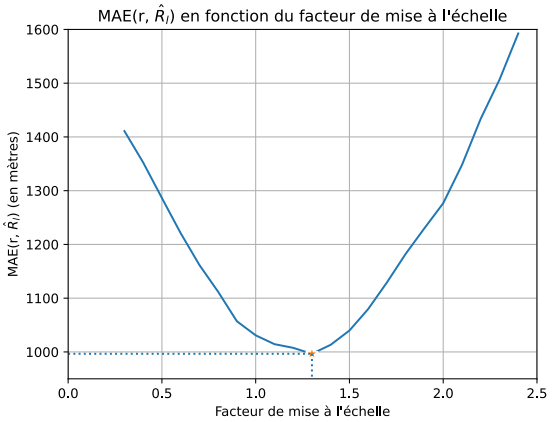
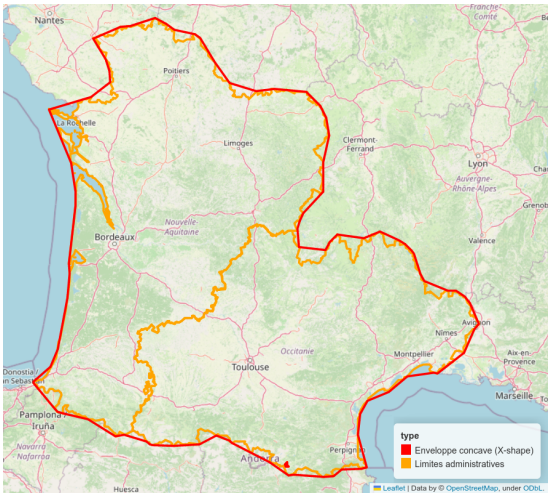
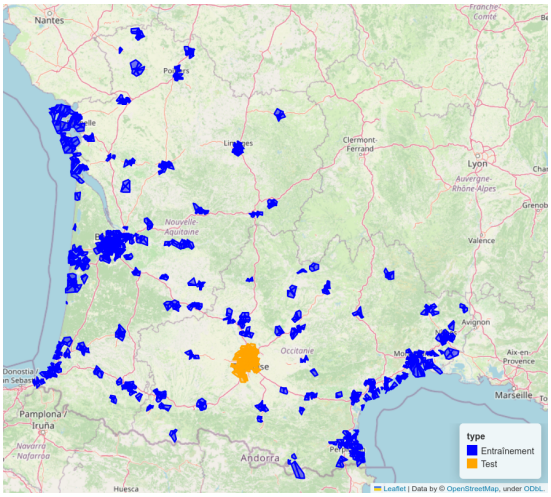


FIGURE 8.1 – Recherche du rapport d'agrandissement minimisant l'écart entre les rayons r des cercles englobants les polygones de Voronoi et les distances limites des utilisateurs $\hat{R}_l$ estimées avec les données du timing advance. La mesure de distance utilisée est l'erreur absolue moyenne.

Effets de bord Le diagramme de Voronoi n'étant pas borné aux extrémités, les couvertures des cellules côtières sont intersectées avec une enveloppe concave (χ-shape) obtenue sur l'union des frontières de la Nouvelle-Aquitaine et de l'Occitanie (Figure 8.2(a)). L'emplacement des sites utilisés dans l'étude et la modélisation des couvertures 5G sont illustrées sur la Figure 8.2(b).



(a) Zone d'étude des sites 4G/5G. En orange : les limites administratives des régions Nouvelle-Aquitaine et Occitanie, en rouge : l'enveloppe utilisée pour découper les couvertures des cellules 5G situées au bord du territoire métropolitain.



(b) Sites 4G/5G utilisés pour l'étude, colorisés par jeu d'entraînement (bleu) ou de test (orange)

FIGURE 8.2 – Frontières administratives et cellules 4G/5G du sud-ouest de la France

Croisement des données Pour compter les éléments du tissu urbain couverts par les cellules 5G, on réalise l'intersection spatiale de leur couverture avec les positions des données externes. Les variables nulles pour plus de 95% des observations sont éliminées, ce qui laisse un total de 380 variables externes scalaires. Le Tableau 8.1 résume les variables d'entrée, de sortie et leur source de provenance. Pour la lisibilité, on ne donne que les clés des attributs OSM recensés, chaque clé étant commune à plusieurs variables (celles-ci sont désignées sous la forme « `clé.valeur` » dans les données d'entraînement).

Source	Noms de variables	Type	Description
Voronoi-sectorisé	area	Entrée	Aire des polygones modélisant la couverture 5G
INSEE	Ind, Ind_snv, Ind_0_3, Ind_4_5, Ind_6_10, Ind_11_17, Ind_18_24, Ind_25_39, Ind_40_54, Ind_55_64, Ind_65_79, Ind_80p	Entrée	Caractéristiques sur les individus
INSEE	Men, Men_pauv, Men_1ind, Men_5ind, Men_prop, Men_fmp, Men_surf, Men_coll, Men_mais	Entrée	Caractéristiques sur les ménages
INSEE	Log_av45, Log_45_70, Log_70_90, Log_ap90, Log_soc	Entrée	Caractéristiques des logements
OpenStreetMap	access, barrier, bridge, man_made, power	Entrée	Petites constructions humaines
OpenStreetMap	construction, covered, landuse, layer, place, surface	Entrée	Lié au sol (caractéristiques, état, aspect,...)
OpenStreetMap	amenity, building, denomination, historic, leisure, office, religion, service, shop, sport, tourism	Entrée	Lié aux bâtiments, édifices et points d'intérêts (caractéristiques, fonctions, types,...)
OpenStreetMap	bicycle, foot, highway, horse, junction, motorcar, oneway, public_transport, railway, route, toll, tracktype, tunnel	Entrée	Lié à la circulation, aux infrastructures de transport
OpenStreetMap	intermittent, natural, water, waterway	Entrée	Décrivant les espaces naturels
Réseau 4G	traffic_LTE700, traffic_LTE800, traffic_LTE1800, traffic_LTE2100, traffic_LTE2600, thr_LTE700, thr_LTE800, thr_LTE1800, thr_LTE2100, thr_LTE2600	Entrée	Trafic et débit 4G
Réseau 5G	traffic_NR3500	Sortie/Cible	Trafic 5G
Réseau 5G	débit_après	Sortie/Cible	Débit moyen considérant les technologies 4G/5G

TABLE 8.1 – Résumé des variables utilisées pour l'entraînement. Les noms utilisés dans les sources de données sont conservés pour nommer les variables. Les variables de sortie/cible sont prédites indépendamment pour deux études différentes.

Augmentation des données pour la prédiction du trafic

La distribution des données de trafic étant très étalée, et les valeurs extrêmes élevées peu représentées, une piste de recherche a été d'augmenter les observations avec l'algorithme SMOTE. Bien que les résultats ne se soient pas avérés concluants, on documente ici la méthode employée pour les travaux souhaitant approfondir cette problématique.

SMOTE (Synthetic Minority Over-sampling Technique) est une technique de sur-échantillonnage des données pour augmenter les données des classes minoritaires [155]. On utilise la classe de plus proches voisins `NearestNeighbors` de scikit-learn [59] et la documentation de la librairie imbalanced-learn [156] pour adapter l'algorithme à générer des valeurs continues de trafic 5G. Au lieu d'avoir une classe minoritaire, on considère que les observations minoritaires sont celles dont le trafic est supérieur au fractile 0.985. On note $D = \{x_0, ..., x_{m-1}\} \subset \mathbb{R}^q$ l'ensemble des m observations minoritaires constituées de q éléments (variables et cibles confondues). Soit $x_{\text{new}} \in \mathbb{R}^q$ le vecteur d'une observation synthétique générée avec SMOTE. Il est construit à partir des éléments suivants :

- une observation $x_i \in \mathbb{R}^q$, tirée aléatoirement de l'ensemble D. Par rapport à l'algorithme initial, on choisit de tirer x_i suivant un vecteur $p \in \mathbb{R}^m$, où chaque élément est la probabilité de choisir l'observation de D de même index. Le vecteur p est obtenu en appliquant la fonction softmax au vecteur composé des distances du voisin le plus éloigné à chaque élément de D. De cette manière, on espère privilégier l'augmentation de données aux endroits peu denses en observations.
- une autre observation $x_{\text{zi}} \in D$, choisie parmi les k voisins les plus proches de x_i. Dans cette étude, on utilise $k = 5$.
- un scalaire, tiré aléatoirement dans l'intervalle $[0, 1]$.

La relation entre x_{new}, x_i, x_{zi} et λ est la suivante :

$$x_{\text{new}} = x_i + \lambda(x_{\text{zi}} - x_i)$$

Le code Python de l'algorithme est donné en page suivante.

Impact sur l'entraînement des modèles Pour les modèles de type GBDT, les modèles sont entraînés avec les mêmes hyperparamètres sur les données initiales et les données augmentées. Les entraînements des lassos font appel à la classe `LassoCV` de scikit-learn, qui trouve automatiquement le paramètre de régularisation adapté aux données d'entraînement augmentées.

```
from scipy.special import softmax
from sklearn.neighbors import NearestNeighbors

def augment_data(data, n_s, k):
    """
    :param data: les données minoritaires sur lesquels l'augmentation des données se base
    :param n_s: le nombre de données à générer
    """

    # Obtention des k voisins les plus proches pour chaque observation minoritaire.
    neigh = NearestNeighbors(n_neighbors=k)
    neigh.fit(data)

    # Obtention de la distance au plus lointain voisin et des index des voisins
    # pour chaque observation (sans inclure l'observation elle-même).
    dist, nb_idx = neigh.kneighbors(data, return_distance=True)
    dist, nb_idx = dist[:, 1:].max(axis=1), nb_idx[:, 1:]

    # Calcul du vecteur des probabilités de choisir chaque observation.
    # Plus la distance au voisin le plus éloigné est grande,
    # plus la probabilité d'être choisi est importante.
    p = softmax(dist/(dist.max() - dist.min()))

    # Choix aléatoire avec remplacement des observations sur lesquelles baser
    # l'augmentation des données.
    x_i = np.random.choice(np.arange(data_aug.shape[0]), size=n_s, p=p, replace=True)

    # Pour chaque observation choisie, on choisit aléatoirement un plus proche voisin
    # parmi les k-1 plus proches.
    x_zi = list(map(np.random.choice, nb_idx[x_i]))

    # Génération d'un vecteur de taille n_s de valeurs entre 0 et 1.
    l = np.random.random(size=len(x_i))

    # Création des données synthétiques
    x_new = data.iloc[xi].to_numpy()
                      + np.diag(l) @ (data.iloc[x_zi].to_numpy() - data.iloc[x_i].to_numpy())

    return x_new
```

8.2.2 Apprentissage

Évaluation des performances

Comme dans le Chapitre 7, on propose deux manières d'étudier les performances des modèles : en mesurant la précision des prédictions par rapport à la réalité, et en mesurant l'utilité des modèles comme systèmes de recommandation.

Pour la première manière, on choisira d'utiliser l'erreur moyenne absolue (MAE) pour comparer les prédictions aux cibles de test. Cette mesure permet de sélectionner le modèle dont les prédictions sont, en moyenne, les plus proches des données réelles. En revanche, cela nous dit peu de choses sur la qualité des prédictions : à quel point les écarts d'erreurs entre les modèles sont-ils réellement significatifs ? Les erreurs sont-elles suffisamment petites pour que l'on puisse se baser sur l'apprentissage automatique pour prendre de bonnes décisions ?

Pour répondre à ces deux interrogations, on évalue le gain d'utiliser les modèles prédictifs pour déployer les cellules 5G dans un ordre priorisant un objectif donné. L'ordre prédictif est comparé à un ordre établi uniquement grâce à la connaissance des données 4G. Les deux objectifs traités par ce chapitre sont la rentabilité et la qualité de service :

— Pour achever un retour sur investissement rapide, on suppose que les cellules 5G devraient idéalement être déployées par volume de trafic mensuel décroissant. En situation réelle, cet ordre n'est pas connu avant que le déploiement soit réalisé. Cependant, on travaille sur des données test pour lesquelles on peut obtenir cet ordre de référence et le comparer à celui obtenu en rangeant les prédictions des modèles par ordre décroissant. Les prédictions peuvent être vues comme des scores de recommandation. Plus le volume de trafic prédit est important, plus le modèle recommande de déployer la cellule associée en priorité.

— Dans le cas de l'amélioration de la qualité de service, on suppose idéalement que les cellules 5G sont déployées par taux de croissance du débit moyen décroissant. L'ordre prédictif est obtenu en calculant le taux de croissance à partir du débit moyen 4G/5G prédit par le modèle d'apprentissage considéré.

Pour comparer le classement idéal au classement prédictif, on utilise la mesure NDCG. Son calcul dépend des vrais scores de recommandation (trafic ou taux de croissance du débit) ainsi que des rangs prédictifs des cellules, mais pas des valeurs des prédictions. On réfère le lecteur au Chapitre 7 pour l'expression de cette mesure.

Choix des modèles

Le traitement des données externes est automatisé et la sélection des variables est très peu supervisée. Il en résulte un nombre important de variables, potentiellement très corrélées, dont certaines ne sont pas nécessairement explicatives de l'activité des réseaux 5G. Les résultats des travaux précédents nous portent à choisir les modèles GBDTs pour l'apprentissage des données tabulaires. Ces modèles sont basés sur les arbres de décision, des composantes qui sont robustes à la grande dimension des données passées en entrée. On évaluera ainsi CatBoost [61], XGBoost [63] et LightGBM [62].

Additionnellement, on choisira comme ligne de base le lasso. Par rapport à la régression linéaire multiple, l'intérêt de ce modèle est l'utilisation d'un terme de régularisation qui restreint le nombre de variables utilisées pour réaliser les prédictions. On réfère le lecteur à la Section 2.2.4 de l'état de l'art pour l'expression mathématique du modèle. Le paramètre de régularisation est déterminé automatiquement grâce à la classe `LassoCV` de scikit-learn. Elle trouve le paramètre de régularisation minimisant les erreurs de prédiction issues d'une validation croisée à k blocs.

Autres modèles de références Selon l'indicateur de performance prédit, on prend également d'autres modèles de références.

Pour l'objectif de rentabilité, on utilise les modèles suivants :

— LightGBM-Fabrics : un modèle apprenant uniquement sur les données du tissu urbain. Il est utilisé lors de l'évaluation des erreurs de prédiction et lors de la priorisation des déploiements.

— Mean-Traffic : un modèle renvoyant la moyenne du trafic des données test. Il est utilisé lors de l'évaluation de la priorisation des déploiements uniquement.

— LTE-Traffic : ce modèle est utilisé lors de l'évaluation de la priorisation des déploiements uniquement. Plus le trafic 4G est important sur un secteur (toutes bandes de fréquences confondues), plus celui-ci est prioritaire pour le déploiement d'une cellule 5G.

Pour l'objectif de l'amélioration de la qualité de service, on utilise les modèles suivants :

— XGBoost-Fabrics : un modèle apprenant uniquement sur les données du tissu urbain. Il est utilisé lors de l'évaluation des erreurs de prédiction et lors de la priorisation des déploiements.
— XGBoost-KPIs : un modèle apprenant uniquement sur les données de trafic et de débit 4G. Il est utilisé lors de l'évaluation des erreurs de prédiction et lors de la priorisation des déploiements.
— Min-MDR : ce modèle est utilisé lors de l'évaluation de la priorisation des déploiements uniquement. Plus le débit moyen 4G est faible, plus le secteur associé est prioritaire pour l'ajout d'une cellule 5G.

Entraînement

La double évaluation des modèles a une conséquence sur le choix des hyperparamètres des modèles, qui est fortement influencé par la qualité de prédiction des valeurs extrêmes.

Validation croisée Les erreurs de mesures rapportées sont des erreurs moyennées sur des répétitions de la validation croisée. On effectue une validation croisée à 10 blocs, répétée 5 fois pour un total de 50 entraînements par modèle.

Recherche des hyperparamètres Le Tableau 8.2 présente les hyperparamètres utilisés par les modèles entraînés à prédire le trafic et le Tableau 8.3 à prédire le débit moyen.

Pour chaque GBDT, les hyperparamètres ont été choisis de manière à avoir des estimateurs simples (arbres peu profonds), et plus nombreux que le nombre proposé par défaut. L'utilisation d'une fraction du nombre total d'observations et des variables à chaque itération de l'entraînement permet d'éviter le surapprentissage.

Le fait de considérer les modèles prédictifs comme des systèmes de recommandation impose d'accorder une attention particulière à la prédiction des valeurs extrêmes, puisque les cellules associées figureront en tête de classement. Or ces valeurs sont assez peu représentées, donc les mesures utilisées pendant l'entraînement peuvent défavoriser l'apprentissage de ces observations. Par extension, les méthodes de recherche automatique d'hyperparamètres se basant sur ces mesures ne sont pas très adaptées. Pour cette raison, on préférera utiliser une approche graphique pour vérifier si :
— les modèles ne font pas du surapprentissage en regardant les nuages de points des prédictions en fonction des cibles de validation associées.
— les prédictions ne sont pas biaisées vers une minimisation des erreurs des valeurs cibles les plus représentées, au détriment des valeurs extrêmes.

	Nombre d'arbres	Fraction de variables	Fraction d'observations	Taux d'apprentissage	Profondeur maximale des arbres	Autres hyperparamètres
CatBoost, CatBoost-Smote	1000	0.5	0.5	0.05	6	l2_leaf_reg=3, min_data_in_leaf=5
LightGBM, LightGBM-Smote, LightGBM-Fabrics	3000	0.8	0.5	0.004	12	min_child_samples=20, num_leaves=32
XGBoost, XGBoost-Smote	1000	0.5	0.5	0.006	5	-
LassoCV	-	-	-	-	-	alpha=735
LassoCV-Smote	-	-	-	-	-	alpha=263

TABLE 8.2 – Hyperparamètres des modèles entraînés pour prédire le trafic 5G

	Nombre d'arbres	Fraction de variables	Fraction d'observations	Taux d'apprentissage	Profondeur maximale des arbres	Autres hyperparamètres
CatBoost	1000	0.5	0.5	0.01	5	l2_leaf_reg=3, min_data_in_leaf=5
LightGBM	3000	0.8	0.5	0.004	12	min_child_samples=20, num_leaves=32
XGBoost, XGBoost-Fabrics	1000	0.3	0.3	0.006	5	-
LassoCV	-	-	-	-	-	alpha=7.6

TABLE 8.3 – Hyperparamètres des modèles entraînés pour prédire le débit moyen 4G/5G

8.3 Résultats et interprétation

Dans cette section, on présente les résultats des modèles entraînés pour la prédiction du trafic 5G et du débit moyen après ajout de la 5G.

8.3.1 Prédiction du trafic 5G

Augmentation des données

L'approche employée pour augmenter les données d'entraînement n'a pas montré d'amélioration notable. Le Tableau 8.4 présente les performances des modèles entraînés à prédire le trafic 5G. Les modèles suffixés « -Smote » sont entraînés sur des données augmentées. On constate que leur MAE n'est pas plus petite que les modèles entraînés sur les données initiales.

Pour plus de détails, on analyse la dispersion des erreurs de prédiction. La Figure 8.3(a) présente les nuages de points des données d'entraînement, validation et test associés au modèle LightGBM, et la Figure 8.3(b) présente les nuages de points associés à LightGBM-Smote. La droite en rouge représente l'équation $y = x$, avec laquelle serait confondus les points si les modèles ne faisaient aucune erreur. On peut observer que le biais de prédiction du modèle est faible pour les volumes de trafic faible et moyen car le nuage est centré autour de la droite. En revanche, les valeurs élevées sont sous-estimées.

L'augmentation des données avec SMOTE sur les valeurs minoritaire corrige le biais des valeurs extrêmes dans le jeu d'entraînement (Figure 8.3(b)). Cependant on peut voir à la dispersion des données de validation et de test que cet apprentissage n'est pas généralisable, et qu'on a contraire du surapprentissage.

Bien que non concluante, il est envisageable d'approfondir l'augmentation des données en envisageant de faire varier le nombre de plus proches voisins ou d'améliorer la manière de choisir les données minoritaires.

Model	**MAE**	**SD**
LightGBM-Smote	642.6	4.0
LightGBM	644.9	3.3
XGBoost	649.8	2.8
XGBoost-Smote	651.7	3.4
CatBoost	660.3	5.3
CatBoost-Smote	663.5	6.5
Lasso	673.0	2.7
Lasso-Smote	753.0	10.0
LightGBM-Fabrics	898.2	5.3

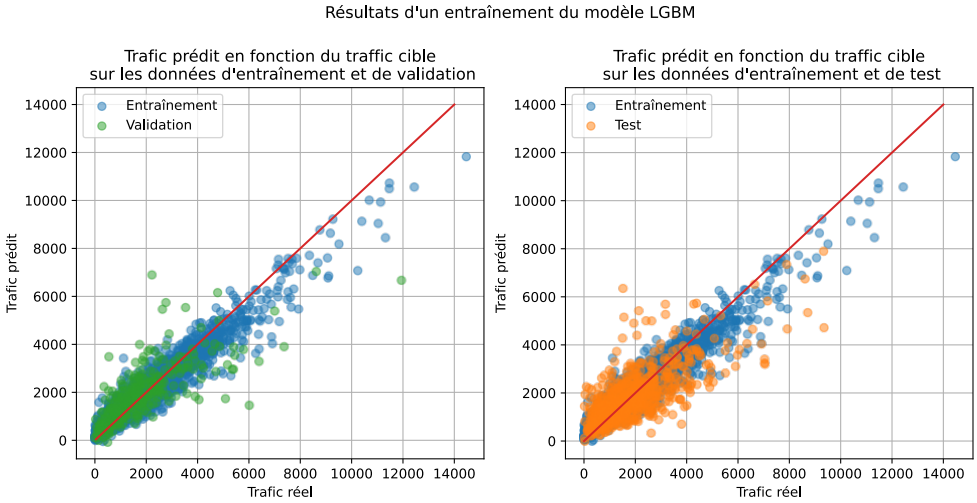
Table 8.4 – MAE des prédictions de trafic 5G sur les données test

Comparaison des modèles

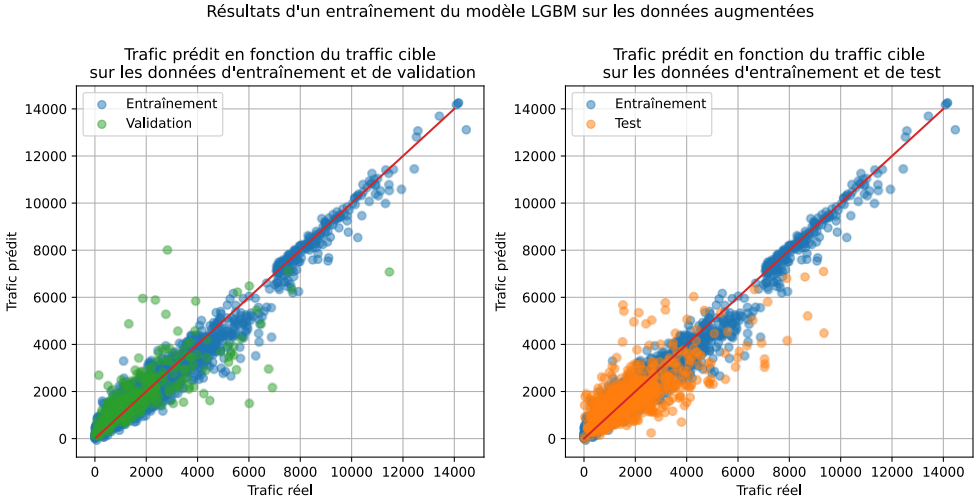
Les modèles sont évalués en mesurant les erreurs de prédiction (Tableau 8.4) ainsi que leur efficacité à ordonner les déploiements par volume de trafic attendu décroissant (Figure 8.4). Les écarts de MAE entre les trois GBDTs sont faibles, ce que le score NDCG confirme également sur la Figure 8.4(b), de sorte qu'on ne puisse clairement conclure de la supériorité de l'un des trois modèles. Le lasso présente également des performances très semblables.

Le modèle LightGBM-Fabrics entraîné uniquement sur les données externes est bien derrière, même si certaines informations sont apprises puisque le classement proposé est mieux noté que celui de MeanTraffic. Comme ce dernier prédit une valeur unique peu importe le secteur considéré (moyenne du trafic 5G des données d'entraînement), tous les secteurs sont de même rang dans le classement. De plus, si on analyse la dispersion des prédictions par rapport aux cibles associée à LightGBM-Fabrics (Figure 8.3(c)), on peut voir que la plus grosse différence par rapport aux autres GBDTs est la difficulté à prédire les valeurs élevées.

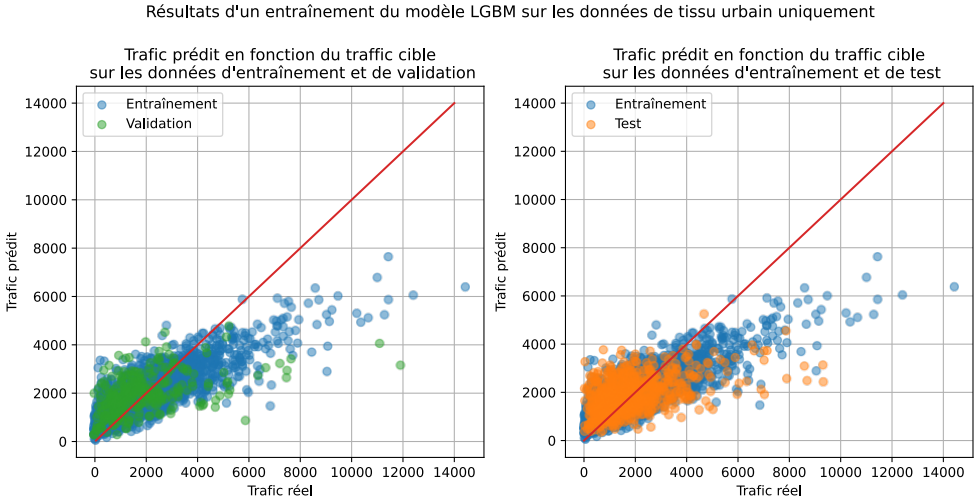
Interprétation des modèles Pour confirmer que les données mobiles sont les variables les plus informatives, on a regardé l'importance des variables pour les modèles LightGBM et lasso. Le premier est interprété avec le graphe beeswarm des valeurs SHAP (Figure 8.5). Une explication sur la manière de la lecture des beeswarms a été fournie dans la 3.4 du Chapitre 3. Le lasso est interprété en ordonnant les



(a) Sans augmentation de données, prédictions de LightGBM



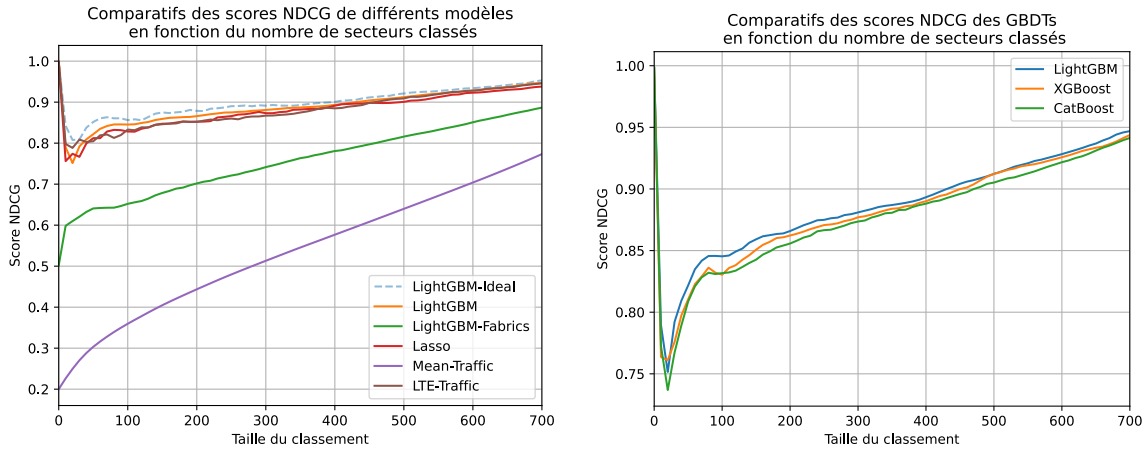
(b) Augmentation des données extrêmes du jeu d'entraînement, prédictions de LightGBM-Smote



(c) Sans augmentation de données, prédictions de LightGBM-Fabrics

FIGURE 8.3 – Nuages de points des prédictions de LightGBM en fonction des cibles des données d'entraînement, de validation et de test. La donnée prédite est le trafic 5G mensuel en gigaoctets. La ligne rouge représente l'équation $y = x$.

coefficients de régression non nuls par ordre décroissant (Tableau 8.5). Plus un coefficient est important, plus la variable associée l'est également.



(a) NDCG avec les lignes de base

(b) NDCG entre GBDTs

FIGURE 8.4 – Scores NDCG médians du classement des secteurs établi en fonction du trafic 5G attendu

Naturellement, le trafic 4G fait partie des variables importantes. Quatre d'entre elles font partie des variables les plus importantes de LightGBM et sont positivement corrélées à la prédiction d'un trafic élevé. En revanche, on voit que les variables décrivant le débit et le tissu urbain utilisées par LightGBM sont pour la plupart différentes de celles retenues par le lasso. Cela montre que beaucoup de variables sont corrélées entre elles et qu'il n'existe pas de sous-ensemble unique pour prédire le trafic 5G. Les variables de l'INSEE qui sont convervées sont principalement des données sur les types de logements. Les données OSM conservées sont principalement liées à des infrastructures de transport, à certains types de commerces et bâtiments, et à l'utilisation résidentielle et naturelle des sols.

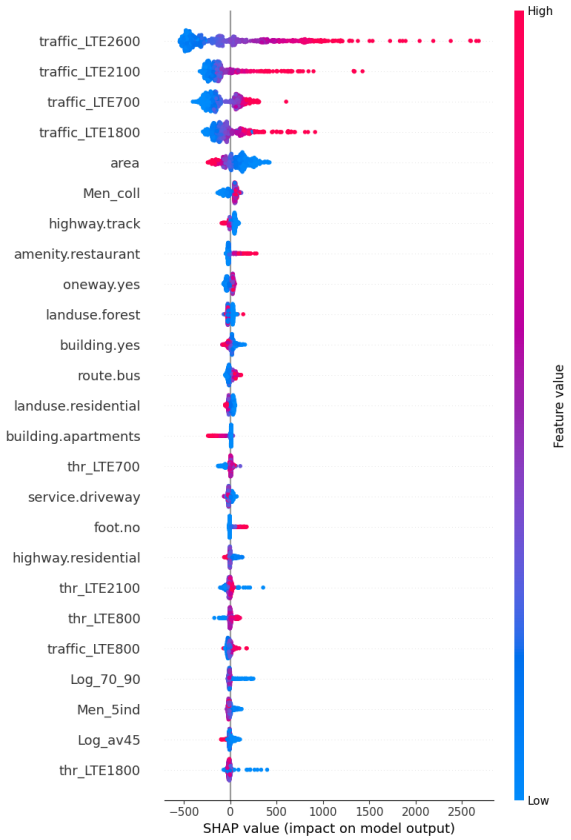


FIGURE 8.5 – Interprétation d'un LightGBM à travers les prédictions du trafic 5G des secteurs toulousains avec SHAP

Variables	Coefficient
highway.crossing	2.96
service.parking_aisle	1.73
oneway.yes	1.30
access.private	0.84
amenity.parking_space	0.69
highway.footway	0.46
Men_1ind	0.42
traffic_LTE700	0.41
Men	0.35
Ind_65_79	0.27
traffic_LTE2600	0.26
Log_soc	0.22
route.bus	0.20
traffic_LTE2100	0.19
Log_av45	0.16
Log_70_90	0.14
traffic_LTE800	0.078
Men_mais	0.070
traffic_LTE1800	0.043

TABLE 8.5 – Interprétation du lasso entraîné à prédire le trafic 5G à l'aide des coefficients non nuls

Enfin, on peut constater que les scores NDCG des modèles prédictifs et de LTE-Traffic sont très proches. Cela signifie que les recommandations de déploiement réalisées par les modèles d'apprentissage ne sont pas significativement meilleures que celles obtenues par la connaissance du trafic 4G de la zone test. Si l'on cherche uniquement à prioriser le déploiement de la 5G sur les sites pouvant générer le plus de nouveau trafic, il est plus direct de se fier aux sites 4G générant le plus de trafic que d'entraînement un modèle automatique. L'apport de ces derniers reste toutefois utile pour connaître le volume de trafic 5G qui serait attendu.

LightGBM-Ideal Pour pousser l'analyse des performances des modèles automatiques plus loin, on a cherché à entraîner un modèle, nommé LightGBM-Ideal, sur un sous-ensemble de variables sélectionnées à l'aide des données test. Le principe est de partir de l'ensemble des variables initiales, puis itérativement, de supprimer la variable la moins importante d'après l'algorithme de permutation des variables appliqué aux données test, puis de ré-entraîner le modèle sur les données d'entraînement. La permutation des variables a pour principe d'évaluer l'importance d'une variable en mesurant la dégradation de performance du modèle lorsque cette variable est permutée aléatoirement avec d'autres. Un dégradation importante suggère que la variable est importante dans la précision du modèle. Le résultat est la conservation des 54 variables présentées dans le Tableau 8.6, ordonnées par ordre alphabétique.

En pratique, cette manière de choisir les variables n'est pas applicable puisqu'on se sert des données test. Elle est seulement réalisée à titre indicatif pour montrer qu'il existe un sous-ensemble de variables pour lesquelles les prédictions peuvent être encore meilleures. Cependant, la méthode pour trouver ce sous-ensemble sans connaître les données cibles du jeu de test reste un point à élucider. Si elle existe, cela implique de devoir spécialiser chaque entraînement en fonction de la zone de déploiement.

Variables (1 à 18)	**Variables (19 à 36)**	**Variables (37 à 53)**
Log_70_90	bicycle.yes	power.cable
Log_soc	building.service	power.generator
amenity.bench	covered.yes	railway.rail
amenity.parking	foot.no	route.bus
amenity.parking_entrance	highway.give_way	service.parking_aisle
amenity.parking_space	highway.path	shop.clothes
amenity.recycling	highway.pedestrian	shop.sports
amenity.restaurant	highway.primary	shop.supermarket
amenity.waste_disposal	junction.roundabout	sport.tennis
area	landuse.grass	surface.sand
area.yes	landuse.residential	surface.sett
barrier.bollard	leisure.pitch	thr_LTE700
barrier.gate	leisure.sports_centre	tourism.hotel
barrier.height_restrictor	natural.grassland	traffic_LTE1800
barrier.lift_gate	natural.wood	traffic_LTE2100
barrier.swing_gate	office.company	traffic_LTE2600
barrier.wall	oneway.no	traffic_LTE700
bicycle.no	oneway.yes	

Table 8.6 – Liste des variables importantes sélectionnées avec LightGBM-Ideal

8.3.2 Prédiction du débit moyen

Les variables utilisées pour l'apprentissage de la prédiction du débit sont les mêmes que pour la prédiction du trafic, c'est pourquoi on suppose que l'augmentation des données aboutirait à des résultats similaires. Par conséquent, cette aspect n'est pas abordé dans cette section.

La MAE est une mesure qui dépend de la grandeur des valeurs prédites. Par rapport à l'étude précédente, le débit moyen présente des ordres de grandeurs et des unités différents. Pour cette raison, les erreurs de prédiction rapportées dans le Tableau 8.7 sont beaucoup plus faibles.

Model	MAE	SD
LightGBM	15.57	0.11
XGBoost	15.63	0.09
CatBoost	15.64	0.07
XGBoost-KPIs	16.24	0.08
Lasso	18.41	0.09
XGBoost-Fabrics	21.12	0.10

TABLE 8.7 – MAE des prédictions de débit moyen 4G/5G sur les données test

Comme précédemment, les GBDTs et le lasso entraînés sur les données mobiles ont des erreurs semblables, qui sont moindres que l'erreur de XGBoost-Fabrics, un modèle entraîné uniquement sur les données externes. La Figure 8.6 illustre les nuages de points des données d'entraînement, de validation et de test associées aux prédictions de XGBoost. On observe que les petites valeurs sont surestimées, et les valeurs élevées sous-estimées, ces dernières présentant un biais plus important. La distribution des données de Toulouse semble un peu différente du reste du sud-ouest car le nuage de points est plus dispersé autour de la droite $y = x$. Cela indique que XGBoost généralise moins bien sur les données test que sur les données de validation.

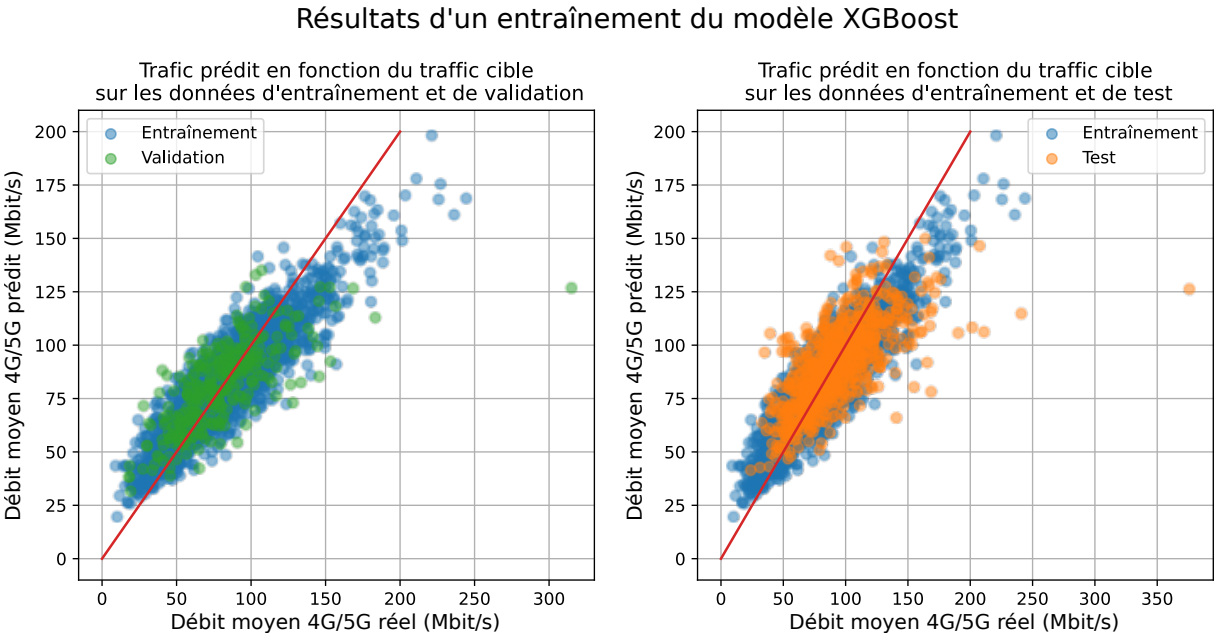
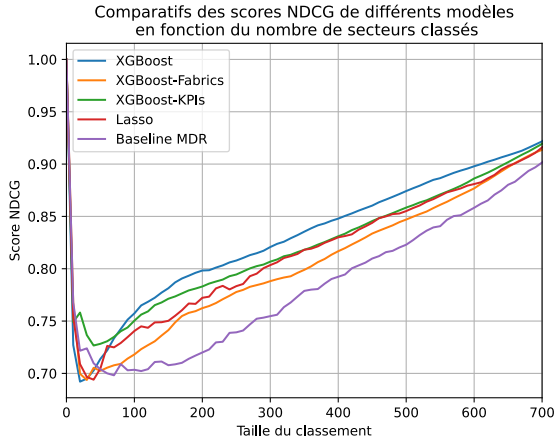
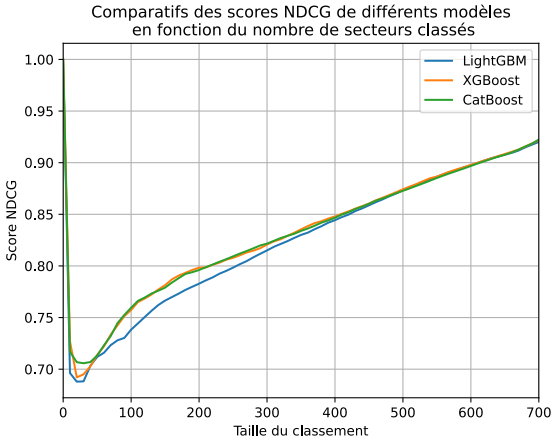


FIGURE 8.6 – Nuage de point des prédictions du modèle XGBoost par rapport aux cibles d'entraînement, de validation et de test. La donnée prédite est le débit moyen 4G/5G en mégabits par seconde. La ligne rouge représente l'équation $y = x$.

Par ailleurs, les modèles avec la plus faible MAE ont aussi les meilleurs scores NDCG (Figure 8.7). L'ordre des meilleurs modèles est globalement conservé d'une mesure d'erreur à l'autre, le profil des scores NDCG des GBDTs étant toujours proches. Le modèle Baseline MDR est moins bon que les autres modèles, avec 0.08 points d'écart par rapport à XGBoost pour la recommandation des 200 premiers sites à mettre à jour. Il y a donc un gain d'apprendre sur les données de déploiement passées plutôt que de déployer en priorité sur les sites de plus faible débit si l'on souhaite améliorer le plus efficacement possible la qualité de service.



(a) Comparaison du NDCG entre plusieurs modèles



(b) Comparaison du NDCG entre les GBDTs

FIGURE 8.7 – Scores NDCG médians du classement des secteurs établi en fonction du débit moyen 4G/5G attendu

Interprétation des modèles On compare les variables importantes pour le modèle XGBoost (Figure 8.8) avec le lasso (Tableau 8.8). Pour les deux modèles, le débit est de loin la donnée la plus importante pour la prédiction du débit moyen futur. Les données externes les plus importantes concernent les caractéristiques des ménages et les infrastructures de transport. Les données liées au trafic disparaissent des variables importantes de lasso, bien qu'elles le restent à des degrés divers pour XGBoost. En plus de cela, les prédictions de ce GBDT sont négativement corrélées à l'utilisation du sol des zones peu urbanisées (ex : `landuse.forest`, `landuse.farmland`). Ici encore, même si des thématiques de variable se dégagent, il n'existe pas de sous-ensemble unique sur lequel réaliser les apprentissages.

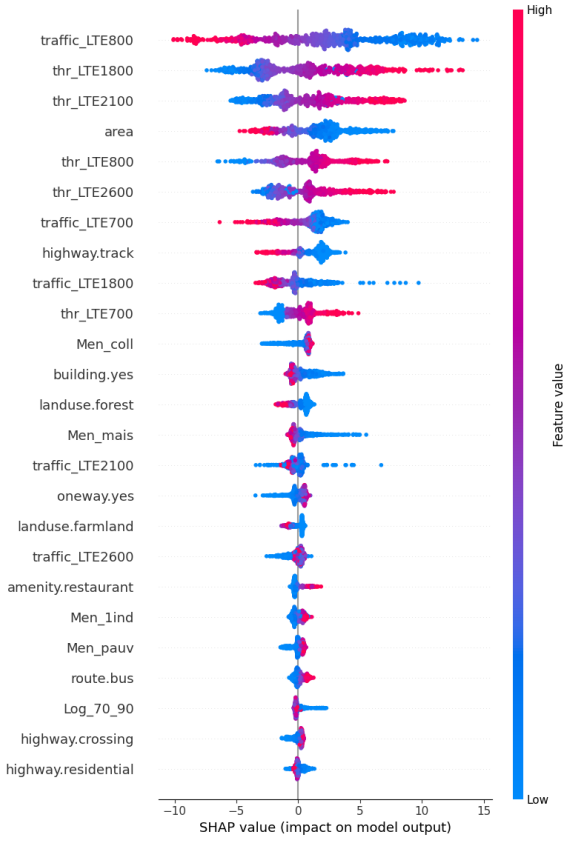


FIGURE 8.8 – Interprétation du modèle LightGBM à travers les prédictions du débit moyen 4G/5G des secteurs toulousains avec SHAP

Variables	**Coefficient**
thr_LTE800	0.72
thr_LTE700	0.25
thr_LTE1800	0.090
service.parking_aisle	0.050
highway.pedestrian	0.042
highway.crossing	0.037
thr_LTE2600	0.031
thr_LTE2100	0.031
Ind_65_79	0.021
highway.residential	0.017
route.bus	0.014
leisure.swimming_pool	0.012
building.roof	0.0066
surface.asphalt	0.0065
highway.service	0.0061
Men	0.0056
Men_coll	0.0053
Log_av45	0.0049
highway.footway	0.0044
bicycle.yes	0.0041

TABLE 8.8 – Interprétation du lasso entraîné à prédire le débit moyen 4G/5G à l'aide des coefficients non nuls. Seuls les 20 plus gros coefficients sont rapportés dans le tableau.

8.4 Conclusion

À travers ces deux études, on propose une deuxième application de l'apprentissage automatique pour aider à densifier le réseau mobile, cette fois dans le cadre de l'ajout d'une nouvelle génération de technologie mobile.

On montre à nouveau le double emploi possible des modèles prédictifs, à la fois pour prédire des indicateurs de performance, et pour prioriser les déploiements selon un critère donné. Les prédictions portent cette fois sur le trafic des nouvelles cellules, et du débit moyen attendu sur un secteur suite à l'ajout de capacité. Cette méthode peut ainsi être combinée aux travaux du Chapitre 7 pour avoir une vision d'ensemble, à la fois sur l'impact d'un déploiement sur l'infrastructure existante mais aussi sur les performances des nouvelles cellules.

En présence de données mobiles dans la zone de déploiement, celles-ci sont de loin les variables les plus importantes pour la bonne précision des modèles prédictifs. Les données externes permettent d'améliorer un peu plus la précision des modèles bien que cela reste marginal car les phénomènes du réseau qu'elles expliquent sont déjà contenues dans les données 4G.

L'usage des modèles d'apprentissage comme système de recommandation pour le déploiement ne fait pas l'unanimité selon la nature de la donnée à prédire. Le trafic est une grandeur extensive pour laquelle la connaissance du volume 4G est suffisante pour déployer en priorité sur les sites présentant une forte demande de connectivité. Dans ce contexte, les modèles prédictifs ne sont intéressants que pour anticiper la quantité de trafic 5G attendue. En comparaison, le débit est une grandeur intensive pour laquelle il est moins évident d'utiliser la connaissance du débit 4G pour déterminer les sites dont l'amélioration de la qualité de service serait la plus importante.

Enfin, les travaux de ce chapitre laissent plusieurs pistes de recherche :

- l'amélioration des prédictions des valeurs extrêmes, en passant par une investigation plus approfondie de l'algorithme SMOTE (choix du nombre de voisins, des valeurs minoritaires,...) ou par l'étude d'autres méthodes d'augmentation de données (par exemple, avec des auto-encodeurs).
- la sélection de variables du tissu urbain « sur mesure » pour le terrain étudié. L'objectif serait de trouver une méthode permettant d'exclure les variables expliquant les mêmes phénomènes que les données mobiles. L'analyse canonique des corrélations pourrait être une piste d'étude.
- le développement d'une validation croisée géographique. Dans ce chapitre, les données test ont été fixées à l'avance. Pour évaluer les performances des modèles sur des zones de test différentes, il serait possible de programmer une forme de validation croisée à k blocs géographique. On pourrait définir une forme géométrique (un cercle de rayon fixe par exemple) qui serait positionnée aléatoirement sur la carte. Les données test seraient constituées des performances des sites ou des cellules situées l'intérieur de la forme. Cette opération pourra être répétée plusieurs fois (jusqu'à ce que toutes les données aient été testées au moins une fois par exemple) pour obtenir des statistiques sur les erreurs de prédiction.

Chapitre 9

Conclusion et perspectives

9.1 Bilan

De nombreuses études socio-économiques sont menées en amont d'un déploiement pour estimer la demande de connectivité des territoires (enquêtes de terrain, analyse de données cartographiques, interprétation de phénomènes, etc.). Grâce aux données massives collectées par l'opérateur sur le fonctionnement de ses réseaux, il est naturel de mettre l'intelligence artificielle au service de ces travaux d'anticipation pour faciliter le traitement des données et estimer plus précisément la demande. Les modèles d'apprentissage mis au point dans la thèse visent à prédire les performances d'un réseau mobile avant qu'il ne soit déployé, et les conséquences d'une densification avant qu'elle ne soit mise en œuvre. Pour rendre ces applications possible, il a fallu constituer des jeux d'entraînement informatifs et penser à des représentations de données adaptées à l'apprentissage automatique. Une approche novatrice a consisté à inverser le sens d'étude classique de la littérature qui exploitait les données mobiles pour étudier l'utilisation des sols. À la place, il a été proposé d'utiliser la représentation exhaustive du tissu urbain fournie par un service cartographique pour estimer les performances du réseau mobile.

Extension de couverture Le premier volet de la thèse a été consacré au développement de modèles d'apprentissage pour estimer le nombre d'utilisateurs connectés à des futures stations de base.

En l'absence de connaissance sur les performances du réseau dans une zone encore non couverte, les données cartographiques et démographiques ont été les seules informations sur lesquelles s'appuyer pour entraîner les modèles. Au départ, les prédictions consistaient simplement à donner la classe d'activité d'une station de base (Chapitre 3). Par la suite, ces travaux ont été étendus pour prédire le profil horaire de l'affluence sur une semaine type (Chapitre 5). Pour cette étude, un grand nombre de modèles d'apprentissage ont été évalués. Les données d'entraînement étant de nature tabulaire et leur volume relativement modeste, les modèles utilisant des arbres de décision (forêt aléatoire, CatBoost, LightGBM, XGBoost) ont été les plus précis. Les temps d'entraînement parfois longs et l'évaluation d'un modèle particulier (transformeur) ont conduit à analyser le fonctionnement de méthodes de réduction et de reconstitution des données mobiles (Chapitre 4).

Ces travaux ont également abouti au développement d'une interface graphique permettant d'interagir avec les données cartographiques pour interroger les modèles prédictifs sur l'usage du réseau à l'emplacement sélectionné par l'utilisateur. Cette démonstration montre le potentiel d'intégrer l'IA comme un outil d'aide au déploiement en permettant de mieux estimer la demande de connectivité d'un territoire.

Modèle de couverture cellulaire Dans les contributions précédentes, la plus petite échelle à laquelle pouvaient être réalisées les prédictions était celle de la station de base. En effet, le croisement des données externes avec les données du réseau nécessite un modèle de couverture, qui jusque là était le diagramme de Voronoi appliqué aux positions des sites.

Dans le Chapitre 6, les régions de ce diagramme ont été redimensionnées puis subdivisées pour proposer un modèle de couverture cellulaire différencié par bandes de fréquences. Le redimensionnement des formes géométriques est paramétré par un rapport d'agrandissement. Celui-ci est choisi de sorte que la couverture modélisée se rapproche le plus de la distribution des positions de rattachement des utilisateurs qui ont été collectés pour l'étude. En l'absence de telles données, les résultats ont montré qu'il était possible d'ajuster la valeur du rapport sur des données décrivant la distribution des distances des utilisateurs agrégées par cellule.

La mise au point d'un modèle de couverture cellulaire a permis de créer des données d'entraînement à cette échelle. Le gain est double : une granularité spatiale plus fine, et un volume de données plus abondant car il n'y a plus de besoin d'agréger les données au niveau des stations de base. Ces résultats ont été utilisés directement pour le deuxième volet de la thèse portant sur la densification du réseau.

Densification des réseaux mobiles Le deuxième volet de la thèse est consacré à l'entraînement des modèles d'apprentissage pour la prédiction d'indicateurs de qualité de service et de trafic suite à la mise à jour des secteurs des sites existants. Deux problématiques ont été étudiés.

La première problématique porte sur la réutilisation de fréquences de technologies antérieures pour les technologies les plus récentes (Chapitre 7). L'usage de l'apprentissage automatique a pour objectif de prioriser les ajouts capacitifs sur les secteurs pouvant bénéficier de la plus forte amélioration de la qualité de service ou de la baisse de congestion. Pour cela, les modèles ont été entraînés à prédire, pour chaque secteur, l'évolution de la disponibilité des ressources et de la proportion des ressources allouée par utilisateur en réaction au déploiement de nouvelles fréquences.

La deuxième problématique porte sur le déploiement d'une nouvelle technologie sur les sites de la technologie dominante actuelle (Chapitre 8). On considère deux approches de priorisation de déploiement : la rentabilité (prédiction du volume mensuel du nouveau trafic) et l'amélioration de la qualité de service (prédiction de l'évolution du débit moyen).

Dans les deux cas, les données du tissu urbain sont complétées par les indicateurs de performance des cellules concernées par la mise à jour. La précision des modèles automatiques a été non seulement évaluée avec les mesures classiques comparant les prédictions aux valeurs réelles, mais aussi avec des mesures empruntées aux systèmes de recommandation pour discuter de leur fiabilité à proposer un ordre de déploiement.

9.2 Perspectives

Deux perspectives majeures se dégagent des résultats de la thèse et pourront faire l'objet de nouveaux travaux de recherche : l'adaptation des méthodes pour le renforcement de la couverture mobile dans les zones rurales et le développement d'une méthode guidée par la donnée pour automatiser la recherche des emplacements et des configurations optimales de déploiement.

9.2.1 Études des zones rurales

Les déploiements étudiés par la thèse se sont principalement portés sur des zones urbaines. Néanmoins, un grand enjeu est placé sur le renforcement de la couverture des zones rurales. En France par exemple, l'Arcep estime que 25% des sites qui seront mis à jour vers la 5G 3500 MHz devront être situés dans une zone rurale ou industrielle. Grâce à ces déploiements et au renforcement de la 4G, le débit des réseaux mobiles devra être quatre fois plus élevé que les obligations actuelles sur tout le territoire [157]. Pour assister le déploiement des zones rurales, les méthodes présentées dans les Chapitres 3 et 5 peuvent être réutilisés pour prédire les performances des futurs sites. Afin d'anticiper correctement tous les effets d'un déploiement, il sera peut-être nécessaire d'adapter le modèle de couverture des sites et/ou des cellules au contexte rural, ainsi que de prédire l'impact de ce déploiement sur les sites voisins.

Adaptation de la couverture des sites/cellules La distribution des clients ruraux étant différente de celle des clients urbains, il pourra être intéressant d'évaluer si le diagramme de Voronoi est toujours pertinent, ou si d'autres géométries sont plus adaptées à la modélisation de la couverture de service dans ces zones. Si de nouveaux modèles devaient être mis au point, les méthodes du Chapitre 6 pourraient être utilisées pour évaluer leur précision sur les positions d'utilisateurs ruraux.

Prédiction de l'impact des déploiements sur les sites voisins Pour les problématiques de densification en zone rurale, une extension du Chapitre 7 serait d'analyser si l'augmentation de la capacité d'un site influence les performances de ses plus proches voisins. Le cas échéant, il pourra être intéressant de prédire ces impacts, qui pourraient être un facteur important pour le choix de l'emplacement des futurs déploiements. Cette partie n'a pas été étudiée en milieu urbain car on suppose que le réseau est tellement sollicité qu'un ajout de capacité est totalement absorbé par les cellules co-sectorisées.

9.2.2 Vers une automatisation des déploiements

Les modèles entraînés dans la thèse peuvent être utilisés pour anticiper les performances attendues d'un déploiement futur, ou bien ordonner chronologiquement des déploiements pour maximiser des objectifs de rentabilité ou de qualité de service. En revanche, il est nécessaire de connaître l'emplacement de la station de base à déployer, ou le nombre et le type de fréquences à ajouter sur un secteur. L'intégration des modèles prédictifs à un système automatisant la recherche des meilleurs emplacements ou configurations de déploiement est la deuxième perspective majeure. Celle-ci reprend les réflexions sur le positionnement de la thèse dans le domaine de la planification réseau, qui sont présentés dans le Chapitre 2, Section 2.4. On reprend dans la Figure 9.1 l'architecture de la méthode de planification guidé par la donnée qui était alors présenté dans l'état de l'art. Les blocs en vert sont les composantes de la méthode qui devront faire l'objet d'études approfondies en lien avec une problématique de déploiement.

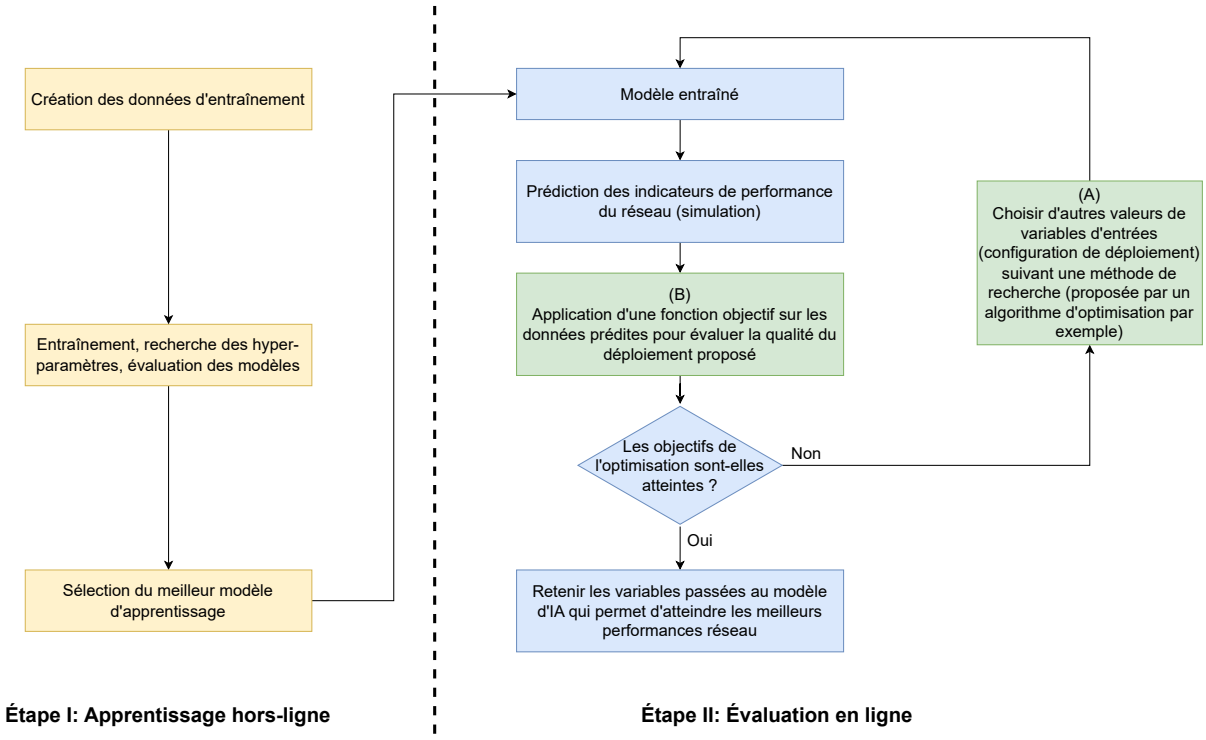


FIGURE 9.1 – Architecture globale d'un système de planification utilisant comme données de simulation les prédictions d'un modèle d'apprentissage. Les blocs en vert font l'objet de perspectives de travaux de recherche futurs.

Recherche des meilleurs emplacements pour le déploiement de nouvelles stations de base

Les Chapitres 3, 5 proposent des méthodes pour prédire de l'affluence et de la charge d'une future station de base, en connaissant à l'avance son emplacement. Il est possible d'aller plus loin en interfaçant ces modèles avec un algorithme de recherche (Figure 9.1, bloc vert A) trouvant automatiquement l'emplacement idéal qui maximise une fonction objectif (Figure 9.1, bloc vert B). Cette fonction pourra prendre en entrée l'activité mobile attendue (prédiction du modèle d'apprentissage) à l'emplacement de déploiement choisi par l'algorithme de recherche.

Si l'on souhaite intégrer dans la fonction objectif l'évolution des performances des sites proches de la nouvelle station de base, il est possible de prendre avantage du modèle de couverture des stations de base. Lors de l'ajout d'un nouveau site, l'actualisation du diagramme de Voronoi modifiera les variables associées aux sites voisins. Il y a donc des chances que les performances des sites voisins soit changées, auquel cas le modèle d'apprentissage pourra prédire le nouveau profil d'activité.

Une implémentation naïve du déploiement est de discrétiser l'espace de recherche en un nombre fini de positions de déploiements possibles, d'itérer sur chaque position de déploiement en calculant à chaque fois le score en fonction des prédictions et de choisir la position maximisant le score. La position initiale du site pourrait être déployée suivant une heuristique (installation à l'endroit le plus densément peuplé

par exemple). Pour rendre l'algorithme plus performant, on pourra filtrer les positions de déploiement selon des critères topographiques. On peut par exemple ne proposer comme emplacements possibles que des zones urbaines, habitées ou touristiques lorsqu'il s'agit de zones peu densément peuplées.

Recherche de configurations optimales

Les Chapitres 7 et 8 permettent de prédire les performances des équipements affectés par l'augmentation de capacité des secteurs existants. Dans ces deux chapitres, il a été proposé de calculer des scores de déploiement reflétant l'amélioration des performances des secteurs, pour réaliser les déploiements dans l'ordre décroissant des scores. Les déploiements considérés sont l'ajout de fréquences, dont les configurations étaient celles des données test. Pour aller plus loin, on pourrait utiliser un algorithme de recherche pour explorer toutes les configurations de fréquences ajoutées possibles, et sélectionner la configuration la plus adaptée à une fonction objectif donnée (maximisation du trafic attendu, minimisation de la congestion, amélioration du débit...).

Fonction objectif Dans les Chapitres 7 et 8, on étudie à plusieurs reprises des scénarios où l'on cherche à prioriser les déploiements améliorant le plus rapidement la qualité de service ou la disponibilité des ressources des secteurs. Ce choix peut être discuté, et des études supplémentaires peuvent être nécessaires pour mettre au point une fonction objectif plus proche des besoins opérationnels. En effet, durant les travaux de recherche, on a rencontré des secteurs présentant une évolution négative des performances suite à l'ajout de nouvelles cellules. Les raisons de ces dégradations mériteraient d'être investiguées : est-ce l'ajout capacitif n'a pas été suffisant par rapport à la demande ? Y a-t-il un problème de configuration de l'antenne (par exemple au niveau de l'orientation des secteurs, de l'inclinaison) ? S'il s'agit de la première raison, elle pourrait être résolue par la méthode de planification en trouvant la configuration apportant suffisamment de capacité (telle que l'évolution soit positive, ou la moins négative possible). En revanche, la deuxième raison est plutôt un problème d'ordre opérationnel.

Ainsi, la continuité des travaux de cette thèse nous semble s'inscrire dans la planification par la proposition d'une solution permettant de simuler certains aspects du comportement du réseau mobile en modélisant implicitement la demande des utilisateurs, au moyen de données reflétant l'activité humaine.

Annexe A

Interface graphique : Déploiement de réseau mobile assisté par l'IA

Pour explorer les application possibles des travaux du Chapitre 5, une interface graphique interactive a été développée pour faciliter le passage des variables d'entrée aux modèles pré-entraînés.

Cas d'étude L'interface a été développée pour montrer l'utilité de l'IA dans un scénario d'extension de couverture. Des études préalables ont permis d'identifier une zone d'absence ou de faible couverture 4G au Sénégal. Les Figures A.1 et A.2 illustrent le fonctionnement de l'outil au sud du Lac Rose, où un pôle urbain en pleine construction fait présager un renforcement nécessaire de la couverture mobile. Grâce à la disponibilité de données historiques du réseau, plusieurs modèles d'apprentissage sont entraînés sur les indicateurs de charge des stations de bases, les données cartographiques et démographiques de la région de Dakar.

Mise en œuvre Une fois entraînés, l'utilisateur peut demander aux modèles de prédire la charge d'une station de base future en interagissant avec l'interface graphique. Elle est composée d'une carte du Sénégal sur laquelle l'utilisateur peut dessiner la forme de la couverture service de la future station. Une fois cette donnée saisie, la page s'actualise et affiche les prédictions de charge pour la station de bases (Figure A.1, droite).

Cette annexe donne une documentation technique de l'outil de démonstration développé de bout en bout, du stockage des données et des modèles au développement de l'application web.

A.0.1 Frontend : Une page web interactive

Le frontend correspond à la partie de l'application visible par l'utilisateur et avec laquelle il peut interagir. Il s'agit ici de l'interface qui permet au client d'envoyer des requêtes au serveur et d'afficher les résultats de celle-ci (Figure A.1). Comme la grande majorité des pages web actuelles, elle est développée en HTML/CSS/JavaScript.

HTML (*HyperText Markup Language*) Le HTML 5 est utilisé pour structurer le contenu de la page en définissant les en-têtes, les titres, les paragraphes, les balises, les boutons, les tableaux, etc.

CSS (*Cascading Style Sheets*) Le CSS est utilisé pour la présentation des éléments HTML. Il permet de définir la fonte, la taille et la couleur du texte et des boutons par exemple. Des bibliothèques sont disponibles pour faciliter la mise en forme ; ici, on utilise Orange Boosted. [1]

JavaScript Le JavaScript est utilisé dans la démonstration pour rendre la page interactive, transmettre les requêtes de l'utilisateur au serveur, récupérer sa réponse et afficher les résultats. Pour cela, on s'appuie sur les bibliothèques suivantes :

— Mapbox GL JS : Mapbox est un service qui fournit des fonctions pour afficher des fonds de cartes ainsi que des bibliothèques pour personnaliser les interactions avec l'affichage, comme Mapbox GL

1. https ://boosted.orange.com/

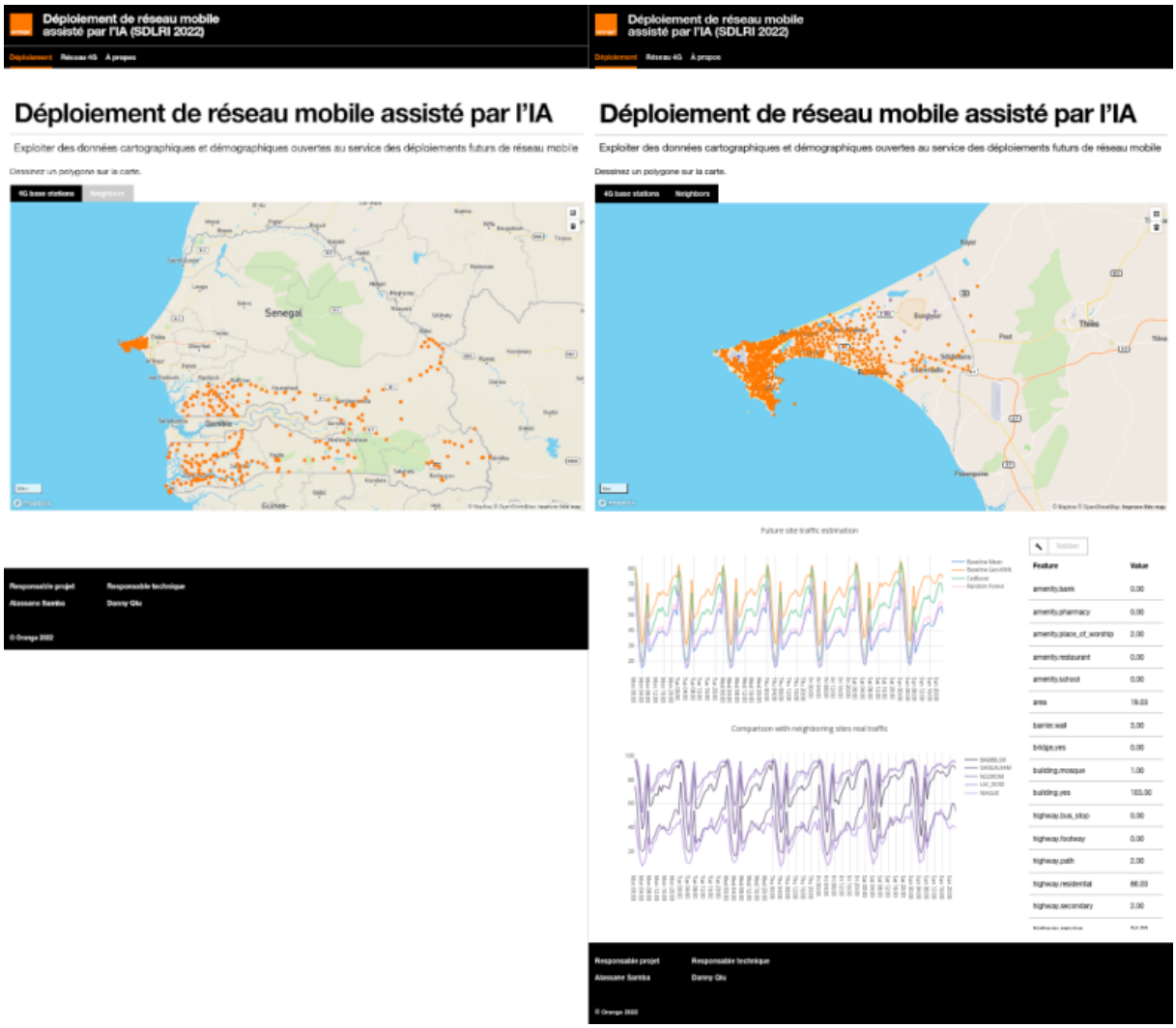


FIGURE A.1 – Captures d'écran de l'interface graphique. À gauche : page d'accueil de l'application. À droite : Affichage des résultats suite à une requête pour connaître le trafic d'une future station de base à l'endroit demandé par l'utilisateur

JS [2]. On utilise mapbox-gl-draw [3] pour permettre aux utilisateurs de dessiner un polygone sur la carte et au serveur de récupérer les coordonnées géographiques des sommets (Figure A.3).

— plotly.js : Cette bibliothèque permet d'afficher interactivement les graphes des signatures hebdomadaires médianes prédites par les modèles pré-entraînés.

L'interaction entre les modules JavaScript et le backend est le suivant :

1. l'utilisateur utilise l'outil proposé par mapbox-gl-draw pour dessiner un polygone sur la carte
2. lorsque le dernier point du polygone est posé, les coordonnées du polygone sont envoyées au serveur
3. le serveur retourne les objets OSM et la population couvertes par le polygone et les affiche dans un tableau (Figure A.4). Ce même tableau est passé en entrée aux modèles prédictifs. Le serveur retourne aussi le données correspondant aux prédictions de CatBoost, de la forêt aléatoire et des lignes de base B2 KNN et B1 AVG (Figure A.3). Elles sont affichées grâce aux fonctions de plotly.js. À titre d'information, on affiche également les signatures réelles des cinq sites les plus proches.
4. Chaque fois que l'utilisateur redessine un nouveau polygone sur la carte, l'affichage est mis à jour pour afficher les prédictions du dernier polygone dessiné.

2. https ://docs.mapbox.com/mapbox-gl-js/guides/
3. https ://github.com/mapbox/mapbox-gl-draw

FIGURE A.2 – Capture d'écran d'un polygone dessiné avec l'outil Mapbox Draw.

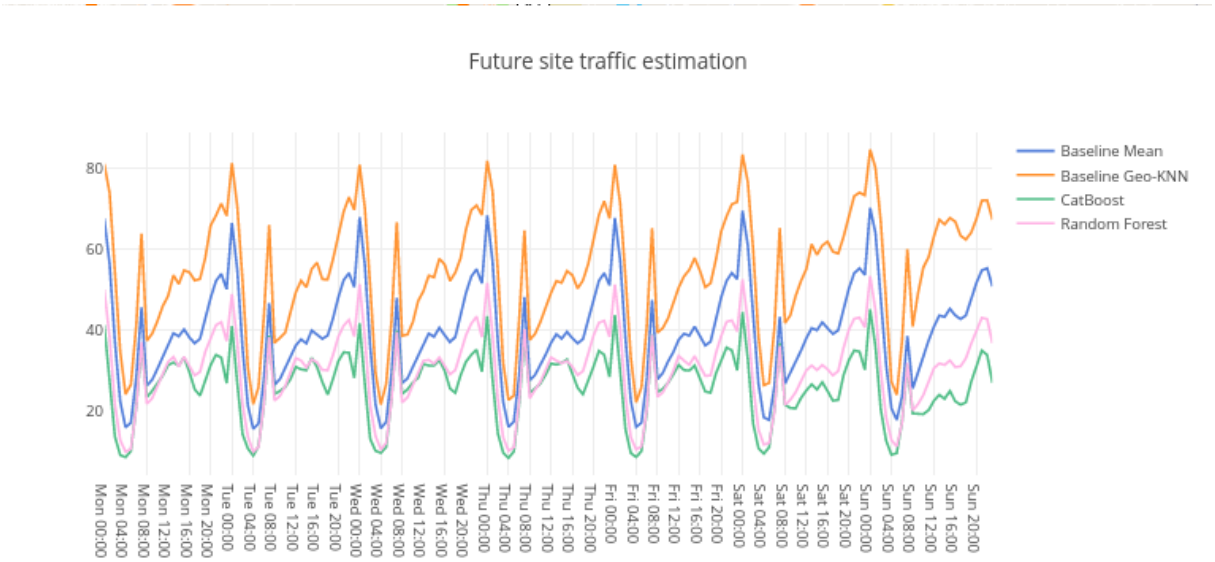


FIGURE A.3 – Graphes des prédictions affichés avec plotly.js. Une courbe correspond à la prédiction d'un modèle

A.0.2 Backend : Stockage, réception des requêtes, traitement et transmission des données

Le backend correspond à la partie de l'application qui n'est pas visible par l'utilisateur. Elle traite les requêtes du client, récupère les données stockées, appelle les modèles de machine learning pour faire les prédictions et renvoie les résultats au frontend.

Les données mobiles et cartographiques sont stockées dans une base de donnée PostgreSQL/PostGIS et le chargement des modèles entraînés est géré par MLflow. Le backend étant majoritairement écrit en Python, la Web Server Gateway Interface utilisée est Flask. Les rôles de ces trois composants sont développés ci-dessous.

PostgreSQL/PostGIS Comme expliqué dans la Section 2.3 de l'état de l'art, PostgreSQL est un système de gestion de base de données et PostGIS est une extension apportant des fonctionnalités géomé-

FIGURE A.4 – Tableau des entrées des modèles prédictifs (données couvertes par le polygone)

triques. Lorsqu'une requête est reçue par le serveur, celle-ci est traduite en requête SQL pour récupérer les données OpenStreetMap stockées dans la base de donnée.

MLflow MLflow propose un ensemble de fonctions permettant aux utilisateurs de sauvegarder et de versionner les modèles entraînés ainsi que des données diverses comme les mesures des performances du modèle, les données d'apprentissage, les hyperparamètres, etc. Les données tabulaires sont stockées dans une base de donnée PostgreSQL et les modèles dans un répertoire de l'espace de stockage.

Flask est une bibliothèque implémentant une *Web Server Gateway Interface* (WSGI). Elle définit les spécifications pour la communication entre le côté serveur et le côté client, notamment au niveau de la gestion des requêtes.

Les requêtes traitées par le backend sont les suivantes :

- au chargement la page, il récupère et envoie les positions des stations de base au frontend qui les affiche sur la carte
- il réceptionne les coordonnées des sommets du polygone tracé par l'utilisateur et envoie requête SQL à la base de données pour :
  - renvoyer le tableau des données cartographiques couvertes par le polygone au frontend.
  - envoyer ces mêmes données aux modèles prédictifs, puis renvoyer au frontend les prédictions à afficher.
- l'interface web offre la possibilité à l'utilisateur d'éditer le tableau des variables. Lorsque la modification est confirmée, le serveur récupère les nouvelles variables du tableau et les passe aux modèles prédictifs, puis renvoie leurs sorties au client.

Bibliographie

[1] *Measuring digital development - Facts and Figures*. ITU, 2022. URL : file:///home/jrjv3136/T%C3%A9l%C3%A9chargements/d-ind-ict_mdd-2022-pdf-e.pdf.

[2] *Urgent, Effective Action Required to Quell the Impact of COVID-19 on Education Worldwide*. The World Bank, jan. 2021. URL : https://www.worldbank.org/en/news/immersive-story/2021/01/22/urgent-effective-action-required-to-quell-the-impact-of-covid-19-on-education-worldwide.

[3] *Achieving universal and meaningful digital connectivity*. ONU, ITU, 2021. URL : https://www.itu.int/itu-d/meetings/statistics/wp-content/uploads/sites/8/2022/04/UniversalMeaningfulDigitalConnectivityTargets2030_BackgroundPaper.pdf.

[4] Phil GOLDSTEIN. *TeliaSonera launches first commercial LTE network*. Déc. 2009. URL : https://www.fiercewireless.com/wireless/teliasonera-launches-first-commercial-lte-networkf.

[5] *Video Traffic Update*. URL : https://www.ericsson.com/en/reports-and-papers/mobility-report/dataforecasts/traffic-by-application.

[6] *La 5G : une nouvelle technologie pour les réseaux mobiles*. ARCEP, 2019. URL : https://www.arcep.fr/fileadmin/cru-1677573101/reprise/dossiers/collectivites/ateliers-TC-2019/atelier-TC-5G-part01-260619.pdf.

[7] *Attribution des fréquences en métropole*. Arcep, juill. 2023. URL : https://www.arcep.fr/la-regulation/grands-dossiers-reseaux-mobiles/la-5g/attribution-des-frequences-en-metropole.html.

[8] *Les services de communications électroniques en France*. Arcep, déc. 2021. URL : https://www.arcep.fr/fileadmin/cru-1677573101/reprise/observatoire/marchan2020/obs-marches-annee-2020-definitif_decembre2021.pdf.

[9] *5G-era Mobile Network Cost Evolution*. GSMA, août 2019. URL : https://www.gsma.com/futurenetworks/wiki/5g-era-mobile-network-cost-evolution/.

[10] *The Mobile Economy*. GSMA, 2023. URL : https://www.gsma.com/mobileeconomy/wp-content/uploads/2023/03/270223-The-Mobile-Economy-2023.pdf.

[11] Isaac K. KASSOUWI. *Le Sénégal déploiera la 5G commerciale cette année*. Agence Ecofin, avr. 2023. URL : https://www.agenceecofin.com/gestion-publique/1504-107446-le-senegal-deploiera-la-5g-commerciale-cette-annee.

[12] *Baromètre du numérique édition 2022*. CGE, Arcep, Arcom, ANCT, 2022. URL : https://www.economie.gouv.fr/files/files/directions_services/cge/barometre-numerique-2022.pdf?v=1684931616.

[13] *Suivi du New Deal Mobile*. Arcep, juin 2023. URL : https://www.arcep.fr/cartes-et-donnees/suivi-du-new-deal-mobile.html.

[14] Frédéric LAUNAY. *La 2G – Un peu de technique*. Août 2011. URL : https://blogs.univ-poitiers.fr/f-launay/2011/08/23/la2g-unpeudetechnique/.

[15] Frédéric LAUNAY. *Et la 3G ?* Sept. 2011. URL : https://blogs.univ-poitiers.fr/f-launay/2011/09/01/etla3g/.

[16] Frédéric LAUNAY. *Chapitre 1 : L'architecture du réseau de mobiles 4G*. Fév. 2021. URL : https://blogs.univ-poitiers.fr/f-launay/2021/02/07/cours-iut-chapitre-1-part-1/.

[17] Frédéric Launay. *1.2.1. Le réseau d'accès radioélectrique NG-RAN*. Sept. 2021. URL : https://blogs.univ-poitiers.fr/f-launay/2021/09/06/livre-5g-nr-chapitre-1-architecture-fonctionnelle/.

[18] *Les attributions de fréquences aux opérateurs de réseaux mobiles ouverts au public*. Arcep, nov. 2020. URL : https://www.arcep.fr/fileadmin/cru-1677573101/reprise/dossiers/frequences/attributions-frequences-operateurs-mobiles-metropole-novembre2020.pdf.

[19] Olivier Rioul. *Qu'est-ce que la théorie de l'information ?* Juin 2021. URL : https://culturemath.ens.fr/thematiques/probabilites/qu-est-ce-que-la-theorie-de-l-information.

[20] Gianni Barlacchi et al. « A multi-source dataset of urban life in the city of Milan and the Province of Trentino ». In : *Scientific data* 2.1 (2015), p. 1-15.

[21] Vincent D Blondel et al. « Data for development : the d4d challenge on mobile phone data ». In : *arXiv preprint arXiv :1210.0137* (2012).

[22] Yves-Alexandre de Montjoye et al. « D4D-Senegal : The Second Mobile Phone Data for Development Challenge ». In : *arXiv :1407.4885 [physics]* (juill. 2014). arXiv : 1407.4885. URL : http://arxiv.org/abs/1407.4885 (visité le 24/09/2021).

[23] Filippo Maria Bianchi et al. « An overview and comparative analysis of recurrent neural networks for short term load forecasting ». In : *arXiv preprint arXiv :1705.04378* (2017).

[24] Chih-Wei Huang, Chiu-Ti Chiang et Qiuhui Li. « A study of deep learning networks on mobile traffic forecasting ». In : *2017 IEEE 28th annual international symposium on personal, indoor, and mobile radio communications (PIMRC)*. IEEE. 2017, p. 1-6.

[25] Qingtian Zeng et al. « Traffic Prediction of Wireless Cellular Networks Based on Deep Transfer Learning and Cross-Domain Data ». In : *IEEE Access* 8 (2020). Conference Name : IEEE Access, p. 172387-172397. ISSN : 2169-3536. DOI : 10.1109/ACCESS.2020.3025210.

[26] Chuanting Zhang et al. « Citywide cellular traffic prediction based on densely connected convolutional neural networks ». In : *IEEE Communications Letters* 22.8 (2018), p. 1656-1659.

[27] Xin Lu et al. « Approaching the limit of predictability in human mobility ». In : *Scientific reports* 3.1 (2013), p. 2923.

[28] Lee Fiorio et al. « Analyzing the effect of time in migration measurement using georeferenced digital trace data ». In : *Demography* 58.1 (2021), p. 51-74.

[29] 3GPP. *Evolved Universal Terrestrial Radio Access (E-UTRA) ; Physical channels and modulation*. Technical Specification (TS) 36.211. Version 17.2.0. 3rd Generation Partnership Project (3GPP), juin 2022.

[30] 3GPP. *Technical Specification Group Radio Access Network ; NR ; Physical channels and modulation*. Technical Specification (TS) 38.211. Version 17.4.0. 3rd Generation Partnership Project (3GPP), jan. 2023.

[31] Mathuranathan Viswanathan. *5G NR Resource block*. Fév. 2022. URL : https://www.gaussianwaves.com/2022/02/5g-nr-resource-block/.

[32] 3GPP. *Technical Specification Group Radio Access Network ; NR ; User Equipment (UE) radio access capabilities*. Technical Specification (TS) 38.306. Version 17.4.0. 3rd Generation Partnership Project (3GPP), mars 2023.

[33] Chaoyun Zhang, Paul Patras et Hamed Haddadi. « Deep learning in mobile and wireless networking : A survey ». In : *IEEE Communications surveys & tutorials* 21.3 (2019), p. 2224-2287.

[34] Ali Yadavar Nikravesh et al. « Mobile Network Traffic Prediction Using MLP, MLPWD, and SVM ». In : *2016 IEEE International Congress on Big Data (BigData Congress)*. Juin 2016, p. 402-409. DOI : 10.1109/BigDataCongress.2016.63.

[35] Md Salik Parwez, Danda B Rawat et Moses Garuba. « Big data analytics for user-activity analysis and user-anomaly detection in mobile wireless network ». In : *IEEE Transactions on Industrial Informatics* 13.4 (2017), p. 2058-2065.

[36] Song Wang et al. « Machine learning in network anomaly detection : A survey ». In : *IEEE Access* 9 (2021), p. 152379-152396.

[37] 3GPP. *Technical Specification Group Radio Access Network ; Radio network planning aspects.* Technical Report (TR) 43.030. Version 17.0.0. 3rd Generation Partnership Project (3GPP), mars 2022.

[38] 3GPP. *Technical Specification Group Radio Access Network ; Study on channel model for frequencies from 0.5 to 100 GHz.* Technical Report (TR) 38.901. Version 17.0.0. 3rd Generation Partnership Project (3GPP), mars 2022.

[39] Han-Shin Jo et al. « Path loss prediction based on machine learning techniques : principal component analysis, artificial neural network, and Gaussian process ». In : *Sensors* 20.7 (2020), p. 1927.

[40] Segun I Popoola et al. « Optimal model for path loss predictions using feed-forward neural networks ». In : *Cogent Engineering* 5.1 (2018), p. 1444345.

[41] Marco Sousa et al. « Analysis and Optimization of 5G Coverage Predictions Using a Beamforming Antenna Model and Real Drive Test Measurements ». In : *IEEE Access* 9 (2021), p. 101787-101808.

[42] Nektarios Moraitis et al. « Performance evaluation of machine learning methods for path loss prediction in rural environment at 3.7 GHz ». In : *Wireless Networks* 27.6 (2021), p. 4169-4188.

[43] Rongrong He et al. « Random forests based path loss prediction in mobile communication systems ». In : *2020 IEEE 6th International Conference on Computer and Communications (ICCC)*. IEEE. 2020, p. 1246-1250.

[44] Zhengqing Yun et Magdy F Iskander. « Ray tracing for radio propagation modeling : Principles and applications ». In : *IEEE access* 3 (2015), p. 1089-1100.

[45] Xin Zhang et al. « Cellular network radio propagation modeling with deep convolutional neural networks ». In : *Proceedings of the 26th ACM SIGKDD International Conference on knowledge discovery & data mining*. 2020, p. 2378-2386.

[46] Leire Azpilicueta et al. « A ray launching-neural network approach for radio wave propagation analysis in complex indoor environments ». In : *IEEE Transactions on Antennas and Propagation* 62.5 (2014), p. 2777-2786.

[47] Larry Medsker et Lakhmi C Jain. *Recurrent neural networks : design and applications*. CRC press, 1999, p. 2-3.

[48] Yann LeCun et al. « Backpropagation applied to handwritten zip code recognition ». In : *Neural computation* 1.4 (1989), p. 541-551.

[49] Wei Wang et al. « End-to-end encrypted traffic classification with one-dimensional convolution neural networks ». In : *2017 IEEE international conference on intelligence and security informatics (ISI)*. IEEE. 2017, p. 43-48.

[50] Sepp Hochreiter et Jürgen Schmidhuber. « Long short-term memory ». In : *Neural computation* 9.8 (1997), p. 1735-1780.

[51] Chaoyun Zhang et Paul Patras. « Long-term mobile traffic forecasting using deep spatio-temporal neural networks ». In : *Proceedings of the Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing*. 2018, p. 231-240.

[52] Dor Bank, Noam Koenigstein et Raja Giryes. « Autoencoders ». In : *arXiv preprint arXiv :2003.05991* (2020).

[53] Chuangfei Liu et al. « Clustering analysis of urban fabric detection based on mobile traffic data ». In : *Journal of Physics : Conf. Series*. T. 1453. 1. 2020, p. 012158.

[54] Ashish Vaswani et al. « Attention is all you need ». In : *Advances in neural information processing systems*. 2017, p. 5998-6008.

[55] scikit-learn developers. *Decision Trees - Multi-output problems*. Accédé le 2023-08-06. url : `https://scikit-learn.org/stable/modules/tree.html#multi-output-problems`.

[56] Yoav Freund et Robert E Schapire. « A decision-theoretic generalization of on-line learning and an application to boosting ». In : *Journal of computer and system sciences* 55.1 (1997), p. 119-139.

[57] Leo Breiman. « Random forests ». In : *Machine learning* 45 (2001), p. 5-32.

[58] Tin Kam Ho. « A data complexity analysis of comparative advantages of decision forest constructors ». In : *Pattern Analysis & Applications* 5 (2002), p. 102-112.

[59] F. Pedregosa et al. « Scikit-learn : Machine Learning in Python ». In : *Journal of Machine Learning Research* 12 (2011), p. 2825-2830.

[60] Nikolai Stepanov et al. « Applying machine learning to LTE traffic prediction : Comparison of bagging, random forest, and SVM ». In : *2020 12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*. IEEE. 2020, p. 119-123.

[61] Liudmila Ostroumova Prokhorenkova et al. « CatBoost : unbiased boosting with categorical features ». In : *NeurIPS*. 2018.

[62] Guolin Ke et al. « Lightgbm : A highly efficient gradient boosting decision tree ». In : *Advances in Neural Information Processing Systems* 30 (2017), p. 3146-3154.

[63] Tianqi Chen et Carlos Guestrin. « Xgboost : A scalable tree boosting system ». In : *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery and Data Mining*. 2016, p. 785-794.

[64] Syahidah Izza Rufaida et al. « Construction of an indoor radio environment map using gradient boosting decision tree ». In : *Wireless Networks* 26 (2020), p. 6215-6236.

[65] Amir Ghasemi. « Data-driven prediction of cellular networks coverage : An interpretable machine-learning model ». In : *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE. 2018, p. 604-608.

[66] Arthur E Hoerl et Robert W Kennard. « Ridge regression : Biased estimation for nonorthogonal problems ». In : *Technometrics* 12.1 (1970), p. 55-67.

[67] Evelyn Fix. *Discriminatory analysis : nonparametric discrimination, consistency properties*. T. 1. USAF school of Aviation Medicine, 1985.

[68] Thomas Cover et Peter Hart. « Nearest neighbor pattern classification ». In : *IEEE transactions on information theory* 13.1 (1967), p. 21-27.

[69] Harris Drucker et al. « Support vector regression machines ». In : *Advances in neural information processing systems* 9 (1996).

[70] Tristan Fletcher. « Support vector machines explained ». In : *Tutorial paper* (2009), p. 1-19.

[71] Lampros Mouselimis. *ClusterR : Gaussian Mixture Models, K-Means, Mini-Batch-Kmeans, K-Medoids and Affinity Propagation Clustering*. R package version 1.3.1. 2023. url : `https://CRAN.R-project.org/package=ClusterR`.

[72] Víctor Soto et Enrique Frías-Martínez. « Automated land use identification using cell-phone records ». In : *Proc. 3rd ACM Int. Workshop MobiArch*. 2011, p. 17-22.

[73] Thiri Aung et al. « Identification and Classification of Land Use Types in Yangon City by Using Mobile Call Detail Records (CDRs) Data ». In : *Journal of the Eastern Asia Society for Transportation Studies* 13 (2019), p. 1114-1133. doi : `10.11175/easts.13.1114`.

[74] Ed Manley et Adam Dennett. « New forms of data for understanding urban activity in developing countries ». In : *Applied Spatial Analysis and Policy* 12.1 (2019), p. 45-70.

[75] Timothy O Hodson. « Root-mean-square error (RMSE) or mean absolute error (MAE) : When to use them or not ». In : *Geoscientific Model Development* 15.14 (2022), p. 5481-5487.

[76] Pavel Senin. « Dynamic time warping algorithm review ». In : *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA* 855.1-23 (2008), p. 40.

[77] Romain Tavenard et al. « Tslearn, A Machine Learning Toolkit for Time Series Data ». In : *Journal of Machine Learning Research* 21.118 (2020), p. 1-6. url : `https://github.com/tslearn-team/tslearn/`.

[78] Chuanting Zhang et al. « Deep transfer learning for intelligent cellular traffic prediction based on cross-domain big data ». In : *IEEE Journal on Selected Areas in Communications* 37.6 (2019), p. 1389-1401.

[79] Jie Feng et al. « Deeptp : An end-to-end neural network for mobile cellular traffic prediction ». In : *IEEE Network* 32.6 (2018), p. 108-115.

[80] Lingcheng Dai et Hongtao Zhang. « Propagation-model-free base station deployment for mobile networks : Integrating machine learning and heuristic methods ». In : *IEEE Access* 8 (2020), p. 83375-83386.

[81] *WGS 84*. EPSG Dataset : v10.094. IOGP Geomatics Committee. URL : https://epsg.org/ellipsoid_7030/WGS-84.html.

[82] *Que sont les données raster ?* Accès le 08-08-2023. ArcGIS Pro. URL : https://pro.arcgis.com/fr/pro-app/2.9/get-started/what-is-raster-data.htm.

[83] ISO/IEC 13249-3 :2016. *Information technology — Database languages — SQL multimedia and application packages — Part 3 : Spatial*. 2023. URL : https://www.iso.org/standard/60343.html.

[84] OpenStreetMap contributors. *Data from http ://download.geofabrik.de/*. 2023. URL : https://www.openstreetmap.org.

[85] QGIS Development Team. *QGIS Geographic Information System*. QGIS Association. URL : https://www.qgis.org.

[86] PostgreSQL Global Development Group. *PostgreSQL : The World's Most Advanced Open Source Relational Database*. 2023. URL : https://www.postgresql.org/.

[87] PostGIS Project Steering Committee et al. *PostGIS, spatial and geographic objects for postgreSQL*. URL : https://postgis.net.

[88] *geopandas : Python tools for geographic data*. Accessed : 2021-02-08. URL : https://github.com/geopandas/geopandas.

[89] *Shapely - Manipulation and analysis of geometric objects in the Cartesian plane.* Accessed : 2023-08-08. URL : https://github.com/shapely/shapely.

[90] *Facebook, Humanitarian Data Exchange*. Accessed : 2021-10-25. URL : https://data.humdata.org/organization/facebook.

[91] INSEE. *Revenus, pauvreté et niveau de vie en 2017 - Données carroyées*. 2022. URL : https://www.insee.fr/fr/statistiques/6215138?sommaire=6215217.

[92] Annisa Sarah. « A multi-dimensions data traffic forecasting model for rural areas ». In : *2018 Int. Conf. ICT Rural Development (IC-ICTRuDev)*, p. 1-6.

[93] Bartłomiej Błaszczyszyn, Miodrag Jovanovicy et Mohamed Kadhem Karray. « How user throughput depends on the traffic demand in large cellular networks ». In : *2014 12th Int. Symp. on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*. IEEE. 2014, p. 611-619.

[94] Jessica Moysen, Lorenza Giupponi et Josep Mangues-Bafalluy. « A machine learning enabled network planning tool ». In : *2016 IEEE 27th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*. IEEE. 2016, p. 1-7.

[95] Mohd Fazuwan Ahmad Fauzi et al. « Mobile Network Coverage Prediction based on Supervised Machine Learning Algorithms ». In : *IEEE Access* (2022).

[96] *Most of the world population is covered by a mobile-broadband signal, but blind spots remain.* ITU. URL : https://www.itu.int/itu-d/reports/statistics/2021/11/15/mobile-network-coverage/.

[97] Jameson L Toole et al. « Inferring land use from mobile phone activity ». In : *Proc. ACM SIGKDD Int. Workshop Urban Computing*. 2012, p. 1-8.

[98] Maxime Lenormand et al. « Comparing and modelling land use organization in cities ». In : *Royal Society open science* 2.12 (2015), p. 150449.

[99] Tao Pei et al. « A new insight into land use classification based on aggregated mobile phone data ». In : *Int. Journal of Geographical Information Science* 28.9 (2014), p. 1988-2007.

[100] Angelo Furno et al. « A tale of ten cities : Characterizing signatures of mobile traffic in urban areas ». In : *IEEE Transactions on Mobile Computing* 16 (2016), p. 2682-2696.

[101] S. E. Hammami et al. « Network planning tool based on network classification and load prediction ». In : *2016 IEEE Wireless Communications and Networking Conference*. 2016, p. 1-6. DOI : 10.1109/WCNC.2016.7565166.

[102] *CatBoost : open-source gradient boosting library*. Accessed : 2021-02-08. URL : https://catboost.ai/.

[103] *Overpass API*. Accessed : 2021-02-08. URL : https://overpass-api.de/.

[104] *overpy : Python Overpass Wrapper*. Accessed : 2021-02-08. URL : https://github.com/DinoTools/python-overpy.

[105] Jeffrey PENNINGTON, Richard SOCHER et Christopher D. MANNING. « GloVe : Global Vectors for Word Representation ». In : *Empirical Methods in Natural Language Processing (EMNLP)*. 2014, p. 1532-1543. URL : http://www.aclweb.org/anthology/D14-1162.

[106] Pauli VIRTANEN et al. « SciPy 1.0 : Fundamental Algorithms for Scientific Computing in Python ». In : *Nature Methods* 17 (2020), p. 261-272. DOI : 10.1038/s41592-019-0686-2.

[107] Scott M LUNDBERG et Su-In LEE. « A Unified Approach to Interpreting Model Predictions ». In : *Advances in Neural Information Processing Systems 30*. Sous la dir. d'I. GUYON et al. Curran Associates, Inc., 2017, p. 4765-4774. URL : http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[108] *GitHub - slundberg/shap : A game theoretic approach to explain the output of any machine learning model*. Accessed : 2021-02-08. URL : https://github.com/slundberg/shap/.

[109] Aditya RAMESH et al. « Zero-shot text-to-image generation ». In : *arXiv preprint arXiv :2102.12092* (2021).

[110] Svante WOLD, Kim ESBENSEN et Paul GELADI. « Principal component analysis ». In : *Chemometrics and intelligent laboratory systems* 2.1-3 (1987), p. 37-52.

[111] Hervé BOURLARD et Yves KAMP. « Auto-association by multilayer perceptrons and singular value decomposition ». In : *Biological cybernetics* 59.4-5 (1988), p. 291-294.

[112] Vinod NAIR et Geoffrey E HINTON. « Rectified linear units improve restricted boltzmann machines ». In : *Proceedings of the 27th international conference on machine learning (ICML-10)*. 2010, p. 807-814.

[113] Andrew L MAAS, Awni Y HANNUN, Andrew Y NG et al. « Rectifier nonlinearities improve neural network acoustic models ». In : *Proc. icml*. T. 30. 1. Atlanta, Georgia, USA. 2013, p. 3.

[114] Bernd JÄHNE. *Digital image processing*. 2005.

[115] Diederik P KINGMA et Max WELLING. « Auto-encoding variational bayes ». In : *arXiv preprint arXiv :1312.6114* (2013).

[116] James LUCAS et al. « Understanding posterior collapse in generative latent variable models ». In : (2019).

[117] Yuhta TAKIDA et al. « Preventing oversmoothing in VAE via generalized variance parameterization ». In : *Neurocomputing* 509 (2022), p. 137-156.

[118] Charlie SNELL. *How is it so good ? (DALL-E Explained Pt. 2)*. Avr. 2021. URL : https://mlberkeley.substack.com/p/dalle2.

[119] *DALL-E in Pytorch*. URL : https://github.com/lucidrains/DALLE-pytorch.

[120] Eric JANG, Shixiang GU et Ben POOLE. « Categorical reparameterization with gumbel-softmax ». In : *arXiv preprint arXiv :1611.01144* (2016).

[121] Noelia CÁCERES, Francisco G. BENÍTEZ et Luis M. ROMERO. « Land use inference from mobility mobile phone data and household travel surveys ». en. In : *Transportation Research Procedia*. 22nd EURO Working Group Transportation Meeting, EWGT 2019, 18th – 20th September 2019, Barcelona, Spain 47 (jan. 2020), p. 417-424. ISSN : 2352-1465.

[122] Ravid SHWARTZ-ZIV et Amitai ARMON. « Tabular data : Deep learning is not all you need ». In : *Information Fusion* 81 (2022), p. 84-90.

[123] Xin HUANG et al. « Tabtransformer : Tabular data modeling using contextual embeddings ». In : *arXiv preprint arXiv :2012.06678* (2020).

[124] Vadim BORISOV et al. « Deep neural networks and tabular data : A survey ». In : *arXiv preprint arXiv :2110.01889* (2021).

[125] Inkit PADHI et al. « Tabular transformers for modeling multivariate time series ». In : *2021 IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, p. 3565-3569.

[126] Hanen BORCHANI et al. « A survey on multi-output regression ». In : *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery* 5.5 (2015), p. 216-233.

[127] Donna Xu et al. « Survey on multi-output learning ». In : *IEEE Transactions on Neural Networks and Learning Systems* 31.7 (2019), p. 2409-2429.

[128] Phil Wang. *DALL-E in Pytorch*. URL : https://github.com/lucidrains/DALLE-pytorch (visité le 20/08/2021).

[129] 3GPP. *Technical Specification Group Radio Access Network ; Radio measurement collection for Minimization of Drive Tests (MDT) ; Overall description ;Stage 2*. Technical Specification (TS) 37.320. Version 17.3.0. 3rd Generation Partnership Project (3GPP), mars 2023.

[130] José N Portela et Marcelo S Alencar. « Cellular coverage map as a voronoi diagram ». In : *Journal of Communication and Information Systems* 23.1 (2008).

[131] Yassine Hmamouche et al. « New trends in stochastic geometry for wireless networks : A tutorial and survey ». In : *Proceedings of the IEEE* 109.7 (2021), p. 1200-1252.

[132] Fengli Xu, Pengyu Zhang et Yong Li. « Context-aware real-time population estimation for metropolis ». In : *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 2016, p. 1064-1075.

[133] Sachit Mishra, Zbigniew Smoreda et Marco Fiore. « Second-level digital divide : A longitudinal study of mobile traffic consumption imbalance in france ». In : *Proceedings of the ACM Web Conference 2022*. 2022, p. 2532-2540.

[134] Freddy Demeersman. « Assessing the Quality of Mobile Phone Data as a Source of Statistics ». In : *European Conference on Quality in Official Statistics* (2016).

[135] Eduardo Graells-Garrido, Oscar Peredo et José García. « Sensing Urban Patterns with Antenna Mappings : The Case of Santiago, Chile ». In : *Sensors* 16.7 (2016). ISSN : 1424-8220. DOI : 10.3390/s16071098. URL : https://www.mdpi.com/1424-8220/16/7/1098.

[136] Orlando E Martínez-Durive et al. « VoronoiBoost : Data-driven Probabilistic Spatial Mapping of Mobile Network Metadata ». In : *2022 19th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE. 2022, p. 100-108.

[137] R Tyrrell Rockafellar. *Convex analysis*. T. 11. Princeton university press, 1997.

[138] GEOS contributors. *GEOS coordinate transformation software library*. Open Source Geospatial Foundation. 2021. URL : https://libgeos.org/.

[139] Ronald L. Graham. « An efficient algorithm for determining the convex hull of a finite planar set ». In : *Info. Proc. Lett.* 1 (1972), p. 132-133.

[140] Herbert Edelsbrunner, David Kirkpatrick et Raimund Seidel. « On the shape of a set of points in the plane ». In : *IEEE Transactions on information theory* 29.4 (1983), p. 551-559.

[141] Matt Duckham et al. « Efficient generation of simple polygons for characterizing the shape of a set of points in the plane ». In : *Pattern recognition* 41.10 (2008), p. 3224-3236.

[142] V. H. Mac Donald. « Advanced mobile phone service : The cellular concept ». In : *The Bell System Technical Journal* 58.1 (1979), p. 15-41. DOI : 10.1002/j.1538-7305.1979.tb02209.x.

[143] Caleb Phillips, Douglas Sicker et Dirk Grunwald. « A survey of wireless path loss prediction and coverage mapping methods ». In : *IEEE Communications Surveys & Tutorials* 15.1 (2012), p. 255-270.

[144] Marceau Coupechoux. *PERFORMANCES 5G : ÉTUDE COMPARÉE EN ZONES RURALES ET URBAINES*. Rapp. tech. Telecom Paris, Institut Polytechnique de Paris, 2021.

[145] Marceau Coupechoux. *Bilans de liaison : de la 2G à la 5G*. Rapp. tech. Telecom Paris, Institut Polytechnique de Paris, 2021. URL : https://marceaucoupechoux.wp.imt.fr/files/2021/05/RIO207-BdL-2G-4G-5G.pdf.

[146] John D Roth, Murali Tummala et James W Scrofani. « Cellular synchronization assisted refinement (CeSAR) : A method for accurate geolocation in LTE-A networks ». In : *2016 49th Hawaii International Conference on System Sciences (HICSS)*. IEEE. 2016, p. 5842-5850.

[147] Boris Delaunay et al. « Sur la sphere vide ». In : *Izv. Akad. Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk* 7.793-800 (1934), p. 1-2.

[148] INSEE. *Découpage communal. Table d'appartenance géographique des communes et tables de passage*. 2023. URL : https://www.insee.fr/fr/information/2028028.

[149] INSEE. *Urban unit*. 2023. URL : https://www.insee.fr/en/metadonnees/definition/c1501.

[150] Fengli Xu et al. « Big data driven mobile traffic understanding and forecasting : A time series approach ». In : *IEEE Transactions on Services Computing* 9.5 (2016), p. 796-805.

[151] Rashmi Korlakai Vinayak et Ran Gilad-Bachrach. « Dart : Dropouts meet multiple additive regression trees ». In : *Artificial Intelligence and Statistics*. PMLR. 2015, p. 489-497.

[152] Takuya Akiba et al. « Optuna : A Next-generation Hyperparameter Optimization Framework ». In : *Proc. 25rd ACM SIGKDD Int. Conf. on Knowl. Discovery and Data Mining*. 2019.

[153] *Communiqué de presse - Qualité de service*. Arcep, oct. 2022. URL : https://www.arcep.fr/actualites/actualites-et-communiques/detail/n/qualite-des-services-mobiles-201022.html.

[154] *Observatoire des déploiements 5G*. Arcep, juin 2023. URL : https://www.arcep.fr/cartes-et-donnees/nos-cartes/5g/derniers-chiffres.html.

[155] Nitesh V Chawla et al. « SMOTE : synthetic minority over-sampling technique ». In : *Journal of artificial intelligence research* 16 (2002), p. 321-357.

[156] Rita P Ribeiro et Nuno Moniz. « Imbalanced regression and extreme value prediction ». In : *Machine Learning* 109 (2020), p. 1803-1835.

[157] *Parlons 5G : toutes vos questions sur la 5G*. Arcep, sept. 2022. URL : https://www.arcep.fr/nos-sujets/parlons-5g-toutes-vos-questions-sur-la-5g.html.

ECOLE
DOCTORALE

Titre : Nouvelles méthodes d'apprentissage automatique pour la planification des réseaux mobiles

Mots clés : apprentissage automatique, réseau mobile, tissu urbain, extension de couverture, densification

Résumé : La connectivité mobile est un moteur important de nos sociétés, c'est pourquoi l'usage des données mobiles n'a cessé de croître à travers le monde. Pour éviter la saturation, les opérateurs mobiles sont amenés à faire évoluer leurs réseaux. Les réseaux mobiles sont renforcés grâce à l'installation de nouvelles stations de base et antennes. Ce chantier étant très coûteux, une grande attention est accordée à l'identification des déploiements les plus rentables et permettant d'être compétitif. L'objectif de la thèse est d'utiliser apprentissage automatique pour proposer des solutions améliorant les décisions de déploiement.
Le premier volet de la thèse est consacré au développement de modèles d'apprentissage pour assister le déploiement des stations de base sur de nouveaux emplacements. En l'absence de connaissance réseau d'une zone non couverte, les modèles sont entraînés en s'appuyant entièrement sur des données du tissu urbain. Au départ, les prédictions consistaient simplement à estimer la classe d'activité majoritaire d'une station de base. Par la suite, ces travaux ont été étendus pour prédire le profil horaire type de l'affluence hebdomadaire. Les temps d'entraînement parfois longs ont conduit à analyser plusieurs méthodes de réduction de données mobiles.
Le deuxième volet de la thèse est consacré au déploiement de nouvelles cellules sur des sites existants afin d'augmenter leur capacité. Pour cela, un modèle de couverture cellulaire a été mis au point en dérivant le diagramme de Voronoi modélisant la couverture des stations de base. La première étude a porté sur la réutilisation de fréquences de générations technologiques antérieures pour les générations les plus récentes. Les modèles entraînés ont pour objectif d'aider à prioriser les ajouts capacitifs sur les secteurs pouvant bénéficier de la plus forte amélioration de la disponibilité des ressources. La deuxième étude a porté sur l'ajout d'une nouvelle génération de réseau en considérant deux axes de déploiement : priorisation de la rentabilité ou de l'amélioration de la qualité de service.
Ainsi, les méthodes développées par cette thèse visent à s'intégrer dans un outil de géo-marketing, en proposant des modèles pouvant prédire la demande de connectivité d'un territoire ainsi que son évolution. Ces informations pourront également servir à rendre le dimensionnement des réseaux plus précis.

Title : New machine learning methods for mobile network planning

Keywords : machine learning, mobile network, urban fabric, coverage extension, densification

Abstract : Mobile connectivity is an important driver of our societies, which is why mobile data consumption has continued to grow steadily worldwide. To avoid global congestion, mobile network operators are bound to evolve their networks. Mobile networks are strengthened through the deployment of new base stations and antennas. As this task is very expensive, a great attention is given to identifying cost-effective and competitive deployments. In this context, the objective of this thesis is to use machine learning to improve deployment decisions.
The first part of the thesis is dedicated to developing machine learning models to assist in the deployment of base stations in new locations. Assuming that network knowledge for an uncovered area is unavailable, the models are trained solely on urban fabric features. At first, models were simply trained to estimate the class of major activity of a base station. Subsequently, this work was extended to predict the typical hourly profile of weekly traffic. Since the train time could be long, several methods for reducting mobile data have been studied.
The second part of the thesis focuses on the deployment of new cells to increase the capacity of existing sites. For this purpose, a cell coverage model was developed by deriving the Voronoi diagram representing the coverage of base stations. The first study examined the spectrum refarming of former generations of mobile technology for the deployment of the newest generations. Models are trained to assist in prioritizing capacity additions on sectors that can benefit from the greatest improvement in resource availability. The second study examined the deployment of a new generation of mobile technology, considering two deployment strategies: driven by profitability or by the improvement of the quality of service.
Therefore, the methods developed in this thesis offer ways to train models to predict the connectivity demand of a territory as well as its evolution. These models could be integrated into a geo-marketing tool, as well as providing useful information for network dimensioning.

Institut Polytechnique de Paris
91120 Palaiseau, France