French Civil Aviation University (ENAC)

7 avenue Édouard Belin - CS 31055 Toulouse Cedex 4
Tél. +33 (0)5 62 17 40 00 - www.enac.fr

IENAC 21

Projet SAT S8

Vérification de l'emplacement avec cartographie du signal WiFi

Défendu le xx mai 2023 par

GOSSIN Antonin
HUGUET Célestin

Encadré par

LESOUPLE Julien
GHIZZO Emile

Table des matières

Introduction

Remerciements

Nous tenons à exprimer nos sincères remerciements à Julien LESOUPLE et Emile GHIZZO pour leur encadrement et leur soutien tout au long de ce projet. Leur expertise, leur patience et leur disponibilité ont grandement contribué à la réussite de ce travail.

Introduction

L'utilisation des technologies satellites est devenue courante pour le repérage sur la surface de la Terre, mais il existe des situations où ces technologies ne sont pas efficaces, notamment en intérieur ou dans des zones géographiques avec de nombreux obstacles. Dans de tels cas, l'utilisation du Wi-Fi comme outil de positionnement peut s'avérer prometteuse. Les routeurs Wi-Fi émettent des signaux puissants dans leur environnement, et grâce à des applications sur smartphones, il est possible de mesurer la puissance de ces signaux émis par différentes stations Wi-Fi dans une zone donnée.

Ce projet se concentre sur l'utilisation des signaux Wi-Fi pour la localisation en intérieur, en exploitant la puissance des signaux des routeurs Wi-Fi présents dans la zone géographique étudiée. Chaque balise Wi-Fi est identifiée par un BSSID unique, et en mesurant les puissances des signaux émises par ces balises, il est possible de retrouver sa position en comparant les observations aux données préalablement mesurées. Afin de construire un modèle de prédiction précis, une base de données a été constituée en enregistrant les niveaux de signal Wi-Fi pour différents BSSID en 103 points répartis de manière équitable dans la cour principale de l'ENAC.

L'algorithme K-Nearest Neighbors (K-NN) est utilisé dans ce projet pour la régression, une technique d'apprentissage supervisé. Cet algorithme, proposé pour la première fois par Fix et Hodges en 1951, se base sur l'idée intuitive que les échantillons similaires se regroupent dans l'espace des caractéristiques. Il attribue une étiquette à une nouvelle observation en se basant sur la classe majoritaire parmi ses k plus proches voisins. Dans le contexte de la localisation par signal Wi-Fi, K-NN s'est révélé efficace et a été largement utilisé.

L'objectif principal de ce projet est de déterminer le secteur où il est le plus probable que l'emplacement se trouve en utilisant les techniques de Bayes et K-NN. L'estimation Bayésienne, une approche statistique permettant de calculer la probabilité d'un événement en fonction d'informations préalables, est utilisée pour estimer la probabilité de la position d'un objet en fonction des observations de puissance de signal Wi-Fi. L'estimation par noyaux, une méthode non paramétrique, est utilisée pour approximer la densité de probabilité à partir des données recueillies.

Dans cette étude sera présenter la construction d'une carte représentant la probabilité de présence par secteurs dans la zone étudiée, en utilisant les algorithmes de Bayes et K-NN. Les performances de ces deux algorithmes sont comparées en termes de probabilité de présence estimée et des

facteurs tels que l'approche de prédiction, la capacité d'apprentissage et la sensibilité aux données sont pris en compte.

Cette introduction met en évidence les motivations du projet, l'utilisation du Wi-Fi comme outil de positionnement en intérieur, l'application de l'algorithme K-NN pour la régression et l'importance de l'estimation Bayésienne. Les objectifs et les méthodes utilisées dans ce projet sont également exposés.

Chapitre 1

Contexte

1.1 Présentation du sujet

Dans de nombreux cas, l'utilisation des technologies satellites est courante pour le repérage sur la surface de la Terre. Cependant, il existe des situations où ces technologies ne sont pas les plus efficaces, notamment dans des bâtiments ou des zones géographiques avec de nombreux obstacles.

L'objectif de ce projet est d'utiliser la puissance des signaux des routeurs Wi-Fi présents dans la zone géographique étudiée pour permettre le positionnement.

En effet, le Wi-Fi devient un outil important pour le positionnement dans ces espaces, en particulier pour les ordinateurs et les smartphones, grâce à son déploiement en intérieur.

Pour plus de clarté, chaque balise émettant du Wi-Fi est attribuée d'un BSSID (Basic Service Set Identifiers), qui sont des identifiants uniques permettant de les différencier.

Enfin, certaines applications sur smartphone permettent de mesurer la puissance des signaux émis par les différentes stations Wi-Fi dans la zone. À partir de ces puissances, il est alors possible de retrouver sa position en comparant les observations aux données mesurées préalablement.

$^{2\ 2\ 2}\text{cours}_{p}rincipale_{v}ierge.png$

FIGURE 1.1 – Cour centrale de l'ENAC : Zone à cartographier
source : www.enac.fr

Une base de données a été constituée à l'aide de l'application Wi-Fi Heat Map et de téléphones portables dans la cour principale de l'ENAC. L'application Wi-Fi Heat Map a été utilisée pour mesurer les niveaux de signal Wi-Fi dans cette zone spécifique.

Les mesures ont été effectuées en 103 points différents répartis de manière équitable dans la cour de l'ENAC, comme illustré dans la Figure **??**. L'objectif était d'obtenir les valeurs de puissance reçues des 31 BSSID différents détectés, associées aux coordonnées géographiques de chaque mesure. Ainsi, les informations sur l'emplacement précis de chaque point de mesure dans la cour (dans le repère (x,y)) ont été enregistrées dans la base de données.

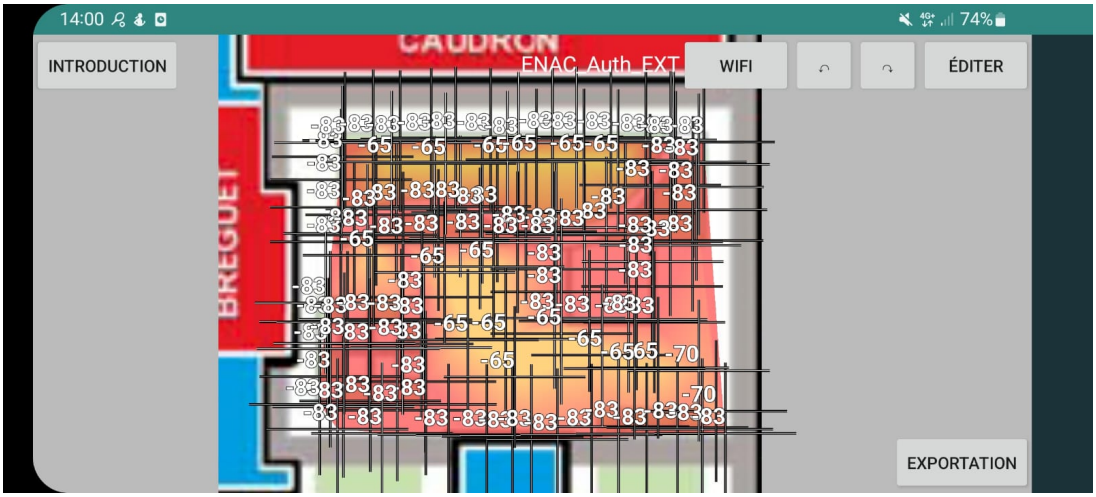

FIGURE 1.2 – Cour centrale de l'ENAC : zone cartographiée avec les points de mesures représentés *source : www.enac.fr*

Pour un point (x, y), correspondant aux coordonnées géographiques dans la zone étudiée, un repère orthonormé est considéré avec l'origine au point noir de la figure **??**. Les composantes selon l'axe x et l'axe y sont représentées par x et y respectivement. La liste des BSSID détectés ainsi que les valeurs de puissance des signaux reçus en dB sont disponibles. Cette phase de mesure permettra de comparer les valeurs obtenues pour une position recherchée avec celles de la base de données.

x	y	00:19:77:64:91:68	00:19:77:64:91:69	00:19:77:64:8d:14	00:19:77:61:37:29	00:19:77:64:8d:16	00:19:77:64:8d:15	00:19:77:61:3d:6a
801.5061	868.9537	-76	-76	-54	-59	-54	-54	-62
746.637	868.9537	-76	-76	-54	-59	-54	-54	-62
754.97156	901.8256	-76	-76	-54	-59	-54	-54	-62

FIGURE 1.3 – Extrait de la base de données. Les colonnes bleues correspondent à des BSSID et les jaune/orange à la position.

1.2 Sectorisation de la zone

La zone géographique a été divisée en secteurs s_k pour $k \in [0; S]$, où S représente le nombre total de secteurs. Pour ce faire, les valeurs les plus extrêmes selon les deux axes ont été prises en compte afin de définir un cadre, puis cette surface a été subdivisée uniformément en secteurs rectangulaires identiques. Le découpage de la zone est illustré dans la figure ci-dessous 1.4.

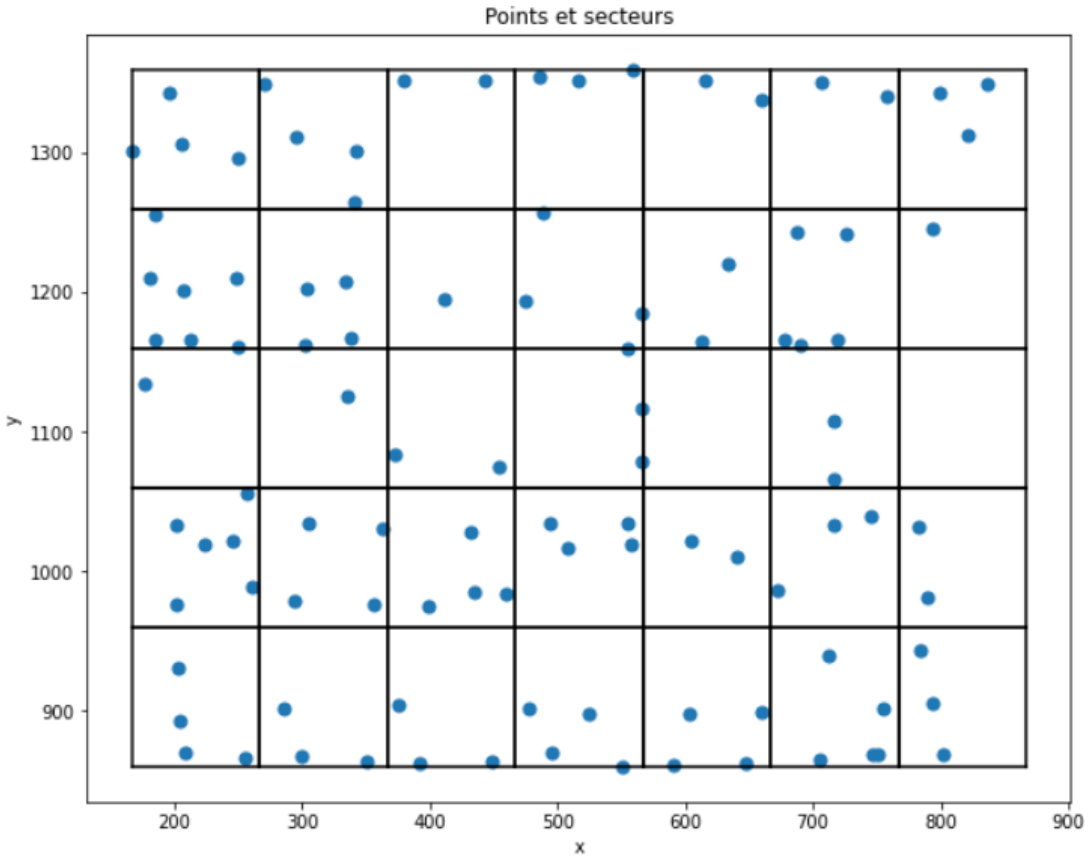

FIGURE 1.4 – Représentation de la cour centrale de l'ENAC décomposée en secteurs et où les points de mesures ont été représentés

L'objectif du projet est de déterminer le secteur où il est le plus probable que l'emplacement se trouve.

Dans la suite du projet, $\boldsymbol{l}$ représente la localisation (x, y) de l'emplacement, et $\boldsymbol{o}$ représente les observations des puissances des signaux Wi-Fi en ce point (x, y) [référence : **??**].

Chapitre 2

État de l'art à propos de l'estimation

2.1 L'algorithme K-Nearest Neighbors

2.1.1 Principe général

L'algorithme K-Nearest Neighbors (*K-NN*) est un algorithme d'apprentissage supervisé utilisé pour la classification et la régression. Cette méthode a été proposée pour la première fois par Fix et Hodges en 1951 [5]. Elle est basée sur l'idée intuitive que les échantillons qui se ressemblent tendent à se regrouper dans l'espace des caractéristiques. Le principe de base consiste à attribuer une étiquette à un nouvel échantillon en fonction de la classe majoritaire parmi ses k plus proches voisins. L'utilisation de techniques de pondération des voisins en fonction de leur distance a été proposée pour donner plus d'importance aux voisins les plus proches [6]. Des extensions spécifiques de *K-NN* ont été développées pour des applications particulières, comme la régression *K-NN* pour la prédiction de valeurs continues [7].
Avant d'en venir à notre cas d'étude il est bon de noter qu'en apprentissage supervisé un algorithme reçoit un ensemble de *données d'entrée*, nommé $\boldsymbol{X}$, qui est étiqueté avec des *valeurs de sortie*, nommées $\boldsymbol{y}$, correspondantes sur lequel il va pouvoir s'entraîner et définir un modèle de prédiction. Cet algorithme pourra par la suite être utilisé sur de nouvelles données afin de prédire leurs valeurs de sorties correspondantes. Ainsi *K-NN* estime la valeur d'une nouvelle observation en fonction de ses k plus proches voisins dans un ensemble d'entraînement (d'où son nom).
La différence entre classification ou régression est la suivante :

— Pour la classification, l'algorithme va comparer sa proximité avec les échantillons existants en utilisant une mesure de distance telle que la distance euclidienne. Les k échantillons les plus proches sont ensuite sélectionnés pour établir la classe à laquelle appartient l'échantillon inconnu.

— Pour la régression, l'algorithme prédit la valeur numérique d'un point de données inconnu en prenant la moyenne des k valeurs les plus proches.

Dans le cadre de notre projet cet algorithme sera utilisé avec la régression. En effet dans le domaine de la localisation par signal Wi-Fi, plusieurs travaux ont été réalisés en utilisant l'algorithme *K-NN*. Liu et Chaudhary ont présenté en 2012 une vue d'ensemble des avancées récentes et des comparaisons des techniques de localisation basées sur les empreintes Wi-Fi [8]. De même, Kaemarungsi et Krishnamurthy ont étudié en 2004 les propriétés de la force du signal reçu en intérieur pour la localisation par empreintes Wi-Fi [9].
Ces travaux antérieurs ont jeté les bases de la localisation WiFi et ont démontré l'efficacité de l'algorithme *K-NN* dans la localisation en intérieur.

L'ensemble de données d'entrée $\boldsymbol{X}$ correspond aux observations $\boldsymbol{o}$ (qui sont les enregistrements de puissance des signaux Wi-Fi) et les valeurs de sorties $\boldsymbol{y}$ correspondent à la localisation $\boldsymbol{l}$ (i.e. les coordonnées (x,y)) des différents points à l'intérieur du lieu étudié.

Enfin *K-NN* à l'avantage d'être simple à mettre en œuvre, mais il peut être sensible aux données bruyantes ou aux valeurs aberrantes. Il est alors important de choisir une valeur de k appropriée, ce choix repose sur un compromis entre la précision et la complexité de l'algorithme :

— Si k est petit : l'algorithme sera plus sensible aux points de mesure individuels et aura tendance à produire des résultats plus précis mais moins robustes car il est plus sensible aux variations de bruit ou d'erreurs dans les données. Cela peut conduire à un surapprentissage (overfitting) où l'algorithme est trop adapté aux données d'entraînement et ne généralise pas bien à de nouvelles données.

— Si k est grand : l'algorithme sera moins sensible aux points de mesure individuels et aura tendance à produire des résultats plus robustes mais moins précis car il y a moins de différenciation entre les différents points de mesure. Cela peut conduire à un sous-apprentissage (underfitting) où l'algorithme n'apprend pas suffisamment des données d'entraînement et ne généralise pas bien à de nouvelles données.

En résumé les grandes étapes de cet algorithme sont les suivantes :

(i) Sélectionner la valeur de k (nombre de voisins les plus proches à prendre en compte) et une distance (euclidienne, de Manhattan, etc.).

(ii) Calculer la distance entre la nouvelle observation et toutes les observations de l'ensemble d'entraînement.

(iii) Sélectionner les k observations les plus proches en fonction de la distance calculée dans l'étape précédente.

(iv) Pour la régression, estimer la valeur moyenne de la variable cible pour les k voisins les plus proches et assigner cette valeur à la nouvelle observation.

(v) Retourner la valeur assignée à la nouvelle observation.

2.1.2 Application de l'algorithme K-NN à la localisation par Wi-Fi

Comme mentionnée précédemment, l'algorithme *K-NN* va servir ici à résoudre des problèmes de localisation par Wi-Fi en utilisant une base de données de puissances de signaux Wi-Fi pour différentes positions connues.

Pour cela il faut procéder comme suit :

(a) Collecter des données de puissance de signal Wi-Fi pour différentes positions connues. Cela peut être fait en utilisant des dispositifs tels que des smartphones ou des ordinateurs portables équipés de récepteurs Wi-Fi. Ici l'application smartphone "Wi-Fi Heat Map" sera utilisée.

(b) Créer une base de données avec les coordonnées de chaque position collectée et les puissances de signaux Wi-Fi mesurées à ces positions.

(c) Lorsqu'une nouvelle position est à déterminer, mesurer la puissance des signaux Wi-Fi cette position.

(d) Trouver les k positions de la base de données qui ont les puissances de signal Wi-Fi les plus proches de celles mesurées à la nouvelle position.

(e) Utiliser la moyenne pondérée des coordonnées des k positions les plus proches pour estimer la localisation de la nouvelle position.

(f) Répéter les étapes 3 à 5 pour chaque nouvelle position à localiser.

2.2 Estimation Bayésienne

2.2.1 La théorie de Bayes

La théorie de Bayes permet de développer un outil statistique qui rend possible le calcul de la probabilité d'un événement en fonction des informations préalables. En effet, elle est un accès direct vers la loi a posteriori. Cette dernière est utile pour la géolocalisation en prenant en compte les incertitudes des mesures. La formule de Bayes permet de mettre à jour les observations en fonction de la probabilité a priori pour estimer la probabilité a posteriori de la position de l'objet.

$$p(\boldsymbol{l}|\boldsymbol{o}) = \frac{p(\boldsymbol{o}|\boldsymbol{l}) \times p(\boldsymbol{l})}{p(\boldsymbol{o})} \tag{2.1}$$

avec

— $p(\boldsymbol{l}|\boldsymbol{o})$ la distribution a posteriori de la localisation sachant les observations.

— $p(\boldsymbol{o}|\boldsymbol{l})$ la distribution conditionnelle de l'observation sachant la position de l'objet, c'est-à-dire la vraisemblance

— $p(\boldsymbol{l})$ la distribution a priori de la position de l'objet

— $p(\boldsymbol{o})$ la distribution marginale de l'observation.

Étant donné l'absence d'expression de la vraisemblance, les noyaux sont utilisés pour s'en approcher.

Les fonctions noyaux expriment la distribution conditionnelle de l'observation connaissant la position de l'objet. Dans l'équation de Bayes, la vraisemblance est supposée être une grandeur continue. Cependant, avec une base de données comme support pour les calculs de probabilités, seules des données discrètes sont disponibles. Par conséquent, la méthode des noyaux est utilisée pour approximer cette vraisemblance continue.

2.2.2 Les fonctions noyaux

L'estimation par noyaux est une méthode couramment utilisée pour estimer la densité de probabilité d'une variable aléatoire à partir d'un échantillon de données. Cette méthode non-paramétrique utilise des fonctions de pondération appelées noyaux, qui interviennent dans l'estimateur par noyau.

Le principe de l'estimation par noyaux consiste à superposer une fonction noyau centrée en chaque point de l'échantillon, puis à sommer ces fonctions pour obtenir une estimation de la densité de probabilité continue. Cette approche permet d'estimer une densité continue à partir d'un ensemble fini d'observations, sans supposer de distribution paramétrique et en offrant une plus grande flexibilité dans la modélisation de la densité de probabilité.

Le choix du noyau est crucial dans cette méthode et doit être symétrique et positif, de sorte que la densité intégrée soit égale à 1. Les fonctions noyau couramment utilisées incluent la fonction gaussienne, la fonction uniforme et la fonction Epanechnikov.

Pour chaque observation dans l'échantillon, on superpose une copie de la fonction noyau centrée en cette observation. La somme de ces fonctions noyau donne une estimation de la densité de probabilité sous-jacente. Le choix du noyau a une influence sur la forme de la densité estimée, par exemple, un noyau Gaussien donnera une estimation lisse de la densité, tandis qu'un noyau uniforme donnera une estimation plus rugueuse.

L'estimation par noyaux est particulièrement utile lorsque la distribution sous-jacente de la variable aléatoire n'est pas connue ou n'est pas facilement modélisable.

Dans le contexte de la problématique de localisation par les signaux Wi-Fi, le choix du noyau revêt une grande importance pour réaliser une estimation précise. Trois des noyaux les plus couramment utilisés ont été comparés : le noyau Gaussien, le noyau triangle et le noyau d'Epanechnikov. Des critères tels que la précision, la simplicité d'implémentation et la robustesse ont été évalués pour chaque noyau. À l'issue de cette analyse comparative, le noyau Gaussien a été sélectionné en raison de sa précision supérieure et de sa simplicité d'implémentation par rapport aux deux autres noyaux.

Le noyau Gaussien est défini par l'équation suivante, qui décrit le noyau gaussien utilisé dans l'estimation par noyau. La distance est calculée entre le vecteur d'observation à une position donnée et chaque observation individuelle dans l'échantillon.

$$K_G(\boldsymbol{o}; \boldsymbol{o_i}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\|\boldsymbol{o}-\boldsymbol{o_i}\|^2}{2\times\sigma^2}} \tag{2.2}$$

[4] $\|v\|$

2.2.3 Application de l'estimation Bayésienne à la localisation par Wi-Fi

La formule de Bayes sera utilisée pour déterminer la loi a posteriori $p(\boldsymbol{l}|\boldsymbol{o})$.

La loi a posteriori est donnée par l'équation (**??**) rappelée ci-dessous.

L'estimateur du maximum a posteriori (MAP), noté $\hat{l}$ dans la suite, sera utilisé. Sa définition consiste à estimer la valeur la plus probable d'une variable aléatoire inconnue en utilisant les informations à priori et les observations disponibles. En d'autres termes, le MAP cherche à trouver la valeur de la variable aléatoire qui maximise la loi a posteriori, c'est-à-dire la loi de la variable sachant les observations [3].

Afin de modéliser mathématiquement cette méthode, la notation précédemment citée **??** sera utilisée pour introduire le vecteur d'observation à une position donnée et chaque observation individuelle :

$$\boldsymbol{o} = \begin{bmatrix} o_1 & \dots & o_p \end{bmatrix} \tag{2.3}$$

avec $o_i \in \mathbb{R}^R$ représentant les puissances en dB des R routeurs détectables du point de mesure d'indice i sur les P points de mesures enregistrés.

$$\boldsymbol{l} = [l_1 ... l_p] \tag{2.4}$$

avec $l_i \in \mathbb{R}^2$ représentant les coordonnées géographiques du point de mesure d'indice i sur les P points de mesures enregistrés.

D'après la formule de Bayes, la probabilité s'exprime de la façon suivante :

$$p(\boldsymbol{l}|\boldsymbol{o}) = \frac{p(\boldsymbol{o}|\boldsymbol{l}) \times p(\boldsymbol{l})}{p(\boldsymbol{o})} \tag{2.5}$$

$\Leftrightarrow$

$$p(\boldsymbol{l}|\boldsymbol{o}) \propto p(\boldsymbol{o}|\boldsymbol{l}) \tag{2.6}$$

La loi de probabilité $p(\boldsymbol{l})$ est choisie comme étant uniforme. Cette loi a priori pourrait également être utilisée si une position GNSS ou toute autre source externe d'information de positionnement était disponible.

En utilisant l'estimation par noyau gaussien, décrite précédemment dans l'équation (2.2), permet d'obtenir :

$$p(\boldsymbol{l}|\boldsymbol{o}) \propto p(\boldsymbol{o}|\boldsymbol{l}) \tag{2.7}$$

$\Leftrightarrow$

$$p(\boldsymbol{o}|\boldsymbol{l_j}) \propto K_{Gauss}(\boldsymbol{o_r}; \boldsymbol{o_{jr}}) \tag{2.8}$$

$\Leftrightarrow$

$\forall l_j \in L$

$$p(o|l_j) = \frac{1}{\sigma\sqrt{2\pi}} \times e^{-\frac{(\boldsymbol{o_r}-o_{jr})^2}{2\sigma^2}} \tag{2.9}$$

$\Leftrightarrow$

$$p(o|l_j) \propto \sum_{r=1}^{R} e^{-\frac{(\boldsymbol{o_r}-o_{jr})^2}{2\sigma^2}} \tag{2.10}$$

De cette façon, il est possible de déduire que la valeur de j pour laquelle $p(\boldsymbol{o}|l_j)$ est maximale correspond à l'indice de la mesure où le vecteur $\boldsymbol{o_j}$ présente la plus grande similarité avec le vecteur observation $\boldsymbol{o}$. C'est donc le point où il est le plus probable de se trouver.

Une fois la mesure avec le maximum de la loi a posteriori déterminée, on peut supposer que le secteur s_j auquel elle appartient est le secteur où il est le plus probable que l'on se trouve. (voir 1.2 pour plus de détails).

Chapitre 3

Carte de distribution de probabilité de présence du signal et comparaison des modèles

3.1 Construction d'une carte représentant la probabilité de présence par secteurs

Dans cette partie, le tracé d'une carte représentant la probabilité de présence par secteurs dans la zone étudiée est abordé, en utilisant les algorithmes de Bayes et *K-NN*. Cette carte permet d'identifier les secteurs où il est le plus probable de se trouver, en se basant sur les observations des puissances des signaux Wi-Fi.

Pour commencer, les algorithmes de Bayes et *K-NN* sont appliqués en utilisant les observations des puissances des signaux Wi-Fi et les probabilités a priori de présence dans chaque secteur. L'estimation des probabilités de présence dans chaque secteur est obtenue à l'aide de la formule de Bayes, comme expliqué précédemment.

Une fois que les probabilités de présence ont été estimées pour tous les secteurs, les secteurs sont classés en fonction de ces probabilités. Les secteurs ayant les probabilités les plus élevées sont considérés comme les plus probables en termes de présence. Pour représenter visuellement ces probabilités, les secteurs sont colorés en rouge pour ceux où la probabilité de présence est la plus élevée, et en orange pour ceux où la probabilité est également élevée.

En superposant cette carte sur la représentation de la zone étudiée, une visualisation graphique des secteurs les plus probables où l'utilisateur pourrait se trouver est obtenue.

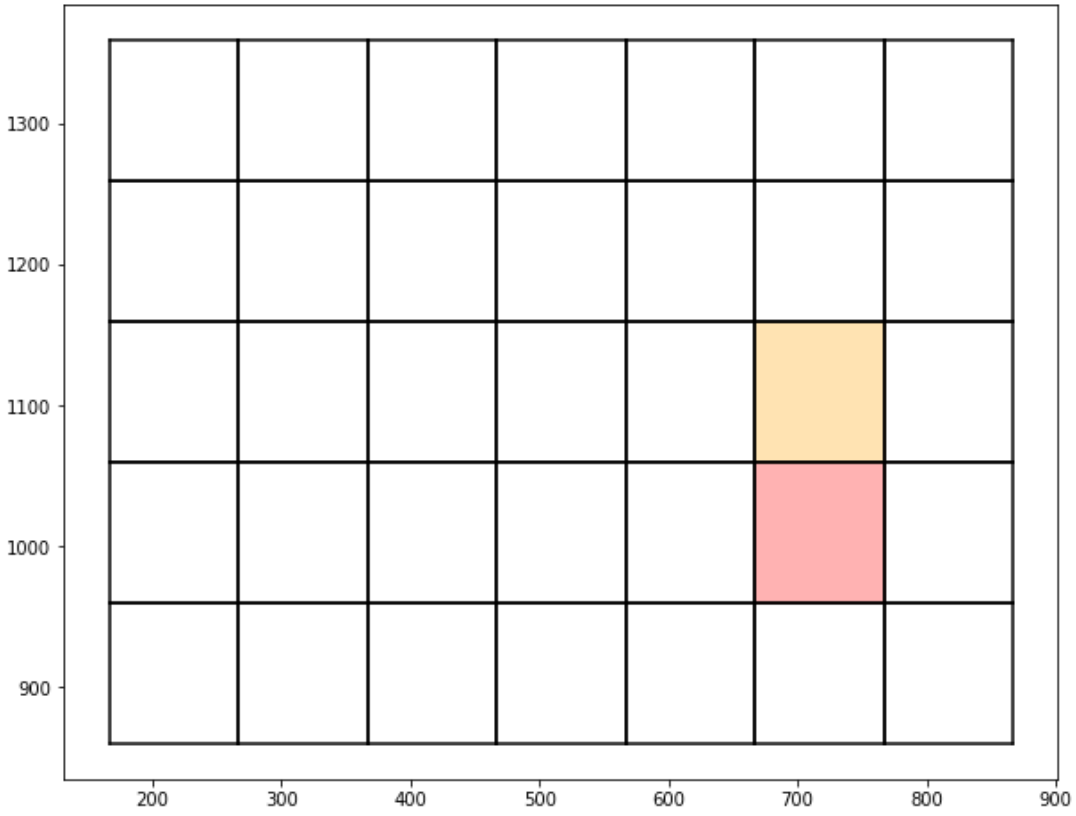

FIGURE 3.1 – Cartographie de la cour principale de l'ENAC sous forme de secteurs, avec le secteur le plus probable en rouge et les secteurs présentant une grande probabilité en orange

La Figure 3.1 présente un exemple de carte de probabilité de présence dans laquelle les secteurs les plus probables sont coloriés en rouge et en orange. Cette carte permet de localiser les zones où il est le plus probable de se trouver en se basant sur les observations des puissances des signaux Wi-Fi.

Dans la prochaine sous-section, une comparaison des performances des algorithmes de Bayes et *K-NN* en termes de probabilité de présence estimée sera réalisée. Enfin, des pistes d'amélioration seront discutées.

3.2 Comparaison des performances

3.2.1 Performances de Bayes

Malheureusement, la similitude des vecteurs d'observations entre les différents points de mesure ne permet pas d'utiliser l'algorithme de Bayes de manière optimale. En d'autres termes, la puissance des signaux WiFi émis par les BSSID ne varie pas suffisamment sur la surface de la cour principale pour permettre une localisation précise en termes de secteurs. Malgré une base de données comprenant plus de cent points de mesure, environ 20 pourcents de succès seulement est obtenu. Cela signifie que dans 80 pourcents des cas, l'algorithme attribue un secteur différent de la position réelle.

En réalisant une interpolation sur la surface de mesure des valeurs caractérisant la vraisemblance entre les signaux reçus et ceux de chaque point mesuré en amont, il a été conclu que les caractéristiques des données disponibles ne permettent pas d'exploiter au mieux cette méthode.

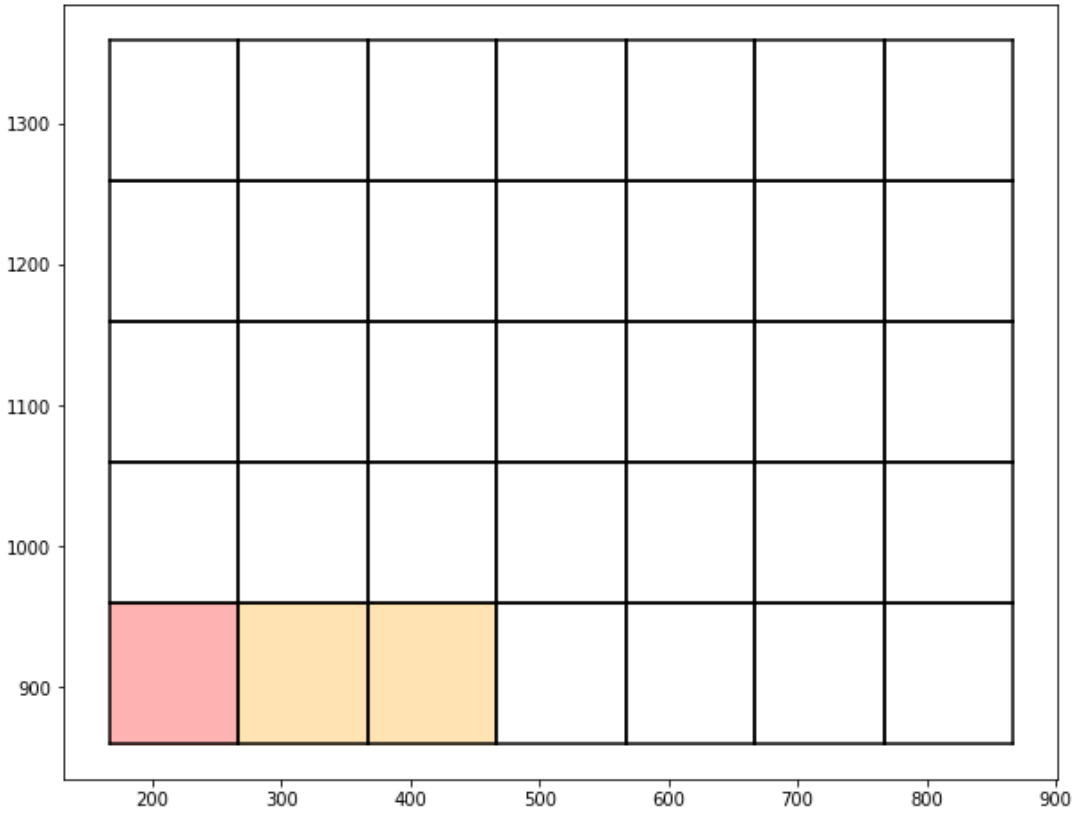

FIGURE 3.2 – Cartographie de la cour principale de l'ENAC sous forme de secteurs, avec le secteur le plus probable en rouge et les secteurs présentant une grande probabilité en orange

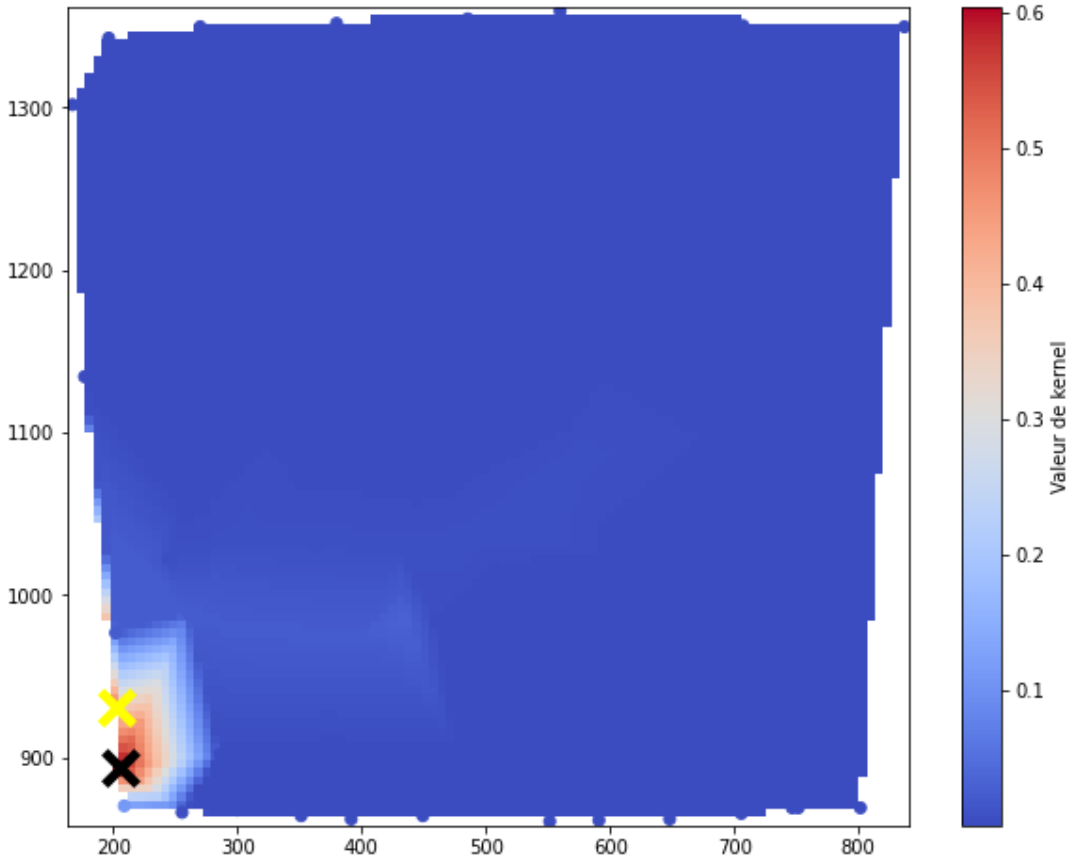

FIGURE 3.3 – Cartographie des valeurs de vraisemblance entre le signal reçu et les signaux pré-enregistrés, rendues continues par interpolation

La croix noire indique la position du maximum de vraisemblance, tandis que la croix jaune représente la position réelle de l'utilisateur de l'algorithme.

On peut observer que dans ce cas, l'algorithme est efficace. Toutefois, ce résultat est rare, car il correspond à une situation extrême où la localisation se situe à une extrémité de la zone cartographiée. Les puissances des signaux varient considérablement en raison de la position spécifique de l'utilisateur en bordure de la zone étudiée.

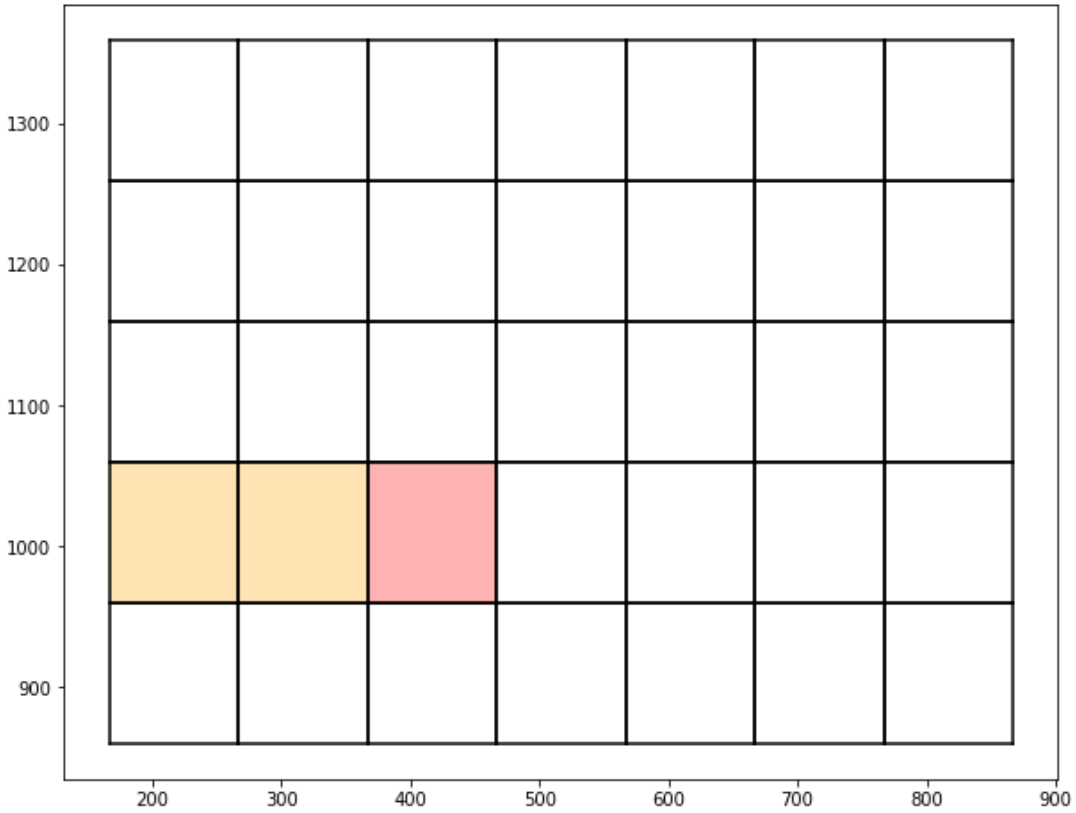

FIGURE 3.4 – Cartographie de la cour principale de l'ENAC sous forme de secteurs, avec le secteur le plus probable en rouge et les secteurs présentant une grande probabilité en orange

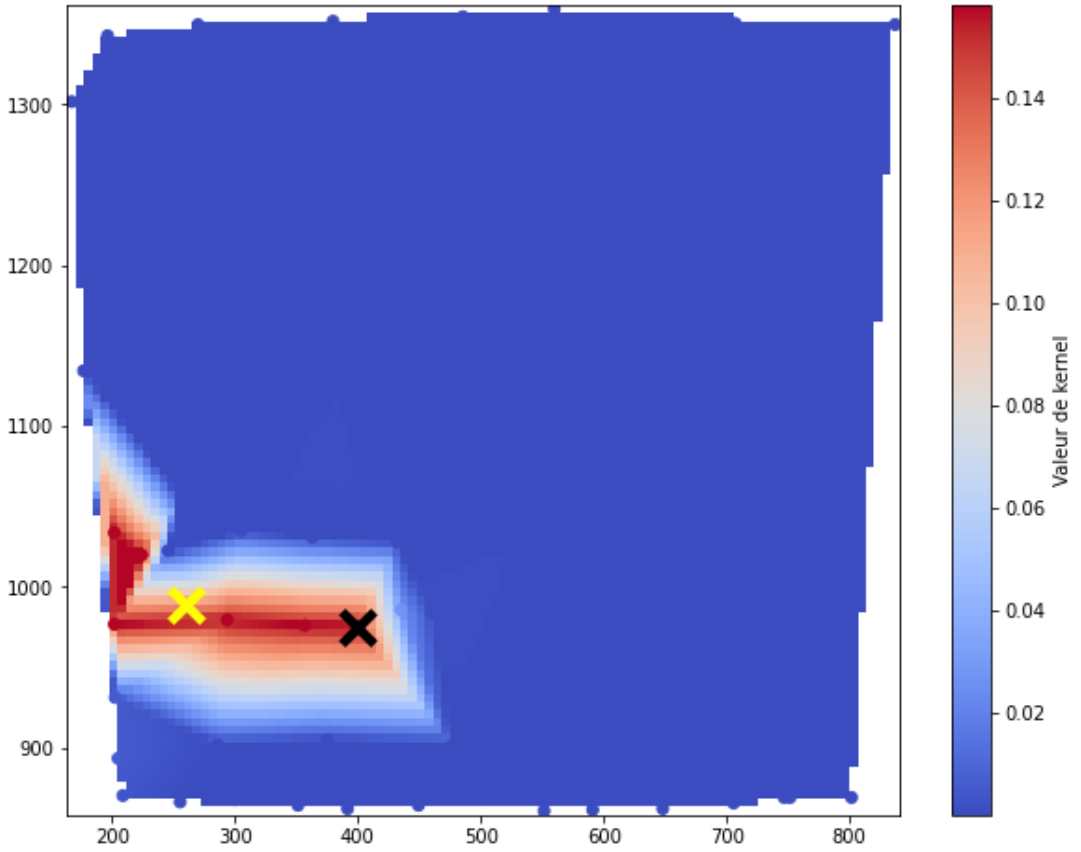

FIGURE 3.5 – Cartographie des valeurs de vraisemblance entre le signal reçu et les signaux pré-enregistrés, rendues continues par interpolation

Voici maintenant un cas plus général où la zone de forte vraisemblance est plus étendue. On remarque que l'algorithme se positionne dans le secteur représenté en rouge, tandis que notre position réelle se trouve dans le secteur orange à gauche (voir Figure 3.4), comme l'indique la croix jaune sur la deuxième carte (voir Figure 3.5). La zone de forte probabilité, indiquée en rouge sur la carte de droite, couvre ainsi six vecteurs.

On peut calculer la distance entre ces deux points (les croix noire et jaune sur la Figure 3.5) sur un échantillon de plusieurs positions d'observations afin d'obtenir un ensemble de cartes similaires à cette deuxième carte, ce qui permettrait d'estimer un cercle de précision autour de la position réelle. C'est grâce à cette approche qu'il a été possible d'estimer que la méthode de l'algorithme de Bayes permettrait d'obtenir une localisation avec une erreur ne dépassant pas 12,7 mètres. Cependant, étant donné que la surface de la cour est de 5200 m^2, soit 84 mètres de long sur 62 mètres de large, il est difficile de se fier à cette méthode dans un contexte où une localisation très précise est nécessaire.

3.2.2 Performances de *K-NN*

L'algorithme *K-NN* ne permet pas d'estimer la probabilité de présence dans chacun des secteurs, il va en revanche directement pouvoir proposer une estimation de l'emplacement du point étudié que l'on peut alors comparé à l'emplacement réel si l'on utilise un point connu de la base de données.

Comme expliqué Chapitre 2, l'algorithme doit entraîner son modèle de prédiction. Les données que l'on a à notre disposition vont alors être scindées en un ensemble d'entraînement sur lequel l'algorithme va apprendre, et un ensemble test qui va permettre d'analyser les performances de ce dernier. En effet en procédant ainsi, il devient possible d'obtenir une comparaison entre les coordonnées réelles et les coordonnées prédites.

Dans un premier temps on choisit arbitrairement 80% de la BDD pour l'entraînement, et un k=3. Sur les 20% on va demander à l'algorithme de prédire la localisation des points à partir des puissances reçues. Les prédictions apparaissent en rouge, on affiche alors également les points réels en bleu que l'on relient à leur prédiction par un simple trait. On obtient le résultat suivant :

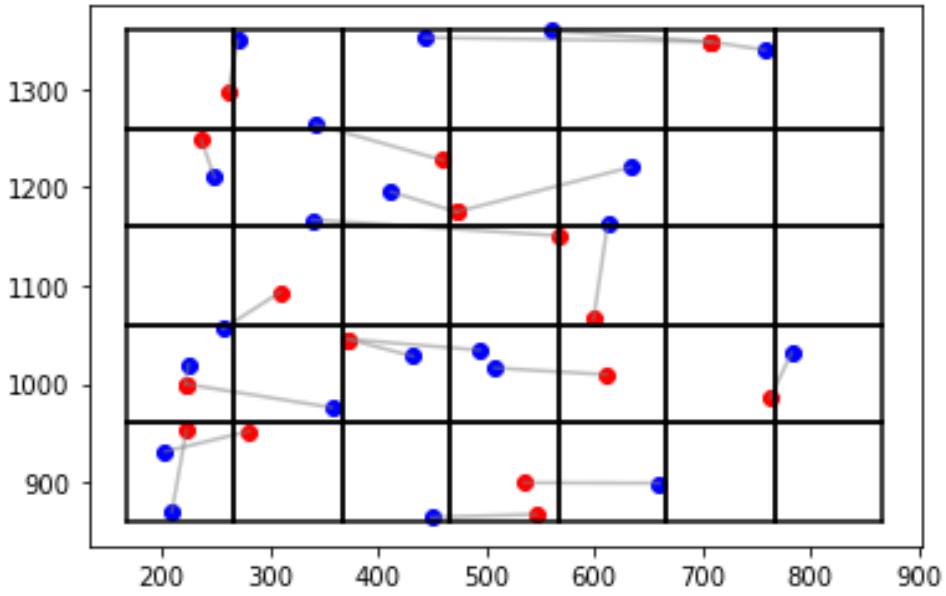

FIGURE 3.6 – Coordonnées prédites (points rouges) relié aux coordonnées réelles (points bleus) pour un échantillon d'entraînement de 80% de la BDD et avec k=3

On peut remarqué que des nombreux points ne sont pas dans le même secteur que le point réel qui leur est associé. En faisant tourner l'algorithme sur une centaine de tirage aléatoire on trouve une probabilité moyenne de 10,2%. Il est également possible d'obtenir le RMSE de tous ces points. De même sur une centaine de tirage aléatoires on trouve un RMSE de 78,4 sur notre grille, ce qui correspond à environ 9,82m en réel.

Il est alors possible de jouer sur le nombre de plus proches voisins ainsi k que sur la taille de l'échantillon test pour établir quels sont les paramètres optimaux pour notre base de données.

On obtient alors les résultats suivants :

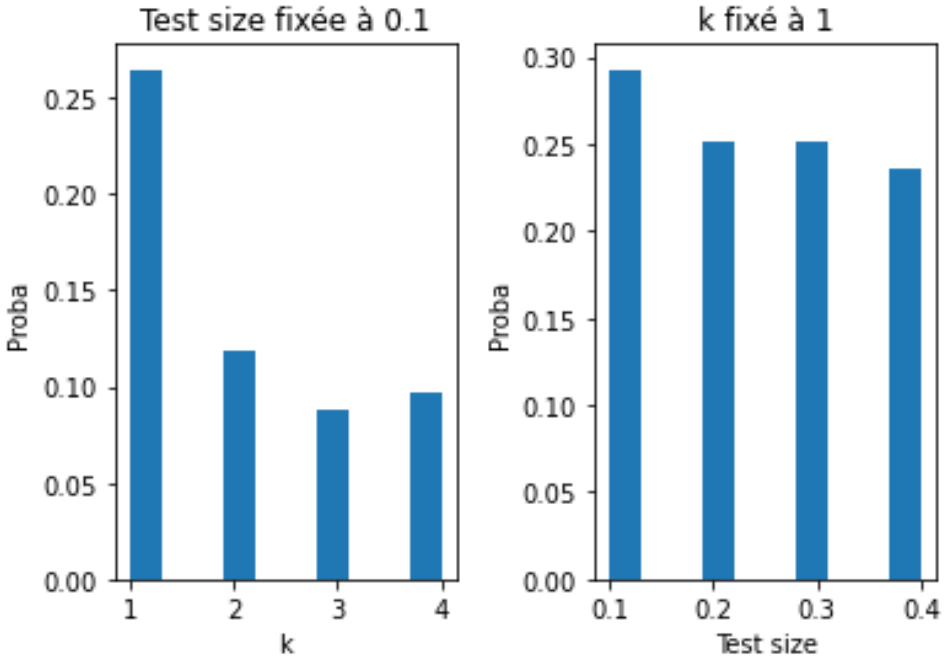

FIGURE 3.7 – Probabilité que les points prédits soient dans le même secteurs que les points réels, à taille d'échantillon test puis à k fixé pour 100 tirages aléatoires différents

Il apparaît nettement que le k optimal est $k = 1$, et également que la taille d'échantillon test optimale est de 10% soit 90% de la BDD pour l'entraînement.

En fixant ses paramètres on obtient alors le graphique suivant :

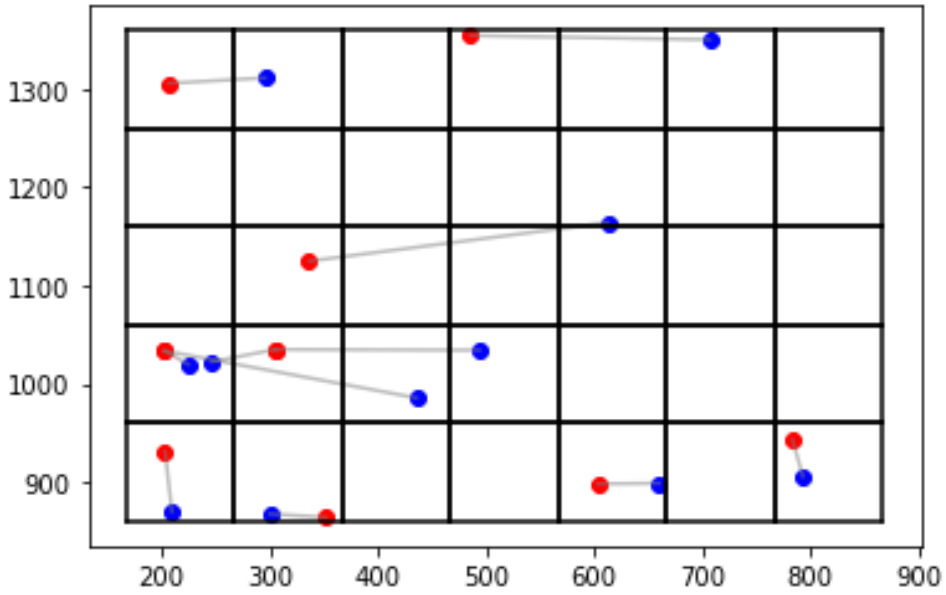

FIGURE 3.8 – Coordonnées prédites (points rouges) relié aux coordonnées réelles (points bleus) pour un échantillon d'entraînement de 90% de la BDD et avec k=1

Enfin en faisant fonctionner l'algorithme sur 1000 tirages aléatoires on obtient une probabilité moyenne d'être dans le bon secteur de 27,4% avec un RMSE moyen de 10,41m (en sachant qu'il oscille entre 4,14m et 17,93m).

Un dernier élément à analyser pour étudier les performances de l'algorithme *K-NN* est la comparaison des valeurs du RMSE pour les différentes valeurs de k et les différentes valeurs de taille d'échantillon d'entraînement.

En plus de la moyenne des RMSE il est également intéressant d'obtenir les valeurs minimales et maximales de ces RMSE pour avoir une meilleur idée des performances de l'algorithme dans le cas de valeurs critiques.

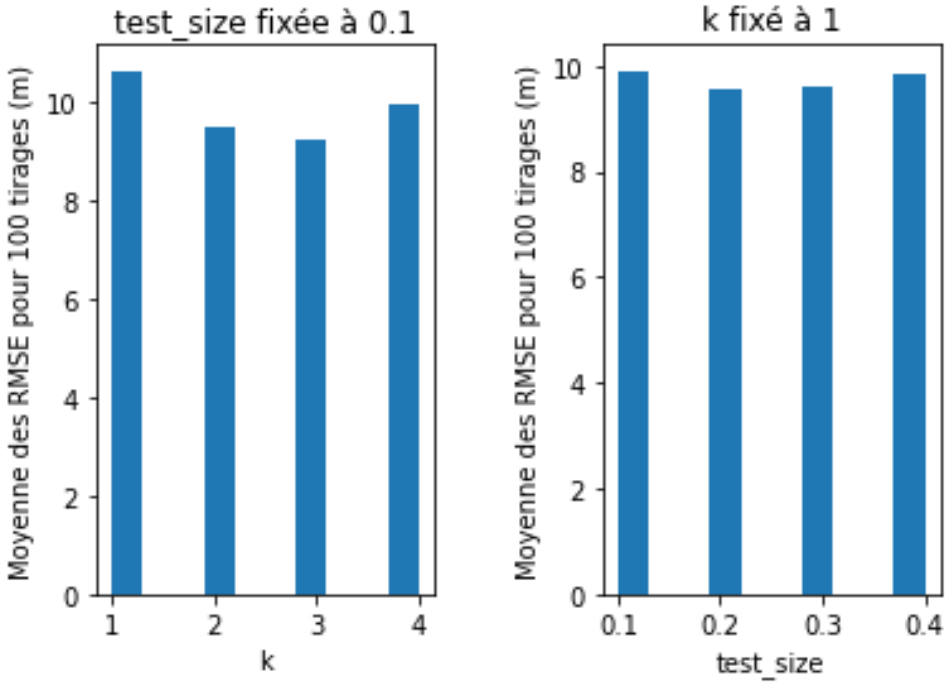

FIGURE 3.9 – RMSE moyen pour 100 tirages aléatoires selon les valeurs de k et des différentes tailles d'échantillons test

On peut donc constater que le choix de ces 2 paramètres à peu d'influence sur le RMSE, i.e. sur la précision en position de l'algorithme.

Ainsi il reste pertinent de maintenir le choix de $k = 1$ et d'un échantillon d'entraînement représentant 90% de la BDD qui a été utilisée. Il faut néanmoins garder en tête que la taille de la BDD peut avoir un impact sur le choix des paramètres.

3.2.3 Comparaison des performances

Ainsi les algorithmes de Bayes et *K-NN* présentent à la fois des similitudes et des différences dans leurs approches et leurs performances.

En effet les performances des deux algorithmes peuvent être mesurées à l'aide de métriques similaires, telles que la précision, le taux d'erreur, le taux de réussite, ou encore le Root Mean Square Error (RMSE) pour évaluer l'erreur de localisation.

Néanmoins ils différents sur de nombreux points :

— **Approche de prédiction :** Dans l'algorithme de Bayes, la prédiction est basée sur la comparaison des observations des signaux Wi-Fi avec des modèles pré-enregistrés pour chaque secteur. L'estimation de la position se fait en identifiant le secteur le plus probable en fonction des observations. En revanche, l'algorithme *K-NN* utilise une approche de voisinage, où la prédiction de la position se fait en trouvant les k points les plus proches dans l'espace des caractéristiques et en attribuant la classe la plus fréquente parmi ces k voisins.

— **Capacité d'apprentissage :** L'algorithme *K-NN* nécessite une étape d'apprentissage préalable où il ajuste son modèle en utilisant les données d'entraînement. En revanche, l'algorithme de Bayes n'a pas de phase d'apprentissage explicite, il estime directement les probabilités en utilisant les informations disponibles.

— **Sensibilité aux données :** Les performances des deux algorithmes peuvent varier en fonction des caractéristiques des données. L'algorithme de Bayes peut être moins performant lorsque les observations des signaux Wi-Fi ne varient pas suffisamment entre les secteurs, ce qui limite sa capacité à distinguer les positions avec précision. En revanche, l'algorithme *K-NN* peut être sensible à la présence de valeurs aberrantes ou à la répartition inégale des observations.

Ces différences peuvent entraîner des performances variables dans différentes situations, et il est important de prendre en compte ces facteurs lors du choix de l'algorithme approprié pour une tâche de localisation spécifique.

Des améliorations supplémentaires peuvent être envisagées pour optimiser davantage les résultats, comme l'ajout de techniques de pré-traitement des données ou l'exploration d'autres algorithmes de localisation.

En conclusion, la comparaison des performances des algorithmes de Bayes et *K-NN* pour la localisation par signaux Wi-Fi met en évidence des limites dans l'utilisation de l'algorithme de Bayes en raison de la similarité des vecteurs d'observations. En revanche, l'algorithme *K-NN* offre des résultats plus prometteurs en termes de localisation, avec une probabilité de présence estimée améliorée en ajustant les paramètres appropriés. Cependant, des efforts supplémentaires sont nécessaires pour améliorer davantage la précision et la fiabilité de la localisation par signaux Wi-Fi.

Conclusion

En conclusion, ce projet se concentre sur l'utilisation des signaux Wi-Fi pour le positionnement en intérieur, en particulier dans des bâtiments ou des zones géographiques avec de nombreux obstacles où les technologies satellites traditionnelles sont moins efficaces. L'algorithme K-Nearest Neighbors (*K-NN*) a été choisi pour la régression dans le cadre de ce projet, en raison de son efficacité prouvée dans le domaine de la localisation par signal Wi-Fi. L'algorithme attribue une étiquette à une nouvelle observation en se basant sur la classe majoritaire parmi ses k plus proches voisins. Le choix de la valeur de k est crucial pour éviter le surapprentissage ou le sous-apprentissage.

L'estimation Bayésienne est également utilisée pour calculer la probabilité de la position d'un objet en fonction des observations de puissance de signal Wi-Fi. L'estimation par noyaux, en utilisant le noyau Gaussien, est une méthode non paramétrique couramment utilisée pour estimer la densité de probabilité d'une variable aléatoire à partir d'un échantillon de données.

Le projet présente la construction d'une carte représentant la probabilité de présence par secteurs dans la zone étudiée en utilisant les algorithmes de Bayes et *K-NN*. Les secteurs sont classés en fonction des probabilités de présence estimées, et les secteurs les plus probables sont visuellement identifiés.

La comparaison des performances des algorithmes de Bayes et *K-NN* révèle que l'algorithme *K-NN* offre des résultats plus prometteurs en termes de prédiction de l'emplacement, avec une probabilité moyenne de présence dans le bon secteur de 26,98% et un RMSE moyen de 10,3 mètres contre 15,38% et un RMSE moyen de 12,7 mètres. Cependant, il est souligné que les performances peuvent varier en fonction des caractéristiques des données, et il est crucial de choisir l'algorithme approprié en fonction de la tâche de localisation spécifique.

En somme, ce projet a exploré l'utilisation des signaux Wi-Fi pour la localisation en intérieur et a comparé les performances des algorithmes de Bayes et *K-NN*. Des améliorations supplémentaires sont suggérées, telles que l'utilisation de techniques de pré-traitement des données ou l'exploration d'autres algorithmes de localisation, afin d'optimiser les résultats et de répondre aux exigences spécifiques de chaque cas d'utilisation.

List of acronyms

ENAC Ecole Nationale de l'Aviation Civile (French Civil Aviation University)
SIGNAV Signal Processing and Navigation Laboratory (Laboratoire de traitement des signaux et de navigation)
BSSID Basic Service Set Identifier (Identifiant de l'ensemble de services de base)

Bibliographie

[1] Michel Crucianu, Marin Ferecatu, Nicolas Thome, Nicolas Audebert - Cnam, "*Méthodes à noyaux"*, `https://cedric.cnam.fr/vertigo/Cours/ml2/coursMethodesNoyaux.html`, 2022

[2] Bruce E. Hansen - University of Wisconsin, "*Lecture Notes on Nonparametrics*", `https://www.ssc.wisc.edu/~bhansen/718/NonParametrics1.pdf`, 2009

[3] Jerry Cain, "MAP", `https://web.stanford.edu/class/archive/cs/cs109/cs109.1216/lectures/22_map.pdf`, 2021

[4] Teemu Roos, Petri Myllymäki, Henry Tirri, Pauli Misikangas, and Juha Sievänen, "*A Probabilistic Approach to WLAN User Location Estimation*", 2002.

[5] E. Fix and J.L. Hodges, "*An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation*", 1951.

[6] Cover T. and Hart P., "*Nearest Neighbor Pattern Classification*" - IEEE Transactions on Information Theory, 13(1), 21-27, 1967.

[7] Altman N. S., "*An introduction to kernel and nearest-neighbor nonparametric regression.*" - The American Statistician, 46(3), 175-185, 1992.

[8] Liu J., Chaudhary S., "*WiFi fingerprint-based indoor positioning : Recent advances and comparisons.*" - IEEE Communications Surveys and Tutorials, 14(2), 515-527, 2012.

[9] Kaemarungsi K., Krishnamurthy P., "*Properties of indoor received signal strength for WLAN location fingerprinting.*" - Proceedings of the 1st international conference on Mobile systems, applications, and services, 14-23, 2004.