Techniques de comptage et cartographie passives de clients WiFi dans l'ère de la protection des données

Feifei Yang

Sorbonne Université

Ecole doctorale EDITE

Institut Langevin, Ondes et Images / Equipe : New concepts for imaging and sensing

Techniques de comptage et cartographie passives de clients WiFi dans l'ère de la protection des données

Par Feifei YANG

Thèse de doctorat d'Electronique

Dirigée par Bruce DENBY

Présentée et soutenue publiquement le 12/10/2023

Devant un jury composé de :

M. Bruce DENBY, professeur, directeur de thèse

M. Emmanuel VIENNET, professeur, rapporteur

Mme. Samia BOUZEFRANE, professeur, rapporteur

M. Aziz BENLARBI-DELAÏ, professeur, examinateur, président du jury

Dédicace

Je dédie humblement cette thèse à :

Ma famille, pour leur soutien indéfectible, leurs encouragements constants et leur amour inconditionnel tout au long de mon parcours académique.

Mon directeur de thèse, le Professeur Bruce DENBY pour son expertise, sa patience et ses conseils éclairés qui ont enrichi ce travail de recherche.

Mes amis et collègues, pour leur amitié, leur soutien moral et leurs échanges fructueux qui ont contribué à rendre cette expérience mémorable.

Enfin, je dédie cette thèse à toutes les personnes qui œuvrent pour l'avancement du savoir et pour un monde meilleur.

Merci à tous.

Feifei YANG

Résumé :

L'avènement de la législation sur la protection des données (GDPR) a présenté d'importants défis pour les opérateurs de réseaux WiFi, entravant leur capacité à suivre et analyser l'activité des clients à des fins commerciales et de sécurité. La randomisation des adresses MAC des appareils clients lors de chaque transmission a rendu difficile l'agrégation de l'activité des clients en utilisant des trames de gestion telles que les probe requests, pratiques courantes par le passé. Par conséquent, le comptage et la cartographie précis de l'activité des clients sont devenus problématiques, nécessitant souvent des mesures supplémentaires à partir de capteurs alternatifs tels que des caméras.

Cependant, nos recherches ont donné des résultats prometteurs pour relever ces défis. Statistiquement, il a été constaté que le nombre de clients maintient une relation proportionnelle avec le nombre brut de probe requests, offrant ainsi un certain soulagement à la problématique du comptage des clients. Malgré ces avancées, la localisation des clients aux adresses MAC randomisées au sein d'un site réseau reste un problème non résolu.

Pour résoudre ce problème, cette thèse propose également une boîte à outils complète comprenant neuf outils. L'objectif principal de cette boîte à outils est d'étendre la proportionnalité entre le nombre de clients et le nombre des probe requests, permettant ainsi la cartographie précise des densités de clients dans les réseaux WiFi extérieurs du monde réel sans recourir à des mesures de référence. En utilisant cette boîte à outils, les probe requests brutes peuvent être transformées en une carte de densité, fournissant des estimations de la population de clients à chaque emplacement.

En combinant la boîte à outils proposée avec les conclusions de la recherche sur le comptage des clients, cette thèse offre une solution globale aux défis auxquels sont confrontés les opérateurs de réseaux WiFi. La boîte à outils permet une cartographie précise des densités de clients dans les réseaux WiFi extérieurs, en tenant compte des réglementations sur la protection des données et de la randomisation des adresses MAC. Cela permet une meilleure compréhension du comportement des clients tout en respectant les préoccupations de confidentialité et en surmontant les complexités techniques.

Mots clés : suivi de l'audience, RGPD, localisation, probe request,Wifi

Abstract :

The emergence of data privacy legislation, such as the GDPR, has posed significant hurdles for WiFi network operators, impeding their ability to track and analyze client activity for commercial and security purposes. The dynamic randomization of client device MAC addresses during each transmission has created obstacles in aggregating client data using management frames like probe requests, which were commonly employed in the past. Consequently, accurately quantifying and mapping client activity has become a complex task, often necessitating supplementary measurements from alternative sensors such as cameras.

Nevertheless, our research has yielded encouraging results in addressing these challenges. Statistically, we have discovered a proportional relationship between the number of clients and the raw count of probe requests, providing some respite to the issue of client counting. Despite these advancements, the localization of clients with randomized MAC addresses within a network site remains an unresolved predicament.

To confront this issue head-on, this thesis proposes a comprehensive toolkit comprising nine tools. The primary objective of this toolkit is to expand the proportionality between the client count and the probe request count, enabling precise mapping of client densities in real-world outdoor WiFi networks without the reliance on ground truth measurements. By harnessing the potential of this toolkit, raw probe requests can be transformed into a density map, offering estimations of the client population at various locations.

By combining the proposed toolkit with the findings derived from the research on client counting, this thesis presents an all-encompassing solution to the challenges confronted by WiFi network operators. The toolkit facilitates accurate mapping of client densities in outdoor WiFi networks while diligently adhering to data privacy regulations and accounting for MAC address randomization. This comprehensive approach enhances the comprehension of client behavior, while ensuring the preservation of privacy and addressing technical intricacies.

Keywords : audience monitoring, GDPR, localization, probe request,Wifi

Table des illustrations

Table des tableaux

Remerciements

Je tiens à exprimer ma profonde gratitude envers toutes les personnes et les institutions qui ont contribué à la réalisation de cette thèse et qui ont rendu cette aventure académique possible.

Tout d'abord, je souhaite remercier chaleureusement mes directeurs de thèse, le Professeur Bruce DENBY et le Docteur Pierre ROUSSEL, pour leur soutien indéfectible, leur expertise et leurs conseils éclairés. Leurs encouragements constants ont été d'une importance cruciale pour mener à bien cette recherche.

Je suis également reconnaissant envers les membres du laboratoire de recherche, Institut Langevin, pour leur collaboration fructueuse, leurs discussions enrichissantes et leur atmosphère de travail conviviale.

Un grand merci à ma famille pour leur amour inconditionnel et leur soutien tout au long de cette aventure académique. Leur présence et leurs encouragements ont été une source d'inspiration permanente.

Mes amis et collègues méritent également mes remerciements pour leur amitié, leur soutien moral et leurs échanges stimulants qui ont rendu cette expérience encore plus mémorable.

Je tiens tout particulièrement à exprimer ma gratitude envers Aleia, qui a généreusement apporté son soutien financier à cette thèse. Leur contribution a permis de réaliser cette recherche dans des conditions optimales et a ouvert de nouvelles perspectives pour cette étude.

Enfin, je voudrais adresser mes remerciements à toutes les personnes qui ont participé de près ou de loin à cette recherche, ainsi qu'à celles qui, par leurs travaux passés, ont jeté les bases de cette étude.

Liste des acronymes utilisés dans le manuscrit

General Data Protection Regulation (GDPR)

Probe Request (PR)

Internet of Things (IoT)

Enhanced Services (ES)

Peer-to-Peer (P2P),

Access Point (AP)

Time Difference Of Arrival (TDOA)

Radio frequency (RF)

Received Signal Strength Indicator (RSSI)

Channel State Information (CSI)

Wireless Local Area Network (WLAN)

Logical Link Control (LLC)

Medium Access Control (MAC)

Service Set Identifier (SSID)

Organizationally Unique Identifier (OUI)

Network Interface Card (NIC)

Information Element (IE)

Global Positioning System (GPS)

Near Field Communication (NFC)

Software Defined Radio (SDR)

Hidden Markov Model (HMM)

GNU Radio Companion (GRC)

VHSIC Hardware Description Language (VHDL)

Direct Sequence Spread Spectrum (DSSS)

Quadrature Phase Shift Keying (QPSK)

Binary Phase Shift Keying (BPSK)

Message Queuing Telemetry Transport (MQTT)

Medium Access Control addresses (MAC)

Carrier Frequency Offset (CFO)

Orthogonal frequency division multiplexing (OFDM)

In-Phase (I)

Quadrature(Q)

Full Width at Half Maximum (FWHM)

Sommaire

I. Introduction

I.A. Contexte de la thèse

Le suivi de la présence et des déplacements des individus ou des foules dans des zones spécifiques joue un rôle crucial en fournissant des informations utiles sur leurs comportements et en révélant des tendances sous-jacentes. Cela revêt une importance significative dans divers domaines tels que la sécurité publique, les transports, l'aménagement urbain, la gestion des catastrophes et des crises, ainsi que l'organisation d'événements à grande échelle. La capacité à surveiller et suivre les individus et les foules permet la mise en place de politiques et de mesures appropriées, ainsi que le développement de services et d'applications avancés adaptés à des besoins spécifiques.

Une méthode couramment utilisée pour la surveillance consiste à utiliser des caméras. Cependant, les caméras soulèvent souvent des préoccupations en matière de vie privée. De plus, elles sont soumises à certaines limitations techniques telles que les obstacles de ligne de mire, les conditions météorologiques défavorables, l'éclairage faible et les scénarios à fort contraste, ce qui peut entraver leur efficacité.

Une autre option viable est l'utilisation de réseaux de capteurs, qui offrent une large gamme de techniques. Par exemple, la densité de foule peut être déduite en utilisant des capteurs de CO_2 [1], bien qu'ils soient sensibles à la circulation de l'air. Alternativement, il est possible d'exploiter les canaux de communication entre les nœuds de mesure et de calculer l'affaiblissement du signal au sein d'un réseau de capteurs pour estimer la densité de foule ou faire la localisation [2, 3, 4] .

Selon les indicateurs de développement utilisés par la Banque mondiale, le nombre d'abonnements mondiaux aux téléphones mobiles par tranche de 100 personnes a atteint un chiffre significatif de 110 en 2021 [5]. Cette donnée met en évidence l'ampleur de l'utilisation des téléphones mobiles à l'échelle mondiale. En effet, ces téléphones sont devenus des outils indispensables dans la vie quotidienne de nombreuses populations. Ces dernières années, les smartphones sont équipés de plusieurs technologies sans fil, notamment le WiFi et le Bluetooth,

en plus de prendre en charge différentes générations de technologies de réseau mobile (c'est-à-dire de la 2G à la 5G). Ces technologies reposent sur l'échange de messages spécifiques de gestion de réseau avec des dispositifs tels que les stations de base et les points d'accès pour établir des connexions et offrir une connectivité transparente aux réseaux disponibles. En exploitant les données de ces messages et en suivant les appareils que les individus transportent, il est possible de détecter leur présence et leurs schémas de déplacement de manière non intrusive. Il convient de noter que les appareils personnels Bluetooth, tels que les téléphones mobiles, ne diffusent pas leur présence par défaut. En revanche, les appareils WiFi effectuent en continu des analyses et diffusent leurs capacités afin d'optimiser la gestion de l'énergie et faciliter des connexions plus rapides. Par conséquent, le coût abordable des cartes d'interface réseau WiFi a joué un rôle majeur dans l'utilisation du WiFi comme base pour les systèmes de collecte de données passives, ce qui a conduit à des recherches approfondies dans ce domaine au cours des 15 dernières années [6] [7].

Bien que l'accès aux données collectées par les technologies de réseau mobile auprès des opérateurs mobiles soit strictement réglementé par la loi, les technologies sans fil sont susceptibles d'être utilisées de manière abusive. Par conséquent, il est nécessaire de mettre en place des mesures de sécurité [8] appropriées pour prévenir tout accès ou abus non autorisé.

Ces dernières années, la question de la confidentialité des données dans les réseaux WiFi est devenue un sujet de préoccupation et de vigilance croissante. Cette attention accrue peut être attribuée à la mise en œuvre de « General Data Protection Regulation » (GDPR) [9], qui est une réglementation de l'Union européenne visant à protéger la vie privée et les données personnelles des individus, y compris les adresses MAC, et à imposer des règles plus strictes sur leur utilisation. Ces réglementations ont entraîné d'importants changements dans la manière dont les réseaux WiFi effectuent le décompte des clients et la surveillance de l'‘audience, car la pratique traditionnelle qui reposait sur les adresses MAC fixes des clients a été perturbée.

Pour se conformer aux réglementations sur la confidentialité, les systèmes d'exploitation des appareils mobiles ont mis en place la randomisation des adresses MAC afin de protéger la vie privée des utilisateurs. Avec cette approche, chaque fois qu'un appareil émet un « Probe Request » (PR), il se voit attribuer une adresse MAC nouvelle et unique. Cette randomisation

est devenue l'outil de facto pour garantir le respect du GDPR dans les réseaux WiFi et est désormais largement adoptée sur divers appareils mobiles. Bien que l'introduction de la randomisation des adresses MAC renforce la confidentialité en rendant plus difficile le lien entre les adresses MAC et des individus spécifiques, elle pose également de nouveaux défis en matière de surveillance de l'audience et de décompte des clients. Outre la randomisation des adresses MAC, dans le but de renforcer la protection de la vie privée des clients, notamment ceux qui sont déjà connectés au Wi-Fi (révélant ainsi leurs véritables adresses MAC), il n'est pas conservé de trace de l'adresse MAC du client lors de la collecte des données. À la place, celle-ci est substituée par une chaîne de hachage [10] anonyme qui transforme les adresses en valeurs uniques et impossibles à inverser. On appelle cela l'anonymisation qui se différencie de la randomisation [11].

En réponse aux limitations posées par la randomisation des adresses MAC [12] [13] [14] [15] [16] [11] [17], les chercheurs explorent activement différentes stratégies et techniques pour surmonter ces obstacles. Certaines études se sont concentrées sur l'identification des failles de confidentialité et des vulnérabilités du processus de randomisation, dans le but de réaliser une dé-randomisation [13] [18] des adresses MAC. Cependant, à mesure que le consensus autour de la protection des données personnelles se renforce, il devient de plus en plus évident que les solutions doivent être résistantes aux techniques de randomisation. L'accent est désormais mis sur le développement de méthodes préservant la confidentialité qui peuvent estimer avec précision les populations de clients et permettre une surveillance fiable de l'audience sans dépendre des adresses MAC fixes.

Ainsi, la question de la confidentialité des données dans les réseaux WiFi a entraîné des changements importants dans la manière de compter les clients et de surveiller l'audience. La mise en œuvre de réglementations telles que le GDPR a nécessité l'adoption de la randomisation des adresses MAC, ce qui pose des défis pour les pratiques de surveillance traditionnelles. Cependant, les recherches en cours visent à développer des techniques préservant la confidentialité qui peuvent estimer avec précision les populations de clients et faciliter une surveillance fiable de l'audience sans compromettre la vie privée individuelle.

I.B. Enoncé de la problématique

La compréhension du comportement des clients et la localisation précise jouent un rôle essentiel dans de nombreux domaines d'étude. Ces deux aspects, le comptage des clients et la localisation, constituent des éléments clés pour appréhender les défis auxquels sont confrontés les chercheurs. Dans cette section, nous présenterons en détail ces deux sujets cruciaux, en explorant leur importance.

I.B.1. Comptage de Clients

L'approche consistant à exploiter les clients dotés d'adresses MAC fixes pour gérer les adresses MAC randomisés peut sembler attrayante. Ces clients à adresse MAC fixe sont généralement ceux qui ont souscrit au service WiFi local ou dont les systèmes d'exploitation n'effectuent pas de randomisation. L'hypothèse sous-jacente à cette approche est que les statistiques d'émission des demandes de PR de ces clients à adresse MAC fixe seraient similaires à celles des clients à adresse MAC randomisé [13] [14]. Cependant, cette hypothèse rencontre deux défis : la prolifération des dispositifs de « Internet of Things » (IoT) et l'utilisation croissante des « enhanced services » (ES) par les clients. Nous expliquons d'abord ces deux défis.

Le premier défi provient du nombre croissant de dispositifs IoT. Ces dispositifs, allant des appareils électroménagers intelligents aux capteurs industriels, ont des comportements uniques en ce qui concerne l'émission de PR [19]. Contrairement aux clients traditionnels, les dispositifs IoT ont souvent des horaires prédéfinis pour l'envoi de PR ou les transmettent uniquement en réponse à des événements spécifiques. Leurs statistiques d'émission diffèrent considérablement de celles des clients réguliers, ce qui rend difficile de s'appuyer uniquement sur les clients à adresse MAC fixe pour estimer précisément l'audience.

Le deuxième facteur contribuant à la complexité est la présence de clients utilisant des « Enhanced Services » (ES) [20]. Ces clients sont engagés dans des activités telles que le partage de fichiers « Peer-to-Peer » (P2P) [21, 22], la diffusion en continu de vidéos ou les appels vocaux sur IP. Comme ils utilisent des protocoles spécifiques et adoptent des modèles de communication plus dynamiques, leurs statistiques d'émission de PR diffèrent à la fois de celles des clients à adresse MAC fixe et des clients réguliers. Par exemple, les applications P2P

génèrent un grand nombre de PR, ciblant souvent simultanément plusieurs appareils [22]. Ce comportement distinct remet en question l'hypothèse de similarité des statistiques de PR entre les clients à adresse MAC fixe et les clients à adresse MAC randomisée.

Face à ces défis, il est devenu de plus en plus nécessaire d'estimer l'audience uniquement à partir des adresses MAC randomisés [19]. Cependant, accomplir cette tâche est loin d'être trivial. L'objectif est de développer des solutions, même approximatives, pouvant fournir des estimations fiables de l'audience en travaillant avec les informations limitées disponibles à partir des adresses MAC randomisés.

En résumé, les défis posés par les adresses MAC randomisés ont suscité une demande de solutions permettant d'estimer la taille et la composition de l'audience sans s'appuyer sur des clients à adresse MAC fixe. La présence croissante de dispositifs IoT et ES complique davantage le processus d'estimation. Néanmoins, les chercheurs explorent activement des approches innovantes, notamment l'analyse de la structure des trames PR, la modélisation statistique et l'intégration de données, pour développer des solutions approximatives. Ces solutions visent à fournir des estimations fiables de l'audience tout en préservant la vie privée et en respectant les réglementations sur la protection des données.

I.B.2. Cartographie de la densité de clients

Ainsi, si le problème du décompte des clients avec des adresses MAC randomisés est progressivement en train d'être résolu, l'attribution de positions à des groupes de clients anonymisés, en particulier dans des scénarios extérieurs avec des architectures réseau diverses et des environnements de propagation particuliers, présente des difficultés supplémentaires. Par exemple, dans la pratique, de nombreux réseaux extérieurs privilégient la connectivité plutôt que la localisation. Dans le contexte de la localisation, la triangulation est une méthode couramment utilisée. Elle consiste à utiliser les signaux provenant de plusieurs points de référence pour déterminer la position d'un objet ou d'un émetteur. La triangulation se base sur la mesure des distances ou des angles entre ces points de référence et l'objet à localiser. En analysant ces mesures, il est possible d'estimer approximativement la position de l'objet. La réalisation de la localisation, avec les techniques existantes (voir section II.D) nécessite une redondance plus élevée des AP, ce qui peut être coûteux. Par conséquent, une partie importante

des PR émises peut être capturée par un ou deux AP seulement, rendant la triangulation impossible. De plus, les émissions ultérieures comporteront de nouvelles chaînes de hachage d'adresse MAC non vus auparavant, ce qui rend impossible de savoir s'ils proviennent du même appareil que le PR précédent ou d'un client différent.

La localisation dans les réseaux où la localisation n'était pas une priorité de conception présente des limites intrinsèques. Avec en plus les réglementations du GDPR, sauf pour les abonnés connectés, il devient presque impossible de réaliser une localisation. Par conséquent, il existe une demande substantielle d'outils capables de surmonter ces défis et de permettre la localisation.

I.C. Les solutions qui seront avancées

Mon projet de doctorat représente une collaboration entre le laboratoire « Institut Langevin – Ondes et Images » qui est Unité Mixte de Recherche de l'ESPCI Paris et du CNRS dédiée à la physique des ondes et à ses applications, comprenant les ondes mécaniques, ondes électromagnétiques et optique (CNRS UMR 7587) et « Aleia » qui fournit les outils d'analyse pour la mobilité et les transports avec une forte expertise en big data et intelligence artificielle. Nous avons eu l'opportunité de bénéficier d'un ensemble de données fourni par Aleia. Ces données comprennent des informations détaillées sur les PRs, collectées à partir de 200 points d'accès de la marque Ruckus [23] déployés dans plus de 20 villes à travers le territoire français. La durée de cette collecte s'étend sur les deux périodes : de juin à aout 2020 et de novembre 2020 à octobre 2021.

Dans ce contexte, ma thèse se propose d'introduire des méthodologies novatrices visant à relever ces défis et à fournir des estimations approximatives du nombre de clients ainsi que de leurs positions géographiques. Ces approches cherchent à enrichir et à optimiser les techniques existantes afin de résoudre les problèmes liés à l'anonymisation et à la randomisation des données, tous deux étant au cœur des préoccupations majeures en matière de protection de la vie privée.

I.C.1. Solution avancée pour le comptage de client

Mon premier travail de recherche présente une approche statistique qui se concentre uniquement sur la découverte des statistiques sous-jacentes des clients, dans le but de fournir des informations pertinentes sans compromettre la confidentialité des utilisateurs. En exploitant les décomptes quotidiens des PRs et en analysant les motifs clés de l'activité des clients et des tendances saisonnières, cette méthode établit un facteur d'échelle approximatif qui facilite la conversion directe des décomptes bruts des PR en estimations fiables de la population des clients. Nous donnons ici un petit résumé de cette étude que sera abordée en détail en section V.

L'un des points forts notables de cette approche réside dans sa dépendance exclusive des clients randomisés. Cette emphase sur les adresses MAC randomisées garantit une préservation de la confidentialité et atténue efficacement les défis associés aux clients IoT et ES. En opérant uniquement sur les clients randomisés, la méthode contourne les problèmes potentiels de confidentialité et s'éloigne des complexités posées par les appareils non randomisés.

De plus, la polyvalence de la méthode proposée s'étend à l'application de jeux de données du monde réel, acquis auprès de réseaux WiFi commerciaux. En utilisant de telle base de données, la méthode démontre son applicabilité pratique et offre des informations sur le comportement des clients au sein des environnements WiFi commerciaux. La méthode proposée est contraire aux approches conventionnelles qui dépendent de techniques de référence supplémentaires, cette méthode innove en exploitant la puissance de l'analyse statistique pour estimer la taille de l'audience et suivre les tendances.

I.C.2. Solution avancée pour la cartographie de la densité de clients

Dans un deuxième volet, qui sera détaillé en section VI, nous proposons un ensemble d'outils pour faciliter la prétraitement, l'affichage, l'interprétation, la localisation et le décompte des clients des réseaux WiFi extérieurs sans avoir recours à une méthode de vérité terrain. Ces outils sont basés sur des comptages bruts de PR avec des adresses MAC randomisées, complétés de localisations des clients ayant émis des PR avec des techniques qui seront décrites

ultérieurement. L'objectif essentiel de cette boîte à outils est de convertir ces comptages de PR en une carte de densité qui représente le nombre de clients à chaque position.

Ce qui rend ce travail unique et innovant, c'est l'assemblage de tous ces éléments nécessaires pour résoudre un problème complexe que de nombreux opérateurs de réseau ont considéré comme insoluble jusqu'à présent. A travers de cette boîte à outils, nous fournissons une solution pratique et efficace pour comprendre et analyser les réseaux WiFi extérieurs et les comportements des clients.

Cette approche novatrice a été appliquée à des données provenant de deux sites de réseau réels en France, fournissant des résultats intéressants. Actuellement, nos prédictions de densité reposent sur une méthode qui utilise une probabilité d'émission de PRs en fonction de leur position.

La boîte d'outils offre une solution pour l'analyse des réseaux WiFi extérieurs, sans compromettre la confidentialité des utilisateurs. Les résultats obtenus témoignent de l'efficacité de notre approche, ouvrant de nouvelles perspectives de recherche et d'application dans le domaine de la surveillance d'audience.

I.C.3. bladeRF

La randomisation des adresses MAC rend plus difficile le comptage des clients et leur localisation, ce qui nécessite l'utilisation de capteurs complémentaires et/ou d'estimations statistiques pour déterminer le nombre de clients à partir des données PR brutes. Cependant, nous avons pu tirer parti de certaines techniques de notre boîte à outils actuelle pour localiser un seul PR envoyé par un client à adresse MAC randomisée, à condition qu'il soit reçu par au moins 3 points d'accès (AP).

Parallèlement, la tendance actuelle en matière de législation sur la sécurité des données vise à renforcer la confidentialité des clients. Dans un avenir envisageable, les systèmes pourraient exiger une randomisation supplémentaire, y compris pour les PR à adresse MAC randomisée reçus par plusieurs AP. Dans ce scénario, il serait judicieux de découvrir de nouveaux outils afin de pouvoir compter et cartographier les clients malgré ces changements, sans toutefois révéler leurs identités. Dans cette section, nous décrivons une technique conçue pour regrouper

les signaux PR reçus en fonction de leurs caractéristiques physiques, en l'occurrence, une mesure des informations d'état de canal (CSI).

I.D. Structure du manuscrit

La thèse se structure comme suit. Tout d'abord, dans la section II, nous examinons la structure des trames de PR dans les réseaux WiFi, ainsi que l'impact du GDPR sur la collecte de données WiFi. Nous explorons également les concepts de localisation basés sur Channel State Information (CSI) et Received Signal Strength Indicator (RSSI), ainsi que l'utilisation des cartes SDR, notamment bladeRF. Dans la section III l'Etat de l'Art récapitulera les travaux déjà réalisés dans le domaine de notre recherche.

Ensuite, dans la section IV nous nous concentrons sur la base de données que nous avons utilisée dans notre recherche. En plus des données de base, nous l'avons enrichi en ajoutant des informations supplémentaires. Nous avons réalisé une analyse approfondie de ces données et présentons les résultats obtenus. Nous mettons également en évidence les défis rencontrés en matière de localisation lors de l'utilisation de cette base de données.

A ce stade, nous abordons le détail de notre travail. Elle se compose de trois parties distinctes :

1. **Comptage de client (section V)** : modèle déterministe afin de révéler les statistiques sous-jacentes des clients dans les données brutes de PR avec clients randomisés.
2. **Localisation (section VI)** : un ensemble des outils permettant de cartographier les densités de clients dans les réseaux WiFi en extérieur du monde réel
3. **BladeRF (section VII)** : mettre au point un moyen de mesurer en temps réel le CSI des PR DSSS et de rendre cette information disponible pour le regroupement des PR susceptibles, dans une fenêtre temporelle à déterminer, d'être émis par le même client.

Enfin, dans la section VIII, nous résumons l'ensemble de notre travail de doctorat et proposons des perspectives pour les futures recherches. Nous mettons en avant les contributions de notre recherche à la compréhension de la localisation basée sur les réseaux WiFi et soulignons les opportunités de recherche futures visant à améliorer les techniques de localisation, optimiser l'utilisation des données WiFi et explorer de nouvelles applications potentielles.

II. Contexte de la recherche

II.A. IEEE 802.11 Wi-Fi

Notre travail principal se concentre sur la technologie Wi-Fi, largement reconnue sous le nom d'IEEE 802.11 Wi-Fi [24]. Il est un ensemble de normes pour le Wireless Local Area Network (WLAN) qui permettent la transmission de données sans fil entre des appareils électroniques. La technologie WLAN permet aux utilisateurs de se connecter à Internet sans avoir à se connecter à un réseau câblé, ce qui offre une plus grande liberté de mouvement et une plus grande flexibilité. Le Wi-Fi se décline en plusieurs protocoles, se distinguant les uns des autres par des variations de fréquence, de débits de données et de techniques de modulation.

Le Wi-Fi est principalement utilisé pour déployer des réseaux sans fil à courte distance, couvrant plusieurs dizaines à plusieurs centaines de mètres, tels que ceux présents dans les entreprises et les domiciles. Avec l'avènement de l'IoT, le Wi-Fi s'est également imposé comme l'une des clés facilitant la communication avec une large gamme de petits appareils ne disposant pas d'une interface câblée, tels que les smartphones. Les avantages de la communication sans fil par rapport aux alternatives câblées sont multiples et incluent notamment la facilité de déploiement, la réduction des coûts d'infrastructure et la mobilité des appareils.

Dans un réseau WLAN (Wireless Local Area Network), la couche de liaison de données (data link layer) et la couche physique (physical layer) sont deux composantes essentielles de son architecture de communication (fig 1).

La couche de liaison de données comprend deux sous-couches : le Contrôle Logique de Liaison (LLC, Logical Link Control) et le Contrôle d'Accès au Support (MAC, Medium Access Control). Le LLC assure la fourniture de services de transmission de données transparents, en traitant la segmentation et l'assemblage des données de la couche supérieure, ainsi que la détection et la correction des erreurs. La sous-couche MAC gère l'accès et la gestion des trames de données, notamment le contrôle de transmission, la résolution des conflits et l'adressage des trames.

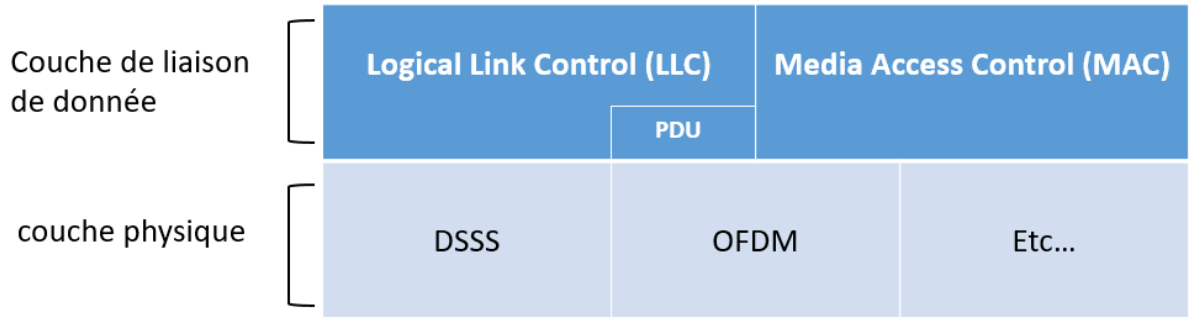


Figure 1 Structure de la pile de protocoles 802.11, la couche de liaison de données comprend LLC et MAC, tandis que la couche physique implique les techniques de modulation OFDM et DSSS.

La couche physique est la couche la plus basse dans un réseau WLAN et elle est responsable de la transmission et de la réception des signaux sans fil. Au sein de la couche physique, deux techniques de modulation courantes sont l'OFDM (Orthogonal Frequency Division Multiplexing) et le DSSS (Direct Sequence Spread Spectrum). Le DSSS est une technique de modulation qui utilise une large bande passante pour transmettre des données en répartissant les bits sur une séquence de codes de *spreading*. Cette technique étend le signal sur une plus grande bande passante que ce qui serait normalement nécessaire pour transmettre les données, ce qui améliore la résistance aux interférences et aux effets de la propagation. Et l'OFDM est une technique de modulation qui divise le spectre de fréquences disponible en plusieurs sous-porteuses orthogonales. Chaque sous-porteuse est modulée indépendamment avec une vitesse de transmission de données plus faible, ce qui améliore l'efficacité spectrale et la tolérance aux interférences.

En ce qui concerne le PR, cette fonctionnalité est implémentée dans la technique de modulation DSSS. Lorsqu'un appareil sans fil envoie un PR pour rechercher les réseaux Wi-Fi disponibles, cette demande est encodée et modulée en utilisant la technique DSSS avant d'être transmise sur le support de communication sans fil.

La PDU (Fig 2), ou *Protocol Data Unit*, est une entité fondamentale dans la communication des données au sein d'un réseau. Elle représente une unité de données encapsulée dans un protocole spécifique. Lorsque nous nous concentrons spécifiquement sur les réseaux Wi-Fi, une

forme courante de PDU est la trame. Les trames Wi-Fi sont utilisées pour l'envoi de données dans les réseaux sans fil. Elles encapsulent les informations essentielles pour la transmission des données, y compris les adresses source et destination, les informations de contrôle, les données elles-mêmes.

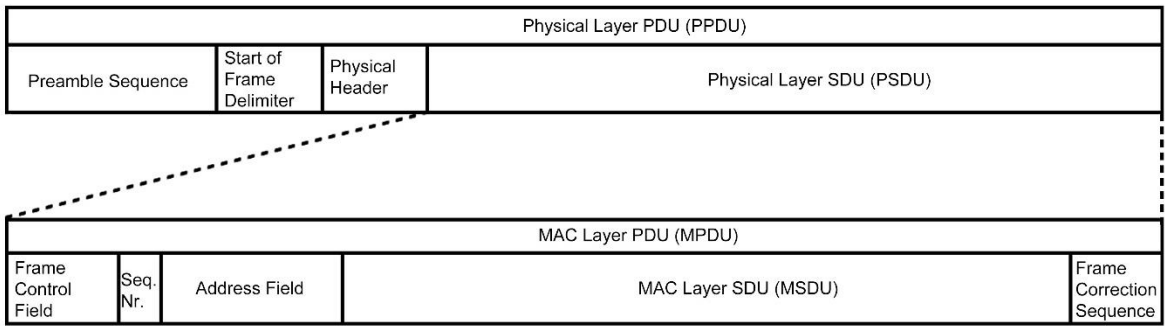


Figure 2 Structure de physical layer protocol data unit

Frame types :

Les trames de données 802.11 sont divisées en plusieurs champs qui contiennent des informations telles que l'adresse MAC de l'émetteur, l'adresse MAC du destinataire, la séquence de numérotation, les données elles-mêmes, et les champs de contrôle et de gestion. Voici un aperçu des différents champs des trames de données 802.11 (fig 3) :

Champs de contrôle : ces champs contiennent des informations sur le type de trame, la durée de transmission, le taux de transmission, le type de cryptage et d'autres informations de contrôle.

Champs de gestion : ces champs contiennent des informations de gestion telles que les identifiants de réseau (SSID), les identifiants de point d'accès (BSSID), les paramètres de canal, les informations de sécurité et autres informations de configuration.

Adresse MAC de l'émetteur : cette adresse identifie l'émetteur de la trame.

Adresse MAC du destinataire : cette adresse identifie le destinataire de la trame.

Séquence de numérotation : ce champ contient un numéro de séquence qui permet au destinataire de reconstituer les trames dans l'ordre approprié.

Données : ce champ contient les données proprement dites qui sont transmises via le réseau sans fil.

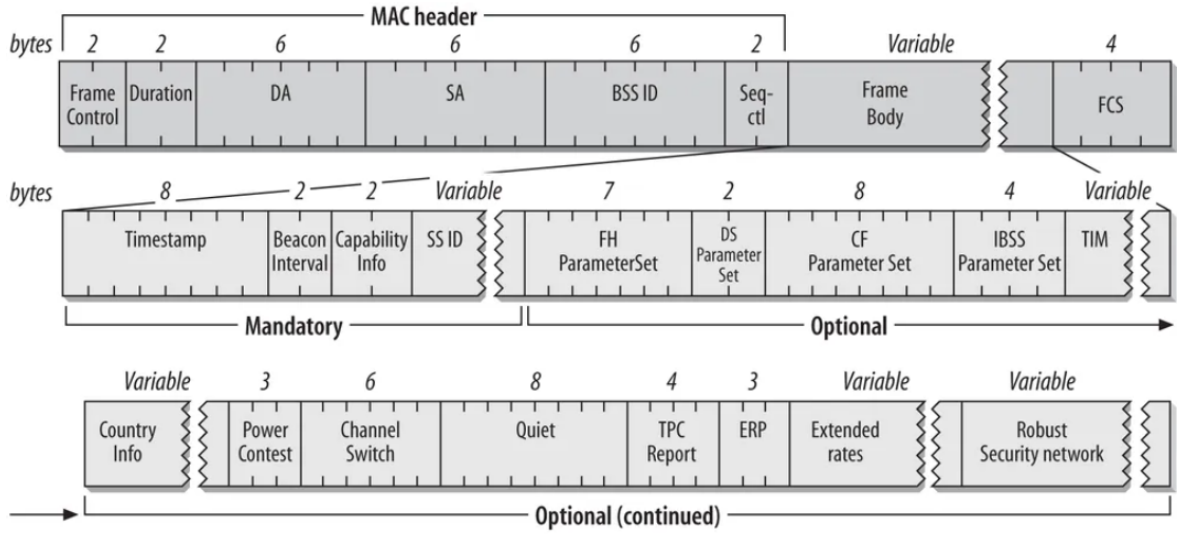


Figure 3 Structure des trames de données 802.11

Les trames de données 802.11 peuvent être utilisées pour différentes fonctions telles que la transmission de données, la découverte de réseaux sans fil, la gestion de la connectivité et la gestion de la sécurité. Parmi ces types, nous nous intéressons particulièrement aux trames de gestion.

Les *Probe Requests* (PRs) sont émises par les stations lors d'une analyse active du réseau. Lorsqu'un PR inclut un *service set identifier* (SSID) non nul, on parle de PR dirigée, et seuls les points d'accès (AP) associés à cet SSID sont censés y répondre. En revanche, lorsque le SSID est nul, le PR est une requête de diffusion, et tous les AP environnants sont supposés y répondre. Probe Requests (PRs) permettent aux périphériques sans fil tels que les ordinateurs portables, les smartphones et les tablettes de rechercher les points d'accès disponibles dans la zone de couverture. Le point d'accès qui reçoit un PR peut répondre en envoyant un message de *Probe Response*, qui contient des informations sur le réseau, telles que la force du signal et les paramètres de sécurité. Les PRs sont envoyées à intervalles réguliers, généralement toutes

les quelques secondes. Cette fréquence peut être modifiée en fonction des réglages du périphérique sans fil et des paramètres de réseau.

Les Beacons, également trames de gestion, sont envoyées périodiquement par les points d'accès pour annoncer leur présence aux stations environnantes. Elles sont utilisées pour une analyse passive du réseau et permettent aux stations de reconnaître les points d'accès disponibles.

La découverte de services :

Lorsqu'une station souhaite se connecter à un réseau Wi-Fi, elle doit suivre un processus d'authentification et d'association. Tant que cette procédure n'est pas entièrement accomplie, le périphérique reste non associé et ne peut échanger que des trames de gestion avec le point d'accès (AP) et d'autres appareils. Pour repérer les AP environnants, les appareils compatibles Wi-Fi utilisent soit un mode de découverte de services actif, soit un mode passif, comme le montre l'illustration de la figure 4. Dans le mode actif, les appareils émettent des trames de gestion PR, auxquelles les AP environnants répondent par des *Probe Responses*. Ces PRs peuvent contenir les noms (SSID) des réseaux auxquels le périphérique souhaite se connecter. Dans le mode passif, les appareils écoutent passivement les Beacons diffusées par les AP, qui annoncent les caractéristiques du réseau Wi-Fi correspondant. La découverte de services active est généralement utilisée par les appareils mobiles en raison de sa consommation d'énergie réduite et de sa vitesse plus rapide par rapport au mode passif. C'est également la seule façon de repérer les AP masqués, c'est-à-dire les AP qui ne signalent pas leur présence en utilisant des Beacons ou en répondant aux PRs diffusées.

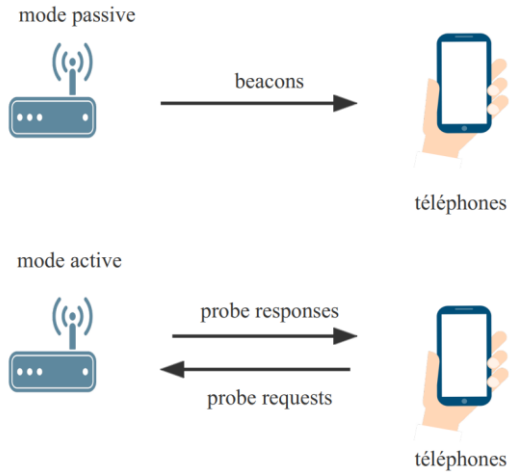


Figure 4 Les deux modes de découverte de service dans IEEE 802.11. Dans le mode passif, les APs propagent des Beacons. Dans le mode actif, la station diffuse des PRs et les APs répondent avec des Probe Response.

Canaux :

Les appareils Wi-Fi fonctionnent généralement sur les bandes de fréquences de 2,4 GHz ou 5 GHz au niveau de la couche physique. En Europe, ces bandes de fréquences sont divisées en 13 bandes chevauchantes allant de 2,400 à 2,4835 GHz, et en 19 bandes non chevauchantes allant de 5,150 à 5,725 GHz. Étant donné qu'un point d'accès fonctionne sur une seule bande, les stations diffusent généralement des PRs sur tous les canaux disponibles afin de découvrir tous les réseaux. Les canaux 1, 6 et 11 sont les plus couramment utilisés.

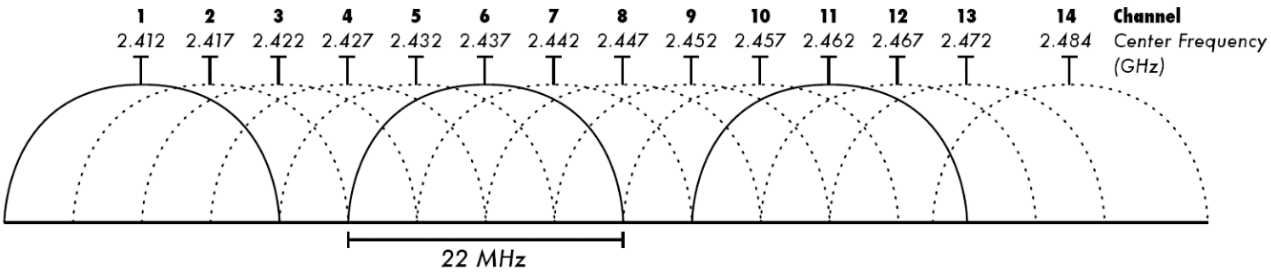


Figure 5 Représentation graphique des canaux WiFi dans la bande des 2.4GHz

Adresse MAC :

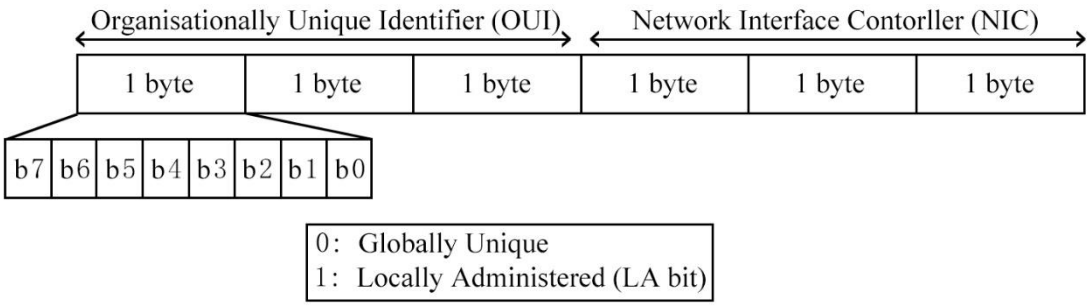


Figure 6 Format d''adresse MAC

Les appareils utilisent une adresse MAC, un identifiant unique de 6 octets, pour communiquer entre eux au niveau de la couche MAC. Comme illustré dans la Figure 6, les trois premiers octets de l'adresse MAC représentent un *Organizationally Unique Identifier* (OUI) que les fabricants doivent obtenir pour garantir l'unicité globale des adresses MAC. Les trois derniers octets correspondent au *network interface card* (NIC). Un bit particulier de l'adresse MAC est le septième bit du premier octet de l'OUI, appelé *Locally Administered Bit* (LA bit). Lorsqu'il est défini à 1, cela indique que l'adresse MAC a été modifiée par l'administrateur de l'appareil et n'est pas garantie d'être unique. Il n'est pas clair que le comportement de configuration du bit LA à 1 soit nécessairement causé par la randomisation de l'adresse MAC, car aucun document ne mentionne explicitement ce cas.

Éléments d'information :

Les PRs transportent des informations supplémentaires dans leur corps de trame sous forme d'*Information Element* (IE). La plupart de ces IE, à l'exception du SSID (nom du réseau) et des débits pris en charge, ne sont pas obligatoires et servent à annoncer la prise en charge de diverses fonctionnalités. Ils sont généralement composés de plusieurs sous-champs dont la taille peut varier, allant d'un simple bit à plusieurs octets. Étant donné leur caractère facultatif, tous les appareils ne les incluent pas, et l'ensemble des IE peut varier d'un appareil à l'autre en fonction de leur configuration et de leurs capacités.

Force du signal :

Lorsque des trames Wi-Fi sont échangées au niveau de la couche physique via des signaux radio, la puissance à laquelle chaque trame est reçue contient des informations pertinentes. Par exemple, elle indique à l'utilisateur la qualité et la stabilité de sa connexion Wi-Fi. La force du signal des trames Wi-Fi reçues est généralement représentée par une mesure appelée *Received Signal Strength Indicator* (RSSI). Cette valeur, exprimée en décibels milliwatts (dBm), varie de -100 à 0 dBm pour WiFi. Une valeur plus élevée en dBm indique un signal plus fort. Le RSSI permet aux utilisateurs et aux appareils de déterminer la force relative des signaux Wi-Fi environnants. Cela peut être utile pour évaluer la qualité de la connexion, identifier les zones à faible couverture, ou la localisation.

II.B. General Data Protection Regulation (GDPR)

De nos jours, la prolifération des appareils mobiles portables émettant en continu des signaux traçables engendre une menace pesant sur la vie privée. Les systèmes de suivi Wi-Fi collectent principalement des données de mobilité sur les individus. De telles données sont très sensibles en termes de confidentialité. Savoir quelques endroits où une personne se trouve peut suffire à l'identifier. Les informations de localisation sont sensibles non seulement parce qu'elles donnent des informations sur la présence d'une personne à un endroit donné à un moment donné, mais aussi en raison de la logique des lieux visités, c'est-à-dire le type de lieux visités et dans quel ordre [25]. Cela peut révéler des informations sur leur consommation (quels magasins fréquentent-ils ?), leur personnalité (quels endroits visitent-ils pour se divertir ?), leurs relations (quelles personnes rencontrent-ils ?), des comportements sensibles spécifiques, etc.

En raison de la préoccupation croissante pour la protection de la vie privée des utilisateurs, l'Europe a adopté la loi GDPR (*General Data Protection Regulation*) en 2016 afin de mettre en place des mesures visant à préserver la confidentialité des données personnelles des utilisateurs et à imposer des contraintes aux entreprises qui les collectent et les gèrent. Par conséquent, il est essentiel de comprendre les principes fondamentaux de la GDPR et ses implications pour garantir la confidentialité et la sécurité des informations sensibles des utilisateurs.

Les principaux objectifs du GDPR sont les suivants :

- Renforcer les droits des individus sur leurs données personnelles : le GDPR donne aux individus le droit d'accéder à leurs données, de les rectifier, de les supprimer et de limiter leur traitement.
- Renforcer la responsabilité des entreprises : les entreprises sont tenues de prendre des mesures pour protéger les données personnelles des individus.
- Renforcer les pouvoirs de contrôle des autorités de protection des données : les autorités de protection des données sont en mesure d'infliger des amendes élevées aux entreprises qui ne respectent pas le GDPR.

Le GDPR a des implications importantes pour toutes les entreprises qui collectent, traitent ou stockent des données personnelles de citoyens européens. Les entreprises qui ne respectent pas le GDPR peuvent faire face à des amendes élevées.

II.C. Randomisation d'adresse MAC

L'adresse MAC est une adresse physique unique qui identifie de manière spécifique chaque appareil réseau. Elle est souvent utilisée pour suivre la localisation des utilisateurs dans des espaces publics tels que les centres commerciaux, les aéroports, les gares, etc. Les appareils mobiles envoient régulièrement des PRs pour rechercher des réseaux Wi-Fi disponibles dans leur environnement.

La randomisation de l'adresse MAC dans les PRs permet de protéger la vie privée des utilisateurs en masquant leur identité réelle. Au lieu d'utiliser leur adresse MAC unique, les appareils utilisent des adresses MAC temporaires qui change régulièrement pour éviter toute détection ou traçabilité. La randomisation de l'adresse MAC ne prévient pas totalement la collecte de données de localisation, car d'autres méthodes comme de-randomisation peuvent être utilisées pour suivre les utilisateurs. Cependant, elle rend plus complexe la collecte de données personnelles, ce qui constitue une étape essentielle dans la protection de la vie privée des utilisateurs. La randomisation de l'adresse MAC est prise en charge par de nombreux systèmes d'exploitation tels qu'iOS, Android et Windows. Cela peut aider les entreprises à se conformer au GDPR en évitant la collecte et le traitement de données personnelles sensibles telles que les adresses MAC. Le GDPR exige également que les entreprises obtiennent le

consentement explicite des utilisateurs avant de collecter et de traiter leurs données personnelles. Si une entreprise souhaite collecter des adresses MAC pour suivre les utilisateurs, elle doit obtenir leur consentement explicite.

II.D. Localisation

La localisation est une discipline de recherche en constante évolution qui vise à déterminer la position d'un objet ou d'un individu dans l'espace. Depuis l'avènement du *Global Positioning System* (GPS) dans les années 1970, les applications de la localisation se sont multipliées, allant de la navigation automobile aux systèmes de surveillance industrielle en passant par les technologies médicales et les réseaux de capteurs sans fil. Cependant, malgré l'utilisation généralisée des systèmes de localisation, la précision et la fiabilité de ces derniers restent des défis importants à relever.

II.D.1. Classification des types de localisation

La localisation peut être classée en différentes catégories en fonction de différents critères. Voici quelques exemples de classification :

II.D.1.a). Critère de la portée:

La localisation peut être basée sur la portée, c'est-à-dire la distance entre l'appareil de localisation et la cible à localiser. Cette catégorie peut être subdivisée en :

- **Localisation à courte portée** : la cible est située à une courte distance de l'appareil de localisation, généralement quelques mètres ou moins. Les technologies utilisées peuvent inclure le Bluetooth, l'ultrason.
- **Localisation à moyenne portée** : la cible est située à une distance plus importante, pouvant aller jusqu'à quelques dizaines de mètres. Les technologies utilisées peuvent inclure le Wi-Fi, le Zigbee ou le RFID (identification par radiofréquence).
- **Localisation à longue portée** : la cible est située à une distance importante, pouvant aller jusqu'à plusieurs kilomètres. La technologie principale utilisée est le GPS.

II.D.1.b). Critère de la précision :

La localisation peut être classée en fonction de la précision requise. Cette catégorie peut être subdivisée en :

- **Localisation de basse précision** : la précision requise est relativement faible, généralement dans une fourchette de quelques mètres ou plus. Les technologies utilisées peuvent inclure le GPS, le Wi-Fi ou le réseau cellulaire.
- **Localisation de haute précision :** la précision requise est très élevée, généralement inférieure à un mètre. Les technologies utilisées peuvent inclure le LiDAR (télédétection par laser) ou la vision par ordinateur.

II.D.1.c). Critère du mode de localisation :

La localisation peut être classée en fonction du mode de localisation. Cette catégorie peut être subdivisée en:

- **Localisation en mode passive** : la localisation se fait sans la participation active de la cible. Les technologies utilisées peuvent inclure la surveillance vidéo, la surveillance acoustique ou la surveillance thermique.
- **Localisation en mode active** : la cible participe activement à la localisation. Les technologies utilisées peuvent inclure le GPS, le Wi-Fi ou le Zigbee.

II.D.2. Choix de la localisation par Wifi

Parmi ces choix, le WiFi répond bien aux exigences de la problématique de la thèse. L'utilisation du Wi-Fi pour la localisation est devenue de plus en plus courante ces dernières années [26, 27]. Les réseaux Wi-Fi sont désormais omniprésents et les appareils mobiles sont équipés de puces Wi-Fi. Cette ubiquité a permis aux développeurs de logiciels de créer des applications qui utilisent les signaux Wi-Fi pour la localisation.

La localisation par Wi-Fi est souvent utilisée en intérieur où le GPS peut ne pas être efficace en raison de la faible pénétration du signal à travers les murs et les obstacles. La localisation par Wi-Fi utilise les signaux Wi-Fi pour déterminer la position d'un appareil mobile ou d'un objet à l'intérieur d'un bâtiment. Les signaux Wi-Fi sont utilisés pour mesurer la force du signal entre

l’appareil mobile et les points d’accès Wi-Fi (AP) dans la zone de couverture. Cette mesure de la force du signal est utilisée pour déterminer la position de l’appareil mobile à l’intérieur de la zone de couverture de ces points d’accès Wi-Fi.

Dans cette méthode, la précision de la localisation peut être améliorée en utilisant plusieurs points d’accès Wi-Fi et en appliquant des techniques de traitement du signal pour filtrer les erreurs de mesure. De plus, l’utilisation de techniques d’apprentissage automatique peut améliorer la précision de la localisation en utilisant des modèles prédictifs pour déterminer la position de l’appareil mobile [28, 29].

La localisation par Wi-Fi est utilisée dans divers domaines, tels que les applications de suivi de flotte de véhicules, la navigation en intérieur, la gestion de l’inventaire. Avec la prolifération de l’Internet des objets (IoT), cette méthode de localisation par Wi-Fi peut être encore plus importante à l’avenir.

Il existe plusieurs techniques couramment utilisées pour la localisation par Wi-Fi. Dans la suite on cite quelques exemples

II.D.2.a) Fingerprinting

Cette technique consiste à mesurer la puissance du signal Wi-Fi à partir de différents points d’accès dans la zone de couverture et à créer une empreinte digitale de la zone. Lorsque l’utilisateur se déplace dans cette zone, les signaux Wi-Fi reçus sont comparés aux empreintes digitales stockées pour déterminer la position de l’utilisateur. La technique de "fingerprinting" (empreinte digitale) est l’une des méthodes les plus courantes utilisées pour la localisation en intérieur à l’aide du Wi-Fi.

Le processus de fingerprinting commence par la collecte de données Wi-Fi à partir de différents points d’accès dans la zone de couverture. Cette collecte de données peut être effectuée à l’aide d’un dispositif mobile équipé d’une antenne Wi-Fi, tel qu’un smartphone ou une tablette, ou d’un dispositif fixe placé dans la zone à couvrir. Les données sont collectées en mesurant la puissance du signal Wi-Fi reçu à partir de chaque point d’accès.

Ces mesures sont ensuite stockées dans une base de données pour créer une carte de couverture Wi-Fi de la zone à couvrir. Cette carte contient une liste de points d'accès Wi-Fi et les niveaux de puissance de signal associés à chaque point d'accès dans chaque zone de la région couverte. Cette carte est ensuite utilisée comme référence pour la localisation future.

Lorsqu'un appareil mobile se déplace dans la zone couverte, il collecte les signaux Wi-Fi des points d'accès à proximité et les compare à la carte de couverture Wi-Fi préalablement créée. À partir de ces comparaisons, l'appareil mobile peut déterminer sa position approximative.

L'un des inconvénients majeurs de la technique de localisation basée sur le fingerprinting est la création et la maintenance d'une base de données. Cette tâche est complexe, chronophage et coûteuse. En outre, la mise à jour régulière de la base de données est essentielle pour maintenir sa précision. Les appareils peuvent être ajoutés, retirés ou modifiés fréquemment, ce qui nécessite une surveillance constante et une mise à jour régulière de la base de données. Cette tâche peut être complexe et exigeante en termes de ressources.

II.D.2.b) Triangulation

La méthode de positionnement par triangulation est une méthode largement étudiée. Cette méthode consiste à former des cercles centrés sur les points d'accès, où le rayon de chaque cercle est déterminé par RSSI. Un point d'intersection se produit lorsqu'il y a trois points d'accès ou plus dans une certaine plage, et le point d'intersection donne la localisation estimée du terminal mobile, comme indiqué dans la figure 7. Dans la pratique, il est presque impossible d'obtenir un seul point d'intersection en raison des erreurs de mesure. La mesure de RSSI peut être affectée par des obstacles et des modèles imparfaits de propagation utilisés.

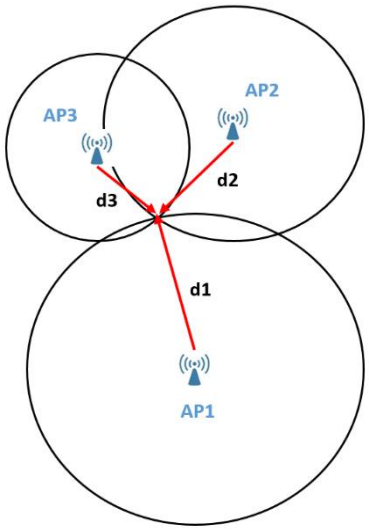


Figure 7 Présentation de la technique de triangulation

II.D.2.c) L'angle d'arrivée (AoA)

L'angle d'arrivée (AoA) est une technique de localisation utilisant des antennes directionnelles pour déterminer la direction d'arrivée des signaux Wi-Fi émis par un appareil mobile. Les mesures d'angle obtenues sont utilisées pour trianguler la position de l'appareil mobile en utilisant plusieurs paires d'antennes directionnelles. Bien que l'AoA puisse offrir une précision de localisation élevée en intérieur, elle nécessite des équipements coûteux et une infrastructure complexe, et peut être affectée par des obstacles physiques.

II.E. Du RSSI au CSI

Au cours des dernières années, de nombreuses applications ont été développées pour exploiter les informations de puissance du signal dans le but de détecter et d'analyser les caractéristiques de l'environnement. Dans le domaine des capteurs sans fil, RSSI est largement utilisée pour le positionnement.

Comme nous avons déjà présenté le RSSI peut être utilisé pour estimer la distance de propagation du signal conformément au modèle de propagation des signaux sans fil. Le RSSI peut aussi être utilisé comme « signature » ou « empreinte digitale » pour exprimer les caractéristiques spécifiques du signal sans fil à un emplacement donné. Cependant, la

propagation par trajets multiples peut engendrer des variations importantes d'amplitude du RSSI. Par exemple, dans un environnement de laboratoire classique, des fluctuations de l'ordre de 5 dB du RSSI ont été observées en moins d'une minute sur un récepteur fixe [30]. Ces fluctuations peuvent entraîner une correspondance inexacte des signatures ou des estimations de position. En effet, le RSSI ne permet pas de distinguer les différents chemins empruntés par le signal.

Afin de surmonter ces limitations, il est possible de recourir à des techniques plus avancées telles que l'utilisation de CSI (*Channel State Information*). Le CSI permet une caractérisation plus précise de la propagation par trajets multiples en fournissant des informations sur l'amplitude et la phase de chaque sous-porteuse du signal. Cette approche permet une estimation plus précise de la qualité du canal, offrant ainsi une meilleure précision dans les applications de positionnement sans fil.

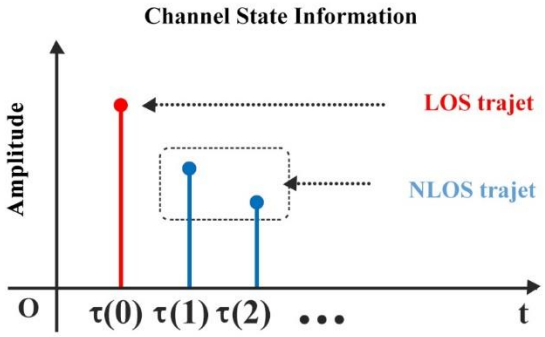


Figure 8 CSI par différents trajets

Afin de caractériser la propagation des trajets multiples, les canaux sans fil sont généralement modélisés en utilisant la réponse impulsionnelle du canal (CIR) également appelé CSI surtout dans les systèmes WiFi. Sous l'hypothèse d'un canal n'évoluant pas dans le temps, le CSI peut être exprimé par :

$$h(\tau) = \sum_{i=1}^{N} \alpha_i \, e^{-j\theta_i} \delta(\tau - \tau_i) \qquad (1)$$

Dans l'équation précédente, où α_i , θ_i et τ_i représentent respectivement l'atténuation d'amplitude, le déphasage et le retard du i-ème trajet, N est le nombre total de trajets multiples

et δ est la fonction delta de Dirac. Chaque terme de l'équation représente l'amplitude, la phase et le retard d'un chemin de propagation dans le domaine temporel.

II.F. Carte SDR et BladeRF

Software Defined Radio (SDR) est un système de communication radio dans lequel les fonctions matérielles sont implémentées à l'aide d'un logiciel. Les systèmes de communication radio traditionnels étaient basés sur un matériel à fonction fixe qui ne pouvait pas être reprogrammé ou modifié. La technologie SDR permet au système radio d'être configuré et contrôlé via un logiciel, ce qui permet une communication radio plus flexible et efficace. La technologie SDR fonctionne en utilisant une plateforme matérielle universelle, qui peut être programmée pour effectuer différentes fonctions. Les fonctions de traitement du signal du système radio, telles que la modulation, la démodulation et la détection, sont effectuées à l'aide d'un logiciel. Le matériel radio peut être modifié en fonction des exigences de l'application, en utilisant différentes configurations logicielles.

La carte BladeRF est une carte de développement SDR offrant une large bande passante, une puissance de traitement élevée et une grande flexibilité. Elle est compatible avec plusieurs logiciels SDR open source et propriétaires tels que GNU Radio et MATLAB. Avec son FPGA Altera Cyclone IV programmable et son processeur ARM Cortex A9, elle permet la conception de systèmes de communication personnalisés. La carte BladeRF est également capable de générer des signaux RF pour les tests et mesures. La raison de notre choix de la carte BladeRF réside dans le fait que son code VHDL est entièrement open source, nous permettant ainsi de le personnaliser et de le comprendre. C'est pour cette raison que nous avons opté pour cette carte afin de continuer nos recherches.

III. L’état de l’art

Dans cette section, nous explorerons les connaissances et les avancées qui sont au cœur de notre recherche. En examinant attentivement les études précédentes, nous visons à situer notre travail dans le contexte académique et à identifier les problèmes non résolus qui justifient notre contribution.

III.A. Détection des PRs

Étant donné que la détection des PRs est notre objet principal de recherche pour les appareils mobiles, nous avons d’abord examiné quelques études explorant les différentes manières dont les téléphones mobiles génèrent ces PRs, tout en observant les différentes caractéristiques des différents appareils.

Conditions d'émission :

Les auteurs de [31] ont été les premiers à remarquer que même si les appareils sont connectés à un point d'accès (AP), ils continuent d'envoyer en permanence des PRs. Freudiger [32] a énuméré plusieurs facteurs influençant le comportement de détection : l'état de l'écran, l'état de charge, le mode avion, si le Bluetooth est activé et la proximité des réseaux connus. Il a effectué des tests approfondis sur plusieurs systèmes d'exploitation populaires en utilisant plusieurs antennes et a calculé le nombre moyen de PR envoyées par chaque appareil. De nombreux appareils déclenchent une série de PR lorsqu'ils activent leur écran. Jamil et al. [33] ont utilisé cette caractéristique pour estimer l'état de l'écran du téléphone en fonction de ses PR.

Fréquence d'émission :

Lim et al. [34] ont constaté que la fréquence de détection dépend du contexte. Ils ont remarqué que certains appareils (utilisant des versions Android antérieures à 2.3.7) n'envoyaient pratiquement aucune PR. Ils ont également calculé le nombre moyen de PR envoyées par appareil sur différentes fenêtres temporelles. Par exemple, sur une plage de 3 minutes, 85% des appareils avaient 80% de probabilité d'envoyer un PR. Les appareils alimentés par batterie peuvent avoir une fréquence de détection plus faible que lorsqu'ils sont connectés à un chargeur

[35]. Abedi et al. [36] ont comparé les taux de détection du Wi-Fi et du Bluetooth et ont conclu que théoriquement, les appareils Wi-Fi diffusent en moyenne 10 fois plus d'adresses MAC que les appareils Bluetooth. De nombreux appareils émettent des PRs à une fréquence élevée, allant de quelques secondes à quelques minutes [32]. En tant que source d'information en constante fuite, ces PRs peuvent être utilisées pour déduire davantage d'informations sur ces appareils et leurs propriétaires. Par exemple, Jamil et al. [33] ont détecté les comportements des utilisateurs concernant l'utilisation de leurs appareils en fonction de la fréquence d'envoi des PRs. Redondi et al. [37] ont proposé une méthode pour déterminer si un appareil émettant des PRs est un smartphone ou un ordinateur portable.

III.B. Localisation en utilisant de PRs

Dans le cadre d'un suivi physique basé sur la radio, des sniffeurs sont utilisés pour intercepter et collecter les PRs émis par les appareils sans fil. Ces PR jouent un rôle essentiel dans la détection de la présence des utilisateurs et dans l'estimation de leur mobilité. Avant la mise en place du GDPR, les réglementations concernant la protection des données personnelles, les adresses MAC étaient transmises en clair. Ainsi, dans ce contexte, le suivi physique utilisait les adresses MAC pour identifier les appareils et suivre leurs déplacements. Le Wi-Fi, étant largement intégré aux appareils portables et facilement détectables, était la technologie radio privilégiée dans le domaine du suivi physique. Il est important de souligner que cette section se réfèrent spécifiquement à la période antérieure à la mise en place du GDPR.

Pour mettre en place un système de suivi passif, il suffit de déployer des capteurs qui interceptent les trames émises par tous les appareils activés à proximité. Ces capteurs peuvent être économiques, car il est possible de convertir de nombreux adaptateurs Wi-Fi en mode moniteur, leur permettant ainsi de collecter les trames [36]. Étant donné que la plupart des appareils Wi-Fi effectuent régulièrement des scans actifs, un système de suivi passif peut détecter la plupart de ces dispositifs.

En utilisant cette approche de suivi physique basé sur le Wi-Fi, il est possible d'obtenir des informations détaillées sur les déplacements des utilisateurs et leur présence dans différentes zones. Ces données peuvent être utilisées dans divers domaines, tels que la gestion des foules,

l'analyse du comportement des utilisateurs et la personnalisation des services basée sur la localisation. Grâce à la disponibilité généralisée du Wi-Fi et à la facilité de déploiement des capteurs passifs, cette méthode offre une solution pratique et économique pour le suivi physique.

Au cours des dernières années, de nombreuses technologies ont été proposées pour surveiller la position des individus à travers leurs appareils en observant le trafic WiFi. Certaines de ces technologies utilisent la reconnaissance de fingerprinting pour une meilleure précision en environnement intérieur [38].

En utilisant la reconnaissance de fingerprinting, Potorti et al. [39] ont obtenu des résultats significatifs en matière de localisation des individus dans des environnements intérieurs tels que des magasins dans des centres commerciaux ou des espaces de bureau ouverts, en exploitant les réseaux WiFi existants et l'analyse du trafic WiFi, qui ne nécessite pas de phase d'enquête et s'adapte automatiquement aux changements de l'environnement.

Les solutions ont été développées pour les applications de surveillance des piétons en utilisant les PRs, comme le travail réalisé par Xu et al. [40]. Ce système utilise la détection sans fil pour obtenir les adresses MAC des smartphones et utilise des méthodes de localisation basées sur RSSI. L'objectif de ce système est de surveiller le trafic des piétons dans les rues et de suivre les smartphones pour estimer la densité des personnes, et d'explorer comment utiliser ces informations pour améliorer les services offerts aux personnes, tels que des horaires de bus plus précis. Des approches similaires ont également été proposées dans [41], en se concentrant sur l'étude de nombreux facteurs qui affectent les performances sans fil dans de tels environnements, tels que le fading lent et le fading rapide.

Dans [42], Traunmueller et al. ont démontré que les données des PRs peuvent être utilisées pour analyser la mobilité externe et les trajectoires des individus dans des zones urbaines densément peuplées, avec une haute résolution spatiale. Ils ont collecté un ensemble de données de PR à partir de 54 points d'accès publics dans le bas de Manhattan à New York, à l'aide de la plateforme de test « Quantified Community ». Ils ont démontré comment utiliser ces données pour analyser les trajectoires courantes et indiquer l'intensité d'activité dans les rues qui varie dans le temps.

La localisation précise des appareils à l'aide du Wi-Fi dans un espace restreint est une tâche difficile. L'estimation de la position en utilisant RSSI est peu fiable car la corrélation entre le RSSI et la distance est peu précise [43] [44] [45]. Elle est influencée par des facteurs tels que le chemin de propagation ou si l'appareil est obstrué (par exemple, dans une poche) [46]. Weppner et al. [46] ont proposé une solution utilisant plusieurs capteurs et un filtrage du bruit pour atteindre une précision d'environ 10 mètres. Une solution plus fiable consiste à utiliser des informations plus précises de Channel State Information (CSI). Un outil développé par Halperin et al. [47] permet d'accéder au CSI à l'aide de cartes réseau existantes. Le CSI est une mesure plus précise que le RSSI car il contient des informations sur le canal pour chaque sous-porteuse plutôt qu'une mesure globale. Cette mesure peut être utilisée pour l'identification de « fingerprinting » individuelles [48].

III.C. Localisation en présence d'adresses MAC randomisé

Dans le domaine du suivi des comportements de foule basée sur l'analyse du trafic WiFi, de nombreuses solutions ont été proposées. Les premières solutions étaient basées sur le suivi des adresses MAC, en supposant que chaque appareil WiFi avait une adresse MAC unique [49]. Avec l'importance croissante de la protection des données dans la collecte et le traitement des informations liées aux utilisateurs, et avec le respect des droits de protection des données en vertu du GDPR, la protection de la vie privée est devenue essentielle. Ces dernières années, la randomisation des adresses MAC a suscité une attention considérable de la part des fournisseurs, qui ont développé et mis en œuvre diverses solutions [12].

Pour différencier les différents appareils WiFi tout en préservant la vie privée des utilisateurs, différentes méthodes et techniques ont été adoptées. Initialement, la distinction des appareils uniques reposait sur l'analyse séquentielle des trames de PR. Dans [50], les auteurs ont mesuré les intervalles de temps entre les PR reçues sur différentes canaux. Étant donné que les délais temporels ne sont pas spécifiés dans la norme et dépendent de la configuration de l'appareil, ils constituent une caractéristique appropriée pour l'identification. La dérandomisation des adresses MAC peut être réalisée en utilisant une attaque temporelle, où les trames provenant du même appareil sont regroupées en utilisant les intervalles de temps entre les trames PR, même si elles utilisent des adresses MAC différentes, comme proposé dans [51].

Cependant, en raison des phénomènes de dispersion et de trajets multiples introduisant des retards aléatoires dans les environnements réels, la temporalité en tant que caractéristique distinctive devient peu fiable [52]. Par conséquent, des solutions plus fiables pour l'identification par fingerprinting basée sur les informations élémentaires (IE) dans les trames PR ont été largement discutées dans la littérature. Dans [53], une étude sur les IE est présentée, identifiant de nouveaux domaines et techniques pour le suivi des utilisateurs. Les expérimentations ont montré que les générateurs de brouillage des émetteurs WiFi commerciaux étaient prévisibles et pouvaient être utilisés pour l'identification des appareils. De plus, deux méthodes d'attaque révélant les véritables adresses MAC des appareils sont également présentées.

En dehors des attaques temporelles, de nombreux articles proposent d'utiliser des dispositifs externes pour obtenir plus d'informations et contourner la randomisation. dans [54], une solution pour compter les participants à des manifestations publiques est proposée, basée sur les trames PR émises par les téléphones portables. Les auteurs ont étudié le comportement des signaux de base en utilisant des filtres de distance basés sur le RSSI et des filtres temporels. Dans [55], une solution pour comprendre le comportement des visiteurs est proposée, exploitant plus de 1,7 million de trames PR, des probabilités de transition historiques et un algorithme d’inférence de trajectoire basé sur un Hidden Markov Model (HMM). Dans [56], une méthode est proposée pour estimer la présence de dispositifs mobiles à des emplacements spécifiques. Cette méthode est basée sur un automate d'état pour détecter les arrivées, les présences et les départs des dispositifs à proximité des capteurs.

Dans [57], une solution d’estimation de la population est développée en utilisant la variation du nombre de trames PR dans des intervalles de temps définis, et en établissant une relation avec le nombre de dispositifs, afin de résoudre le problème de la randomisation des adresses MAC. Les auteurs de [11] ont développé un système efficace de surveillance des foules basé sur les trames PR. L'algorithme utilise des estimateurs statistiques pour compter les dispositifs à partir du taux de mesures des transmissions PR en rafale (le nombre de trames PR reçues en 10 ms).

Certains chercheurs utilisent des modèles d'apprentissage automatique (ML) pour estimer le nombre de personnes [58] [59] et ont obtenu de bons résultats. Cependant, dans les villes à forte

densité de population, il est difficile d'obtenir des informations sur la position de chaque individu ou le nombre total de personnes, ce qui constitue les « étiquettes » dans l'apprentissage automatique. C'est pourquoi le clustering, une forme d'apprentissage non supervisé, suscite un grand intérêt. Certains articles utilisent la longueur des données [13] ou l'IE [60] comme caractéristiques, puis appliquent l'algorithme de clustering DBSCAN pour différencier les différents appareils et contourner la randomisation.

En résumé, la surveillance des comportements de foule basée sur l'analyse du trafic WiFi fait face à des défis tels que la randomisation des adresses MAC, la protection de la vie privée des utilisateurs et le comptage précis des individus. Différentes solutions ont été proposées, notamment l'analyse temporelle, la reconnaissance d'empreintes basée sur les IE, la mesure de puissance et le comptage des piétons à partir des trames PR, etc. Chaque méthode présente ses avantages et ses inconvénients, et le choix de la solution appropriée dépendra des besoins spécifiques. Dans ma thèse, étant donné que notre projet est en collaboration avec une entreprise, nous préférons développer une technologie qui ne nécessite pas l'ajout de dispositifs externes supplémentaires, qui soit facile à déployer et simple d'utilisation. Cependant, les chercheurs ont néanmoins fourni de bonnes idées et perspectives pour notre travail.

IV. Base de données

Dans ce chapitre, nous présenterons en détail les données utilisées dans le cadre de notre thèse, ainsi que les méthodes que nous avons mises en œuvre pour enrichir et analyser ces données. Cette section joue un rôle crucial dans notre étude, car elle fournit les fondements nécessaires pour notre analyse approfondie.

IV.A. Présentation de la base de données

La base de données fournie par l'entreprise a été constituée sur les deux périodes : de juin à aout 2020 et de novembre 2020 à octobre 2021, permettant ainsi une collecte des données. Cette collecte s'est déroulée sur un vaste réseau WiFi qui compte plus de 200 points d'accès de la marque Ruckus, installés dans diverses villes afin de garantir une couverture géographique.

Pour recueillir les informations sur les activités des clients, les outils de gestion sophistiqués mis en place exploitent un serveur MQTT (Message Queuing Telemetry Transport), qui agit comme un véritable centre pour la réception et l'émission des données. Grâce à ce protocole de communication efficace, les mesures sont sollicitées auprès des clients connectés au réseau, offrant ainsi une vision complète des interactions et des comportements observés.

Les mesures collectées sont ensuite organisées de manière méthodique et ordonnée dans des fichiers spécifiques, où chaque enregistrement est accompagné d'un horodatage précis. Cette approche chronologique permet d'établir une chronologie des événements qui se sont produits sur le réseau WiFi au fil du temps.

Grâce à l'utilisation des points d'accès Ruckus et du serveur MQTT, l'entreprise a pu accumuler une base de données riche en informations sur les activités des clients sur son réseau WiFi. La collecte des données sur une durée prolongée et leur organisation chronologique offrent une opportunité d'analyse et de compréhension des comportements des utilisateurs.

Les données recueillies sont organisées en deux fichiers, l'un pour le mode « présence » et le second pour le mode « connecté ».

IV.A.1. Mode « presence »

Également connu sous le nom de mode monitor, c'est une configuration spécifique des points d'accès qui leur permet de capturer individuellement les trames des clients sur l'ensemble des canaux WiFi. Dans ce mode, les points d'accès sont principalement utilisés pour récupérer des informations provenant des PRs émis régulièrement par les appareils lorsque ces clients ne sont pas encore connectés au réseau.

Ces données récupérées dans ce mode sont généralement limitées aux niveaux de RSSI des appareils clients et à quelques autres éléments de base. Cela signifie que les informations collectées sont relativement simples et ne fournissent qu'une indication de la présence des appareils à proximité du réseau.

En raison des limitations inhérentes à ce mode, le nombre de points d'accès configurés en mode « présence » est souvent restreint. En effet, ce mode ne permet pas de gérer les connexions au réseau ni de fournir des fonctionnalités avancées telles que l'authentification ou l'attribution des adresses IP.

Cependant, malgré ses limitations, le mode « présence » reste un outil précieux pour obtenir des informations sur la présence des appareils et pour analyser la densité du trafic WiFi dans une zone donnée. Il est souvent utilisé dans des scénarios où une surveillance passive et non intrusive est nécessaire.

IV.A.2. Mode « connecté »

Le mode « managed », également connu sous le nom de mode connecté, est le mode de fonctionnement le plus courant pour la plupart des points d'accès d'un réseau. Contrairement au mode « présence » ou « monitor », le mode « managed » permet de gérer le trafic des clients connectés de manière active et dynamique.

Lorsqu'un client se connecte à un point d'accès en mode « managed », des informations détaillées sur la connexion sont enregistrées. Ces données de connexion comprennent généralement le profil du client, tel que son adresse MAC, son adresse IP, et d'autres

informations d'identification, ainsi que des informations sur la session elle-même, telles que la durée de la session, le débit de la connexion, la quantité de données échangées, etc.

Grâce à ces données, il est possible de suivre et d'analyser le comportement des clients connectés au réseau WiFi. Cela permet, par exemple, de générer des statistiques sur l'utilisation du réseau, d'identifier les clients les plus actifs ou les plus gourmands en bande passante, et de détecter d'éventuels problèmes de performance ou de sécurité.

IV.B. Protection de données

Au sein de cette infrastructure réseau, la protection de la confidentialité des données personnelles des utilisateurs est une priorité absolue. Dans cet objectif, des mesures rigoureuses sont mises en place pour garantir la sécurité des informations relatives aux clients.

Tout d'abord, les données des clients sont systématiquement cryptées pour empêcher toute interception ou utilisation non autorisée. Cela garantit que seules les personnes habilitées peuvent accéder aux informations confidentielles.

De plus, une attention particulière est accordée à la protection des adresses MAC des clients. Les adresses MAC, qui sont des identifiants uniques associés à chaque appareil, sont rigoureusement protégées pour éviter toute traçabilité intrusive. il n'est pas conservé de trace de l'adresse MAC du client lors de la collecte des données. À la place, celle-ci est substituée par une chaîne de hachage [10] anonyme qui transforme les adresses en valeurs uniques et impossibles à inverser. Il convient également de mentionner les efforts des principaux fabricants de téléphones mobiles pour préserver la confidentialité des utilisateurs. Ils ont développé une technique novatrice consistant à générer des adresses MAC réelles aléatoires sur les appareils mobiles. Cette approche permet aux téléphones de générer et de modifier continuellement des adresses MAC fictives de manière aléatoire lorsqu'ils ne sont pas connectés à un réseau WiFi. Ainsi, toute tentative de suivi ou de traçage basée sur l'adresse MAC est rendue inefficace, préservant ainsi activement la vie privée des utilisateurs.

Cependant, l'application de ces mesures de protection de la vie privée présente des défis importants en matière d'anonymisation et de randomisation des données. Ces techniques

compliquent considérablement la tâche de suivi des localisations et de l'estimation du nombre de clients. Néanmoins, ces défis sont acceptés et relevés avec diligence pour garantir la confidentialité des utilisateurs et respecter les normes de protection des données en vigueur.

IV.C. L'enrichissement des données

En comptant le nombre de fois où une adresse MAC anonymisée distincte est observée au cours d'une période d'une journée et en la comparant à un seuil, tout en tenant compte du fichier des détails de connexion, nous pouvons classer les PR d'une journée dans trois catégories mutuellement exclusives :

Client Randomisé (CR) : L'adresse MAC est observée de 1 à « seuil » fois.

Client Fixe (CF) : L'adresse MAC est observée plus de « seuil » fois.

Client Connecté (CC) : L'adresse MAC est présente dans le fichier des détails de connexion.

Notre projet de recherche vise à analyser les enregistrements PR obtenus à partir de plusieurs réseaux WiFi publics en extérieur en France. Les ensembles de données fournis à notre groupe ne sont pas accessibles au public et ont été anonymisés en ce qui concerne les noms des sites.

Conformément au GDPR, les adresses MAC des clients n'ont pas été conservées lors de l'acquisition des données. Au lieu de cela, les fichiers fournis contiennent des chaînes de hachage anonymisées en remplacement des adresses MAC d'origine. Il est important de noter qu'une adresse MAC fixe produira toujours la même chaîne de hachage, tandis que pour les adresses MAC randomisés, le code de hachage sera différent pour chaque PR émis par un client.

Malgré la randomisation de l'adresse MAC, ses trois premiers octets, le OUI déjà évoqué, sont conservés et font partie de l'entrée PR. L'OUI est censé fournir des informations sur le fabricant de la carte réseau WiFi qui génère le PR, disponibles en interrogeant un référentiel web public des OUI. Cependant, l'utilisation de l'OUI est non réglementée et volontaire, et dans la pratique, la demande d'OUI renverra souvent NaN, c'est-à-dire non défini. Dans les ensembles de

données étudiés, le pourcentage de réponses d'OUI NaN s'est avéré dépendre fortement de la classe de client définie, comme illustré dans la Figure 9.

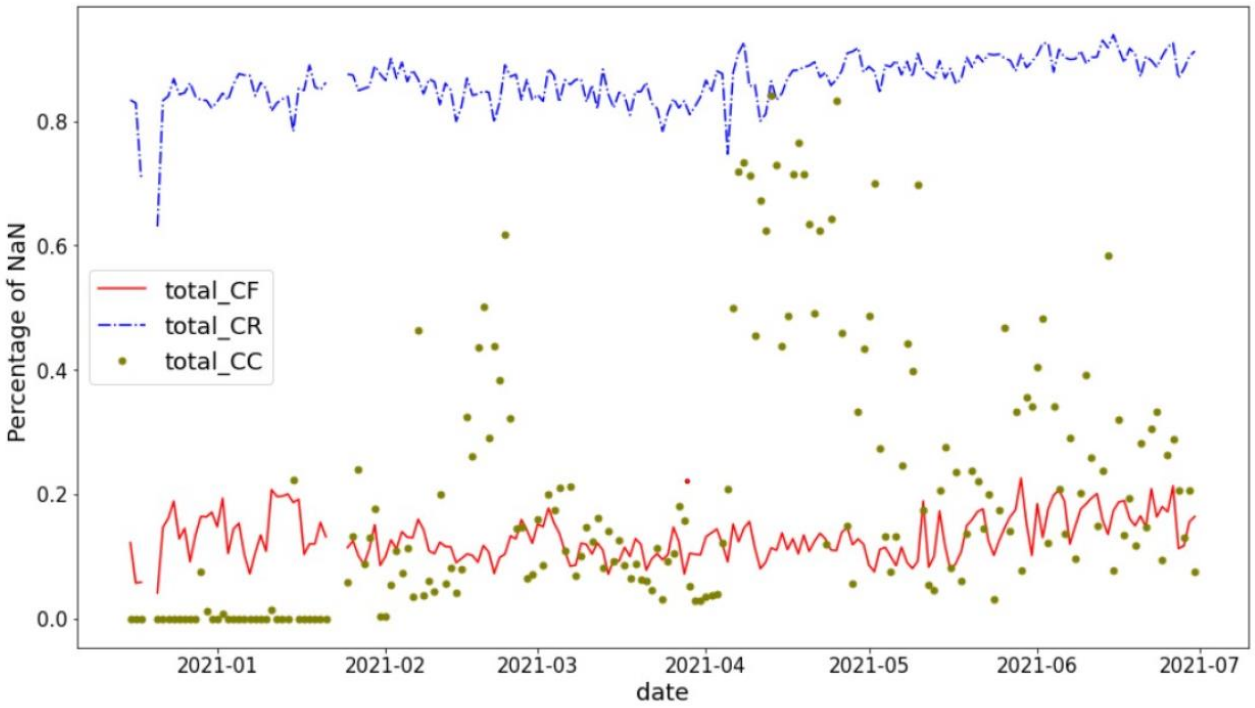


Figure 9 Pourcentage de NaN de OUIs pour différents classes de clients

La figure 9 montre les pourcentages de réponses NaN d'OUI pour les classes de clients CR, CF et CC pour les PR de la ville 1 sur une période de 6 mois en 2021. On peut observer que les PR des clients randomisés, c'est-à-dire provenant de systèmes d'exploitation conformes au GDPR, tendent également à masquer les informations sur le fabricant de l'appareil. Les OUI des clients fixes sont le plus souvent disponibles, comme prévu. Cependant, cette observation doit être nuancée. La Figure 10 montre que les réponses d'OUI de la classe CF incluent un éventail déconcertant de fabricants, parfois difficiles à identifier clairement. Une enquête détaillée suggère que jusqu'à 80 % sont des fabricants d'appareils IoT tels que l'éclairage, le chauffage, les caméras, etc., qui ont un comportement réseau assez distinct des véritables appareils clients. Les clients connectés, dans la figure 9, présentent un comportement OUI mixte. De plus, la classe CC est susceptible de contenir des clients ES, comme nous l'avons déjà signalé. Ces observations fournissent une forte motivation pour baser exclusivement les solutions d'estimation de foule conformes au GDPR sur des clients MAC randomisés.

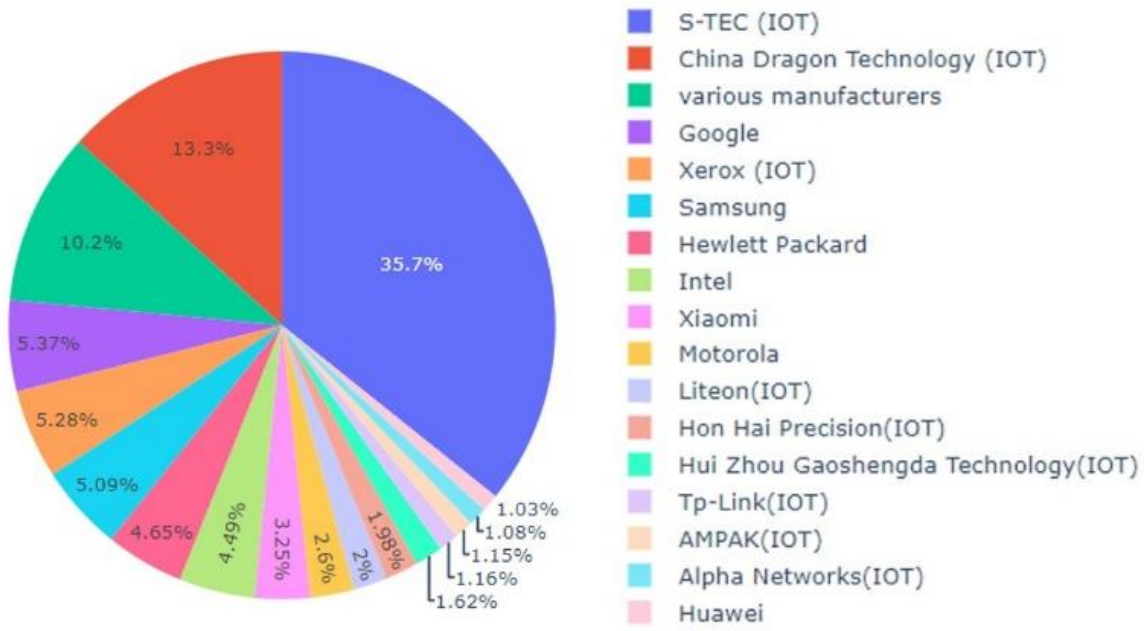


Figure 10 Pourcentage des fabricants de CF OUI différents lors d'une journée typique

IV.D. Difficultés rencontrées

IV.D.1. Impact de la randomisation des adresses MAC sur la détection

D'après la figure 11, on peut constater que la majorité des clients (environ 67 %) dans le fichier de « présence » ne sont détectés qu'une seule fois. Cette situation est principalement due à la randomisation des adresses MAC des appareils, ce qui pose un défi majeur en termes de localisation. En effet, étant donné que chaque fois qu'un client émet, il utilise une adresse MAC différente, il devient difficile d'associer les différentes observations d'un même client au fil du temps. Par conséquent, il est complexe de construire un historique de localisation fiable pour ces clients.

D'autre part, bien que le nombre de clients détectés plus de 100 fois soit relativement faible, ils représentent néanmoins une part significative, soit environ 30 % à 40 %, du nombre total de PR

enregistrés. Ces clients spécifiques n'effectuent pas systématiquement une randomisation de leurs adresses MAC et conservent une adresse identique lors de leurs connexions successives.

Date du fichier

	2020-06-30	2020-07-01	2020-07-02	2020-07-03	2020-07-04	2020-07-05	2020-07-06	2020-07-07	2020-07-08	2020-07-09	2020-07-10
nb_clients_differents	428440	248093	442831	468349	360080	292188	445476	477317	509855	482652	441728
clients_vus_1_fois	286756	167376	298486	320481	247924	196103	313198	323063	354834	352826	310544
%_de_1_fois	66.93%	67.47%	67.40%	68.43%	68.85%	67.12%	70.31%	67.68%	69.60%	73.10%	70.30%
clients_vus_(2_10)_fois	136443	77404	139036	142170	107843	92681	127080	148648	149310	124122	125941
%_de_(2_10)_fois	31.85%	31.20%	31.40%	30.36%	29.95%	31.72%	28.53%	31.14%	29.28%	25.72%	28.51%
clients_vus_(11_50)_fois	3910	2455	3887	4188	3130	2243	3664	3990	4088	4148	3790
%_de_(11_50)_fois	0.91%	0.99%	0.88%	0.89%	0.87%	0.77%	0.82%	0.84%	0.80%	0.86%	0.86%
clients_vus_(51_100)_fois	651	370	676	709	570	479	703	766	742	744	719
%_de_(51_100)_fois	0.15%	0.15%	0.15%	0.15%	0.16%	0.16%	0.16%	0.16%	0.15%	0.15%	0.16%
clients_vus_>100_fois	680	488	746	801	613	682	831	850	881	812	734
%_de_>100_fois	0.16%	0.20%	0.17%	0.17%	0.17%	0.23%	0.19%	0.18%	0.17%	0.17%	0.17%

Figure 11 Statistiques client dans les différents fichiers « presence »

IV.D.2. Analyse des clients vus par au moins 3 AP

Selon nos observations comme la figure 12 dans l'ensemble du système, la grande majorité des clients (environ 98,8 %) ne sont détectés que par un seul point d'accès. Cela est dû à la distance importante entre les différents points d'accès, ce qui limite les clients à se connecter uniquement au point d'accès le plus proche. À partir de nos données, il est clair que seule une infime fraction des clients (environ 0,2 %) présente un potentiel de triangulation. Cependant, même parmi cette petite fraction de clients potentiellement localisables (présenté la partie basse de la figure 12), seuls environ 6 % d'entre eux disposent d'un nombre suffisant de PRs (>100) pour une localisation précise, c'est à dire, pour une journée entière de données, seulement 0,2 % * 6 % des données nous permettent de suivre les clients et de les localiser. Cela signifie que la grande majorité des clients ne peuvent pas être localisés efficacement et que même parmi ceux qui peuvent être localisés, seuls quelques-uns obtiennent des résultats fiables. En conclusion, sans résoudre le problème de la randomisation des adresses MAC, il est pratiquement impossible de

suivre et de localiser les clients de manière efficace. La randomisation des adresses MAC limite notre capacité à effectuer un suivi persistant des clients et à analyser leur position, rendant la localisation précise difficile.

Date du fichier

	2020-06-30	2020-07-01	2020-07-02	2020-07-03	2020-07-04	2020-07-05	2020-07-06	2020-07-07	2020-07-08	2020-07-09	2020-07-10
nb_clients_differents	428440	248093	442831	468349	360080	292188	445476	477317	509855	482652	441728
clients_vus_>3_AP	726	513	781	918	772	723	808	825	773	333	324
%_>3_AP	0.17%	0.21%	0.18%	0.20%	0.21%	0.25%	0.18%	0.17%	0.15%	0.07%	0.07%

	2020-06-30	2020-07-01	2020-07-02	2020-07-03	2020-07-04	2020-07-05	2020-07-06	2020-07-07	2020-07-08	2020-07-09	2020-07-10
clients_vus_>3_AP	726	513	781	918	772	723	808	825	773	333	324
>3_AP_(ligne<10)	576	393	594	680	589	541	602	584	593	281	267
%_(ligne<10)	79.34%	76.61%	76.06%	74.07%	76.30%	74.83%	74.50%	70.79%	76.71%	84.38%	82.41%
>3_AP_(10<ligne<50)	113	87	135	175	135	128	154	169	133	39	43
%_(10<ligne<50)	15.56%	16.96%	17.29%	19.06%	17.49%	17.70%	19.06%	20.48%	17.21%	11.71%	13.27%
>3_AP_(ligne>100)	37	33	52	63	48	54	52	72	47	13	14
%_(ligne>100)	5.10%	6.43%	6.66%	6.86%	6.22%	7.47%	6.44%	8.73%	6.08%	3.90%	4.32%

Figure 12 Statistiques des clients vus par au moins 3 AP dans les fichiers « présence »

IV.D.3. Corrélation de Pearson pour différents sites

Le coefficient de corrélation de Pearson (ρ) est un indicateur utilisé pour évaluer la configuration spatiale des AP en termes de coordonnées x-y. Lorsque le coefficient de corrélation est proche de zéro, cela indique une bonne triangulation, c'est-à-dire une disposition spatiale favorable des AP. À l'inverse, un coefficient de corrélation élevé suggère une mauvaise triangulation (fig 13). Cependant, malheureusement, la géométrie des points d'accès détectant un même client est souvent défavorable, en particulier lorsque les AP sont alignés. Cette configuration peut entraîner des difficultés dans la précision de la localisation, car la distinction entre les AP devient plus difficile lorsque leur disposition est linéaire.

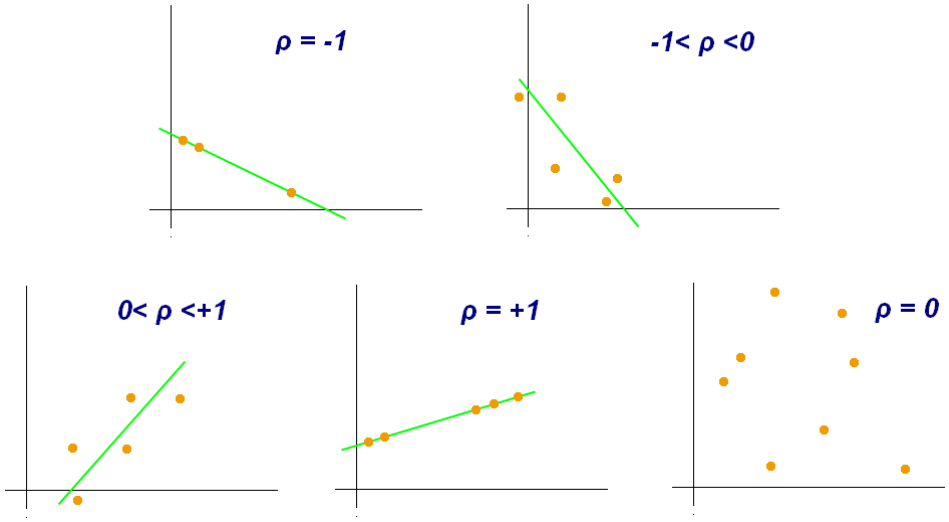


Figure 13 Exemples de coefficient ρ de corrélation de Pearson pour des points x-y (Wikipedia)

Dans notre ensemble de données, nous avons calculé les coefficients de corrélation de Pearson entre différentes villes et les avons affichés dans le tableau 1. Comme expliqué précédemment, des coefficients de corrélation de Pearson plus élevés indiquent une moins bonne possibilité de triangulation précise. Les résultats que nous avons obtenus varient approximativement entre 0,4 et 0,9, ce qui signifie que la plupart villes de nos données ne remplissent pas les conditions d'une bonne triangulation. Cela représente sans aucun doute un défi majeur pour le développement de techniques de localisation que nous envisageons.

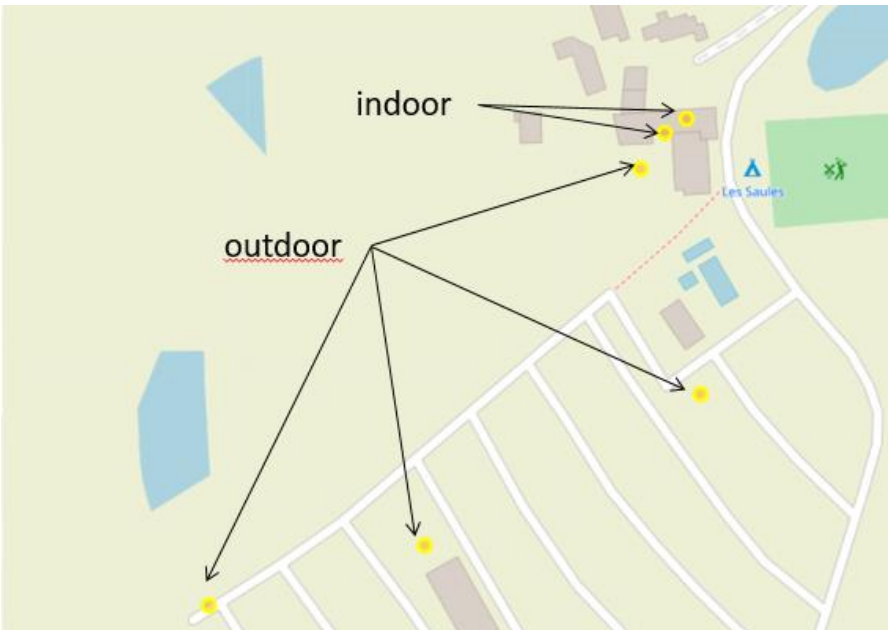
Cependant, en ce qui concerne nos 2 sites touristiques, appelés Sites 1 et 2, que nous rencontrerons dans le volet localisation de la thèse, leur coefficient de corrélation de Pearson est autour de 0,4 ; aussi nous verrons qu'il a été possible de mettre au point une méthode de de localisation pour ces lieux.

Tableau 1 Calcul de Coefficient de Pearson des x-y d'AP pour les différentes villes

Lieu	Coefficient de Pearson des x-y d'AP
Ville A	0,88
Ville B	1
Ville C	1
Site 1	0,44
Site 2	0,40

IV.D.3. Impact de la présence simultanée d’AP à l'intérieur et à l’extérieur.

Lorsqu'il s’agit de localiser des AP dans un réseau, il est crucial de prendre en compte leur emplacement, qu'ils soient installés en extérieur ou en intérieur. Dans notre ensemble de données, en particulier, certains AP se trouvent en intérieur tandis que d’autres sont en extérieur présenté dans la figure 14, ce qui pose un défi important en termes de localisation.



*

Figure 14 Exemple de camping illustre la présence simultanée d’AP à l’intérieur et à l’extérieur.

V. Approche de comptage de client

V.A. introduction

Au fil des années, l'utilisation des adresses MAC fixes des clients présentes dans les PRs pour compter les clients est devenue une pratique courante dans l'industrie, à des fins commerciales et de sécurité. Cependant, l'introduction du GDPR en Europe en 2016, ainsi que des réglementations similaires dans d'autres pays, a classé les adresses MAC comme des informations personnelles. Cela a entraîné la randomisation des adresses MAC des clients par les systèmes d'exploitation des appareils mobiles, ce qui a rendu nécessaire la recherche de méthodes alternatives de suivi des audiences.

La randomisation des adresses MAC est devenue la principale mesure de mise en conformité avec le GDPR dans les réseaux WiFi. Bien que certaines études [12] [32] [51] aient exploré des techniques telles que la dé-randomisation [13] [18] des adresses MAC, il y a un consensus croissant pour considérer que la protection des données est essentielle et que les solutions doivent être résistantes à la randomisation [11] [15] [16] [17] [56].

Une approche consiste à se fier aux clients ayant des adresses MAC fixes, en supposant que leurs statistiques d'émission de PR seraient similaires à celles des clients ayant des adresses MAC randomisées [15] [16]. Cependant, cette hypothèse est remise en question en raison de l'augmentation du nombre de dispositifs de l'Internet des objets (IoT) et de clients utilisant des Enhanced Services (ES) tels que le Peer-to-Peer (P2P), dont les statistiques PR diffèrent considérablement des profils clients authentiques, rendant difficile l'estimation de l'audience.

Bien des approches proposées [16] [11] [17] [61] nécessitent une analyse détaillée des enregistrements PR dans le domaine temporel, ce qui demande beaucoup de calculs. Ces approches s'appuient souvent sur un matériel réseau spécialisé et nécessitent une collaboration étroite entre les équipes de développement matériel et logiciel. De plus, elles exigent généralement une technique de calibrage, telle que des systèmes de caméras ou des capteurs spéciaux, pour établir une référence fiable permettant de compter les clients, ce qui peut être coûteux à installer et à entretenir.

Dans cette étude, nous proposons une approche statistique simple qui repose exclusivement sur la découverte de statistiques sous-jacentes des clients. En analysant les compteurs PR quotidiens, les modèles d'activité des clients et les tendances saisonnières, nous déduisons un facteur d'échelle approximatif, X, qui permet de convertir directement les compteurs PR bruts en estimations de la population des clients. Cette méthode préserve la confidentialité des données car elle fonctionne exclusivement avec des clients ayant des adresses MAC randomisées et évite les problèmes liés aux clients IoT et ES. De plus, elle peut être appliquée à des ensembles de données réels provenant de réseaux WiFi commerciaux qui ne disposent pas d'un système de référence supplémentaire. À notre connaissance, notre méthode proposée est la première à permettre la surveillance de l'audience basée uniquement sur les statistiques PR des adresses MAC randomisées, sans nécessiter de technique de référence supplémentaire.

V.B. Matériels et méthodes

V.B.1. Ensembles de données et définitions

Dans le cadre de cette étude, nous avons examiné des bases de données provenant de trois villes et d'un camping. Ces données ont été collectées au cours des années 2020 et 2021. Pour faciliter notre analyse, nous avons divisé les données urbaines en deux périodes en fonction de la présence ou de l'absence de mesures de confinement liées à la COVID-19. Au total, nous avons créé 6 fichiers à partir de la base de données d'origine, comme décrit en détail dans le Tableau 2 En incorporant des villes de tailles et de populations différentes, nous pouvons évaluer l'efficacité de notre approche sur une large gamme de comptages PR, couvrant presque un ordre de grandeur. De plus, l'inclusion des données du camping, qui présente des caractéristiques logistiques différentes de celles d'une ville, nous permet de garantir l'applicabilité générale de la méthode proposée.

Tableau 2 Résumé des sites étudiés dans ce chapitre

Site	population	Densité	Confinement de Covid
Ville-A	3720	194/km^2	Non
Ville-B1	1870	148/km^2	Non
Ville-B2	1870	148/km^2	Oui
Ville-C1	136,250	3954/km^2	Non
Ville-C2	136,250	3954/km^2	Oui
Camping*	144 terrains de camping, 11 cabanes, 9 hectares.		Non

*juste pour validation

Dans ce chapitre, nous avons choisi d'agréger les PR en totaux quotidiens, bien que les horodatages de chaque enregistrement permettent une sélection flexible de l'intervalle de temps. En comptant le nombre de fois où une adresse MAC anonymisée distincte est observée au cours d'une période d'une journée et en la comparant à un seuil, tout en tenant compte du fichier des détails de connexion, nous pouvons classer les PR d'une journée dans trois catégories mutuellement exclusives :

- Client Randomisé (CR) : L'adresse MAC est observée de 1 à "seuil" fois.
- Client Fixe (CF) : L'adresse MAC est observée plus de "seuil" fois.

- Client Connecté (CC) : L'adresse MAC est présente dans le fichier des détails de connexion.

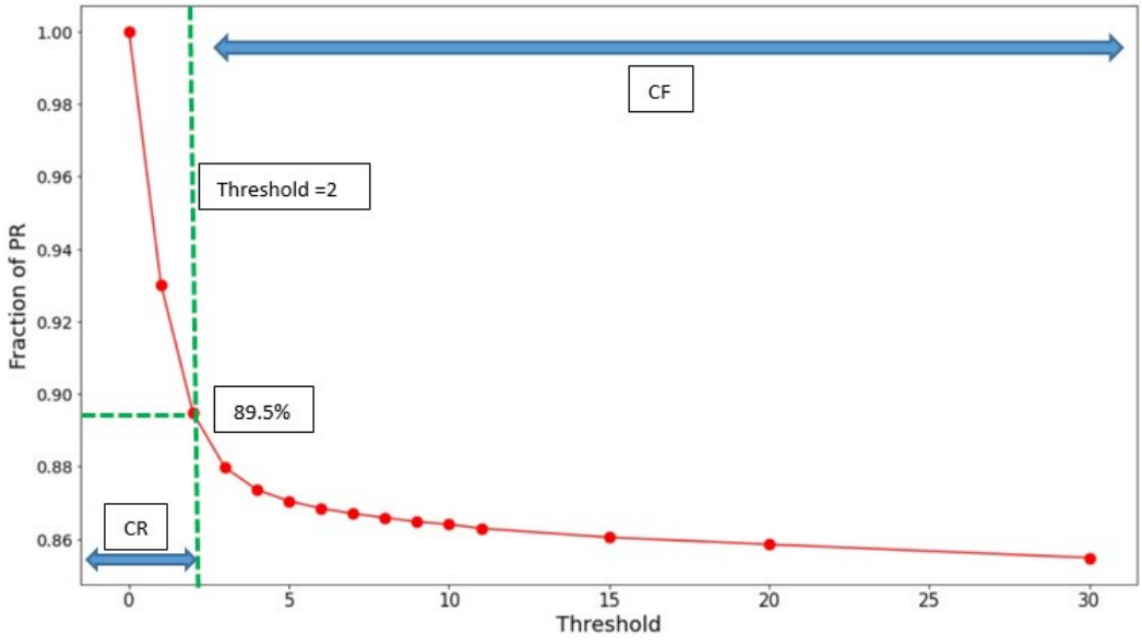


Figure 15 Comportement du nombre de PR conservés en fonction du seuil utilisé pour définir les classes CF et CR. Un seuil de 2 est choisi de sorte que les adresses MAC anonymisées vues 1 ou 2 fois soient considérées comme appartenant à la classe des clients randomisés, ou CR, tandis que celles apparaissant plus fréquemment sont considérées comme des adresses MAC fixes, ou CF.

Initialement, nous avons supposé que les adresses MAC des clients CR seraient totalement aléatoires, ce qui conduirait à un seuil de 1. Cependant, notre étude préliminaire a révélé une situation plus complexe. La Figure 15 illustre la fraction de PR totale conservée en fonction du seuil pour un ensemble de données typique. À partir de seuils de 30 jusqu'à environ 4, le pourcentage de PR conservées reste relativement constant, ce qui indique que la plupart des adresses MAC sont observées plusieurs fois. Cela correspond à la classe CF. Les seuils de 2 ou 3 se situent dans une zone grise où les PR conservées pourraient appartenir à des clients CR dont la randomisation des adresses MAC n'est pas parfaite, ou à des clients CF individuels contribuant peu au trafic. Fixer le seuil trop haut risque d'inclure des appareils IoT dans la classe CR, comme expliqué dans la section suivante. À l'inverse, fixer le seuil trop bas pourrait réduire inutilement l'efficacité de l'identification des clients CR. Il est probable que les deux scénarios

contribuent aux PR dans cette plage. Pour parvenir à un compromis, nous avons choisi un seuil de 2 pour distinguer les classes CR et CF. Augmenter le seuil a un impact minime sur les résultats numériques obtenus, mais choisir un seuil de 1, que nous considérons comme trop restrictif, réduit les facteurs de conversion mesurés (X) d'environ 25 %. Cela introduit une certaine incertitude, mais n'altère pas fondamentalement la signification des résultats qui seront présentés.

V.B.2. Description de la méthode

Dans la Figure 16, nous pouvons observer les décomptes quotidiens de PR (enregistrements de MAC aléatoires) dans les trois villes sur une période de plusieurs semaines. La figure révèle la présence de schémas hebdomadaires avec des pics distincts en semaine et des compteurs plus bas le week-end. De plus, des irrégularités correspondant aux jours fériés et à des périodes occasionnelles d'écart par rapport à la régularité peuvent être observées. On peut également remarquer des changements d'échelle notables pendant les périodes de confinement liées à la COVID-19, ainsi que de légères déviations à la hausse probablement liées aux saisons.

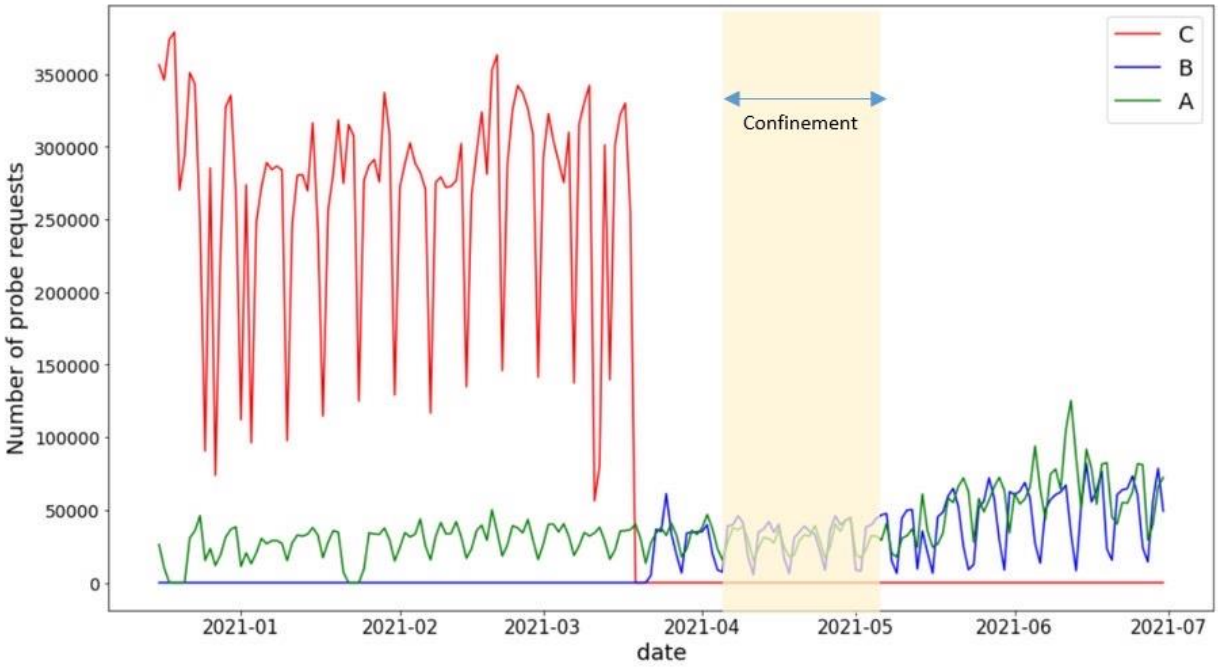


Figure 16 Comptage quotidien des PR CR.

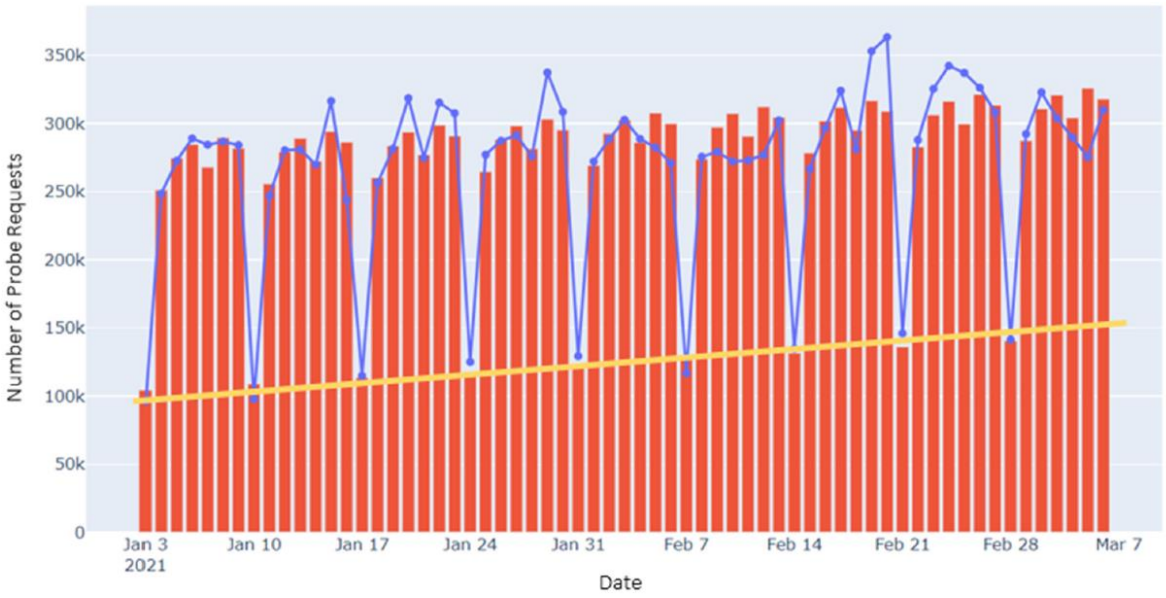


Figure 17 Exemple de graphique des données CR en fonction du temps, avec le modèle superposé, pour la ville-C2. La ligne bleue représente les données PR brutes, et les barres rouges représentent le modèle ajusté. La tendance linéaire du modèle est indiquée par une ligne jaune.

Pour mettre en œuvre la procédure, nous commençons par choisir un intervalle où la périodicité propre des données PR est évidente pour un site particulier, et nous considérons les données comme un ensemble d'expériences répétées sur plusieurs semaines. Un modèle des données est ensuite construit, sous la forme d'un modèle hebdomadaire répété composé du nombre moyen de PR pour chacun des 7 jours de la semaine, auquel nous ajoutons un terme de tendance linéaire pour tenir compte des variations saisonnières possibles. Les paramètres du modèle sont déterminés en minimisant, par rapport à ces paramètres, les écarts quadratiques sommés des données par rapport au modèle, c'est-à-dire en effectuant un ajustement des moindres carrés du modèle aux données. Un exemple du modèle (barres rouges) superposé aux comptages PR bruts (courbe bleue) pour la ville-C est montré dans la Figure 17, où le terme de tendance linéaire est également indiqué explicitement par une ligne jaune. L'algorithme proposé, décrit ci-dessous, est basé sur l'interprétation des écarts quadratiques des points de données par rapport au modèle, ainsi que sur certaines hypothèses supplémentaires.

Pour élaborer l'algorithme, nous considérons un site équipé de points d'accès WiFi capables de capturer les PR émis par les téléphones portables des visiteurs, également appelés clients. La

durée de l'expérience est notée T, généralement fixée à 1 jour dans cet article. Le nombre total de visiteurs pendant la période spécifiée est indiqué par A.

Pour chaque client, identifié comme client *b* avec *b* allant de 1 à *A*, leur séjour sur le site a une durée t_b, où t_b est inférieur à *T*. De plus, x_b représente le nombre moyen de PR émis par le téléphone du client *b* pendant la durée de *T*. Il est bien connu que les taux d'émission de PR des téléphones portables peuvent varier considérablement en fonction de facteurs tels que le système d'exploitation et l'état actuel de l'appareil, comme discuté dans des études antérieures [32].

Nous aurons alors, pour le nombre total de PR émis, *P* :

$$P = \sum_{b=1}^{A} \frac{x_b}{T} t_b = \frac{A}{T} \sum_{b=1}^{A} \frac{x_b t_b}{A} = \frac{A}{T} \langle x_b t_b \rangle, \qquad (V-1)$$

Le symbole < > représente la valeur moyenne de la quantité entre parenthèses. Dans ce contexte, nous supposons que la durée du séjour d'un client sur un site n'est pas influencée par le taux d'émission de PR de son téléphone. Sur la base de cette hypothèse, nous pouvons appliquer la loi de la moyenne du produit de deux distributions indépendantes pour établir la relation suivante :

$$P = \frac{A}{T} \langle x_b \rangle \langle t_b \rangle = \left(\frac{\langle x_b \rangle}{T} \right) (A \langle t_b \rangle) \qquad (V-2)$$

$$P \equiv XC \qquad (V-3)$$

Dans ce contexte, nous supposons que *T* correspond à une seule journée, et nous interprétons *X* comme le nombre moyen de PR émis par un client sur une journée complète. *C* représente le nombre effectif de journées-client observées lors de l'expérience. Par exemple, si *C* = 2, cela signifie qu'il y avait deux clients présents pendant toute la journée, ou quatre clients présents pendant une demi-journée chacun, et ainsi de suite. Un autre scénario pourrait être de prédire 500 clients pour une journée, dont 250 restent toute la journée tandis que les 250 autres sont progressivement remplacés par 250 clients différents. Dans ce cas, nous avons toujours *C* = 500, tandis que *A* serait de 750.

On peut se demander pourquoi nous choisissons de prédire *C* en termes de journées-client plutôt qu'*A*, qui représente le nombre réel de clients distincts. En raison de la randomisation des adresses MAC, nous ne pouvons pas identifier individuellement les clients et n'avons accès qu'à des données agrégées représentées par *C*. Le choix entre A et C dépend de l'application spécifique. Pour les applications liées à la sécurité, telles que la détermination de la capacité d'occupation d'un bâtiment ou d'un site, le facteur crucial est C, qui représente le nombre de personnes présentes dans une fenêtre temporelle spécifique, plutôt que le nombre total d'individus traversant le site pendant la période. En revanche, si l'objectif est de déterminer le nombre de billets vendus sur un site, indépendamment du temps que chaque visiteur y passe, *A* serait la mesure la plus appropriée. Estimer *A* à partir de *C* en moyenne est possible lorsque des données de référence locales, telles que les recettes des guichets, sont disponibles, mais ces méthodes ne relèvent pas du cadre de cet article.

À cette étape, nous avons établi une méthode pour estimer le nombre de journées-client à partir du nombre de PRs en utilisant un simple facteur multiplicatif X. La prochaine étape consiste à proposer une technique pour obtenir la valeur de X à partir des données. La technique proposée repose sur l'analyse des écarts quadratiques des données PR par rapport au modèle. Les écarts par rapport au nombre moyen de PR pour un jour donné peuvent être dus à des fluctuations aléatoires dans le nombre de clients (C) ainsi qu'à des fluctuations potentiellement dépendantes du temps dans la valeur de X. Pour quantifier cela, nous utilisons l'expression de la variance d'une distribution de produit, dans ce cas, P = XC, qui peut être dérivée comme suit :

$$P = XC \qquad (V-4)$$

$$\sigma_P^2 = \sigma_C^2 X^2 + \sigma_C^2 \sigma_X^2 + \sigma_X^2 C^2 \qquad (V-5)$$

$$\frac{\sigma_P^2}{P} = \frac{\sigma_C^2 X}{C} + \frac{\sigma_C^2 \sigma_X^2}{CX} + \frac{\sigma_X^2 C}{X} \qquad (V-6)$$

Dans cette équation, les termes σ^2 représentent les variances des distributions parent et produit, et *P*, *C* respectivement les moyennes des distributions P et C. Étant donné que *C* résulte d'une expérience de comptage sur un intervalle de temps fixe, nous nous attendons à ce qu'elle suive une distribution de Poisson. Dans une distribution de Poisson, la variance est égale à la moyenne. Nous pouvons donc réécrire l'équation de la manière suivante :

$$\frac{\sigma_P^2}{P} = X + \frac{\sigma_X^2}{X} + \frac{\sigma_X^2 C}{X} \qquad (V-7)$$

$$\frac{\sigma_P^2}{P} \approx X + \frac{\sigma_X^2 C}{X} \qquad (V-8)$$

$$\frac{\sigma_P^2}{P} \approx X + \frac{\sigma_X^2}{X^2} P \qquad (V-9)$$

En supposant que $C >> 1$ dans l'équation (V-8), nous avons utilisé une approximation pour simplifier l'expression. Ensuite, dans l'équation (V-9), nous avons remplacé C par P/X afin d'obtenir une équation ne dépendant que de X et P. Le côté gauche de l'équation (V-9) correspond au rapport de la variance à la moyenne de P, également connu sous le nom de facteur de Fano. Ce facteur permet de détecter les corrélations dans les écarts par rapport à la valeur moyenne et est particulièrement pertinent dans le cas d'une population de clients émettant plusieurs PRs chacun.

Lorsque le facteur de Fano est supérieur à un, cela indique l'existence de corrélations significatives dans les écarts par rapport à la moyenne. Dans notre contexte, cela suggère la présence de corrélations dans les fluctuations du nombre de PRs émis par les clients. Le facteur de Fano nous fournit ainsi une mesure simple du facteur de conversion moyen X, qui représente le nombre moyen de PRs émis par client.

Il convient de noter que l'équation (V-9) tient également compte d'un terme supplémentaire qui prend en compte la possible variabilité de la valeur de X. Cette variabilité peut être liée à des facteurs tels que les fluctuations dans le comportement des clients, les variations temporelles dans les PRs émis ou d'autres sources de variation. Il est important de prendre en compte cette variabilité potentielle, car elle peut augmenter avec le nombre de PRs émis.

Dans notre étude, nous allons maintenant appliquer cette équation à nos données de PR de MAC aléatoires, également appelées CR. En analysant le facteur de Fano et en estimant la valeur de X à partir de nos données, nous pourrons obtenir des informations précieuses sur le comportement des clients et les taux d'émission de PRs. Cela nous permettra de mieux comprendre et estimer le nombre de client-jours à partir du nombre de PRs observés.

V.B.3. Résultat

En utilisant un modèle similaire à celui illustré dans la figure 17, nous procédons maintenant au calcul de la variance de P à partir de la dispersion des points de données par rapport au modèle construit, afin de déterminer un X pour chaque site. Cependant, étant donné que la variance est sensible aux valeurs aberrantes, il est nécessaire de traiter ce problème de manière appropriée. Pour ce faire, nous avons mis en place une méthode de détection des valeurs aberrantes en combinant les distributions de X de tous les sites pour former un seul graphique, puis en choisissant un seuil adéquat. La distribution résultante est présentée dans la figure 18.

Une limitation de cette approche réside dans le fait que les distributions de X peuvent varier d'un site à l'autre. Cependant, étant donné que nos échantillons statistiques sont quelque peu limités - une seule mesure par jour sur une période de quelques semaines - nous avons opté pour l'utilisation de la distribution combinée, considérée comme plus fiable pour notre analyse. La coupure résultante, fixée à X = 4000, est située au-delà du corps principal de la distribution, marquant le début de la "queue". Cette coupure nous permet de conserver 95% des données dans notre analyse.

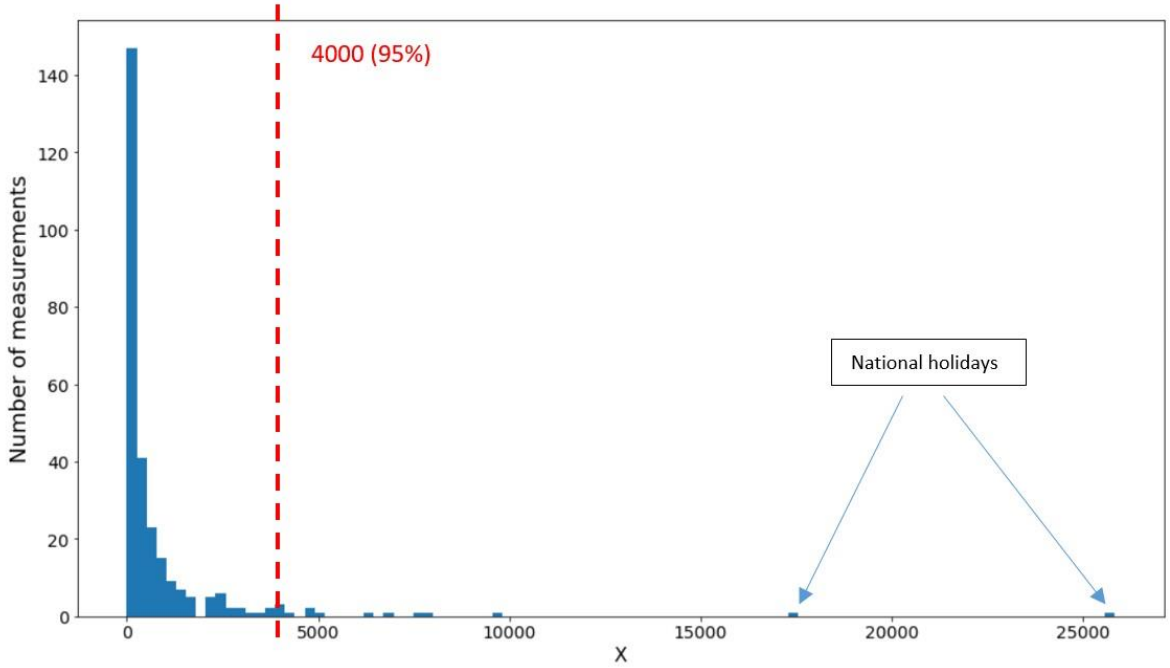


Figure 18 La coupure utilisée pour exclure les valeurs aberrantes de la distribution de X.

Dans la figure 18, nous pouvons observer deux valeurs aberrantes correspondant aux jours fériés nationaux en France, qui doivent clairement être exclues de notre échantillon. Il est également possible que d'autres sources de fluctuations importantes et corrélées de la population des clients, moins "formelles" mais tout aussi significatives, contribuent à la "queue" de la distribution. Parmi ces sources, nous pourrions citer des événements tels que des grèves ou des pannes de transport, des mouvements sociaux, et d'autres facteurs externes susceptibles d'influencer le nombre de clients présents.

Il est essentiel de prendre en compte ces facteurs lors de l'analyse des résultats afin de s'assurer que notre modèle et nos estimations de X soient robustes et représentatifs de la réalité. Dans la Figure 19, nous présentons les valeurs moyennes de X résultantes pour les cinq ensembles de données de villes, en fonction de P, accompagnées d'une barre d'erreur indiquant l'écart type de chaque moyenne mesurée de X. Cette représentation graphique nous permet d'observer le comportement de X sur l'ensemble de la plage de valeurs de P couverte par nos ensembles de données, qui s'étend sur près d'un ordre de grandeur.

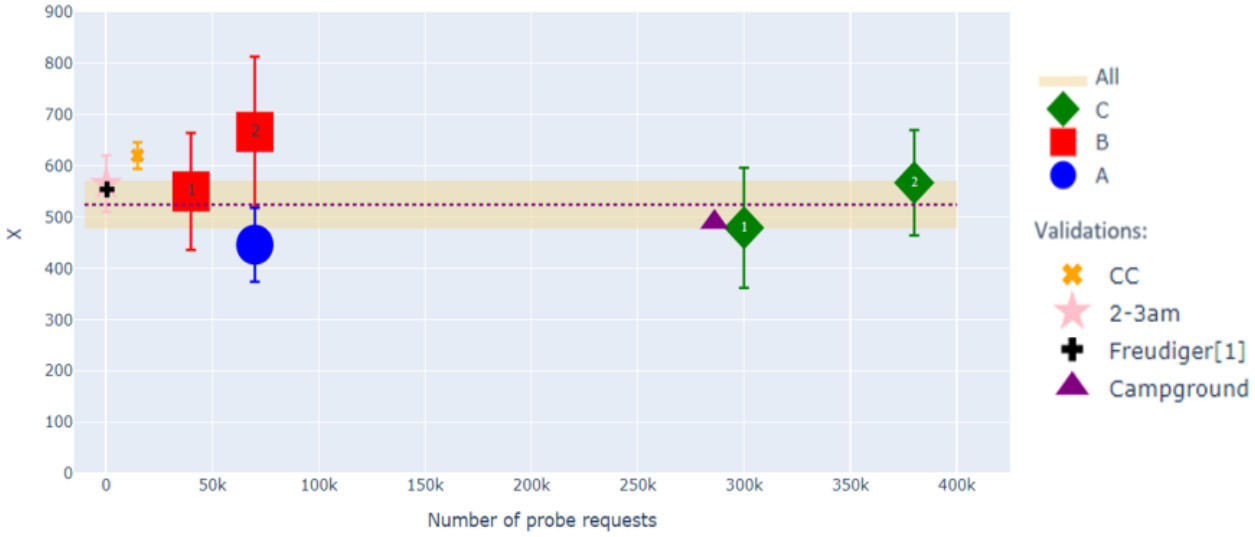


Figure 19 La valeur de X pour les 5 ensembles de données des villes A, B et C, en fonction de P. Les points de validation indiqués dans la figure sont discutés dans la section 3 ; pour une meilleure lisibilité, une validation supplémentaire utilisant [11] n'est pas affichée sur le graphique mais est discutée en détail dans la section 3.

L'analyse de la Figure 19 révèle que, dans la de précision de notre méthode, les valeurs de X restent relativement constantes sur l'ensemble de la gamme de valeurs de P. Cela indique que, pour l'instant, nous pouvons négliger le terme linéaire en P dans l'équation V-9 sans introduire de biais significatif dans nos estimations. Il est important de souligner que les valeurs de X peuvent varier d'un site à l'autre en raison de différentes caractéristiques propres à chaque site, telles que le nombre de points d'accès (AP), la disposition des AP et les effets de propagation spécifiques. Cependant, les incertitudes statistiques actuelles semblent masquer ces différences potentielles entre les sites. Par conséquent, dans un souci de représentativité et de robustesse, nous avons choisi de combiner les données pour obtenir une valeur moyenne globale de X, ainsi qu'un écart type représentatif de l'ensemble des sites.

Dans la Figure 19, la valeur moyenne globale de X est indiquée par une ligne en pointillés à l'intérieur de la bande jaune. Cette valeur moyenne globale, égale à 524 ± 47, est obtenue en tenant compte de la dispersion observée dans les données de chaque site. En combinant les données de cette manière, nous obtenons une estimation plus fiable et représentative de la

conversion moyenne de clients en PR, tout en prenant en compte les variations potentielles entre les sites.

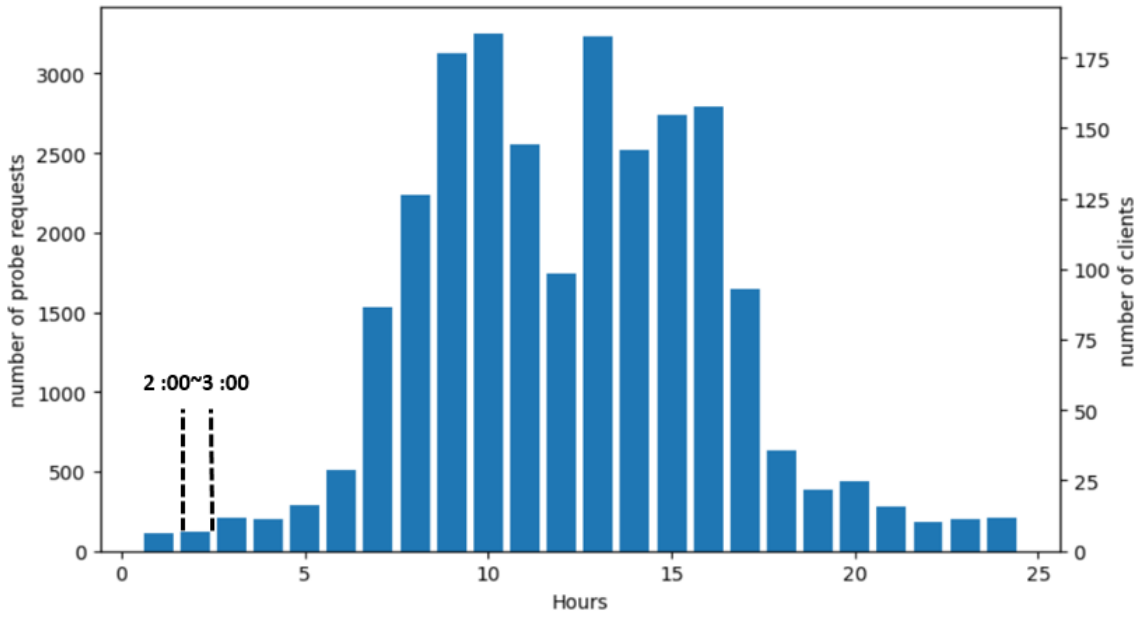


Figure 20 Valeurs typiques horaires de CR P (axe de gauche) et C (axe de droite) pour la ville-A. L'axe de droite est calibré en heures-client. La période surlignée de 02:00 à 03:00 sera utilisée pour une validation dans la section 3.

Une fois que nous avons obtenu une valeur satisfaisante de X, le nombre prédit de journées-client pour une date particulière peut être simplement calculé en utilisant l'équation C = P/X. Si l'on souhaite étudier l'occupation d'un site tout au long de la journée, les horodatages des PR peuvent être utilisés pour diviser le nombre de journées-client en heures-client, en divisant X par 24, comme illustré pour une journée typique de la ville A dans la Figure 20.

La période mise en évidence dans la figure, de 02:00 à 03:00, sera utilisée dans l'une des validations de la section suivante. Cette subdivision en heures-client nous permet d'analyser plus en détail l'occupation du site pendant des périodes spécifiques de la journée, ce qui peut être particulièrement pertinent pour certaines études et évaluations.

En utilisant cette approche, nous pouvons obtenir des estimations du nombre de clients présents sur le site à des intervalles horaires spécifiques, ce qui peut contribuer à une meilleure compréhension des modèles d'occupation et des schémas de fréquentation. Ces informations peuvent être utilisées dans divers contextes, tels que la planification de la capacité, l'optimisation des ressources et l'évaluation des performances.

V.C. Validation des résultats

Les données utilisées dans cette étude proviennent de systèmes WiFi en extérieur en fonctionnement, qui ne sont pas équipés de systèmes auxiliaires de référence au sol tels que des caméras, etc. Par conséquent, il n'est pas possible de valider directement nos résultats par rapport à un tel système. Cependant, nous avons pu valider la raisonnabilité des valeurs de X produites par notre méthode en utilisant les statistiques de notre CC ; en étudiant notre CR avec une fenêtre temporelle alternative ; et en dérivant une valeur de X pour un camping de nos ensembles de données, pour lequel nous avons pu obtenir des comptages de personnes pendant une courte période fixe. De plus, nous comparons nos résultats avec deux études de la littérature qui sont également basées sur l'obtention d'un taux global d'émission de PR WiFi par les téléphones. Au total, cinq validations distinctes sont présentées dans la suite de l’étude.

Ces validations multiples renforcent la validité et la crédibilité de notre approche et confirment que les valeurs de X que nous avons obtenues sont raisonnables et fiables. Bien que nous ne puissions pas valider directement nos résultats par rapport à un système de référence auxiliaire, ces validations alternatives nous permettent de démontrer la solidité de notre méthodologie et d'apporter une assurance supplémentaire quant à la validité de nos conclusions.

V.C.1. Validation avec CC

Lorsque les adresses MAC sont fixes, comme pour les classes CC et CF, le nombre de clients réels peut être compté, et la durée de leur séjour peut être obtenue à partir des horodatages. Ainsi, bien que nous ne croyions pas que CC ou CF seront utiles pour estimer l'audience, nous pouvons néanmoins mesurer les valeurs de X pour ces classes et les comparer à celles obtenues

avec notre méthode pour la classe CR. En raison de la difficulté de distinguer les clients réels des objets connectés dans CF, nous choisissons de valider en utilisant CC.

Comme mentionné précédemment, CC peut utiliser la connexion WiFi souscrite comme point d'ancrage pour des services améliorés, par exemple, le P2P [22] ; les protocoles Spanning Tree [62] ; les *chats* ; les réseaux sociaux ; les talkies-walkies ; les jeux en ligne ; et les répéteurs WiFi, etc., ce qui crée un trafic PR élevé en utilisant des PR pour la découverte de la topologie ; le routage ; ou le transfert de messages encapsulés dans des PR. Dans notre étude, nous avons effectivement observé un comportement cohérent avec ES chez un faible pourcentage de nos clients CC, ce qui a pour effet de fausser les mesures de X vers des valeurs plus élevées. Le prétendu trafic ES était sporadique dans le sens où les clients concernés avaient tendance à être exceptionnellement actifs dans des blocs continus d'une durée de quelques heures. Pour résoudre le problème, une "ES-cut" a été développée, définie comme l'émission de plus de 80 PR par bloc de 6 heures, en moyenne sur toute la période de mesure (c'est-à-dire en incluant également les blocs vides). La figure 21 pour la ville B explique comment la valeur de coupure a été déterminée. Bien qu'il ne soit pas possible de confirmer absolument que les clients exclus étaient engagés dans ES, pour les villes A et B, où le problème était le plus prononcé, nous avons pu confirmer, en utilisant les valeurs enregistrées de l'indicateur de force du signal reçu, ou RSSI, des PR, pour localiser leurs points d'émission, que ces clients fréquentaient exclusivement un petit nombre de lieux fixes dans ou près des bâtiments administratifs de la ville. Cela constitue une autre preuve de la nature non standard de ces clients.

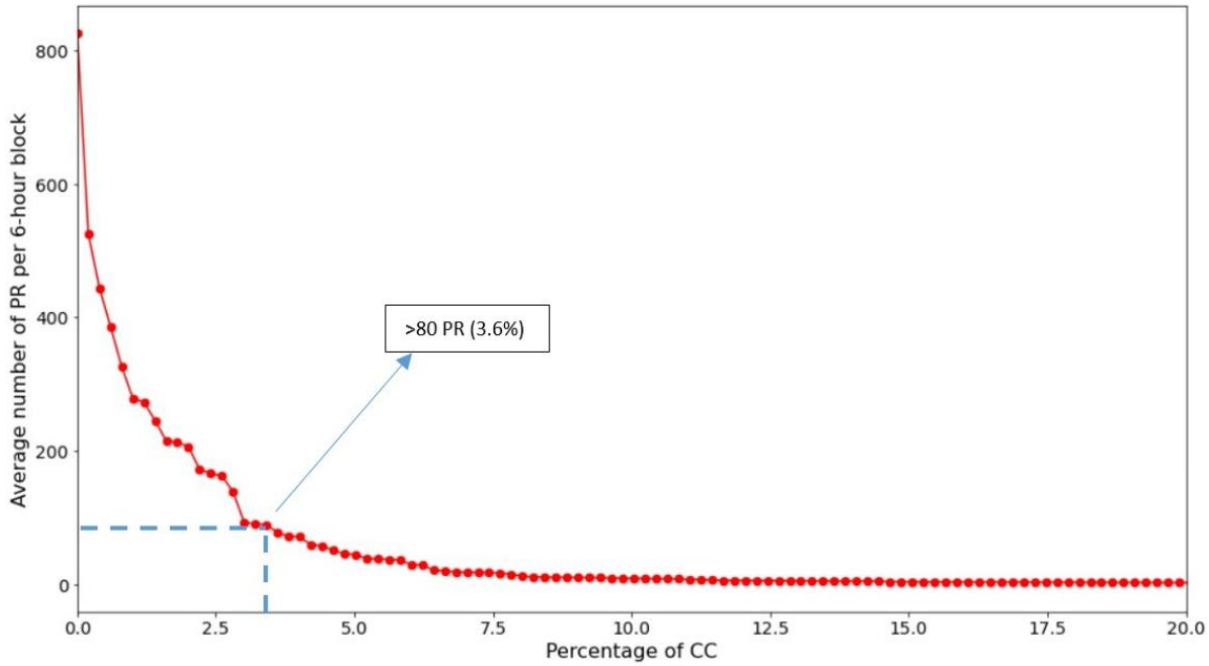


Figure 21 La coupure ES élimine les clients CC qui émettent un nombre anormalement élevé de PR par blocs de six heures, ce qui est cohérent avec des services P2P ou d'autres services étendus. Les données concernent la ville B.

Pour valider notre méthode basée sur la classe CC, nous utilisons la quantité X_{CC} qui représente le nombre moyen de PR par CC au cours d'une journée, corrigé en fonction de l'occupation quotidienne moyenne des CC calculée à partir des horodatages des PR. Ainsi, nous obtenons $X_{CC} = P/(A_{CC}<t_{CC}>)$, où A_{CC} correspond au nombre de CC et $<t_{CC}>$ à leur durée moyenne de séjour. Les résultats de cette validation sont présentés dans le Tableau 2, en incluant les cas avec et sans l'ES-cut, ainsi que le pourcentage de clients conservés après l'application de la coupe. Étant donné que la classe CC ne contribue qu'à un petit nombre de PR, les valeurs de X_{CC} dans le tableau sont rapportées sous la forme d'un seul point représentant la valeur moyenne de X_{CC}, situé tout à gauche dans la Figure 21.

Il est important de noter que, contrairement à la méthode basée sur CR qui mesure la variance, la mesure de X_{CC} est basée sur une expérience de comptage, ce qui entraîne des écarts-types légèrement plus faibles par rapport aux valeurs de X basées sur CR. Cela est dû à la nature du calcul de X_{CC}, qui ne prend pas en compte les variations individuelles des CC, mais plutôt une mesure agrégée sur l'ensemble de la classe. Cependant, même avec cette différence, les valeurs

de X_{CC} fournissent une validation cohérente et raisonnable de notre méthode, en confirmant que les estimations de X obtenues à partir de la classe CR sont fiables et représentatives de la réalité.

En résumé, la validation basée sur la classe CC nous permet de comparer les résultats de X obtenus à partir de cette classe avec ceux obtenus à partir de la classe CR. Cela nous donne une perspective supplémentaire pour évaluer la performance et la précision de notre méthode, renforçant ainsi la confiance dans les estimations de X que nous utilisons pour prédire le nombre de jours-client et étudier l'occupation des sites.

Tableau 3 X_{CC} values for the different sites.

Site	X_{CC} **raw**	X_{CC} **ES-cut**	**% clients**
city-A	802	468	99.1%
city-B1	2329	778	96.4%
city-B2	2253	948	96.4%
city-C1	743	696	99.91%
city-C2	405	378	99.97%

Nous tenons à souligner que X et X_{CC} ne sont pas a priori censés être identiques en raison des natures différentes des clients CR et CC, des statistiques réduites de CC par rapport à CR, et du fait que les taux d'émission de PR sont connus pour dépendre de l'état de l'appareil [32]. Il ne serait pas surprenant de constater que X et X_{CC} soient similaires, comme c'est effectivement le cas ici, ce qui renforce l'idée que la technique proposée pour obtenir X est effectivement sensible aux statistiques sous-jacentes des clients CR.

V.C.2. Validation sur la base d'un modèle alternatif : 02:00 - 03:00

La deuxième validation consiste à utiliser CR dans la ville A et se concentre sur une période spécifique de 02h00 à 03h00 chaque jour (comme indiqué dans la figure 20). Cette plage horaire

est choisie car elle présente généralement un nombre plus faible de PR et est censée être relativement indépendante des variations saisonnières et des périodicités liées à la semaine de travail, en raison de sa nature nocturne. Pour effectuer cette validation, un nouveau modèle est créé spécifiquement pour la période de 02h00 à 03h00. La méthode du facteur de Fano, similaire à celle utilisée dans la validation précédente, est appliquée à ce modèle. Les valeurs de X obtenues à partir de cette analyse sont soumises à la même limite de 4000, introduite dans la section 2, mais adaptée à la fenêtre temporelle plus réduite.

Les résultats provenant des différents sites de la ville A sont combinés pour obtenir une seule mesure, qui est affichée près de zéro PR dans la figure 19. La figure montre que la valeur obtenue pour $X_{02\text{-}03}$ est cohérente avec les valeurs de X obtenues à partir des modèles hebdomadaires, ce qui confirme la fiabilité de notre méthode. Cette validation confirme que notre approche est robuste pour différentes périodes de temps et fournit des résultats cohérents, renforçant la confiance dans l'exactitude de notre technique d'estimation de X.

V.C.3. Validation de la littérature I

Freudiger et al. [32] ont mesuré les périodes d'interarrivée des PR pour les appareils Samsung, iPhone et Nexus dans différents états - écran allumé, navigateur ouvert, etc. On peut utiliser ces mesures pour déduire une période d'interarrivée moyenne globale de 156 s, ce qui permet d'obtenir $X_{Freudiger} = 554$, qui est représenté par le symbole "+" à P = X du côté gauche de la Figure 19. Cette valeur n'est pas techniquement directement comparable à notre X, qui mesure le nombre de PR reçus d'un client dans un réseau déployé réel, et non le nombre émis, qui peut dépendre du nombre et de la disposition des points d'accès locaux. De plus, la variété de téléphones explorée dans l'étude de [32] est probablement quelque peu limitée par rapport à celle de nos données du monde réel. Cependant, le fait que $X_{Freudiger}$ s'aligne raisonnablement avec les résultats obtenus avec notre méthode est encourageant.

V.C.4. Validation de la littérature II

Dans [11], une technique d'estimation de l'audience à partir de PR aléatoires utilisant un facteur de conversion fixe est présentée. Il est donc intéressant de comparer ce travail avec notre

technique. Cependant, il convient de noter que les deux méthodes ne sont pas directement comparables pour plusieurs raisons. Tout d'abord, [11] traite des réseaux intérieurs, tandis que tous nos ensembles de données sont extérieurs. Deuxièmement, la technique décrite dans [11] nécessite une procédure d'accord préalable basée sur le RSSI au niveau du point d'accès, utilisée pour éviter le double comptage des PR, ce qui n'est pas possible dans notre cas car nos ensembles de données ont été fournis tels quels, sans possibilité d'effectuer un tel accord. Troisièmement, bien qu'il ne soit pas précisé dans [11] si les clients sont connectés ou non, l'environnement de type bibliothèque universitaire étudié dans ce travail semblerait propice à ce que les utilisateurs se connectent réellement au WiFi, ce qui, selon notre estimation, pourrait donner des résultats différents de ceux des clients de la classe CR. La principale similitude entre [11] et notre travail réside dans l'utilisation d'un facteur de conversion constant des clients en PR, appelé β dans [11]. Cependant, cela représente également la plus grande différence par rapport à notre méthode, car dans [11], β est ajusté à une valeur de vérité terrain obtenue à partir d'un système de surveillance par caméra. En revanche, dans notre cas, X est obtenu uniquement à partir d'arguments statistiques. Néanmoins, placer la valeur ajustée β sur la même échelle que X nous permet de valider si les deux approches semblent mesurer le même phénomène. Pour ce faire, il est nécessaire de spécifier un facteur de couverture supplémentaire, κ, défini dans [11], qui vise à prendre en compte la géométrie du site étudié et l'efficacité résultante de la détection des PR. Les expériences présentées dans [11] ont des valeurs de κ allant de 1, une couverture totale, à 3, une couverture partielle, en raison des zones non couvertes par le WiFi. Étant donné que nos PR sont comptées indépendamment par tous les points d'accès, et que nous n'avons effectué aucun accord préalable, il n'est pas possible de spécifier une valeur appropriée de κ pour nos données, même si nous pourrions nous attendre à ce que cette valeur soit dans la même plage que dans [11]. En faisant cette hypothèse et en mettant β à l'échelle sur une période d'une journée, on obtient une plage de valeurs $960 > X > 320$, qui englobe assez bien les valeurs obtenues sur nos ensembles de données. La conclusion est qu'un système ajusté à une vérité terrain basée sur une caméra donne des valeurs X assez similaires à ce que nous obtenons uniquement à partir des statistiques. En ce qui concerne la précision, les tailles de foule prédites dans [11], également ajustées à une vérité terrain, suivent cette vérité terrain au fil du temps avec une précision de quelques pourcents, ce qui est bien meilleur que les 10 à 20 % obtenus ici. Ce résultat suggère néanmoins que si la précision de notre méthode statistique peut être

améliorée, X devrait pouvoir donner une bonne représentation de la réalité sans avoir besoin d'un système de vérité terrain.

V.C.4. Validation dans un camping

Obtenir des données précises sur le nombre de visiteurs d'une ville est difficile sans disposer d'un système coûteux de caméras ou d'autres capteurs. Cependant, notre ensemble de données comprend également des enregistrements d'un terrain de camping régional comprenant des piscines, des sentiers naturels, des restaurants et d'autres activités, pour lesquels des décomptes quotidiens sont enregistrés, même s'ils ne sont généralement pas disponibles au public. Bien que nous n'ayons pas de détails précis sur les allées et venues des clients sur le terrain de camping, une demande spéciale a permis d'obtenir des informations sur une période de 10 jours, incluant un long week-end de mai 2021, auprès du directeur du site. Il a rapporté la présence de 584 visiteurs en séjour pendant cette période. Au cours de cette période, ces clients ont généré environ 286 000 PR, ce qui donne une valeur de $X_{campground}$ égale à 490. Cette valeur est également représentée dans la figure 19 et correspond de manière satisfaisante aux autres cas étudiés. Dans ce cas, nous constatons que CC au camping ne produit qu'une petite fraction de ce nombre de PR. En revanche, des centaines de milliers de PR supplémentaires sont produits par CF, la majorité étant identifiables comme des objets connectés (IoT), soulignant une fois de plus l'importance de mesurer la taille de la foule exclusivement avec CR.

V.D. Conclusion

L'utilisation des signaux de réseau radio pour identifier et suivre les appareils sans fil a été étudiée pendant de nombreuses années en utilisant une grande variété de techniques [63] [64]. Récemment, cela s'est transformé en l'utilisation généralisée de systèmes de surveillance de l'audience basés sur les adresses MAC des clients contenues dans les PR. Avec l'arrivée des préoccupations concernant la vie privée des clients et du GDPR, ainsi que la croissance de l'IoT et des services clients gourmands en bande passante, cette approche devient rapidement intenable, ce qui crée un besoin de nouvelles approches préservant la vie privée.

Ici, une analyse du facteur de Fano des données PR provenant de trois villes françaises, associée à un modèle de template déterministe, nous permet de produire une mesure approximative, X, du nombre moyen de PR produits par jour par les clients WiFi utilisant des adresses MAC aléatoires. Ces valeurs sont relativement stables sur différents sites et sur une plage de volumes de PR d'environ un ordre de grandeur. Les validations basées sur les clients connectés, un modèle de template alternatif, une comparaison avec deux mesures pertinentes de la littérature et un test de vérité terrain sur un site de camping montrent que les prédictions basées sur X sont raisonnables.

La technique présente cependant quelques inconvénients. En se basant sur le calcul de la variance, sa sensibilité aux valeurs aberrantes entraîne une incertitude statistique qui doit être améliorée. L'utilisation d'une statistique de déviation absolue médiane (MAD) plutôt que d'une variance pourrait être une piste à suivre, à condition que la relation entre les valeurs MAD obtenues et la valeur X nécessaire puisse être découverte. De plus, X ne peut actuellement être évalué que sur des données PR présentant des périodicités quotidiennes et/ou hebdomadaires. La validation basée sur des fenêtres horaires présentée ici peut fournir des indices pour appliquer la méthode à des données dépourvues de périodicités à plus grande échelle. Enfin, on peut se demander si les clients des services améliorés mentionnés concernant CC pourraient également être présents dans les groupes CR. Bien que cela semble peu probable, car une connexion Internet fixe est un élément essentiel de ces services, c'est une éventualité à prendre en compte. D'autres perspectives incluent une analyse théorique plus rigoureuse des statistiques des PR produits dans les réseaux WiFi publics, une analyse de la variabilité de X pour différents sites et populations de clients, et une étude plus approfondie de la procédure de seuillage pour définir les classes CR et CF.

L'estimation de la taille de l'audience à partir des adresses MAC aléatoires est encore un domaine nouveau. L'efficacité ultime de la technique proposée viendra probablement avec le temps grâce à des tests sur le terrain et des opportunités de validation. En attendant, en raison de sa simplicité et de son applicabilité facile dans des systèmes réels, cela devrait susciter un intérêt substantiel.

VI. Densité

VI.A. Introduction

Depuis l'introduction des réseaux WiFi commerciaux il y a plus de vingt ans, il existe un intérêt considérable pour localiser les utilisateurs du service à des fins commerciales, de sécurité et de surveillance. La localisation est généralement effectuée en utilisant les propriétés physiques des signaux WiFi émis par les clients, tels que la RSS à plusieurs AP. La plupart des systèmes exploitent les trames PR émises régulièrement par les appareils clients, même lorsqu'ils ne sont pas connectés à un réseau. Aujourd'hui, les interfaces sophistiquées de localisation des clients WiFi sont largement répandues dans le monde entier. Les sections VI.A.1, VI.A.2, VI.A.3, VI.A.4 et VI.A.5 fournissent des informations historiques et technologiques sur l'état actuel de ce domaine.

VI.A.1. Surveillance des clients pour les mesures commerciales

Le flux de personnes, également connu sous le nom de comptage des clients ou simplement de trafic, est une mesure du nombre de visiteurs entrant dans un site commercial, tel qu'un magasin de détail, un centre commercial, un musée, etc. Depuis des décennies, le comptage des flux de personnes est utilisé pour aider les détaillants et les gestionnaires de sites à évaluer l'attrait de leurs offres et à améliorer l'expérience globale des clients. Initialement réalisé à l'aide de compteurs manuels, de tourniquets ou de tapis sensibles au poids placés dans les entrées, le comptage des flux de personnes a évolué pour devenir une métrique commerciale clé obtenue généralement à l'aide de techniques plus sophistiquées telles que la vidéosurveillance, les détecteurs thermiques infrarouges passifs (PIR), les imageurs infrarouges actifs de temps de vol (ToF), le WiFi, etc. Les données des capteurs sont généralement transférées via un protocole Internet pour être stockées et analysées sur un serveur. Les statistiques en temps réel sur le flux de personnes sont maintenant devenues une donnée précieuse pour l'analyse commerciale, utile non seulement pour l'analyse du commerce de détail et du marketing, mais aussi dans les applications de "Smart Building" et de "Smart City", y compris la surveillance de la consommation d'énergie, les flux de personnes, la sécurité, etc.

VI.A.2. Surveillance des clients sans fil

Alors que les appareils clients basés sur la radio, tels que les audio-guides, sont utilisés depuis plusieurs années, c'est avec l'essor des téléphones portables dans les années 1990 que l'intérêt pour l'exploitation de la connectivité de ces appareils à des fins de comptage des flux de personnes a commencé à croître. Cependant, avec l'arrivée des smartphones et leur adoption quasi universelle par les consommateurs, la perspective d'une carte d'interface réseau (NIC) WiFi/Bluetooth active dans la poche de chaque client a conduit à un changement de paradigme dans la surveillance des clients mobiles. En effet, bien que de nombreux outils de comptage des flux de personnes mentionnés dans la section précédente soient encore utilisés, l'utilisation de technologies sans fil pour collecter les données d'activité des clients a connu une véritable explosion. Au-delà du simple comptage des clients, le sans-fil permet également de localiser les clients en analysant la puissance de leurs signaux radio. Ainsi, en exploitant les appareils clients omniprésents connectés à Internet et un réseau sans fil existant, la surveillance des clients sans fil peut être beaucoup moins dépendante de matériels spécialisés, d'expertise nouvelle et de déploiements, calibrations, opérations et maintenances intensifs en main-d'œuvre, par rapport aux caméras de surveillance, aux capteurs infrarouges ou aux solutions de comptage des flux de personnes traditionnelles.

Motivés par le désir d'une connectivité omniprésente, les réseaux locaux sans fil (WLAN) utilisant le WiFi, ou dans certains cas, les technologies Bluetooth, sont aujourd'hui présents dans presque tous les espaces publics commerciaux, et la couverture WiFi à l'échelle de la ville devient rapidement la norme. De nombreux services de localisation personnels bien connus, tels que Google Maps et d'autres, exploitent les réseaux WiFi et Bluetooth existants, ainsi que les signaux GPS, pour la navigation et l'orientation dans les espaces intérieurs et extérieurs. Cependant, du point de vue de la surveillance des clients, ce sont généralement les signaux WiFi générés par les smartphones eux-mêmes, capturés par les points d'accès réseau (AP), qui suscitent un intérêt principal pour le comptage et la localisation.

En effet, pour assurer un accès rapide aux réseaux WiFi ambiants, les NIC WiFi des smartphones émettent régulièrement certaines trames de contrôle qui établissent une liste des réseaux disponibles et annoncent la présence du dispositif ainsi que ses capacités. Les PR,

émises environ 500 fois en 24 heures par un NIC typique [19] [32], même lorsqu'il n'est pas connecté à un réseau, identifie le client auprès du fournisseur de services via l'adresse d'accès au support unique du NIC, ou adresse MAC. De nos jours, tous les principaux fabricants de réseaux sans fil, tels que Cisco/Meraki et CommScope/Ruckus, pour n'en citer que deux, commercialisent des solutions sophistiquées clés en main de surveillance des clients WiFi et d'analyse commerciale pour les entreprises, les centres commerciaux, les Smart Cities, etc., basées sur ce modèle. Ces systèmes offrent des interfaces graphiques en temps réel qui affichent des estimations du nombre de clients, des statistiques de trafic, des paramètres de gestion, etc., pour les différentes zones couvertes par le réseau sous-jacent, fournissant ainsi des données sur le flux de personnes ainsi que d'autres analyses commerciales intéressantes pour les responsables d'entreprise qui déploient le logiciel.

VI.A.3. Protection des données

Cependant, depuis 2016, avec l'avènement du Règlement général sur la protection des données (GDPR) européen et de législations similaires ailleurs dans le monde, à l'instar du comptage de clients déjà discuté, bon nombre des opportunités de surveillance et de localisation des clients WiFi ont aussi été remises en question. Afin de préserver l'identité des clients, les adresses de contrôle d'accès au support sont désormais considérées comme des informations privées et doivent être anonymisées à l'aide d'une fonction de hachage non réversible. Étant donné qu'une adresse MAC donnée est toujours hachée en une même valeur de jeton, pour une protection supplémentaire, tous les appareils WiFi des utilisateurs fabriqués aujourd'hui attribuent également une nouvelle adresse MAC aléatoire à chaque transmission. Ces techniques ont déjà été introduites dans le premier volet de ce manuscrit. Par conséquent, les fournisseurs de services WiFi sont maintenant confrontés à une multitude de PR sans aucune possibilité de les corréler avec les différents clients, rendant ainsi invalide le modèle de surveillance des clients WiFi basé sur la norme de facto décrite précédemment.

VI.A.4. Comptage des clients WiFi avec adresses MAC aléatoires

Des solutions pour contourner ce problème existent. L'une d'entre elles consiste à utiliser des indices dans le contenu numérique des trames reçues pour dérandomiser efficacement les PRs,

comme nous l'avons déjà cité. Une autre consiste à se concentrer sur les clients connectés qui, en s'abonnant au service WiFi, ont donné leur consentement à l'utilisation de leur adresse MAC fixe et privée. Ces deux approches présentent toutefois des inconvénients importants, comme nous avons déjà vu dans le premier volet sur le comptage.

Au-delà de ces solutions de contournement, comme a déjà été discuté, pour des populations de clients suffisamment grandes, le nombre de PRs s'avère proportionnel au nombre réel de clients. Le problème qui subsiste alors consiste à déterminer la constante de proportionnalité appropriée, ce qui peut être réalisé soit par comparaison à une vérité terrain obtenue à partir d'autres capteurs tels que des caméras, des moniteurs d'entrée, etc., comme dans [11] [17] [57], soit sans vérité terrain, mais avec une précision réduite, en exploitant les périodicités observées dans les graphiques représentant les comptages de PRs en fonction du temps, comme explicité dans la section précédente ainsi que dans [19]. Ce type d'approche constitue le point de départ des techniques de localisation des clients mobiles qui seront développées dans ce qui suit.

VI.A.5. Localisation des clients WiFi anonymisés

Si le problème du comptage des clients avec des adresses MAC aléatoires est peut-être en train d'être résolu, comme le suggèrent les observations de la Section VI.A.1.4, attribuer un emplacement à un groupe de clients anonymisés présente plusieurs difficultés supplémentaires. La localisation dans les réseaux sans fil est un domaine actif depuis plusieurs années, avec une grande variété de techniques proposées [65] [64] [66] [67] [68]. En utilisant le WiFi sans assistance, des précisions de positionnement d'environ 2 m en intérieur ou de 10 m en extérieur sont possibles lorsqu'un nombre adéquat de points d'accès sont impliqués. Une méthode simple, populaire et efficace de localisation dans les réseaux WiFi extérieurs, qui s'adapte facilement à la localisation WiFi des smartphones, est la triangulation, dans laquelle la position d'un utilisateur est calculée à partir des valeurs de RSS reçues simultanément par trois points d'accès ou plus. La triangulation est en réalité possible pour une seule PR avec une adresse MAC aléatoire, car tous les points d'accès recevant PR produisent des jetons de hachage identiques. Cependant, en pratique, dans de nombreux réseaux extérieurs aujourd'hui, l'accent est principalement mis sur la fourniture de connectivité, tandis que la capacité à effectuer également une localisation, qui est plus coûteuse car elle nécessite une redondance plus élevée

des points d'accès, reste d'importance secondaire. Dans un tel cas, une fraction importante des PRs émises peut être capturée par un seul ou deux points d'accès, rendant la triangulation impossible. De plus, les émissions suivantes comporteront de nouveaux jetons d'adresse MAC non vus auparavant, ce qui rend impossible de savoir s'ils proviennent du même appareil que PR précédente ou d'un client différent.

En effet, bien que la localisation soit par nature suboptimale dans les réseaux où la localisation n'était pas une priorité de conception initiale, sous le GDPR, les solutions standard deviennent a priori impossibles pour tous sauf les abonnés connectés ayant des adresses MAC fixes. Par conséquent, des outils permettant d'effectuer une localisation malgré ces difficultés sont aujourd'hui largement demandés. L'objectif de ce chapitre de la thèse est de proposer un ensemble d'outils destinés à être un premier pas vers le rétablissement de la surveillance des clients à des fins commerciales et de sécurité, à une époque où de nombreuses approches traditionnelles ne sont plus disponibles.

VI.A.4. Boîte à outils proposée

Dans ce chapitre, nous vous présentons un ensemble de 9 outils spécifiquement conçus pour répondre aux défis liés à la prétraitement, l'affichage, l'interprétation, la localisation et le décompte des clients des réseaux WiFi extérieurs. Ces outils ont été développés afin de surmonter les limitations imposées par les adresses MAC aléatoires et les réglementations telles que le GDPR en Europe et des législations similaires dans le reste du monde.

Notre objectif principal avec cette trousse à outils est de transformer les décomptes bruts des trames PR avec des adresses MAC aléatoires en une carte de densité calibrée en nombre de clients à chaque position. Pour atteindre cet objectif, nous avons rassemblé les éléments essentiels de plusieurs méthodes statistiques et mathématiques, et nous les avons intégrés de manière cohérente dans un ensemble d'outils qui permettent de résoudre un problème complexe que de nombreux opérateurs de réseaux considèrent comme insoluble.

- Un outil de Prétraitement pour sélectionner les populations de clients à étudier.
- Un outil de visualisation appelé "Bowl" pour afficher les densités de probabilité de localisation.

- Un outil de Renormalisation pour condenser les probabilités distribuées en régions de décision.
- Un outil de Décompte pour estimer la proportionnalité entre les trames PR et les clients.
- Un outil de Filtrage pour supprimer les pics de densité qui obscurcissent les structures à plus petite échelle.
- Un outil de Localisation pour la localisation des appareils et le calibrage du modèle de propagation.
- Un outil Fiducial pour délimiter les régions de densité de clients.
- Un outil de Drill-down pour l'étude des densités de clients à différentes échelles.
- Un outil de Réassemblage pour créer la carte finale calibrée en clients plutôt qu'en trames PR.

Certains de ces procédures sont des méthodes simples où le terme "outil" peut sembler excessif ; pour d'autres, plus complexes, l'étiquette s'applique effectivement. Par exemple, les outils "Bowl", Filtrage et Drill-down utilisent des méthodes graphiques relativement standard, tandis que les outils de Fiducial, Localisation, Comptage et Renormalisation sont novateurs et incarnent une grande partie de l'originalité du présent travail.

VI.B. Ensembles de données et définitions

La boîte à outils proposée est présentée tout d'abord en présentant les jeux de données utilisés et quelques définitions nécessaires.

VI.B.1. Jeux de données du réseau WiFi extérieur étudiés

Les jeux de données étudiés comprennent des enregistrements détaillés et horodatés des PRs provenant de réseaux WiFi extérieurs publics dans un site touristique extérieur, Site 1, et un terrain de camping, Site 2, tous deux en France. Ils font partie d'un ensemble de données plus vaste, comprenant les données de la ville française introduites dans le chapitre précédent et en [19]. Notez que, pour le comptage des clients, il existe quelques différences importantes entre les présentes données et les données de la ville de [19] (qui seront discutées dans la Section VI.B.2 et la Section VI.E.1). Ainsi, les deux études sont complémentaires. Les données présentées ici ont été accumulées de la dernière semaine de mars à la fin juin 2021, donnant un total de 7,0 millions de PRs pour le Site 1 et 5,3 millions pour le Site 2. Un fichier distinct stocke les enregistrements de détails de connexion des clients qui se sont à un moment donné connectés au service WiFi. Les agencements des points d'accès des sites apparaissent dans les

figures des sections suivantes. Conformément à la législation européenne actuelle sur la protection des données, toutes les adresses MAC des clients sont remplacées par des chaînes de hachage anonymisées.

VI.B.2 Classes de clients

En comptant le nombre de fois où une adresse MAC anonymisée distincte est détectée dans une fenêtre de temps sélectionnée et en la comparant à un seuil, et en tenant également compte du fichier de détails de connexion, nous pouvons identifier les trames PR appartenant à l'une des trois classes suivantes :

- Client Aléatoire, CR : La chaîne MAC est détectée de 1 à "seuil" fois.
- Client Fixe, CF : La chaîne MAC est détectée plus de "seuil" fois.
- Client Connecté, CC : La chaîne MAC apparaît dans le fichier de détails de connexion.

Les clients connectés sont séparés en premier lieu, de sorte que les classes soient mutuellement exclusives. Ces définitions de classes de clients sont les mêmes que celles introduites dans [19]. Une différence importante est qu'ici, un seuil de 3 est choisi, plutôt que 2 comme dans [19], ce qui s'avérera être un élément clé dans certains des outils présentés.

VI.B.3. Une mise en garde : OUI et IoT

Malgré la randomisation de l'adresse MAC, ses trois premiers octets, appelés OUI, est censé fournir des informations sur le fabricant de la carte réseau WiFi qui produit PR, mais en pratique, les valeurs de l'OUI, souvent, ne sont pas fiables et peuvent correspondent aux dispositifs IoT, comme nous l’avons déjà signalé. Dans cet article, toutes les estimations de densité sont basées sur les clients de classe CR.

VI.B.4. Outil de prétraitement

Pour préparer les données aux étapes restantes de la boîte à outils, il est nécessaire d'effectuer une étape de prétraitement basée sur les points exposés dans la Section VI.B.1, la Section VI.B.2 et la Section VI.B.3. Dans cette thèse, une fenêtre temporelle de base de trois heures a été choisie, suffisamment longue pour fournir un échantillon statistiquement riche de milliers de PRs, mais suffisamment courte pour permettre de suivre le comportement des clients tout au

long de la journée. La catégorie des clients CR est d'abord sélectionnée en appliquant un seuil de 3 au nombre de fois où l'adresse MAC apparaît dans l'échantillon de données sélectionné. Cela élimine les clients abonnés ainsi que les dispositifs IoT, produisant un échantillon de P PRs avec des adresses MAC aléatoires qui ont chacune été vues une, deux ou trois fois. Comme mentionné, P est généralement un nombre dans les milliers ici, ce qui signifie que la grande majorité des adresses MAC sont toutes différentes en raison de la randomisation à chaque nouvelle émission. Ensuite, on compte le nombre de PRs vues trois fois dans cet échantillon, P_3, après avoir vérifié que leurs horodatages sont cohérents avec une réception quasi-simultanée. Nous rappelons que l'adresse MAC d'un PR reçue par plusieurs points d'accès produira le même résultat de hachage à tous les points d'accès concernés. Les quantités P et P_3 seront importantes ultérieurement dans les étapes de Localisation, de Comptage et de Reconstitution.

VI.C. Introduction tutorielle à la boîte à outils proposée

Dans cette section, nous présentons une introduction de type tutoriel à certaines des difficultés rencontrées dans la localisation des réseaux WiFi en extérieur dans le monde réel, ainsi que la manière dont les outils proposés peuvent être utilisés pour y remédier. Nous notons que l'outil de prétraitement a déjà été introduit dans la Section VI.B; quant aux outils de fiducie, de Drill-down et de reconstitution, ils seront abordés dans les Sections VI.D et VI.E La section commence par une introduction à la triangulation en utilisant les valeurs de réception du signal RSS des clients.

VI.C.1. Les bases de la triangulation et de la précision de positionnement attendue

Les réseaux WiFi en extérieur couvrant de vastes zones, il est difficile d'obtenir une couverture complète à la fois pour le trafic et la localisation. De nombreuses approches basées sur le RSSI, ToA et ses variantes, ou les informations sur CSI sont possibles, que ce soit en mode empreinte digitale ou en mode de mesure de distance. Cependant, en raison de sa simplicité et de son utilisation répandue, nous nous concentrons ici sur la triangulation, également appelée trilatération.

L'origine physique de l'approche de triangulation est incarnée dans l'équation de Friis bien connue qui relie le RSS à la distance :

$$RSS(dBm) = A(dBm) + 10n\,log_{10}\big(d(m)\big) \qquad (\text{VI} - 1)$$

Ici, le RSS est mesuré en dB par rapport à 1 milliwatt (dBm), tandis que la distance entre l'appareil client et le point d'accès (AP) d est en mètres. La constante A, également en dBm, est la puissance émise à 1 mètre de l'appareil client - selon les normes de l'industrie, généralement prise comme -30 dBm. La constante n est appelée exposant d'affaiblissement. L'équation de Friis peut être dérivée à partir des premiers principes pour la propagation dans l'espace libre, où la puissance diminue selon l'inverse du carré de la distance, donnant n = 2. Il a également été démontré que la propagation sur une surface plate et dégagée produit n = 4. Pour les situations intermédiaires avec des effets d'encombrement et de trajets multiples, comme on peut s'attendre dans les réseaux WiFi extérieurs, des valeurs dans la plage de 2 à 3 sont rencontrées, selon l'environnement particulier. En ce qui concerne la précision de la localisation, les réseaux WiFi intérieurs, où le nombre d'AP est plus élevé, offrent généralement une précision de quelques mètres. En revanche, dans un réseau WiFi extérieur bien calibré, une précision de +/- 10 mètres est déjà considérée comme un bon résultat. Cette constatation doit être prise en compte lors de l'évaluation de la faisabilité des outils proposés dans ce chapitre.

VI.C.2. L'outil Bowl

Commençons par considérer une configuration de base avec 3 AP disposés en formation triangulaire, comme indiqué dans la Figure 22a. À l'intérieur de ce triangle, il y a un appareil client qui émet un PR. Ce signal se propage vers l'extérieur et est capturé successivement par chacun des trois AP, ce qui donne les valeurs RSS1, RSS2 et RSS3. En inversant l'équation de Friis, nous pouvons calculer les distances R1, R2 et R3 entre l'appareil client et chaque AP. En traçant des cercles de ces rayons comme indiqué dans la figure, nous localisons l'appareil client au point où ils se croisent, indiqué par une étoile noire.

Cette approche de la triangulation nous permet d'approximer la position de l'appareil client en se basant sur les mesures de RSS provenant de plusieurs AP. En utilisant les propriétés

géométriques des cercles qui s'intersectent, nous pouvons déterminer l'emplacement probable de l'appareil client dans la zone de couverture.

Dans la Figure 22a, le point d'intersection des cercles est représenté par une étoile noire, ce qui indique la position probable de l'appareil client à l'intérieur du triangle. Ainsi, en analysant les forces du signal reçu par les différents AP et en utilisant la méthode de triangulation, nous pouvons estimer la localisation de l'appareil client dans une configuration de réseau WiFi.

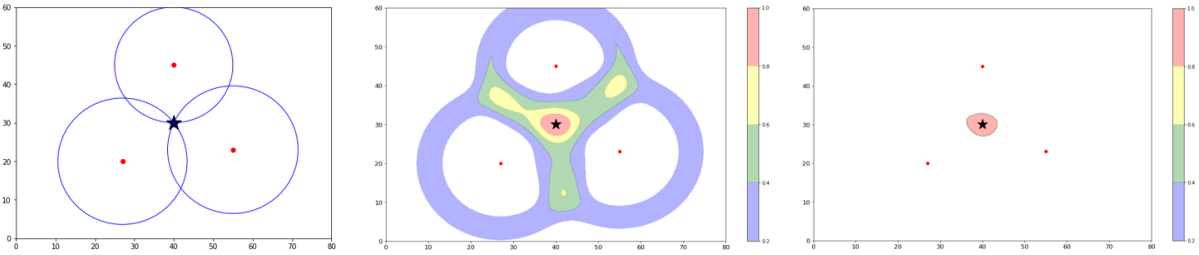


Figure 22 a) Triangulation d'un seul PR en utilisant des cercles déterminés à partir du RSS. La solution est indiquée par l'étoile noire. b) Méthode équivalente utilisant la somme des "bols" de probabilité (voir le texte) pour localiser le PR. L'étoile noire coïncide avec la zone de probabilité élevée en rouge. c) Réincorporation de toutes les probabilités dans la "zone de décision" rouge.

Une simulation équivalente, mais légèrement plus réaliste, est présentée dans la Figure 22b. Dans ce cas, nous tenons compte de l'incertitude anticipée des valeurs de RSS mesurées, généralement considérée comme suivant une distribution gaussienne avec une variance de 2 dB. Ainsi, pour une valeur de RSS donnée, un PR produit une distribution de densité de probabilité spatiale sous la forme d'un "bol" dont le rayon est calculé à l'aide de l'équation de Friis, et dont l'"épaisseur des bords" est dictée par l'incertitude de 2 dB de la valeur de RSS, traduite à nouveau par l'équation de Friis en une distance physique. La formule pour le bol est alors donnée par :

$$\hat{f}_p(x,y) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(d-d_{rssi})^2}{2\sigma^2}} \qquad (\text{VI}-2)$$

$$d = \sqrt{\left(x - x_{ap}\right)^2 + \left(y - y_{ap}\right)^2} \qquad (\text{VI} - 3)$$

Cette approche tient compte de l'incertitude inhérente aux mesures de RSS en attribuant une distribution probabiliste à la position spatiale du client. Ainsi, au lieu d'avoir une localisation précise, nous obtenons une estimation de la position sous la forme d'une distribution de probabilité. Cela reflète la réalité où les mesures de RSS peuvent varier en raison de facteurs tels que l'atténuation du signal, les obstacles environnants, etc.

En utilisant cette approche, nous pouvons obtenir une estimation plus réaliste de la localisation du client en prenant en compte l'incertitude des mesures de RSS. Cela nous permet d'avoir une meilleure compréhension de la probabilité de la position réelle du client dans une configuration donnée du réseau WiFi.

Dans cette équation, f_p représente la fonction de densité du bol, σ est un paramètre de variance de distance dérivé de la variance de 2 dBm dans le RSS, d_{rssi} est la distance obtenue en inversant l'équation de Friis, x et y sont les coordonnées d'un point dans la carte de densité, et x_{ap} et y_{ap} sont les coordonnées de l'AP. La forme de la fonction de bol est illustrée en projection 2D et en 3D dans la figure 23. L'intégrale d'un bol provenant d'un seul PR sur le plan x-y est, bien sûr, normalisée pour obtenir une probabilité totale de 1.

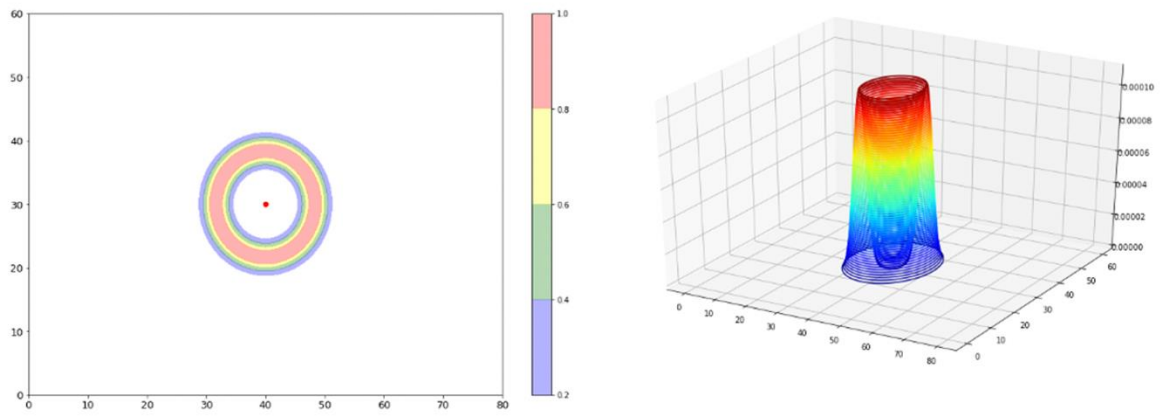


Figure 23 Illustration de la formule de la coupe en forme de bol a) en projection 2D et b) en 3D.

Pour illustrer l'utilisation de l'outil du bol, revenons à l'exemple du modèle simplifié, en superposant dans la figure 22b les densités de bol de chacun des 3 AP qui ont reçu le PR émis. Dans la figure, la somme des probabilités résultantes est codée en fonction d'une barre de couleurs. Nous remarquons qu'une zone rouge correspondant à la région où les "lèvres" des 3 bols se chevauchent est apparue. En pratique, nous sommes libres de choisir le seuil considéré comme circonscrivant au mieux cette région. Nous décidons maintenant, de manière analogue à la figure 1a, que le PR a été localisé à l'intérieur de la région rouge.

VI.C.3. L'outil Renormalisation

Pour exprimer cette décision, nous mettons en œuvre l'outil de Renormalisation, en intégrant toutes les probabilités à l'extérieur de la région de décision et en les réinjectant à l'intérieur de celle-ci. La zone sélectionnée, une fois renormalisée, se manifeste en deux dimensions dans la Figure 22c. Ainsi, nous avons délimité une petite région que nous croyons contenir un client WiFi. La probabilité intégrée au sein de cette région est de 3. Sachant que, dans ce cas, les 3 PR ont été reçus simultanément à partir d'une seule émission, nous effectuons une division par 3 afin d'obtenir le nombre de clients, C = 1. De plus, la Figure 22c présente l'étoile noire qui indique la position déterminée à partir de l'intersection des trois cercles figurant dans la Figure 22a. Les deux méthodes de localisation, à savoir la méthode d'accumulation de densité du bol et la méthode d'intersection des cercles, concordent, comme cela devrait être le cas. Cette concordance entre la mesure de la densité du bol et une localisation par triangulation constitue un élément clé de notre approche, comme cela sera détaillé dans les sections VI.D et VI.E Par souci de précision, nous pouvons qualifier la zone rouge de "zone Fiducial" pour deux raisons majeures : a) elle représente la zone avec la probabilité la plus élevée de contenir le client ; et b) elle est confirmée par la triangulation. Dans la VI.D, nous aborderons la notion d'une zone fiducial plus généralisée. Il convient de noter incidemment que nous aurions pu envisager de multiplier les probabilités des trois AP au lieu de les additionner. Cependant, dans la pratique, ce choix s'avère problématique lorsqu'il faut traiter un grand nombre d'AP, certains pouvant être hors de portée les uns des autres.

Dans la Figure 24, nous présentons un nouvel exemple illustrant certaines des difficultés rencontrées dans les réseaux WiFi extérieurs du monde réel. Nous considérons un groupe de 11

clients dont les positions sont réparties de manière aléatoire sur une petite région à l'intérieur du triangle formé par les AP. Nous allons réaliser une expérience sur une fenêtre de temps pendant laquelle un appareil client moyen émettrait un seul PR. Implicitement, nous faisons ici référence à une probabilité moyenne d'émission de PR, disons X, moyennée sur tous les types d'appareils et les activités des clients, c'est-à-dire X = 1 pour notre modèle de base pour la fenêtre de temps choisie. Pour cet exemple, supposons que, en raison des effets de propagation, la probabilité moyenne pour un AP donné de recevoir un PR émis par un client à cette distance soit d'environ 1 sur 3, de sorte que les réceptions par un seul AP dominent notre expérience, mais que parfois 2 ou 3 AP pourraient être impliqués. Supposons en outre que, pendant que notre fenêtre de temps évolue, 5 clients émettent chacun un PR qui est reçu uniquement par l'AP le plus à gauche, et que 5 autres émettent un PR reçu uniquement par l'AP le plus à droite. En superposant les bols correspondant à ces 10 PR, nous obtenons la distribution représentée dans la Figure 24a.

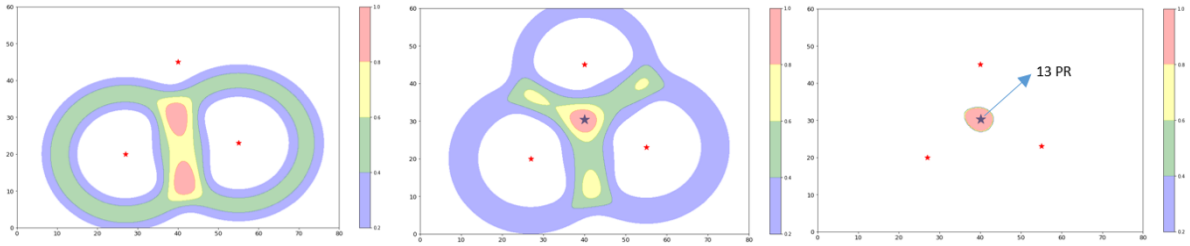


Figure 24 a) Somme des bols de probabilité à partir de 10 PR émis par 10 clients, dont 5 reçus par l'AP de gauche et 5 par l'AP de droite. Le choix de l'emplacement des clients est ambigu (deux zones rouges). b) Un 11e client émet un PR qui est reçu par les 3 AP, ce qui résout l'ambiguïté. Une étoile noire indique la position de ce PR obtenue par triangulation. Sa position correspond à celle de la zone rouge. c) Renormalisation des probabilités dans la région de décision centrale, qui a une probabilité intégrée de 13 PR (6 de l'AP de gauche, 6 de l'AP de droite et 1 de l'AP du haut).

La distribution de probabilité dans la Figure 24a, à ce stade, présente deux zones rouges. Bien que les clients aient des positions légèrement différentes et que le bruit dans leurs valeurs de RSS génère des rayons de "bol" légèrement différents, les amplitudes de ces deux pics sont presque identiques, ce qui rend impossible de choisir celui correspondant au regroupement de

clients. Supposons cependant qu'avant la fin de la fenêtre expérimentale, le onzième client émette un PR qui, par hasard, soit capté simultanément par les trois AP. En ajoutant les trois "bols" correspondant à ces nouvelles réceptions, nous obtenons maintenant la distribution illustrée dans la Figure 24b. Il est clair d'après la figure que le PR à 3 AP, bien qu'étant une occurrence minoritaire, est suffisant pour lever l'ambiguïté concernant le pic à choisir pour le regroupement de clients. La Figure 24b montre également l'étoile noire obtenue en effectuant la triangulation à partir des 3 PR reçus simultanément. Sa position confirme la décision prise en utilisant la méthode de densité, comme prévu. À titre de précaution, il est clair qu'il existe une multitude de configurations différentes des 11 clients qui pourraient avoir produit les distributions présentées dans la Figure 24. Notre hypothèse est que, dans la limite de grands nombres de clients, comme discuté dans la VI.D, et étant donné la tendance naturelle des clients à se regrouper plutôt qu'à être organisés en cercles, en patchs symétriques ou répartis uniformément sur toute la surface d'un site, l'interprétation choisie est la plus plausible.

VI.C.4. Outil de Comptage

Bien que le groupe de 11 clients ait été localisé dans l'exemple, il n'a pas encore été tenté de les compter. Pour ce faire, l'outil de Normalisation peut être utilisé pour obtenir la Figure 24c, où la somme des probabilités de toutes les PRs reçues a été réinjectée dans la zone rouge sélectionnée. La probabilité intégrée dans la zone est alors, par construction, de 13. Si on utilise la probabilité d'émission moyenne pour la fenêtre temporelle, X = 1, mentionnée ci-dessus, le nombre estimé de clients est C_{est} = 13/X = 13. Comme le nombre réel de clients est C = 11, on pourrait dire que l'estimation est déjà assez bonne. En réalité, il y a deux choix possibles. Comme première possibilité, si l'on est certain que la probabilité des PRs provenant de trois points d'accès est faible par rapport à celle des demandes provenant d'un seul point d'accès, on peut choisir de simplement tolérer l'erreur de comptage qu'elles introduisent en échange de la propriété de résolution d'ambiguïté que fournissent les PRs provenant de trois points d'accès. Une deuxième possibilité consisterait à mesurer la fraction de PRs provenant de trois points d'accès dans les données et à l'utiliser d'une manière ou d'une autre pour corriger X afin d'obtenir une estimation plus précise. L'outil de comptage et la possibilité d'introduire une telle correction de X seront abordés plus en détail dans la Section VI.D et la Section VI.E

VI.C.5. Outil de Filtrage

Avant de passer aux applications dans la section suivante, il est important de discuter d'une configuration supplémentaire simple mais problématique fréquemment rencontrée dans les vrais jeux de données WiFi extérieurs. Dans certains réseaux, certains points d'accès peuvent être situés à l'intérieur des bâtiments ou dans d'autres zones favorisant une proximité très étroite avec le point d'accès, quelques mètres par exemple. En raison de la géométrie fermée et/ou de la distance accrue par rapport aux autres points d'accès, la triangulation n'est généralement pas possible. Dans ces cas, le bol associé au point d'accès sera très fortement pointu en raison de son petit rayon. Tant que le rayon n'est pas trop grand, les conséquences pour la localisation sont en fait minimes, car on peut simplement répertorier le nombre total de PRs reçues, presque exclusivement provenant d'un seul point d'accès, et les associer au bol étroit entourant l'emplacement du point d'accès (voir Figure 25). La difficulté avec l'utilisation de l'outil de bol ici est que le pic extrême obscurcit souvent la structure plus fine de la carte de densité des clients plus éloignés du point d'accès. Pour cette raison, il est pratique, lors de l'analyse d'un site avec l'outil de bol, d'exclure les points d'accès intérieurs, ainsi que tout PR ayant un RSS > -60 dBm, de la carte avant de poursuivre l'étude. Cette procédure est appelée l'outil de Filtrage.

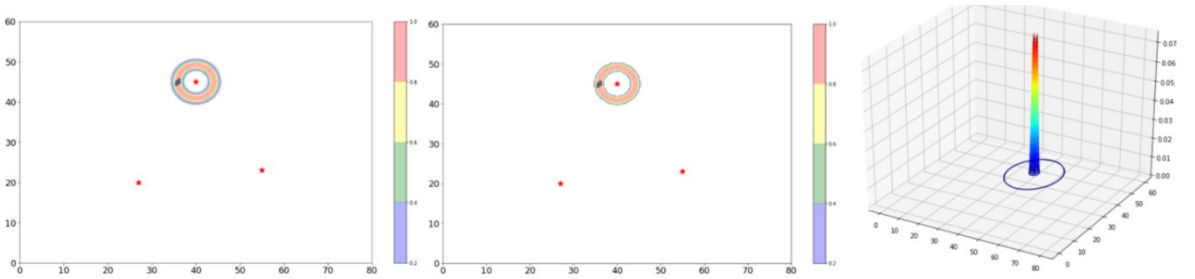


Figure 25 Un point d'accès (AP) avec des caractéristiques de propagation "intérieure". Dans a), cinq clients se trouvent à proximité immédiate de l'AP, créant un bol fortement pointu. b) Un seuil peut être appliqué pour renormaliser la densité dans une région plus compacte. c) Représentation 3D du bol fortement pointu. Le nombre total de PR intégré est enregistré et les PR sont supprimées du graphique pour rendre plus visible la structure à plus petite échelle (indiquée schématiquement par un cercle bleu dans le graphique 3D).

VI.C.1. L'outil Localisation

Lors de notre discussion sur la triangulation, nous avons supposé que les valeurs des paramètres de propagation A et n étaient connues. Cependant, dans la pratique, il est plus judicieux de mesurer réellement la valeur de A pour un site spécifique, car même si la valeur normalisée de A est généralement fixée à -30 dBm dans l'industrie, elle peut varier en fonction des caractéristiques propres à chaque environnement. De plus, la valeur de n, qui représente l'exposant de propagation, est fortement dépendante de l'environnement de propagation étudié. Utiliser des valeurs incorrectes pour A ou n peut entraîner des estimations de distance considérablement différentes et donc affecter la précision de la localisation. Malheureusement, réaliser des études de propagation précises et fiables sur site peut s'avérer complexe et coûteux, ce qui limite souvent leur disponibilité dans des scénarios du monde réel.

Cependant, il existe une façon de réduire la dépendance aux paramètres de propagation en exploitant une observation simple basée sur l'équation de Friis. Cette observation concerne la configuration d'un triangle formé par trois points d'accès (AP) : AP1, AP2 et AP3, comme illustré dans la Figure 26 pour un exemple provenant du Site 2. Les distances entre ces AP sont notées D_{12}, D_{13}, et D_{23}. Lorsqu'un client se trouve à proximité de ce triangle, voire à l'intérieur, il émet un paquet reçu (PR) qui est simultanément capté par les 3 AP avec des valeurs de puissance RSS1, RSS2 et RSS3 respectives. En inversant l'équation (VI-1) et en considérant d_1, d_2, et d_3 comme les distances entre le client et les trois AP, nous pouvons formuler :

$$\alpha_{ij} = \frac{d_i}{d_j} = 10^{\frac{RSS_i - RSS_j}{10n}} \quad \mathrm{i} \neq \mathrm{j} = 1, 2, 3. \qquad (\mathrm{VI} - 4)$$

La dépendance des ratios de distance par rapport au paramètre A a été annulée. L'équation (VI-4) nous dit que le rapport des distances entre le client et n'importe quelles deux AP est constant. Il s'avère que cela implique que, pour chaque paire ij, la position du client se trouve sur un cercle appelé "Cercle d'Apollonius" en référence au géomètre grec du 2ème siècle qui a découvert cette relation. Ce cercle a son centre sur une ligne tracée à travers les deux AP et un rayon donné par :

$$R_{ij} = \frac{\alpha_{ij}}{{\alpha_{ij}}^2 - 1} D_{ij} \quad \mathrm{i} \neq \mathrm{j} = 1, 2, 3. \qquad (\mathrm{VI} - 5)$$

Les cercles apparaissent en vert, bleu et orange sur la Figure 26. Nous tenons à souligner que ceux-ci ne sont pas les mêmes que les cercles de triangulation introduits dans la section VI.C.1.

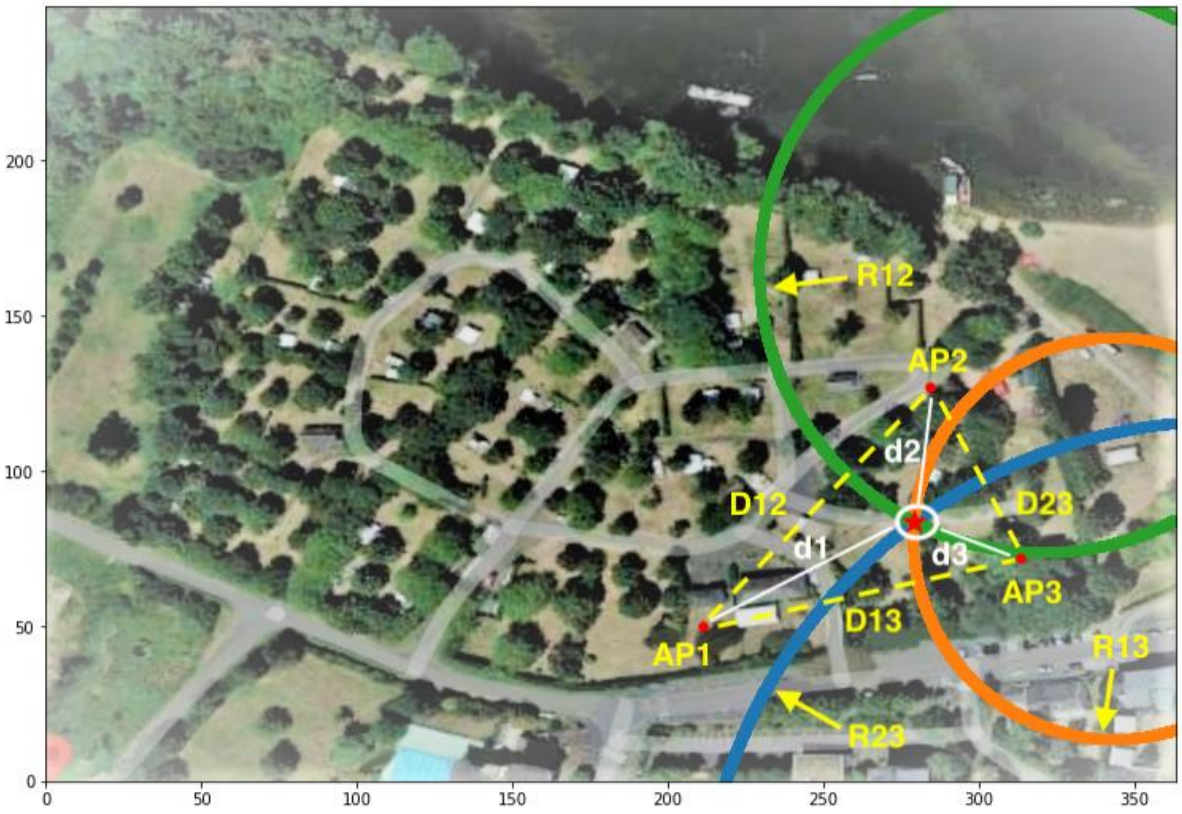


Figure 26 Les cercles d'Apollonius (vert, bleu, orange) pour les AP1, AP2, AP3, séparés par les distances D12, D13, D23. Un client situé à l'intérieur du cercle blanc se trouve à des distances d1, d2, d3 des trois AP. Les rayons des cercles sont R12, R13, R23. La solution pour la localisation du client déterminée à partir des valeurs de RSS aux trois AP est indiquée par l'étoile rouge à l'intersection des cercles.

Si nous considérons le scénario où la relation de Friis est une expression exacte plutôt qu'un modèle, et supposons qu'il n'y ait aucun bruit dans les valeurs de RSS, ainsi que des triangles AP qui ne sont pas excessivement obtus, une observation intéressante se présente. Dans cette situation idéale, n'importe quels deux cercles d'Apollonius se croiseraient en deux emplacements potentiels pour le client. Cependant, il est généralement possible de rejeter l'un de ces emplacements comme étant peu probable, ce qui nous laisse avec un candidat plus plausible.

Dans des scénarios plus réalistes, où des solutions exactes peuvent ne pas exister, les solveurs numériques viennent à notre secours. Ces solveurs peuvent être utilisés pour trouver la meilleure solution non exacte, souvent en utilisant des techniques d'optimisation des moindres carrés. Bien que non exactes, ces solutions peuvent néanmoins offrir un haut niveau de précision. L'utilisation des trois cercles d'Apollonius au lieu de seulement deux peut également renforcer la robustesse du processus de localisation dans de tels cas, en ajoutant une couche supplémentaire de redondance. Cependant, il est crucial de souligner que, dans la pratique, selon le site étudié, le taux de localisations avec des solutions d'Apollonius médiocres ou inexistantes peut atteindre plusieurs dizaines de pour cent.

L'importance de la solution de localisation d'Apollonius réside non seulement dans son indépendance vis-à-vis du paramètre A, mais également dans sa faible dépendance à l'égard de l'exposant de propagation n pour la plage de valeurs $2 < n < 3$. Une analyse simple révèle que, dans cette plage, le rayon d'un cercle d'Apollonius varie seulement d'environ 16 % de la distance Dij entre deux AP, ce qui est du même ordre de grandeur que la résolution intrinsèque obtenue dans les réseaux WiFi extérieurs. Cette relative indépendance des positions des clients d'Apollonius vis-à-vis des paramètres de propagation nous ouvre la voie à leur utilisation pour calibrer les paramètres A et n. Cela représente une proposition très intéressante dans les situations réelles courantes où les mesures de propagation fiables ne sont pas disponibles.

Dans cette optique, il est possible d'utiliser les solutions d'Apollonius pour effectuer une calibration des paramètres A et n. Cette approche devient pertinente dans les cas où les mesures de référence de propagation font défaut. Une fois les solutions d'Apollonius de bonne qualité identifiées, elles peuvent servir à établir une relation entre les niveaux de RSS et les distances réelles. Une illustration concrète est fournie par la Figure 27, qui présente un graphique de RSS en fonction du logarithme de la distance pour l'ensemble de données du Site 1. Les solutions d'Apollonius considérées comme fiables ont été utilisées pour estimer la distribution des valeurs, qui s'avère bien correspondre à la formule de Friis avec des valeurs calibrées de A = 36,2 dBm et n = 2,24. Ces valeurs calibrées peuvent ensuite être utilisées pour ajuster la formule de Friis spécifiquement pour le site étudié. Cette approche permet de pallier l'absence de mesures de propagation fiables et offre une méthode pratique pour calibrer les paramètres de propagation dans les environnements réels.

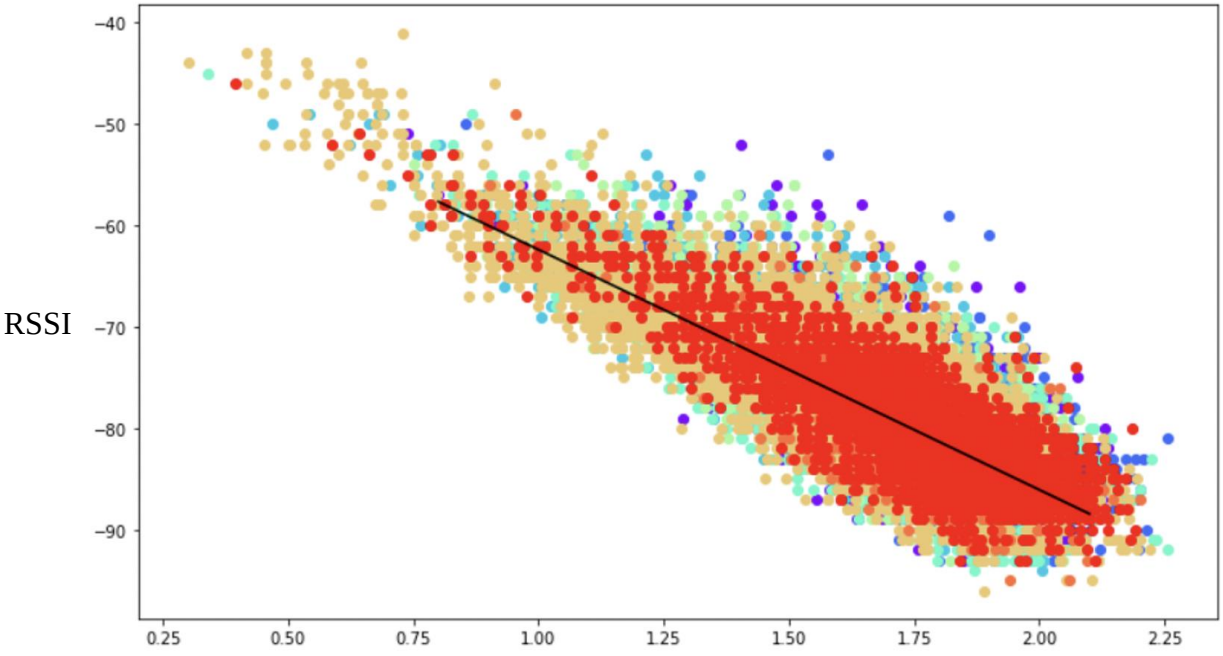


Figure 27 Graphique du RSS en dBm en fonction du logarithme de la distance en mètres. Mesure des paramètres de propagation A et n pour l'ensemble de données du Site 1 en utilisant les solutions d'Apollonius comme positions des clients. Les points de différentes couleurs correspondent aux différents AP du site. Un ajustement par moindres carrés à la formule de Friis sur l'ensemble des AP donne A = 36.2 dBm et n = 2.42 pour le Site 1. Des résultats similaires sont obtenus pour le Site 2.

VI.D. Application aux jeux de données réels

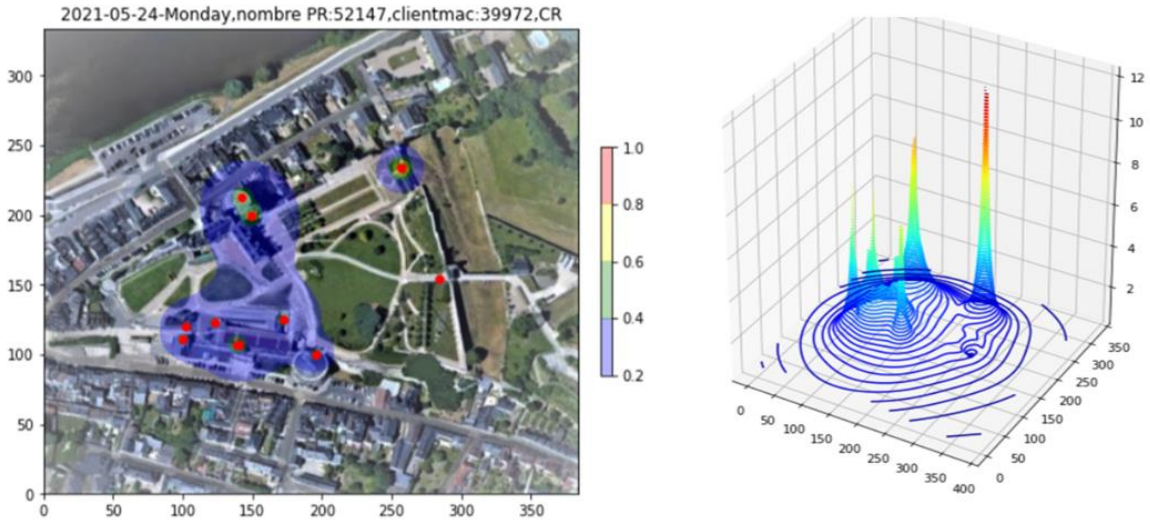
Dans cette section, les outils présentés sont appliqués à des exemples tirés de jeux de données WiFi réels. De plus, les outils Fiducial et Drill-Down sont présentés, ce qui permet une interprétation plus détaillée et quantitative du processus de réduction des données par rapport à ce qui a été présenté jusqu'à présent. Pour chaque site, une carte de densité des PRs sur une période de trois heures est sélectionnée, suffisante pour garantir un nombre suffisamment élevé de PRs accumulées tout en permettant une analyse temporelle de l'activité des clients. Les

exemples choisis sont typiques, mais ils illustrent également certaines configurations particulières.

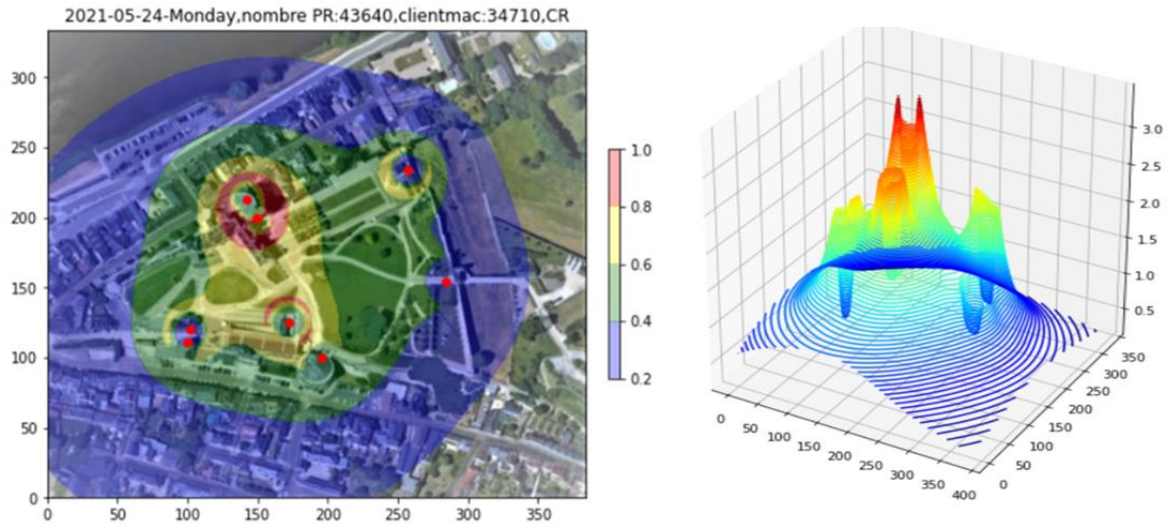
VI.D.1. Scénario complet de réduction des données appliqué à Site 1

VI.D.1.1. Application de l'outil de Filtrage

La densité brute des PRs sur la période de trois heures de l'après-midi, de 12 h à 15 h, est présentée dans la Figure 28, à la fois en 2D et en 3D. À partir du graphique 3D, il est clair que les pics prononcés causés par les clients situés à quelques mètres d'un point d'accès obscurcissent les données à plus petite échelle. L'effet est si prononcé que le graphique 2D semble presque vide. Pour résoudre ce problème, le nombre total de PRs provenant des points d'accès intérieurs, ainsi que toutes les PRs ayant un RSS > -60 dBm, est enregistré, puis ces PRs sont supprimées du graphique (étape de Filtrage). La densité après filtrage apparaît dans la Figure 28b), où la structure sous-jacente de la densité sur l'ensemble du site commence à devenir apparente.



(a)



(b)

Figure 28 Graphique de densité de PR du Site 1. a) Données brutes. b) Après filtrage des AP intérieurs et des PR à RSS élevé.

VI.D.1.2. L'outil Fiducial

Les régions rouges dans la Figure 28b correspondent à des probabilités plus élevées basées sur le "bowl" de la présence de clients. En principe, nous pourrions appliquer un seuil et renormaliser les probabilités, comme dans la Section 4.3, mais notre problème est maintenant plus complexe. Il y a 8 AP et plus de 43 000 PR représentés dans le graphique. De plus, il reste encore certaines structures "pic" apparentes. Quel est le meilleur seuil à appliquer ici ? Afin de répondre avec plus de confiance, nous complétons les informations de densité du "bowl" avec une confirmation indépendante basée sur des localisations basées sur les PR reçus simultanément à partir de 3 AP. Bien que ceux-ci ne représentent qu'une fraction du total des PR, et bien que l'efficacité de la localisation d'Apollonius réussie, comme discuté dans la Section VI.E.6, soit loin d'atteindre 100%, s'il y a suffisamment de cas de PR à partir de 3 AP, leurs localisations spatiales seront utiles pour confirmer les informations fournies par la densité elle-même. En utilisant les localisations basées sur les PR reçus simultanément à partir de 3 AP,

nous pouvons obtenir une confirmation indépendante des informations de densité fournies par le "bowl". Bien que ces localisations ne représentent qu'une partie des PR totales et qu'il puisse y avoir des erreurs dans la localisation d'Apollonius, si nous avons un nombre suffisant de cas de PR à partir de 3 AP, nous pouvons utiliser ces informations pour confirmer ou valider les données de densité. Cela nous permet d'avoir une approche plus robuste pour déterminer les seuils optimaux à appliquer et pour identifier les zones où la présence de clients est plus certaine.

La Figure 29a montre le graphique de densité de la Figure 28b avec les localisations réussies d'Apollonius superposées sous forme d'étoiles noires. Il y en a environ 700, dont certaines semblent se trouver en dehors du site couvert. Nous souhaitons caractériser la région occupée par ces localisations afin de pouvoir l'utiliser comme entrée pour l'Outil de Renormalisation. Afin de sélectionner une région qui représente la majeure partie des localisations tout en éliminant les valeurs aberrantes, nous utilisons l'Ensemble de Couches Convexes de la distribution des points de localisation. La première couche de cet ensemble est l'Enveloppe Convexe ; la deuxième est l'enveloppe convexe des points restants après avoir supprimé la première enveloppe, et ainsi de suite. Étant donné qu'il y a de nombreuses localisations ici et un nombre significatif de valeurs aberrantes en raison de solutions d'Apollonius de mauvaise qualité, nous choisissons la cinquième Couche Convexe pour délimiter les points, comme le montre la Figure 29 b. Cette utilisation des localisations d'Apollonius et des Couches Convexes pour caractériser l'étendue spatiale réelle de la densité des clients constitue notre Outil Fiducial.

En utilisant les localisations réussies d'Apollonius et les Couches Convexes, nous pouvons délimiter la région qui représente la majeure partie des localisations et éliminer les valeurs aberrantes. Cela nous permet d'obtenir une représentation plus précise de l'étendue spatiale réelle de la densité des clients. Cette information est essentielle pour utiliser l'Outil de Renormalisation, car elle nous permet de sélectionner une zone significative pour réinjecter les probabilités dans le processus de renormalisation.

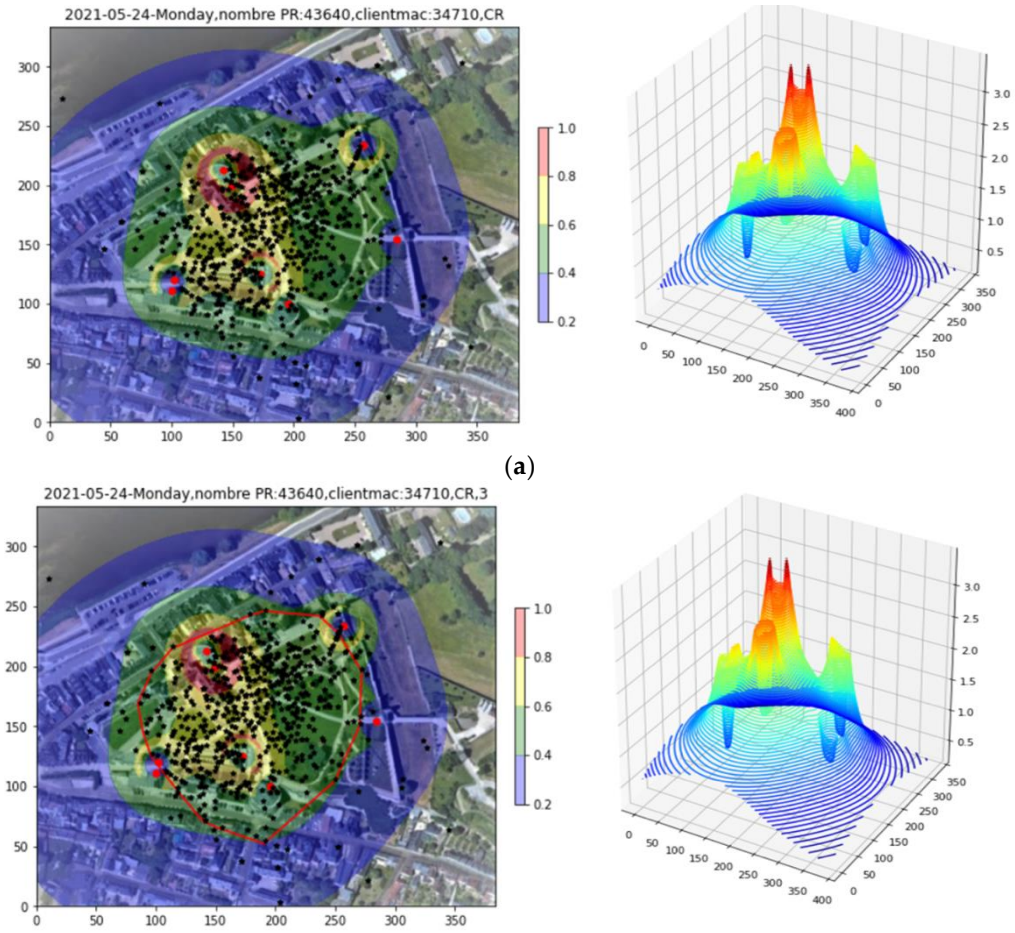


Figure 29 a) Les localisations réussies d'Apollonius sont ajoutées au graphique de densité sous forme d'étoiles noires. Ensuite, les couches convexes sont testées comme régions Fiducial potentielles. b) La cinquième coquille convexe est la plus appropriée pour éliminer les valeurs aberrantes tout en conservant la concentration principale de points.

VI.D.1.3. Appliquer l'outil de renormalisation

Avec la région Fiducial désormais définie, nous renormalisons toutes les probabilités dans la région. Le résultat est présenté dans la Figure 30, à la fois en 2D et en 3D. Toutes les densités

en dehors de la région Fiducial ont disparu, ce qui nous permet de nous concentrer sur la région où les clients se trouvent réellement. À l'intérieur de la région Fiducial, dans le graphique en 2D, nous remarquons deux régions de probabilité accrue, visibles sous forme de pics dans le graphique en 3D. Cependant, ces améliorations basées sur les "bols" ne sont pas confirmées par une densité accrue de points d'Apollonius en dessous d'elles. En effet, la nature cylindrique des pics en 3D indique une réception de PR principalement par des AP individuels.

Cela suggère qu'il existe des zones spécifiques dans la région Fiducial où la probabilité de présence de clients est plus élevée, mais cette augmentation de probabilité n'est pas corroborée par un nombre accru de localisations d'Apollonius. Il semble que ces zones soient principalement influencées par la réception de PR provenant d'AP individuels plutôt que de PR simultanés provenant de plusieurs AP.

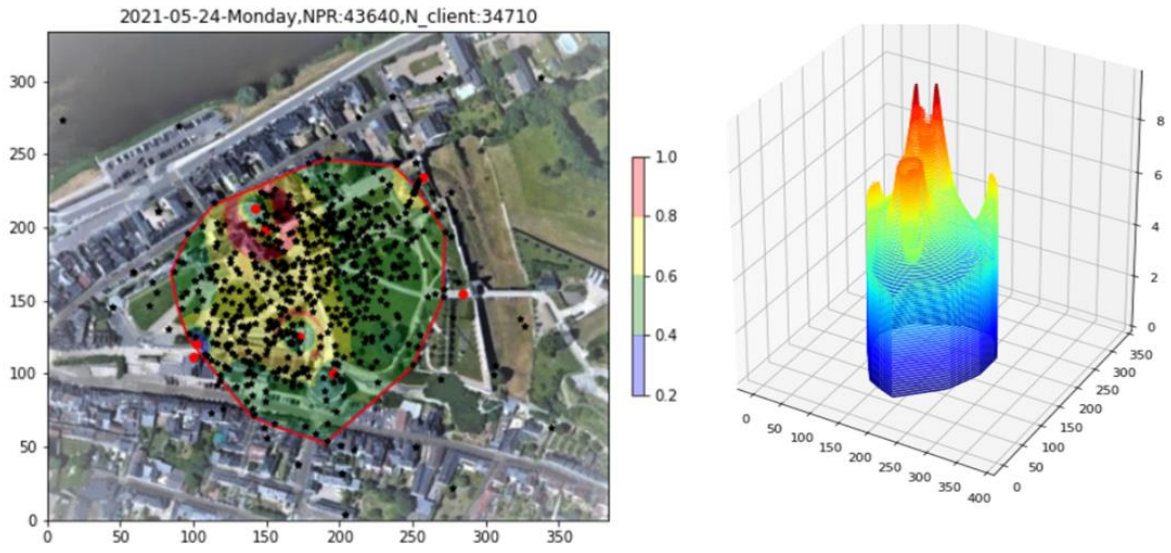


Figure 30 Renormalisation des probabilités à l'intérieur de la région Fiducial.

VI.D.1.4. L’outil de Drill-down

Afin de minimiser l'effet de ces pics et d'étudier plus en détail le comportement du graphique de densité à plus petite échelle, nous diminuons progressivement un seuil - facilement contrôlé, par exemple, avec la molette d'une souris - qui permet de creuser jusqu'à un niveau où la

structure plus fine de la densité devient apparente. Ainsi, en abaissant progressivement un seuil de manière contrôlée, nous parvenons à étudier la structure plus fine de la densité de PR sans être perturbé par les pics causés par certains clients proches des AP. Les PR situés au-dessus du seuil sont temporairement écartés, mais conservés pour une réincorporation ultérieure. Le résultat obtenu, voir la figure 31, est une densité en forme de bol presque uniforme, qui est en accord avec la distribution uniforme des points de localisation. Il est intéressant de noter que malgré la diminution du seuil, nous conservons encore 90% du total des PR, ce qui garantit une représentativité statistique élevée.

Avec l'achèvement de l'analyse de l'exemple du site 1, il ne reste plus qu'à assembler les différentes parties pour créer une carte cohérente. Cette étape sera abordée dans la Section 6, après la présentation de l'exemple du Site 2. En résumé, en ajustant le seuil progressivement, nous parvenons à étudier la structure fine de la densité de PR et à éliminer les effets des pics. Cela permet une analyse plus détaillée de la densité à plus petite échelle et garantit une représentation cohérente des données.

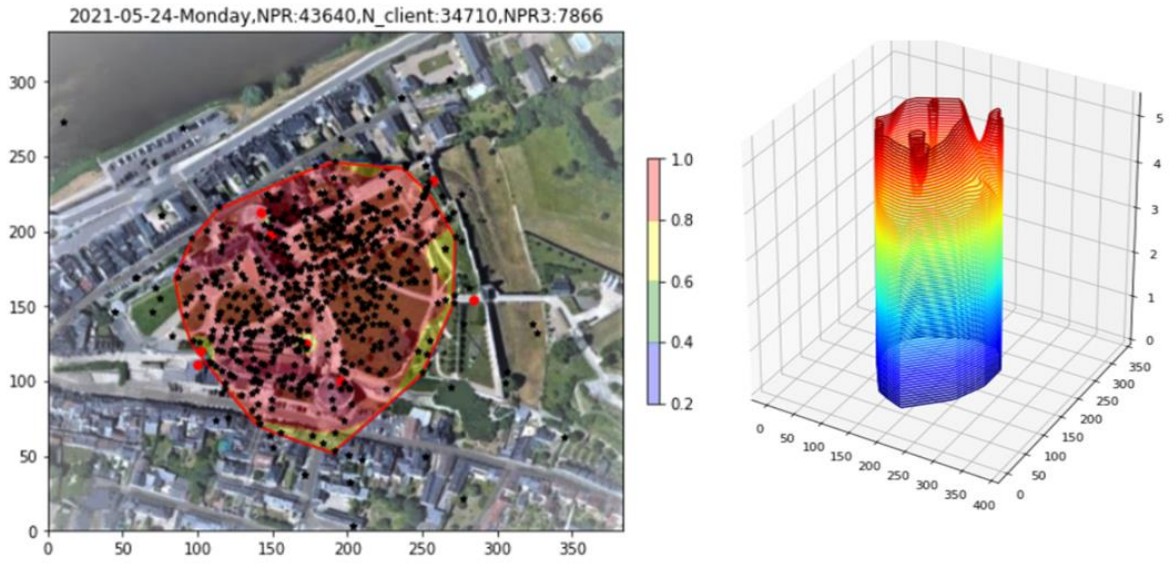


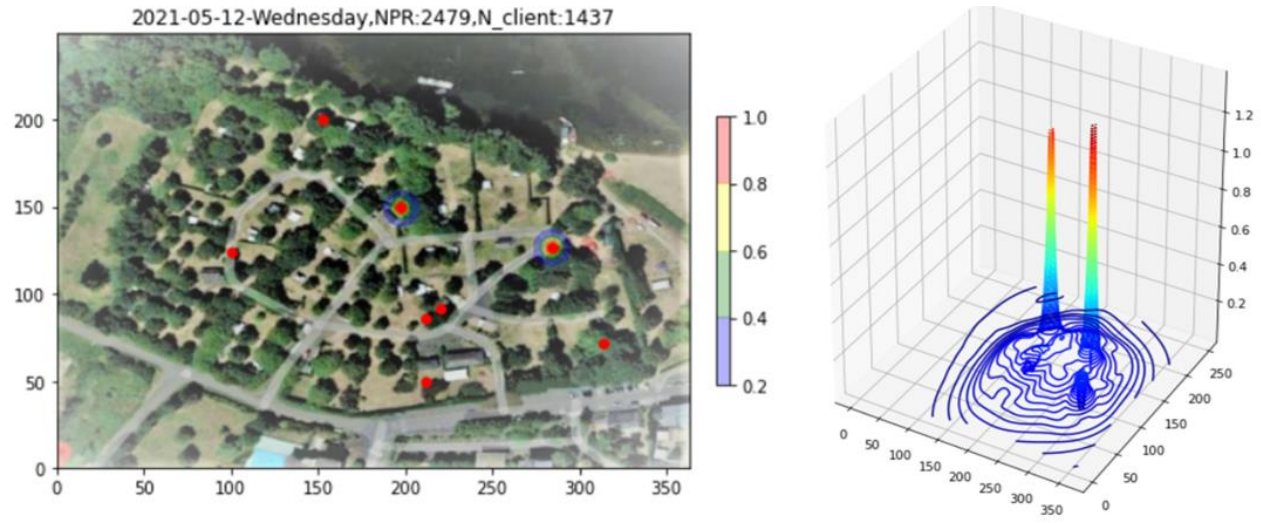
Figure 31 Drilling down du Site 1 pour révéler la structure de densité à une échelle plus fine. La distribution uniforme de densité est corroborée par la distribution uniforme des localisations (étoiles noires). Le seuil appliqué conserve encore 90 % du total des PR. Les PRs supprimés par le seuillage seront réintégrés ultérieurement.

VI.D.2. Application complète du scénario de réduction de données à l'exemple du Site 2

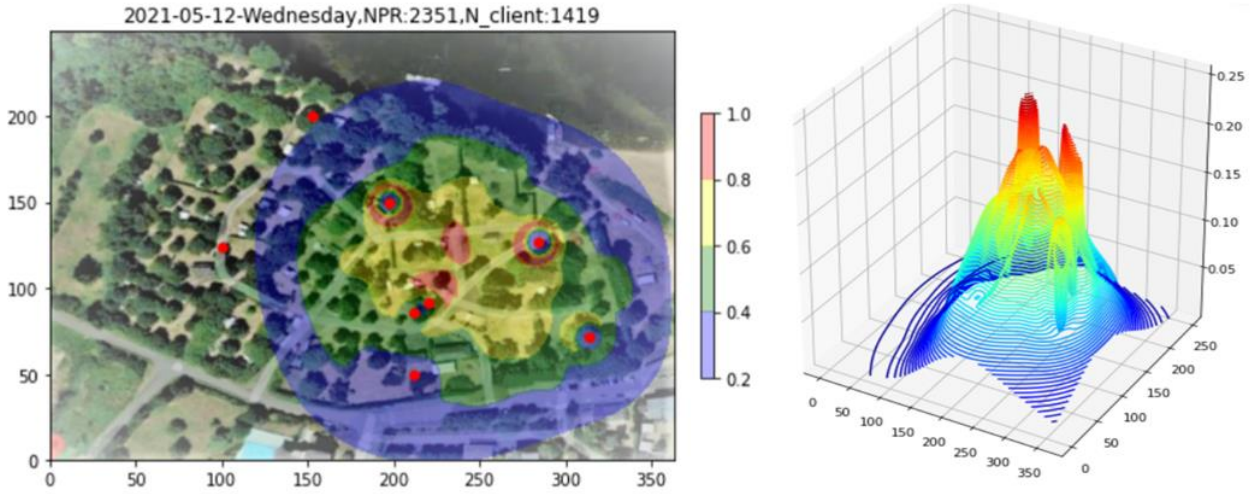
VI.D.2.1. Application de l'outil de filtrage

Dans le panneau a) de la Figure 32, nous pouvons observer que la période de trois heures étudiée ici, qui correspond à la période du soir de 18h00 à 21h00, présente un nombre beaucoup moins élevé de PR par rapport à l'exemple du Site 1. En effet, nous avons seulement 2479 PR, soit près d'un vingtième de la quantité de PR dans l'exemple précédent. De plus, nous constatons que le graphique est principalement composé de pics étroits, au nombre de deux dans ce cas, qui sont dus exclusivement aux PR ayant une puissance RSS supérieure à -60 dBm.

Dans le panneau b) de la Figure 32, nous utilisons l'outil de filtrage pour éliminer ces PR et ainsi mettre en évidence la structure plus fine de la densité. En retirant ces PR, nous sommes en mesure d'observer les détails plus subtils de la répartition des clients sur le site.



(a)

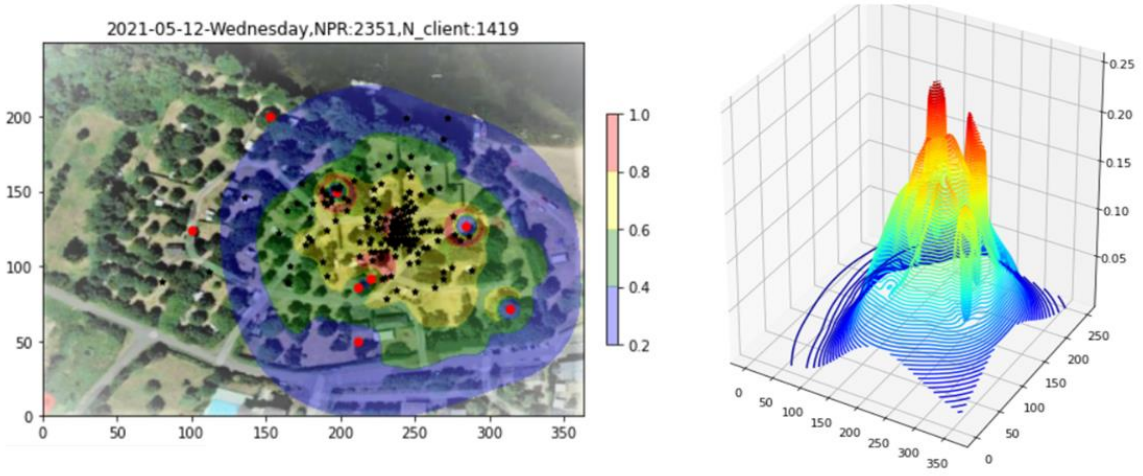


(b)

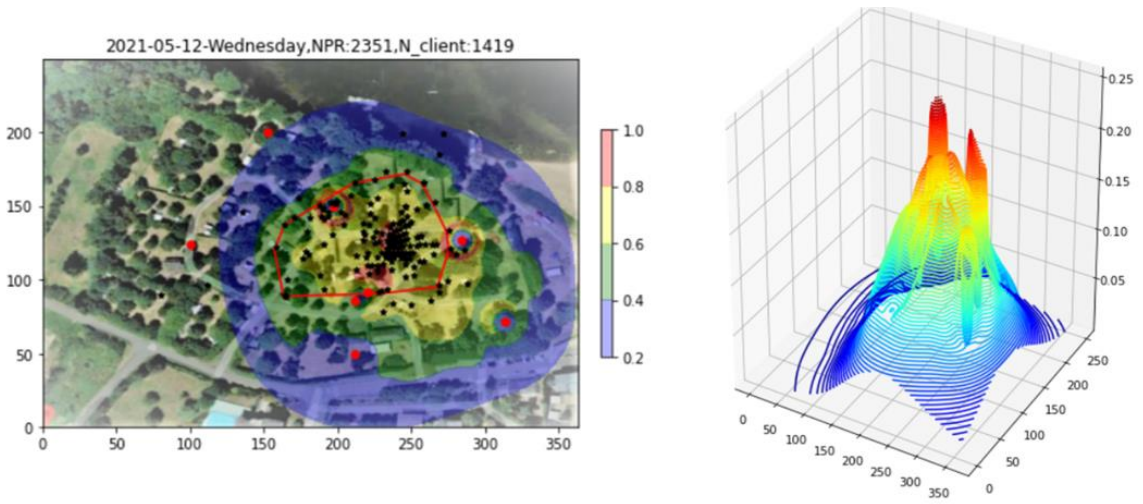
Figure 32 Application de l'outil de filtrage à la représentation de densité du Site 2. Dans a), deux pics dus aux PRs à RSS élevé dominent la représentation. En les supprimant, b), la structure plus fine de la densité devient apparente.

VI.D.2.2. L’outil Fiducial

Encore une fois, nous souhaitons restreindre notre étude à une zone Fiducial susceptible de représenter la région où la majorité des clients sont réellement situés. La Figure 33 illustre le graphique de densité du Site 2 avec les points d'Apollonius, au nombre d'environ 200, superposés. Dans cet exemple, avec un nombre de PR beaucoup plus faible, il est suffisant d'utiliser la troisième Couche Convexe pour délimiter la zone d'intérêt et exclure les valeurs aberrantes, comme le montre le panneau b) de la figure 33.



(a)



(b)

Figure 33 a) Représentation de densité du Site 2 avec les points d'Apollonius superposés. b) Le troisième Convex Layer est appliqué pour délimiter la région fiduciale.

VI.D.2.3. Appliquer l'outil de Renormalisation

Dans la Figure 34, l'outil de Renormalisation est appliqué à l'exemple du Site 2. Trois caractéristiques clés dans les graphiques de forme de bol résultants méritent d'être notées, peut-être plus facilement visibles sous forme de petites zones rouges dans la représentation 2D. La première est un pic cylindrique associé à un seul AP dans la partie centrale supérieure gauche

de la région fiduciaire. Les deuxième et troisième sont deux petites zones rouges, près du centre et du centre inférieur de la région fiduciaire, qui coïncident avec des densités améliorées de points de localisation - contrairement aux résultats du Site 1, où la densité d'Apollonius était relativement constante dans toute la région fiduciaire.

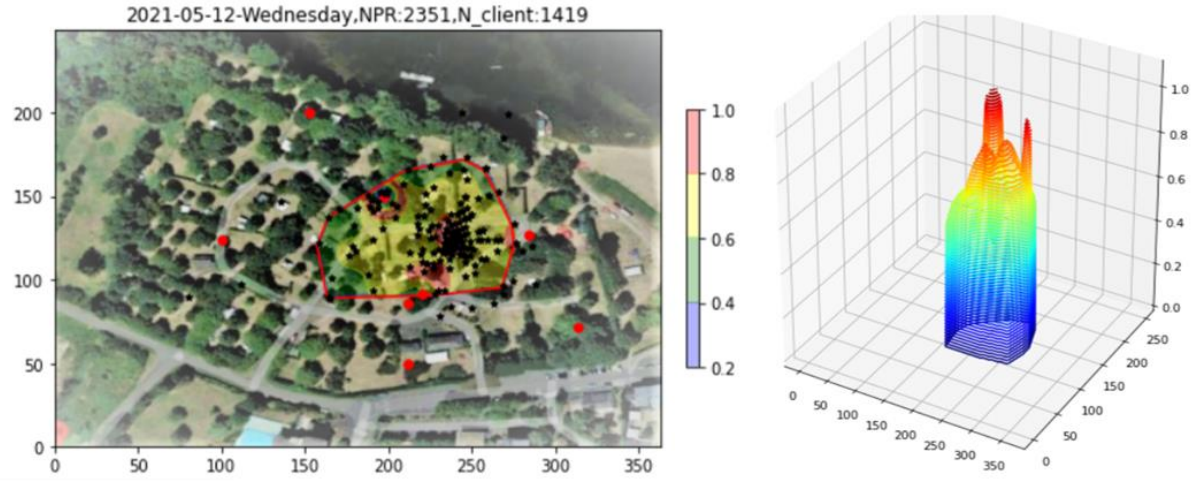


Figure 34 Renormalisation de la densité du Site 2 dans la région fiduciale. Les caractéristiques prédominantes sont un pic associé à un seul point d'accès (AP), situé dans le coin supérieur gauche de la région fiduciale, ainsi que deux autres petites régions au centre et en bas du centre qui coïncident avec une densité accrue de localisations d'Apollonius.

VI.D.2.4. l'outil Drill-Down

Pour approfondir l'étude de la relation entre la densité basée sur les "bowls" et la densité des localisations d'Apollonius pour ce site, nous allons maintenant mettre en œuvre l'outil Drill-Down. Le résultat est illustré dans la Figure 35. Nous notons que le processus de drill-down réalisé conserve 99,7% du nombre total de PR ; son effet se traduit par une légère réduction du pic le plus élevé dans le graphique. Cependant, avec ce seuil, dans le graphique en 2D, les densités d'Apollonius les plus élevées sont désormais contenues dans une zone rouge de densité accrue, tandis que les densités d'Apollonius plus faibles se situent dans la partie jaune de la zone Fiducial. Cette observation nous alerte sur une possible source d'erreur si nous utilisions un seul facteur de proportionnalité X pour convertir les PR en clients, car la région rouge contient un pourcentage plus élevé de clients dont les PR sont détectés par 3 AP. Afin de prendre en compte

cet effet, nous introduisons, dans la Section VI.E, certaines améliorations à l'outil de comptage, avant de détailler l'outil de reconstitution, qui permet d'obtenir des cartes finales du site calibrées en clients plutôt qu'en PR.

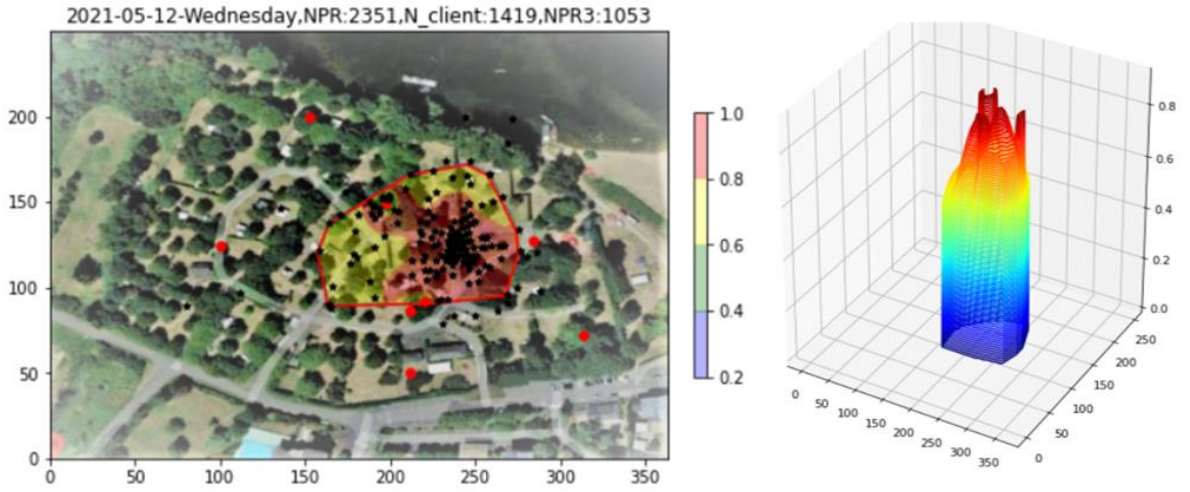


Figure 35 Application de l'outil Drill-Down à la densité renormalisée du Site 2. Nous remarquons une zone rouge qui coïncide avec une densité élevée de localisations, et une zone jaune qui coïncide avec une densité de localisation plus faible.

VI.E. Outil de Reconstitution

Comme indiqué précédemment, l'objectif de l'outil de reconstitution est de rassembler les différentes parties de la carte de densité des PRs en une carte calibrée pour les clients plutôt que pour les PRs. L'idée selon laquelle le nombre de clients est obtenu à partir du nombre de PRs en divisant par un facteur de proportionnalité X approprié à la fenêtre temporelle en question a déjà été introduite. Cependant, afin de suivre l'activité des clients dans le temps, l'accent est mis sur des blocs de trois heures tout au long de la journée, qui peuvent nécessiter différentes valeurs de X en fonction du type d'activité (matin, après-midi, week-end, activité, repos, etc.) prédominant dans chaque bloc. Des variations de X sont également visibles, dans l'exemple du Site 2, en fonction de la position, en particulier en ce qui concerne le pourcentage de PRs détectées simultanément par trois points d'accès et pouvant donc être localisées. Afin de

reconstituer correctement les différentes parties de la carte de densité des PRs en une carte cohérente de densité des clients, il est donc nécessaire de disposer de deux éléments clés : une valeur de référence X de base pour le site étudié et une technique de correction de X en fonction du temps et de la position sur la carte de densité. Dans la prochaine section, certaines extensions de l'outil de comptage conçues pour faciliter le processus de reconstitution sont proposées.

VI.E.1. Outil de reconstitution

VI.E.1.1. Détermination d'une valeur X de référence

Comme mentionné précédemment, les données utilisées dans ce travail font partie d'un ensemble plus large qui inclut les données des ville discutées dans la section précédente. Dans ce cas, nous avons mis au point un moyen de dériver les valeurs de référence de X pour un site à partir des propriétés statistiques de ses données de PRs, en particulier les périodicités hebdomadaires dans les populations en jours ouvrables par rapport aux week-ends présentes dans les données de la ville. Pour rappel, pour nous trois villes françaises, sur deux périodes de temps distinctes en 2020 et 2021, les valeurs X mesurées quotidiennement variaient de 420 à 670 (où une barre d'erreur typique est de 10 à 15 %), avec une moyenne sur tous les sites de X = 524 ± 47, encore une fois, pour une période de 1 jour. Cependant, comme les données du présent travail ont été acquises sur un site touristique historique (Site 1) et un terrain de camping (Site 2), leurs comptages de PRs ne manifestent pas les périodicités hebdomadaires claires qui étaient la norme pour les données de la ville et, en tant que telles, ne peuvent pas être directement utilisées pour mesurer X pour les deux nouveaux sites en utilisant notre méthode statistique.

Il est cependant possible de démontrer que les valeurs de X pour les deux sites actuels, après certaines adaptations, sont cohérentes avec celles mesurées dans les données de la ville. En effet, les valeurs de X mesurées pour les villes ont été validées par rapport à plusieurs références indépendantes, y compris des comparaisons avec des mesures similaires dans la littérature, ainsi qu'un calcul basé sur les clients connectés (CC) présents dans ces données. L'idée est que puisque les adresses MAC des CC sont fixes, on peut directement compter le nombre de PRs par CC. À titre de réserve, on ne s'attend pas a priori à ce que le comportement des CC soit identique à celui des CR. En particulier, les CC utilisent souvent la connexion WiFi fixe pour

mettre en place des services étendus ou ES, tels que des réseaux Peer-to-Peer, qui peuvent produire un nombre extrêmement élevé de PRs. Cependant, en éliminant les clients considérés comme participant à ES, on s'attend à ce que les valeurs de X obtenues pour CC et CR soient relativement similaires, comme cela a été le cas pour les villes.

Pour les données des villes, les valeurs de X estimées pour CC ont été obtenues en utilisant la relation $X_{CC} = P/(A_{CC}<t_{CC}>)$, où X_{CC} est la valeur estimée de X pour les CC, P est le nombre de PRs, A_{CC} est le nombre de CC, et $<t_{CC}>$ est leur durée moyenne de séjour sur le site pendant la fenêtre de temps choisie. Pour les données villes, les valeurs X_{CC} résultantes, allant de 378 à 948, sont en accord raisonnable avec la plage de valeurs X (villes) estimées pour CR mentionnées dans le paragraphe précédent, à savoir de 420 à 670. On peut dire que pour les villes, les CC ont été utilisés pour valider les valeurs de X obtenues à partir de CR dans les villes. Pour valider les données CR des Sites 1 et 2 du présent article, en utilisant les CC du présent article, on procède de la même manière. Les résultats sont donnés dans les colonnes 1 et 2 du Tableau 3.

Tableau 4 Résultats de validation de X en utilisant CC, pour le présent article, évalué sur une période de 1 jour pour permettre une comparaison directe avec [19] ,Colonne 1 : pourcentage de CC conservés après la coupure ES (voir texte). Colonne 2 : valeur de référence X calculée en utilisant $P/(A_{CC} <t_{CC}>)$ comme discuté dans le texte. La valeur pour le Site 1 est en accord avec les résultats pour les villes dans [19], mais celle pour le Site 2 est trop élevée et présente une grande variance (voir discussion dans le texte). Colonne 3 : plage approximative des valeurs, ou P_3/P dans les données CR des deux sites. Colonne 4 : valeur moyenne de P_3/P dans les données CR des deux sites. Colonne 5 : correction de la valeur de la colonne 2 en utilisant la formule dérivée dans le texte à partir de P_3 et P. Colonne 6 : valeurs de X_{12} prédites directement comparables aux valeurs X des données de la ville dans [19]. Après correction, les valeurs moyennes de la colonne 6 sont en accord avec les données de la ville dans [19]. La grande variance de la valeur du Site 2 est discutée dans le texte.

Site	% de CC après ES-Cut	X_{base} (pour CC)	Range $\frac{P_3}{P}$ (pour CR)	$<\frac{P_3}{P}>$ (pour CR)	Correction $1-\frac{2}{3}<\frac{P_3}{P}>$	X_{12} prédit (pour CC)
1-château	99.8%	369 ± 36	0.0–0.15	0.07	0.953	352 ± 34
2-camping	92%	1139 ± 394	0.2–0.7	0.35	0.767	873 ± 302

La première colonne du tableau donne le pourcentage de CC conservés après la suppression des ES. Pour le site 1, très peu de clients avaient des taux d'émission de PR anormalement élevés, tandis que pour le site 2, le pourcentage était beaucoup plus élevé. Cela est supposé être dû aux caractéristiques différentes des deux sites. Dans un site touristique historique comme le Site 1, la plupart des clients ne restent que peu de temps et passent la majeure partie de leur temps à se déplacer d'un point d'intérêt à l'autre, ce qui les incite peu à se connecter au WiFi local pour des situations particulières. En revanche, le Site 2 est un camping qui comprend des bungalows. Les clients y séjournent souvent plusieurs jours dans des quartiers qui peuvent ne pas disposer de l'environnement sans fil et des outils auxquels ils sont habitués dans leur domicile habituel, ce qui les incite à utiliser le WiFi local avec son potentiel de services à fort débit de PR. De plus, sur les deux sites, les valeurs OUI des CC avec un nombre élevé de PR, lorsqu'elles sont exploitables, indiquent souvent des dispositifs de type IoT plutôt que des téléphones utilisateurs. Enfin, il convient de souligner que le nombre total de CC est une fraction réduite de celui des CR, ce qui le rend moins exploitable statistiquement pour une étude détaillée, un point qui sera important dans les discussions à venir.

La deuxième colonne du tableau donne les valeurs moyennes de référence X des CC dans le présent travail, calculées à l'aide de la formule présentée précédemment, $X_{CC} = P/(A_{CC}<t_{CC}>)$, ainsi que leur écart absolu moyen (EAM). On constate que la valeur moyenne de 369 pour le Site 1 correspond bien à la plage des valeurs X basées sur CR (420-670) ainsi qu'aux valeurs X_{CC} (378-948) obtenues pour les villes. En ce qui concerne le Site 2, la situation

est quelque peu différente. La valeur X obtenue à partir des CC ici, 1139, est assez élevée par rapport aux autres. De plus, son MAD de 394 semble curieusement élevé par rapport à la plage normalement rencontrée de 10 à 15 %, mentionnée précédemment. La section suivante montrera que ces différences résultent du comportement de P_3, le nombre de PR reçus par trois AP, dans les présentes données par rapport aux données des villes.

VI.E.1.2. Un Strawman: Xbase dépendant du site et du temps

Comme mentionné dans la section 3, la classe CR pour ces données a été définie en exigeant au maximum trois observations d'une adresse MAC hachée dans la fenêtre de temps sélectionnée. Ce seuil garantit que la plupart des adresses MAC CR sont effectivement aléatoires, c'est-à-dire qu'elles ne conservent pas la même adresse MAC lors de plusieurs émissions de PR, mais il permet toujours le cas où le PR d'un client CR est observé sur trois AP, ce qui permet de localiser ce PR par triangulation. En revanche, pour les données urbaines analysées, un seuil de 2 a été choisi. Bien que ce choix garantisse également une séparation nette entre CR et CC, il élimine la possibilité de triangulation. En effet, il a été constaté empiriquement que les PR détectés sur trois AP étaient très rares dans les données urbaines : seulement environ 2 % des PR ont été détectés deux fois au cours d'une journée, et une proportion supplémentaire minime sur trois AP. Cela contraste fortement avec ce qui est observé pour les Sites 1 et 2 du présent travail, où les PR détectés sur trois AP peuvent constituer une part importante du total des AP. En effet, s'il n'y avait pas au moins une fraction de CR détectée simultanément sur trois AP, il serait considérablement plus difficile de parler de localisation sur les deux sites. Cette différence fondamentale entre les données urbaines et celles des Sites 1 et 2 peut être attribuée à la disposition géométrique des AP dans les deux cas. Dans les données urbaines, les réseaux déployés consistaient invariablement en un nombre limité d'AP espacés largement et disposés selon des configurations principalement linéaires, comme nous l’avons présentés dans le chapitre sur les données, et donc peu adaptées à la triangulation.

Afin de mettre les données urbaines et les données des Sites 1 et 2 sur un pied d'égalité pour les comparer, ainsi que de prendre en compte les valeurs de P_3 plus élevées sur les sites 1 et 2, nous proposons la procédure suivante « Strawman », à titre d'approximation de premier ordre. On peut supposer que la valeur X rapportée pour les villes, que nous appellerons désormais X_{12},

est valable pour les sites où les PR sur plusieurs AP sont rares, puis essayer de corriger cette valeur pour les sites ayant une disposition plus dense des AP et donc plus de PR sur trois AP. Nous posons les hypothèses suivantes :

$$P \sim P_{12} + P_3 \sim C_{12}X_{12} + 3C_3X_{12} = (C + 2C_3)X_{12*} \quad (\text{VI} - 6)$$

$$X_{base} = \frac{P}{C} = (1 + \frac{2C_3}{C})X_{12} = (1 + \frac{2\frac{P_3}{3X_{12}}}{C})X_{12} = (1 + \frac{2\frac{P_3}{3X_{12}}}{\frac{P}{X_{base}}})X_{12} \quad (\text{VI} - 7)$$

$$X_{base} = X_{12} + \frac{2}{3}\frac{P_3}{P}X_{base} \quad (\text{VI} - 8)$$

$$X_{base} = \frac{X_{12}}{1 - \frac{2}{3}\frac{P_3}{P}} \quad (\text{VI} - 9)$$

où P est le nombre total de PR, C est le nombre total de clients, P_3 est le nombre de PR détectés simultanément par trois AP, C_{12}est le nombre de clients produisant un ou deux PR par émission, et C_3 est le nombre de clients dont les PR ont été détectés sur trois AP. Il convient de noter que dans la dernière ligne de l'équation (VI-9), X_{base}, la valeur de référence corrigée pour le site, est exprimée entièrement en termes de quantités mesurables P et P_3, ainsi que X_{12} provenant des données urbaines, supposé également valable pour les sites étudiés ici. Pour $P_3 = 0$, nous retrouvons $X_{base} = X_{12}$, comme requis, tandis que pour $P_3 =$ P, le modèle prévoit $X_{base} = 3X_{12}$, ce qui est raisonnable compte tenu des hypothèses de départ.

On peut se faire une idée de l'applicabilité du modèle en se référant à nouveau au Tableau 4. La colonne 3 présente la plage de valeurs du rapport P_3/P pour les Sites 1 et 2. Il est évident que les nombres absolus de CC sur les sites 1 et 2 sont trop faibles pour fournir une mesure statistiquement stable de P_3/P. Pour remédier à cela, dans les colonnes 3, 4 et 5 du tableau, les plages et les valeurs moyennes de P_3/P mesurées pour les CR sont citées, qui devraient être similaires à celles des CC puisque P_3/P est une quantité qui dépend principalement de la

disposition du réseau. Les colonnes 3 et 4 du tableau montrent que la fraction P_3/P est en effet significative et qu'elle est beaucoup plus élevée sur le site 2 tant en magnitude qu'en plage. Cette observation est cohérente avec la valeur MAD beaucoup plus élevée sur le Site 2, c'est-à-dire que la fraction P_3/P varie de manière significative tout au long de la journée, entraînant de grandes variations autour de la valeur moyenne. Dans la colonne 5 du tableau, cette valeur moyenne de P_3/P est utilisée pour calculer le réciproque de la fraction du modèle, créant ainsi un facteur de correction ad hoc qui permet de mettre les valeurs quotidiennes de X des sites 1 et 2 sur un pied d'égalité avec celles des données urbaines, pour la comparaison. Dans la colonne 6, on constate que les valeurs résultantes sont effectivement cohérentes avec les plages mentionnées précédemment, c'est-à-dire 420-670 pour les valeurs mesurées en ville et 378-948 pour les valeurs de validation basées sur les CC.

Il convient de noter que le modèle prescrit l'utilisation de la valeur instantanée de P_3/P, et non sa moyenne, pour corriger X_{12}. La colonne 6 du tableau vise à montrer la tendance générale de la correction. Lorsqu'il est réellement utilisé dans les outils de comptage et de reconstruction, les valeurs instantanées de P_3/P dérivées des CR doivent être utilisées. C'est également pour cette raison que, par la suite, nous préférons utiliser X_{12}= 524 à partir des données urbaines comme facteur de conversion canonique, plutôt que les valeurs corrigées ad hoc dérivées des CC rapportées dans la colonne 6 du tableau 1. Il est clair que P et P_3 sont des quantités spécifiques à chaque site, de sorte que X_{base} s'adapte automatiquement à chaque site. De plus, dans le traitement proposé, la fenêtre de temps utilisée n'est pas spécifiée explicitement. En effet, la formule résultante pour X_{base} est valable pour P et P_3 de n'importe quelle fenêtre de temps. En particulier, elle est directement appliquée aux fenêtres de 3 heures choisies pour cet article, de sorte que X_{base} s'adapte également à chaque fenêtre de temps individuelle. Il convient de noter que la valeur constante X_{12} dans les équations ci-dessus doit être ajustée pour correspondre à la fenêtre de temps de 3 heures choisie pour les tracés de densité de clients dans le présent travail, c'est-à-dire X_{12}= 524 × (8/24) = 65.

En créant le modèle, certaines approximations simplificatrices ont été faites. Il est clair qu'un traitement plus rigoureux est possible, par exemple en incluant explicitement un terme pour P_2 ou P_4, etc. Cependant, compte tenu de l'incertitude statistique initiale de X_{12} et des incertitudes systématiques liées à son application à de nouveaux ensembles de données, cela peut ne pas

être la priorité majeure à ce stade. Le modèle fournit une estimation informée de premier ordre des résultats qui peuvent être attendus pour estimer les densités numériques des clients à partir des PR brutes dans les réseaux WiFi extérieurs pour lesquels la vérité terrain n'est pas disponible. Nous l'utiliserons dans l'outil de reconstitution, le cas échéant, pour estimer les décomptes globaux de clients en fonction de la position.

VI.E.1.3. X_{base} adapté à la position

Dans la section VI.D.1.4, nous avons identifié des zones rouges et jaunes dans la densité en forme de bol qui coïncident avec différentes densités de localisations d'Apollonius, c'est-à-dire une augmentation du nombre de localisations dans la zone rouge et un déficit relatif dans la zone jaune. Au cours de cette discussion, il a été suggéré que la densité en forme de bol dans la zone rouge, qui compte les PR, peut avoir été artificiellement augmentée en raison de l'abondance de PR détectés sur trois AP. En raison de la manière dont la densité en forme de bol est construite, il n'est pas possible de connaître les nombres de PR d'un seul AP et de trois AP dans les zones rouges et jaunes. Cependant, en guise de substitut, il est facile de compter le nombre de points de localisation dans les deux zones. Il est proposé d'utiliser cette information pour corriger X_{base}, zone de couleur par zone de couleur, lors de la reconstitution des différentes parties de chaque site, comme décrit dans la section VI.E.2ci-dessous pour les deux sites étudiés. La correction dépendante de la position à X_{base} peut être appliquée selon la formule suivante :

$$X_{color} = \frac{X_{12}}{1 - \frac{2}{3}\frac{P_3'}{P}} \quad P_3' = \frac{N_{Apollo}(color\ zone)}{N_{Apollo}} P_3 \quad (\text{VI} - 10)$$

où N_{Apollo}(zone de couleur) est le nombre de localisations d'Apollonius dans chaque zone de couleur de la densité en forme de bol, et N_{Apollo} est le nombre total de telles localisations.

Une difficulté avec cette correction telle qu'elle est présentée est que, bien qu'elle adapte X aux différentes régions de couleur, elle modifiera le nombre total estimé de clients. Ce problème peut être contourné en normalisant les valeurs $X_{couleur}$ avec la formule suivante :

$$\langle \frac{1}{X_{color}} \rangle = \frac{1}{P} \sum_i (P_i \cdot \frac{1}{X_i}) \qquad \text{(VI} - 11)$$

$$X_{color} \leftarrow X_{color} \cdot X_{base} \cdot \langle \frac{1}{X_{color}} \rangle \qquad \text{(VI} - 12)$$

Un exemple, avec les valeurs obtenues pour le Site 1 dans la Figure 35, est présenté dans le Tableau 4.

Tableau 5 Répartition de X_{base} par zone de couleur et nombre correspondant de clients pour le site 2 pour le seuil de Drill-down de la Figure 35. Les colonnes sont %pr : pourcentage du total des PR ; Nb_pr : nombre de PR ; Nb_apollo : nombre de localisations d'Apollonius (étoiles noires) ; les trois colonnes suivantes sont Nb_client : estimation du nombre de clients en utilisant trois estimations de X ; et dans la dernière colonne, les valeurs X_{color} spécifiques à chaque couleur issues du graphique.

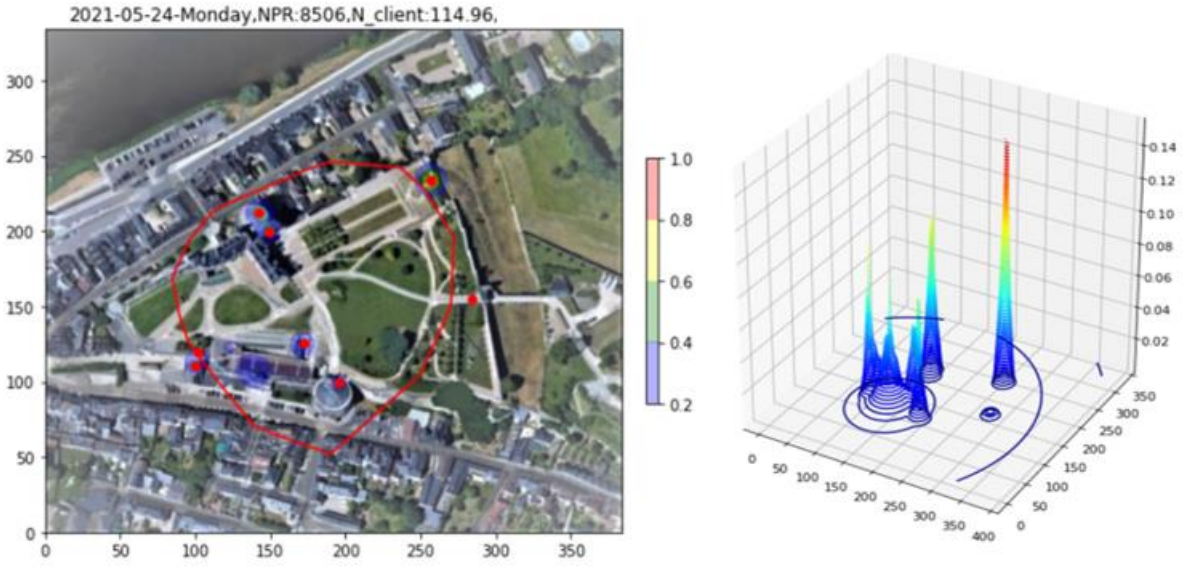
	%pr	Nb_pr	Nb_apollo	Nb_client (X_{12}=65)	Nb_client (X_{base}=93)	Nb_client (X_{color})	X_{color}
total	99.67	2343.27	163	36.05	25.2	25.2	77.49
red	59.21	1391.99	135	21.42	14.97	13.43	86.36
yellow	38.45	903.97	20	13.91	9.72	11.16	67.47
green	2.01	47.3	8	0.73	0.51	0.6	65.97
blue	0	0	0	0	0	0	65
transparent	0	0	0	0	0	0	65

La septième et dernière colonne du tableau donne les valeurs de X_{color} obtenues à partir de la Figure 35, tandis que les colonnes 4, 5 et 6 donnent respectivement le nombre estimé de clients en utilisant la valeur X fixe héritée des villes normalisée sur une période de 3 heures (65) ; une valeur X corrigée globale du site en fonction de P_3 (93) ; et une correction zone par zone de couleur (valeurs apparaissant dans la dernière colonne). Les estimations du nombre de clients varient en fonction de la méthode X utilisée, comme prévu. Par exemple, la prédiction pour la zone jaune est de 13,9 clients pour le X hérité non corrigé (colonne 4), de 9,7 clients pour une valeur X corrigée globale du site (colonne 5) et de 11,2 clients lorsque X_{color} est utilisé pour corriger la position (colonne 6). Ce qui a été fait, en utilisant X_{base} et X_{color}, est de corriger l'augmentation artificielle de la densité du bol, causée par la présence de PR à trois AP proportionnellement plus élevés.

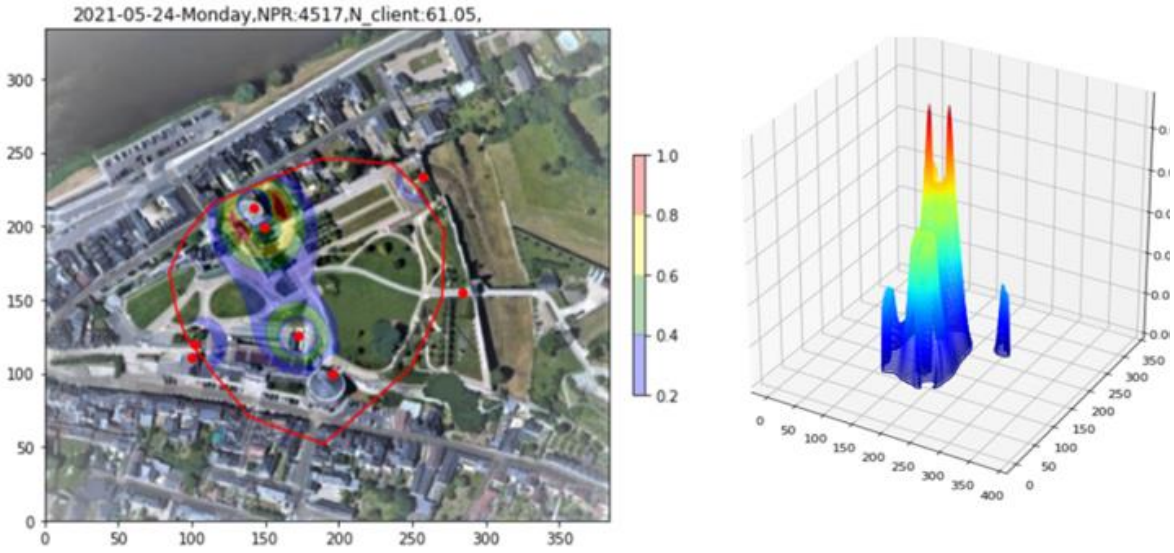
VI.E.2. Outil de reconstitution

VI.E.2.1. Reconstitution du site 1

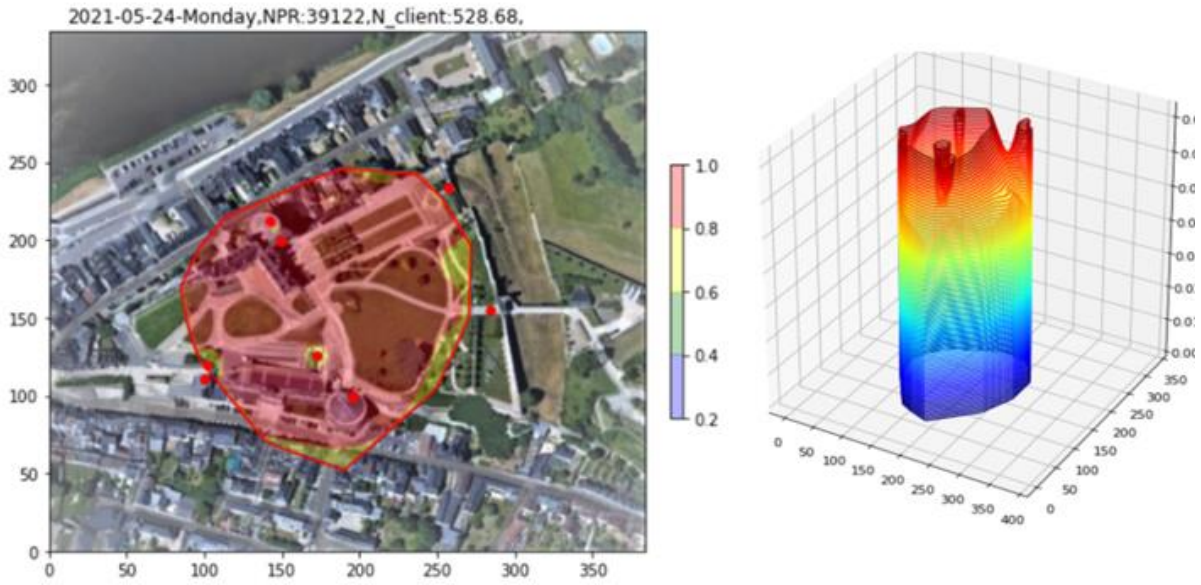
Les densités de PR du Site 1, après avoir été filtrées et soumises à la technique du Drill Down, ainsi que la densité de PR obtenue après le Drill Down, sont présentées graphiquement dans la Figure 36, récapitulant certaines des étapes que nous avons déjà vues dans la Figure 31. Comme la densité de PR après le Drill Down ne présente aucune structure particulière, il n'est pas nécessaire d'appliquer un X dépendant de la position. La densité des clients est ensuite obtenue en divisant toutes les composantes par le facteur X global du site pour la fenêtre de temps considérée, puis en sommant les résultats obtenus.


2021-05-24-Monday,NPR:8506,N_client:114.96,

(**a**)


2021-05-24-Monday,NPR:4517,N_client:61.05,

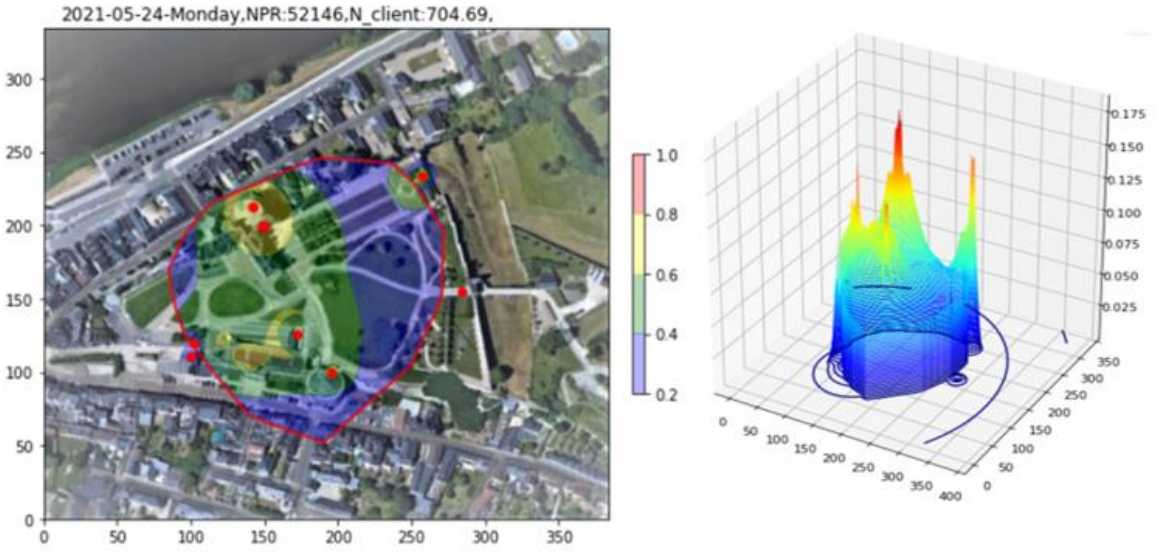
(**b**)



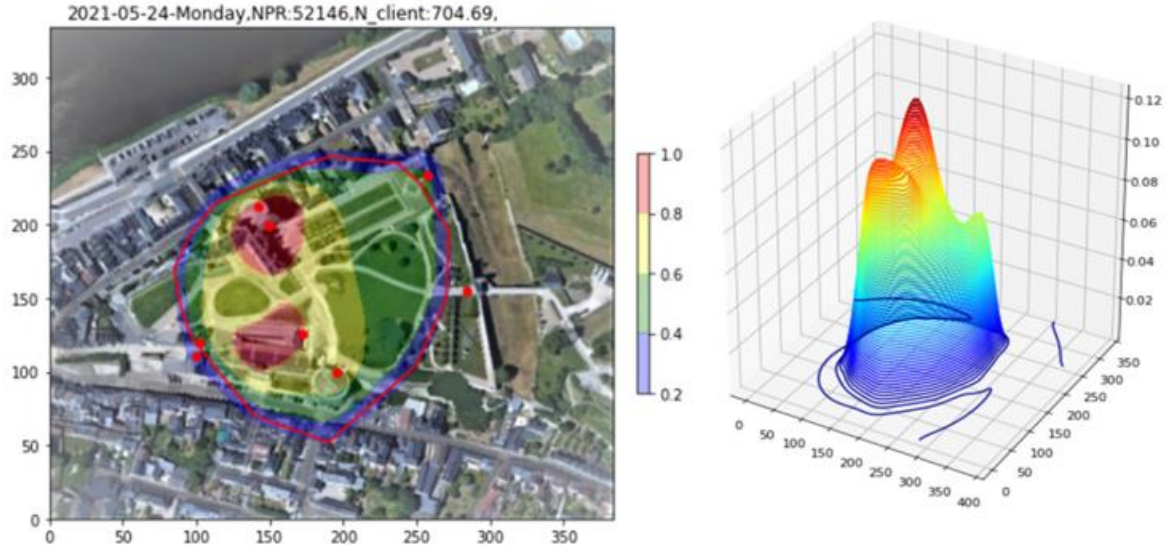
(**c**)

Figure 36 Site 1. a) APs intérieurs filtrés et PR à RSS élevé ; b) structures en pic éliminées par le Drill Down ; c) Densité conservée après le Drill Down.

La carte de densité des clients résultante est présentée dans la Figure 37 a). Afin de rendre le graphique plus lisible, nous l'avons lissé avec un noyau gaussien d'écart-type de 10 mètres, de l'ordre de la résolution de position intrinsèque, afin d'éliminer les caractéristiques à une échelle sub-résolution tels que les pics étroits et les frontières entre les régions, comme illustré dans la Figure 37 b).



(**a**)



(**b**)

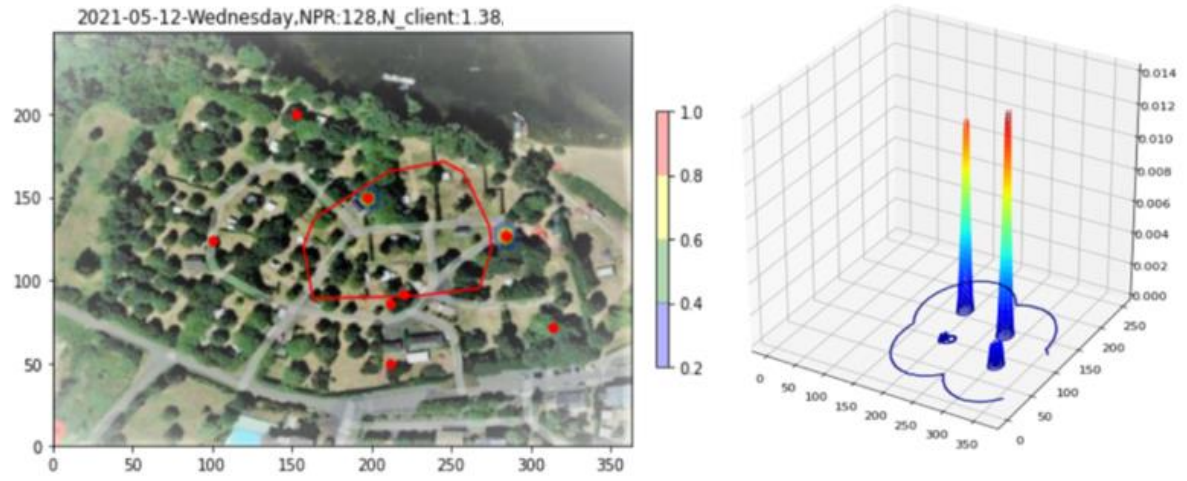
Figure 37 Site 1. a) Somme pondérée par X des différentes parties. Dans ce cas, toutes les parties sont pondérées par la valeur globale de X du site pour la fenêtre temporelle. b) Somme pondérée par X après lissage par un noyau gaussien ayant un écart-type de 10 m, afin de lisser les artefacts sub-résolution (pics, contours, ...).

Dans le graphique 2D, deux zones rouges à forte occupation sont visibles. Notre estimation pour la zone inférieure, qui correspond à la zone d'entrée du site, est de 68 clients sur une surface

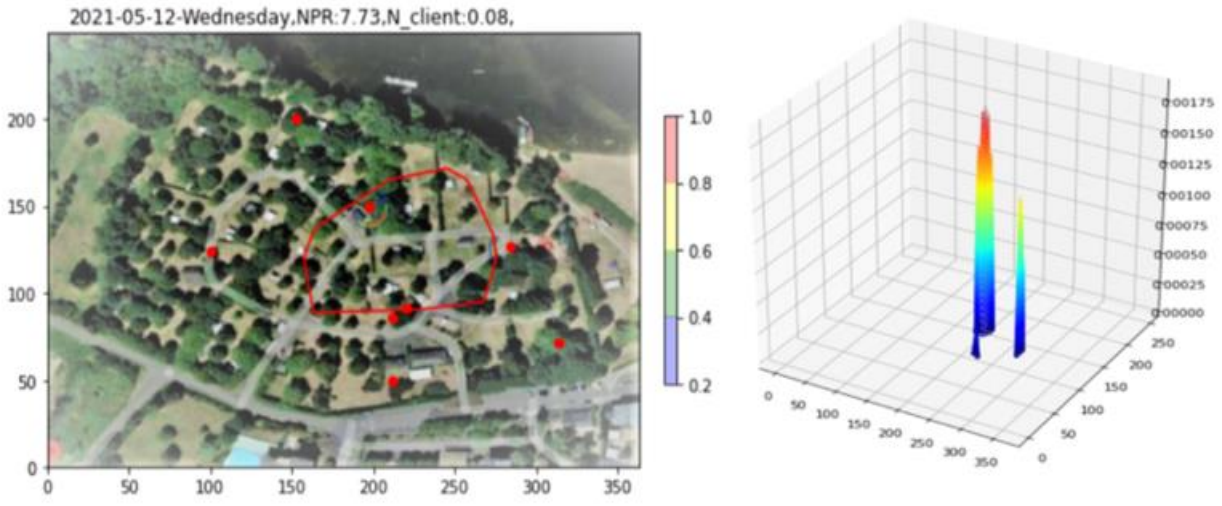
de 1990 m2. La zone supérieure, correspondant à l'une des principales attractions touristiques du site, contient, selon notre estimation, 48 clients sur une superficie de 1482 m2, tandis que pour la zone jaune environnante, nous estimons 245 clients sur une superficie de 8954 m2. Le nombre total estimé de clients dans toutes les zones du graphique s'élève à 704.

VI.E.2.2. Reconstitution du site 2

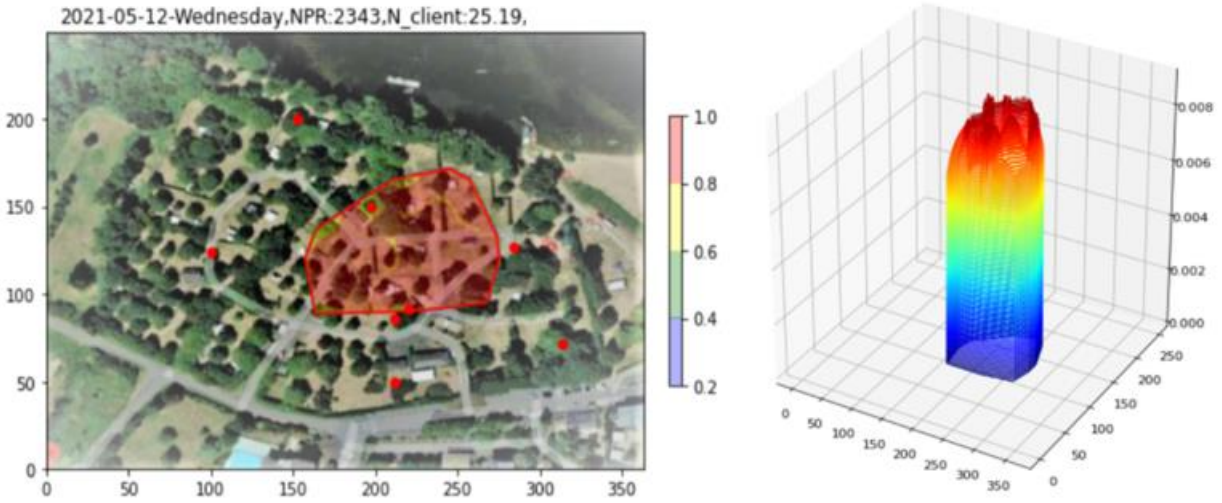
La même procédure est suivie pour le site 2, comme le montre la figure 38, à l'exception du fait que la densité après la phase de "drill-down" (c) a été corrigée par un X dépendant de la position, comme expliqué dans la section 6.1. Nous notons que la densité après la phase de "drill-down" dans la figure 38 c) est maintenant plus uniforme, comparée à celle de la figure 35, grâce à cette correction X dépendante de la densité. La densité totale obtenue par sommation est présentée dans la figure 39, avant et après lissage avec un noyau gaussien d'écart type de 10 m afin de supprimer les caractéristiques de sous-résolution telles que les pics étroits et les lignes de frontière.



(**a**)

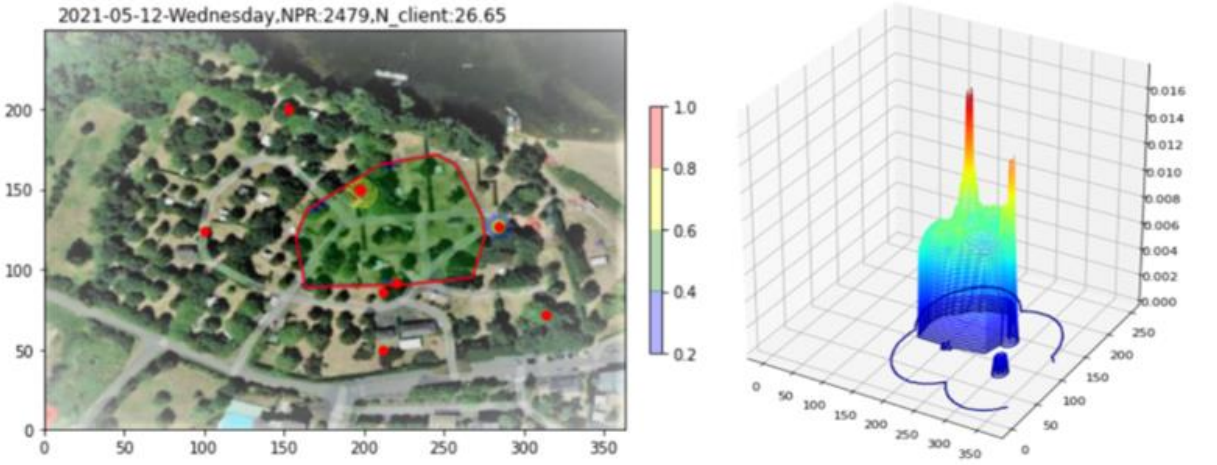


(**b**)

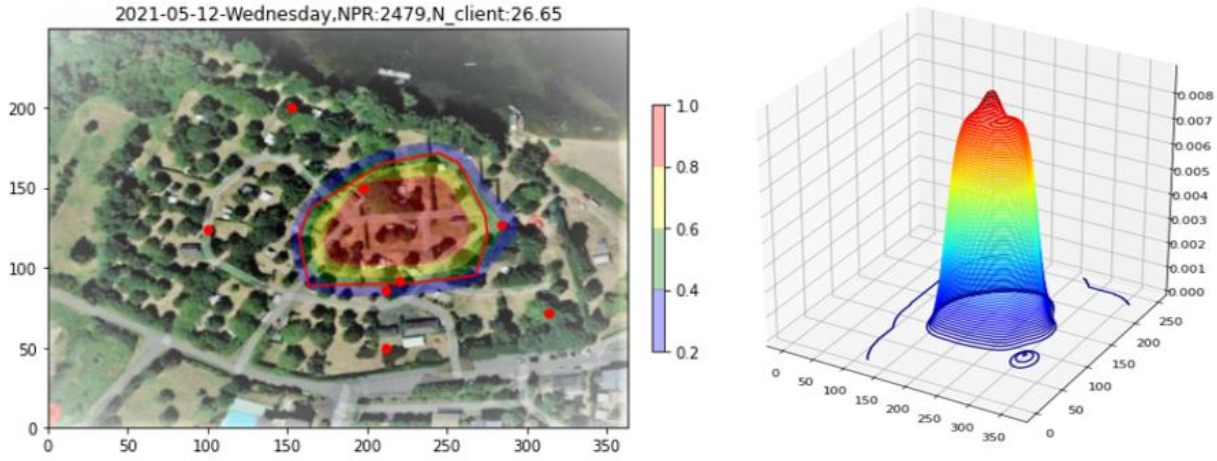


(**c**)

Figure 38 Site 2. a) Suppression des PR à RSS élevé par filtrage ; b) Élimination des structures en pic lors de la descente progressive ; c) Densités restantes après la descente progressive pondérées par X dépendant de la position et sommées.



(**a**)



(**b**)

Figure 39 Site 2. a) Somme pondérée des morceaux par X. Les morceaux sont pondérés par le X du site pour la fenêtre de temps et corrigés en fonction de la position sur la carte. b) Somme pondérée par X après lissage avec un noyau gaussien de déviation standard 10 m.

Notre technique estime, dans la zone rouge de la figure 39 a), la présence de 16 clients sur une surface de 3741 m², tandis que pour la zone environnante, il y aurait 7 clients répartis sur 2124 m². L'estimation du nombre total de clients dans l'ensemble du graphique est de 33. Un organigramme récapitulant les différentes étapes de l'application de l'outil est présenté dans la figure 40.

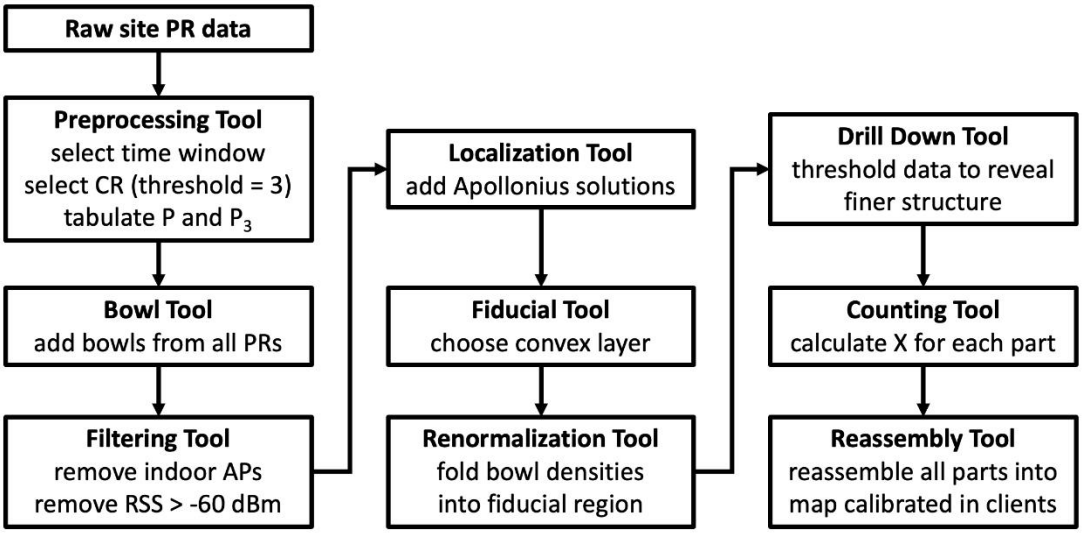


Figure 40 Organigramme résumant l'ensemble du protocole de réduction des données utilisant l'ensemble d'outils.

Comme nous l'avons vu, les outils "Bowl", de Filtrage et de Drill-down sont des techniques graphiques utilisées pour rendre plus claire la représentation spatiale de la densité des clients. L'outil de Reconstitution, quant à lui, en collaboration avec les autres outils, est un outil de comptabilité qui permet de récupérer les différentes parties dans lesquelles le problème a été divisé et de les reconstituer en un tout cohérent et facilement interprétable. C'est l'outil Fiduciaire, qui tire le meilleur parti des informations de localisation limitées disponibles dans les réseaux à ressources limitées ; l'outil de Localisation, calibré dans une procédure de démarrage ; l'outil de Comptage à plusieurs étapes permettant l'adaptation temporelle et spatiale du facteur de proportionnalité X ; et les outils de Renormalisation permettant de redistribuer les probabilités de PR dans la région fiduciaire déduite, qui incarnent la majeure partie de l'originalité et garantissent le succès de l'approche présentée.

VI.F. Conclusions

Avec l'émergence des préoccupations concernant la confidentialité des clients et la croissance rapide de l'Internet des objets (IoT), la pratique courante d'utiliser les adresses MAC des clients contenues dans les PR pour surveiller et cartographier l'activité des clients est devenue intenable, créant aujourd'hui un besoin de nouveaux outils. Ici, un ensemble de neuf outils est présenté pour transformer directement les comptages bruts de PR de MAC aléatoires provenant de réseaux WiFi extérieurs du monde réel en cartes de densité calibrées pour les clients, le tout sans utiliser de référence au sol. La technique a été appliquée à des données provenant de deux sites de réseau réels en France, avec des résultats intéressants.

À l'heure actuelle, les prédictions de densité de la technique utilisent une probabilité d'émission de PR de base, X, adoptée à partir d'une étude similaire, augmentée de quelques corrections raisonnables mais approximatives. Une perspective de haute priorité est donc de découvrir des méthodes d'estimation plus précises et de les utiliser pour évaluer notre technique sur des données actuelles. Une voie importante pour atteindre cet objectif est de favoriser les interactions avec les fournisseurs de services WiFi extérieurs et les gestionnaires de sites afin d'évaluer et d'étalonner les outils proposés ici.

En effet, le personnel qui est familiarisé avec les détails des sites surveillés est susceptible de posséder des informations expertes précieuses qui ne sont pas facilement accessibles aux scientifiques des données, munis d'un fichier Excel et d'une vue satellite Google Maps du site, chargés d'extraire les métriques nécessaires. Voici quelques exemples :

- Les responsables du réseau local conserveront probablement des enregistrements des dispositifs IoT du site, tels que les caméras, l'éclairage, les haut-parleurs, etc., utilisés régulièrement, ainsi que des équipements utilisateur attribués au personnel du site et potentiellement utilisés pour des services de P2P ou similaires. Ces informations seront précieuses pour séparer les comptages de PR correspondants de ces dispositifs de ceux dérivés des clients légitimes.

- Les gestionnaires de site posséderont également probablement des informations sur les capacités d'accueil des auditoriums, des attractions, etc. ; les capacités d'hébergement et de

restauration ; et les capacités d'accueil des attractions, des événements et du site lui-même. Ces limites peuvent être utilisées pour contraindre les estimations du facteur X local du site.

- Dans les cas où des équipements complémentaires de comptage de fréquentation sont présents à certains endroits du site, par exemple, les décomptes, les capteurs de présence, etc., leurs sorties peuvent également être utilisées pour conditionner les estimations du facteur de conversion PR-client X.

- Le personnel familiarisé avec la configuration et la maintenance du site aura également connaissance de la répartition de l'occupation des terres. Les zones inaccessibles, clôturées, restreintes d'accès ou simplement situées en dehors des limites officielles détaillées du site peuvent être exclues de la zone Fiducial lors de l'application de l'outil de renormalisation. De même, les zones d'utilisation spéciale, telles que les jardins, les chemins, les parkings, les terrains de jeux et les terrains de golf, afficheront probablement des comportements de clients propres aux usages spéciaux de ces zones et peuvent être traitées séparément des zones d'accès utilisateur plus générales lors de la prédiction des densités de clients.

- Lorsqu'ils sont disponibles, les recettes des guichets, les décomptes de billets, etc. peuvent également, dans certaines circonstances, servir de référence indirecte à laquelle les comptages de clients dérivés du WiFi peuvent être comparés.

- Enfin, les gestionnaires de site et les organisateurs auront également accès aux informations sur les horaires quotidiens, les jours fériés et les heures d'ouverture saisonnières pour l'ensemble du site ainsi que pour les installations individuelles, notamment les piscines, les restaurants, le golf, etc., et pour les événements spéciaux ainsi que les fermetures temporaires, occasionnelles ou non planifiées. Ces horaires seront précieux pour établir une cohérence entre les comptages dérivés du WiFi et les attentes basées sur la référence au sol.

La cartographie de l'activité des clients basée sur les adresses MAC aléatoires est encore une nouvelle entreprise. La simplicité et l'indépendance par rapport à la référence au sol de la trousse d'outils proposée en font une nouvelle contribution intéressante aux travaux en cours dans le domaine, tant pour les chercheurs que pour le personnel chargé de la gestion des sites étudiés.

VII. bladeRF – un retour à la dérandomisation

Les outils développés dans la section précédente s'appuient fortement sur les localisations d'Apollonius, qui, à leur tour, surviennent du fait que le MAC d'un PR vu par 3 AP produira un *hash token* identique à chacun de ces AP. Ce constat nous expose à la possibilité, lors d'une mise à jour du RGPD, que pour mieux protéger les identités et les positions des clients, ces 3 *hash tokens* aussi soient randomisés. Pour prévenir cette possibilité, nous nous voyons contraints de reconsidérer la possibilité d'un type de dérandomisation, afin de sauvegarder la possibilité de faire des triangulations.

Nous avions écarté, en début de cette thèse, la dérandomisation car la plupart des méthodes avancées réposent sur le contenu numérique de la trame d'un PR, tels que les Information Elements IE et leur répartition, des informations de séquencement, etc., qui peuvent facilement être modifiés par les constructeurs afin de confondre la dérandomisation.

On trouve pourtant dans la littérature la notion d'effectuer une dérandomisation plutôt sur les propriétes physiques, soit de la liaison, comme par exemple le CSI, soit du device lui-même, comme le Carrier Frequency Offset ou CFO [69], car ce serait difficile, voire impossible, pour un fabricant d'intervenir sur ces quantités.

La dérandomisation à base de propriétés physiques, tel que c'est proposé dans la littérature actuellement, manifeste néanmoins des difficultés importantes. D'abord, en WiFi sous OFDM, le CSI s'utilise couramment, en mode connecté, dans l'adaptation de la liaison, étant calculée avec des séquences d'apprentissage. Ces valeurs de CSI peuvent être obtenues et exploitées directement par certains NICS [70]. La difficulté principale avec une application de la CSI dans la dérandomisation des PR est que les PR, sont transmises, dans la bande prédominante à 2,4 GHz, non pas avec OFDM mais avec des trames en DSSS, qui, de plus, ne contiennent pas de séquences d'apprentissage. En outre, les PR sont envoyés par définition en mode non-connecté. Quant à la CFO, qui mesure les variations dans la fréquence de l'oscillateur local des *devices* individuels, il s'agit d'un calcul sophistique nécessitant un matériel sophistiqué et couteux, disponible seulement en OFDM et en mode connecté [69].

L'objectif de ce volet est donc de proposer une méthode de dérandomisation des PR DSSS réalisable avec du matériel WiFi standard, basé sur des CSI et CFO calculés en temps réel, en encapsulant ces grandeurs dans la trame PR elle-même via des outils de *packet capture*. Le projet se compose en 6 étapes :

1. Manipulations et études préliminaires
    a. Captage et analyse des PR DSSS avec différents outils
    b. Méthodes d'encapsulation de nouveaux champs dans une trame WiFi
2. Choix de la plateforme bladeRF
    a. Etude *hardware* (FPGA) et *software* (VHDL) de la plateforme
    b. Démodulation de trames PR en simulation (Modelsim)
    c. Familiarisation avec Commande Line Interface CLI pour captage de trames
3. Campagne de mesures
    a. Environnement *outdoor* – une vingtaine de positions
    b. Téléphones iPhone et Androïde
4. Dépouillement des données et développement d'algorithmes pour le calcul de CSI et CFO
5. Preuve de principe – démonstration que les valeurs obtenus sont exploitables
6. Application à plus grande échelle

Au moment de la rédaction de ce manuscrit, les 5 premières étapes ont été réalisées. Elles sont décrites dans ce qui suit.

VII.A. Manipulations préliminaires

VII.A.1. OpenWRT et WireShark

Nous avons entrepris des manipulations visant à récupérer les trames Wifi à partir d'un point d'accès. Pour cela, nous avons utilisé l'outil logiciel Wireshark ainsi que le HackRF SDR, une carte de radio logicielle populaire.

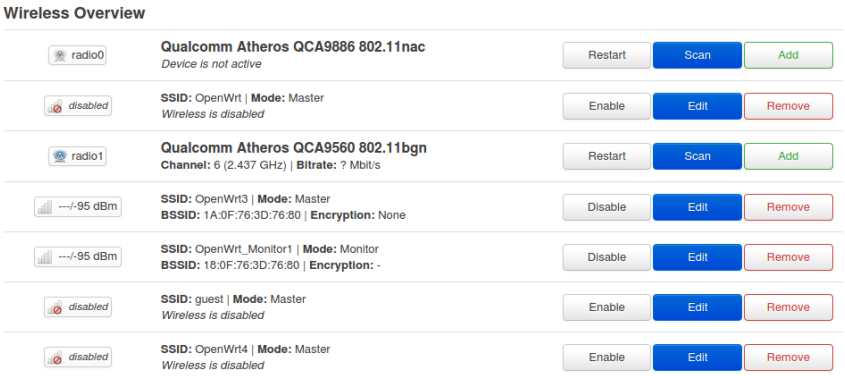


Figure 41 Interface de paramètre d'OpenWRT

Lorsque nous remplaçons le système d'exploitation (OS) d'un point d'accès Wi-Fi ordinaire par OpenWRT, nous bénéficions d'une flexibilité accrue pour configurer et contrôler les fonctionnalités du point d'accès. OpenWRT est une distribution Linux open source spécialement conçue pour les routeurs et les points d'accès Wi-Fi, offrant une multitude d'options de personnalisation et de configuration avancées.

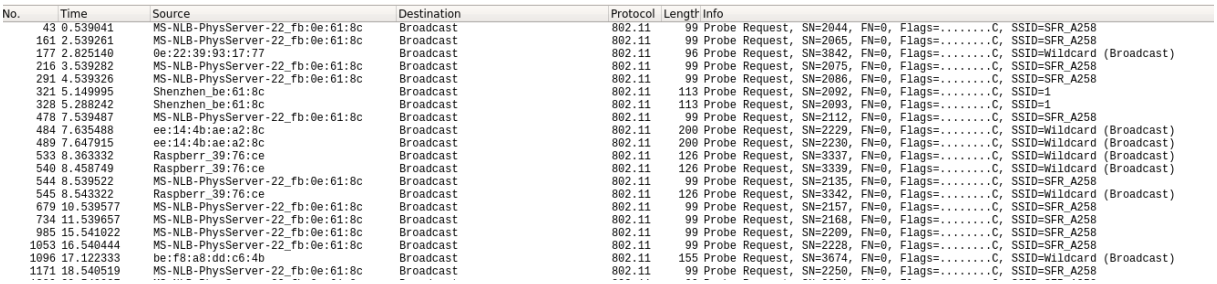
L'un des avantages majeurs d'OpenWRT est la possibilité de configurer le point d'accès en mode connecté et en mode monitor simultanément. Le mode connecté est le mode opérationnel standard dans lequel le point d'accès agit comme une passerelle sans fil, permettant aux appareils de se connecter au réseau et d'accéder à Internet. Dans ce mode, le point d'accès gère les connexions sans fil, attribue des adresses IP aux appareils, assure la sécurité des communications.

En plus du mode connecté, OpenWRT offre également la possibilité d'activer le mode monitor. Ce mode spécial permet au point d'accès de capturer et d'analyser toutes les trames Wi-Fi

échangées dans son environnement, même si elles ne sont pas destinées spécifiquement au point d'accès lui-même. En activant le mode monitor, le point d'accès peut écouter toutes les trames Wi-Fi émises par d'autres appareils à proximité, offrant ainsi une visibilité étendue sur le trafic sans fil du réseau.

Pour exploiter cette fonctionnalité de capture de trames, nous pouvons utiliser l'outil TCPDump disponible dans OpenWRT. TCPDump est un outil en ligne de commande puissant qui permet de capturer et d'analyser le trafic réseau. En utilisant TCPDump sur OpenWRT, nous pouvons enregistrer toutes les trames Wi-Fi détectées par le point d'accès, qu'il s'agisse de trames destinées au point d'accès lui-même ou à d'autres appareils du réseau.

Une fois les trames capturées par TCPDump, nous pouvons les rediriger vers un autre outil d'analyse plus convivial appelé Wireshark. Wireshark est un logiciel d'analyse de trames réseau largement utilisé qui offre une interface graphique permettant de visualiser et d'inspecter les trames en détail. En établissant un canal de communication entre OpenWRT et Wireshark, nous pouvons transférer les données capturées par TCPDump vers Wireshark pour une analyse approfondie présenté comme la figure 42.



No.	Time	Source	Destination	Protocol	Length	Info
43	0.539041	MS-NLB-PhysServer-22_fb:0e:61:8c	Broadcast	802.11	99	Probe Request, SN=2044, FN=0, Flags=........C, SSID=SFR_A258
161	2.539261	MS-NLB-PhysServer-22_fb:0e:61:8c	Broadcast	802.11	99	Probe Request, SN=2065, FN=0, Flags=........C, SSID=SFR_A258
177	2.825140	0e:22:39:93:17:77	Broadcast	802.11	96	Probe Request, SN=3842, FN=0, Flags=........C, SSID=Wildcard (Broadcast)
216	3.539282	MS-NLB-PhysServer-22_fb:0e:61:8c	Broadcast	802.11	99	Probe Request, SN=2075, FN=0, Flags=........C, SSID=SFR_A258
291	4.539326	MS-NLB-PhysServer-22_fb:0e:61:8c	Broadcast	802.11	99	Probe Request, SN=2086, FN=0, Flags=........C, SSID=SFR_A258
321	5.149995	Shenzhen_be:61:8c	Broadcast	802.11	113	Probe Request, SN=2092, FN=0, Flags=........C, SSID=1
328	5.288242	Shenzhen_be:61:8c	Broadcast	802.11	113	Probe Request, SN=2093, FN=0, Flags=........C, SSID=1
478	7.539487	MS-NLB-PhysServer-22_fb:0e:61:8c	Broadcast	802.11	99	Probe Request, SN=2112, FN=0, Flags=........C, SSID=SFR_A258
484	7.635488	ee:14:4b:ae:a2:8c	Broadcast	802.11	200	Probe Request, SN=2229, FN=0, Flags=........C, SSID=Wildcard (Broadcast)
489	7.647915	ee:14:4b:ae:a2:8c	Broadcast	802.11	200	Probe Request, SN=2230, FN=0, Flags=........C, SSID=Wildcard (Broadcast)
533	8.363332	Raspberr_39:76:ce	Broadcast	802.11	126	Probe Request, SN=3337, FN=0, Flags=........C, SSID=Wildcard (Broadcast)
540	8.458749	Raspberr_39:76:ce	Broadcast	802.11	126	Probe Request, SN=3339, FN=0, Flags=........C, SSID=Wildcard (Broadcast)
544	8.539522	MS-NLB-PhysServer-22_fb:0e:61:8c	Broadcast	802.11	99	Probe Request, SN=2135, FN=0, Flags=........C, SSID=SFR_A258
545	8.543322	Raspberr_39:76:ce	Broadcast	802.11	126	Probe Request, SN=3342, FN=0, Flags=........C, SSID=Wildcard (Broadcast)
679	10.539577	MS-NLB-PhysServer-22_fb:0e:61:8c	Broadcast	802.11	99	Probe Request, SN=2157, FN=0, Flags=........C, SSID=SFR_A258
734	11.539657	MS-NLB-PhysServer-22_fb:0e:61:8c	Broadcast	802.11	99	Probe Request, SN=2168, FN=0, Flags=........C, SSID=SFR_A258
985	15.541022	MS-NLB-PhysServer-22_fb:0e:61:8c	Broadcast	802.11	99	Probe Request, SN=2209, FN=0, Flags=........C, SSID=SFR_A258
1053	16.540444	MS-NLB-PhysServer-22_fb:0e:61:8c	Broadcast	802.11	99	Probe Request, SN=2228, FN=0, Flags=........C, SSID=SFR_A258
1096	17.122333	be:f8:a8:dd:c6:4b	Broadcast	802.11	155	Probe Request, SN=3674, FN=0, Flags=........C, SSID=Wildcard (Broadcast)
1171	18.540519	MS-NLB-PhysServer-22_fb:0e:61:8c	Broadcast	802.11	99	Probe Request, SN=2250, FN=0, Flags=........C, SSID=SFR_A258

Figure 42 l’interface principale de Wireshark , nous pouvons voir timestamp, émetteur (source), récepteur (destination) , Protocol et des informations plus précises

Grâce à Wireshark, nous pouvons examiner les trames individuelles capturées, afficher les détails des paquets comme la figure 43, les adresses MAC source et destination, les en-têtes de protocole, les données de charge utile et bien plus encore. Cela nous donne une visibilité précieuse sur le trafic Wi-Fi dans notre réseau.

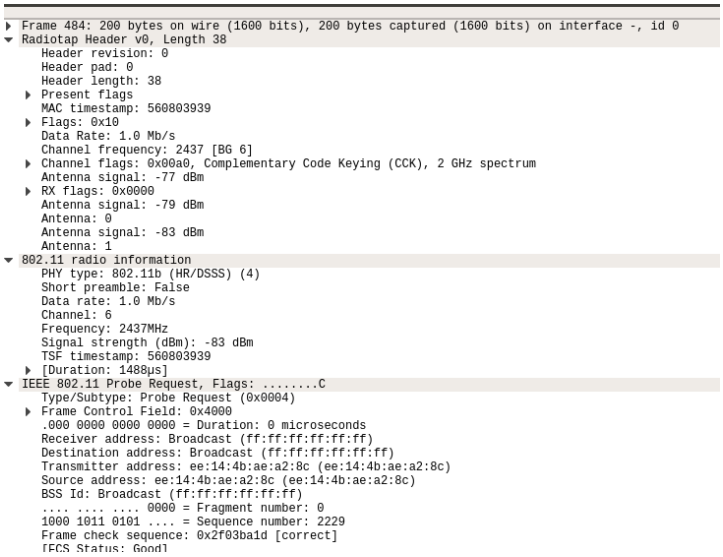


Figure 43 tous les détails pour une trame PR

En résumé, en remplaçant l'OS d'un point d'accès Wi-Fi ordinaire par OpenWRT, nous pouvons exploiter les fonctionnalités de configuration simultanée en mode connecté et en mode monitor. En utilisant TCPDump pour capturer les trames et Wireshark pour analyser leur contenu, nous pouvons obtenir une visibilité approfondie sur le trafic Wi-Fi dans notre réseau.

VII.A.2. démodulation en utilisant de Matlab

Nous avons réalisé une étape préliminaire dans notre recherche en utilisant MATLAB, où nous avons réussi à obtenir la trame complète numérique à partir de la démodulation. Ce processus nous a permis de comparer cette trame avec le protocole Wi-Fi et d'extraire des informations personnelles pertinentes avec succès. Cependant, nous sommes conscients que le processus actuel, qui est effectué manuellement, n'est pas viable pour gérer un grand nombre de trames à l'avenir. En raison de l'utilisation de HackRF, nous devons utiliser URH pour collecter les

données, puis manuellement extraire les PR à partir de l'interface visuelle d'URH. Malheureusement, nous n'avons pas encore trouvé de méthode pour automatiser cette étape dans URH.

Afin de remédier à cette situation, nous avons entrepris une réflexion pour automatiser

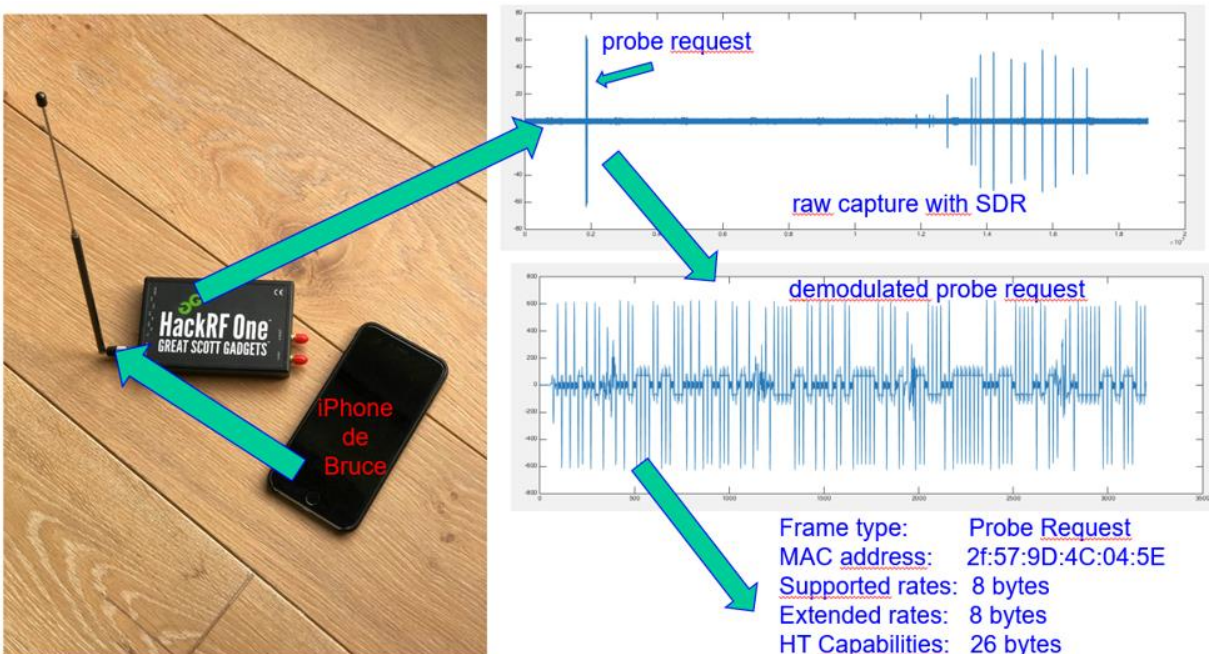


Figure 44 Recevoir un PR par la carte SDR, et puis on obtient I et Q de cette trame, ensuite faire la démodulation et comparer avec le standard pour trouver les informations de PR.

l'ensemble du processus. Notre objectif est de mettre en place une solution qui nous permettra de traiter de manière efficace et précise un volume important de trames Wi-Fi. Pour ce faire, nous envisageons d'utiliser une carte BladeRF (présenté dans la partie background), qui offre une flexibilité et une personnalisation accrues, ainsi que de développer une encapsulation appropriée pour obtenir de nouvelles variables.

L'automatisation de ce processus revêt une importance pour notre projet, car elle permettrait de rationaliser les opérations et de garantir des résultats fiables à grande échelle. En exploitant les avantages offerts par la carte WiFi OpenSource et en mettant en œuvre une encapsulation adaptée, nous serions en mesure de gérer efficacement la collecte et l'analyse de trames Wi-Fi, tout en minimisant les erreurs potentielles liées aux interventions manuelles.

VII.A.3. Encapsulation des données

Automatiser le processus de capture, de démodulation et d'analyse des trames Wi-Fi est certainement une étape cruciale pour gérer efficacement un grand volume de données. Dans cette partie, nous avons cherché des moyens pour encapsuler nos données.

L'utilisation de GNU Radio et du protocole RFtap pour encapsuler nos données Wi-Fi est une approche prometteuse. GNU Radio est une suite de logiciels open source dédiés au traitement du signal, offrant des fonctionnalités avancées pour la capture, la manipulation et l'analyse des signaux radio.

En utilisant GNU Radio Companion (GRC), un outil graphique fourni avec GNU Radio, nous avons pu créer un organigramme détaillé représentant les différentes étapes du processus, de la réception de la trame Wi-Fi à l'encapsulation des données et à leur envoi via un socket UDP. GRC nous permet de concevoir visuellement le flux de traitement du signal, en connectant des blocs fonctionnels pour effectuer des opérations spécifiques.

L'un des avantages clés de GRC est sa capacité à faciliter l'ajout de nouvelles variables sans avoir à écrire une seule ligne de VHDL (VHSIC Hardware Description Language). GRC propose la traduction de notre organigramme en un script Python, ce qui nous permet de réutiliser facilement notre travail pour envoyer de nouvelles trames avec des variables supplémentaires. Cela nous offre une grande flexibilité pour adapter notre système à de nouveaux scénarios ou pour expérimenter différents paramètres.

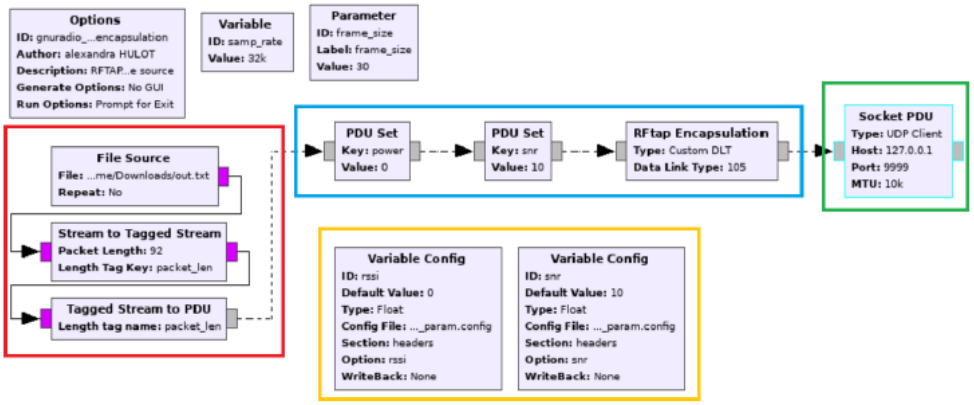


Figure 45 Organigramme rfrap_encapsulation.grc créé par GRC

En utilisant le protocole RFtap dans GNU Radio, nous pouvons encapsuler nos données dans un format standardisé, permettant ainsi leur analyse avec des outils compatibles RFtap, tels que Wireshark. RFtap fournit un moyen d'extraire les trames encapsulées et d'analyser leur contenu, ce qui facilite l'interprétation et la compréhension des données capturées.

En utilisant un autre script Python pour le traitement de nos données, nous avons pu automatiser la détection et l'envoi de nouvelles trames. Cette automatisation nous permet de gérer un flux continu de trames Wi-Fi, de les encapsuler avec les nouvelles variables souhaitées, puis de les transmettre via le socket UDP. Nous pouvons ensuite analyser ces trames encapsulées à l'aide de Wireshark ou d'autres outils d'analyse adaptés.

En résumé, en utilisant GNU Radio, nous avons pu encapsuler nos données Wi-Fi en utilisant le protocole RFtap. GRC facilite la création d'un organigramme décrivant les étapes de réception, d'encapsulation et d'envoi des trames. La traduction de l'organigramme en script Python nous permet de réutiliser facilement notre travail et d'ajouter de nouvelles variables sans avoir à écrire du code VHDL. En combinant ces capacités avec d'autres scripts Python pour le traitement et l'analyse des données, nous avons pu automatiser le processus de détection et d'envoi de nouvelles trames encapsulées.

VII.B. Carte BladeRF

VII.B.1. processus de démodulation de DSSS dans BladeRF

Le protocole IEEE 802.11b est utilisé dans les réseaux Wi-Fi pour les PR dans la bande 2,4 GHz ainsi que dans d'autres trames, et présente certaines caractéristiques clés. Il offre une portée de transmission d'environ 100 mètres et une vitesse maximale de 11 Mbps. Pour la modulation, il utilise la technique DSSS ainsi que les modulations QPSK (Quadrature Phase Shift Keying) ou BPSK (Binary Phase Shift Keying) [71] avec une bande passante de 22 MHz. L'utilisation des séquences de 11 chip par bit pour l'étalement de spectre en DSSS rend cette modulation intéressante pour les mesures de CSI. .

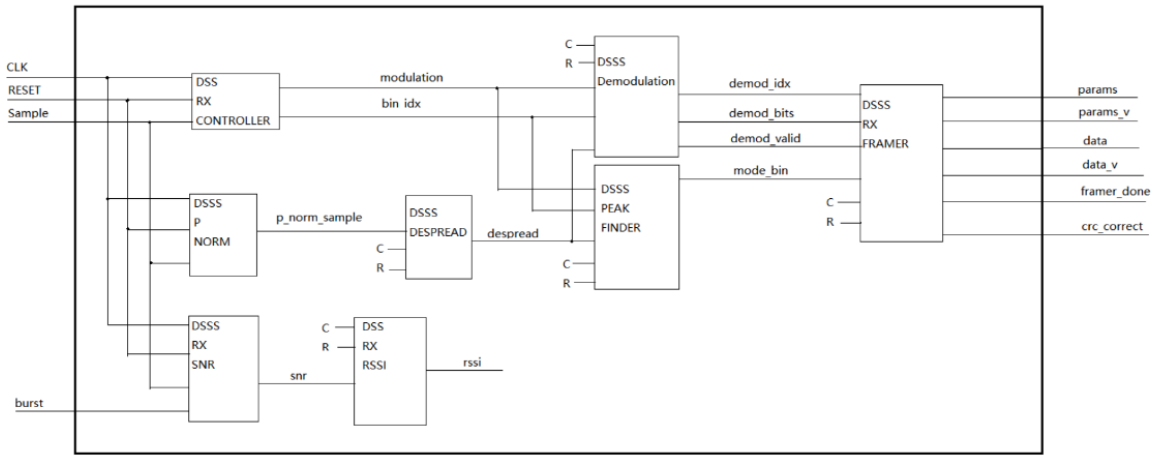


Figure 46 Structure de module U_dsss_rx dans la carte BladeRF

Le fichier U_dsss_rx fait partie du code VHDL dans le projet bladeRF et est spécifiquement lié à la réception des signaux Wi-Fi utilisant la modulation DSSS. Pour mieux comprendre sa structure et son fonctionnement, une analyse plus détaillée du code VHDL est nécessaire.

Dans le cadre du module global U_dsss_rx, plusieurs étapes clés sont implémentées :

Le module "wlan_dsss_despreader" est responsable de la désétalement des chips DSSS. Une séquence de chips connue, ou code, présente une forte corrélation avec sa conjuguée complexe lorsque la fenêtre de comparaison aligne parfaitement les deux séquences de chips. Même un

léger décalage temporel entre un code DSSS reçue et sa conjuguée complexe connue entraîne une corrélation très faible. La corrélation est également très faible lorsque la conjuguée complexe du code DSSS connue est corrélée avec du bruit ou des échantillons non DSSS. Le processus de comparaison des valeurs de corrélation entre les échantillons IQ reçus et les conjuguées complexes des codes DSSS pour trouver un bit est appelé désétalement. Un modem IEEE 802.11 DSSS utilise un préambule appelé "Start of Frame Delimited" (SFD) composé de 144 puces alternant entre -1 et +1 pour récupérer la synchronisation des symboles. Les corrélations positives et négatives sont utilisées pour représenter les bits "1" et "0". Les bits sont ensuite concaténés, convertis en octets, et stockés dans un FIFO pour recréer le PDU.

Le module "wlan_dsss_peak_finder" est chargé d'estimer la synchronisation des symboles en identifiant l'offset temporel qui présente la plus grande corrélation. L'indice de bin qui correspond à la corrélation maximale est considéré comme la synchronisation correcte. Il est ensuite sélectionné comme l'indice utilisé pour reconstruire progressivement le PDU, bit par bit.

Le module "wlan_dsss_demodulator" est responsable de la démodulation des bits PDU utilisant la modulation différentielle de phase binaire (DBPSK). Un bit "0" est démodulé lorsque le bit actuel a la même valeur que le précédent, tandis qu'un bit "1" est démodulé lorsque le bit actuel diffère du précédent.

Le module "wlan_dsss_rx_framer" joue un rôle essentiel en capturant le moment approprié pour enregistrer l'indice de bin correspondant à la synchronisation optimale. Cette estimation est ensuite transmise aux autres composants de la chaîne de réception DSSS. Le module "wlan_dsss_demodulator" est finalement responsable de la concaténation des bits démodulés provenant du module "wlan_dsss_demodulator" pour reconstituer les octets PDU. Il vérifie également que le CRC dans l'en-tête du paquet et le FCS dans le paquet correspondent pour garantir l'intégrité des données. Les octets PDU sont temporairement stockés dans un tampon dans le module "wlan_rx_packet_buffer" jusqu'à ce que le FCS du paquet puisse être validé.

VII.B.2. Processus de collection de Donnée

BladeRF CLI (Command Line Interface) est une interface en ligne de commande qui permet de contrôler le BladeRF et d'exécuter diverses opérations, y compris la capture de signaux RF.

Voici une approche générale pour utiliser BladeRF CLI afin d'obtenir des informations CSI :

Configuration du BladeRF : Assurez que le BladeRF est correctement connecté à l'ordinateur et que les pilotes appropriés sont installés. On peut utiliser les outils fournis par Nuand, le fabricant du BladeRF, pour effectuer la configuration initiale.

Ouvrir une session BladeRF CLI : Lancez une session BladeRF CLI en ouvrant une fenêtre de terminal et en exécutant la commande appropriée selon le système d'exploitation. Par exemple, sur Linux, on peut utiliser la commande bladeRF-cli.

Configuration des paramètres RF : Utilisez les commandes BladeRF CLI pour configurer les paramètres RF tels que la fréquence, la bande passante et le gain. Par exemple, la commande set frequency pour définir la fréquence de fonctionnement souhaitée.

Capture du signal : Utilisez la commande rx config pour configurer les paramètres de capture du signal. Cela comprend des éléments tels que la fréquence d'échantillonnage, la durée de capture et la destination des échantillons. Par exemple, la commande rx config file=capture.iq format=bin n=10000000 pour capturer un fichier binaire contenant 10 millions d'échantillons.

Traitement du signal : Une fois avoir capturé le signal, on peut utiliser des outils de traitement du signal pour extraire les informations CSI. Cela peut inclure des techniques telles que la démodulation, la synchronisation, l'estimation du canal et l'analyse spectrale. On peut faire appal aux bibliothèques telles que GNU Radio ou des langages de programmation tels que Python pour effectuer ces opérations.

VII.C. Campagne de mesures

Une campagne de mesures a été réalisée sur un terrain extérieur vallonné, ayant une couverture d'arbres, sans structures à part une maison isolée servant de base d'opérations. L'AP bladeRF

a été positionné sur un petit espace ouvert devant cette maison. Cet environnement a été choisi afin de s'approcher du type d'environnement rencontrés avec les données fournies par Aleia, SA, pour le cas des sites touristiques. Des trames PR ont été lancés à partir d'une vingtaine de positions autour de l'AP, avec un iPhone et un téléphone sous Androïde, comme illustrées dans la figure 47. Pour chaque émission, trois *bursts* de PR ont été créés en rentrant et sortant le téléphone du mode avion rapidement. Les PR ont été enregistrés dans des fichiers d'analyse avec le bladeRF en se servant du CLI. Le bladeRF échantillonne à 20 MHz, et nous l'avons calé sur le canal WiFi 11 pour ces tests.

VII.D. Dépouillement des données et développement des algorithmes CFO et CSI

Par la suite, les PR sont extraits en repérant les *bursts* dépassant le plancher de bruit et de durée cohérent avec un PR. En opération normale, le bladeRF se sert d'un algorithme VHDL permettant de démoduler les trames WiFi sans prétraitement. Transposant ce code en python, nous avons pu vérifier rapidement que les trames sélectionnées étaient bien des PR et non pas d'autres types de signaux (par exemple, compteur EdF de type « Linky ») et de vérifier le numéro de canal WiFi de l'émission. Un exemple des données bruts est également présenté en figure 47, où on peut voir également les bases des calculs de CFO et de CSI que nous allons effectuer. Le CFO est une mesure de porteuse résiduelle dans les valeurs de I et de Q du signal. C'est une mesure de décalage fréquentiel entre les oscillateurs locaux du téléphone et du bladeRF. Ainsi ce n'est pas une mesure du CFO *absolue* du téléphone mais plutôt du couple (télépone, AP). Néanmoins, comme nous nous intéressons à la dérandomisation de groupes de PR susceptibles de provenir d'un même client, notre CFO est bel et bien une grandeur d'intérêt. Quant à la CSI, comme indique la figure, nous allons nous baser sur la largeur du pic représentant un bit, après désétalement du signal DSSS, qui est fonction de la réponse impulsionnelle du canal multi-trajet emprunté par le PR.

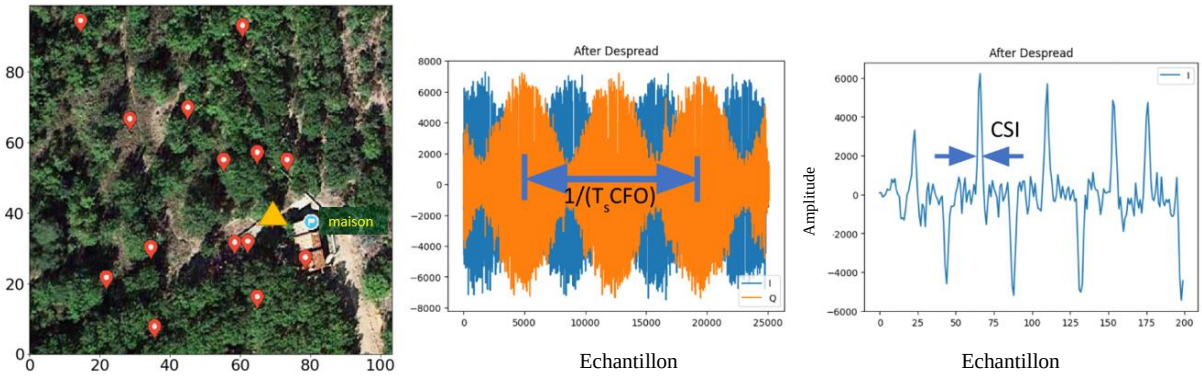


Figure 47 A gauche, terrain vallonné de la campagne de mesures, à couverture d'arbres, avec les positions des envois des PR. La position du bladeRF est indiquée par un triangle jaune à côté d'une maison isolée. Au centre, I et Q brut du signal, indiquant la présence d'une porteuse ayant un décalage CFO. A droite, indication de la source des mesures dites CSI, basées sur la largeur du pic représentant un bit après désétalement DSSS.

Afin d'obtenir une mesure de CSI plus précise, nous moyennons les pics bit sur la trame entière. Pour ce faire nous nous munissons d'un modèle de la trame, composée d'un échantillon unitaire, avec le même signe que le bit en question, à chaque position bit. Le moyennage se fait ensuite en calculant la convolution entre la trame reçue et ce modèle. Afin d'obtenir un résultat plus claire, nous nous limitons aux séquences de 3 bits ayant le même signe, car la fonction d'autocorrélation de la séquence Barker [72] (celle utilisé pour créer la modulation DSSS) est quasiment plate entre des bits de même signe. La technique est illustrée dans la figure 48. Un exemple de CSI brute pour une position sur le site est présenté en figure 49. On remarque, dans la figure, que le résultat est relativement stable pour les trois envois de PR.

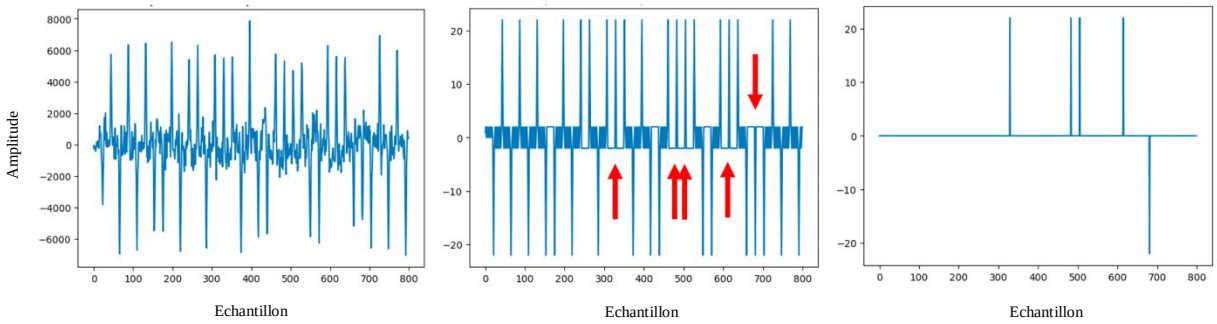


Figure 48 Gauche : section du signal I du PR après désétalement et correction porteuse. Centre : Modèle Barker du signal indiquant les positions où la corrélation théorique est presque plate. Droite : modèle utilisé pour moyenner le CSI.

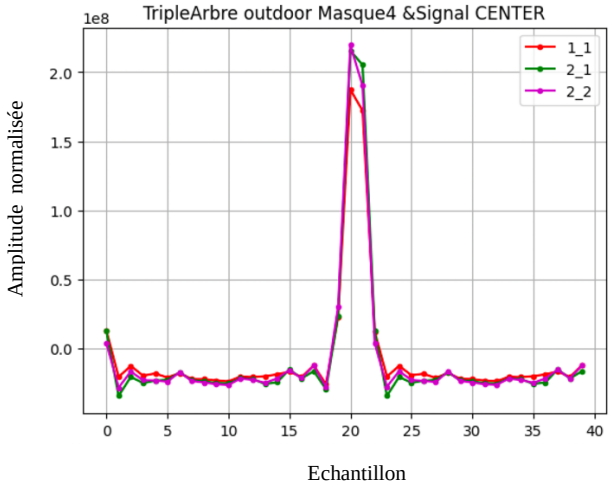


Figure 49 Exemple de CSI obtenu avec la procédure de convolution décrite dans le texte, pour 3 envois de PR de la position dite TripleArbre. Les résultats sont relativement stables pour les 3 envois.

Afin d'extraire des traits caractéristiques des CSI obtenus aux différentes positions, dans l'optique d'utiliser ceux-ci pour une dérandomisation, deux méthodes ont été expérimentées. La première, proposée dans [73], est une technique de déconvolution à partir d'un *pulse shape* de base. Bien que fournissant des *fits* aux courbes CSI très fidèles, cette technique a finalement été jugé inappropriée pour nos données car nous n'avons pas la possibilité de distinguer des trajets individuels, étant donné notre fréquence d'échantillonnage de 20 MHz. Ainsi, la déconvolution souvent proposait des solutions très différentes pour des CSI visuellement très similaires. Un example est présenté en figure 50.

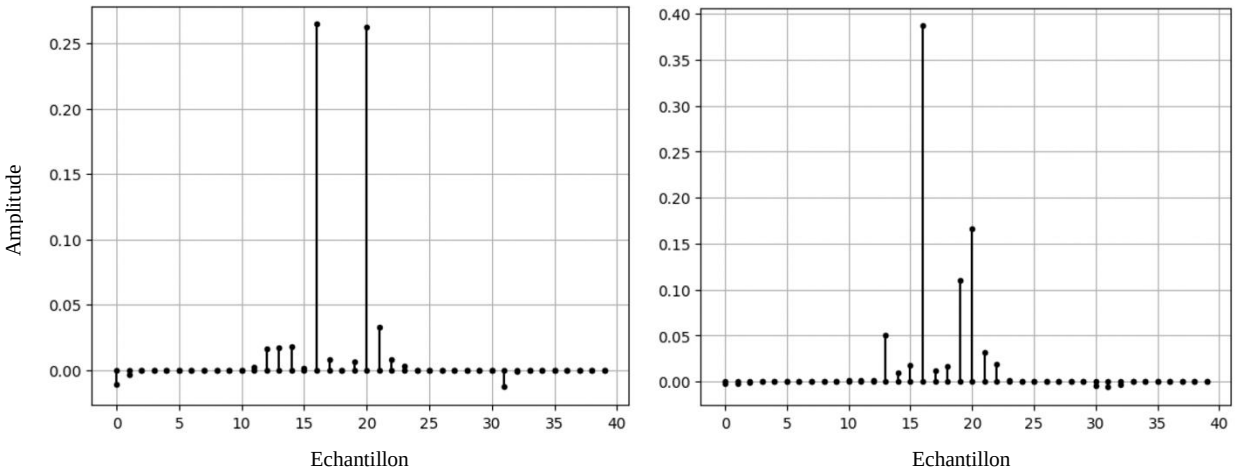


Figure 50 Résultat de déconvolution. Les échantillons sont censés représenter des répliques d'un *pulse shape* de base, ainsi s'assimilant à des trajets individuels. La figure montre deux résultats assez différents pour des CSI, qui, visuellement, étaient très similaires. La technique a ainsi été écartée.

Pour avancer, nous nous somme rabattus, dans un premier temps, sur un trait caractéristique plus simple, le *Full Width at Half Maximum* ou FWHM.

VII.E. Preuve de principe

Un résultat préliminaire, pour des PR provenant de 3 positions différentes, est présenté en Figure 51, à gauche, où les CSI ont été normalisés en amplitude et alignés dans le temps. Dans le panneau droite de la figure, on présente un *scatter plot* des FWHM des CSI versus leur RSSI. On constate, dans la figure, que les valeurs des traits caractéristiques obtenus pour les 3 lieux sont à la fois stables dans le temps, et différents entre elles. Ceci est un très bon résultat qui sert de première preuve de principe de notre technique.

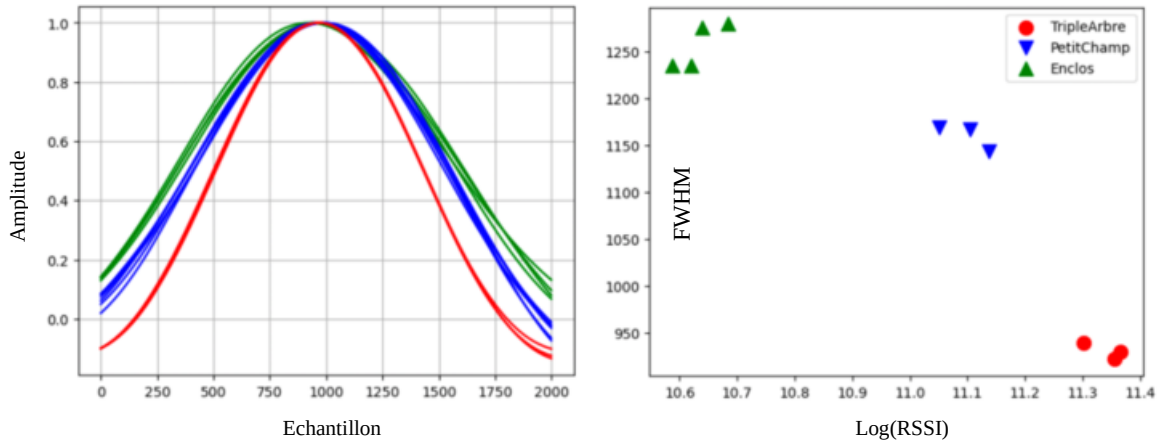


Figure 51 A gauche, les CSI venant des positions TripleArbre, PetitChamp, et Enclos, alignées temporellement et normalisées en amplitude. On constate qu'il est possible d'identifier le lieu à partir de la forme de la courbe. A droite, les FWHM versus le RSSI correspondant. Là encore, nous voyons des petits clusters de points pour les différentes répétitions de PR à chaque position.

Les résultats des mesures de CFO sont présentés en figure 52. Comme indiqué plus haut, pour les propos de ce manuscrit, le CFO mesure le décalage entre les oscillateurs locaux du téléphone et du bladeRF. La figure montre que ce décalage est suffisamment divers, selon le type de téléphone, et stable dans le temps, pour pouvoir distinguer un PR appartenant à un couple (téléphone, AP) très facilement. Toutefois, les mesures, présentées pour l'ensemble des données, ne correspondent pour l'heure qu'à un seul iPhone et un seul téléphone Androïde.

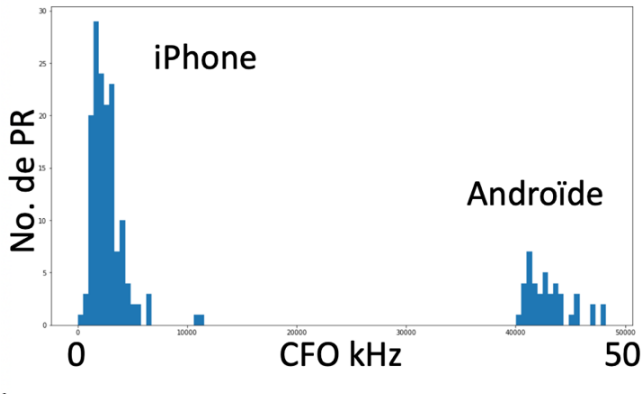


Figure 52 Histogramme des valeurs de CFO en kHz obtenues pour l'iPhone et le téléphone Androïde, sur l'ensemble des PR reçus sur le canal WiFi 11.

VII.F. Application à plus grande échelle

L'évaluation des résultats CSI est toujours en cours pour l'ensemble des données. Certains positions présentent des résultats plus bruités que ceux présentés ici. Nous pensons pouvoir améliorer les résultats en extrayant des traits caractéristiques supplémentaire des courbes CSI, et en écartant des cas ou la forme d'onde de départ du signal PR semble aberrant. Nous pensons pouvoir nous approcher à un résultat plus intéressant d'un point de vue pratique, à travers des simulations incorporant des éléments provenant de notre campagne de mesures. Concernant CFO, nous comptons effectuer des mesures sur un parc plus étendu de téléphones afin de mieux évaluer la saillance de ce trait caractéristique pour la dérandomisation, encore une fois, via des simulations. Lorsque des résultats exploitables seront obtenus, il sera question de traduire certains des algorithmes python développés en VHDL et les implanter dans le FPGA du bladeRF. L'encapsulation des traits caractéristiques extraits au sein de la trame PR reste également à implémenter.

VIII. Conclusion

Dans cette thèse, nous avons abordé plusieurs problématiques liées à la localisation et au comptage des clients à l'aide de signaux de réseau radio. Nous avons réalisé des avancées dans trois domaines clés, dont nous présentons un résumé et une perspective.

Tout d'abord, nous avons exploré l'utilisation de données de PRs émis par des clients avec des adresses MAC randomisés. Face aux préoccupations croissantes concernant la confidentialité des clients et les contraintes légales comme le GDPR, l'approche traditionnelle basée sur les adresses MAC s'est avérée insoutenable. Pour pallier cette problématique, nous avons développé une méthode novatrice basée sur l'analyse du facteur de Fano des données PR provenant de réseaux WiFi dans différentes villes françaises.

La méthode nous a permis de produire une mesure approximative, X, du nombre moyen de PR émis par jour par des clients WiFi avec des adresses MAC aléatoires. Ainsi, en divisant le nombre total de PRs par X, nous obtenons le nombre de clients quotidiens. Les résultats obtenus ont montré que les valeurs de X étaient relativement stables pour différents sites et volumes de PR, avec une variation d'environ un ordre de grandeur. De plus, les prédictions basées sur X se sont avérées raisonnables lorsqu'elles ont été validées en utilisant différentes approches. Nous avons également effectué un test de vérification sur le terrain dans un site de camping pour établir la concordance entre les prédictions basées sur X et les données réelles. Ces résultats encourageants ouvrent la voie à des approches de préservation de la confidentialité dans la collecte de données des clients WiFi, tout en fournissant des estimations fiables du nombre de PR produits par des clients à adresses MAC randomisées.

Cependant, notre technique a également révélé certaines limites. Notamment, sa sensibilité aux valeurs aberrantes a entraîné une incertitude statistique qui doit être améliorée. En outre, notre méthode actuelle ne permet d'évaluer X que sur des données PRs présentant des périodicités quotidiennes et/ou hebdomadaires. La validation basée sur des fenêtres horaires que nous avons présentée pourrait fournir des pistes pour étendre cette méthode à des données sans périodicités à plus grande échelle.

Ensuite, nous avons introduit une boîte à outils composée de neuf méthodes pour transformer directement les comptages bruts de PRs avec des adresses MAC aléatoires en cartes de densité de clients dans des réseaux WiFi extérieurs, sans recourir à des informations de référence au sol. Les résultats prometteurs obtenus en appliquant cette boîte à outils à des données réelles de deux sites français soulignent son intérêt potentiel pour les chercheurs et les gestionnaires de sites.

Cependant, nous reconnaissons que les prédictions de densité de notre technique actuelle se basent sur un taux de probabilité d'émission de PR, X, adopté à partir d'une étude similaire et ajusté par quelques corrections raisonnables mais approximatives. Une perspective essentielle est donc de découvrir des méthodes d'estimation plus précises pour la densité des clients, afin de calibrer notre boîte à outils avec plus de rigueur sur les données actuelles.

Enfin, notre travail se poursuit dans le domaine de la localisation de clients WiFi utilisant des adresses MAC aléatoires. Nous cherchons à exploiter les CSI et les CFO pour regrouper les signaux PRs en fonction de leurs caractéristiques physiques. Cette approche représente une perspective passionnante pour surmonter les défis posés par la randomisation des adresses MAC et renforcer la confidentialité des clients.

En résumé, cette thèse a apporté des contributions à la compréhension et à la résolution des problèmes liés à la localisation et au comptage des clients dans un contexte de confidentialité accrue des données clients. Nos résultats ouvrent des perspectives intéressantes pour de futures recherches dans ce domaine en pleine évolution. Nous espérons que notre travail sera utile aux chercheurs, aux responsables de sites pour une gestion plus efficace et respectueuse de la vie privée des utilisateurs dans les réseaux sans fil.

IX. Références

[1] R. Agarwal, S. Kumar et R. M. Hegde, «Algorithms for crowd surveillance using passive acoustic sensors over a multimodal sensor network,» *IEEE Sensors Journal*, vol. 15, n° %1 3, p. 1920–1930, 2014.

[2] O. Fatemieh, R. Chandra et C. A. Gunter, «Secure collaborative sensing for crowd sourcing spectrum data in white space networks,» *IEEE Symposium on New Frontiers in Dynamic Spectrum (DySPAN),* pp. 1-12,Singapore, 2010.

[3] H. Mohammadmoradi, S. Yin et O. Gnawali, «Room occupancy estimation through WiFi, UWB, and light sensors mounted on doorways,» *Proceedings of the 2017 International Conference on Smart Digital Environment,* pp. 27-34, 2017.

[4] O. Kaltiokallio, M. Bocca et N. Patwari, «Enhancing the accuracy of radio tomographic imaging using channel diversity,» *IEEE 9th International Conference on Mobile Ad-Hoc and Sensor Systems (MASS 2012),* pp. 254-262, 2012.

[5] «Mobile cellular subscriptions,» International Telecommunication Union, https://data.worldbank.org/indicator/IT.CEL.SETS.P2, 2023.

[6] P. Bahl et V. N. Padmanabhan, «Radar: An in-building rf-based user location and tracking system,» *Proceedings IEEE INFOCOM, Conference on computer communications,* vol. 2, pp. 775-784, 2000.

[7] M. Kotaru, K. Joshi, D. Bharadia et S. Katti, «Spotfi: Decimeter level localization using wifi,» *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*, pp. 269-282, 2015.

[8] H. Peng, «WIFI network information security analysis research,» chez *International Conference on Consumer Electronics, Communications and Networks (CECNet)*, Yichang, China, 2012.

[9] P. Voigt et A. Von dem Bussche, «The EU general data protection regulation (GDPR),» *A Practical Guide, 1st Ed Cham: Springer International Publishing*, n° %13152676, 2017.

[10] Y. Ouyang, Z. Le, Y. Xu, N. Triandopoulos, S. Zhang, J. Ford et F. Makedon, «Providing anonymity in wireless sensor networks,» *IEEE international conference on pervasive services*, pp. 145-148, 2007.

[11] J.-F. Determe, S. Azzagnuni, U. Singh, F. Horlin et P. D. Doncker, «Monitoring large crowds with WiFi: A privacy-preserving approach,» *IEEE Systems Journal*, vol. 16, n° %1 2, pp. 2148-2159, 2022.

[12] J. Martin, T. Mayberry, C. Donahue, L. Foppe, L. Brown, C. Riggins, E. C. Rye et D. Brown, «A study of MAC address randomization in mobile devices and when it fails,» *Proceedings on Privacy Enhancing Technologies*, 2017 arXiv preprint arXiv:1703.02874 (2017).

[13] U. Marco, R. Cossu, E. Ferrara, O. Bagdasar, A. Liotta et L. Atzori, «Wifi probes sniffing: an artificial intelligence based approach for mac addresses de-randomization,» *IEEE 25th*

International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), pp. 1-6, 2020 Sep 14.

[14] L. Oliveira, D. Schneider, J. D. Souza et W. Shen, «Mobile device detection through WiFi probe request analysis,» *IEEE Access,* vol. 7, pp. 98579-98588, 2019.

[15] T. Terje, B. Soundararaj et J. Cheshire, «Using Wi-Fi probe requests from mobile phones to quantify the impact of pedestrian flows on retail turnover,» *Computers, Environment and Urban Systems,* vol. 87, p. 101601, 2021.

[16] G. Sonja, P. Rutten, J. Amoraal, E. Ranguelova, R. Bakhshi, B. L. d. Vries, M. Lees et S. Klous, «Detecting high indoor crowd density with Wi-Fi localization: A statistical mechanics approach,» *Journal of Big Data,* n° %11, pp. 1-23, 2019.

[17] J.-F. Determe, U. Singh, F. Horlin et P. D. Doncker, «Forecasting crowd counts with Wi-Fi systems: Univariate, non-seasonal models,» *IEEE Transactions on Intelligent Transportation Systems,* n° %110, pp. 6407-6419, 2020.

[18] S. Balamurugan, J. Cheshire et P. Longley, «Estimating real-time high-street footfall from Wi-Fi probe requests,» *International Journal of Geographical Information Science,* vol. 34, n° %12, pp. 325-343, 2020.

[19] F. Yang, I. Ahriz et B. Denby, «Statistical Approach to Estimating Audience from MAC-Randomized WiFi Probe Requests,» *Sensors,* n° %122, p. 8679, 2022.

[20] R. Lamia, Q. Ni et T. Turletti, «Adaptive EDCF: enhanced service differentiation for IEEE 802.11 wireless ad-hoc networks,» *IEEE Wireless Communications and Networking,* vol. 2, pp. 1373-1378, 2003.

[21] K. Thomas, A. Broido, M. Faloutsos et K. C. Claffy, «Transport layer identification of P2P traffic,» *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, pp. 121-134, 2004.

[22] S. Kim, J. Chun et C. H. O. HanGyu., «Method for finding instrument for wi-fi direct P2P (peer to peer) communication and apparatus therefor.». Brevet U.S. Patent 9,736,766, 15 8 2017.

[23] «official website of Ruckus,» [En ligne]. Available: https://www.ruckusnetworks.com/.

[24] «IEEE Standard for Information technology--Telecommunications and information exchange between systems Local and metropolitan area networks--Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications,» *IEEE Std 802.11-2012 (Revision of IEEE Std 802.11-2007)* , 2012.

[25] B. Ağır, K. Huguenin, U. Hengartner et J.-P. Hubaux, «On the privacy implications of location semantics,» *Proceedings on Privacy Enhancing Technologies*, n° %14, 2016.

[26] F. Liu, J. Liu, Y. Yin, W. Wang, D. Hu, P. Chen et Q. Niu, «Survey on WiFi-based indoor positioning techniques,» *IET communications*, vol. 14, n° %19, pp. 1372-1383, 2020.

[27] V. Q. Duy et P. De, «A survey of fingerprint-based outdoor localization,» *IEEE Communications Surveys & Tutorials*, vol. 18, n° %11, pp. 491-506, 2015.

[28] S. A. H., M. Tamazin, M. A. Sharkas et M. Khedr, «An enhanced WiFi indoor localization system based on machine learning,» *International conference on indoor positioning and indoor navigation (IPIN)*, pp. 1-8 IEEE, 2016.

[29] X. Wang, L. Gao, S. Mao et S. Pandey, «DeepFi: Deep learning for indoor fingerprinting using channel state information,» *IEEE wireless communications and networking conference (WCNC)*, pp. 1666-1671. IEEE, 2015.

[30] Z. Yang, Z. Zhou et Y. Liu, «From RSSI to CSI: Indoor localization via channel response,» *ACM Computing Surveys*, vol. 46, n° %12, pp. 1-32, 2013.

[31] F. Jason, D. McCoy, P. Tabriz, V. Neagoe, J. V. Randwyk et D. Sicker, «Passive Data Link Layer 802.11 Wireless Device Driver Fingerprinting,» *USENIX Security Symposium*, vol. 3, n° %12006, pp. 16-89, 2006.

[32] J. Freudiger, «How talkative is your mobile device? An experimental study of Wi-Fi probe requests,» *Proceedings of the 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks* , pp. 1-6, 2015.

[33] S. Jamil, S. Khan, A. Basalamah et A. Lbath., «Classifying smartphone screen ON/OFF state based on wifi probe patterns,» *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, pp. 301-304, 2016 Sep 12.

[34] L. Roman, M. Zimmerling et L. Thiele, «Passive, privacy-preserving real-time counting of unmodified smartphones via zigbee interference,» *2015 International Conference on Distributed Computing in Sensor Systems*, pp. 115-126, 2015.

[35] O. Waltari et J. Kangasharju., «The wireless shark: Identifying wifi devices based on probe fingerprints,» *Proceedings of the First Workshop on Mobile Data*, pp. 1-6, 2016.

[36] A. Naeim, A. Bhaskar et E. Chung, «Bluetooth and Wi-Fi MAC address based crowd collection and monitoring: Benefits, challenges and enhancement,» *Australasian*

Transport Research Forum 2013 Proceedings, pp. 1-17, Australasian Transport Research Forum, 2013..

[37] A. E. C. Redondi, D. Sanvito et M. Cesana, «Passive classification of Wi-Fi enabled devices,» *Proceedings of the 19th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, pp. 51-58, 2016.

[38] E. Martin, O. Vinyals, G. Friedland et R. Bajcsy, «Precise indoor localization using smart phones.,» *Proceedings of the 18th ACM international conference on Multimedia*, pp. 787-790, 2010 Oct 25.

[39] F. Potortì, A. Crivello, M. Girolami, P. Barsocchi et E. Traficante, «Localising crowds through Wi-Fi probes,» *Ad Hoc Networks,* vol. 75, pp. 87-97, 2018 Jun 1.

[40] Z. Xu, K. Sandrasegaran, X. Kong, X. Zhu, B. Hu, J. Zhao et C. Lin., «Pedestrain monitoring system using Wi-Fi technology and RSSI based localization,» *International Journal of Wireless & Mobile Networks*, pp. 17-35, 2013.

[41] L. Schauer, M. Werner et P. Marcus, «Estimating crowd densities and pedestrian flows using wi-fi and bluetooth,» *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pp. 171-177, 2014 Dec 2.

[42] M. W. Traunmueller, N. Johnson, A. Malik et C. E. Kontokosta, «Digital footprints: Using WiFi probe and locational data to analyze human mobility trajectories in cities,» *Computers, Environment and Urban Systems*, vol. 72, pp. 4-12, 2018 Nov.

[43] P. Sapiezynski, R. Gatej, A. Mislove et S. Lehmann, «Opportunities and challenges in crowdsourced wardriving,» *Proceedings of the 2015 Internet Measurement Conference*, pp. 267-273, 2015.

[44] Y. Fukuzaki, M. Mochizuki, K. Murao et N. Nishio, «A pedestrian flow analysis system using Wi-Fi packet sensors to a real environment,» *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pp. 721-730, 2014 Sep 13.

[45] D. Dardari, P. Closas et P. M. Djurić, «Indoor tracking: Theory, methods, and technologies,» *IEEE Transactions on Vehicular Technology*, vol. 64, n° %14, pp. 1263-1278, 2015.

[46] J. Weppner, B. Bischke et P. Lukowicz, «Monitoring crowd condition in public spaces by tracking mobile consumer devices with wifi interface,» *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, pp. 1363-1371, 2016 Sep 12.

[47] H. Daniel, W. Hu, A. Sheth et D. Wetherall, «Tool release: Gathering 802.11 n traces with channel state information,» *ACM SIGCOMM computer communication review*, vol. 41, n° %11, pp. 53-53, 2011.

[48] T. Xin, B. Guo, Z. Wang, M. Li, Z. Yu et X. Zhou, «Freesense: Indoor human identification with Wi-Fi signals,» *IEEE Global Communications Conference (GLOBECOM)*, pp. 1-7, 2016 Dec 4.

[49] B. Bram, A. Barzan, P. Quax et W. Lamotte, «WiFiPi: Involuntary tracking of visitors at mass events,» *IEEE International Symposium on" A World of Wireless, Mobile and Multimedia Networks"(WoWMoM)*, pp. 1-6, 2013.

[50] F. Jason, D. McCoy, P. Tabriz, V. Neagoe, J. V. Randwyk et D. Sicker, «Passive Data Link Layer 802.11 Wireless Device Driver Fingerprinting,» *USENIX Security Symposium*, vol. 3, n° %12006, pp. 16-89, 2006.

[51] C. Matte, M. Cunche, F. Rousseau et M. Vanhoef, «Defeating MAC address randomization through timing attacks.,» *Proceedings of the 9th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, pp. 15-20, 2016.

[52] M. Uras, R. Cossu, E. Ferrara, O. Bagdasar, A. Liotta et L. Atzori, «Wifi probes sniffing: an artificial intelligence based approach for mac addresses de-randomization.,» *2020 IEEE 25th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, pp. 1-6, 2020 Sep 14.

[53] M. Vanhoef, C. Matte, M. Cunche, L. S. Cardoso et F. Piessens, «Why MAC address randomization is not enough: An analysis of Wi-Fi network discovery mechanisms,» *Proceedings of the 11th ACM on Asia conference on computer and communications security*, pp. 413-424, 2016.

[54] C. Groba, «Demonstrations and people-counting based on Wifi probe requests,» *2019 IEEE 5th World Forum on Internet of Things (WF-IoT)*, pp. 596-600, 2019.

[55] H. Hong, G. D. D. Silva et M. C. Chan, «Crowdprobe: Non-invasive crowd monitoring with wi-fi probe,» *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, pp. 1-23, 2018 Sep 18.

[56] L. Oliveira, D. Schneider, J. D. Souza et W. Shen, «Mobile device detection through WiFi probe request analysis,» *IEEE Access*, vol. 7, pp. 98579-98588., 2019.

[57] Y. Furuya, H. Asahina, M. Yoshida et I. Sasase, «Indoor crowd estimation scheme using the number of wi-fi probe requests under mac address randomization,» *IEICE TRANSACTIONS on Information and Systems 104*, pp. 1420-1426, 2021.

[58] A. Guillen-Perez et M.-D. Cano, «Counting and locating people in outdoor environments: a comparative experimental study using WiFi-based passive methods.,» *ITM Web of Conferences*, vol. 24, p. 01010, EDP Sciences, 2019.

[59] K. Gebru, «A privacy-preserving scheme for passive monitoring of people's flows through WiFi beacons,» *IEEE Annual Consumer Communications & Networking Conference (CCNC)*, pp. 421-424, 2022 Jan 8.

[60] L. Pintor et L. Atzori, «Analysis of Wi-Fi Probe Requests Towards Information Element Fingerprinting,» *GLOBECOM 2022-2022 IEEE Global Communications Conference*, pp. 3857-3862, 2022.

[61] K.-J. Djervbrant et A. Häggström, «A Study on Fingerprinting of Locally Assigned MAC-Addresses.,» *Bachelor's Thesis,* Halmstad, Sweden, 2019.

[62] S. I. Rahman et N. W. Finn, «Spanning tree protocol for wireless networks,» *U.S. Patent 7,653,011*, 2010.

[63] Y. Tian, B. Denby, I. Ahriz, P. Roussel et G. Dreyfus, «Robust indoor localization and tracking using GSM fingerprints,» *EURASIP Journal on Wireless Communications and Networking*, pp. 1-12, 2015.

[64] H. Obeidat, W. Shuaieb, O. Obeidat et R. Abd-Alhameed, «A review of indoor localization techniques and wireless technologies,» *Wireless Personal Communications*, pp. 289-327, 2021.

[65] A. Küpper, Location-based services: fundamentals and operation, John Wiley & Sons, 2005.

[66] R. J. Thompson, E. Cetin et A. G. Dempster, «Unknown source localization using RSS in open areas in the presence of ground reflections,» *Proceedings of the 2012 IEEE/ION Position, Location and Navigation Symposium*, pp. 1018-1027, 2012.

[67] G. Deak, K. Curran et J. Condell, «A survey of active and passive indoor localisation systems,» *Computer Communications 35*, pp. 1939-1954, 2012.

[68] S.-C. Yeh, W.-H. Hsu, M.-Y. Su, C.-H. Chen et K.-H. Liu, «A study on outdoor positioning technology using GPS and WiFi networks,» *International Conference on Networking,Sensing and Control*, pp. 597-601, 2009.

[69] C. G. Wheeler et D. R. Reising, «Assessment of the impact of CFO on RF-DNA fingerprint classification performance,» *International Conference on Computing, Networking and Communications*, pp. 110-114, 2017.

[70] D. Halperin, W. Hu, A. Sheth et D. Wetherall, «Tool release: Gathering 802.11 n traces with channel state information,» *ACM SIGCOMM computer communication review*, n° %11, p. 53, 2011.

[71] D. Maas, M. H. Firooz, J. Zhang, N. Patwari et S. K. Kasera, «Channel sounding for the masses: Low complexity GNU 802.11 b channel impulse response estimation,» *IEEE transactions on wireless communications*, vol. 11, n° %11, pp. 1-8, 2011.

[72] N. Levanon et E. Mozeson, Radar Signals, Eyrolles, 2004.

[73] J.-J. Fuchs, «Multipath time-delay detection and estimation,» *IEEE transactions on signal processing*, vol. 47, n° %11, pp. 237-243, 1999.

[74] Y. Li et T. Zhu, «Gait-based wi-fi signatures for privacy-preserving,» chez *Proceedings of the 11th ACM on asia conference on computer and communications security*, 2016.

X. Publications issues du travail de thèse

Yang, Feifei, Iness Ahriz, and Bruce Denby. 2022. "Statistical Approach to Estimating Audience from MAC-Randomized WiFi Probe Requests" *Sensors* 22, no. 22: 8679. https://doi.org/10.3390/s22228679

Yang, Feifei, Iness Ahriz, and Bruce Denby. 2023. "Tools for Ground-Truth-Free Passive Client Density Mapping in MAC-Randomized Outdoor WiFi Networks" *Sensors* 23, no. 13: 6142. https://doi.org/10.3390/s23136142