# Location Fingerprinting for Enhanced Positioning of Mobile Terminals in Cellular Networks

Azin Arya

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2012-ENST

EDITE - ED 130

# Doctorat ParisTech

# T H È S E

**pour obtenir le grade de docteur délivré par**

# TELECOM ParisTech

## Spécialité “Informatiques et Réseaux”

*présentée et soutenue publiquement par*

**Azin ARYA**

Paris, le 30 septembre 2011

# Localisation à base d’Empreintes Radios: Méthodes Robustes de Positionnement pour les Terminaux Cellulaires

Directeur de thèse: **Philippe GODLEWSKI**

**Jury**

| | |
|---|---|
| **M. Dirk SLOCK**, Professeur, Eurecom - France | Rapporteur |
| **M. Bijan JABBARI**, Professeur, George Mason University - USA | Rapporteur |
| **M. François BACCELLI**, Professeur, INRIA - France | Examinateur |
| **M. Stéphan CLÉMENÇON**, Professeur, Télécom ParisTech - France | Examinateur |
| **M. Alexandre CAMINADA**, Professeur, UTBM - France | Examinateur |
| **M. François VINCENT**, SFR - France | Invité |

**TELECOM ParisTech**
école de l’Institut Télécom - membre de ParisTech

THÈSE



# Location Fingerprinting for Enhanced Positioning of Mobile Terminals in Cellular Networks

by

**Azin ARYA**

Submitted for the qualification of

**Doctor of Philosophy of Télécom ParisTech**

**Examining committee:**

| | |
|---|---|
| **Prof. Dirk SLOCK** | **Reviewer** |
| **Prof. Bijan JABBARI** | **Reviewer** |
| **Prof. François BACCELLI** | **Examiner** |
| **Prof. Stéphan CLÉMENÇON** | **Examiner** |
| **Prof. Alexandre CAMINADA** | **Examiner** |
| **Mr. François VINCENT** | **Invited** |
| **Prof. Philippe GODLEWSKI** | **Thesis Adviser** |

Paris, 30 September 2011

*To My Lovely Family*

# Acknowledgement

It is a pleasure to thank those who have helped and inspired me during my doctoral study.

First and foremost, I would like to thank my adviser professor Philippe Godlewski, for his guidance, advice, and patience, throughout these three years. His truly scientist intuition inspired and enriched my growth as a student, researcher, and an independent thinker. Philippe, I thank you.

I would like to gratefully thank my industrial advisers in SFR, Mr. Philippe Mellé and Mr. François Vincent, who helped me to enrich this thesis thanks to their valuable viewpoints and experiences.

I was honored that Professor Bijan Jabbari at George Mason University in USA and Professor Dirk Slock at Eurecom in France, accepted to review and evaluate this study. I sincerely thank Professor François Baccelli at INRIA, Professor Stéphan Clémençon at Télécom ParisTech, and Professor Alexandre Caminada at l'Université de Technologie de Belfort-Monbéliard, for their participation in the jury as examiners. In particular, I would like to dedicate special thanks to Professor Clémençon, for sharing valuable insights and constructive ideas to the field of Machine Learning. I am also thankful to Professor Caminada, for offering me the occasion to present my thesis findings at l'Université de Technologie de Belfort-Monbéliard.

I was delighted to interact with Dr. Marine Campedel, at Signal and Image department of Télécom ParisTech. I thank her for granting me kindly her time, and sharing valuable insights in the relevance of this study to modern fields of data analysis.

I would like to acknowledge professor Philippe Martins and professor Ahmed Serhrouchni at Télécom Paris, for their permanent advice and support during my study.

The years spent in Télécom Paris would not have been so wonderful, without my good friends at the Network and Computer Science Department. I will surely miss all our coffee breaks, and our extra-scholar activities! I take the opportunity to thank in particular Michael, Fabien, and Simon, for their help in editing the French abstract.

I would like to thank also all my friends and colleges in the Research and Development department of SFR, whose friendship and humor made it a convivial place to work.

Thanks to Atoosa.

Ten years ago we were friends ; now we are sisters. Thank you for every nice moment that we shared.

At last but not least, my deepest gratitude goes to my lovely family, in particular my parents. I had the chance to be born in a family that offered me endless love and care; I believe that this was the greatest chance of my life.

I love you all so much.

# Table of contents

# Résumé

Ces dernières années, les services basés sur la position (*Location Based Services*, LBS) ont attiré l'attention des opérateurs mobiles et autres acteurs des télécommunications. Ce genre de services peut s'appliquer dans différents contextes, par exemple la navigation, la géo-publicité, les réseaux sociaux, etc.([1]).

Différentes technologies de localisation peuvent être utilisées dans les LBS. Deux classes majeures des infrastructures de localisation consistent en les systèmes satellitaires et les réseaux cellulaires. Les systèmes satellitaires (comme GPS, ou plus généralement GNSS) peuvent fournir une localisation assez précise dans les environnements ouverts, avec une précision de l'ordre de quelques mètres. Néanmoins, ces systèmes possèdent des inconvénients comme des performances dégradées dans les zones urbaines denses, où il n'y a pas de vue directe vers le ciel. De plus les systèmes satellitaires exigent une consommation élevée d'énergie au niveau du mobile, ce qui diminue considérablement l'autonomie du terminal. Afin de surmonter ces problèmes, sont développé des méthodes de localisation basées sur les réseaux cellulaires. C'est dans ce contexte que s'inscrit cette thèse.

### Localisation dans les réseaux cellulaires

Dans un réseau cellulaire, la trace de tous les terminaux allumés est suivie en permanence par le réseau. Pendant la communication, les terminaux sont suivis à la cellule-près (quelques centaines de mètres dans les zones urbaines, et quelques kilomètres dans les zones rurales). Lorsque les mobiles sont en mode veille, ces derniers sont suivis au niveau de la zone de localisation (Location Area). La zone de localisation consiste en un groupe de quelques dizaines de cellules, définie et configurée par l'opérateur.

En résumé, tous les terminaux allumés dans un réseau cellulaire sont intrinsèquement localisés, avec une précision qui dépend de leur statut. Cette précision est suffisante pour certains services, mais elle ne l'est pas pour des applications critiques comme le positionnement des appels d'urgence. Par conséquence, des méthodes plus avancées sont développées pour positionner les terminaux dans les réseaux cellulaires, comprenant les techniques *Uplink Time-difference Of Arrival* (U-TDOA), *Enhanced Observed Time Difference* (E-OTD), etc.

## Localisation basée sur les empreintes radios

L'un des intérêts des opérateurs mobiles dans le contexte de LBS est d'offrir aux abonnés une localisation précise, durable et d'un coût peu élevé. Actuellement, le GPS et le Cell-ID sont les méthodes les plus utilisées dans les LBS. Cependant ces techniques ne satisfont pas les critères mentionnés ci-dessus. Des méthodes plus avancées sont envisageables en combinant plusieurs techniques. C'est la stratégie adoptée par certains acteurs de LBS, comme Ericsson et TrusPosition. Ces entreprises proposent des solutions qui intègrent des méthodes variées, comme le GPS, Cell-ID, U-TDOA, etc. Néanmoins, ces solutions n'ont pas été adoptées par beaucoup d'opérateurs dans le monde, à cause du coût élevé d'implémentation.

Une méthode alternative pour offrir une localisation durable et abordable est la « localisation basée sur les empreintes radios » (*Location Fingerprinting*, LFP). La méthode LFP exploite les réseaux radios existants, comme les réseaux cellulaires, ou les WLANs. La méthode profite des mesures génériques qui sont disponibles à partir des interfaces radios, et permet donc une localisation à bas coûts.

Le système de LFP consiste en deux phases. Tout d'abord, pendant une "phase d'apprentissage" (*training phase*), une base de données radio est constituée sur la région considérée. Une fois que la base est construite, les mobiles peuvent entrer dans la "phase de localisation" (*localization phase*). Ici, un mobile fait des mesures de test, et sera localisé en associant ces mesures aux éléments qui sont déjà enregistrés dans la base. Pour le cas des réseaux cellulaires, la méthode de LFP permet une localisation plus précise que Cell-ID. La méthode n'exige pas une grande consommation d'énergie, car elle profite des mesures radios génériques qui se font régulièrement au sein du terminal. Etant données ces caractéristiques, la technique de LFP peut être considérée comme une solution potentielle pour fournir une localisation durable et à bas-coûts, et constitue l'axe principale de cette thèse.

## Terminologie et modélisations

Dans cette section, nous précisons les terminologies et les modélisations utilisées ci-après dans cet ouvrage.

Une "base de données radio" est un ensemble "d'enregistrements". Dans ce contexte, chaque enregistrement est constitué de deux parties: la partie de position (*location part*), et la partie radio (*radio part*). La partie de position décrit la position d'un point spécifique; la partie radio décrit quant à elle la mesure radio effectuée à cette position spécifique. La mesure radio contient plusieurs types de paramètres, disponibles à partir des interfaces radios (par exemple *Received Signal Strength* (RSS), *Timing Advance* (TA), etc.). On représente la mesure radio par le vecteur $\underline{s} \in \mathbb{R}^{D_R}$ comprenant $D_R$ éléments réels. De la même façon, la partie de position est représenté par un vecteur $\underline{x} \in \mathbb{R}^{D_G}$. Un enregistrement est donc donné par $\underline{r} = (\underline{x}, \underline{s}) \in \mathbb{R}^D$, où $D = D_G + D_R$.

Dans le cadre de cette thèse, on s'intéresse au cas des systèmes LFP basés sur les mesures RSS. Ainsi dans le texte ci-après, le vecteur $\underline{s}$ représente toujours un vecteur des valeurs de RSS.

Afin de modéliser les mesures RSS on a besoin d'un modèle de propagation radio. Un modèle classique est le modèle OSLN (*One Slop Log-Normal model*), qui décrit la perte radio (*pathloss*) comme suivan:

$$Pl_a(d) = -k + 10\alpha \log(d) + X_{sh}, \tag{1}$$

où $Pl_a$ est la perte moyenne (en dB), $k$ est un constant, $d$ est la distance, $\alpha$ est le paramètre de propagation exponentielle (*propagation exponent*), et $X_{sh}$ est une variable aléatoire log-Normal qui représente l'effet de shadowing. Le modèle OSLN ne considère aucune corrélation géographique pour l'effet de shadowing.

Le modèle radio que nous utilisons dans le cadre de cette thèse (nommé "Mondrian"), est un modèle log-Normal qui considère un certain niveau de corrélations pour la propagation dans un voisinage géographique ([2]). Cette effet est réalisé en introduisant un certain nombre de masques $\{\mu\}$ dans la région considérée $\mathcal{A}$, comme illustré dans la figure 1. Un masque est un segment associé avec un paramètre d'atténuation $a(\mu)$, qui est tiré à partir d'une distribution log-Normal.

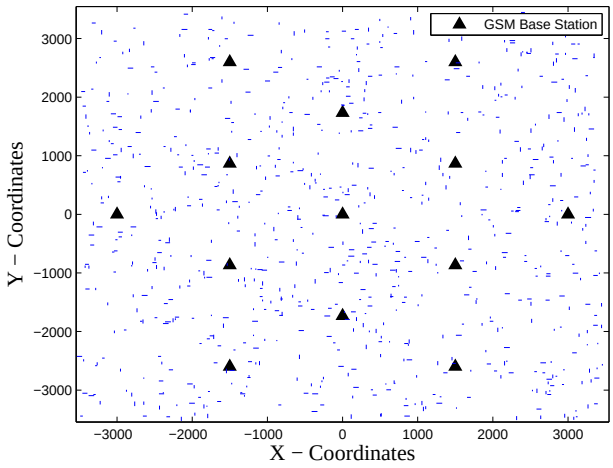



Figure 1: Les masques imitant l'effet de shadowing pour le modèle Mondrian

Pour un lien radio donné, considérons la trajectoire directe (*line of sight*) $\pi$ ; selon notre modèle, la perte radio correspondant à ce lien est modélisée comme suit:

$$PL_a(d) = -k_0 + 20 \log(d) + \sum_{\mu \in \mathcal{M}(\pi)} a(\mu) w(\mu, \pi), \tag{2}$$

où les deux premiers termes donnent la perte dans l'espace libre, $\mathcal{M}(\pi)$ est l'ensemble des masques qui croisent $\pi$, et $w(\mu, \pi)$ est un facteur de pondération. Nous remarquons qu'une fois que les masques sont tirés, le modèle devient déterministe et $PL_a(d)$ prend une valeur fixe.

Une fois que la perte radio est calculée par le modèle Mondrian, une mesure RSS instantanée s'obtiendrait comme $s = P_T - PL_a + X_m$ où $X_m \sim \mathcal{N}(0, \sigma_m^2)$ est un variable aléatoire Gaussienne, qui modélise les variations temporelles du signal.

## Une première analyse de performance

On suppose que le service de localisation est offert sur la région géographique $\mathcal{A}$, qui est couvert par le system GSM. Les cellules GSM sont considérées hexagonales, d'un rayon de 1 Km. La région $\mathcal{A}$ couvre une surface comprenant $B = 13$ cellules (comme illustré dans la figure 1).

Afin de construire la base de données radio, la région $\mathcal{A}$ est couverte par un quadrillage, comprenant $M$ zones carrées. La résolution du quadrillage $g$ est définit comme le côté de chaque carré. La base de données $\mathcal{R}$ est donc l'ensemble de $M$ enregistrements, décrit comme suit :

$$\mathcal{R} = \{(\underline{x}_m \in \mathbb{R}^{D_G}, \underline{s}_m) \in \mathbb{R}^{D_R}\}_{m=1...M}. \tag{3}$$

avec un enregistrement pour chaque zone (où $D_G = 2, D_R = B = 13$); cet enregistrement étant obtenu en moyennant plusieurs mesures brutes effectuées sur la zone correspondante..

Pendant la phase de localisation, le mobile effectue une mesure $\underline{s}'$ à la position $\underline{x}'$. Afin de localiser le terminal, on utilise la méthode de classification de KNN (*K Nearest Neighbours*). Deux types de métrique sont considérés pour implémenter la classification KNN: la distance Euclidien, et le coefficient de corrélation.

Pendant la phase de localisation, le mobile fait une mesure $\underline{s}'$ à la position $\underline{x}'$. Afin de localiser le terminal, on utilise la méthode de classification de KNN (*K Nearest Neighbours*). Deux types de métrique sont considérés pour implémenter la classification KNN: la distance Euclidien, et le coefficient de corrélation.

Une fois que le mobile est localisé, l'erreur de la localisation est donnée par $\varepsilon(\underline{x}') = \|\underline{x}' - \widehat{\underline{x}}\|$, où $\widehat{\underline{x}}$ est la position estimée pour le terminal.

Les figures 2 et 3 montrent certains résultats obtenus pendant cette première analyse de performance. En résumé, on remarque que la localisation est meilleure dans les régions denses urbaines (avec alpha fort) par rapport aux régions rurales (avec un alpha faible). De plus, on observe que l'enrichissement de la base de données (en affinant la résolution $g$) n'améliore pas forcément la précision de la localisation.

## L'analyse de cluster pour la compression de la base radio

Dans la section précédente, on a observé que l'enrichissement de la base de données n'améliore pas forcément la qualité de la localisation. D'autre part, la taille de la base est un facteur important (surtout dans les approches *mobile-based*), comme elle influence la charge de calcule, la charge de transmission, et dans un sens plus général, l'autonomie énergétique du terminal. Par conséquence, dans la littérature sont proposées des méthodes

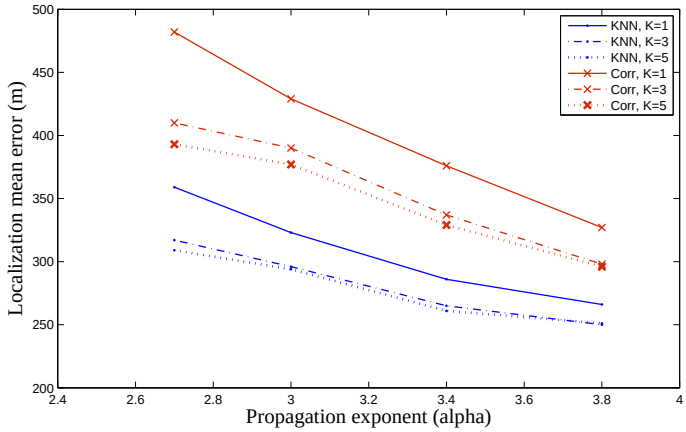
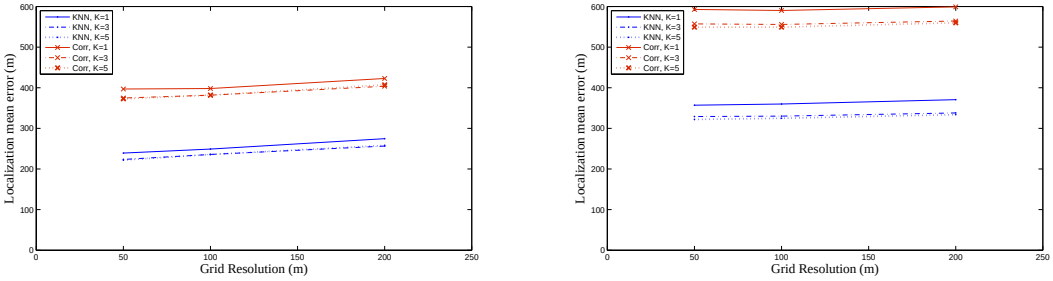



Figure 2: La précision de localisation en fonction de "propagation exponent"




(a) Scénario bruit faible

(b) Scénario bruit fort

Figure 3: Précision de la localisation en fonction de la résolution du quadrillage

qui visent à compresser la base de données radio. La plupart de ces méthodes comme *Principal Component Analysis* (PCA) et *Kernel Canonical Correlation Analysis* (KCCA), essaie de réduire la dimension de la base, en utilisant les contraintes de covariance ([3], [4] and [5]). Dans ce travail, nous proposons de réduire le nombre des enregistrements grâce aux techniques de *clustering*.

## Concept and notations

Conformément aux sections précédentes, une base de données $\mathcal{R}$ est un ensemble d'enregistrements; chaque enregistrement est représenté par $\underline{r} = (\underline{x}, \underline{s}) \in \mathbb{R}^D$. Supposons que les positions inclues dans la base sont données par l'ensemble $\chi$ :

$$\chi = \{\underline{x}_1, ..., \underline{x}_m, ..., \underline{x}_M\}. \tag{4}$$

La base radio est donc donnée par $\mathcal{R} = \{\underline{r}_m\}_{m=1...M}$.

La base de données finale $\mathcal{R}$ pourrait s'obtenir en traitant les éléments d'une base de données « initiale » ; une base radio constituée selon les mesures terrains brutes est nommée une base de données initiale $\mathcal{R}^\circ$. Un enregistrement de $\mathcal{R}^\circ$ est donné par $\underline{r}^\circ = (\underline{x}^\circ, \underline{s}^\circ)$.

Le but de ce travail est de compresser une base de données initiale $\mathcal{R}^\circ = \{\underline{r}_n^\circ\}_{n=1...N}$ en utilisant des techniques de clustering, afin d'obtenir une base de données plus compacte $\mathcal{R} = \{\underline{r}_m\}_{m=1...M}$ $(M < N)$. Pour faire cela, nous proposons une architecture comme celle illustrée dans la Figure 4, où une étape de clustering a été ajoutée dans la phase d'apprentissage. L'indice de compression $\eta$ dans ce contexte est défini par $\eta = M/N$.

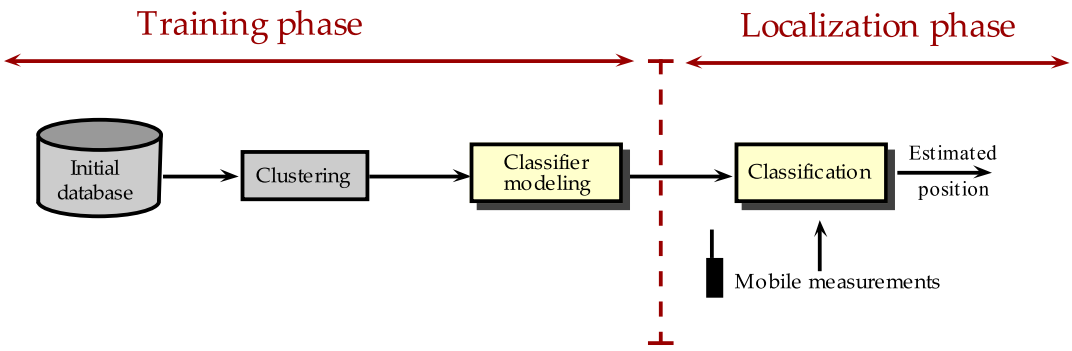



Figure 4: L'architecture proposée, comprenant l'étape de clustering

## Clustering algorithms

Supposons une base comprenant $N$ data-points $\mathcal{R}^\circ = \{\underline{r}_n^\circ\}_{n=1...N}$, dans une espace de dimension $D$ ($\underline{r}_n^\circ \in \mathbb{R}^D$). Une technique de clustering essaie de diviser $\mathcal{R}^\circ$ en $M$ $(M < N)$ sous-ensembles ou clusters, telle que les points dans chaque cluster soient similaires dans certain sens.

A ce stade de notre travail, on considère deux types d'algorithme pour réaliser l'étape de clustering pour un LFP système. Tout d'abord, on adopte un algorithme de k-means basé sur le critère de minimum de la variance intra-cluster (*minimum intra-cluster variance*). Dans cette technique, l'algorithme de clustering essaie de trouver une partition des données qui minimise la somme des variances intra-cluster. Pour l'ensemble de $\mathcal{R}^\circ$ donne, une partition pourrait être représentée par une matrice $M \times N$, $U = [u_{mn}]$, qui satisfait les critères suivants ([6]):

$$u_{mn} \in \{0, 1\}, \tag{5a}$$

$$\sum_{m=1}^{M} u_{mn} = 1; \text{ for } 1 \leq n \leq N, \tag{5b}$$

$$\sum_{n=1}^{N} u_{mn} > 0; \text{ for } 1 \leq m \leq M, \tag{5c}$$

Etant donnée la définition ci-dessus, l'algorithme de k-means essaie de minimiser la fonction d'objective suivante :

$$J_2(U, \mathcal{R}) = \sum_{n=1}^{N} \sum_{m=1}^{M} u_{mn} d^2_{E(\underline{w})}(\underline{r}^\circ_n, \underline{r}_m). \tag{6}$$

où $\mathcal{R} = \{\underline{r}_m\}_{m=1,\ldots,M}$ est l'ensemble de $M$ vecteur représentant les centroides des clusters ; $d_{E(\underline{w})}$ est une distance Euclidien pondérée, que l'on a adopté pour calculer les variances.

Comme pour le deuxième algorithme, on considère une technique hiérarchique agglomérative. Dans cette technique, on essaie de minimiser la même fonction d'objective que celle de k-means; mais l'optimisation se fait dans une façon hiérarchique. Partant de l'hypothèse que chaque vecteur dans $\mathcal{R}^\circ$ constitue un cluster, on procède en fusionnant les deux cluster qui minimise la variation dans $J_2$ à chaque étape de la procédure.

Selon les simulations effectuées, les algorithmes proposés sont assez efficaces pour compresser la base de données radio, dans le contexte des systèmes LFP. La figure montre l'erreur moyenne de la localisation en fonction de l'indice de compression, pour un scénario simulé. Comme pour la simulation décrite dans la section précédente, celle-ci est également basée sur le modèle radio Mondrian.

Selon les résultats, on observe que les techniques de clustering sont plus efficaces que la méthode simple du quadrillage pour la compression de la base de donnés radio. D'ailleurs, pour un large intervalle des valeurs d'eta, la compression de la base ne cause pas une énorme dégradation de la qualité de la localisation (ce qui n'est pas le cas pour la méthode simple de quadrillage).

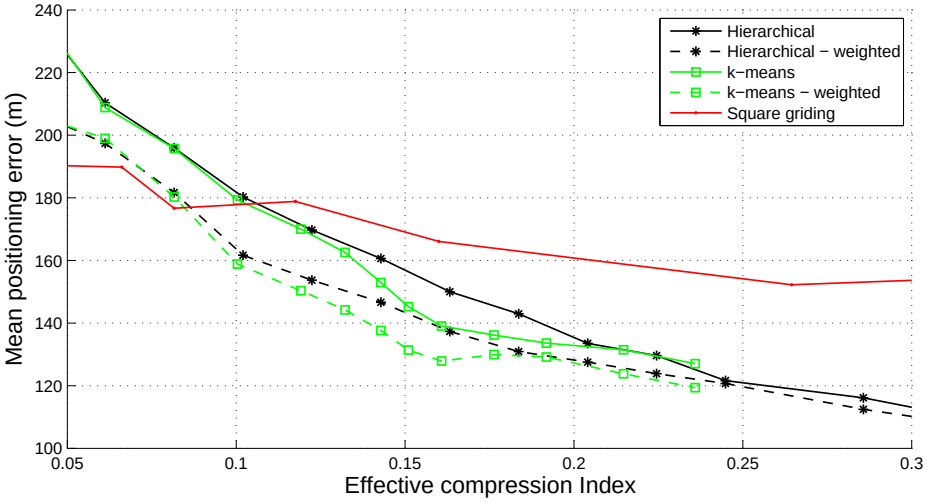



Figure 5: La performance des techniques de clustering en fonction de l'indice de compression

## Clustering de la base radio: la technique BWC (*Bloc-based Weighted Clustering*)

Dans la dernière étape de ce travail, nous avons implémenté quelques algorithmes standards de clustering pour les systèmes de LFP. Dans cette étape, on développe la méthode de *Bloc-based Weighted Clustering* (BWC); celle-ci consiste en un algorithme pondéré, adapté à la structure de la base de données radio. Avant de développer cette méthode, on formalise le concept de *feature type*.

Supposons une base de données $\mathcal{R} = \{\underline{r}_m\}_{m=1...M}$ ; on remarque que tous les éléments dans un enregistrement $\underline{r}_m\}$ n'appartiennent pas à la même nature. On définit un feature type comme l'ensemble de tous les paramètres qui appartiennent à la même nature. Dans le cas le plus simple, il existe au moins deux feature types dans la base : le feature type « position », et le feature type « RSS ». On peut envisager des cas plus compliqués, où il existe des feature types variés (RSS 2G et 3G, TA, etc.). Donc, un enregistrement peut être représenté aussi comme suivant :

$$\underline{r} = (\underline{\rho}_1, ..., \underline{\rho}_h, ..., \underline{\rho}_{N_f}),$$

où $N_f$ est le nombre total des features types, and $\underline{\rho}_h$ est le sous-vecteur correspondant au $h$ème feature type.

Basé sur cette définition, on propose la fonction objective suivante pour l'étape de clustering :

$$J_5(U, \mathcal{R}, \underline{\omega}) = \sum_{n=1}^{N} \sum_{m=1}^{M} \sum_{h=1}^{N_f} u_{mn} \omega_h^{\beta} \|\underline{\rho}_{n,h}^{\circ} - \underline{\rho}_{m,h}\|^2. \tag{7}$$

où $\underline{\omega} = [\omega_1, ..., \omega_{N_f}] \in \mathbb{R}^{N_f}$ est le vecteur comprenant les poids, sous la contrainte $\sum_{h=1}^{N_f} w_h = 1$.
Cette algorithme est sensé être plus efficace que les algorithmes standards de clustering, car il prend en compte la diversité des types des éléments dans chaque enregistrement.

Les tests effectués confirment l'efficacité de la méthode BWC, dans le contexte des systèmes LFP. La figure illustre l'évaluation de la performance pour plusieurs méthodes de clustering, selon des tests simulés et les tests réels. Dans la figure, on voit l'erreur moyenne de la localisation en fonction de l'indice de compression. On observe que pour le scenario simulé, la méthode BWC est largement plus performante que les autres techniques. Par contre dans le scenario réel, la méthode simple du quadrillage manifeste une performance pas trop dégradée par rapport à BWC. Ce résultat montre que sur la zone considérée pour les tests réels, la propagation radio est plus homogène que prévu dans les simulations, et donc les méthodes de clustering ne sont pas très efficaces dans ce cas.

## Traitement des données manquantes dans les systemes LFP

Un problème important concernant les systèmes LFP qui n'a pas été bien examiné dans la littérature consiste en problématique des données manquantes.

Comme mentionné précédemment, dans cette thèse on s'intéresse aux systèmes basés sur les mesures de RSS; ces mesures sont normalement obtenues par une procédure nommée *scanning process*. Le scanning process est une procédure essentielle dans les réseaux radios mobiles, où chaque terminal mesure le niveau de RSS des cellules en voisinage. Mais certaines stations de base ne peuvent pas être détectées pendant cette procédure, à cause des raisons variées :

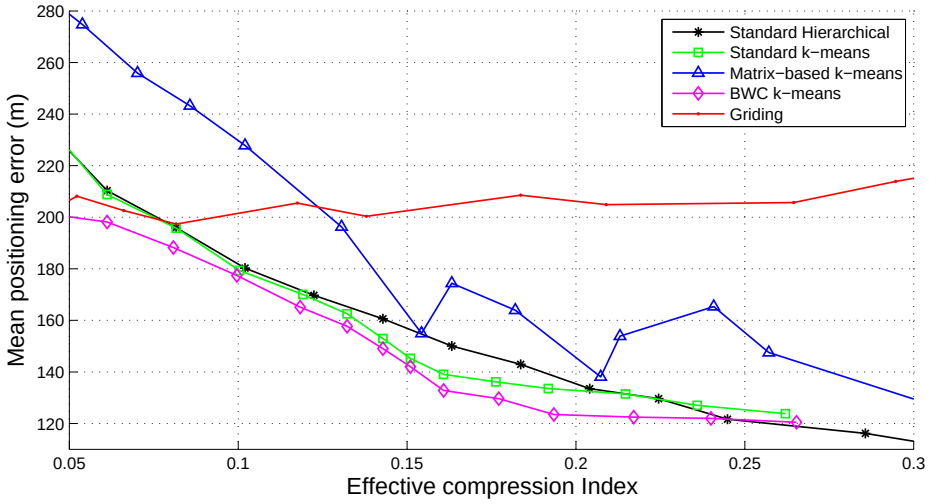
- le signal reçu pourrait être plus faible que le seuil de la sensibilité du terminal,
- le signal reçu pourrait être perdu dans la forte interférence,
- le nombre des stations de base mesurable pourrait être limité au niveau du terminal,
- certaines stations de base pourrait être éteintes.

Dans le cadre de notre étude, on considère tous les signaux non-mesurés comme les données manquantes.

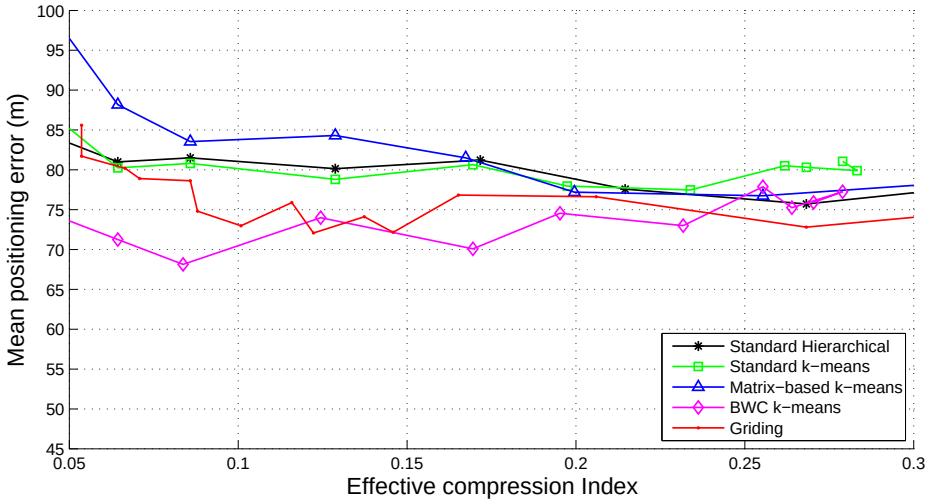
Les méthodes statistiques pour le traitement des données manquantes ont considérablement évolué depuis les dernières années. ([7]). Nous prenons le cadre théorique proposé par [8] et [9], où on distingue le modèle des données complètes (qui modélise l'ensemble des données complètes) et le mécanisme d'effacement (qui rend une partie des donnée inaccessible).

Considérons les mesures RSS effectuées par un terminal mobile, sur une région $\mathcal{A}$ où se trouve $B$ stations de base. Une mesure complète à la position $\underline{x}$ peut être représentée par le vecteur $\underline{s} = (s_1, ..., s_b, ..., s_B) \in \mathbb{R}^B$.

Dans ce travail, nous modélisons le mécanisme d'effacement par deux paramètres : $\lambda$, le seuil de sensibilité du terminal, et $B_{max}$, le nombre maximum des stations de base mesurable




(a) Les résultats concernant le scenario simulé




(b) Les résultats concernant le scenario réel

Figure 6: L'erreur de localisation moyenne en fonction de l'indice de compression

au niveau du terminal ($B_{max} \leq B$). Ensuite, pour formaliser le concept d'effacement, on définit le vecteur indicateur d'effacement $\underline{i} \in \{0,1\}^B$ correspondant à chaque mesure $\underline{s}$, comme suivant. Supposons que $\underline{\sigma} = (\sigma(1), \sigma(2), ..., \sigma(B))$ est une permutation d'indices des stations de base, tel que $s_{\sigma(1)} \geq s_{\sigma(2)} \geq ... \geq s_{\sigma(B)}$; le vecteur $\underline{i}$ correspondant à $\underline{s}$ est alors défini comme:

$$\forall b, 1 \leq b \leq B,\ i_b = \begin{cases} 1 & \text{if } b \in \{\sigma(1), \sigma(2), ..., \sigma(B_{max})\}, \\ & \text{and } s_b \geq \lambda\ , \\ 0 & \text{otherwise.} \end{cases}$$

On définit l'ensemble $\Psi$ comme l'ensemble de tous les paramètres qui modélisent le mécanisme d'effacement ($\Psi = \{\lambda, B_{max}\}$). Finalement, pour une position donnée $\underline{x}$, $\mathcal{B}(\underline{x}) = \{b :\ i_b = 1\}$ représente l'ensemble des stations de base observées à la position $\underline{x}$.

Dans les systèmes de LFP, les données manquantes pourront arriver pendant les deux phases d'apprentissage et localisation. Dans ce travail, nous traitons le problème en deux étapes. Dans la première étape, on suppose que le mécanisme d'effacement est présent exclusivement pendant la phase de localisation ; autrement dit, on suppose que l'on a une base de données complète, mais les mesures du terminal incomplètes.

Pendant la deuxième étape, on enlève l'hypothèse d'une base de données complète ; le mécanisme d'effacement est sensé etre présent pendant les deux phase d'apprentissage et localisation.

### Etape 1: localisation basée sur le maximum de vraisemblance (*Maximum Likelihood*, ML)

Les algorithmes de localisation basés sur le maximum de vraisemblance (Maximum Likelihood, ML) sont déjà proposés dans le contexte des systèmes de LFP. La méthode de ML que nous proposons dans ce travail est différente dans le sens qu'elle prend en compte l'effet d'effacement et les données manquantes.

Supposons que la mesure du terminal pendant la phase de localisation $\underline{s}'$ peut etre décomposé en une partie observée $\underline{s}'^{(obs)}$ et une partie manquante $\underline{s}'^{(mis)}$, ayant pour résultat un vecteur indicateur d'éffacement $\underline{i}'$. Notre algorithm de ML estime la position du terminal comme suivant :

$$\widehat{\underline{x}} = \underline{x}_{\widehat{m}},\ \ \widehat{m} = \text{argmax}_m\ \ p(\underline{s}'^{(obs)}, \underline{i}' |\ m, \Theta_L, \Psi) \tag{8}$$

où $\widehat{\underline{x}}$ est la position estimée du terminal, et l'ensemble $\Theta_L$ modélise la distribution des mesures RSS complètes sur les clusters (précisé ci-dessous). Etant donné le mécanisme d'effacement on peut écrire :

$$p(\underline{s}'^{(obs)}, \underline{i}' |\ m, \Theta_L, \Psi)\ =\ \int_{\xi} p(\underline{s}'^{(obs)}, \underline{s}'^{(mis)} |\ m, \Theta_L) d\underline{s}'^{(mis)}$$

où $\xi$ est un évènement définit par:

$$\xi = \left\{\underline{s}' :\ \forall b \notin \mathcal{B}(\underline{x}'),\ \ s'_b \leq \lambda'(\underline{x}')\right\}, \tag{9}$$

avec

$$\lambda'(\underline{x}') = \begin{cases} \lambda & \text{if } |\mathcal{B}(\underline{x}')| < B_{max} \\ \min\{\underline{s}'^{(obs)}\} & \text{if } |\mathcal{B}(\underline{x}')| = B_{max} \end{cases} \tag{10}$$

Supposant une distribution Gaussienne pour les mesures autours des centroids, on peut écrire:

$$p(\underline{s}'|\ m, \Theta_L) \sim \mathcal{N}(\underline{s}_m, \Gamma_m), \tag{11}$$

avec

$$\Theta_L = \{(\underline{s}_m, \Gamma_m)\}_{m=1,\ldots,M},$$

où $\underline{s}_m$ et $\Gamma_m$ sont respectivement le centroid et la covariance matrice du $m$-ème cluster. En prenant une hypothèse d'independence parmi les signaux des différentes stations de base, on obtient:

$$p(\underline{s}'^{(obs)}, \underline{i}'|\ m, \Theta_L, \Psi) =$$

$$\prod_{b \in \mathcal{B}(\underline{x}')} p_b(s'_b|\ m, \Theta_L) \prod_{b \notin \mathcal{B}(\underline{x}')} F_b(\lambda'(\underline{x}')|\ m, \Theta_L) \tag{12}$$

où $F_b(.|\ m, \Theta_L)$ représente CDF de la distribution Gaussien, correspondant au $b$-ème composant radio.

### Etape 2 : Multiple Imputation

Ce niveau du problème suppose la présence du mécanisme d'effacement pendant toutes les deux phases d'apprentissage et localisation. Afin de traiter les données manquante au niveau de la phase d'apprentissage, on propose une méthode de "*Multiple Imputation*" (MI), qui essaie de remplir les valeurs manquantes dans la base. Une fois que la base radio est complémentée, le traitement des données manquantes pendant la phase de localisation revient à la même problématique étudiée dans l'étape précédente. La figure 7 illustre la méthodologie proposée.

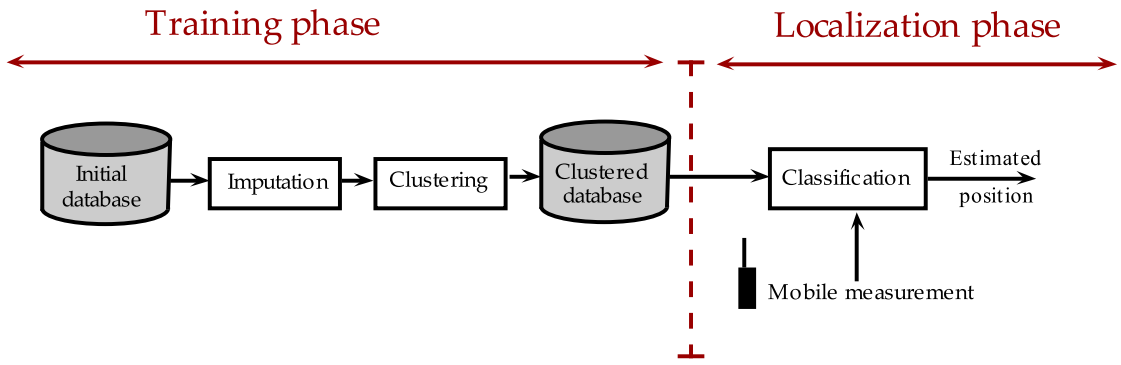



Figure 7: L'architecture proposée comprenant l'étape d'imputation

### Le modèle des données complètes

Le modèle des données complètes, à cette étape, modélise la base de données radio complète $\mathcal{S}^{\circ}$. Prenant un modèle classique log-Normal, chaque mesures de RSS pourrait être modélisée comme suivant :

$$p(\underline{s}_n^{\circ}|\Theta_T) \sim \mathcal{N}([\mu_{n,1}, ..., \mu_{n,B}], \Sigma), \tag{13}$$

où $\Theta_T$ inclut les paramètres du modèle log-Normal, qui permettent de calculer $\mu_{n,1}, ..., \mu_{n,B}, et \Sigma$. Prenant une hypothèse d'indépendance parmi les différentes stations de base, on obtiendra:

$$p(\underline{s}_n^{\circ}|\Theta_T) = \prod_{b=1}^{B} p_b(\underline{s}_{n,b}^{\circ}|\Theta_T), \tag{14}$$

où $p_b(.|\Theta_T)$ est la densité marginale du $b$ème composant, pour $1 \leq b \leq B$.

### Multiple Imputation (MI)

La méthode MI essaie de remplir chaque valeur manquante par une liste de plusieurs valeurs alternatives. Ainsi, la méthode fournit plusieurs versions de la base de données complètes. La combinaison de ces plusieurs version est ensuite exploitée par les methodes d'apprentissage standardes, pour estimer des paramètres et les données manquantes.

Dans ce travail, la méthode de MI est implémentée par l'algorithm de *Monte Carlo Expectation Maximisation* (MCEM). L'algorithme avance en répétant des étapes d'Expectation (E-step) et Maximization (M-step) d'une façon itérative, comme présentée par l'Algorithm 1.

**Algorithm 1** L'algorithm de MCEM pour Multiple Imputation

(Initialisation)

Démarrez avec une valeur initiale pour les paramètres $\Theta_T^{(0)}$.

(Iterations)

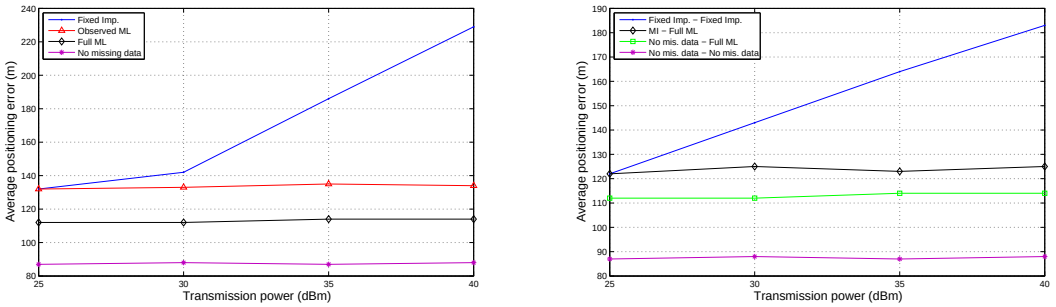
Pour $t \geq 0$, répétez jusqu'à la convergence :

**Monte Carlo E step:** A partir de la distribution conditionnelle $p(\mathcal{S}^{(mis)}|\mathcal{S}^{(obs)}, \mathcal{I}, \Theta_T^{(t)})$, tirez $Q$ échantillons $\{\mathcal{S}^{(mis),(q)}\}_{q=1,...,Q}$.

**M step:** Optimisez les paramètres du modèle, en maximisant l'expectation suivante :

$$\begin{aligned}\Theta_T^{(t+1)} &= \text{argmax}_{\Theta_T}\, \mathbb{E}[\log P(\mathcal{S}^{(obs)}, \mathcal{S}^{(mis)}|\Theta_T)] \\ &= \text{argmax}_{\Theta_T}\, \frac{1}{Q} \sum_q \log P(\mathcal{S}^{(obs)}, \mathcal{S}^{(mis),(q)}|\Theta_T).\end{aligned}$$

(Condition de la convergence) L'algorithme converge quand la variation de $\Theta_T^{(t)}$ est insignifiante.




(a) Résultats pour la première étape

(b) Résultats pour la deuxième étape

Figure 8: L'erreur de localisation moyenne en fonction de la puissance d'emmission

### L'évaluation de la performance

Les performances de la méthode proposée sont évaluées par les simulations. La propagation radio dans les simulations est toujours basée sur le modèle Mondrian. Un environnement similaire à celui de la partie précédente (voir l'analyse de cluster) a été créé.

La Figure 8 illustre les résultats correspondant à la première et deuxième étape de l'analyse.
On peut observer que les méthodes proposées (Full ML, et MI-ML) sont plus performantes que les méthodes naïves existantes dans la littérature.

## Conclusions

Avec l'émergence des services basés sur la position (LBS), les operateurs mobiles souhaitent offrir aux abonnés une localisation précise, durable et d'un coût peu élevé. Cette thèse a été initialisée dans un tel contexte industriel. L'axe principale de la thèse consiste en localisation basée sur les empreintes radios (*Location Fingerprinting*, LFP). Cette méthode exploite les réseaux radios existants, comme les réseaux cellulaires, ou les WLANs. La méthode profite des mesures génériques qui sont disponibles à partir des interfaces radios, et permet donc une localisation à bas coûts.

Dans le cadre de cette thèse, on s'est focalisé sur les systèmes de LFP basés sur les réseaux cellulaires, qui profitent des mesures de RSS pour construire la base de données radio. Dans une première étape, nous avons effectué une analyse de performance qui démontre

certaines caractéristiques intéressantes des systèmes de. Les simulations effectuées utilisent le modèle Mondrian pour modéliser la propagation radio.

La première partie principale de cette thèse concerne la compression de la base de données radio, dans les systèmes de LFP. Nous avons proposé d'effectuer cette compression en appliquant une technique de *clustering* pendant la phase d'apprentissage. Dans une première étape, nous avons utilisé des algorithmes classiques de clustering, y compris l'algorithme de k-means. Dans une étape plus avancée, nous avons développé un algorithme de clustering, bien adapté à la structure des empreintes radios dans la base. Les simulations théoriques et les tests réels ont démontré l'efficacité de la méthode proposée.

Dans la deuxième partie principale de cette thèse, nous avons abordé le sujet du traitement des données manquantes dans les bases de données radio. Une approche systématique a été développée, où on distingue le modèle pour les données complètes, et le modèle pour le mécanisme d'effacement. Le problème des données manquantes a été ensuite analysé dans deux étapes. Dans la première étape, le mécanisme d'effacement est supposé d'être présent seulement pendant la phase de localisation. Ici, un algorithme de localisation basé sur le maximum de vraisemblance a été développé, qui prend en compte le mécanisme d'effacement pour calculer les vraisemblances.

Dans une deuxième étape, le mécanisme d'effacement est supposé d'être présent pendant les deux phases d'apprentissage et localisation. Ici, un algorithme de Multiple Imputation a été développé, qui complémente les éléments manquants dans la base de données radio. Une fois que la base est complétée, le traitement des données manquantes pendant la phase de localisation revient au même problème traité en première étape. Les simulations théoriques ont démontré que les méthodes proposées améliorent la qualité de localisation dans les systèmes de LFP.

# Bibliography


[1] S. Wang, J. Min, B.K. Yi, "Location based services for mobiles: Technologies and standards," in *Tutorial in IEEE International Conference on Communications (ICC)*, June 2008.

[2] Philippe Godlewski, "The Mondrian propagation simulation model," in *Vehicular Technology Conference*, May 2011.

[3] S. Fang, T. Lin, P. Lin, "Location fingerprinting in a decorrelated space," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 20, no. 5, pp. 685 – 691, May 2008.

[4] Y. Chen, J. Yin, X. Chai, and Q. Yang, "Power-efficient access-point selection for indoor location estimation," in *IEEE Transactions on Knowledge and Data Engineering*, July 2006, pp. 877–888.

[5] M. A. Youssef, A. Agrawala, A. Udaya Shankar, "WLAN location determination via clustering and probability distributions," in *Proceedings of the Conference on Pervasive Computing and Communications*, March 2003, pp. 143– 150.

[6] J . C. Bezdek, "A convergence theorem for the fuzzy ISODATA clustering algorithms," pp. 1–8, January 1980.

[7] J. L. Schafer, J. W. Graham , "Missing data: our view of the state of the art," in *Psychological Methods*, vol. 7, no. 2, June 2002, pp. 147–177.

[8] D. B. Rubin, "Inference and missing data," *Journal of Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[9] R. J. A. Little, and D. B. Rubin, *Statistical Analysis with Missing Data.* Wiley, New York, 1987.

[10] A. Kupper, *Location based services.* Wiley, 2005.

[11] J. Figueiras, S. Frattasi , *Mobile Positioning and Tracking: From Conventional to Cooperative Techniques.* Wiley, 2010.

[12] K. Pahlavan, *Wireless information networks.* Wiley, 2005.

[13] M.O. Gheorghita, A. Solanas, and J. Forne, "Location privacy in chain-based protocols for location-based services," in *Proceedings of the Third International Conference on Digital Telecommunications*, 2008, pp. 64–69.

[14] A. M. Bernardos, J. R. Casar, and P. Tarrío, "Building a framework to characterize location-based services," in *Proceedings of the International Conference on Next Generation Mobile Applications, Services and Technologies*, 2007, pp. 110–118.

[15] C. Takenga, K. Kyamakya, "A low-cost fingerprint positioning system in cellular networks," in *Proceedings of the International Conference on Communications and Networking*, August 2007, pp. 915 – 920.

[16] C. Dessiniotis, "Motive project deliverable 2.1," European Project FP6-IST 27659, Tech. Rep., September 2006.

[17] P. Bahl, V.N Padmanabhan, "RADAR: an in-building RF-based user location and tracking system," *Proceedings of the Joint Conference of the IEEE Computer and Communications Societies*, vol. 2, pp. 775 – 784, March 2000.

[18] B.D.S. Lakmali, W.H.M.P. Wijesinghe, K.U.M. De Silva, "Design, implementation and testing of positioning techniques in mobile networks," in *Proceedings of the International Conference on Information and Automation for Sustainability*, December 2007, pp. 94–99.

[19] A.K.M. Mahtab Hossain, H. Nguyen Van, Y. Jin, W.S. Soh, "Indoor localization using multiple wireless technologies," in *Proceedings of the IEEE Internatonal Conference on Mobile Adhoc and Sensor Systems*, 2007, pp. 1 – 8.

[20] M. Anisetti, V. Bellandi, E. Damiani, S. Reale, "Advanced localization of mobile terminal," in *Proceedings of the International Symposium on In Communications and Information Technologies*, 2007, pp. 1071–1076.

[21] Widyawan, M. Klepal, D. Pesch, "Influence of predicted and measured fingerprint on the accuracy of RSSI-based indoor location systems," in *Proceedings of the 4th workshop on positioning, navigation and communication*, 2007, pp. 145–151.

[22] S. Fang, T. Lin, "A dynamic system approach for radio location fingerprinting in wireless local area networks," *Trans. Comm.*, vol. 58, pp. 1020–1025, April 2010.

[23] M.M. El-Said, A. Kumar, A.S. Elmaghraby, "Pilot pollution interference reduction using multi-carrier interferometry," in *Proceedings of the 8th IEEE International Symposium on Computers and Communication*, vol. 2, 2003, pp. 919 – 924.

[24] J. Niemela, J. Lempiainen, "Mitigation of pilot pollution through base station antenna configuration in WCDMA," in *Proceedings of the IEEE 60th Vehicular Technology Conference*, vol. 6, 2004, pp. 4270 – 4274.

[25] C. Takenga, W. Quan, K. Kyamakya, "On the accuracy improvement issues in GSM location fingerprinting," in *Proceedings of the IEEE 64th Vehicular Technology Conference*, September 2006, pp. 1–5.

[26] B.D.S. Lakmali, D. Dias, "Database correlation for GSM location in outdoor and indoor environments," in *Proceedings of the 4th International Conference on Information and Automation for Sustainability*, December 2008, pp. 42 – 47.

[27] A. Arya, P. Godlewski, P. Mellé, "Performance analysis of outdoor localization systems based on RSS fingerprinting," in *Proceedings of the International Conference on Wireless Communication Systems*, September 2009, pp. 378 – 382.

[28] ——, "A hierarchical clustering technique for radio map compression in location fingerprinting systems," in *Proceedings of the International Conference on Vehicular Technology*, May 2010, pp. 1 – 5.

[29] A. Arya, P. Godlewski, Marine Campedel, Ghislain du Chéné, "Radio database compression for accurate energy-efficient localization in fingerprinting systems," *to appear in IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 2011.

[30] A. Arya, P. Godlewski, Stéphan Clémençon, François Vincent, "Handling missing data in cellular localization systems based on RSS fingerprinting," *submitted to IEEE Global Communications Conference*, 2011.

[31] ——, "A MI-ML method to handle missing data in RSS-based location fingerprinting systems," *submitted to IEEE Transactions on Mobile Computing (TMC)*, 2011.

[32] A. Arya, P. Godlewski, "An analysis of radio fingerprints behavior in the context of RSS-based location fingerprinting systems," *to appear in proceedings of IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications*, 2011.

[33] J.D. Gibson, *The mobile communications handbook*. Springer, 1999.

[34] P. Bellavista, A. Kupper, and S. Helal, "Location-based services: Back to the future," in *IEEE Transactions on Pervasive Computing*, vol. 7, no. 2, 2008, pp. 85 – 89.

[35] A.J. Blumberg, P. Eckersley, "On locational privacy, and how to avoid losing it forever," Electronic frontier foundation, Tech. Rep., August 2009.

[36] T. Ming, Q. Wu, Z. Guoping, H. Lili, and Z. Huan-guo, "A new scheme of LBS privacy protection," in *Proceedings of the 5th International Conference on Wireless communications, networking and mobile computing*, ser. WiCOM'09, 2009, pp. 5219–5524.

[37] European Parliament, "Directive 95/46/ec on the protection of individuals with regard to the processing of personal data and on the free movement of such data," Tech. Rep., October 1995.

[38] ——, "Directive 2002/58/ec on privacy and electronic communications," Tech. Rep., July 2002.

[39] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," SRI International, Tech. Rep., 1998.

[40] P. Samarati, "Protecting respondents identities in microdata release," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 13, no. 6, p. 1010–1027, 2001.

[41] L. Sweeney, "k-anonymity: a model for protecting privacy," *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, vol. 10, pp. 557–570, October 2002.

[42] W. Foy, "Position location solutions by Taylor series estimation," *IEEE Transactions on Aerospace Electron System*, pp. 187–193, 1976.

[43] L. Cong, W. Zhuang, "Non-line-of-sight error mitigation in TDOA mobile location," in *Proceedings of Global Telecommunications Conference (GLOBECOM)*, vol. 1, 2001, pp. 680 – 684.

[44] Y. Qi, H. Kobayashi, H. Suda, "Analysis of wireless geolocation in a non-line-of-sight environment," *IEEE Transactions on Wireless Communications*, vol. 5, no. 3, pp. 672 – 681, 2006.

[45] N. Bhagwat, L. Kunpeng, B. Jabbari, "Robust bias mitigation algorithm for localization in wireless networks," in *Proceedings of the IEEE International Conference on Communications*, May 2010, pp. 1–5.

[46] C.E. Cook and M. Bernfeld, *Radar Signals: An Introduction to Theory and Applications.* New York: Academic, 1970.

[47] Y. Qi; H. Kobayashi, "On relation among time delay and signal strength based geolocation methods," in *Proceedings of Global Telecommunications Conference (GLOBECOM)*, 2003, pp. 4079 – 4083 vol.7.

[48] D. Almodóvar, "Location technologies white paper," VF Group R&D Enablers Research, Tech. Rep., 2008.

[49] F. Duquenne, *GPS, localisation et navigation par satellite.* Hermes Science, 2005.

[50] G. M. Djuknic, R.E. Richton, "Geolocation and assisted GPS," *Computer*, vol. 34, pp. 123–125, February 2001.

[51] Y. Zhao, "Standardization of mobile phone positioning for 3G systems," *IEEE Transactions on Communications*, vol. 40, no. 7, pp. 108–116, July 2002.

[52] J. Wang, "Pseudolite applications in positioning and navigation: Progress and problems," *Journal of Global Positioning Systems*, vol. 1, pp. 48–56, 2002.

[53] H. S. Cobb, "GPS pseudolites: Theory, design and applications," Ph.D. dissertation, Stanford University, 1997.

[54] J.L. Dornstetter, "Brevet européen, numéro 90401043.6," MATRA communication, Tech. Rep., 1990.

[55] "3GPP TS 44.071," 3rd Generation Partnership Project; Technical Specification Group GSM/EDGE Radio Access Network; Location Services (LCS); Mobile radio interface layer 3 LCS specification (Release 10), Tech. Rep., 2011.

[56] "3GPP TS 44.035," 3rd Generation Partnership Project; Technical Specification Group GSM/EDGE Radio Access Network; Location Services (LCS); Broadcast network assistance for Enhanced Observed Time Difference (E-OTD) and Global Positioning System (GPS) positioning methods (Release 10), Tech. Rep., 2011.

[57] "3GPP TS 44.031," 3rd Generation Partership Project, Technical Specification Group GSM/EDGE Radio Access Network; Location Services (LCS); Mobile Station (MS) - Serving Mobile Location Centre (SMLC) Radio Resource LCS Protocol (RRLP) (Release 10), Tech. Rep.

[58] H. Laitinen, S. Ahonen, S. Kyriazakos, J. Lähteenmäki, R. Menolascino, S. Parkkila, "Cello project deliverable: Cellular location technology," European Project on Cellular network optimisation based on mobile location, IST-2000-25382-CELLO, Tech. Rep., November 2001.

[59] H Laitinen, J Lahteenmaki, T Nordstrom, "Database correlation method for GSM location," in *Proceedings of the International Conference on Vehicular Technology*, vol. 4, May 2001, pp. 2504–2508.

[60] M. Khalaf-Allah, K. Kyamakya, "Database correlation using bayes filter for mobile terminal localization in GSM suburban environments," in *Proceedings of the Conference on Vehicular Technology*, vol. 2, May 2006, pp. 798–802.

[61] D. Zimmermann, J. Baumann, M. Layh, F.M. Landstorfer, R. Hoppe, "Database correlation for positioning of mobile terminals in cellular networks using wave propagation models," in *Proceedings of the Conference on Vehicular Technology*, vol. 7, September 2004, pp. 4682 – 4686.

[62] R.S. Campos, L. Lovisolo, "Location methods for legacy GSM handsets using coverage prediction," in *Proceedings of the IEEE 9th Workshop on Signal Processing Advances in Wireless Communications*, July 2008, pp. 21 – 25.

[63] P. Wijesinghe, D. Dias, "Novel approach for RSS calibration in DCM-based mobile positioning using propagation models," in *Proceedings of the 4th International Conference on Information and Automation for Sustainability*, December 2008, pp. 29 – 34.

[64] F. Evennou, "Techniques et technologies de localisation avancées pour terminaux mobiles dans les environnements indoor," Ph.D. dissertation, L'université Joseph Fourier, 2007.

[65] H. D. Chon, "Using RFID for accurate positioning," in *Proceedings of the International Symposium on GNSS/GPS*, December 2004, pp. 32–39.

[66] H. Liu, H. Darabi, P. Banerjee, J. Liu, "Survey of wireless indoor positioning techniques and systems," *IEEE Transactions on systems, man and cybernetics*, vol. 37, no. 6, pp. 61 067–1080, 2007.

[67] S. Gezici, Z. Tian, G. B. Giannakis, H. Kobayashi, A. F. Molisch, H. V. Poor, and Z. Sahinoglu, "Localization via ultra-wideband radios, a look at positioning aspects of future sensor networks," *IEEE Transactions on Signal Processing*, vol. 22, no. 4, pp. 70–84, 2005.

[68] T. Gigl, G. J.M. Janssen, V. Dizdarevi, K. Witrisal and Z. Irahhauten, "Analysis of a UWB indoor positioning system based on received signal strength," in *Proceedings of the 4th Workshop on Positioning, Navigation and Communication*, 2007, pp. 97 – 101.

[69] O. Baala, A. Caminada, "Wlan-based indoor positioning system: experimental results for stationary and tracking MS," in *Proceedings of the International Conference on Communication Technology*, November 2006, pp. 1–4.

[70] T. Mitchell, *Machine Learning.* McGraw Hill, 1997.

[71] D.J.C. MacKay, *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, 2003.

[72] H. Zang, F. Baccelli, J. Bolot, "Bayesian inference for localization in cellular networks," in *Proceedings of the 29th Conference on Information Communications*, March 2010, pp. 1963–1971.

[73] D. Fox, J. Hightower, L. Liao, D. Schulz, "Bayesian filters for location estimation," in *IEEE Transactions on Pervasive Computing*, vol. 2, July-Septembre 2003, pp. 24 – 33.

[74] F. Evennou, F. Marx, E. Novakov, "Map-aided indoor mobile positioning system using particle filter," in *Proceedings of the IEEE Conference on Wireless Communications and Networking*, vol. 4, March 2005.

[75] C. Contopoulos, "Motive project deliverable 5.1," European Project FP6-IST 27659, Tech. Rep., 2007.

[76] M. Brunato and R. Battiti, "Statistical learning theory for location fingerprinting in wireless lans," *Comput. Netw. ISDN Syst.*, vol. 47, pp. 825–845, April 2005.

[77] C. Takenga, K. Kyamakya, "A hybrid neural network-data base correlation positioning in GSM network," in *Proceedings of the IEEE International Conference on Communication Systems*, October 2006, pp. 1 – 5.

[78] Z. Wu, C. Li, J. Kee-Yin Ng, and K. R.P.H. Leung, "Location estimation via support vector regression," *IEEE Transactions on Mobile Computing*, vol. 6, no. 3, pp. 311 – 321, march 2007.

[79] J. Junfeng Pan, J. T. Kwok, Q. Yang, and Y. Chen, "Multidimensional vector regression for accurate and low-cost location estimation in pervasive computing," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 18, no. 9, pp. 1181–1193, September 2006.

[80] J. Yang, Y. Chen, "A theoretical analysis of wireless localization using RF-based fingerprint matching," *Proceedings of the IEEE International Symposium on Parallel and Distributed Processing*, 2008.

[81] K. Kaemarungsi, P. Krishnamurthy, "Modeling of indoor positioning systems based on location fingerprinting," in *Proceedings of the twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 2, March 2004, pp. 1012–1022.

[82] T.S. Rappaport, *Wireless Communications, principles and practice.* Prentice Hall PTR, 2002.

[83] M.B Kjaergaard, "A taxonomy for radio location fingerprinting," in *Proceedings of the International Symposium on Location and Context Awareness*, vol. 4718, October 2007, pp. 139–156.

[84] A Kushki, KN Plataniotis, AN Venetsanopoulos, CS Regazzoni, "Radio map fusion for indoor positioning in wireless local area networks," in *Proceedings of the International Conference on Information Fusion*, vol. 2, July 2005, pp. 1311–1318.

[85] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, 2009.

[86] E. Backer and A. Jain, "A clustering performance measure based on fuzzy set decomposition," *IEEE Transactions on Pattern Analysis and Machine Intellegence*, vol. 3, no. 1, p. 66–75, January 1981.

[87] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645–678, May 2005.

[88] P.N. Tan, M. Steinbach, V. Kumar, *Introduction to Data Mining.* Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, 2005.

[89] A.C. Rencher, *Methods of multivariate analysis.* Wiley, 2002.

[90] L. Lebart, A. Morineau, K.M. Berry, *Multivariate descriptive statistical analysis.* Wiley, 1984.

[91] H.H. Bock, "Origins and extensions of the k-means algorithm in cluster analysis," in *Electronic journal for history of probability and statistics*, vol. 4, no. 2, December 2008.

[92] A. Likas, N. A. Vlassis and Jakob J. Verbeek, "The global k-means clustering algorithm," *The journal of the Pattern Recognition Society*, 2003.

[93] "3GPP TS 23.032," 3rd Generation Partership Project, Technical Specification Group Services and System Aspects, Universal Geographical Area Description (GAD) (Release 9), Tech. Rep., 2009.

# Chapter 1

# Mobile localization via Location Fingerprinting

Over the last decade, Location Based Services (LBS) are grabbing the attention of mobile service providers. Such services can be used in a variety of contexts, such as advertising, social networking, personal tracking, etc ([1]). Two major infrastructures for positioning methods involved in the LBS field, are satellite systems and cellular networks (GSM, UMTS or CDMA 2000, and soon LTE). Satellite-based positioning (via GPS, or more generally GNSS) provides accurate localization in the outdoor open environments, with an accuracy of about few meters ([1]). However, it has limitations such as poor performance in built-up areas, where there is no direct line of sight between the satellites and the receiver. Satellite-based positioning suffers also from a heavy power consumption at the receiver ([2], [3]). To overcome these problems, positioning methods based on cellular networks have been developed. It is in such a context, that this thesis proceeds.

## 1.1 Localization in cellular networks

Any powered-on mobile terminal in a cellular system is continuously tracked by the network. During calls, the network follows the mobile users at a cell level. Thereby, a subscriber in communication mode, is localized by the network typically within a few hundred of meters in cities, and a few kilometers in rural areas. During the idle mode, the location of the mobile is followed at the Location Area (LA) level. The location area consists of a group of few tens of cells, defined and configured by the network operator. Supposing that a location area may include a number of 25 to 100 cells, localization during the idle mode will be about 5 to 10 times less accurate than it is during the connected mode. In a nutshell, any powered-on mobile terminal in a cellular network is continuously localized with an accuracy depending on its status.

### Emergency call and the E-911 mandate

The inherent localization of mobile terminals discussed above may be accurate enough for certain applications; but it is not appropriate for challenging situations, like positioning the emergency calls. In 1996, the Federal Communications Commission (FCC) of United States introduced the E-911 mandate, which obliged the wireless service providers to locate mobile callers in emergency situations with a specified accuracy ([2]). Regarding the coarseness of cell-based provided positions, more advanced positioning methods were developed for wireless networks. The FCC E-911 mandate may be considered as a foundation stone of the Location Based Services (LBS) ([4]).

### Location Based Services

Location based services (LBS) permit users to receive highly personalized information and services based on their location ([5]). In a commercial viewpoint, the LBS are expected to provide new sources of revenue, by offering services tailored to the special needs of mobile users ([2]). Such services can be used in a variety of contexts, such as advertising, social networking, personal tracking, etc ([1]).

The 3GPP (3rd Generation Partnership Project) has extensively contributed to the standardization of LBS. Regarding technological aspects, a major problem in the LBS field is that no single positioning method can provide an accurate and continuous positioning in all environments (outdoor, indoor, rural, urban, etc) ([2], [1]). However, while certain services require highly accurate positioning, there exist plenty of applications that do not need very precise localization. Positioning technologies with low power consumption and vast availability (indoor-outdoor) could allow to offer continuous and seamless location based services ([6]).

### Precise localization everywhere?

One of the interests of the service providers in the LBS field, is the ability to offer accurate, low-cost, and continuous location based services to the mobile users. In this regards, the mobile operators desire to provide a precise localization of terminals to the subscribers, or to other legitime commercial parties. At present, GPS and the Cell-ID are the positioning techniques widely adopted to offer LBS. However, these methods can not satisfy all the criteria mentioned above (accuracy, low cost, and continuity). More advanced solutions may be envisaged by using a combination of techniques instead of a single method. This strategy is adopted by certain localization technology vendors; Ericsson Mobile Positioning System or TruePosition Location Platform are examples of this kind, that incorporate multiple location technologies including GPS, Cell-ID (and enhancements), Uplink Time

Difference Of Arrival (U-TDOA), etc. Such advanced positioning solutions may be used as a support for the national security (by offering government agencies solutions with the power to defend against criminal activities), and the public safety (by providing emergency systems that accurately obtain the location of wireless callers, e.g. E-911). However, these solutions are not adopted by many operators, since they are too costly to implement.

### Localization via fingerprinting

One alternative positioning method which is notably less expensive than TDoA-based techniques, is Location Fingerprinting (LFP). Location fingerprinting is a positioning method that exploits the already existing infrastructures such as cellular networks ([7], [8]) or WLANs ([9], [10], [11]). The method takes advantage of standard radio measurements that are available from the radio interface technologies, and hence, may be considered as a low cost positioning method.

Within LFP method, at first during a "training phase", a radio map is constructed over the area of interest. Then a mobile terminal may enter a "localization phase", where its position is determined by matching its local measurements to the radio map entries. The method could be available over the whole coverage area of the underlying network. However, the major implementation drawback is to construct and upgrade the radio database. We notice that the radio database may be used also for other applications, such as radio resource management, network optimization, etc.

In the case of cellular networks, location fingerprinting provides a positioning more accurate than that of cell-ID. The method does not induce a notable power consumption on the target mobile terminals, since it exploits the standard measurements of radio access technologies. Although in time-based positioning approaches (like TDoA) multi-path or non-line-of-sight effects degrade the performance, they do not necessarily cause limitations in the case of LFP systems. Therefore, location fingerprinting may allow to provide a continuous localization of the mobile terminals, in all types of environments (outdoor, indoor, etc.). The method could be a strategic solution to satisfy the interest of mobile operators in providing low-cost, continuous, and rather accurate localization to the subscribers, and constitutes the main axes of this thesis.

## 1.2 Location fingerprinting, Some challenging aspects

Here, we present some challenging aspects of LFP systems, which will be addressed in the framework of the thesis.

In the context of location fingerprinting systems, the radio database may be constructed

by using either empirical measurements, or theoretical modeling tools, or a hybrid approach where a limited number of empirical measurements are performed to calibrate the theoretic propagation models ([8]). The empirical data may be obtained by conducting specific measurement campaigns, or by using the databases that are already at the disposal of the network operators (e.g. databases arisen from network monitoring tools). A radio database may be updated over the time, by incorporating new measurement reports.

The "size" of the radio database is an important aspect regarding the database construction, specially in mobile-based fingerprinting systems. Generally, an under-trained database (containing a low number of measurements) leads to a degraded performance in fingerprinting systems ([12]). In works such as [12], [13] and [14] some methods are proposed to enrich the database, by predicting theoretically the signal values at some new locations. On the other hand, regarding the chaotic nature of radio signal, an over-trained database does not bring further improvement to the positioning accuracy.

It is noteworthy that in LFP systems, the size of the radio database is an influential factor in regards to issues such as *computation load* of the positioning algorithm, during the localization phase. In mobile-based fingerprinting systems, the computation load affects directly the terminal power consumption. Regarding the recent demand for energy efficient networks and the emergence of issues like green networking, reduction of the computation load may be a figure of merit in fingerprinting systems.

The radio database in a fingerprinting system may involve various types of information such as Received Signal Strength (RSS), Timing Advance (TA), path loss profiles, etc. Contrary to time-based measurements that need a synchronization over the network, RSS measurements do not require any additional constraints on the system; hence, today, RSS information is widely used as the adopted parameter in fingerprinting systems ([15]). The works conducted in the framework of this thesis, focus also on the case of RSS-based fingerprinting systems.

In a cellular network, any legacy mobile terminal performs RSS measurements, during the so-called "scanning process". The scanning process is an essential function in cellular networks, where the mobile terminals scan the reference signals of the serving and a number of neighbor cells (the reference signal concerns the BCCH frequency in GSM, or CPICH bit sequence in UMTS). The list of the neighbor cells is already declared on the network side, and is transmitted to the mobile terminals through broadcast messages (case of GSM and UMTS systems). We notice that in RSS-based fingerprinting systems, the mobile measurements during the localization phase usually come from the scanning process; on the other hand, the radio database is not necessarily constructed based on the classic scanning process (since the network operator may use proprietary tools to perform the measurements).

One important issue here, concerns the missing character of the radio measurements. The performed RSS or RSCP measurements during the scanning process may contain some non-detected (missing) values, because of various reasons:

• the target signal may be received with a signal level lower than a minimum threshold,

• the target signal may be lost (or jammed) in severe interference; it might happen in the case of pilot pollution in CDMA (see [16] and [17]),

• the set of measured signals may be incoherent between several field measurement campaigns,

• some base stations may be temporarily switched off (either accidentally, or intentionally for energy saving purposes); this switch-off causes missing values for the corresponding components in the radio measurements,

• the number of base stations to be measured in practice is limited by an upper bound; hence at a given point, there might be some detectable base stations who are not measured because of this limitation.

In RSS-based fingerprinting systems, the mobile measurements during the localization phase arise from the scanning process, and hence include missing data due to the reasons given above. On the other hand, the radio database is not necessarily constructed based on the classic scanning process. As a result, the database records are not necessarily subject to the same degree of missingness.

The methodological difficulties raised by the missing character of RSS measurements have not received much attention in the literature of location fingerprinting systems. A possible heuristic approach, proposed in works such as [8], [18], and [19], consists in replacing all the missing elements by a single reference value. Development of systematic statistical methods to deal with missing data in fingerprinting systems remains as an open problem.

This thesis proposes several contributions concerning these important issues in location fingerprinting systems.

## 1.3 Thesis outline

In the present chapter, we provided a brief review of the location fingerprinting method, and we pointed out some relative challenging aspects of this technique. The next chapter gives a more detailed review for localization techniques in wireless networks. We present the existing positioning techniques in the literature, and we precise the perimeters of the

thesis vis-à-vis the previous works. The chapter is finalized by a detailed discussion on location fingerprinting, where the method is analyzed in a *Machine Learning* perspective.

Chapter 3 presents a performance analysis of RSS-based cellular fingerprinting systems, based on computer simulations.

The RSS measurements are modeled using the "Mondrian" radio propagation model (already developed at Télécom ParisTech). The Mondrian model may provide a standard propagation environment with a homogeneous propagation exponent over the whole area. It allows to introduce a correlated shadowing effect, which is not the case for the classic one-slop log-normal shadowing models. Using the Mondrian radio model, impacts of certain parameters (e.g. measurements error, database density, radio propagation exponent) are examined on the positioning accuracy. The conducted analysis reveals some general characteristics of RSS-based fingerprinting systems in the context of cellular networks and outdoor environments. On this topic we have published [20].

Chapters 4 and 5 concern the development of a "database clustering" step, in the training phase of LFP systems. This clustering step is proposed to compress the initial radio database, and hence to improve the power consumption of mobile terminals during the localization phase. To develop the cluster analysis, a working framework is defined where the input data points and the associated distance metric are determined. At a first step in chapter 4, standard clustering algorithms such as k-means and the minimum variance-based hierarchical method are examined in the context of LFP systems. Next at a second step in chapter 5, a clustering algorithm well-tailored to the structure of the radio database is proposed. Here, we define the concept of *feature types* in association with database records. A feature type is defined as all the stored parameters in a record that belong to the same nature. Based on this definition, a *Block-based Weighted Clustering* (BWC) scheme is proposed, which imposes equal weights to blocks of components belonging to the same feature type, in the clustering cost function; the weight factors associated to feature types are optimized during the clustering process.

On the topic of cluster analysis in LFP systems, we have published [21] and [22].

The next part of this thesis is devoted to treating the problem of *missing data* in the RSS-based fingerprinting systems. In chapter 6, statistical methods are developed to deal with missing data, within the framework of a well-defined missing mechanism. Our modeled missing mechanism proceeds based on two parameters: the receiver minimum sensitivity for signal detection, and the maximum number of base stations to be measured in the radio measurements. The proposed modeling is well tailored to missingness occurring in RSS measurements, issued from the 3GPP-defined scanning process (as in 2G and 3G). Once the missing mechanism is defined, statistical methods are developed at two different levels. At the first level, the missing mechanism is assumed to be present exclusively during the

localization phase; the radio database is supposed to contain no missing elements. Here, a localization algorithm based on Maximum Likelihood (ML) method is proposed, which takes into account the missing mechanism, to compute the likelihoods. At the second level of modeling, the missing mechanism is assumed to be present during both the training and localization phases. Here, a Multiple Imputation (MI) method is proposed to fill in the missing items in the radio database, during the training phase. Once the database is completed, dealing with missing data in the localization phase sends us back to the problem mentioned at the first level.

On this topic, we have submitted [23] and [24].

Finally, the conclusions and perspective for future works are given in chapter 7.

## 1.4 Publications

The academic publications during the thesis include:

A. Arya, P. Godlewski, P. Méllé, "*Performance analysis of outdoor localization systems based on RSS fingerprinting*", in Proc. of International Conference on Wireless Communication Systems, September 2009, pp. 378 - 382 ([20]).

A. Arya, P. Godlewski, P. Méllé, "*A hierarchical clustering technique for radio map compression in location fingerprinting systems*", in Proc. of International Conference on Vehicular Technology, May 2010, pp. 1 - 5 ([21]).

A. Arya, P. Godlewski, "*An analysis of radio fingerprints behavior in the context of RSS-based location fingerprinting systems*", to appear in Proc. of IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC) 2011 ([25]).

A. Arya, P. Godlewski, Marine Campedel, Ghislain du Chéné, "*Radio Database Compression for Accurate Energy-Efficient Localization in fingerprinting Systems*", to appear in IEEE transactions on Knowledge and Data Engineering (TKDE) ([22]).

A. Arya, P. Godlewski, Stéphan Clémençon, François Vincent, "*Handling Missing Data in Cellular Localization Systems based on RSS Fingerprinting*", submitted to Global Communications Conference (Globecom) 2011 ([23]).

A. Arya, P. Godlewski, Stéphan Clémençon, François Vincent, "*A MI-ML Method to Handle Missing Data in RSS-based Location Fingerprinting Systems*", submitted to IEEE Transactions on Mobile Computing (TMC) ([24]).

# Bibliography

[1] S. Wang, J. Min, B.K. Yi, "Location based services for mobiles: Technologies and standards," in *Tutorial in IEEE International Conference on Communications (ICC)*, June 2008.

[2] A. Kupper, *Location based services.* Wiley, 2005.

[3] J. Figueiras, S. Frattasi , *Mobile Positioning and Tracking: From Conventional to Cooperative Techniques.* Wiley, 2010.

[4] K. Pahlavan, *Wireless information networks.* Wiley, 2005.

[5] M.O. Gheorghita, A. Solanas, and J. Forne, "Location privacy in chain-based protocols for location-based services," in *Proceedings of the Third International Conference on Digital Telecommunications*, 2008, pp. 64–69.

[6] A. M. Bernardos, J. R. Casar, and P. Tarrío, "Building a framework to characterize location-based services," in *Proceedings of the International Conference on Next Generation Mobile Applications, Services and Technologies*, 2007, pp. 110–118.

[7] C. Takenga, K. Kyamakya, "A low-cost fingerprint positioning system in cellular networks," in *Proceedings of the International Conference on Communications and Networking*, August 2007, pp. 915 – 920.

[8] C. Dessiniotis, "Motive project deliverable 2.1," European Project FP6-IST 27659, Tech. Rep., September 2006.

[9] P. Bahl, V.N Padmanabhan, "RADAR: an in-building RF-based user location and tracking system," *Proceedings of the Joint Conference of the IEEE Computer and Communications Societies*, vol. 2, pp. 775 – 784, March 2000.

[10] B.D.S. Lakmali, W.H.M.P. Wijesinghe, K.U.M. De Silva, "Design, implementation and testing of positioning techniques in mobile networks," in *Proceedings of the International Conference on Information and Automation for Sustainability*, December 2007, pp. 94–99.

[11] S. Fang, T. Lin, P. Lin, "Location fingerprinting in a decorrelated space," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 20, no. 5, pp. 685 – 691, May 2008.

[12] A.K.M. Mahtab Hossain, H. Nguyen Van, Y. Jin, W.S. Soh, "Indoor localization using multiple wireless technologies," in *Proceedings of the IEEE Internatonal Conference on Mobile Adhoc and Sensor Systems*, 2007, pp. 1 – 8.

[13] M. Anisetti, V. Bellandi, E. Damiani, S. Reale, "Advanced localization of mobile terminal," in *Proceedings of the International Symposium on In Communications and Information Technologies*, 2007, pp. 1071–1076.

[14] Widyawan, M. Klepal, D. Pesch, "Influence of predicted and measured fingerprint on the accuracy of RSSI-based indoor location systems," in *Proceedings of the 4th workshop on positioning, navigation and communication*, 2007, pp. 145–151.

[15] S. Fang, T. Lin, "A dynamic system approach for radio location fingerprinting in wireless local area networks," *Trans. Comm.*, vol. 58, pp. 1020–1025, April 2010.

[16] M.M. El-Said, A. Kumar, A.S. Elmaghraby, "Pilot pollution interference reduction using multi-carrier interferometry," in *Proceedings of the 8th IEEE International Symposium on Computers and Communication*, vol. 2, 2003, pp. 919 – 924.

[17] J. Niemela, J. Lempiainen, "Mitigation of pilot pollution through base station antenna configuration in WCDMA," in *Proceedings of the IEEE 60th Vehicular Technology Conference*, vol. 6, 2004, pp. 4270 – 4274.

[18] C. Takenga, W. Quan, K. Kyamakya, "On the accuracy improvement issues in GSM location fingerprinting," in *Proceedings of the IEEE 64th Vehicular Technology Conference*, September 2006, pp. 1–5.

[19] B.D.S. Lakmali, D. Dias, "Database correlation for GSM location in outdoor and indoor environments," in *Proceedings of the 4th International Conference on Information and Automation for Sustainability*, December 2008, pp. 42 – 47.

[20] A. Arya, P. Godlewski, P. Mellé, "Performance analysis of outdoor localization systems based on RSS fingerprinting," in *Proceedings of the International Conference on Wireless Communication Systems*, September 2009, pp. 378 – 382.

[21] ——, "A hierarchical clustering technique for radio map compression in location fingerprinting systems," in *Proceedings of the International Conference on Vehicular Technology*, May 2010, pp. 1 – 5.

[22] A. Arya, P. Godlewski, Marine Campedel, Ghislain du Chéné, "Radio database compression for accurate energy-efficient localization in fingerprinting systems," *to appear in IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 2011.

[23] A. Arya, P. Godlewski, Stéphan Clémençon, François Vincent, "Handling missing data in cellular localization systems based on RSS fingerprinting," *submitted to IEEE Global Communications Conference*, 2011.

[24] ——, "A MI-ML method to handle missing data in RSS-based location fingerprinting systems," *submitted to IEEE Transactions on Mobile Computing (TMC)*, 2011.

[25] A. Arya, P. Godlewski, "An analysis of radio fingerprints behavior in the context of RSS-based location fingerprinting systems," *to appear in proceedings of IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications*, 2011.

# Chapter 2

# Localization technologies in wireless networks

In this chapter we give a brief review of wireless localization techniques. At first the context of Location Based Services (LBS) is introduced. Afterwards, we present a state of the art of positioning techniques and technologies. Our state of the art study is finalized by a detailed discussion on location fingerprinting, where the method is analyzed in a *machine learning* perspective.

## 2.1 Location Based Services (LBS) context

### 2.1.1 Origins, evolution

Localization is a process to obtain the spatial position of a user. The problem of positioning radio stations became more relevant with the military operations during Second World War, when it was critical to locate the soldiers in the emergency situations ([1]). A couple of years later, the US Department of Defence launched the Global Positioning System satellites to support localization in military operations. In 1990, the system was made accessible to the public for commercial applications ([1]). Gradually, by advances achieved in GPS receiver technology during the recent years, it has been possible to build low-cost and low-power GPS receivers ([2]), and today the GPS is perhaps the most popular commercial positioning system ([3]).

In 1996, the Federal Communications Commission (FCC) of United States introduced the E-911 mandate, which obliged the wireless service providers to locate mobile callers in emergency situations with a specified accuracy. Therefore, positioning methods were also developed for wireless networks . The FCC E-911 mandate laid the foundation stone of

the Location Based Services (LBS) ([1]).

Location based services permit users to receive highly personalized information and services based on their location ([4]). The LBS are expected to provide new sources of revenue, by offering services tailored to the special needs of mobile users ([2]). Such services can be used in a variety of contexts, such as advertising, social networking, personal tracking, etc ([5]). The 3GPP (3rd Generation Partnership Project) has extensively contributed to the standardization of LBS ([2]).

The first generation of LBS launched in the early 2000's, did not gain a significant success among wireless subscribers ([2], [6]). Recently, the emergence of the mobile handsets equipped by GPS capability have given a rise to the growth of Location Based services. This rise has been coupled also with the emergence of vendor-neutral operating systems (e.g. Windows Mobile, Android, etc.), that allow to create device-independent client applications for mobile devices. A prominent example is the location-based application proposed by Google, "Google Latitude", that combines the GPS positioning with WiFi access point sensing, and the cellular Cell-ID positioning. One may envisage that the combination of GPS-equipped mobile devices and open mobile operating systems might significantly contribute to the evolution of LBS in the next years.

### 2.1.2 Operational actors

The LBS operational actors consist of entities that are directly involved in the practical operation and functioning of LBS ([7]). There is no standard model to describe the operational actors in the LBS supply chain. However, here we present some actors mentioned in [2], which are selected in consensus with most LBS approaches:

• Target: it concerns the mobile individual or object that is to be located.

• Position originator: this is the actor that calculates the position of the target (may be the target itself, the network operator, etc.).

• Location provider: this is an intermediate role between position originator and LBS provider, concerning the mere delivery of location data. The location provider gathers the positions provided by one or several position originators, refines them, and returns them to the LBS provider. This service is referred to as Location Service (LCS).

• LBS provider: it is the central actors that offer the LBS and maintain the users subscriptions. They collect location data, combine it with other geographic contents, and transfer the resulting application to the LBS users.

• Content providers: these support the LBS provider by offering geographic content such as maps, points of interest, etc.

• LBS users: these are the final consumers of the LBS.

### 2.1.3 Privacy

#### Mechanisms for privacy protection

The location information is directly related to the privacy of the users. "Location privacy" is the ability of an individual to move in public space with the expectation that under normal circumstances their location will not be systematically and secretly recorded for later use ([8]). By the arrival of LBS in the recent years, people have realized that these kind of services expose extraordinary threats to users location privacy ([9], [4]). Once the location information is exposed, a third party may misuse that privacy information for other benefits. The problem of how to protect user space-time privacy has been becoming an increasingly urgent study ([4]).

Based on [2], the underlying mechanisms for privacy protection may be classified as secure communications, legal privacy policies, and anonymization. In secure communications, the information is not accessed or altered by unauthorized parties, and furthermore, the involved parties are authenticated to verify that they are really the parties they claim to be. Secure communications prevent the intruders from accessing or falsifying the location information during transmission.

A secure communication protocol protects privacy of location information against the unauthorized intruders; however, a user should be able to control the flow of location information by legitime actors. This control may be achieved by privacy policies imposed by governments or international entities, providing a legal framework for data protection. The European Union directives on "Personal data" (Directive 95/46/EC, see [10]) and "Personal data in electronic communications" (Directive 2002/58/EC, see [11]) are examples of such legal policies.

Policies may be efficient for protecting privacy, if all the LBS actors are trustworthy. However, in practice there may be some rule violations (may be not necessarily on purpose). One suggested mechanism to cope with such violations is anonymization. Anonymization-based methods rely on inserting a "trusted third party" or "anonymizer" between the target and the LBS service providers, to hide the real identity of the target ([4]). Various anonymization methods have been proposed in the literature (see [12], [13], [14], [4], [9]).

#### Mobile-based and network-based positioning approaches

One proposed classification for localization techniques, which may concern the privacy issue, is based on the entity that performs the measurements and calculates the position. In this context, the positioning techniques are categorized as "network-based" or "mobile-

based" ([2], [7]). In terminal-based methods, measurements and positioning process are performed on the terminal side (here the target and the position originator coincide). In network-based methods, the required measurements and the positioning algorithm are performed by the network (position originator is the network). On the other hand, there exists also a hybrid approach in which the measurements are done by the terminal and according to these measurements, the position is calculated by the network. This method is called mobile-assisted. In a privacy view-point, the terminal-based methods are the most appropriate localization techniques to be implemented in LBS.

### 2.1.4 Databases in LBS

In the context of LBS, various types of databases may be involved; the provided information may be used either as assistance data for the localization process (at LCS level), or as additional data to provide practical services (as content providers). The main involved databases in LBS are as follows.

- Satellite ephemeris and almanac.

In satellite-based positioning systems, one needs to calculate the current position of the satellites in the space. To this end, each satellite continuously broadcasts the ephemeris (its highly accurate orbital data) and the almanac (approximate orbital data for all other satellites).

- Base stations positions and the time-offsets.

Analogously to satellite systems, for terrestrial lateration-based positioning methods in cellular systems, one needs to know the exact geographical position of the base stations. The base stations clock offset is another important information to compute the signals propagation delay, since the base stations in cellular networks are not synchronous. These information are provided by a special center called the Serving Mobile Location Center (SMLC).

- Radio databases.

The location fingerprinting localization method requires a radio database, consisting of location-dependent radio parameters. This radio database may be obtained by conducting specific measurement campaigns, or by using the databases that are already at the disposal of the network operators. We notice that the radio database may be used also for other applications, such as Radio Resource Management (RRM), network optimization, etc.

- Geographical Information System (GIS).

The computed location of a target may not be directly meaningful for the LBS user; it would be more convenient for the user to receive the target position in combination with a map and additional navigation assistance. Geographic Information Systems (GIS) are

the essential technologies for mapping spatial location onto meaningful descriptive location information (see [2] for more details).

## 2.2 Fundamental concepts

### 2.2.1 Basic localization techniques

The basic techniques of positioning an object may be categorized as follows.

- Triangulation.

Triangulation is a technique that makes use of angle of arrival measurements. Here, the position is calculated by using the measured angles between the target and a number of reference points.

- Lateration.

In lateration, either the range or the range difference between a target and a number of at least three reference points are used to calculate the target position. There exist two different methods of lateration. If positioning is based on the range measurements, we will have "circular lateration", while range difference measurements will lead to "hyperbolic lateration" (for more details see [2]).

- Proximity sensing.

Proximity-sensing techniques do not use any measured quantity to determine the position; here the localization is based on the presence of the target in a particular area, within the range of a specific emitter ([7]).

- Cell ID.

Cell-ID is a derivation of the proximity-sensing method. Here, the location of the mobile is determined according to the identity of the serving radio antenna.

- Location Fingerprinting.

Here, a database of location dependent parameters is constructed over a radio network. Later a moving terminal performs measurements of the same parameters; these measurements are matched to those values in the database, in order to yield position estimates. This approach may be considered as a "radio cognitive" method.

Any positioning technology is based on one of the basic localization methods, given above. In the following sections, we review the various positioning technologies.

### 2.2.2 Time-based versus RSS-based range measurements

All lateration-based positioning techniques use the range or range difference measurements to localize a target. The range measurements may be obtained either by performing "time" measurements, or by measuring the "RSS" level of the radio signals. In the following subsections, we explain some properties of these measurements.

#### Major error sources

Errors in the range measurements leads to poor accuracy for localization process. The main sources of error for time-based range measurements include:

• Clocks inaccuracy: inaccurate and instable clocks directly lead to errors in the range or range difference measurements ([2]). As a result, mechanisms of clock synchronization are used in all positioning methods which use time-based range measurements.

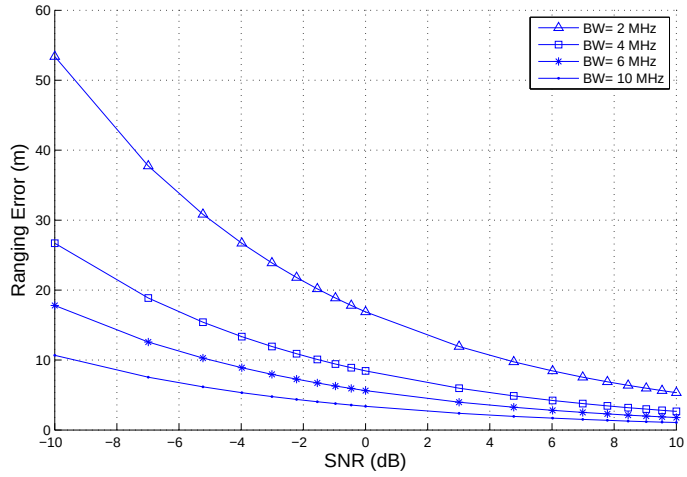
• Non Line of Sight (NLoS) propagation: the NLoS is a major error source in time-based localization methods ([15], [16], [17], [18]). Under NLoS propagation, the signal arriving at the receiver from a transmitter is reflected and diffracted and takes a path that is longer than direct path. So, the corresponding computed locus will lie far from the true position of the terminal. This is a particular problem for positioning in cellular networks since in satellite-based systems, line of sight is always required ([2]). Various methods are proposed in the literature to mitigate the error due to NLoS propagation ([16], [17]).

The main error sources for RSS-based range measurements include:

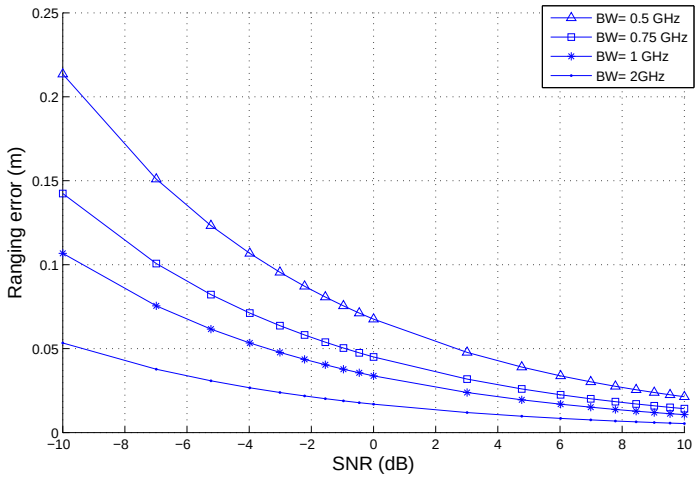
• Non-calibration of radio propagation models: the signal propagation between the transmitter and receiver may be subject to NLoS and the shadowing effect, due to the complicated surrounding environment. Consequently, it is difficult to characterize the relationship between distance and RSS by theoretical propagation models. As an example, taking a classic log-Normal propagation model with a propagation exponent of 3, a typical variation of 6 dB for shadowing effect leads to a variation of 60 % for range measurements.

• Inaccuracy of RSS measurements: the instantaneous RSS measurements are subject to fluctuations due to factors like fast fading. These variations in signal value can degrade significantly the positioning accuracy.

In general, the error potential of RSS measurements is much higher than that of timing measurements, and hence in most cases, the ranges are generally derived from the latter measurements ([2]).




(a) Wide Band systems




(b) Ultra Wide Band systems

Figure 2.1: CRLB for ranging error, based on time of arrival measurements

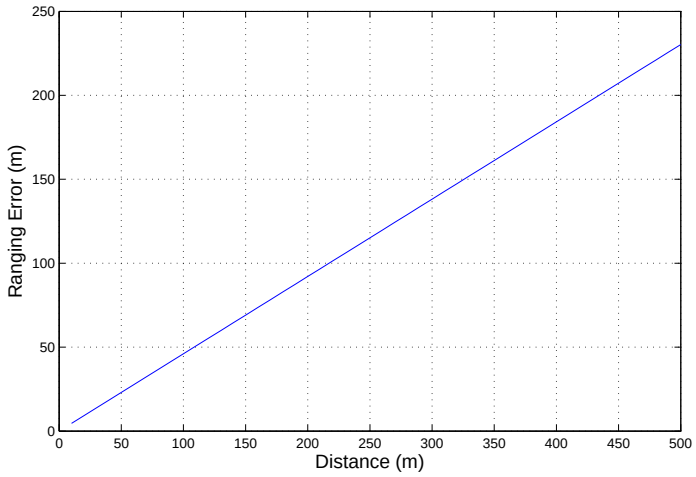



Figure 2.2: CRLB for ranging error, based on RSS measurements

**Theoretical lower bounds for range estimates**

Some theoretical lower bounds are derived for range estimates in the literature. In [19] and [20], the Cramér-Rao lower bound (CRLB) of the range estimates $\widetilde{d}$ derived from time measurements, for a single-path additive white Gaussian noise (AWGN) channel is presented. Based on these references, the standard deviation of range estimates $\sigma(\widetilde{d})$ in this case may be described as follows:

$$\sigma(\widetilde{d}) \geq c/(2\sqrt{2}\pi\sqrt{SNR}\,BW)$$

where $c$ is the speed of light, $SNR$ is the signal-to-noise ratio, and $BW$ is the effective signal bandwidth. The derived expression assumes that the transmitter and receiver have the same reference clock.

Now, assume the case of RSS-based ranging systems. Suppose the classic one-slop log-Normal radio propagation model. Based on this model, the CRLB associated to range estimates $\widetilde{d}$ from RSS measurements is shown to be as follows ([20] ):

$$\sigma(\widetilde{d}) \geq (\ln 10/10).(\sigma_{Sh}/\alpha)\; d$$

where $\sigma(\widetilde{d})$ is the standard deviation of range estimates, $d$ is the actual distance in meters, $\alpha$ is the path loss exponent, and $\sigma_{Sh}$ is the standard deviation (in dB) of the zero mean Gaussian random variable representing the log-normal shadowing effect.

From the above result, we may deduce that the accuracy of a time-based approach can be improved by increasing the $SNR$ or the effective signal bandwidth, as in UWB systems.

Figure 2.1 gives some numerical values for CRLB in time-based ranging systems. On the other hand, in RSS-based approach, the best achievable limit depends on the channel parameters and the distance between the transmitter and the receiver. The CRLB in this case can not be improved by configuring the system parameters. Figure 2.2 gives some numerical values for CRLB in RSS-based ranging systems, under a log-Normal channel with $\alpha = 3$ and $\sigma_{Sh} = 6$ dB.

## 2.3 Satellite positioning

### Global Navigation Satellite Systems

Satellite-based positioning systems are globally subsumed under the title Global Navigation Satellite Systems (GNSS) ([21]). The most prominent example of these infrastructures is the Global Positioning System (GPS). Similar systems are the Russian GLONASS and the European Galileo.

The Global Positioning System (GPS) is a satellite-based positioning system designed by the Department of Defence of United States ([22]). The GPS applies a terminal-based positioning method which lies on circular lateration; the lateration is performed by measuring the the propagation delay of signals coming from the satellites (Time of Arrival (ToA)-based lateration). A GPS receiver must capture the signals coming from at least four GPS satellites, in order to determine its 3-D position ([2]).

The most important advantages of satellite-based positioning systems are their global availability and high accuracy ([2]). But there are also certain drawbacks concerning satellite infrastructures. First of all, satellite signals are very sensitive to shadowing effects. This is because the microwave frequencies that GPS satellites broadcast may easily bounce or be absorbed by buildings, walls, etc. It is especially the case in dense urban areas and other places where there are many large obstructions in the receiver's horizon. So the system works well if a direct line of sight exists between the satellite and the receiver, which is not the case particularly for "indoor" situations. Another major disadvantage of satellite based systems is their high power consumption at the receiver side. This problem restricts usage of positioning applications, especially in battery-operated mobile devices ([2]).

The time it takes the GPS receiver to calculate its position when turned on is called "Time To First Fix" (TTFF). Classic GPS receivers have a long TTFF when starting "cold" (i.e. without any knowledge about the GPS constellation state). The TTFF in cold start can take from 30 seconds to few minutes, which is not acceptable for many applications such as emergency calls ([23], [24]).

### Assisted-GPS (A-GPS)

Assisted GPS describes a system where an assistance server helps the GPS receiver perform the tasks required to make range measurements and position solutions. Thanks to this assistance, a set of tasks that the receiver would normally handle itself, is shared with the assistance server.

The assistance server communicates with the GPS receiver via a wireless link (such as 3G). The basic types of data that the assistance server provides to the GPS receiver are the precise GPS satellite orbit information (ephemeris) and the initial position and time estimates. These assistant data allow a much narrower signal search bandwidth for ToA calculations, and hence reduce notably the TTFF and the power consumption of the receiver ([24], [2]).

The Assisted-GPS technique may be implemented by a mobile-based or mobile-assisted approach ([24]).

#### Pseudolites

We notice that the lateration-based positioning as in GNSS, may also be done by using the ranging signals coming from ground-based transmitters; these transmitters may be called pseudo-satellites or in short "pseudolites" ([25], [26]). Pseudolites may be regarded to be used either as an augmentation tool of existing satellite-based systems, or as an independent system for indoor positioning applications ([25]).

## 2.4 Cellular positioning

Cellular positioning refers to the positioning mechanisms that are implemented in cellular networks like GSM or UMTS. Such techniques can be used for both outdoor and indoor situations.

### 2.4.1 Cell-ID and enhancements

In cellular networks, the proximity sensing can be implemented by using the Cell-ID information.

#### Cell-ID

The easiest way to estimate the location of a mobile terminal is to use its Cell-Identity (Cell-ID). According to the Cell-ID, we can identify the mobile serving cell, and this way we can provide an approximate estimation of the mobile position. Since the mobile terminal can be anywhere in the cell coverage area, the accuracy of the Cell-ID method depends on the size of the cell. The best performance is achieved in the urban areas where the cell sizes are the smallest (micro and pico cells) ([27]).

#### Cell-ID combined with TA/RTT

The accuracy of the pure Cell-ID technique can be enhanced by incorporating timing information, such as Time Advance (TA) in GSM or Round Trip Time (RTT) in UMTS ([27]). Based on this timing information, it will be possible to identify a ring of potential positions of the mobile with the serving base station in its center. However in the case of GSM systems, regarding the poor resolution of the TA parameter (about 550 m), the method is beneficial only in case of large cells ([27]). We note that the resolution of RTT in UMTS is much better than TA parameter in GSM (80 m in UMTS against 550 m in GSM), because of wider bandwidths used in UMTS networks.

#### RSS-based lateration

In the RSS-based lateration technique, the RSS measurements are exploited to calculate the distance between the mobile terminal and the reference base stations. Then a lateration method is used to determine the position of the terminal. In [27] the method is also called CGI++ (Cell Global Identity).

Although there exists various statistical path loss models in the literature, but they are not very accurate due to shadowing and fading effects ([1]). As mentioned before, in general, the error potential of RSS-based range measurements is much higher than that of time-based measurements. Consequently, the approach is mostly used in indoor environments, where the signal traveling time is hard to measure due to extremely short distances between the target and the transmitters ([2]).

### 2.4.2 Time Difference of arrival (TDoA)

In TDoA method, the time measurements are exploited to calculate the range differences between the mobile terminal and the reference base stations. A lateration method is then used to determine the position of the terminal. In order to provide an accurate positioning, the base stations must be synchronized among eachother. In other words they must constitue a pseudolite-like network, which is not the case in practice. Additional mechanisms are considered to provide a posterior synchronization over the network.

#### Downlink TDoA (E-OTD and OTDoA)

#### Basic concept of downlink TDoA

In downlink TDoA positioning, the measured time periods between the arrival of data bursts from different base stations at the terminal are used for localization. However, owing

to the absence of synchronization between the base stations, the observed time difference of arrival at the terminal is not the actual value.

Suppose that the reference time is represented by $\theta$; the moment of occurrence of any event $\pi$ may be denoted by $\theta(\pi)$. All other clocks are plesiochronous with the virtual clock, generating $\theta$ with different offsets, drifts or jitters. We may assume that during the time intervals involved in TDoA measurements, the drift and jitter effects are negligible. Therefore, the inexact generated time $\theta'$ may be described by $\theta' = \theta + T$, where $T$ is a time offset. Based on this model, the local measured time at any base station $BS_k$ is given by $\theta'_k = \theta + T_k$, where $T_k$ represents the corresponding offset. Different values of $T_k$ for different base stations reflect the lack of synchronization in the network. For a pair of base stations $BS_i$ and $BS_j$, the Real Time Difference (RTD) is defined as $T_{R,i,j} = T_j - T_i$. The knowledge of RTD values $\{T_{R,i,j}\}$ would allow to synchronize the base stations.

Figure 2.4 illustrates a mobile terminal in a cellular network, performing time measurements for downlink TDoA positioning (based on [28]). The quantity Observed Time Difference (OTD) refers to the time period observed at the terminal between the arrival of data bursts from different base stations. Assuming that the base stations emit their reference signals at $\theta = 0$, and assuming that $\theta'_m$ denotes the timing measured by the terminal clock, we have:

$$\begin{aligned} T^{DL}_{m,i,j} &= \theta'_m(\text{signal arrival from } BS_i) - \theta'_m(\text{signal arrival from } BS_j) \\ &= T_i + t_{m,i} - (T_j + t_{m,j}) \\ &= T_{R,i,j} + (t_{m,i} - t_{m,j}) \end{aligned}$$

where $T^{DL}_{mij}$ is the OTD between the downlink signals coming from $BS_i$ and $BS_j$, and $t_{m,i}$ and $t_{m,j}$ are the propagation delays between the terminal and the base stations $BS_i$ and $BS_j$, respectively. We note that $t_{m,i}$ and $t_{m,j}$ are temporal intervals independent of any clock measurement. We notice that without removing the base stations offset term $T_{R,i,j}$, the raw OTD measurements do not provide accurate positioning.

The same OTD measurements may be made at special units called *Location Measurement Units* (LMUs), installed at well known positions in the network. In this case the measurements are called Radio Interface Timing (RIT) measurements. The knowledge of the exact coordinates of LMUs and the base stations allows to compute the actual RTD values. The measurement reports from all LMUs are collected by a special center called the *Serving Mobile Location Center* (SMLC), which compiles this assistance data and passes them to the target terminals in order to configure them for OTD measurements. The location can be calculated either in the mobile terminal (mobile-based solution) or in the

network side (mobile-assisted solution). The data delivery to a target terminal may be done by means of either a dedicated signaling channel, or broadcast signaling.

### Positioning process

Having introduced the TDoA concept and the OTD measurements, the overview of positioning process in downlink TDoA method can be summarized as follows. The SMLC collects the measurements reports from all LMUs it is responsible for and compiles assistance data (the detailed message format and information elements exchanged between the LMU and SMLC are described in document 3GPP TS 44.071 [29]). The positioning is initialized upon arrival of a "Location Request" message at the SMLC, which contains the identifier of the target terminal and its rough position. The latter is needed to identify the serving and neighbor base stations that are located around the terminal and are supposed to be used for OTD measurements. With knowledge of this rough position, the SMLC compiles the relevant assistance data and passes them to the target terminal. The assistance data includes the coordinates and RTD values corresponding to relative base stations, and also determines the type of localization: mobile-based or mobile-assisted (the content of assistance data is detailed in the document 3GPP TS 44.035 [30]). In the mobile assisted scenario, the terminal measures the OTD values and send them back to the SMLC. In the mobile-based scenario, the terminal measures the OTD values and by using the assistance data performs a self-localization (the detailed message format and information elements exchanged between the mobile and SMLC are described in document 3GPP TS 44.031 [31]).

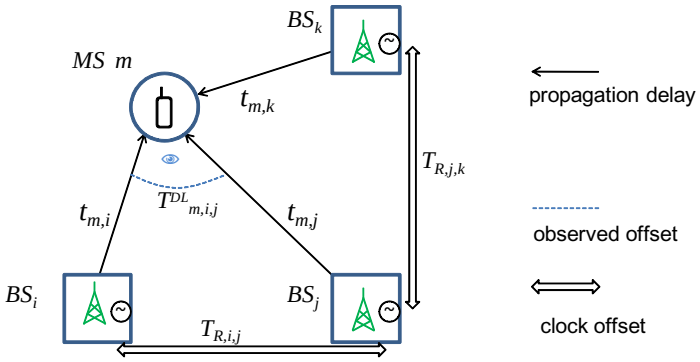



Figure 2.3: Time measurements of a mobile terminal in downlink TDoA

### E-OTD versus OTDoA

The downlink TDoA positioning is called E-OTD (Enhanced Observed Time Difference) in the context of GSM networks; OTDoA (Observed Time Difference of Arrival) is its

counterpart in UMTS networks. In E-OTD, the $OTD$ and $RIT$ measurements are generally performed on the BCCH channel by using the synchronization bursts. In OTDoA, timing measurements are based on CPICH observations. A basic concern in the case of OTDoA method is the "hearability" problem. In CDMA-based systems like UMTS, if a terminal stays close to the serving base station it can not properly receive the signals from farther base stations. Thus the terminal might be unable to detect a sufficient number of neighbor base stations for $OTD$ measurements. To overcome this problem, each base station must switches off its transmitter for all channels (common and dedicated) for short periods of time, during which the terminal is able to detect the CPICH of neighbor cells. These periods are called *idle periods*, and the method to coordinate them in a base station is called Idle Period DownLink (IPDL). The document "3GPP TS 25.214" (section 8: "Idle periods for IPDL location method") is the primary 3GPP specification reference for this functionality. The control of IPDL is with SMLC, which configures base stations for inserting the idle periods; the parameters of the IPDL configuration are then passed to the target terminal as a part of assistance data.

## Uplink TDoA (U-TDoA)

### Basic concept of U-TDoA

Like E-OTD and OTDoA, the Uplink Time difference of Arrival (U-TDoA) method uses hyperbolic lateration to localize the terminal. The difference is that it is a network-based method, which exploits the uplink transmissions of the terminals. In U-TDoA the uplink transmissions of a busy terminal are observed by the serving base station and also a number of LMUs (since it can be heard by other base stations). Assume that the terminal emits a unique signal at instant $\theta = 0$; for two base stations $BS_i$ and $BS_j$, the time difference of arrival is derived:

$$\begin{aligned} T^{UL}_{m,i,j} &= \theta'_i(\text{signal arrival at } BS_i) - \theta'_j(\text{signal arrival at } BS_j) \\ &= T_i + t_{m,i} - (T_j + t_{m,j}) \\ &= T_{R,i,j} + (t_{m,i} - t_{m,j}) \end{aligned}$$

which may be calculated by using LMUs measurements.

An important requirement of U-TDoA is to have a sufficient number of LMUs in proximity of the terminal. Another requirement is that the terminal is in communication, since the LMUs do not detect idle terminals. In the case of an idle terminal, the network must stimulate it to transmit data.

### Positioning procedure

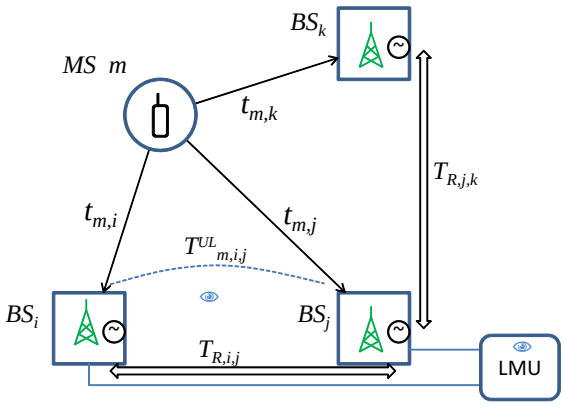



Figure 2.4: Time measurements of a mobile terminal in uplink TDoA

In a first step, the SMLC discovers the serving base station and the physical uplink channel of the mobile. With the knowledge of the rough position of the terminal, a set of LMUs close to the terminal are identified (at least three). After selecting the LMUs, they are configured for measurements, e.g. by defining the physical channel to be monitored. The LMUs then listen to the incoming bursts from the terminal and record their time of arrivals. In the last step, the results are returned to the SMLC, which derives the time differences of arrival and estimates the position of the mobile.

### 2.4.3 Angle of Arrival (AOA)

In AoA, the position of the mobile terminal is determined by considering the angles between the terminal and a number of base stations ([2], [7]). If there is not a line of sight between the mobile and the base station, the signal will be subject to one or multiple reflections making the signal arrival direction random. Consequently, it is not the method of choice in dense urban areas.

In order to implement the AoA method, either the base stations or the terminals should be equipped with antenna arrays, depending on whether the positioning is network or terminal based. With the technologies available today, a terminal based solution is not practicable (because of economic and technical reasons). So, in the systems implementing AoA, the array antennas are usually arranged at the base stations (leading to an uplink solution). In 3G and 4G systems, AOA method may become available without separate hardware if adaptive antennas (e.g. MIMO) are widely deployed in base stations ([32]).

### 2.4.4 Location Fingerprinting (LFP)

Location Fingerprinting (LFP) is an enhancement of Cell-ID method, where the information of neighbor cells are also incorporated to localize the mobile terminal. It also appears in the literature under the names "database correlation", "pattern recognition" and "pattern matching" ([27]). Adopting the Machine Learning terminology, localization by fingerprinting systems may be described as follows. At first, during a preliminary *training phase*, a radio map is constructed over the area where the mobiles are to be located. Once the radio map is constructed, mobile terminals may enter a *localization phase*. Here a mobile can be localized by matching its received signal to the radio map entries.

#### Training phase

During the training phase a radio database is constructed over the considered area. The radio database consists of a set of radio measurements performed at a number of known locations over the area. A radio measurement contains a number of parameters available from the radio interface technologies, such as Received Signal Strength (RSS) from different base stations, Timing Advance (TA) or Round Trip Time (RTT), or the results of more complex processing such as path loss profiles. It is also possible to consider measurements from several radio systems (GSM, UMTS, etc.). The radio measurements kept in the database are called fingerprints ([33], [7]). Each fingerprint stored in the database may be obtained by averaging several radio measurements performed at the same location but at different moments (as in [34]), or by averaging several measurements performed at different locations (as in [27], [35]).

The database is created by using either empirical measurements or theoretical modeling tools ([27]). The latter approach is used specially in the case of outdoor positioning where large surfaces are to be covered by the fingerprinting system. In works such as [36], [37], [38] and [39] various radio propagation models are used to predict the radio database measurements theoretically. In [40] and [41] hybrid methods are proposed, where a limited number of empirical measurements are performed to calibrate the theoretic propagation models.

The training phase may also include some additional processing steps, in order to elaborate the raw database for various purposes. In section 2.7, a review of database processing methods proposed for location fingerprinting systems, will be presented.

In general, the training phase is cumbersome and time-consuming. In the extreme case of global positioning, it is not possible to cover the whole world ([7]). Once the database is created, another major effort consists of its upgrading and maintenance ([33]).

### Localization phase

During the localization phase, a moving mobile terminal collects measurements to be compared with the values in the database. A "positioning algorithm" or "matching algorithm" is then used to determine the position of the mobile by associating the actual measurement (fingerprint) to the ones stored in the database. At this stage, various filtering methods may also be used to incorporate the mobile motion history or the area map information, in order to estimate more realistic trajectories for the mobile ([42]). The positioning algorithms used in fingerprinting systems are presented thoroughly in section 2.7. In general, the main challenges of localization phase include improving the quality of matching algorithms, and reducing the complexity of processing the data ([1]).

## 2.5 Indoor/WLAN positioning

Indoor positioning includes localization techniques that are intended to be used inside buildings, on university campuses, etc. It is generally based on radio, infrared or ultrasound technologies.

In the following we review briefly the techniques based on radio infrastructures, i.e. positioning techniques developed for Radio Frequency Identification (RFID) networks, Ultra Wide Band (UWB) systems, and Wireless Local Area Networks (WLANs).

### 2.5.1 RFID positioning

An RFID (Radio Frequency Identification) system consists of tags, a scanner (reader), and software such as a driver and middleware. The main function of the RFID system is to retrieve information (ID) from a tag (also known as a transponder). A tag can include additional information other than the ID, which opens up opportunities to new application areas ([43]).

RFID positioning is merely based on proximity sensing technique. RFID systems determine the position of a target based on the presence of that target in a particular area, within the range of a RFID scanner.

Deployment of an RFID system over a large campus or company area is very expensive because of the need for installing a multitude of scanners. Also, changing the layout of a manufacturing plant or moving walls in an office requires remounting and rewiring of the RFID readers. Besides, RFID positioning needs proprietary hardware; such proprietary hardware is usually only available from a single vendor, making equipment prices higher than standard-based solutions.

### 2.5.2 UWB positioning

Ultra Wide Band (UWB) is a radio technology based on using ultrashort pulses (typically ¡1 ns). On the spectral domain, the system enables transmission of data over a large bandwidth ($>$ 500 MHz) ([44]).

UWB positioning systems, similar to most other positioning solutions, have proprietary scanners that continuously monitor UWB radio transceivers attached to clients. Positioning approaches for UWB are either based on lateration (by using time or RSS measurements), or angulation (AoA) ([45], [46]). According to [45], due to the high time resolution of UWB signals, time-based location estimation schemes usually provide better accuracy than the others. The lateration based on RSS measurements in UWB suffers from the same problems as in cellular networks. The AoA approache is not suitable neither, since it demands use of antenna arrays, increasing notably the system cost. More importantly, due to the large bandwidth of a UWB signal, the number of paths may be very large, especially in indoor environments. Therefore, accurate angle estimation becomes a very challenging issue.

### 2.5.3 WiFi positioning

The major problem of indoor positioning technologies discussed so far is their proprietary nature, which demands a separate infrastructure to perform the localization. This attribute makes these techniques costly to deploy, scale, and support. Integrated solutions are certainly preferable in order to reduce these costs and operational support risks.

Over the past few years, WiFi has been adopted as the primary standard for wireless LANs in company facilities and homes worldwide. Based on the IEEE 802.11 standards, WiFi addresses needs for secure, high performance mobile data networking. With the widespread adoption of wireless LANs, WiFi is an ideal infrastructure for positioning technologies. The WiFi signal does not contain any exploitable temporal information. Thus, in order to design an integrated positioning technique, we must rely on the received power measurements. There exist some commercial WiFi positioning solutions in the market that use temporal methods such as TDoA. But all these solutions demand certain modifications to the actual WiFi sructure (such as modified WiFi access points) ([42]).

Since the available information in the WiFi signal is the received power level, the proposed positioning approaches are: proximity sensing, RSS-based trilateration, location fingerprinting. Proximity sensing is equivalent to Cell-ID method in GSM and UMTS. The position of the terminal is simply determined by considering its serving access point. In the RSS-based trilateration, the position is determined by lateration with respect to three or more access points. The distance between the mobile and the access points is calculated by using a radio propagation model. The main problem in this technique is the lack of a pre-

cise radio propagation model for the complex indoor environments ([42]). Fingerprinting method requires a database containing the signal strength records. The position is determined by comparing the measurements of the mobile terminal with the database stored fingerprints. The main constraint in this method is building and upgrading the database ([42]). There exists another problem in fingerprinting technique which stems from the received signal temporal variations. As the received signal strength fluctuates over the time, the position extracted from the database will fluctuate as well. Several filtering methods have been introduced in the literature to improve WiFi positioning techniques (see [42]).

## 2.6 A performance comparison

The wireless positioning technologies were presented in the previous sections. We saw that any technology uses is based on one of the basic localization methods: angulation, lateration, proximity sensing and fingerprinting.

The angulation-based methods like AoA are not widely implemented in practice, since they demand special hardware requirements (e.g. array antenna). Moreover, all the lateration-based methods require a temporal synchronization among the emitters in the network. This requirement is obviously met in satellite-based positioning systems like GPS. However, in cellular networks such a synchronization is not necessarily guaranteed. Because of several reasons, a "synchronous network" was not the solution adopted by the mobile system designers. Hance for methods like E-OTD and U-TDoA, additional equipments should be installed in the cellular networks to provide a posterior synchronization, and a pseudolite-like system. Considering the cost of these additional equipments, the future of cellular lateration-based methods is not evident in the LBS filed.

We saw that all the proximity sensing methods (like WiFi access point sensing and cellular Cell-ID) can provide a rough positioning of the terminal, which is not a refined positioning w.r.t. the system working range. However they are widely applied in the LBS context, since they are easy to implement. As an example, the recent localization tool of Google ("Google Latitude"), uses a combination of GPS and the proximity sensing methods. While it does not require any hardware modifications on the terminal, it combines the GPS positioning with WiFi access point sensing, and the cellular Cell-ID positioning. To do that, Google constructs a database of WiFi acces points and cellular antennas locations. These tens of millions of fixed locations enable Google Latitude to localize the mobile terminals, even in the absence of GPS.

The fingerprinting method may be considered as an enhancement of proximity sensing methods. In the context of cellular networks, fingerprinting provides a positioning accuracy notably improved w.r.t. Cell-ID, but yet it is quite inferior w.r.t. that of satellite position-

ing. However, the method is not completely without interest, since in some applications a rough ("public") localization may be preferable versus an accurate ("private") positioning.

The Table 2.1 gives some representative values for positioning accuracy of various technologies, for an outdoor urban context. The data are based on references [2], [27], and [5].

| Positioning method | accuracy | Net. or mobile based |
|---|---|---|
| A-GPS | 5 m - 30 m | Mobile based or assisted |
| Cell-ID | 100 m - 1 Km | Net. based |
| E-OTD | 50 m - 300 m | Mobile based or assisted |
| OTDoA | 50 m - 300 m | Mobile based or assisted |
| U-TDoA | 40 m - 50 m | Net. based |
| AoA | 100 m - 200 m | Net. based |
| Cellular LF | 50 m - 300 m | Net. based |

Table 2.1: Some represtative values for positioning accuracy in different methods ([2], [27], and [5])

## 2.7 Location fingerprinting in a machine learning viewpoint

Location fingerprinting is a positioning method that exploits the already existing infrastructures such as cellular networks ([47], [27]) or WLANs ([48], [35], [49], [50]). The principle of location fingerprinting systems consists in approximating the location of a mobile terminal based on its radio measurements, assuming that a training database is at disposal. This task may be cast in terms of statistical inference of a mapping function between the signal space and the location space, based on training data. The localization issue can thus be viewed as a typical *supervised learning* problem, in the *Machine Learning* viewpoint.

Machine Learning refers to the study of algorithms that improve automatically through experience. Applications range from data mining programs that discover general rules in large data sets, to information filtering systems that automatically learn users' interests ([51], [52]). Although the machine learning viewpoint is not widely adopted in the previous works in location fingerprinting literature, it is adopted in this thesis since we found it as an original framework. Figure 2.5 illustrates an overview of learning-based methods in fingerprinting literature.

From a machine learning perspective, location fingerprinting systems may be designed by using a *classification-based* or a *regression-based* approach. In the former case, the

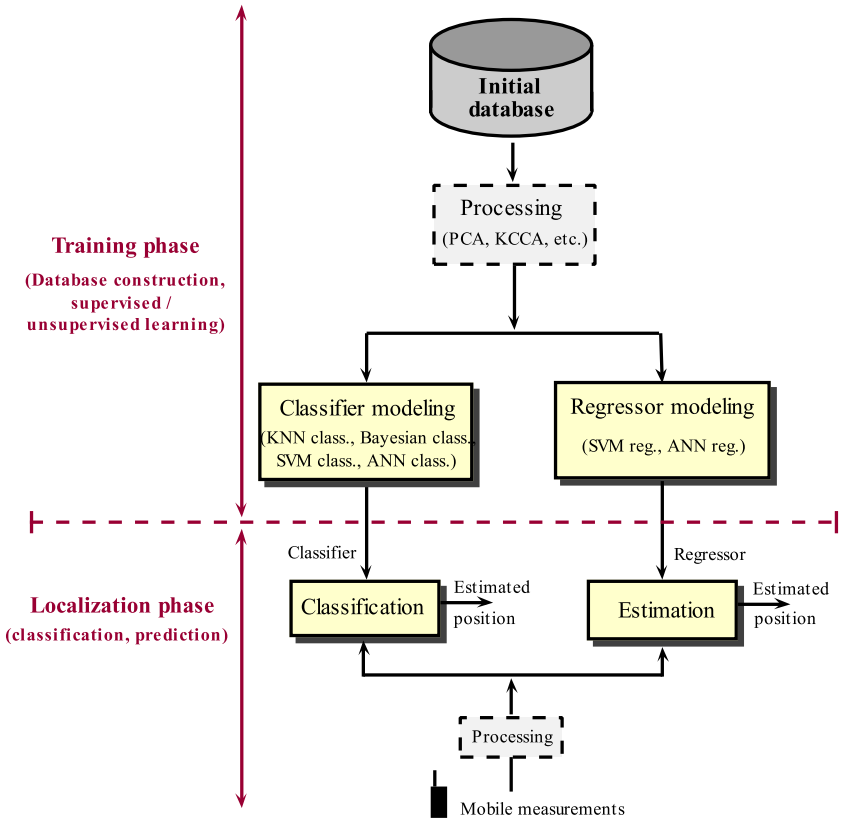



Figure 2.5: A schematic overview of learning-based methods for location fingerprinting systems

localization problem is treated as a classification task. A classification model is developed during the training phase (the "classifier modeling" step). The developed classifier is then used during the localization phase, to associate mobile measurements to one (or several) stored location(s) in the database. In the regression-based situation, a function that maps the radio signal space to the location space is learnt (the "regressor modeling" step) and then is used in the localization phase, to estimate the mobile position based on its radio measurements.

One basic classifier widely used in fingerprinting literature is the K-Nearest-Neighbors (KNN) method ([48], [36], [27]). In the KNN method, based on a pre-defined distance metric, any measurement in the localization phase is associated to the $K$ closest measurements stored in the training database. The interpolation of the corresponding location components is returned as the mobile position. The adopted distance metric and the number of interpolated neighbors $K$ are fixed during the classifier modeling step, beforehand. The probabilistic Bayesian classifier is another technique commonly used in fingerprinting sys-

tems ([34], [53], [49], [54]). The Bayesian approach treats the radio signal values as random variables that are statistically dependent on the location ([49]), and so may be modeled by a probability distribution function. Based on the developed model, for any measurement in the localization phase, the Maximum a Posteriori probability (MAP) estimate, i.e. the stored location corresponding to the highest value for the likelihood function, is returned as the mobile position. The Bayesian approach may include also a filtering process, which takes into account the mobile motion history or the area map, in order to provide a more coherent localization; yhis is done in works such as [55], [56], [34], and [57], where various filtering methods (e.g. Kalman, extended Kalman and particle filtering) have been implemented.

More advanced learning techniques are presented in works such as [47] and [58], where Support Vector Machines (SVMs) and Artificial Neural Networks (ANNs) are exploited. Both SVMs and ANNs are supervised learning methods, which may be used for regression or classification tasks. In [47], [59], [58] and [60], SVM and ANN regressors are implemented in the context of location fingerprinting. In both methods, the regressor learns a mapping function between the radio signal space and the location space, based on the provided radio database. The learned mapping function is then used during the localization phase to estimate the position of mobile terminals. Besides regression, the classification-based approach has also been implemented in [47] and [58]. To function as classifier, both SVMs and ANNs must be fed by some prior classes during the training phase. In the works mentioned above, the prior classes are provided by partitioning the radio database according to simple geographical patterns.

Some recent works such as [49] and [61] take up basic techniques like KNN, but they insert an extra "processing" step in the training phase of fingerprinting systems. This processing could be performed for various purposes. In [49] the authors propose to process the initial radio database by projecting it into a compact decorrelated signal space. The main advantage of this approach is the compression of the radio database by reducing the dimension of the radio signal space. Methods such as Principal Component Analysis (PCA) and Independent Component Analysis (ICA) have been exploited to project the radio signals into decorrelated spaces. While in these methods only the radio components are used to construct the new radio feature space, the authors in [61] propose a projection space which is constructed based on both radio and location components. Here a Kernel Canonical Correlation Analysis (KCCA) is exploited to project the vectors in a new space, where the correlation coefficient between the signal and location vectors is maximized. It is noteworthy that all the above techniques require an extra processing step for the mobile measurements, during the localization phase.

## Bibliography

[1] K. Pahlavan, *Wireless information networks*. Wiley, 2005.

[2] A. Kupper, *Location based services*. Wiley, 2005.

[3] J.D. Gibson, *The mobile communications handbook*. Springer, 1999.

[4] M.O. Gheorghita, A. Solanas, and J. Forne, "Location privacy in chain-based protocols for location-based services," in *Proceedings of the Third International Conference on Digital Telecommunications*, 2008, pp. 64–69.

[5] S. Wang, J. Min, B.K. Yi, "Location based services for mobiles: Technologies and standards," in *Tutorial in IEEE International Conference on Communications (ICC)*, June 2008.

[6] P. Bellavista, A. Kupper, and S. Helal, "Location-based services: Back to the future," in *IEEE Transactions on Pervasive Computing*, vol. 7, no. 2, 2008, pp. 85 – 89.

[7] J. Figueiras, S. Frattasi , *Mobile Positioning and Tracking: From Conventional to Cooperative Techniques*. Wiley, 2010.

[8] A.J. Blumberg, P. Eckersley, "On locational privacy, and how to avoid losing it forever," Electronic frontier foundation, Tech. Rep., August 2009.

[9] T. Ming, Q. Wu, Z. Guoping, H. Lili, and Z. Huan-guo, "A new scheme of LBS privacy protection," in *Proceedings of the 5th International Conference on Wireless communications, networking and mobile computing*, ser. WiCOM'09, 2009, pp. 5219–5524.

[10] European Parliament, "Directive 95/46/ec on the protection of individuals with regard to the processing of personal data and on the free movement of such data," Tech. Rep., October 1995.

[11] ——, "Directive 2002/58/ec on privacy and electronic communications," Tech. Rep., July 2002.

[12] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," SRI International, Tech. Rep., 1998.

[13] P. Samarati, "Protecting respondents identities in microdata release," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 13, no. 6, p. 1010–1027, 2001.

[14] L. Sweeney, "k-anonymity: a model for protecting privacy," *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, vol. 10, pp. 557–570, October 2002.

[15] W. Foy, "Position location solutions by Taylor series estimation," *IEEE Transactions on Aerospace Electron System*, pp. 187–193, 1976.

[16] L. Cong, W. Zhuang, "Non-line-of-sight error mitigation in TDOA mobile location," in *Proceedings of Global Telecommunications Conference (GLOBECOM)*, vol. 1, 2001, pp. 680 – 684.

[17] Y. Qi, H. Kobayashi, H. Suda, "Analysis of wireless geolocation in a non-line-of-sight environment," *IEEE Transactions on Wireless Communications*, vol. 5, no. 3, pp. 672 – 681, 2006.

[18] N. Bhagwat, L. Kunpeng, B. Jabbari, "Robust bias mitigation algorithm for localization in wireless networks," in *Proceedings of the IEEE International Conference on Communications*, May 2010, pp. 1–5.

[19] C.E. Cook and M. Bernfeld, *Radar Signals: An Introduction to Theory and Applications*. New York: Academic, 1970.

[20] Y. Qi; H. Kobayashi, "On relation among time delay and signal strength based geolocation methods," in *Proceedings of Global Telecommunications Conference (GLOBECOM)*, 2003, pp. 4079 – 4083 vol.7.

[21] D. Almodóvar, "Location technologies white paper," VF Group R&D Enablers Research, Tech. Rep., 2008.

[22] F. Duquenne, *GPS, localisation et navigation par satellite*. Hermes Science, 2005.

[23] G. M. Djuknic, R.E. Richton, "Geolocation and assisted GPS," *Computer*, vol. 34, pp. 123–125, February 2001.

[24] Y. Zhao, "Standardization of mobile phone positioning for 3G systems," *IEEE Transactions on Communications*, vol. 40, no. 7, pp. 108–116, July 2002.

[25] J. Wang, "Pseudolite applications in positioning and navigation: Progress and problems," *Journal of Global Positioning Systems*, vol. 1, pp. 48–56, 2002.

[26] H. S. Cobb, "GPS pseudolites: Theory, design and applications," Ph.D. dissertation, Stanford University, 1997.

[27] C. Dessiniotis, "Motive project deliverable 2.1," European Project FP6-IST 27659, Tech. Rep., September 2006.

[28] J.L. Dornstetter, "Brevet européen, numéro 90401043.6," MATRA communication, Tech. Rep., 1990.

[29] "3GPP TS 44.071," 3rd Generation Partnership Project; Technical Specification Group GSM/EDGE Radio Access Network; Location Services (LCS); Mobile radio interface layer 3 LCS specification (Release 10), Tech. Rep., 2011.

[30] "3GPP TS 44.035," 3rd Generation Partnership Project; Technical Specification Group GSM/EDGE Radio Access Network; Location Services (LCS); Broadcast network assistance for Enhanced Observed Time Difference (E-OTD) and Global Positioning System (GPS) positioning methods (Release 10), Tech. Rep., 2011.

[31] "3GPP TS 44.031," 3rd Generation Partership Project, Technical Specification Group GSM/EDGE Radio Access Network; Location Services (LCS); Mobile Station (MS) - Serving Mobile Location Centre (SMLC) Radio Resource LCS Protocol (RRLP) (Release 10), Tech. Rep.

[32] H. Laitinen, S. Ahonen, S. Kyriazakos, J. Lähteenmäki, R. Menolascino, S. Parkkila, "Cello project deliverable: Cellular location technology," European Project on Cellular network optimisation based on mobile location, IST-2000-25382-CELLO, Tech. Rep., November 2001.

[33] H Laitinen, J Lahteenmaki, T Nordstrom, "Database correlation method for GSM location," in *Proceedings of the International Conference on Vehicular Technology*, vol. 4, May 2001, pp. 2504–2508.

[34] M. Khalaf-Allah, K. Kyamakya, "Database correlation using bayes filter for mobile terminal localization in GSM suburban environments," in *Proceedings of the Conference on Vehicular Technology*, vol. 2, May 2006, pp. 798–802.

[35] B.D.S. Lakmali, W.H.M.P. Wijesinghe, K.U.M. De Silva, "Design, implementation and testing of positioning techniques in mobile networks," in *Proceedings of the International Conference on Information and Automation for Sustainability*, December 2007, pp. 94–99.

[36] D. Zimmermann, J. Baumann, M. Layh, F.M. Landstorfer, R. Hoppe, "Database correlation for positioning of mobile terminals in cellular networks using wave propagation models," in *Proceedings of the Conference on Vehicular Technology*, vol. 7, September 2004, pp. 4682 – 4686.

[37] Widyawan, M. Klepal, D. Pesch, "Influence of predicted and measured fingerprint on the accuracy of RSSI-based indoor location systems," in *Proceedings of the 4th workshop on positioning, navigation and communication*, 2007, pp. 145–151.

[38] R.S. Campos, L. Lovisolo, "Location methods for legacy GSM handsets using coverage prediction," in *Proceedings of the IEEE 9th Workshop on Signal Processing Advances in Wireless Communications*, July 2008, pp. 21 – 25.

[39] M. Anisetti, V. Bellandi, E. Damiani, S. Reale, "Advanced localization of mobile terminal," in *Proceedings of the International Symposium on In Communications and Information Technologies*, 2007, pp. 1071–1076.

[40] A.K.M. Mahtab Hossain, H. Nguyen Van, Y. Jin, W.S. Soh, "Indoor localization using multiple wireless technologies," in *Proceedings of the IEEE Internatonal Conference on Mobile Adhoc and Sensor Systems*, 2007, pp. 1 – 8.

[41] P. Wijesinghe, D. Dias, "Novel approach for RSS calibration in DCM-based mobile positioning using propagation models," in *Proceedings of the 4th International Conference on Information and Automation for Sustainability*, December 2008, pp. 29 – 34.

[42] F. Evennou, "Techniques et technologies de localisation avancées pour terminaux mobiles dans les environnements indoor," Ph.D. dissertation, L'université Joseph Fourier, 2007.

[43] H. D. Chon, "Using RFID for accurate positioning," in *Proceedings of the International Symposium on GNSS/GPS*, December 2004, pp. 32–39.

[44] H. Liu, H. Darabi, P. Banerjee, J. Liu, "Survey of wireless indoor positioning techniques and systems," *IEEE Transactions on systems, man and cybernetics*, vol. 37, no. 6, pp. 61 067–1080, 2007.

[45] S. Gezici, Z. Tian, G. B. Giannakis, H. Kobayashi, A. F. Molisch, H. V. Poor, and Z. Sahinoglu, "Localization via ultra-wideband radios, a look at positioning aspects of future sensor networks," *IEEE Transactions on Signal Processing*, vol. 22, no. 4, pp. 70–84, 2005.

[46] T. Gigl, G. J.M. Janssen, V. Dizdarevi, K. Witrisal and Z. Irahhauten, "Analysis of a UWB indoor positioning system based on received signal strength," in *Proceedings of the 4th Workshop on Positioning, Navigation and Communication*, 2007, pp. 97 – 101.

[47] C. Takenga, K. Kyamakya, "A low-cost fingerprint positioning system in cellular networks," in *Proceedings of the International Conference on Communications and Networking*, August 2007, pp. 915 – 920.

[48] P. Bahl, V.N Padmanabhan, "RADAR: an in-building RF-based user location and tracking system," *Proceedings of the Joint Conference of the IEEE Computer and Communications Societies*, vol. 2, pp. 775 – 784, March 2000.

[49] S. Fang, T. Lin, P. Lin, "Location fingerprinting in a decorrelated space," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 20, no. 5, pp. 685 – 691, May 2008.

[50] O. Baala, A. Caminada, "Wlan-based indoor positioning system: experimental results for stationary and tracking MS," in *Proceedings of the International Conference on Communication Technology*, November 2006, pp. 1–4.

[51] T. Mitchell, *Machine Learning.* McGraw Hill, 1997.

[52] D.J.C. MacKay, *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, 2003.

[53] M. A. Youssef, A. Agrawala, A. Udaya Shankar, "WLAN location determination via clustering and probability distributions," in *Proceedings of the Conference on Pervasive Computing and Communications*, March 2003, pp. 143– 150.

[54] H. Zang, F. Baccelli, J. Bolot, "Bayesian inference for localization in cellular networks," in *Proceedings of the 29th Conference on Information Communications*, March 2010, pp. 1963–1971.

[55] D. Fox, J. Hightower, L. Liao, D. Schulz, "Bayesian filters for location estimation," in *IEEE Transactions on Pervasive Computing*, vol. 2, July-Septembre 2003, pp. 24 – 33.

[56] F. Evennou, F. Marx, E. Novakov, "Map-aided indoor mobile positioning system using particle filter," in *Proceedings of the IEEE Conference on Wireless Communications and Networking*, vol. 4, March 2005.

[57] C. Contopoulos, "Motive project deliverable 5.1," European Project FP6-IST 27659, Tech. Rep., 2007.

[58] M. Brunato and R. Battiti, "Statistical learning theory for location fingerprinting in wireless lans," *Comput. Netw. ISDN Syst.*, vol. 47, pp. 825–845, April 2005.

[59] C. Takenga, K. Kyamakya, "A hybrid neural network-data base correlation positioning in GSM network," in *Proceedings of the IEEE International Conference on Communication Systems*, October 2006, pp. 1 – 5.

[60] Z. Wu, C. Li, J. Kee-Yin Ng, and K. R.P.H. Leung, "Location estimation via support vector regression," *IEEE Transactions on Mobile Computing*, vol. 6, no. 3, pp. 311 – 321, march 2007.

[61] J. Junfeng Pan, J. T. Kwok, Q. Yang, and Y. Chen, "Multidimensional vector regression for accurate and low-cost location estimation in pervasive computing," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 18, no. 9, pp. 1181–1193, September 2006.

# Chapter 3

# Location fingerprinting: A performance study for cellular systems

In the context of location fingerprinting, the Received Signal Strength (RSS) is widely used as the adopted parameter in the radio database. However there are only few studies that analyze the performance of RSS-based fingerprinting systems, as a function of physical parameters of the underlying environment.

Here we present an analysis based on simulated experiments. The analysis is based on the Mondrian radio propagation model. This propagation model enables us to include the layout of the clutters over the considered area, and hance to introduce certain degree of correlation for the shadowing effect. Based on the Mondrian model, we perform a performance analysis for an outdoor RSS-based fingerprinting system, implemented over a GSM or UMTS network.

## 3.1 Background and basic definitions

In location fingerprinting, a database of location dependent radio parameters is constructed during a training phase. Later, during the localization phase, a mobile terminal performs measurements to be compared with the values in the radio database, in order to yield position estimates. Various radio parameters may be used to construct the radio database. Today RSS information is widely used as the adopted parameter in location fingerprinting, since it does not require any additional hardware neither on the network nor on the terminal side. However there are few studies that analyze the performance of RSS-based fingerprinting systems, as a function of physical parameters of the underlying

environment. The analytic modeling of fingerprinting systems is a difficult task, due to the complexity of radio propagation. In some previous works such as [1] and [2], the authors try to develop some analytic approaches; however the presented models are based on many simplifying assumptions. For example, they do not take into account the shadowing effect and the relating properties, such as shadowing correlation.

Believing that the shadowing effect has an important impact on the performance of fingerprinting systems, here we present an analysis which allows us to take it into account. The analysis is based on the Mondrian radio propagation model; this propagation model enables us to include the layout of the clutters over the considered area, and hance to introduce certain degree of correlation for the shadowing effect. Based on the Mondrian model, we perform a performance analysis for an outdoor RSS-based fingerprinting system, implemented over a GSM or UMTS network. We examine the influence of physical parameters of the system on the achieved accuracy. Based on our obtained results, we provide a framework which may be useful for the design and implementation of location fingerprinting systems.

In this regard, we precise the terminology that we use hereafter in this thesis. In the following, a *radio database* is a set of *records*. A record (in this context) consists of two parts: a location part and a radio-system part. The "location part" describes the position of a specific point, and may contain geographical coordinates, floor labels, or some context information (e.g. indoor/outdoor); the "radio-system part" describes the radio measurement performed at this specific point. The radio measurement contains a number of parameters available from the radio interface technologies, such as RSS (from different base stations), TA or RTT, or the results of more complex processing such as path loss profiles. The radio measurement may be denoted by a vector $\underline{s} \in \mathbb{R}^{D_R}$, consisting of a number of $D_R$ real-valued scalar components. Similarly, the location part may be represented by a vector $\underline{x} \in \mathbb{R}^{D_G}$. As a result, a record may be described by $\underline{r} = (\underline{x}, \underline{s}) \in \mathbb{R}^D$, where $D = D_G + D_R$. The database density may be defined as the average number of records per surface unit.

## 3.2 System model

This section describes the simulation environment by introducing the adopted models for the radio propagation, measurements error and the fingerprinting system.

### 3.2.1 Propagation environment

We assume that the localization service is offered over a geographical area $\mathcal{A}$, which is covered by a GSM cellular network. The GSM cells are considered to be omnidirectional hexagonal with a radius of 1 km. The area $\mathcal{A}$ covers a surface of $B = 13$ cells (one reference central cell and two rings of neighbor cells).

To model the RSS measurements at an arbitrary location in the area, a radio propagation model is required. A general-purpose model gives the radio propagation as a one slope model with log-normal shadowing, as follows ([3]):

$$Pl_a(d) = -k + 10\alpha \log(d) + X_{sh}, \tag{3.1}$$

where $Pl_a$ is the average path loss (in dB), $k$ is a constant, $d$ is the transmitter-receiver distance, $\alpha$ is the path loss exponent, and $X_{sh}$ is a log-normal variable which models the shadowing effect. This model does not introduce any geographical consistency in the considered area. We use a log-normal shadowing model which induces some local correlations between the neighboring locations ([4]). This is done by considering a certain number of masks $\{\mu\}$ in the area $\mathcal{A}$ (as illustrated in figure 3.1). A mask is a line segment associated with an attenuation parameter $a(\mu)$ which is randomly drawn according to a log-normal law. For any transmitter-receiver link, the direct path $\pi$ is considered; the corresponding

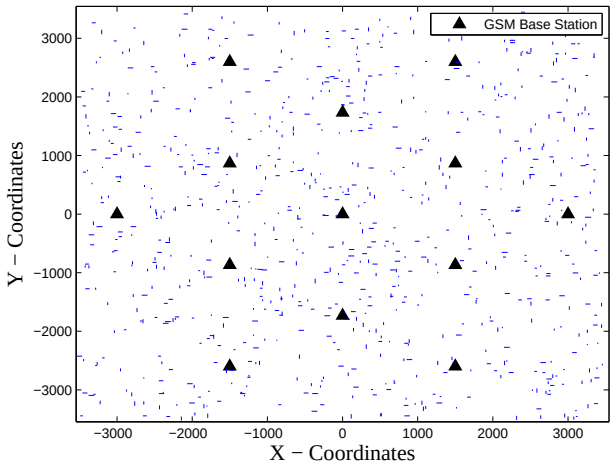



Figure 3.1: Random masks introducing shadowing effect

path loss in dB is modeled by:

$$PL_a(d) = -k_0 + 20 \log(d) + \sum_{\mu \in \mathcal{M}(\pi)} a(\mu) w(\mu, \pi), \tag{3.2}$$

where the first two terms give the free space path loss, $\mathcal{M}(\pi)$ is the set of masks intersecting the path $\pi$, and $w(\mu, \pi)$ is a weighting factor which may be deduced from the harmonic series. It is noteworthy that once the masks are drawn in the area, the shadowing effect becomes deterministic and the $PL_a$ gets a fixed value for any two points in the area. Computer simulations confirm that the statistics of $PL_a$ over a circle of radius $d$ around a fixed transmitter show a log-normal behavior. Besides the average path loss in terms of the distance $d$ demonstrates a one slope behavior (where the slope correspond to the path loss exponent $\alpha$). Therefore the adopted model is quite consistent with the traditional one presented in relation (3.1).

### 3.2.2 Measurements error

Once the path loss is determined according to Equation (3.2), the average received signal power is given by $s = P_T - PL_a$, where $P_T$ represents the transmitted power. However the instantaneous measurements performed by a mobile terminal are not equal to this average value. The deviation, which we call it the measurements error, is mainly due to the signal temporal variations and the receiver hardware uncertainty ([5]).

The RSS temporal variations are traditionally modeled by a log-normal distribution ([6]). Accordingly, in this work we use a log-normal random variable to model these variations. Concerning the hardware uncertainty, on the other hand, there is not any general model to be used. Here we assume that the major impact of the terminal hardware is to add up a constant offset value to the measurements, and we use a constant term to represent this offset. Putting all together, we model the mobile equipment measured RSS by:

$$s^{(ME)} = P_T - PL_a + X_{ME} + c_{ME}, \tag{3.3}$$

where $X_{ME} \sim \mathcal{N}(0, \sigma^2_{ME})$ is a gaussian random variable (in dB) with standard deviation of $\sigma_{ME}$ which denotes the measurements temporal variations, and $c_{ME}$ is a constant term that stands for mobile offset value.

It is noteworthy that in general the offline measurements are more accurate than that of the online phase, since they are usually the average of several samples taken at the same place. Therefore we model the database RSS measurements by:

$$s^{(DB)} = P_T - PL_a + X_{DB} + c_{DB}, \tag{3.4}$$

where $X_{DB} \sim \mathcal{N}(0, \sigma^2_{DB})$ is a gaussian random variable (in dB) which denotes the measurements temporal variations, and generally $\sigma_{DB} < \sigma_{ME}$; $c_{DB}$ stands for database measurements offset.

### 3.2.3 Fingerprinting system

Assume that the localization service is offered over an area $\mathcal{A}$, where a number of $B$ base stations are present. To construct the data base, the geographic area is covered by a uniform grid which consists of $M$ square zones. The grid *resolution* $g$ is defined as the length of a side of each square zone. The radio database $\mathcal{R}$ is then a set of $M$ records, given by:

$$\mathcal{R} = \{(\underline{x}_m, \underline{s}_m)\}_{m=1...M}. \tag{3.5}$$

with one record for each grid zone. Any geographical location $\underline{x}_m$ in the area can be described by a 2-dimensional vector (i.e. $D_G = 2$). A radio measurement vector $\underline{s}_m$ in this context is described by:

$$\underline{s}_m = (s_{m1}, ..., s_{mb}, ..., s_{mB}) \in \mathbb{R}^B, \tag{3.6}$$

where $s_{mb}$ is the RSS level of the $b$-th base station at location $\underline{x}_m$ (i.e. $D_R = B$).

During the training phase, one common method to obtain the fingerprints $\underline{s}_m$ is to perform $N_P$ raw measurements at $N_P$ different points in the corresponding grid zone, and to consider their average as a single fingerprint. Although this method is not necessarily the most optimized approach, but it has been adopted here, since it is simple enough and allows us to make our desired comparisons. Here the database fingerprints have been constructed by averaging $N_P = 5$ different measurements in each grid zone. The single measurements are generated according to the Equation (3.4).

During the localization phase, the mobile terminal performs a sample RSS measurement $\underline{s}'$ at location $\underline{x}'$, which is generated according to the Equation (3.3). In order to localize the mobile terminal, the basic $K$ Nearest Neighbors (KNN) algorithm is adopted. As mentioned in section 2.7, in the KNN method, based on a pre-defined distance metric, any measurement in the localization phase is associated to the $K$ closest measurements stored in the training database. The interpolation of the corresponding location components is returned as the mobile position. Here, two types of metrics have been used to implement the KNN method. The first type is the common Euclidian distance. Here, we compute the Euclidian distance between the terminal measurement and the stored fingerprints $d_m = \|\underline{s}' - \underline{s}_m\|$. The $K$ fingerprints with smallest distances are selected, and the average of their corresponding locations is returned as the mobile location.

The second type of metric used in this work is the normalized correlation coefficient. Here we compute the normalized correlation coefficient between the terminal measured RSS vector and the stored fingerprints, as follows:

$$\varrho_m = \frac{< \underline{s}' . \underline{s}_m >}{\|\underline{s}'\| . \|\underline{s}_m\|},$$

where $< . >$ denotes the inner product operator. Then, the $K$ fingerprints with the largest correlation coefficients are used to estimate the mobile location. We note that the normalized correlation coefficient is not mathematically a distance metric, since it does not satisfy all the required conditions (e.g. the triangle inequality). Once the mobile terminal is localized, the localization error is defined as $\varepsilon(\underline{x}') = \|\underline{x}' - \widehat{\underline{x}}\|$, where $\widehat{\underline{x}}$ denotes the estimated position of the terminal.

In the actual mobile terminals the maximum number of scanned base stations in a measurement vector is restricted by an upper bound $B_{max}$ (in GSM standard $B_{max}$ = 7). Moreover, the terminal receiver can detect only the RSS values that are higher than a predefined threshold $\lambda$ (in GSM standard $\lambda$ = -110 dBm). Therefore, in practice, a measurement vector $\underline{s}$ does not contain the signal components concerning all the $B$ base stations in the area. The undetected components may be considered as *missing data*. To simulate this effect in this work, we compute the components concerning the seven strongest base stations according to our radio model, and we consider all the other components as unknown values. The optimal handling of missing data in radio measurements will be treated later in chapter 6. Here, as a simple approach to treat the problem, we set all the missing values at the receiver minimum detectable level $\lambda$.

For both types of distance metrics, the input RSS vectors may be expressed in dBm or in their natural unit (Watt). Here we have adopted the dBm implementation since it showed a better performance during our pre-computations.

## 3.3 Performance analysis

In this section we analyze the influence of several physical parameters on the performance of the positioning system. The parameters of interest are the path loss exponent, the measurements error and the grid resolution. Initially, some reference values are assigned to these parameters. We consider a path loss exponent of 3.7, a grid resolution of 200 m and offset values of zero for both mobile and database measurements. Moreover, we assume $\sigma_{DB}$ = 0 dB, $\sigma_{ME}$ = 3 dB as the reference values. These initial values are modified later during the simulations, in order to study their impact on the system performance.

### 3.3.1 Impact of the path loss exponent

Here we investigate the accuracy of the positioning system for environments with different values of path loss exponent ($\alpha$). In order to control the value of $\alpha$ in the simulations, we adjust the number of masks in the simulated area $\mathcal{A}$. The grid resolution and the offset

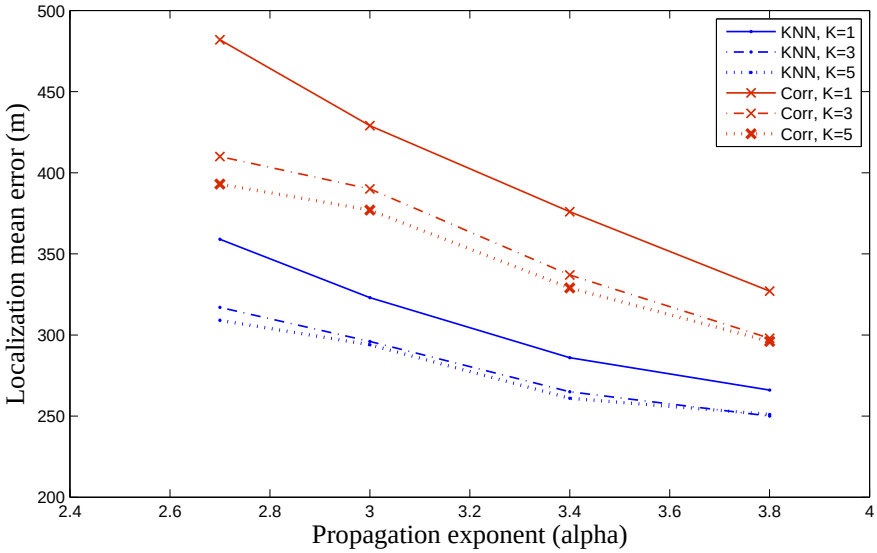
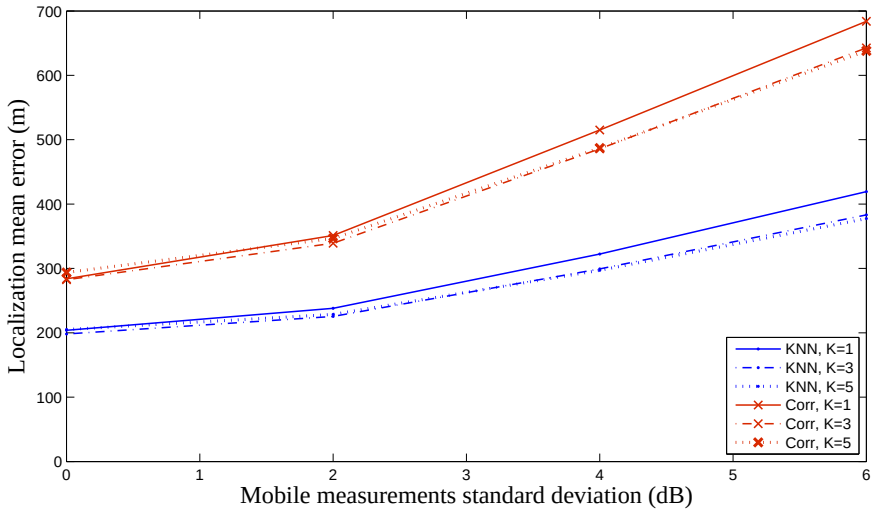



Figure 3.2: Positioning accuracy versus the propagation exponent

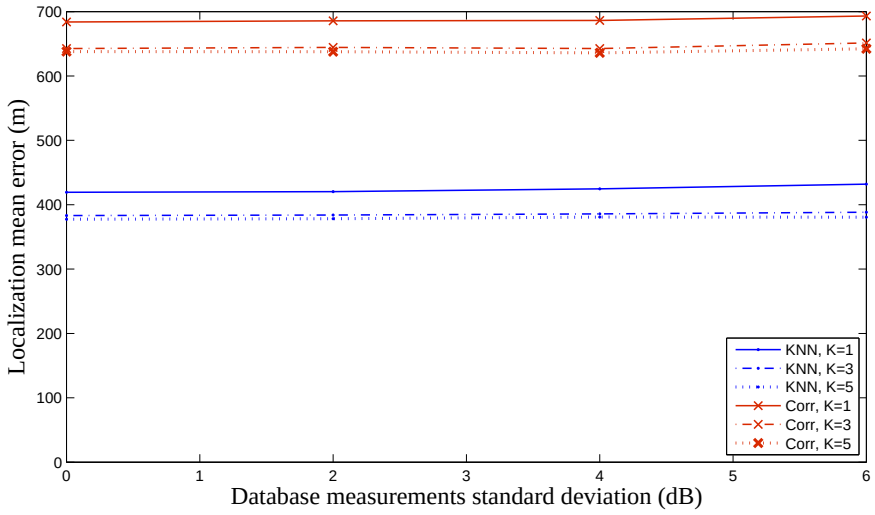
parameters are set at their reference values and for the measurements variation the context of scenario $b$ has been adopted. Figure 3.2 depicts the corresponding results.

We observe that in general, a higher value of $\alpha$ leads to a lower positioning error. According to figure 3.2, as $\alpha$ grows from 2 to 3.5 the performance of all the algorithms improves significantly. For higher values of $\alpha$ the improvement is not so notable, such that we can see in the case of KNN algorithms the positioning error remains almost constant at the tail of the curves. Based on these results, we may deduce that the fingerprinting system provides a more accurate position in the dense urban areas with respect to the sparse rural regions. This can be justified intuitively by the fact that in the former case, the severe multipath effect makes the RSS vectors more diversified and distinguishable, and hence this leads to a more accurate localization.

Another remark according to the figure 3.2 is that the correlation-based technique is quite outperformed by the KNN method. We will observe the same effect during the experiments in the next sections. The low performance of the correlation-based algorithm stems from the fact that it does not conserve the signals energy level. Here, we see that the energy conservative methods like KNN have a better performance for matching the radio signals in the fingerprinting context.




(a) Effect of $\sigma_{ME}$ ($\sigma_{DB} = 0$ dB)




(b) Effect of $\sigma_{DB}$ ($\sigma_{ME} = 6$ dB)

Figure 3.3: Positioning accuracy versus the measurements variations

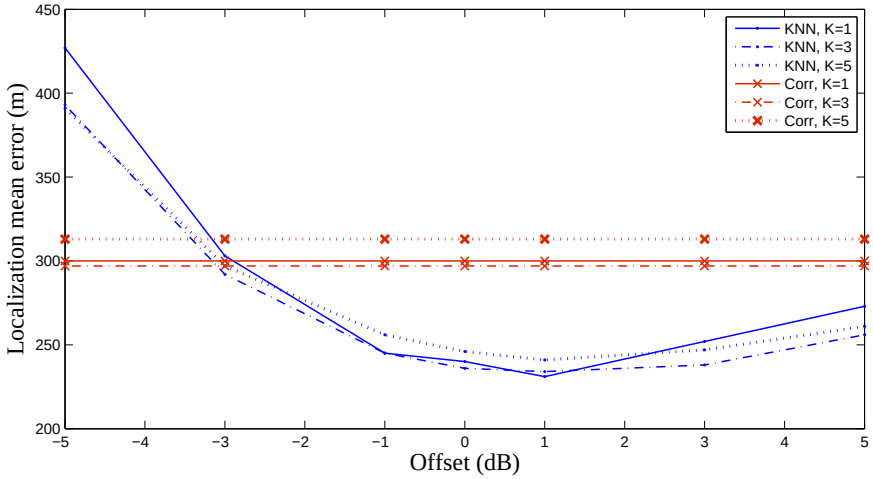


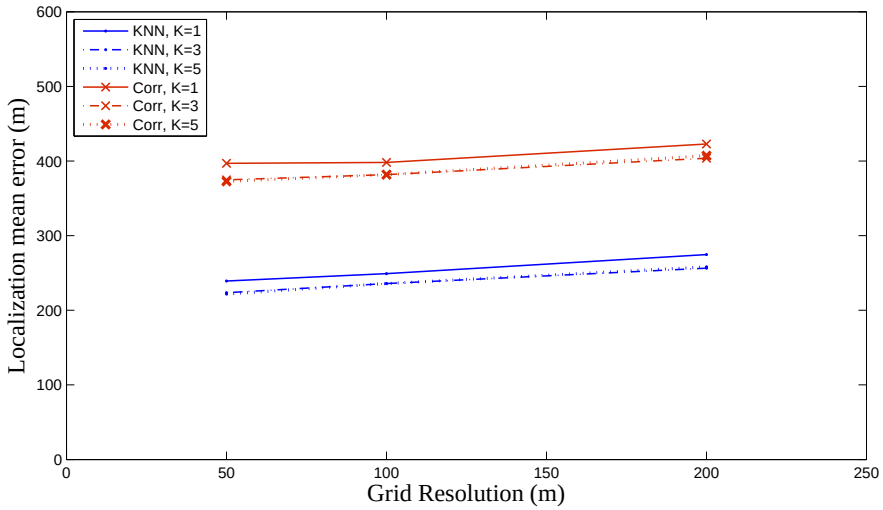

Figure 3.4: Positioning accuracy versus the measurements offset

### 3.3.2 Impact of the measurements error

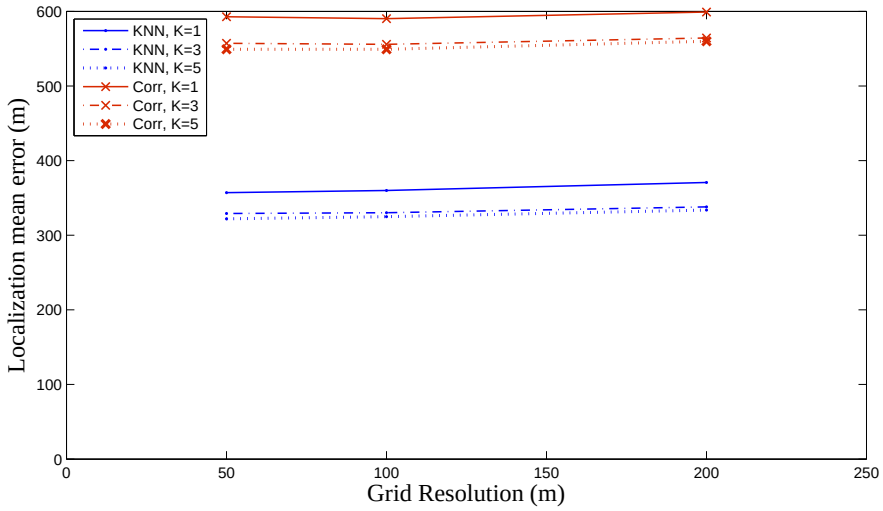
According to our model, the measurements error consists of two elements: temporal variations and the hardware offset. At the first time we examine the former, while setting the latter equal to zero. Other parameters are kept at their reference values. We present the effect of the database variations ($\sigma_{DB}$) and the mobile equipment variations ($\sigma_{ME}$) separately, in figures 3.3-a and 3.3-b.

We note that in the case of mobile equipment, increasing $\sigma_{ME}$ degrades the performance dramatically. On the other hand effect of $\sigma_{DB}$ is rather negligible, since an average of $N_P = 5$ measurements has been used to construct each database fingerprint. In the same way, we can mitigate effect of $\sigma_{ME}$ by averaging several measurements during the online phase. But this solution is not suitable for real-time navigation systems which need a rapid calculation of the position.

At the second time, we consider the effect of the measurements offset on the localization process. Since offset is a relative concept, we fix the value of $c_{DB}$ at 0 dB and we only consider $c_{ME}$ as the varying parameter. The other parameters are set at their reference values and the scenario $b$ has been adopted for the measurements variations. Figure 3.4 gives positioning error for different values of $c_{ME}$. It is apparent that the correlation-based algorithms completely eliminate the offset effect. In the case of KNN algorithms, an offset value of $\pm 1$ dB has a slight impact, but higher values of offset degrade the performance. However, although the correlation-based algorithms are insensitive to offset, but yet in most cases they are outperformed by the KNN method (we can see that for offset values




(a) Scenario $a$ : $\sigma_{DB} = 0$ dB, $\sigma_{ME} = 3$ dB




(b) Scenario $b$ : $\sigma_{DB} = 2$ dB, $\sigma_{ME} = 5$ dB

Figure 3.5: Positioning accuracy versus grid resolution

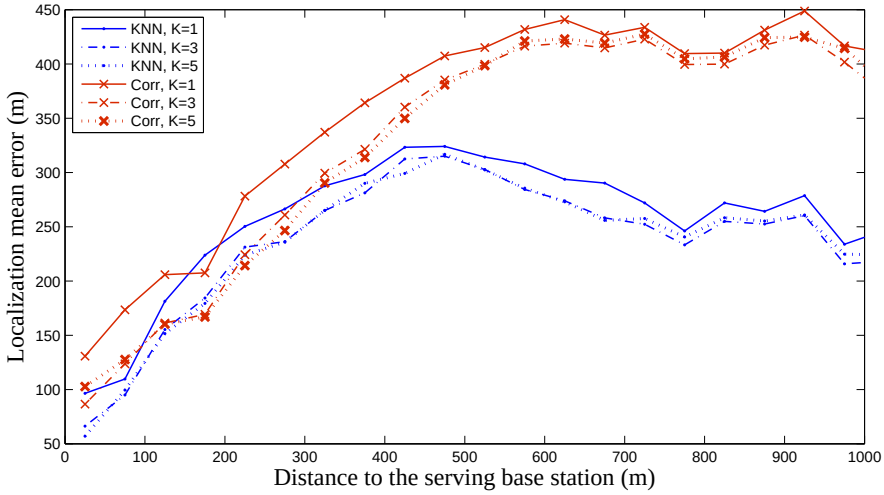



Figure 3.6: Positioning accuracy versus distance to the serving base station (cell radius = 1 km)

up to ±3 dB, the KNN method is always more effective). We deduce that applying a correlation-based method is not preferable unless in situations where a large value of offset is presumed to exist.

### 3.3.3 Impact of the grid resolution

The parameter of interest in this section is the grid resolution. We examine the system performance for grid resolutions of 50 m, 100 m and 200 m. In order to have more general results we repeat the experiment for two scenarios:

- Scenario a : $\sigma_{DB}$ = 0 dB, $\sigma_{ME}$=3 dB (where the reference values are adopted),
- Scenario b : $\sigma_{DB}$ = 2 dB, $\sigma_{ME}$=5 dB.

The parameter $\alpha$ and the offsets are set at their reference values. Figure 3.5 illustrates the obtained results.

It is notable that enhancing the resolution improves the accuracy in situations where the measurements are not too erroneous. As we see by enhancing the resolution in the scenario $a$ (which is a less noisy scenario), the accuracy improves. In the scenario $b$ (which is a more noisy case), enhancing the resolution does not improve the performance at all. We may conclude that in practice the influence of the grid resolution is dominated by the

measurements error level, and enhancing the resolution necessarily does not improve the system performance.

In the figure 3.5 we presented the average positioning error of the mobile terminals, regardless of their position in the GSM cell. Now we examine the results in more details by taking into account the distance of the mobile terminals to the serving base station. In figure 3.6, the positioning error is displayed as a function of the terminal distance to the serving base station. The figure presents the results corresponding to scenario $a$ ($\sigma_{DB} =$ 0 dB, $\sigma_{ME} = 3$ dB) with a grid resolution of 200 m (other cases are not presented, since they show the same trends).

As we see in the figure, the same style is followed in all the three cases. At first the positioning error grows almost linearly until it reaches a maximum value. In this interval the KNN method and the correlation-based technique demonstrate comparable results. The increase of error does not hold for further distances, such that the error remains almost stable at the end of the curves. Here the correlation-based techniques are quite outperformed by the KNN method. In the latter case, the positioning error is almost limited to $\frac{1}{3}$ of the cell radius. In accordance with our previous observation, we note that the maximum positioning error does not change considerably for different values of grid resolution.

## 3.4 Conclusion

In this chapter through extensive simulations, some characteristics of cellular RSS-based fingerprinting systems are presented.

We observed that the fingerprinting system provides more accurate positions in the dense urban areas with respect to sparse rural regions, since in the latter case the received RSS vectors are not sufficiently distinguishable. The experiments showed that the measurements accuracy is a key factor in the localization process. In environments with severe signal fluctuation the positioning accuracy degrades significantly. In such environments even enhancing the grid resolution does not improve the system performance. Moreover, we noted that in a GSM cell, by moving away from the serving base station the error grows linearly reaching a maximum, and then it becomes almost stable.

In most cases KNN algorithm outperformed the correlation-based technique. However since the latter eliminates any values of offset, one may imagine a "combined" algorithm in future studies, in order to take advantage of both techniques.

Among the provided results, the one that constructs the axis of our next works, is the impact of the grid resolution. According to the obtained results, we deduced that a finer

database does not necessarily improve the performance. Therefore, one may demand for methods that allow to provide a database of an optimal size, while keeping an optimal level of performance. This issue builds up the next steps of the thesis, presented in the next chapters.

## Bibliography

[1] J. Yang, Y. Chen, "A theoretical analysis of wireless localization using RF-based fingerprint matching," *Proceedings of the IEEE International Symposium on Parallel and Distributed Processing*, 2008.

[2] K. Kaemarungsi, P. Krishnamurthy, "Modeling of indoor positioning systems based on location fingerprinting," in *Proceedings of the twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 2, March 2004, pp. 1012– 1022.

[3] T. Rappaport, *Wireless Communications, principles and practice.* Prentice Hall PTR, 2002.

[4] Philippe Godlewski, "The Mondrian propagation simulation model," in *Vehicular Technology Conference*, May 2011.

[5] M.B Kjaergaard, "A taxonomy for radio location fingerprinting," in *Proceedings of the International Symposium on Location and Context Awareness*, vol. 4718, October 2007, pp. 139–156.

[6] A Kushki, KN Plataniotis, AN Venetsanopoulos, CS Regazzoni, "Radio map fusion for indoor positioning in wireless local area networks," in *Proceedings of the International Conference on Information Fusion*, vol. 2, July 2005, pp. 1311–1318.

# Chapter 4

# Cluster analysis for radio database compression

## 4.1 Introduction: cluster analysis for location fingerprinting

In the context of location fingerprinting systems, the radio database may be constructed by using either empirical measurements, or theoretical modeling tools, or a hybrid approach where a limited number of empirical measurements are performed to calibrate the theoretic propagation models ([1]). The empirical data may be obtained by conducting specific measurement campaigns, or by using the databases that are already at the disposal of the network operators (e.g. databases arisen from network monitoring tools).

The "size" of the radio database is an important aspect regarding the database construction, specially in mobile-based fingerprinting systems. Generally, an under-trained database (containing a low number of measurements), leads to a degraded performance in fingerprinting systems ([2]). In works such as [2], [3] and [4] some methods are proposed to enrich the database, by predicting theoretically the signal values at some new locations. On the other hand, regarding the chaotic nature of radio signal, an over-trained database does not bring further improvement to the positioning accuracy.

It is noteworthy that, the size of the radio database is an influential factor in regards to issues such as *computation* and *transmission* loads. In mobile-based fingerprinting systems the computation load is of great importance since it affects directly the terminal autonomy and the CPU computing load. Regarding the recent demand for energy efficient networks and the emergence of issues like green networking, reduction of the computation load may be a figure of merit in fingerprinting systems. Moreover in a mobile-based approach, the

radio database and its updated versions are transferred to the terminal through the cellular network. In the case of a unicast solution for signaling between the network and the mobile, a lower transmission load could be clearly desirable. This issue may be also considered as a memory saving problem on the terminal side.

Regarding the above issues, some methods have been proposed in the literature of location fingerprinting systems, which aim to compress the radio database ([5], [6], [7], [1], [8]); the compression quality in this context is evaluated by the resulting positioning error. As the database statistical properties may depend on the underlying radio system, one may expect different methodologies for different radio networks. One suggested approach for database compression, specially in the context of WLAN fingerprinting, is to reduce the dimension of the radio feature space ([5], [6] and [7]). Various techniques such as Principal Component Analysis (PCA) and Kernel Canonical Correlation Analysis (KCCA) have been proposed to implement this approach.

An alternative solution might be envisaged by reducing the number of records, i.e. to reduce the database density. One simple way to reduce the database density (used in [1], in the context of cellular systems), is to cover the considered area by a uniform grid, and to perform an averaging function over all the measurements which fall in the same grid zone. The grid resolution is defined as the length of a side of each square zone. We notice that in the griding method, selection of the gathered measurements depends only on their location parts; the radio parts do not intervene in the grouping procedure. However, the method has the advantage that it allows to reconstruct the geographical coordinates of the grid, just by knowing a single grid node (or a single zone center) and the grid resolution.

In this chapter we use some *clustering* techniques for radio database compression, which take into account both the location and the radio parts of the recorded measurements. Clustering is an unsupervised learning method ([9]). In cluster analysis, a set of objects is split into a number of homogeneous subsets, based on an often subjectively chosen measure of similarity ([10], [11]). Cluster analysis has been used since long in various fields such as image processing, biology, business and economy, etc. ([12]). In the context of our problem, clustering techniques are expected to extract consistent geographic zones which are homogenously covered by the radio signal. The clustering quality here is evaluated by the resulting positioning error for the fingerprinting system. The considered techniques include the classic *k-means* and the *minimum-variance based hierarchical clustering* algorithms. The algorithms are applied in a concatenated "location-radio" signal space. A generalized form of Euclidian distance is adopted, which allows to attribute different weight factors to the location and radio parts in the concatenated vectors. It is noteworthy that although in this work we focus on the radio database clustering for location fingerprinting, the proposed method may be used in more general applications, e.g. network planning, cognitive radio,

etc.

## 4.2 Cluster analysis

*Clustering* is an unsupervised learning method ([9]). In cluster analysis, a set of objects is split into a number of homogeneous subsets, based on an often subjectively chosen measure of similarity ([10], [11]). The clustering process is expected to create the subsets, such that the similarity between objects within a subset is larger than the similarity between objects belonging to different subsets ([9], [10], [11]).

As a mathematical description, given a set of $N$ data points $\mathcal{R}^\circ = \{\underline{r}_n^\circ\}_{n=1\ldots N}$ in a $D$-dimensional feature space ($\underline{r}_n^\circ \in \mathbb{R}^D$), a clustering technique attempts to divide $\mathcal{R}^\circ$ into $M$ ($M < N$) subsets or clusters, so that the members in the same cluster are similar in some sense. Two major classes of clustering techniques are the hierarchical and the partitional techniques ([13], [14]).

Hierarchical clustering is a technique that constructs a tree-like nested structure of clusters ([10]). Hierarchical algorithms may be implemented as agglomerative or divisive. In the agglomerative variant, one starts by considering each data point ($\underline{r}_n^\circ$) as a single cluster, and follows by merging two neighboring clusters at each step of the process ([13], [14]). The neighboring clusters are chosen based on a *linkage* criterion, that determines the distance between two clusters. On the other hand, in divisive variant one considers all the data points in a single cluster, and follows by splitting recursively the existing clusters as moving down in the hierarchy.

In the second class of clustering techniques (the partitional methods), there is no notion of hierarchy concerning the provided clusters. The clustering algorithm provides a partitioning of $\mathcal{R}^\circ$ into $M$ clusters, represented by a $M \times N$ matrix $U = [u_{mn}]$ that satisfies the following conditions ([15]):

$$u_{mn} \in \{0, 1\}, \tag{4.1a}$$

$$\sum_{m=1}^{M} u_{mn} = 1; \text{ for } 1 \leq n \leq N, \tag{4.1b}$$

$$\sum_{n=1}^{N} u_{mn} > 0; \text{ for } 1 \leq m \leq M, \tag{4.1c}$$

Here, condition 4.1c avoids existence of empty clusters; the first and second conditions ensure that a single data point belongs only to a single cluster ("hard partitioning"). Altering the condition 4.1a to

$$u_{mn} \in [0, 1],$$

allows a "soft" or "fuzzy" partitioning of $\mathcal{R}^\circ$ with partial memberships, which is not in the scope of this study (for more details see [15]).

The partitions are generally provided by optimizing a pre-defined objective function. One common objective function in the literature is the sum of square errors function ([16]), defined as follows:

$$J_1(U, \mathcal{R}) = \sum_{n=1}^{N} \sum_{m=1}^{M} u_{mn} d^2(\underline{r}_n^\circ, \underline{r}_m). \tag{4.2}$$

where $\mathcal{R} = \{\underline{r}_m\}_{m=1,\ldots,M}$ is a set of $M$ vectors representing the centroids of the $M$ clusters, and $d(\underline{r}_n^\circ, \underline{r}_m)$ is a distance or dissimilarity measure between the $n$-th data point and the $m$-th centroid. The clustering algorithm tries to find the pair $(U, \mathcal{R})$ that minimizes the above objective function. The input data points to the algorithm are generally normalized, so that all components vary on the same dynamic scale. Although this criterion does not perform well for all kinds of data, it is widely used in clustering literature ([16]). We note that by taking Euclidian metric as the adopted distance, the function $J_1$ becomes a measure of total intra-cluster variance of the provided partitions ([15]). Some other proposed distance metrics are presented in [10].

A famous method for approximating the minimum of $J_1$, in the case of Euclidian distance, is the *k-means* algorithm ([10], [16]). The k-means algorithm solves the optimization problem by iterating the partial minimization steps ([16]). It is noteworthy that the k-means algorithm does not provide a global solution of the optimization problem, but it converges to a local solution. The local minimum provided by k-means algorithm is strongly dependent on the choice of the initial centers at the beginning of the algorithm. Recently some enhanced algorithms have been proposed (e.g. "global k-means" algorithm in [17]), to improve the search properties of the classic k-means method.

## 4.3 Radio database clustering

In this section, at first we present the definitions and the notations. Afterwards, we go on by introducing the adopted clustering algorithms.

### 4.3.1 Concept and notations

In coherence with our definitions in the previous chapter, a radio database $\mathcal{R}$ is a set of *records*. A record is a vector $\underline{r} = (\underline{x}, \underline{s}) \in \mathbb{R}^D$, where $\underline{x} \in \mathbb{R}^{D_G}$ represents a geographical position, and $\underline{s} \in \mathbb{R}^{D_R}$ is the corresponding measurement vector in radio feature space, and

we have $D = D_G + D_R$. The included positions in the database may be numbered from 1 to $M$ and given by the set:

$$\chi = \{\underline{x}_1, ..., \underline{x}_m, ..., \underline{x}_M\}. \tag{4.3}$$

The radio database is then given by $\mathcal{R} = \{\underline{r}_m\}_{m=1...M}$. It is noteworthy that the parameters stored in the records may belong to different natures and be measured in different units. We define a *feature type* as all the stored parameters in a record that belong to the same nature. In the simplest case, there are two different feature types in each record: location feature type, and a single-RAT (Radio Access Technology) RSS feature type. One may imagine more complicated cases where a larger number of feature types are included in the records, e.g. multi-RAT RSS measurements, time advance information, etc. Therefore a record may also be described as follows:

$$\underline{r} = (\underline{\rho}_1, ..., \underline{\rho}_h, ..., \underline{\rho}_{N_f}),$$

where $N_f$ denotes the number of feature types in each record, and $\underline{\rho}_h$ is the sub-vector concerning the $h$-th feature type.

The database $\mathcal{R}$ may be provided by processing the elements of an *initial* radio database; a radio database constructed according to raw field measurements is called an initial database $\mathcal{R}^\circ$. A record of $\mathcal{R}^\circ$ is given by $\underline{r}^\circ = (\underline{x}^\circ, \underline{s}^\circ)$. The raw measurements are performed at geographical positions called *elementary points*, given by the set $\chi^\circ = \{\underline{x}_1^\circ, ..., \underline{x}_n^\circ, ..., \underline{x}_N^\circ\}$. The aim of this study is to compress the initial database $\mathcal{R}^\circ = \{\underline{r}_n^\circ\}_{n=1...N}$ by applying a clustering technique, in order to gather the elementary points in $\chi^\circ$ into homogeneous zones or clusters, in a more compact radio database $\mathcal{R} = \{\underline{r}_m\}_{m=1...M}$ $(M < N)$.

Figure 4.1 illustrates our proposed architecture for database compression in location fingerprinting systems. By inserting a clustering step in the training phase, the initial database could be divided into $M(M < N)$ subsets or clusters with the $m$-th cluster described by:

$$C_m = \{\underline{x}_n^\circ\}_{n \in \mathcal{N}(m)}, \tag{4.4}$$

where $\mathcal{N}(m)$ is the set of associated elementary points. The centroids of the clusters may be used to construct the compressed database $\mathcal{R}$. We define the compression index $\eta$ as the ratio between the size of the compressed and the initial databases. "Size" is defined as the number of records in the database times the corresponding dimension; thus in the case of database clustering we have $\eta = M/N$.

The proposed clustering step may be categorized as a "processing" method, in comparison to the state-of-the-art architectures in Figure 2.5; the difference is that it does not require any additional processing for the mobile measurements during the localization phase. In this work we propose to apply the clustering techniques in an extended

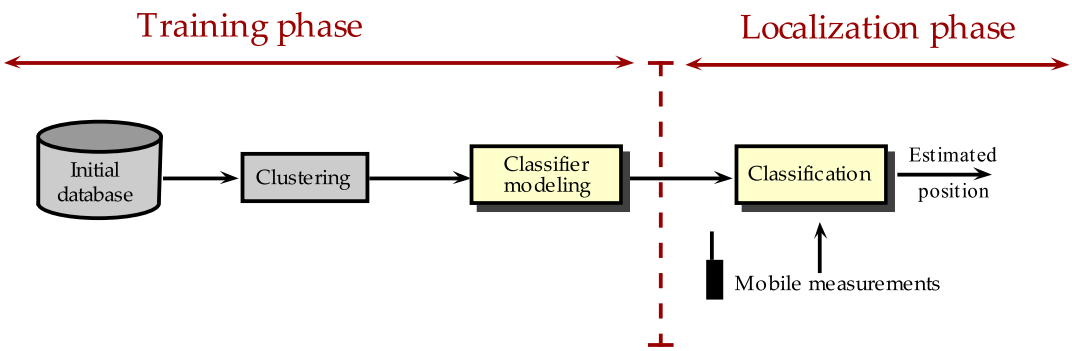



Figure 4.1: The proposed architecture for the fingerprinting system, introducing the clustering process

location-radio signal space. The adopted clustering algorithms are presented in the next section.

### 4.3.2 Clustering algorithms

Different clustering techniques consider different criteria to perform the partitioning of a given set. In the context of location fingerprinting we adopt the common criterion of sum of square errors minimization, which aims to minimize the objective function given by Equation (4.2). One important issue at this point is the choice of the distance metric $d(.,.)$.

In this work, we rely on the widely-used Euclidian distance. More precisely, we use a generalized weighted form of Euclidian distance, given as follows:

$$d^2_{E(\underline{w})}(\underline{r}^\circ_n, \underline{r}_m) = \sum_{h=1}^{D} w_h \|r^\circ_{n,h} - r_{m,h}\|^2, \tag{4.5}$$

where $\underline{w} = [w_1, ..., w_D] \in \mathbb{R}^D$ is the vector containing the weight factors. Adopting this distance metric, the objective function to be minimized is given by:

$$J_2(U, \mathcal{R}) = \sum_{n=1}^{N} \sum_{m=1}^{M} u_{mn} d^2_{E(\underline{w})}(\underline{r}^\circ_n, \underline{r}_m). \tag{4.6}$$

The function $J_2$ may be considered as a measure of the total intra-cluster variance of the provided partitions. The used weighted distance $d_{E(\underline{w})}$ allows to control (via vector $\underline{w}$) the relative importance of different components in the total variance. We note that here

the weight factors are considered as input variables, and are expected to be defined before performing the clustering.

As we mentioned in the previous section, one common technique to solve the above optimization problem is the k-means algorithm. The k-means algorithm tries to find the solution by iterative partial minimization steps, the so-called *alternating optimization method* ([16]). Although it has been shown to be efficient in many cases, the k-means algorithm faces two major problems. The first one is that the k-means algorithm does not provide a global solution of the above optimization problem, but it finds a local solution. The local minimum provided by k-means algorithm is strongly dependent on the choice of the initial centers at the beginning of the algorithm. The second problem of k-means is the generation of empty clusters. Generally for large values of $k$ there may be many empty clusters in the results, and this phenomenon degrades the clustering performance.

Regarding the problems of k-means clustering, we propose to use also a hierarchical clustering technique. As we mentioned before, hierarchical clustering is a technique that constructs a hierarchy of clusters. In the "agglomerative" hierarchical methods, one starts by considering each data point as a cluster, and follows by merging two neighboring clusters at each step of the process ([13], [14]). The neighboring clusters are chosen based on a *linkage* criterion. The linkage criterion determines the distance between two clusters, as a function of pairwise distances between their corresponding data points. Common criteria include single-linkage, complete-linkage and average-linkage; they determine the neighboring clusters according to the minimum, maximum or average pairwise distances respectively (for more details see [13] and [14]).

To be consistent with our objective function given by equation (4.6), we adopt the *minimum variance* criterion that tries to minimize the total sum of square errors (total intra-cluster variance) at each step of the process. This is done by combining two clusters, whose combination results in the smallest increase of the total variance. It can be shown that in the case of Euclidian (or generalized Euclidian) metric, the increase of the total variance due to merging the $m$-th and $j$-th clusters ($\Delta_{mj}$) depends only on the centroid of the merged clusters and their cardinalities. For the weighted Euclidian distance given by Equation (4.5), we will have ([14]):

$$\begin{aligned} \Delta_{mj} &= SSE_{mj} - (SSE_m + SSE_j) \\ &= \frac{(\sum_{n=1}^{N} u_{mn}).(\sum_{n=1}^{N} u_{jn})}{(\sum_{n=1}^{N} u_{mn}) + (\sum_{n=1}^{N} u_{jn})} d^2_{E(\underline{w})}(\underline{r}_m, \underline{r}_j) \end{aligned} \tag{4.7}$$

where $SSE_m$, $SSE_j$ and $SSE_{mj}$ are the intra-cluster variance (sum of squared errors) for

the $m$-th, $j$-th and the resulting merged clusters, defined as follows:

$$SSE_i = \sum_{n=1}^{N} u_{in} d^2_{E(\underline{w})}(\underline{r}^{\circ}_n, \underline{r}_i),$$

and $\underline{r}_m$ and $\underline{r}_j$ represent the centroids of the $m$-th and $j$-th clusters. Therefore at each step of the clustering process, one merges two clusters that minimize the above criterion.

## 4.4 Complexity analysis

In this section we present a complexity analysis for the proposed clustering technique, in the context of a mobile-based fingerprinting system. The complexity analysis investigates two aspects: the induced transmission load on the network to transmit the database to the terminal, and the on-board computation load on the terminal during the localization phase. The performed complexity analysis examines the clustering method, the PCA, and the KCCA method. The reference case of a non-processed database is analyzed as well.

In the rest of the text, we consider the following notations:

• $N_{pca}$ and $N_{cv}$ stand for the number of adopted principle components and canonical vectors in the PCA and KCCA methods, respectively,
• $b_G$ is the number of bits required to code a single geographical coordinate (according to [18] equal to 24),
• $b_R$ is the average number of bits required to code a single radio parameter (for RSS measurements in GSM system equal to 6),
• $b_{R,pca}$ and $b_{R,cv}$ denote the number of bits required to code a radio parameter projected on principal components and canonical vectors, respectively,
• $b_{Eig}$ is the number of bits used to code a single element in eigenvectors in the PCA method,
• $b_{Tot}$ is the total number of bits required to code the database.

### 4.4.1 Transmission load

In this section for each processing method, we analyze the induced transmission load on the network to transmit the database to the terminal. At the first step of the analysis, we assume that the initial database is processed on the network side; the processed database along with the necessary parameters, are then sent to the terminal. The required transmission load for the considered processing techniques will be as follows:

1. Clustering technique: here the clustered database is transferred to the terminal. Simply we have:
   $b_{Tot} = M(D_G b_G + D_R b_R)$.

2. PCA technique: here the projected database along with a matrix of eigenvectors are transferred to the terminal. The matrix of eigenvectors allows to project the mobile measurements during the localization phase. So we have:
   $b_{Tot} = N(D_G b_G + N_{pca} b_{R,pca}) + D_R N_{pca} b_{Eig}$. As principal components provide an orthonormal basis for the signal space, we may consider that the projected radio components require as many bit as original radio components, i.e. $b_{R,pca} = b_R$. Hance, we have: $b_{Tot} = N(D_G b_G + N_{pca} b_R) + D_R N_{pca} b_{Eig}$.

3. KCCA technique: again the projected database is transferred to the terminal. Moreover all the initial radio measurements are needed at the terminal, in order to enable kernel computation during the localization phase. So we have:
   $b_{Tot} = N(D_G b_G + N_{cv} b_{R,cv}) + N D_R b_R$

4. No processing: here the initial database is directly transferred to the terminal. The resulting transmission load is given by:
   $b_{Tot} = N(D_G b_G + D_R b_R)$.

Now in a second step of the analysis, we assume that the initial database is directly transmitted to the terminal, and then all the processing is performed on the terminal side. In this case the required transmission load for all techniques will be equal to that of a non-processed database, given in item 4 above. The optimum transmission load may be considered as the minimum between the two argued cases. Clearly, for clustering technique it is more efficient to send the clustered database; while for KCCA technique, transmission of initial database is more efficient. Taking into account that in general $N_{PCA} \ll D_R$, and assuming that $b_{Eig}$ is in the same order as $b_R$, we may deduce that in the case of PCA technique it is more efficient to send the processed database. Table 4.1 summarizes the performed analysis for the transmission load.

### 4.4.2 Computation load

In this section we analyze the on-board computation load during the localization phase, corresponding to the considered processing techniques. The standard KNN method with $K = 1$ is assumed to be used as the classifier in the localization phase.

1. Clustering technique: here KNN is applied on a database consisting of $M$ samples in a $D_R$-dimensional space, which requires a complexity of $O(D_R M)$ for distance

| Proc. meth. | Database optimum transmission load |
|---|---|
| Clustering technique | $b_{Tot} = M(D_G b_G + D_R b_R)$ |
| PCA | $b_{Tot} = N(D_G b_G + N_{pca} b_R) + D_R N_{pca} b_{Eig}$ |
| KCCA | $b_{Tot} = N(D_G b_G + D_R b_R)$ |
| No processing | $b_{Tot} = N(D_G b_G + D_R b_R)$ |

Table 4.1: Analysis of transmission load to transmit the radio database

evaluations, and a complexity of $O(M \log M)$ for sorting. As sorting involves only comparison operations, the corresponding complexity may be neglected against that of distance evaluation part; thus the total computational complexity may be approximated by $O(D_R M)$.

2. PCA technique: Here the computation complexity consists of two parts. The first part is the computation cost due to decomposing a new measurement into the principle components. This decomposition requires a matrix multiplication with a complexity of $O(N_{pca} D_R)$. The next step is to perform a KNN in the $N_{pca}$-dimensional space, resulting a complexity of $O(N_{pca} N)$. The total complexity is given by $O(N_{pca} D_R) + O(N_{pca} N)$.

3. KCCA technique: Again the computation complexity consists of two parts. The first part is due to the projection of test measurement onto the canonical vectors, where each projection requires $N$ kernel evaluations; this requires a complexity of $O(N_{cv} N D_R)$. The next part is to perform a KNN in the $N_{cv}$-dimensional space. So the total complexity is given by $O(D_R N_{cv} N) + O(N_{cv} N)$.

4. No processing: here KNN method is applied on the initial database consisting of $N$ samples in a $D_R$-dimensional space, which requires simply a complexity of $O(D_R N)$.

Table 4.2 summarizes the computational complexity of different techniques. Obviously, the clustering technique can reduce the computation cost with respect to the reference case of no-processing; the reduction is directly proportional to the compression index $\eta$. At the same compression index, considering that generally $N_{PCA} \ll N$ and $D_G \ll D_R$, we may deduce that the PCA method induces a computational complexity close to that of the clustering technique. On the other hand, KCCA processing demands a much higher complexity w.r.t. a non-processed database. As a conclusion, the clustering technique could be used to reduce the computation cost of localization phase in fingerprinting systems.

| Proc. meth. | Computation complexity |
|---|---|
| Clustering technique | $O(D_R M)$ |
| PCA | $O(N_{pca}N) + O(N_{pca}D_R)$ |
| KCCA | $O(N_{cv}N) + O(D_R N_{cv}N)$ |
| No processing | $O(D_R N)$ |

Table 4.2: Analysis of computation complexity for localization phase

## 4.5 Positioning performance evaluation

In this section, we examine the performance of the proposed clustering algorithms, in the context of a cellular fingerprinting system. We saw that the clustering techniques are effective enough for reduction of computation and transmission loads. Here, we examine how they impact the positioning accuracy in fingerprinting systems. The positioning performance criterion here is the positioning error (in meters), which may be evaluated by its average value and standard deviation. All the performed evaluations are based on computer simulations.

### 4.5.1 Simulations setup

#### Radio propagation model

Similarly to computer simulations in the previous chapter, we assume that the localization service is offered over a geographical area $\mathcal{A}$, which is covered by a GSM cellular network. The GSM cells are again considered to be omnidirectional hexagonal with a radius of 1 km. The area $\mathcal{A}$ covers a surface of $B = 13$ cells (one reference central cell and two rings of neighboring cells). To model the RSS measurements at an arbitrary location in the area, we adopt the Mondrian model introduced in the previous chapter (see section 3.2.1).

For any transmitter-receiver link, the path loss $PL_a$ is determined according to Equation (3.2). The average received signal power is then given by $P_T - PL_a$, where $P_T$ represents the transmitted power. However the instantaneous measurements performed by a mobile terminal are not equal to this average value. The RSS temporal variations are traditionally modeled by a log-normal distribution ([19]). Accordingly, here we use a log-normal random variable ($X_{Noise}$) to model these variations. Thus, the measured RSS may be modeled as

follows:

$$s = P_T - PL_a + X_{Noise}, \tag{4.8}$$

where $X_{Noise} \sim \mathcal{N}(0, \sigma^2_{Noise})$ is a gaussian random variable (in dB) with standard deviation $\sigma_{Noise}$.

We note that the shadowing effect is an important factor which influences the radio signal behavior in any environment. Thus the performance of our proposed clustering algorithms may be affected by the configurations of the shadowing in the area $\mathcal{A}$. In order to analyze the shadowing effect more elaborately, two different environments have been simulated. The following table gives the adopted parameters for each environment. In order to model an urban environment, both areas are generated with an equivalent $\alpha$ of 3.8.

| Parameter | Environment1 | Environment2 |
|---|---|---|
| Masks density (per km$^2$) | 32.5 | 325 |
| Masks length (m) | $U(250-750)^*$ | $U$(50-150) |
| Masks att. coef.(dB) | $\mathcal{N}(10, 3^2)^*$ | $\mathcal{N}(10, 3^2)$ |
| Equivalent $\alpha$ | 3.8 | 3.8 |

* The symbols $U$ and $\mathcal{N}$ represent the uniform and Gaussian distributions, respectively.

Table 4.3: Masks configuration for the simulated environments

**Fingerprinting system configurations**

The simulated experiments are based on the system architecture given in Figure 4.1. The test area is limited to the surface of the central cell, in order to eliminate the border effects. Each geographical location in the area may be described by a 2-dimensional vector ($D_G = 2$). All the RSS measurements are simulated according to Equation (4.8), with $\sigma_{Noise} = 0$ dB (noiseless scenario). Any measurement is assumed to contain the signal components concerning all the $B$ base stations in the area (i.e. $D_R = B = 13$). Therefore, for the concatenated location-radio space we have $D = 15$.

In practical implementations, RSS measurements of actual terminals do not contain the signal components concerning all the $B$ base stations of the area. As mentioned in the previous chapter, the maximum number of scanned base stations in a measurement vector in actual terminals is restricted by an upper bound $B_{max}$ (in GSM standard $B_{max} = 7$).

Moreover, the terminal receiver can detect only the RSS values that are higher than a predefined threshold $\lambda$ (in GSM standard $\lambda$ = -110 dBm). The undetected components may be considered as *missing data*.

In order to include the missing data in the simulations, similarly to chapter 3, we compute the components concerning the seven strongest base stations according to our radio model, and we consider all the other components as missing data with unknown values. The optimal handling of missing data in RSS measurements will be treated later in chapter 6. Here, as a simple approach, we fill in all the missing RSS values with the minimum detectable signal level $\lambda$.

Concerning the training phase, an initial database $\mathcal{R}^\circ = \{(\underline{x}_n^\circ, \underline{s}_n^\circ)\}_{n=1\ldots N}$ is constructed by simulating a number of $N = 1225$ RSS measurements. The measurements are simulated according to a regular pattern over the central cell. By applying a clustering technique, the initial database is divided into a certain number of clusters $M$ ($M < N$). The cluster centroids are then used to construct the compressed radio database $\mathcal{R} = \{(\underline{x}_m, \underline{s}_m)\}_{m=1\ldots M}$.

During the localization phase, the mobile terminal performs a sample RSS measurement $\underline{s}'$ at location $\underline{x}'$. In order to localize the mobile terminal, the basic K-Nearest-Neighbors method has been adopted, with $K = 1$. We note that in this case KNN serves as a classifier that assigns each RSS measurement to the closest cluster center. The estimated position of the mobile $\widehat{\underline{x}}$ is given by:

$$\widehat{\underline{x}} = \underline{x}_{\widehat{m}}, \qquad \widehat{m} = \text{argmin}_m \, \|\underline{s}' - \underline{s}_m\|. \tag{4.9}$$

The corresponding positioning error in meters is then defined as:

$$\varepsilon(\underline{x}') = \|\underline{x}' - \widehat{\underline{x}}\|.$$

A number of 1000 test measurements have been simulated concerning the localization phase. These measurements are uniformly distributed over the central cell.

### 4.5.2 Parameters setting

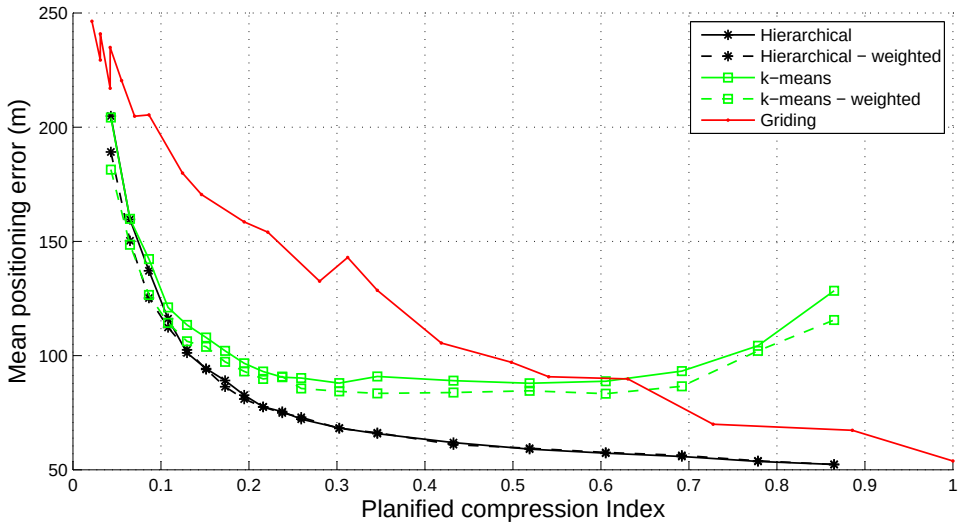
The only parameters to be set in the simulations are the weight factors introduced in the generalized Euclidian distance in Equation (4.5). We recall that here, the weight factors are considered as input variables that must be fixed before running the clustering algorithm. In this work, we consider equal weight factors for all the components belonging to the same feature type; in other words we have:

$$w_1 = w_2 = w_G,$$

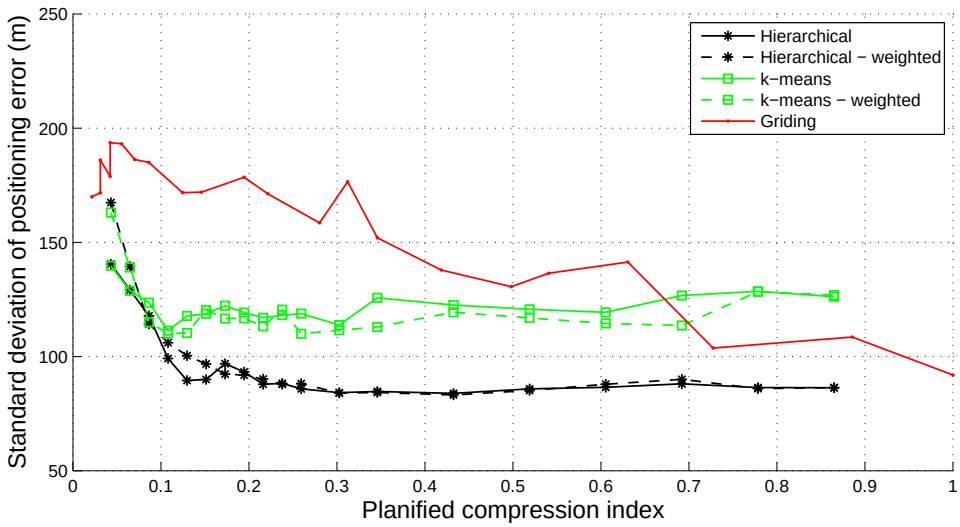
and

$$\begin{aligned} w_3 &= \ldots = w_{15} = w_R \\ &= 1 - \omega_G. \end{aligned}$$

Now, we make a heuristic assumption to fix the weight factors. We propose to set $w_G/w_R = D_R/D_G$. There is no mathematical justification behind this assumption; however we will see in the next chapter that this chosen ratio is close to the optimized ratio, provided by a weighted clustering algorithm.

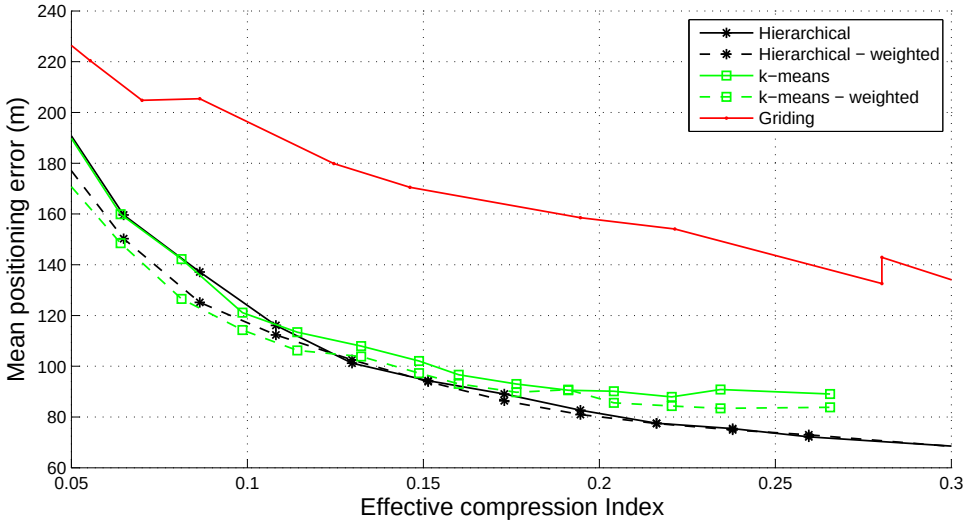



(a) Mean positioning error with respect to the planified compression index

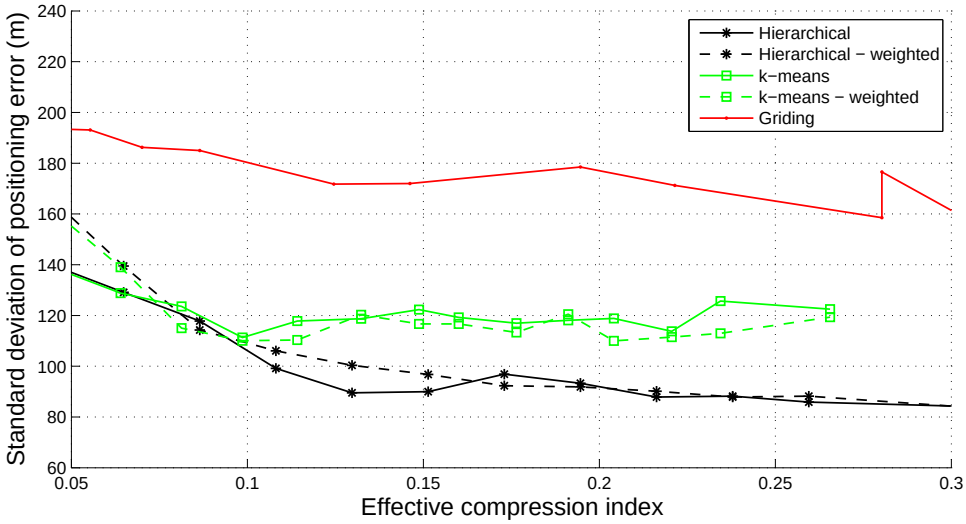



(b) Standard deviation of positioning error with respect to the planified compression index

Figure 4.2: Performance of clustering techniques versus planified compression index, for environment 1




(a) Mean positioning error with respect to the effective compression index




(b) Standard deviation of positioning error with respect to the effective compression index

Figure 4.3: Performance of clustering techniques versus effective compression index, for environment 1

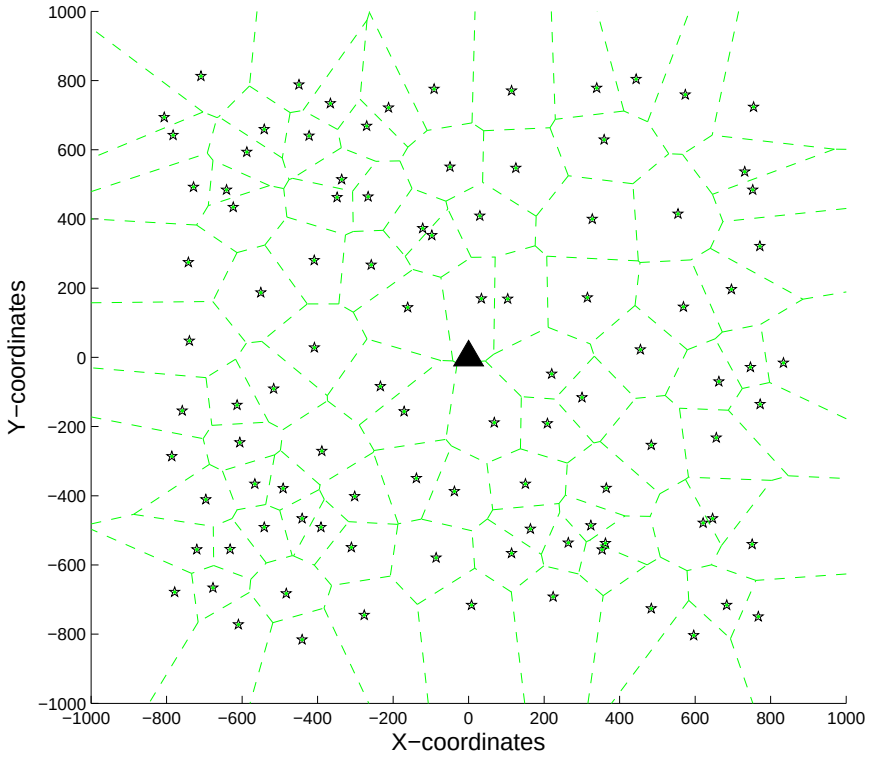
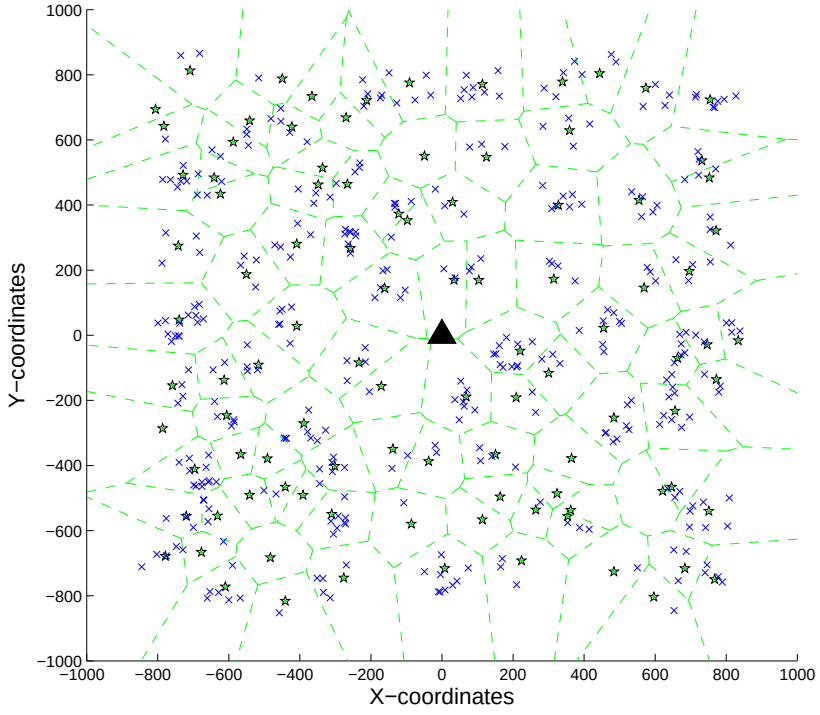



Figure 4.4: Distribution of clusters over the central cell, for $M = 100$. The star signs represent the clusters geographical centers.

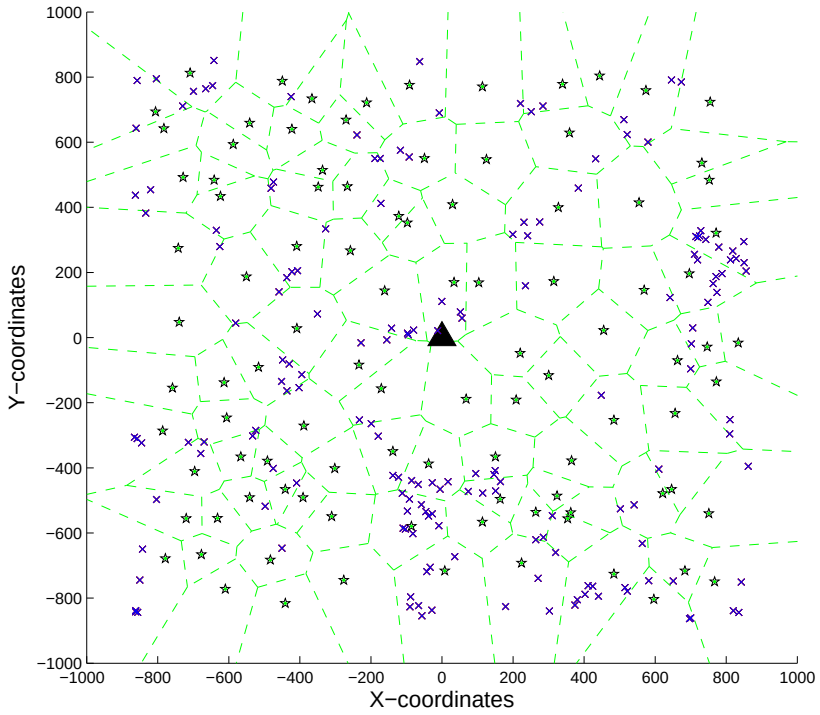
### 4.5.3 Simulations results

We have simulated a location fingerprinting system by adopting all the configurations explained in the previous sections. The performance of the positioning system has been evaluated for the proposed clustering techniques and the conventional griding method. The performance indexes consist of the average positioning error and the corresponding standard deviation. Each clustering algorithm has been implemented by using both types of normal and weighted Euclidian distances. The k-means algorithm is everywhere initialized by the results of a primary hierarchical clustering.

We first discuss the obtained results for environment 1, which are illustrated in Figures 4.2, 4.3, 4.4 and 4.5. Figures 4.2 and 4.3 demonstrate the average positioning error and the corresponding standard deviation versus the compression index. As we mentioned before in section 4.2, the k-means algorithm may provide a large number of empty clusters. There-




(a) Mobile positions resulting in a low positioning error (error <76 m)




(b) Mobile positions resulting in a high positioning error (error > 176 m)

Figure 4.5: Distribution of high and low positioning errors over clusters for M=100 (average positioning error = 126 m)

fore, the results are presented with respect to both "planified" and "effective" compression indexes; the planified compression index ($\eta$) is computed by adopting the predefined number of clusters $M$, while the effective compression index ($\eta_{Eff}$) is computed by taking the number of non-empty clusters ($\eta_{Eff} \leq \eta$).

Figure 4.2 illustrates the results as a function of the planified compression index. We observe that in general, the clustering algorithms provide a more accurate positioning w.r.t the conventional griding method. The hierarchical and the k-means methods provide close performance for small compression indexes ($\eta < 0.1$); but for large values of $\eta$ the hierarchical method outperforms the k-means algorithm, in both average error and the standard deviation. The reason stems from generation of plenty of empty clusters in the k-means method, for large values of $\eta$. We can observe that the hierarchical algorithms provide an almost stabilized performance for $0.3 < \eta$ (with an average positioning error of about 50 - 60 m) . We may deduce that in the simulated scenario, a compressed database with $\eta \simeq 0.3$ provides a positioning performance close to that of a non-compressed database. We note that the error standard deviation for all techniques is in the same order as the error average value. Therefore, we may expect large variations of positioning error during the localization phase. As a last point, the weighted and the non-weighted Euclidian distances here do not lead to considerable differences in positioning performance.

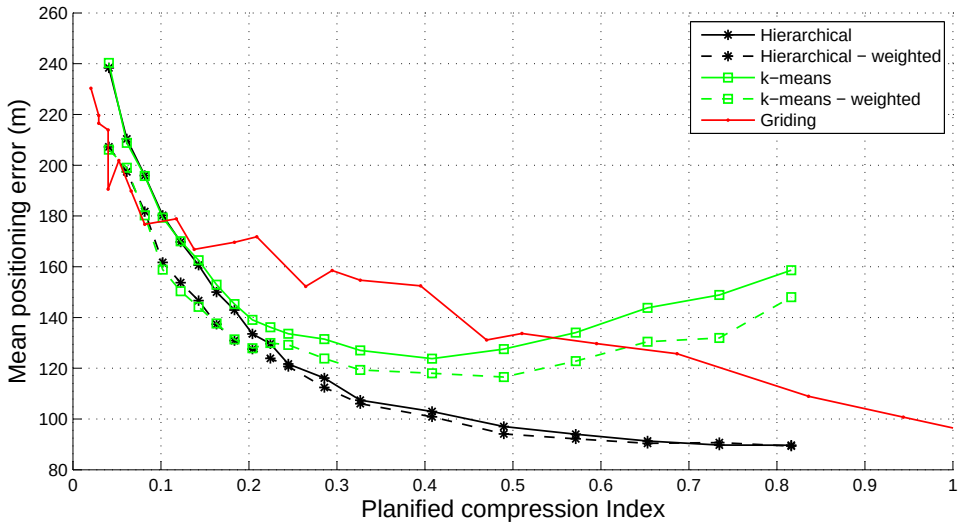
To provide a more accurate comparison, in Figure 4.3 the performance of clustering methods is illustrated versus the effective compression index $\eta_{Eff}$. As we see, all the clustering techniques outperform significantly the simple griding method. We observe that the k-means algorithm provides an average positioning error competitive to that of the hierarchical method. The main difference is the higher standard deviation of error for k-means clustering. According to the figure, using the weighted Euclidian distance is slightly more effective than the non-weighted distance.

Finally, in order to provide a visual presentation of the clustering process, the distribution of clusters over the central cell for $M = 100$ is illustrated in Figure 4.4. The figure depicts the perimeters and the geographical centers of the clusters provided by the k-means algorithm and the weighted Euclidian distance. Afterwards, Figure 4.5 illustrates the distribution of high and low positioning errors over the central cell, and over the clusters. We can observe that, low errors in general arise when the mobile terminal is in the central part of the clusters. The large errors happen generally when the mobile is in the outer part of clusters, close to the perimeters.

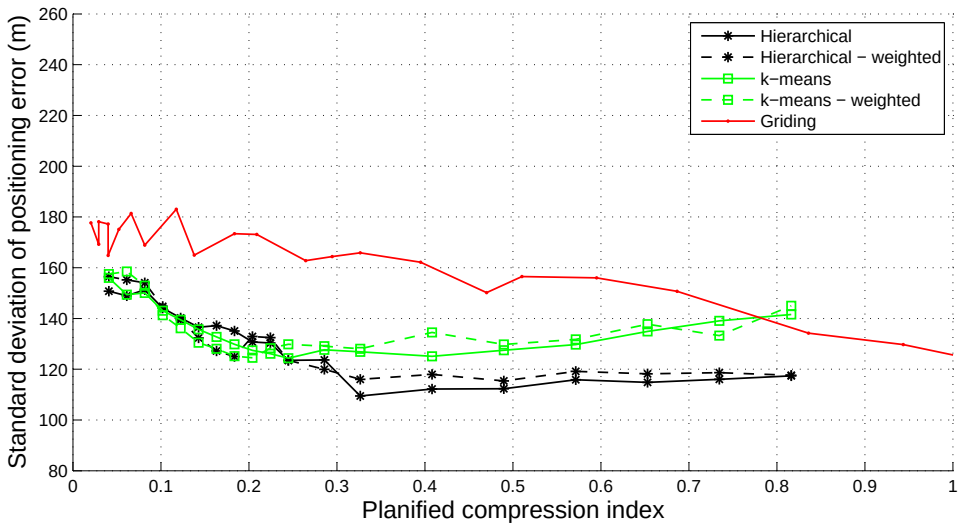
Now, we consider the results of the simulations for environment 2, which are illustrated in Figures 4.6 and 4.7. As a general remark, we note that the performance of all the techniques is degraded with respect to environment 1. In fact, the large shadowing masks in environment 1 induce considerable changes in the measured RSS values, and this fact

makes the clustering techniques effective enough in this context. The small shadowing masks in environment 2 do not induce as much signal variations as the large masks in the previous environment. Thus here the RSS patterns are more homogenous over the simulated area, and this fact makes the clustering and the positioning tasks more difficult.

Figure 4.6, illustrates the performance as a function of planified compression index. In general, one may observe the same trends as those of environment 1. In general, the clustering algorithms provide a more accurate positioning w.r.t the conventional griding method (except for $\eta < 0.1$). Again, the hierarchical method outperforms the k-means algorithm, in both average error and the standard deviation. One may observe that the hierarchical algorithms provide an almost stabilized performance for $0.3 < \eta$ (with an average positioning error of about 90 - 100 m). As a final remark, we note that at the presence of RSS homogeneity, the weighted Euclidian distance provides a lower average error w.r.t the non-weighted distance, by attributing more importance to the location parts.

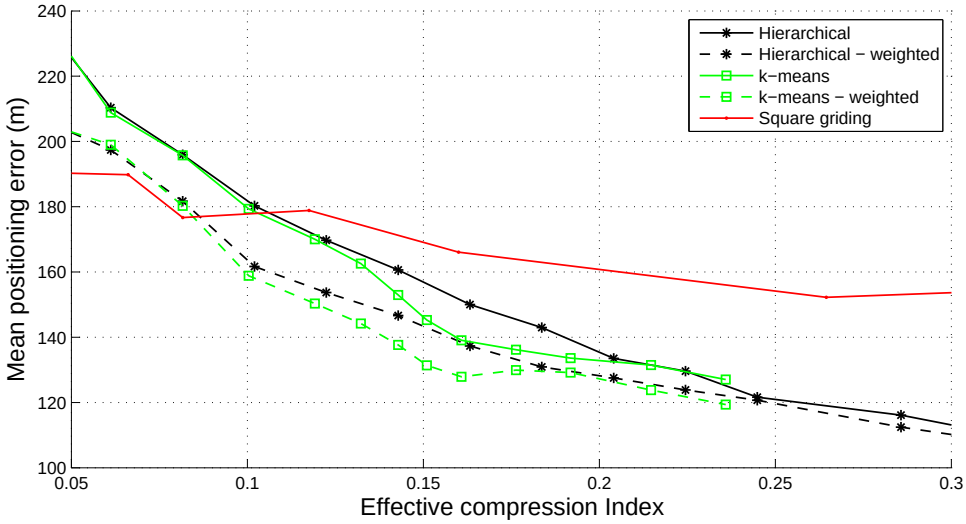



(a) Mean positioning error with respect to the planified compression index

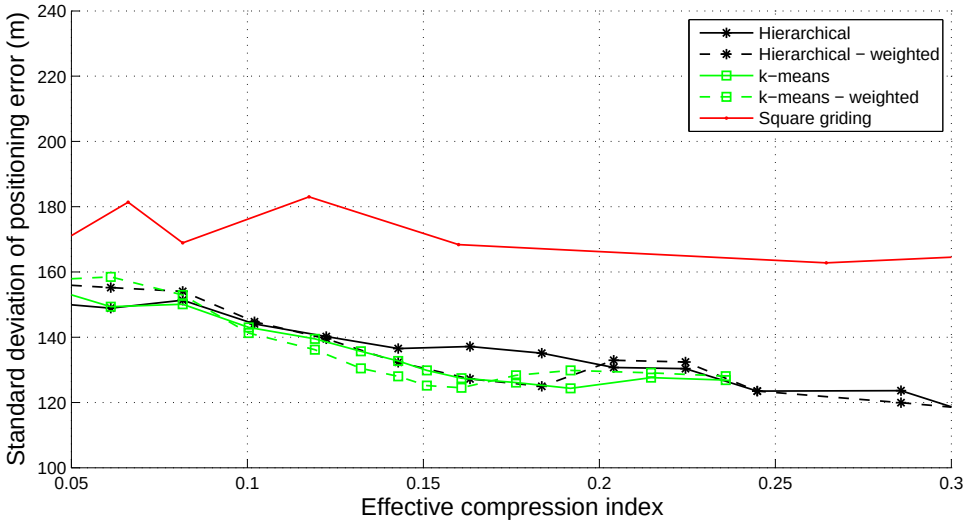



(b) Standard deviation of positioning error with respect to the effective compression index

Figure 4.6: Performance of clustering techniques versus planified compression index, for environment 2




(a) Mean positioning error with respect to the planified compression index




(b) Standard deviation of positioning error with respect to the effective compression index

Figure 4.7: Performance of clustering techniques versus effective compression index, for environment 2

Figure 4.6 illustrates the positioning performance as a function of the effective compression index. Similarly to environment 1, the comparison of the methods versus the effective compression index shows that the k-means algorithm provides a competitive performance w.r.t the hierarchical method. The visual presentation of clusters and the distribution of positioning error for environment 2 is neglected here, since it shows the same trends as those of environment 1.

## 4.6 Conclusion

In this chapter, we tackled the problem of radio database compression by reducing the number of records, in the context of location fingerprinting systems. We proposed to perform the compression by applying the standard clustering techniques, including the k-means and the minimum-variance hierarchical clustering. The algorithms are applied in a concatenated "location-radio" signal space. The adopted algorithms take into account both location and radio parts of the stored records, and hence are expected to provide a more accurate positioning than that of naive griding method (which proceeds only based on records location parts).

The performance of the standard clustering algorithms based on the Euclidian distance, was examined by computer simulations. Two different environments were examined: a first area with large shadowing masks (average masks length about 500 m), and a second area with small shadowing masks (average masks length about 100 m). The average positioning error obtained by clustering algorithms was stabilized at about 50-60 m for the first area, and about 90-100 m for the second area. The cluster analysis was more effective in the area with large masks than it was in the area with small masks, since the radio propagation was more homogeneous in the latter case. We observed a superior positioning performance of clustering techniques with respect to the naive griding method in both simulated scenarios.

The cluster analysis in this work was proposed with the goal of decreasing the terminal power consumption in mobile-based LFP systems, by reducing the computation and transmission loads. A complexity analysis was performed to confirm this claim by evaluating the computation and transmission loads issued from clustering techniques. The presented analysis included also a comparison with other compression methods proposed in the literature, such as Principle Component Analysis (PCA) and Kernel Canonical Correlation Analysis (KCCA). Based on the performed analysis, the clustering techniques outperform the other compression methods in the complexity viewpoint.

Finally, it was shown that using the weighted Euclidian distance is advantageous with

respect to the non-weighted distance, since the former works well even under homogenous propagation conditions. However, the choice of the weight factors in this chapter is done quite heuristically. One may envisage a systematic method to choose the weight factors in the clustering process. This subject will be treated in the next chapter, where a weighting scheme will be proposed for the clustering algorithm. A comparative framework will be also presented, which evaluates the clustering techniques versus other compression techniques, like PCA and KCCA.

## Bibliography

[1] C. Dessiniotis, "Motive project deliverable 2.1," European Project FP6-IST 27659, Tech. Rep., September 2006.

[2] A.K.M. Mahtab Hossain, H. Nguyen Van, Y. Jin, W.S. Soh, "Indoor localization using multiple wireless technologies," in *Proceedings of the IEEE Internatonal Conference on Mobile Adhoc and Sensor Systems*, 2007, pp. 1 – 8.

[3] M. Anisetti, V. Bellandi, E. Damiani, S. Reale, "Advanced localization of mobile terminal," in *Proceedings of the International Symposium on In Communications and Information Technologies*, 2007, pp. 1071–1076.

[4] Widyawan, M. Klepal, D. Pesch, "Influence of predicted and measured fingerprint on the accuracy of RSSI-based indoor location systems," in *Proceedings of the 4th workshop on positioning, navigation and communication*, 2007, pp. 145–151.

[5] S. Fang, T. Lin, P. Lin, "Location fingerprinting in a decorrelated space," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 20, no. 5, pp. 685 – 691, May 2008.

[6] Y. Chen, J. Yin, X. Chai, and Q. Yang, "Power-efficient access-point selection for indoor location estimation," in *IEEE Transactions on Knowledge and Data Engineering*, July 2006, pp. 877–888.

[7] M. A. Youssef, A. Agrawala, A. Udaya Shankar, "WLAN location determination via clustering and probability distributions," in *Proceedings of the Conference on Pervasive Computing and Communications*, March 2003, pp. 143– 150.

[8] A. Arya, P. Godlewski, P. Mellé, "Performance analysis of outdoor localization systems based on RSS fingerprinting," in *Proceedings of the International Conference on Wireless Communication Systems*, September 2009, pp. 378 – 382.

[9] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, 2009.

[10] E. Backer and A. Jain, "A clustering performance measure based on fuzzy set decomposition," *IEEE Transactions on Pattern Analysis and Machine Intellegence*, vol. 3, no. 1, p. 66–75, January 1981.

[11] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645–678, May 2005.

[12] P.N. Tan, M. Steinbach, V. Kumar, *Introduction to Data Mining.* Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, 2005.

[13] A.C. Rencher, *Methods of multivariate analysis.* Wiley, 2002.

[14] L. Lebart, A. Morineau, K.M. Berry, *Multivariate descriptive statistical analysis.* Wiley, 1984.

[15] J . C. Bezdek, "A convergence theorem for the fuzzy ISODATA clustering algorithms," pp. 1–8, January 1980.

[16] H.H. Bock, "Origins and extensions of the k-means algorithm in cluster analysis," in *Electronic journal for history of probability and statistics*, vol. 4, no. 2, December 2008.

[17] A. Likas, N. A. Vlassis and Jakob J. Verbeek, "The global k-means clustering algorithm," *The journal of the Pattern Recognition Society*, 2003.

[18] "3GPP TS 23.032," 3rd Generation Partership Project, Technical Specification Group Services and System Aspects, Universal Geographical Area Description (GAD) (Release 9), Tech. Rep., 2009.

[19] A Kushki, KN Plataniotis, AN Venetsanopoulos, CS Regazzoni, "Radio map fusion for indoor positioning in wireless local area networks," in *Proceedings of the International Conference on Information Fusion*, vol. 2, July 2005, pp. 1311–1318.

# Chapter 5

# Block-based Weighted Clustering (BWC) scheme for radio database clustering

In the previous chapter, we presented clustering techniques for radio database compression, which take into account both the location and the radio components of the recorded measurements. The algorithms were applied in a concatenated "location-RSS" space; a generalized form of Euclidian distance with heuristic fixed weights was adopted to attribute different weight factors to the location and radio parts in the concatenated vectors. The main idea of this chapter is to consider the weight factors as variables that would be optimized during the clustering process. We propose a method which may be called *Block-based Weighted Clustering* (BWC) technique: a weighted clustering algorithm is applied in a concatenated location-radio signal space, and the weight factors are optimized during the clustering process. This might be considered as providing a refined distance metric, adapted to the specific structure of records in the database.

## 5.1 Weighted variants of k-means algorithm

The standard variant of k-means clustering has been presented in chapter 4. The k-means algorithm tries to minimize the objective function given by Equation (4.2), where the Euclidian metric is adopted as the dissimilarity measure.

Assume a set of $N$ data points $\mathcal{R}^\circ = \{\underline{r}_n^\circ\}_{n=1...N}$ in a $D$-dimensional feature space ($\underline{r}_n^\circ \in \mathbb{R}^D$). The classic k-means clustering for the set $\mathcal{R}^\circ$ tries to minimize the objective function given by Equation (4.2). Coming back to this equation, we observe that all the input components contribute to the total intra-cluster variance with equal weights. Some

extensions of Equation (4.2) have been developed that associate different importance to different components, such as the one proposed in [1] as follows:

$$J_3(U, \mathcal{R}, \underline{w}) = \sum_{n=1}^{N} \sum_{m=1}^{M} \sum_{h=1}^{D} u_{mn} w_h^{\beta} d^2(r^{\circ}_{n,h}, r_{m,h}). \tag{5.1}$$

where $N$ and $M$ denote the number of the data points and the clusters respectively, $\beta$ is a fixed constant parameter ($\beta > 1$), $d(r^{\circ}_{nh}, r_{mh})$ is the distance between the $h$-th component of the $n$-th data point and the $m$-th centroid, and $\underline{w} = [w_1, ..., w_D] \in \mathbb{R}^D$ is the vector including the weight factors, subject to the constraint $\sum_{h=1}^{D} w_h = 1$. An extension of k-means algorithm is afterwards developed in [1] to provide the optimized values of $U, \mathcal{R}$, and $\underline{\omega}$.

We note that the defined weight factors in $J_3$ are the same for all the provided clusters. A more flexible weighting scheme may be deployed by defining different weight factors for different clusters, as in [2]. Here, the weights may be represented by a $M \times D$ matrix $W = [w_{mh}]$, subject to the constraint $\sum_{h=1}^{D} w_{mh} = 1$ for $1 \leq m \leq M$. The resulting objective function may be described as follows:

$$J_4(U, \mathcal{R}, W) = \sum_{n=1}^{N} \sum_{m=1}^{M} \sum_{h=1}^{D} u_{mn} w_{mh}^{\beta} d^2(r^{\circ}_{n,h}, r_{m,h}), \tag{5.2}$$

It is noteworthy that the above weighting scheme involves a large number of variables to be optimized by the algorithm. The following table summarizes the number of unknown variables involved in each of the objective functions $J_3$ and $J_4$; the simple objective function $J_1$ is also considered as a reference case.

| Objective function | Number of variables |
|---|---|
| $J_1$ | $N + MD$ |
| $J_3$ | $N + (M + 1)D$ |
| $J_4$ | $N + 2MD$ |

Table 5.1: Number of variables to be optimized for different objective functions

Concerning all the three objective functions, there are $N$ cluster membership variables and $MD$ centroid coordinates to be found. The functions $J_3$ and $J_4$ involve $D$ and $MD$ additional weighting variables, respectively. Thus, we see that on the whole, $J_4$ needs to optimize much more variables w.r.t. other objective functions.

## 5.2 Block-based Weighted Clustering (BWC) scheme

Assume an initial radio database $\mathcal{R}^\circ = \{\underline{r}_n^\circ\}_{n=1...N} \subset \mathbb{R}^D$, constructed according to raw field measurements. Application of standard k-means algorithm for the initial database compression was examined in the previous chapter. Here, we are interested in weighted variants of k-means clustering, with an objective function like the one given by Equation (5.1); but instead of this generic form, we propose a Block-based Weighted Clustering scheme (BWC) for the objective function. In this weighting scheme, we impose equal weight factors to all the components that belong to the same feature type. Thus, the weight vector $\underline{w}$ consists of blocks of equal values. We recall from the previous chapter that any record $\underline{r}_n^\circ$ may be considered as follows:

$$\underline{r}_n^\circ = (\underline{\rho}_{n,1}^\circ, ..., \underline{\rho}_{n,h}^\circ, ..., \underline{\rho}_{n,N_f}^\circ),$$

, where $\underline{\rho}_{nh}^\circ$ is the vector denoting the $h$-th feature type, and $N_f$ is the number of feature types in each record. By adopting the Euclidian distance and the proposed weighting scheme, the objective function of BWC technique can be represented as follows:

$$J_5(U, \mathcal{R}, \underline{\omega}) = \sum_{n=1}^{N} \sum_{m=1}^{M} \sum_{h=1}^{N_f} u_{mn} \omega_h^\beta \|\underline{\rho}_{n,h}^\circ - \underline{\rho}_{m,h}\|^2. \tag{5.3}$$

where $\underline{\omega} = [\omega_1, ..., \omega_{N_f}] \in \mathbb{R}^{N_f}$ is the vector including the weight factors, subject to the constraint $\sum_{h=1}^{N_f} w_h = 1$.

It is noteworthy that it is possible to choose a distance metric other than the Euclidian metric. In this work we adopt simply the Euclidian distance; choosing a more appropriate distance metric might be the subject of further studies.

The objective function $J_5$ has the same structure as that of Equation (5.1); the only difference appears in the distance evaluation term, where vectors replaced scalars. Based on the presented algorithm in [1] to optimize the generic objective function of Equation (5.1), it is straightforward to show that the update equations for the BWC scheme will be as follows:

- updated memberships:

$$u_{mn} = \begin{cases} 1 & \text{if } \forall m \neq t, \quad d_\omega(\underline{r}_n^\circ, \underline{r}_m) \leq d_\omega(\underline{r}_n^\circ, \underline{r}_t) \\ 0 & \text{otherwise} \end{cases}$$

  where $d_\omega(.,.)$ is the weighted distance defined as:

$$d_\omega^2(\underline{r}_n^\circ, \underline{r}_m) = \sum_{h=1}^{N_f} \omega_h^\beta \|\underline{\rho}_{n,h}^\circ - \underline{\rho}_{m,h}\|^2$$

- updated centers:

$$\underline{r}_m = \frac{\sum_{n=1}^{N} u_{mn}\underline{r}_n^{\circ}}{\sum_{n=1}^{N} u_{mn}}$$

- updated weights:

$$w_h = \begin{cases} 0 & \text{if } D_h = 0 \\ \frac{1}{\sum_{t=1}^{N_f}[\frac{D_h}{D_t}]^{\frac{1}{\beta-1}}} & \text{otherwise} \end{cases}$$

where

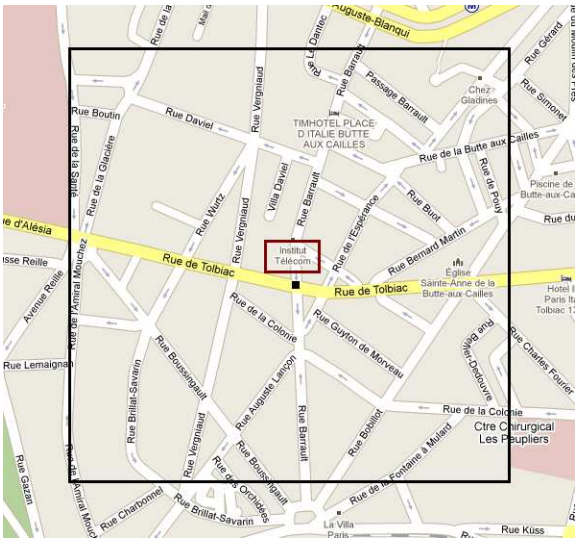
$$D_h = \sum_{n=1}^{N}\sum_{m=1}^{M} u_{mn}\|\underline{\rho}_{n,h}^{\circ} - \underline{\rho}_{m,h}\|.$$

We note that the generic objective function given by Equation (5.1) is not really appropriate to be used for a database of RSS measurements. In a cellular network the importance of an RSS measurement generally depends on the corresponding transmitter-receiver distance. Hence the importance of a single base station can vary from one location to another. The objective function in Equation (5.1) imposes a single set of weights to the base stations over all the considered area, which is in contrast to location-dependency of weight factors. On the other hand, a matrix-structured weighting scheme (as in Equation (5.2)) may be consistent with the properties of RSS measurements, since it attributes different sets of weights to different clusters. But in this method there exist a large number of weights to be learnt (see table 5.1). One of our goals in this work is to demonstrate whether such a heavy weighting is advantageous w.r.t. the proposed feature type-weighted scheme.

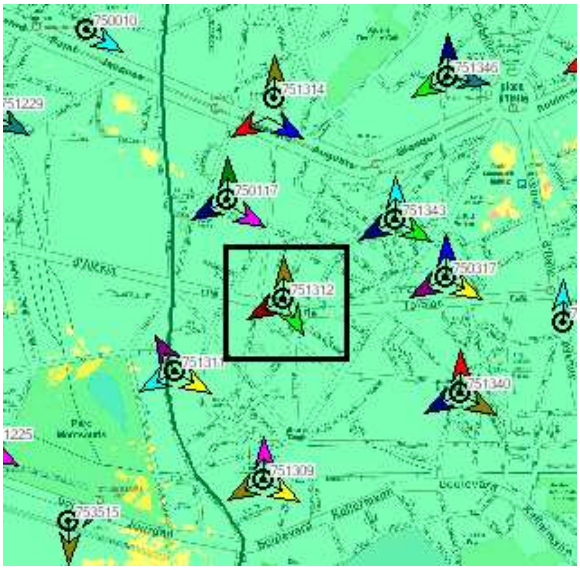
## 5.3 Positioning performance evaluation

In this section, the performance of the proposed BWC scheme is evaluated by simulated and real experiments, in the context of a cellular fingerprinting system. At a first step, the performance is evaluated versus the compression index for different clustering algorithms; the considered clustering algorithms consist of the proposed BWC technique, the matrix-based weighted k-means, the standard k-means, and the standard hierarchical clustering with a minimum-variance linkage criterion. The performance criteria include the average positioning error and the corresponding standard deviation.

Returning to the state-of-the-art architectures presented in Figure 2.5, we recall that processing methods such as PCA and KCCA could also be applied to compress a radio database. In these methods, compression is done by reducing the dimension of records of the initial database. Hence, as a second step, we provide a comparative evaluation of these techniques w.r.t. the proposed BWC method.




(a) The adopted test zone in Paris




(b) The base stations placement (the test zone limited by the black square)

Figure 5.1: The selected test area for the real experiments

### 5.3.1 Experiments setup

Concerning the simulated experiments, we take up the simulation setup introduced in the previous chapter, in section 4.5.1. Previously, we examined two different areas (with large, and small shadowing masks), to evaluate the clustering techniques. Here we adopt the second area (with small masks) to run the experiments, since it was shown to be the worse case. The fingerprinting system configurations are adopted exactly as in section 4.5.1.

Concerning the real experiments, a limited urban area around the Télécom ParisTech building, in the southern part of Paris has been considered. The RSS measurements have been performed over this area, in cooperation with the French mobile operator SFR. The test zone is illustrated in Figure 5.1-a, as the surface limited by the black square. The considered surface is about 500 m $\times$ 500 m, and covers approximately a single GSM cell. This area has been chosen to conduct the tests, since the corresponding placement of GSM antennas is consistent with that of simulations. The location of GSM antennas around the test zone is depicted in Figure 5.1-b. We can observe clearly a ring of neighbor cells around the considered central cell.

The RSS measurements are performed in outdoor, over all the streets present in the test zone with a resolution of 10 m. A total number of 750 test points have been picked out. At each test point 60 successive measurements are performed over 60 intervals of one second. The average of all the 60 measurements is considered as a single measurement at a single test point. We attributed 450 out of 750 test measurements to database construction ($N = 450$), and the remaining 300 measurements were used in the localization phase to evaluate the positioning error.

A total number of $B = 25$ different GSM antennas are detected over the test zone. This large value of $B$ is a result of using three-sector antennas over the area. Considering that in actual terminals any single measurement contains only 7 signal components, there are plenty of missing RSS values in the initial database, which are all replaced by $\lambda = -110$ dB (similarly to simulated experiments).

In both simulated and real experiment, the records in the database consist of two feature types: location and RSS. Thus we have $N_f = 2$; concerning the BWC scheme we may write: $\omega = [\omega_G, \omega_R]$, where $\omega_G$ and $\omega_R$ are the weights attributed to location and RSS feature types, respectively.

### 5.3.2 Parameters setting

There are several parameters to be tuned at this step. Concerning the clustering techniques one must set:

- $\beta$ for the matrix-based weighted k-means,
- $\beta$ for the proposed BWC technique.

In order to set the above parameters, we use the cross-validation method. One round of cross-validation method involves splitting the initial training data set into two subsets. The first subset (here a fraction of $\frac{4}{5}$ of the training data) is used to build the model, and the remaining data (here a fraction of $\frac{1}{5}$) is used as a validation set. For each parameter we consider a variation interval, and empirically pick up the value that performs well in the validation set. Multiple rounds of the procedure are performed by permuting the validation set over the five possibilities, and then the validation results are averaged over the five rounds. Once parameters are determined, we recombine the two parts of data to build the model for the localization phase. The configured values for the parameters are turned out to be similar for the simulated and the real scenarios. We obtained $\beta = 20$ for the matrix-based weighted k-means, and $\beta = 10$ for the BWC technique.

There are two parameters concerning the KCCA method to be fixed, before running the experiments. These parameters include:

- the standard deviation of the Gaussian kernel $\sigma_G$,
- the regularization factor $\kappa$.

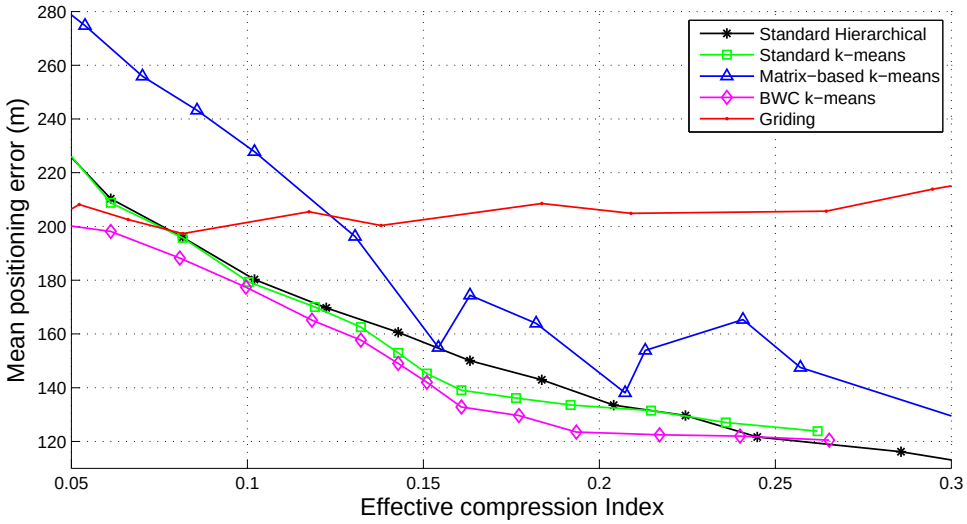
Again, by using a cross validation method, we obtained: $\sigma_G = 40$ for the Gaussian kernel, and $\kappa = 1.5$ for the regularization factor.

### 5.3.3 Evaluation of clustering techniques

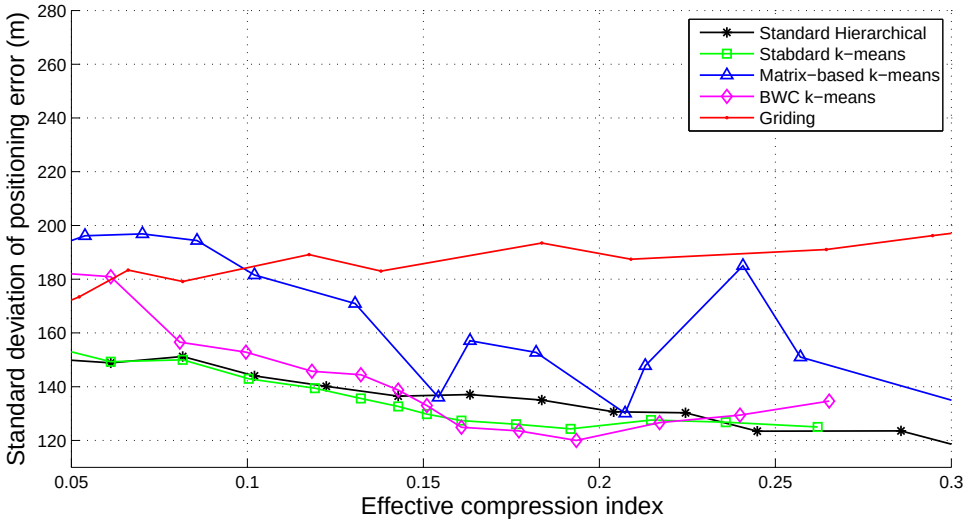
In this section, we examine the positioning performance of various clustering techniques. The performance criteria include the average positioning error and the corresponding standard deviation, which are evaluated for a range of compression index.

#### Simulated scenario

In this section, we examine the positioning performance in the context of the simulated experiments. The simulation results are illustrated in Figure 5.2. As we see in the figure, the performance is evaluated versus the effective compression index for different clustering algorithms, since in the variants of k-means algorithm there exist a certain number of empty clusters in the provided results. The clustering considered algorithms include the proposed BWC technique, the matrix-based weighted k-means, the standard k-means, and




(a) Average positioning error versus effective compression index




(b) Standard deviation of positioning error versus effective compression index

Figure 5.2: Performance of different clustering algorithms (simulated data)

the standard hierarchical clustering with a minimum-variance linkage criterion. All the k-means-type algorithms are initialized by the results of a primary hierarchical clustering.

We observe that for the whole range of $\eta_{Eff}$, the proposed BWC method provides a lower average positioning error w.r.t. other clustering technique and also the griding method; nevertheless the resulted standard deviation is not the most efficient. As a general remark, we note that the error standard deviation for all techniques is in the same order as the error average value. Therefore, we may expect large variations of positioning error during the localization phase.

As we can see in the figure, with a compression index of $\eta_{Eff} \approx 0.25$ an average positioning error of about 120 m could be obtained. Considering the size of GSM cells (1 km), there is a notable improvement with respect to a classic cell-ID method; however, the obtained accuracy is not sufficient for all types of services.

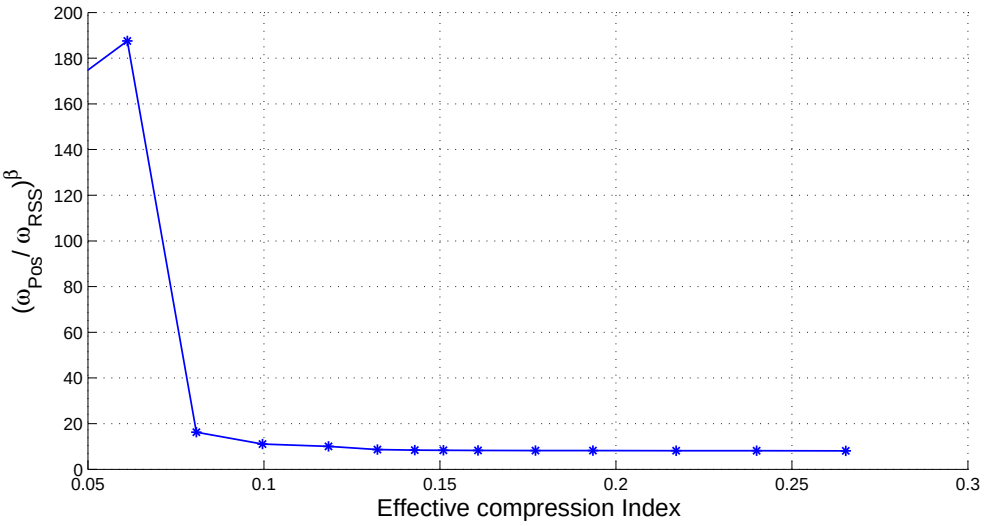



Figure 5.3: The relative weight of position to RSS feature types (simulated data)

In order to study the proposed BWC technique in more details, we examine the relative weight of position to RSS feature types provided by the algorithm. The provided relative weights are illustrated in Figure 5.3; the curve traces the quantity $(\omega_G/\omega_R)^{\beta}$ (to be consistent with the weighting scheme proposed by Equation 4.5). Based on the figure, we note that for $\eta < 0.1$ (i.e. high compression), the position feature type becomes more important w.r.t. RSS feature type (a relative weight of above 100). For $\eta > 0.2$ (lower compression), the relative weight of position to RSS become almost stable at a value of about 8. We note that this relative value is close to the heuristically proposed value in previous chapter

$w_G/w_R = D_R/D_G = 6.5$.

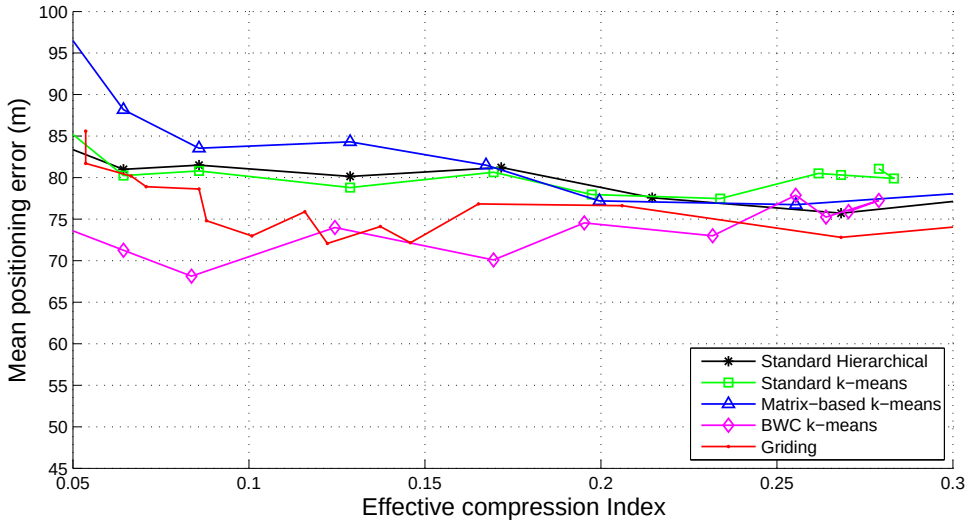
**Real scenario**

Here we examine the performance of various clustering techniques in the context of the real experiments. As in section 5.3.3, the performance is evaluated versus the compression index for different clustering algorithms. The results are illustrated in Figure 5.4.

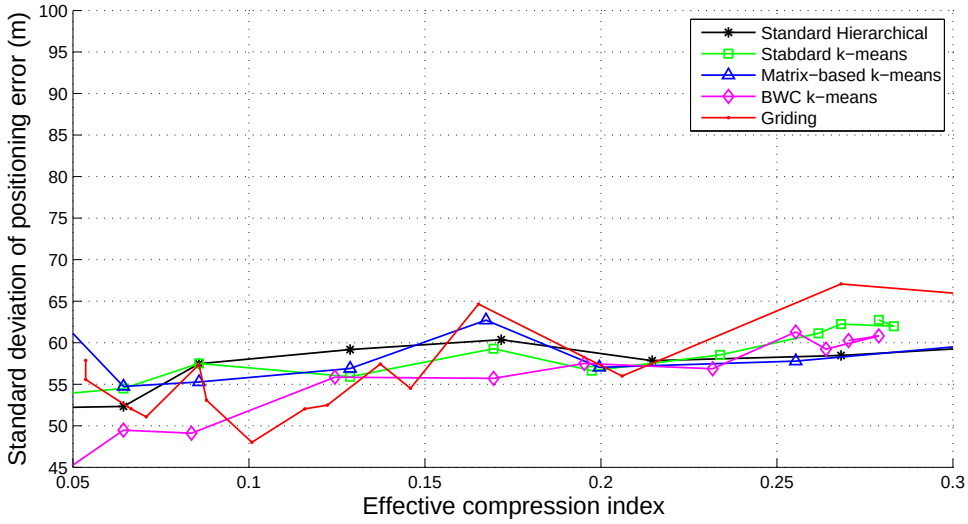
We observe that in general, the proposed BWC method provides a lower average positioning error w.r.t. other methods. However, the simple griding method provides a better performance w.r.t. other clustering techniques. Again, we note that the error standard deviation for all techniques is in the same order as the error average value. Therefore, we may expect large variations of positioning error during the localization phase.

As we can see in the figure, with a compression index of $\eta_{Eff} \approx 0.25$ an average positioning error of about 75 m could be obtained, which is less than $\frac{1}{5}$ of the GSM cell size (here 500 m); so there is a notable improvement with respect to a basic cell-ID method.

Similarly to section 5.3.3, we examine the relative weight of position to RSS feature types provided by the BWC algorithm. The quantity $(\omega_G/\omega_R)^\beta$ is traced versus the compression index $\eta$ in Figure 5.5. We observe that for $\eta < 0.1$ (i.e. high compression), the position feature type gets more important w.r.t. RSS feature type (a relative weight of above 150). For $\eta > 0.2$ (lower compression), the relative weight of position to RSS become almost stable at a value of about 16. Again, we note that this value is close to the heuristically proposed value in previous chapter $w_G/w_R = D_R/D_G = 12.5$




(a) Average positioning error versus effective compression index




(b) Standard deviation of positioning error versus effective compression index

Figure 5.4: Performance of different clustering algorithms (real data)

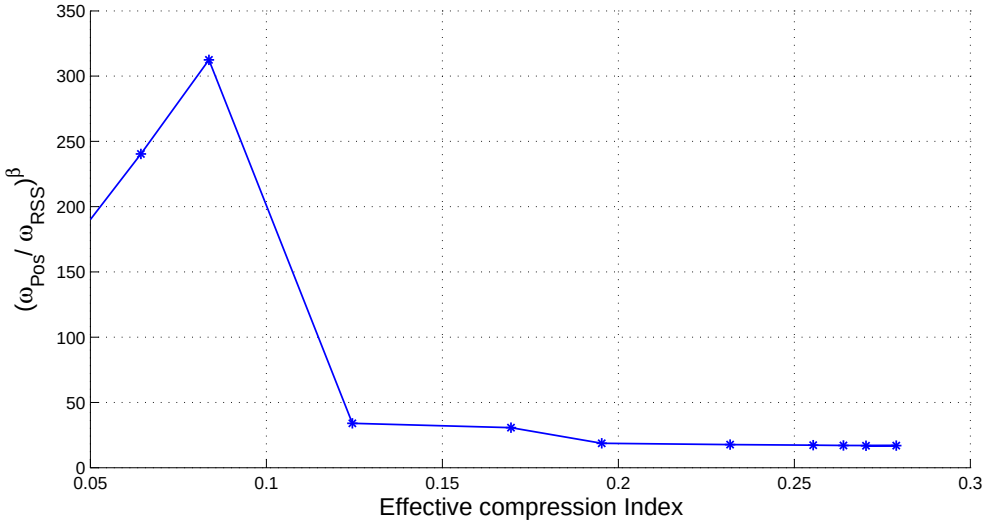



Figure 5.5: The relative weight of position to RSS feature types (real data)

### 5.3.4 Comparison with other compression techniques

The second part of simulations compares the performance of the proposed BWC technique w.r.t. other compression methods: PCA and KCCA. To provide a comparative framework, we trace the Cumulative Distribution Functions (CDF) of positioning error which is defined as follows:

$$F_{PosEr}(a) = p\{\varepsilon(\underline{x}') < a\},$$

where the defined probability is estimated according to empirical experiments.

All the three techniques are implemented at the same compression index $\eta = 0.2$. The case of a non-processed database is also examined to provide a reference performance level.

#### Simulated scenario

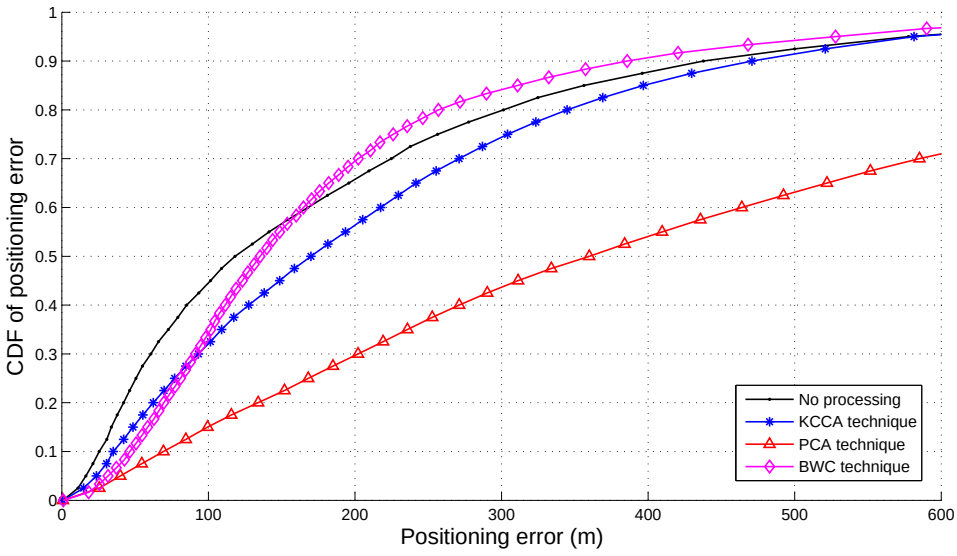



Figure 5.6: Comparison with state-of-the-art compression techniques at $\eta = 0.2$, (simulated data)

The corresponding simulation results are illustrated in Figure 5.6. As we can see, the clustering technique is more efficient than the other techniques; the provided accuracy is competitive to that of a non-compressed database. On the other hand, the PCA technique is the less efficient among the three techniques. In fact at a compression index of $\eta = 0.2$,

the dimension of radio signal space reduces form 13 to 3 in the PCA method; the resulting loss of information degrades considerably the performance. We can see that under the same situation, the KCCA method outperforms the PCA technique.

According to the simulation results, we may deduce that clustering technique provides a positioning performance higher than that of other compression methods, while simultaneously it reduces the transmission and computation costs (see section 4.4). To provide some representative values, the transmission loads corresponding to different techniques are computed for the simulated experiment at $\eta = 0.2$; the realistic values given in section 4.4 are adopted for $b_G$ and $b_R$. The results are presented in table 5.2.

| Proc. meth. | Database transmission load |
|---|---|
| BWC | 17 kbits |
| PCA | 37 kbits |
| KCCA | 88 kbits |
| No processing | 88 kbits |

Table 5.2: Analysis of transmission load for the simulated experiment (at $\eta = 0.2$)

**Real scenario**

Similarly to section 5.3.4, a comparative performance evaluation is presented in this part of experiments. The CDFs of positioning error for various processing techniques are traced in Figure 5.7. The case of a non-processed database is also examined to provide a reference performance level. As in section 5.3.4, the techniques are compared at a compression index of $\eta = 0.2$.

According to the figure, we observe that there is no significant difference between performance of clustering and KCCA techniques; they both provide accuracies close to a non-compressed database. Again, PCA technique is the less efficient compression method. Recalling the high computation complexity of the KCCA method, we may conclude that the clustering technique provides a competitive positioning performance while it needs the lowest computational complexity. Similarly to the simulated experiments in section 5.3.4, we provide some representative values for the transmission loads corresponding to different techniques. The computed values at $\eta = 0.2$ are presented in table 5.3.

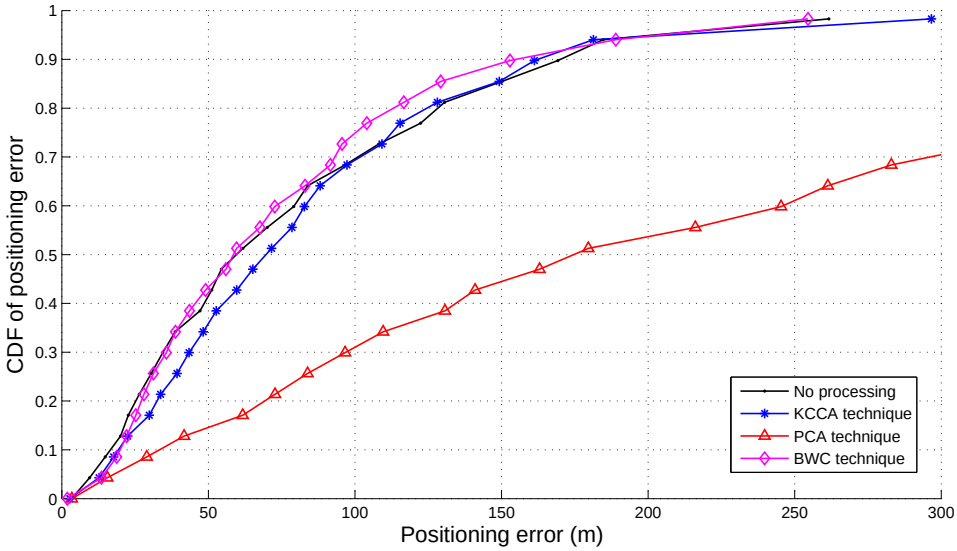



Figure 5.7: Comparison with state-of-the-art compression techniques at $\eta = 0.2$, (real data)

## 5.4 Conclusion

In this chapter, we developed the Block-based Weighted Clustering (BWC) scheme, which is well-tailored to the structure of the records in the radio database.

We defined the concept of *feature types* in association with the records; a feature type is defined as all the stored parameters in a record that belong to the same nature. Based on this attribution, the BWC scheme was proposed; this scheme imposes equal weights to blocks of components belonging to the same feature type, in the clustering cost function. The weight factors associated to feature types are optimized during the clustering process. This might be considered as providing a refined distance metric, adapted to the specific structure of records in the database.

Computer simulations and real experiments were conducted to evaluate the performance of the proposed BWC technique in the context of a cellular fingerprinting system. The results of both simulated and real experiments show that the proposed BWC technique outperforms the standard clustering algorithms and also other compression methods, like PCA and KCCA, taking the positioning accuracy as the performance criterion. In the simulated experiments all the clustering techniques provided a performance superior to that of the naive griding method. Although this superiority was incontrovertible in the simulated scenario, in the real experiments the griding method showed a performance competitive enough, and it outperformed the standard clustering methods. This phenomenon needs

| Proc. meth. | Database transmission load |
|---|---|
| Clustering | 17 kbits |
| PCA | 30 kbits |
| KCCA | 89 kbits |
| No processing | 89 kbits |

Table 5.3: Analysis of transmission load for the real experiment (at $\eta = 0.2$)

further investigations, to interpret the reasons of the degraded performance of the cluster analysis, in the context of the real experiments. One possible reasoning is that the real measurements are picked up on close points (with a distance of about 10 m), and along pavements on street level. One may say that the measurements are performed over some homogenous sub-areas, and hence the clustering techniques are less effective in this context.

It is noteworthy that in this work, minimization of the positioning error was not taken into account as an explicit criterion for optimizing the weight factors. Enhanced clustering methods may be envisaged by considering this criterion. Another important issue which is not treated in this work is the problem of missing data in the radio database. As we mentioned before, a radio database may contain a large number of missing data corresponding to non-detected base stations. In this work, we replaced all these missing data by a single reference value. One may expect an improved performance for clustering algorithms if the missing data are taken into account more intelligently. The optimal handling of radio missing data will be treated in the next chapter.

# Bibliography

[1] J. Z. Huang, M. K. Ng, H. Rong, and Z. Li, "Automated variable weighting in k-means type clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 657–668, May 2005.

[2] H. Frigui., O. Nasraoui, "Unsupervised learning of prototypes and attribute weights," *The journal of the Pattern Recognition Society*, vol. 37, no. 3, pp. 567–581, 2004.

# Chapter 6

# Handling missing data in RSS-based location fingerprinting

## 6.1 Introduction:

### Missing data in location fingerprinting

An important issue in location fingerprinting systems, for which no fully satisfactory solution has been proposed in the literature yet, lies in the missing character of the radio measurements. Missingness is concerned with the incompleteness of the set of measured parameters (e.g. missing RSS or TA/RTT values) in a radio measurement performed at a specific location; it may also concern the inconsistency of different radio measurements performed at different moments, at a given point.

Hereafter in this thesis, we focus on data missingness issued from the "scanning" process, occurring in RSS-based cellular fingerprinting systems. The scanning process is an essential function in cellular networks, which is performed by any mobile terminal in order to select the serving cell. All mobile terminals in cellular networks scan regularly the reference signals of the serving and a number of neighbor cells (the reference signal concerns the BCCH frequency in GSM, or the CPICH bit sequence in UMTS). The list of the neighbor cells is already declared on the network side, and is transmitted to the mobile terminals through broadcast messages (case of GSM and UMTS systems). The scanning process then proceeds by making measurements on the indicated reference signals (RSS measurements in GSM, RSCP measurements in UMTS). All the mobility management procedures in cellular networks such as cell (re)selection, location updating, or handover are based on this process.

However, the performed RSS or RSCP measurements during the scanning process may

contain some non-detected (missing) values, because of various reasons:

• the target signal may be received with a signal level lower than a minimum threshold,

• the target signal may be lost (or jammed) in severe interference; it might happen in the case of pilot pollution in CDMA (see [1] and [2]),

• the set of measured signals may be incoherent between several field measurement campaigns,

• some base stations may be temporarily switched off (either accidentally, or intentionally for energy saving purposes); this switch-off causes missing values for the corresponding components in the radio measurements,

• the number of base stations to be measured in practice is limited by an upper bound; hence at a given point, there might be some detectable base stations who are not measured because of this limitation.

In RSS-based fingerprinting systems, the mobile measurements during the localization phase arise from the scanning process, and hance include missing data due to the reasons given above. In contrast, the radio database is not necessarily constructed based on the classic scanning process (since the network operator may use proprietary tools to perform the measurements). As a result, the database records are not necessarily subject to the same degree of missingness.

Statistical learning procedures for handling missing data have vastly improved since the last decades ([3]). To develop a systematic learning method for incomplete data sets, a theoretical framework should be stated. Adopting the framework proposed in [4], [5], at a first step one defines a *complete data model* which generates the complete data set, and a *missing mechanism* which renders a subset of the complete data set unobservable to the learner. Next, learning methods are developed, which benefit from knowledge on both complete data model and missing mechanism. The goal of the learning process may be to approximate a mapping function between some inputs and target variables, or to extract some statistical description of the inputs ([6]). In this regards, different approaches are proposed in the missing data literature, e.g. missing data imputation methods, maximum likelihood-based estimation methods, etc. ([3]).

The methodological difficulties raised by the missing character of the RSS measurements have not received much attention in the literature of location fingerprinting systems, and the application of statistical data completion techniques in this context is not well-documented. A possible heuristic approach, proposed in works such as [7], [8] and [9], consists in replacing all the missing elements by a single reference value. Another proposed approach, as in [10] and [8], is to discard all the missing elements, and to exploit the observed parts of

measurements only.

It is the purpose of this chapter to develop a statistical method to deal with missing data in RSS-based fingerprinting systems, within the framework of a limited well-defined missing mechanism. Our defined missing mechanism proceeds based on two parameters: the receiver minimum sensitivity for signal detection, and the maximum number of base stations to be measured in the radio measurements. In future works, one may envisage an extended modeling for the missing mechanism, such that it includes more missingness factors. Once the missing mechanism is defined, statistical methods are developed at two different levels. At the first level, the missing mechanism is assumed to be present exclusively during the localization phase; the radio database is supposed to contain no missing elements. Here, a localization algorithm based on Maximum Likelihood (ML) method is proposed, which takes into account the missing mechanism, to compute the likelihoods (see section 6.4.2). At the second level of modeling, the missing mechanism is assumed to be present during both the training and localization phases. Here, a Multiple Imputation (MI) method is proposed to fill in the missing items in the radio database, during the training phase (see section 6.4.3). Once the database is completed, dealing with missing data in the localization phase sends us back to the problem mentioned at the first level.

## 6.2 Inference from missing data

In many practical situations, statistical learning must be performed from incomplete data. In this section, we briefly review the problem of inference based on partially observed data. We adopt the statistical framework proposed by [5] and make a distinction between the *complete data model* which generates the complete data set, and the *missing mechanism* which renders a subset of the complete data set unobservable to the learner ([6]).

The problem of learning from incomplete data may constitute a "supervised" or an "unsupervised" problem. In the case of supervised learning, the goal consists of approximating a mapping function between inputs and the target variables. The unsupervised learning generally consists of extracting some statistical description of the inputs. In both cases, the learner may benefit from knowledge on both complete data model and missing mechanism ([6]).

### 6.2.1 The framework

Consider a complete data set given by $\mathcal{S}^{\circ} = \{\underline{s}_n^{\circ}\}_{n=1,\ldots,N} \subset \mathbb{R}^B$, subject to a missing mechanism. Assume that the set $\Theta$ represent the parameters of the complete data model and $\Phi$ those of the missing mechanism.

Due to the missing mechanism, one cannot observe all the elements of $\mathcal{S}^{\circ}$ in practice. Each data vector $\underline{s}_n^{\circ}$ can be decomposed into an observed part $\underline{s}_n^{\circ(obs)}$ and a missing part $\underline{s}_n^{\circ(mis)}$. Notice that each vector $\underline{s}_n^{\circ}$ may have a different pattern of missing features. At the population level, the complete dataset $\mathcal{S}^{\circ}$ can be similarly divided into the observed part $\mathcal{S}^{(obs)}$, and the missing part $\mathcal{S}^{(mis)}$. To formalize the notion of missing mechanism we define the missingness indicator matrix $\mathcal{I}$ as follows:

$$\mathcal{I} = \{\underline{i}_n\}_{n=1,\ldots,N}, \tag{6.1}$$

where $\underline{i}_n$ is the missingness indicator vector corresponding to $\underline{s}_n^{\circ}$ and given by:

$$\underline{i}_n \in \{0,1\}^B,\ i_{n,b} = \begin{cases} 1 & \text{if } s_{n,b}^{\circ} \text{ is observed} \\ 0 & \text{otherwise} \end{cases} \quad 1 < b < B$$

In modern missing data procedures, missingness is regarded as a *probabilistic* phenomenon ([4], [3]). The missingness indicator matrix $\mathcal{I}$ is treated as a set of random variables having a joint probability distribution ([4], [3]). In this probabilistic viewpoint, two general classes are defined for the missingness:

1. *Ignorable Missingness*: Ignorable missingness describes the situation where the missingness may depend on the observed part of data, but not on the missing part. In other words, we have in this case:

$$p(\mathcal{I}|\mathcal{S}^{\circ}, \Phi) = p(\mathcal{I}|\mathcal{S}^{(obs)}, \Phi). \tag{6.2}$$

2. *Non-ignorable missingness*: When condition (6.2) is not fulfilled, we are said to be in the situation of non-ignorable missingness.

The type of missingness is critical in evaluating learning algorithms for incomplete data ([3]). Estimates of the parameters $\Theta$ and $\Phi$ can be obtained by maximizing the likelihood function of the observed data:

$$p(\Theta, \Phi|\mathcal{S}^{(obs)}, \mathcal{I}) \propto p(\mathcal{S}^{(obs)}, \mathcal{I}|\Theta, \Phi). \tag{6.3}$$

Computing this probability in general is a difficult task. However, it can be shown that, in the case of ignorable missingness, the density factorizes as follows:

$$\begin{aligned} p(\mathcal{S}^{(obs)}, \mathcal{I}|\Theta, \Phi) &= p(\mathcal{I}|\mathcal{S}^{(obs)}, \Phi) \int p(\mathcal{S}^{\circ}|\Theta) d\mathcal{S}^{(mis)} \\ &= p(\mathcal{I}|\mathcal{S}^{(obs)}, \Phi) p(\mathcal{S}^{(obs)}|\Theta). \end{aligned} \tag{6.4}$$

Equation 6.4 reveals that, in the case of ignorable missingness, the parameters of the missing mechanism can be ignored for the purpose of estimating $\Theta$, making computations much more easier.

### 6.2.2 Methods for handling missing data

A wide range of methods have been proposed in the literature to cope with missing data. Among the older methods, the most popular is the simplest, the *case deletion* technique, that consists in discarding all units containing missing elements ([3]), producing in general a strong bias in the estimation. More advanced procedures rely on Maximum Likelihood computation, following a possible data-completion stage ("imputation-based methods", [3], [6]).

**Maximum Likelihood (ML) method.** The principle of drawing inferences from a likelihood function is widely accepted ([3]). The Maximum Likelihood (ML) method, can be utilized for estimating the parameters of the complete data model $\Theta$. As we saw in the previous subsection, in the case of ignorable missingness the marginal distribution of the observed data provides the correct likelihood for the unknown parameters $\Theta$. For non-ignorable missingness, there exist no general approaches ([3]).

**Data imputation.** Imputation is the practice of filling in missing items. Imputation is potentially more efficient than case deletion, because no unit is sacrificed. In addition, if the observed data contain useful information for predicting the missing values, an imputation procedure can exploit it and maintain high precision. Imputation also produces an apparently complete data set that may be analyzed by standard methods and softwares ([3]). There exist two main classes of imputation methods:

1. *Single Imputation* (SI): Single imputation methods assign a single value to each missing item. For instance, *unconditional mean imputation*, a naive SI method, simply substitutes all missing values for a given variable with the average of the observed values. More interesting, SI may be performed by drawing new data from a proper conditional distribution. Suppose that we have the data set $\mathcal{S}^{\circ} = (\mathcal{S}^{(obs)}, \mathcal{S}^{(mis)})$ from distribution $p(\mathcal{S}^{(obs)}, \mathcal{S}^{(mis)}|\Theta)$. Imputing from the conditional distribution means simulating a draw from $p(\mathcal{S}^{(mis)}|\mathcal{S}^{(obs)}, \mathcal{I}, \widehat{\Theta})$, where $\widehat{\Theta}$ is an estimate of $\Theta$ based on the observed data ([3]).

2. *Multiple Imputation* (MI): In MI procedures, each missing value is is replaced by a list of $Q > 1$ simulated values; this leads to produce $Q$ plausible alternative versions of the complete data. Each of the $Q$ data sets is next analyzed by a complete-data method. The results are then combined to obtain overall estimates, that reflect the missing data uncertainty ([3]).

## 6.3 Missing data in RSS measurements

We now describe at length the nature of the data involved in location fingerprinting systems and introduce incidentally the notations that shall be used in the subsequent analysis.

### 6.3.1 Complete RSS measurements

Let us consider the Received Signal Strength (RSS) radio measurements performed by a mobile terminal over an area $\mathcal{A}$, where $B \geq 1$ base stations are present. The vector of complete RSS measurements at an arbitrary location $\underline{x}$ is denoted by:

$$\underline{s} = (s_1, ..., s_b, ..., s_B) \in \mathbb{R}^B, \tag{6.5}$$

where $s_b$ is the RSS level of the $b$-th base station at location $\underline{x}$.

### 6.3.2 Missing mechanism for RSS measurements

As above, consider an area $\mathcal{A}$ where a number of $B$ base stations are present. Let $\underline{s}$ be a vector of RSS measurements performed at location $\underline{x}$. The missing mechanism in this context can be described by means of two parameters: $\lambda$, the minimum sensitivity of the terminal receiver, and $B_{max}$, the maximum number of observable base stations in the measurements of actual terminals ($B_{max} \leq B$). In practice, in a measurement vector, the $B_{max}$ largest measures (corresponding to the $B_{max}$ best received base stations at the mobile position) are observed, provided they are all above the threshold value $\lambda$.

To describe the missing mechanism mathematically, sort the components of $\underline{s}$ by increasing order of magnitude: let $\underline{\sigma} = (\sigma(1), \sigma(2), ..., \sigma(B))$ be the permutation of the base station indexes such that $s_{\sigma(1)} \geq s_{\sigma(2)} \geq ... \geq s_{\sigma(B)}$. In accordance with section 6.2.1, the missingness indicator vector $\underline{i} \in \{0,1\}^B$ corresponding to $\underline{s}$ is determined as follows:

$$\forall b, 1 \leq b \leq B,\ i_b = \begin{cases} 1 & \text{if } b \in \{\sigma(1), \sigma(2), ..., \sigma(B_{max})\}, \\ & \text{and } s_b \geq \lambda\ , \\ 0 & \text{otherwise.} \end{cases}$$

We denote $\Psi$ the set of missing mechanism parameters in the specific context of RSS measurements, i.e. $\Psi = \{\lambda, B_{max}\}$. Accordingly, the set of observed base stations at any location $\underline{x}$ is given by $\mathcal{B}_\Psi(\underline{x}) = \{b :\ i_b = 1\}$.

It is noteworthy that the missing mechanism in the context of radio fingerprinting is categorized as non-ignorable, which leads to a complicated situation. As a final remark, we notice that the missing mechanism here is not a random process; the parameters in $\Psi$

are deterministic, and known by the system engineer. Hence, for the sake of simplicity, we shall omit the subscript in $\mathcal{B}_\Psi(\underline{x})$, i.e. $\mathcal{B}(\underline{x}) = \mathcal{B}_\Psi(\underline{x})$, throughout the rest of the paper.

## 6.4 Handling missing data in fingerprinting systems

In fingerprinting systems, data missingness may occur frequently in RSS measurements of both training and localization phases. In this study, we develop statistical methods tailored for partially observed data, at two different levels. At the first level, the missing mechanism is assumed to be present exclusively during the localization phase; in other words, the radio database is supposed to contain no missing elements ("complete database - incomplete mobile measurements"). This level of analysis may be considered as a simplified version of the problem. Here, a localization algorithm based on Maximum Likelihood (ML) method is proposed, which deals with missing data in RSS measurements of the localization phase.

At the second level of modeling, we drop the assumption of "complete database"; in other words, the missing mechanism is assumed to be present during both the training and localization phases ("incomplete database - incomplete mobile measurements"). Now, a Multiple Imputation (MI) method is proposed to fill in the missing items in the radio database, during the training phase. Once the database is completed, dealing with missing data in the localization phase sends us back to the problem mentioned at the first level.

### 6.4.1 Notations

As defined in our preceding works, a radio database $\mathcal{R}$ is a set of *records*. A record is a vector $\underline{r} = (\underline{x}, \underline{s})$ where $\underline{x}$ represents a geographical position and $\underline{s}$ is the corresponding measurement vector in the radio signal space. The included positions in the database may be numbered from 1 to $M$ and given by the set:

$$\chi = \{\underline{x}_1, ..., \underline{x}_m, ..., \underline{x}_M\}.$$

The radio database is then given by $\mathcal{R} = \{\underline{r}_m\}_{m=1...M}$.

As mentioned in previous chapters, the database $\mathcal{R}$ may be provided by processing the elements of an *initial* radio database; a radio database constructed according to raw field measurements is called an initial database $\mathcal{R}^\circ$. A record of $\mathcal{R}^\circ$ is given by $\underline{r}^\circ = (\underline{x}^\circ, \underline{s}^\circ)$. The raw measurements are performed at geographical positions called *elementary points*, given by the set $\chi^\circ = \{\underline{x}_1^\circ, ..., \underline{x}_n^\circ, ..., \underline{x}_N^\circ\}$. The corresponding RSS measurements may be given by $\mathcal{S}^\circ = \{\underline{s}_1^\circ, ..., \underline{s}_n^\circ, ..., \underline{s}_N^\circ\}$. The initial database is then described by $\mathcal{R}^\circ = \{\underline{r}_n^\circ\}_{n=1,...,N}$.

Figure 6.4.1 illustrates the architecture of the fingerprinting system adopted in this work. The adopted architecture is the one proposed in our preceding works, where a "clustering" step is inserted in the training phase to process the initial database. The clustering step has no special impact in the context of missing data; it is inserted for operational purposes as described in [9] and [11]. By applying the clustering technique, an initial raw data base $\mathcal{R}^\circ = \{\underline{r}_n^\circ\}_{n=1...N}$ may be compressed into a more concise radio database $\mathcal{R} = \{\underline{r}_m\}_{m=1...M}$ $(M < N)$. The clustering is performed by using a k-means type algorithm (for more details see [9] and [9]).

During the localization phase, the mobile terminal performs a radio measurement $\underline{s}'$ at location $\underline{x}'$, Once the location of the terminal is estimated based on a positioning algorithm, the corresponding positioning error in meters is given by:

$$\varepsilon(\underline{x}') = \|\underline{x}' - \widehat{\underline{x}}\|,$$

where $\underline{x}'$ is the mobile estimated position.

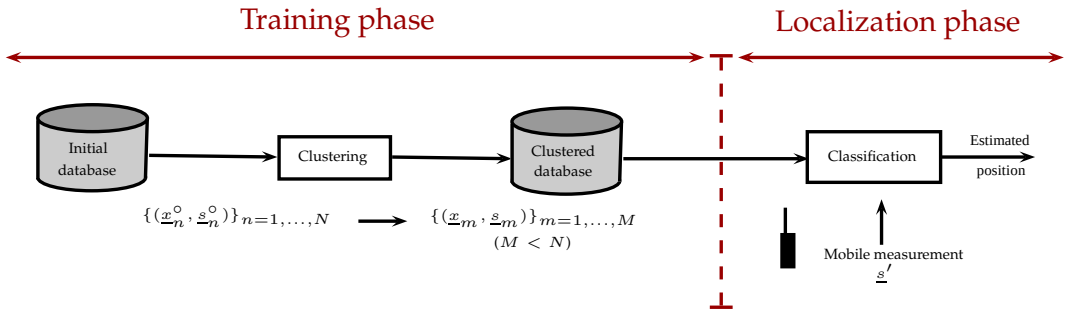



Figure 6.1: The adopted architecture for the fingerprinting system

### 6.4.2 Complete database - Incomplete mobile measurements

During the localization phase, the mobile terminal performs a radio measurement $\underline{s}'$ at location $\underline{x}'$ which is subject to the missing mechanism. Assuming that the complete clustered database $\mathcal{R}$ is provided, the goal is to localize the mobile terminal based on $\underline{s}'^{(obs)}$. We go on by formalizing the complete data model, and then developing a maximum likelihood-based positioning algorithm.

#### Complete data model

The complete data model here must describe the complete measurement vector $\underline{s}'$. We assume that the data points in each cluster are Gaussian. The Gaussianity hypothesis for

RSS measurements around a central geographic point (*i.e.* the cluster centroid) has been proposed in previous works, such as [12] and [13]. Here, as the clusters are constructed through the k-means algorithm, the Gaussianity of the data lying in each cluster is less questionable. Under this assumption, the vector of measurements in the $m$-th cluster is assume to be distributed as:

$$p(\underline{s}'|\, m, \Theta_L) \sim \mathcal{N}(\underline{s}_m, \Gamma_m), \tag{6.6}$$

with

$$\Theta_L = \{(\underline{s}_m, \Gamma_m)\}_{m=1,\ldots,M},$$

where $\underline{s}_m$ and $\Gamma_m$ are respectively the centroid and the covariance matrix of the $m$-th cluster, provided upon the clustering process; $\Theta_L$ denotes the set of complete data model parameters, for the localization phase. If there is no cross correlation between the signals coming from different base stations (i.e. $\Gamma_m$ is diagonal), we have:

$$p(\underline{s}'|\, m, \Theta_L) = \prod_{b=1}^{B} p_b(s'_b|\, m, \Theta_L), \tag{6.7}$$

where $p_b(.|\, m, \Theta_L)$ denotes the marginal density of the $b$-th component, $1 \leq b \leq B$.

**Maximum Likelihood (ML) positioning**

Maximum likelihood-based methods are already used in the literature as localization algorithms ([14], [15], [16]). The proposed ML method here is different in the sense that it takes into account the missing mechanism to compute the likelihoods (as in Equation (6.3)).

Assume that the RSS measurement $\underline{s}'$ may be decomposed into an observed part $\underline{s}'^{(obs)}$ and a missing part $\underline{s}'^{(mis)}$, giving a missingness indicator vector $\underline{i}'$. The localization algorithm is proposed to return the class with the highest likelihood, as follows:

$$\widehat{\underline{x}} = \underline{x}_{\widehat{m}}, \;\; \widehat{m} = \text{argmax}_m \;\; p(\underline{s}'^{(obs)}, \underline{i}'|\, m, \Theta_L, \Psi) \tag{6.8}$$

where $\widehat{\underline{x}}$ is the estimated location of the terminal. According to the described missing mechanism, we may write:

$$p(\underline{s}'^{(obs)}, \underline{i}'|\, m, \Theta_L, \Psi) \;=\; \int_{\xi} p(\underline{s}'^{(obs)}, \underline{s}'^{(mis)}|\, m, \Theta_L) d\underline{s}'^{(mis)}$$

where $\xi$ is the event defined by:

$$\xi = \left\{\underline{s}' :\; \forall b \notin \mathcal{B}(\underline{x}'), \;\; s'_b \leq \lambda'(\underline{x}')\right\}, \tag{6.9}$$

with

$$\lambda'(\underline{x}') = \begin{cases} \lambda & \text{if } |\mathcal{B}(\underline{x}')| < B_{max} \\ \min\{\underline{s}'^{(obs)}\} & \text{if } |\mathcal{B}(\underline{x}')| = B_{max} \end{cases} \tag{6.10}$$

If the components of the measurement vector are independent as in Equation (6.7), we have the following closed-form expression:

$$p(\underline{s}'^{(obs)}, \underline{i}' |\ m, \Theta_L, \Psi) =$$
$$\prod_{b \in \mathcal{B}(\underline{x}')} p_b(s'_b |\ m, \Theta_L) \prod_{b \notin \mathcal{B}(\underline{x}')} F_b(\lambda'(\underline{x}') |\ m, \Theta_L) \tag{6.11}$$

where $F_b(.|\ m, \Theta_L)$ denotes the marginal cumulative distribution function of the Gaussian distribution given by Equation (6.6), corresponding to the $b$-th radio component.

### 6.4.3 Incomplete database - Incomplete mobile measurements

In this section, the missing mechanism is supposed to be present during both training and localization phases. Figure 6.2 illustrates the fingerprinting system, in the context of this scenario.

Here, we propose to apply an imputation technique to fill in the missing values in the initial radio database. Once the database is imputed, dealing with missing data in localization phase leads to the problem treated in the last subsection.

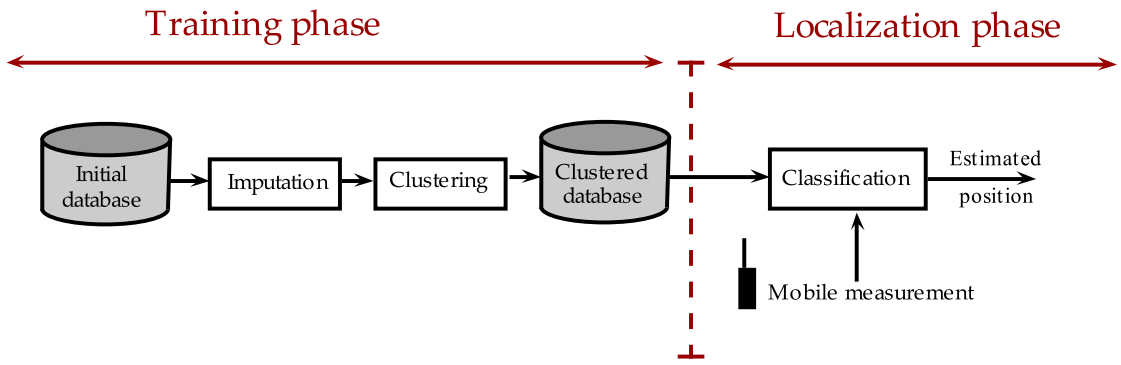



Figure 6.2: The architecture of fingerprinting system with database imputation

#### Complete data model

The complete data model here must describe the set of complete measurements $\mathcal{S}^{\circ}$. Adopting the one-slope log-normal model for the shadowing effect, a measurement vector

$\underline{s}(\underline{x}_n^\circ)$ may be considered as a multi-variate Gaussian variable as follows:

$$p(\underline{s}_n^\circ|\Theta_T) \sim \mathcal{N}([\mu_{n,1}, ..., \mu_{n,B}], \Sigma), \tag{6.12}$$

with:

$$\Theta_T = \{a_1, \alpha_1, ..., a_B, \alpha_B, \sigma_{Sh}\}, \tag{6.13}$$

$$\mu_{n,b} = a_b - 10\alpha_b \log(d_b(\underline{x}_n^\circ)), \tag{6.14}$$

$$\Sigma = \sigma_{Sh}\, \mathbf{I}_B, \tag{6.15}$$

where $a_b$ is a constant, $\alpha_b$ is the propagation exponent, $d_b(\underline{x}_n^\circ)$ is the distance between location $\underline{x}_n^\circ$ and the $b$-th base station, $\sigma_{Sh}$ is the standard deviation of the shadowing effect, and $\mathbf{I}_B$ is the $B \times B$ identity matrix; the set $\Theta_T$ includes all the complete data model parameters for the training phase. If there is no cross correlation between the signals coming from different base stations (i.e. $\Sigma$ is diagonal), we may write:

$$p(\underline{s}_n^\circ|\Theta_T) = \prod_{b=1}^{B} p_b(\underline{s}_{n,b}^\circ|\Theta_T), \tag{6.16}$$

where $p_b(.|\Theta_T)$ is the marginal density of the $b$-th component, for $1 \leq b \leq B$. Additionally, in absence of autocorrelation for the shadowing effect, the complete data set $\mathcal{S}^\circ$ may be described as follows:

$$p(\mathcal{S}^\circ|\Theta_T) = \prod_{n=1}^{N} p(\underline{s}_n^\circ|\Theta_T). \tag{6.17}$$

### Multiple Imputation (MI)

The idea of Multiple Imputation (MI) was introduced in section 6.2.2. It is performed here by means of the Monte Carlo EM algorithm (MCEM), see [17] and [18]. It is implemented in an iterative fashion, repeating the E- and M- steps (consisting respectively in the approximate computation of the expected log-likelihood based on the current estimates for the unobserved variables and in its maximization), as described by Algorithm 2.

**Monte Carlo E-step: Sampling from the predictive distribution.** This step requires to sample new data from the current conditional distribution estimate, $P(\mathcal{S}^{(mis)}|\mathcal{S}^{(obs)}, \mathcal{I}, \Theta_T^t)$, in order to approximate the conditional expectation of the log-likelihood through a Monte-Carlo scheme, as one faces obvious computational difficulties when trying to perform numerical integration. Given the very high complexity of this distribution, data are here generated by means of a simple accept-reject sampling algorithm. Assuming zero auto correlation and cross correlation for the shadowing effect, the accept-reject method can be implemented iteratively, as in Algorithm (3).

**Algorithm 2** MCEM algorithm for Multiple Imputation

(INITIALIZATION)

Start with an initial guess for the parameters: $\Theta_T^{(0)}$.

(ITERATIONS)

For $t \geq 0$, iterate until "convergence:

**Monte Carlo E step:** From the conditional distribution $p(\mathcal{S}^{(mis)}|\mathcal{S}^{(obs)}, \mathcal{I}, \Theta_T^{(t)})$, draw $Q$ samples $\{\mathcal{S}^{(mis),(q)}\}_{q=1,\ldots,Q}$.

**M step:** maximize the resulting conditional expectation and set

$$\begin{aligned}\Theta_T^{(t+1)} &= \text{argmax}_{\Theta_T}\, \mathbb{E}[\log P(\mathcal{S}^{(obs)}, \mathcal{S}^{(mis)}|\Theta_T)] \\ &= \text{argmax}_{\Theta_T}\, \frac{1}{Q}\sum_q \log P(\mathcal{S}^{(obs)}, \mathcal{S}^{(mis),(q)}|\Theta_T).\end{aligned}$$

(STOPPING RULE) The algorithm terminates when the change in value of $\Theta_T^{(t)}$ is negligible.

**M-step: Analysis of complete database.** In the Gaussian linear regression model, it is well-known that log-likelihood maximization and least-square minimization lead to the same estimate, see [19] for instance. Here we thus estimate the parameters of the model given by Equation (6.12) by the Ordinary Least Squares method (OLS) .

We point out that Monte-Carlo simulations involved in the E-step could possibly be performed by means of Markov chain Monte Carlo (MCMC) techniques or by using sequential importance sampling methods (see Chapters 6-8 in [20] for instance). However, the very high complexity of the distribution of the latent/unobserved variables renders the design of such sampling schemes very complicated here and significant progress in the analysis of its structure is required for implementing them efficiently in practice. Statistical inference is tackled from the angle of ML estimation in this work; notice in addition that Bayesian versions of the procedures described here could also be considered if prior information is available, leading to view $\Theta_T$ as a latent variable too.

**Algorithm 3** Accept-reject algorithm

For each $n$, $1 \leq n \leq N$, repeat the following steps:

1. In accordance with Equation (6.10), compute $\lambda'(\underline{x}_n^\circ)$ as follows:

$$\lambda'(\underline{x}_n^\circ) = \begin{cases} \lambda & \text{if } |\mathcal{B}(\underline{x}_n^\circ)| < B_{max} \\ \min\{\underline{s}_n^{\circ(obs)}\} & \text{if } |\mathcal{B}(\underline{x}_n^\circ)| = B_{max} \end{cases}$$

2. For each $b \in \mathcal{B}(\underline{x}_n^\circ)$, repeat:

   - a) Draw from the distribution $P(s_{n,b}^\circ|\Theta_T^t)$.
   - b) If $s_{n,b}^\circ \leq \lambda'(\underline{x}_n^\circ)$ accept the sample. If not, go to a.

## 6.5 Simulations setup

Computer simulations are conducted to evaluate the performance of the proposed approaches. In this section, we introduce the adopted configurations to perform the simulations.

### 6.5.1 Modeling the radio propagation

Similarly to computer simulations in the previous chapters, we assume that the localization service is offered over a geographical area $\mathcal{A}$, which is covered by a GSM cellular network. The GSM cells are again considered to be omnidirectional hexagonal with a radius of 1 km. The area $\mathcal{A}$ covers a surface of $B = 13$ cells (one reference central cell and two rings of neighboring cells). To model the RSS measurements at an arbitrary location in the area, we adopt the Mondrian model introduced in the previous chapter (see section 3.2.1).

For any transmitter-receiver link, the path loss $PL_a$ is determined according to Equation (3.2). The average received signal power is then given by $P_T - PL_a$, where $P_T$ represents the transmitted power. An instantaneous RSS measurement may be modeled as follows:

$$s = P_T - PL_a + X_{Noise}, \tag{6.18}$$

where $X_{Noise} \sim \mathcal{N}(0, \sigma_{Noise}^2)$ is a gaussian random variable (in dB) with standard deviation $\sigma_{Noise}$ denoting the measurements noise.

### 6.5.2 Fingerprinting system configurations

In order to model an urban environment, a geographical area with an equivalent $\alpha$ of 3.8 has been generated. The test area is limited to the surface of the central cell, in order to eliminate the border effects. Each geographical location in the area may be described by a 2-dimensional vector. A complete RSS measurement is assumed to contain the signal components concerning all the $B$ base stations in the area (here $B = 13$). Any single RSS component is simulated according to Equation (6.18), with $\sigma_{Noise} = 0$ dB (noiseless scenario). The missing mechanism is then applied on the simulated data, based on realistic operational constraints. According to the GSM standard, the missing mechanism is given by $\lambda = -110$ dBm and $B_{max} = 7$.

The simulated experiments are based on the system architecture given in Figure 6.1. Concerning the training phase, an initial database is constructed by simulating a number of $N = 1225$ measurements. The measurements are simulated according to a regular pattern over the central cell. Then by applying a clustering technique, a compressed database with a reduced number of records $M = 100$ is provided. Concerning the localization phase, a number of 1000 test measurements have been simulated. These measurements are uniformly distributed over the central cell.

## 6.6 Simulation results

The performance criterion here is the positioning error (in meters), which may be evaluated by its average value, or the corresponding Cumulative Distribution Function (CDF). In the conducted simulations, the performance is traced versus the base stations transmission power $P_T$, since it may be an influential factor in the context of missing data. A higher transmission power implies a higher percentage of missing data due to $B_{max}$, while a lower $P_T$ implies a higher percentage of missing data due to $\lambda$. Figures 6.3 and 6.4 show the obtained results for the simulated experiments.

### 6.6.1 Complete database-Incomplete mobile measurements

Four different approaches are implemented in the simulations, to deal with the missing data:

- Fixed Imputation: here all the missing items in the mobile measurement are replaced by a fixed reference value ($\lambda$). A simple nearest neighbor method is then used to return the cluster with the closest centroid as the mobile position.

- Observed ML: in this approach the ML method is implemented by ignoring the missing items. In other words, we compute only the likelihood of the observed data. The estimated position is given by:

$$\widehat{\underline{x}} = \underline{x}_{\widehat{m}}, \;\; \widehat{m} = \text{argmax}_m \;\; p(\underline{s}'^{(obs)} | \; m, \Theta, \Psi).$$

- Full ML: this scenario implements the proposed Full ML localization algorithm, given by Equation (6.8) and computed in Equation (6.11), which takes into account the missing mechanism.

- No missing data: in this reference case, no missing mechanism is applied during the localization phases. The complete mobile measurements are used to localize the mobile terminal, by using a nearest neighbor method.

The first and second scenarios above implement the classic approaches to handle the missing data in fingerprinting systems. The forth scenario serves as a reference case, and provides an upper bound of performance.

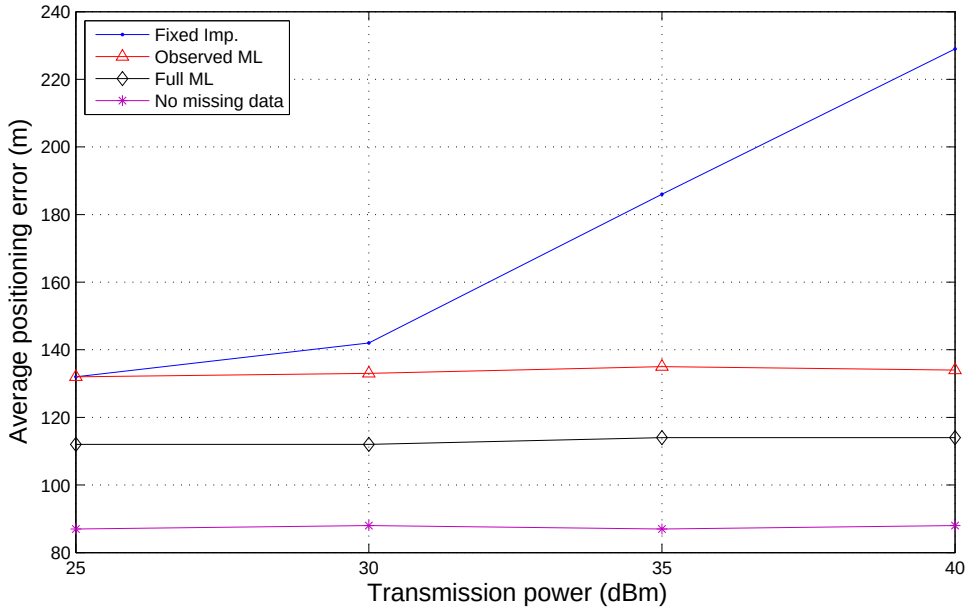
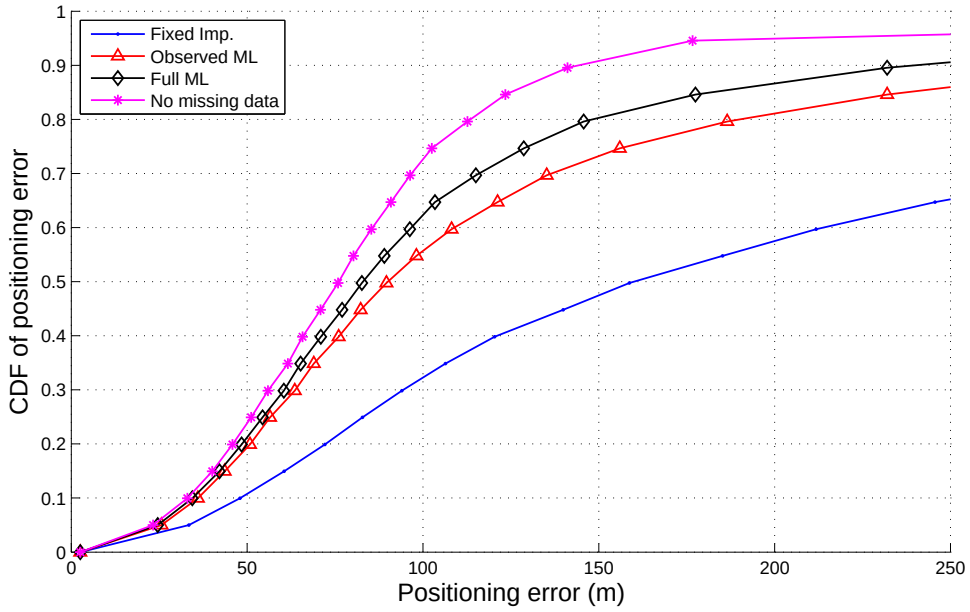
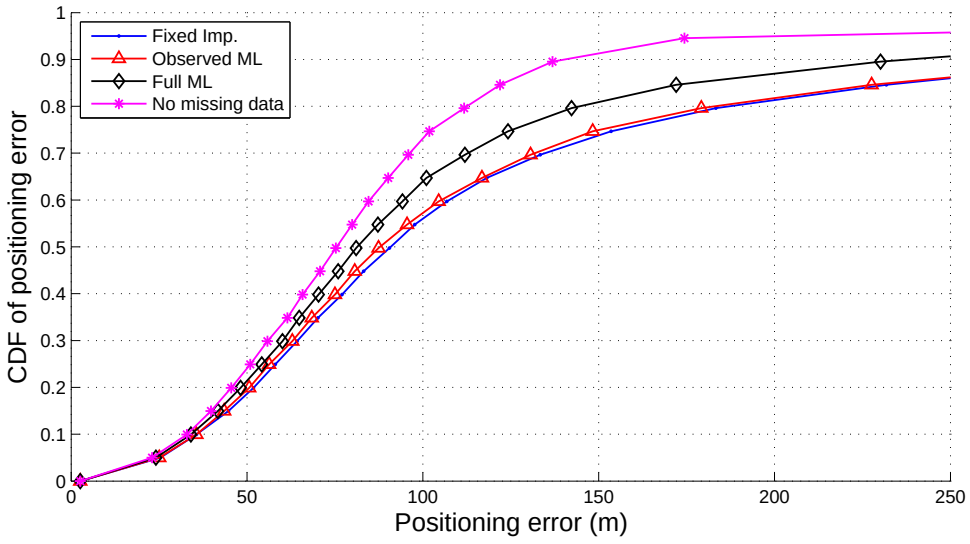



Figure 6.3: Average positioning error versus transmission power




(a) Pt=40 dBm




(b) Pt=25 dBm

Figure 6.4: CDF of positioning error for $P_T = 40$ dBm and $P_T = 25$ dBm

Figure 6.3 illustrates the average positioning error versus the transmission power $P_T$, for the four defined approaches. As we can see, the proposed full ML method outperforms the Fixed Imputation and the Observed ML methods. The improvement is about 15% w.r.t the Observed ML method, and varies between 15% and 45% w.r.t the Fixed Imputation method. We note that, the Observed and the Full ML methods provide a constant performance level for the whole interval of transmission power. On the contrary, the Fixed Imputation method degrades as the transmission power increases. The reason is that a higher value of $P_T$, leads to higher values for components in $\underline{s}'$. In such a situation, the percentage of missing data due to $B_{max}$ increases, while that of $\lambda$ decreases. On the sequel, filling out all the missing items by $\lambda$ leads to larger errors.

The same tendencies are observed in figure 6.4, where the CDF of positioning error is depicted for $P_T = 40$ dBm and $P_T = 25$ dBm. We see that, the curves corresponding to Observed and the Full ML methods provide close performances in both cases. On the contrary, the Fixed Imputation method degrades significantly for $P_T = 40$.

### 6.6.2 Incomplete database-Incomplete mobile measurements

Four different scenarios are examined in the simulations:

- Fixed Imp. - Fixed Imp.: This is the classic scenario that applies a deterministic imputation method for both training and localization phases. Here, in both phases, all the missing items are replaced by a deterministic reference value ($\lambda$). Then the ML localization algorithm given by Equation (6.8) is applied on the imputed database and imputed mobile measurements. We note that using the ML method in this scenario is equivalent to applying the ordinary KNN method on the imputed measurements.

- MI - Full ML: This scenario implements our proposed method. Here, during the training phase, the database is imputed by a Multiple Imputation (MI) method. Then the Full ML localization algorithm given by Equation (6.8) is applied on the incomplete mobile measurements.

- No missing data - Full ML: This scenario applies no missing mechanism during the training phase; so it requires no imputation. To handle the incomplete measurements during the localization phase, the proposed ML localization algorithm is implemented. This scenario serves as a reference case, allowing to evaluate the quality of database imputation.

- No missing data - No missing data: This scenario applies no missing mechanism during neither the training, nor the localization phases; so it does not require imputation

in either phases. This scenario serves as a reference case, and provides an upper bound of performance.

We point out that the observed ML method has not been examined here, since we can not apply the clustering step on an incomplete database.

Figures 6.5 and 6.6 show the obtained results for the simulated experiments. Figure 6.5 illustrates the means positioning error versus the transmission power $P_T$, for the four defined scenarios. We observe that in most cases, the proposed MI-Full ML method outperforms the naive Fixed Imp. method. Although at $P_T = 25$ dBm no improvement is provided by applying the MI-Full ML method, the performance improvement goes up to 30 %, as the transmission power increases. The reason of this phenomenon is that for a higher transmission power, a higher percentage of missing data is due to $B_{max}$. The information of this portion of missing data may be retrieved by using the proposed treating methods. On the other hand, the other portion of missing data due to $\lambda$ is naturally not retrievable. Therefore, the performance of proposed method improves as the transmission power increases.

More details could be observed in figure 6.6, where the CDF of positioning error is traced for $P_T = 40$ and $P_T = 25$. Again, we note that for the case $P_T = 40$ the curves of "MI-Full ML" and "No mis. data-Full ML" are very close, and they outperform the classic Fixed Imputation method. On the other hand, at $P_T = 25$ the three curves are very close, which means that the treatment of missing data in this situation is not really advantageous.

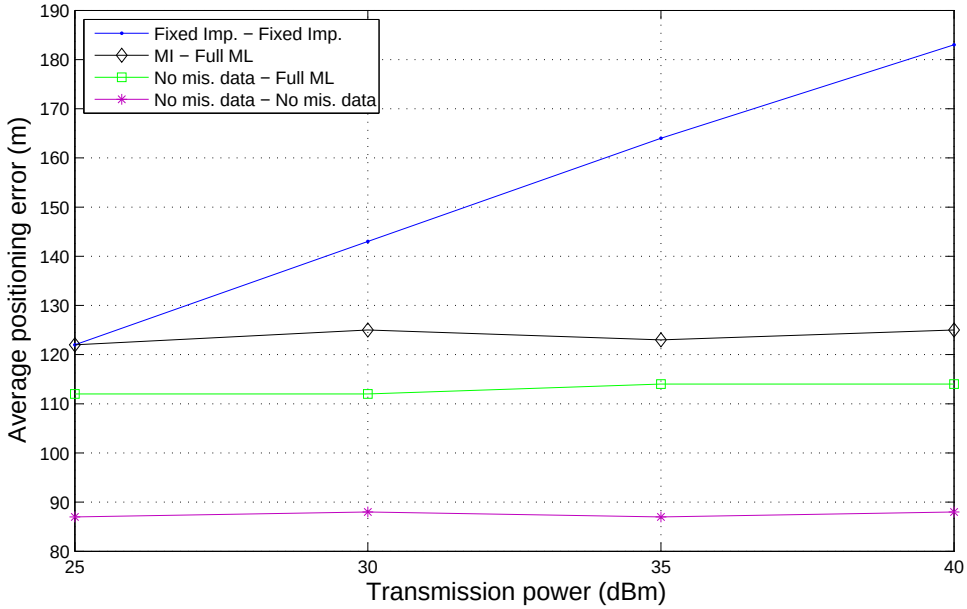
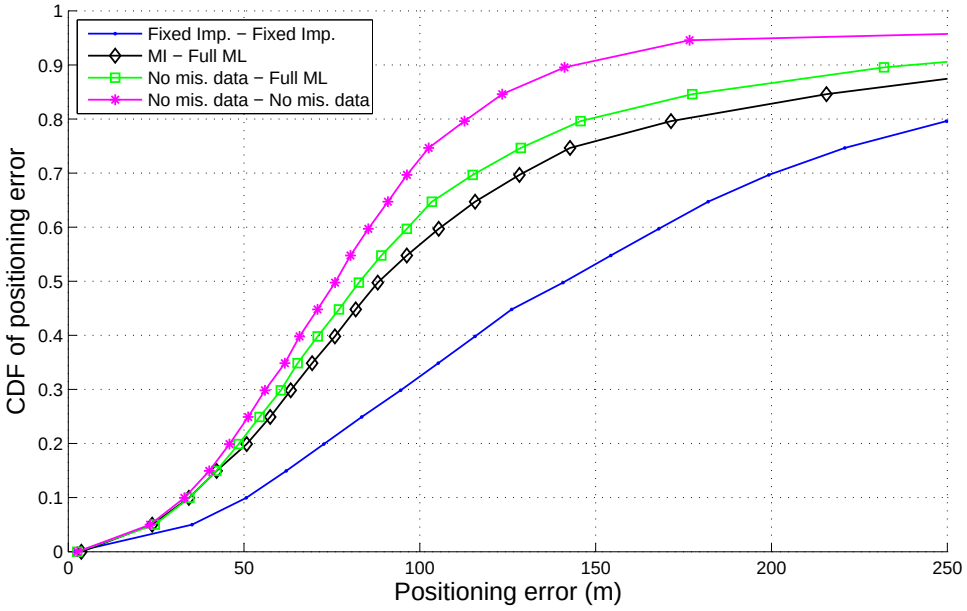
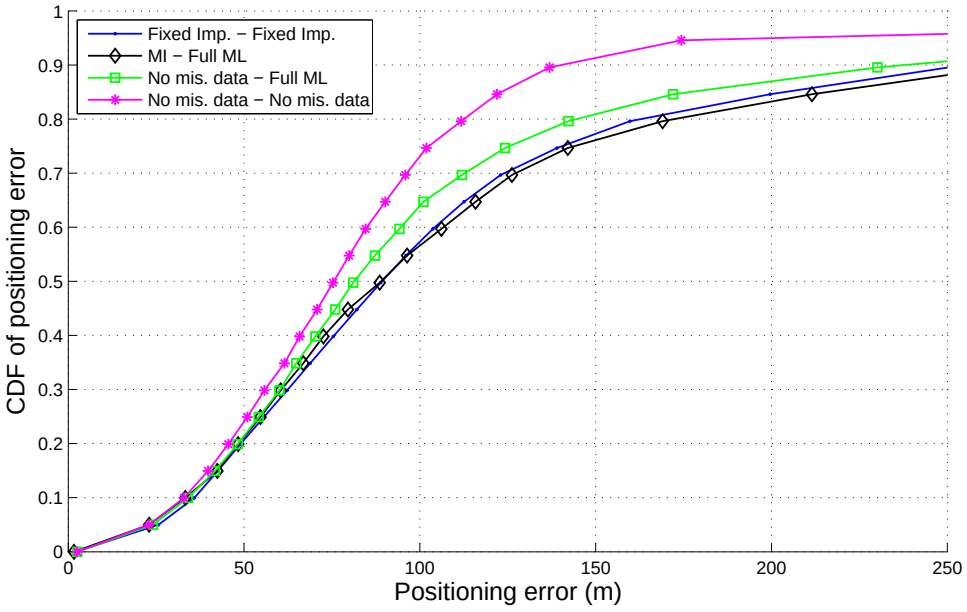



Figure 6.5: Average positioning error versus transmission power




(a) Pt=40 dBm




(b) Pt=25 dBm

Figure 6.6: CDF of positioning error for $P_T = 40$ dBm and $P_T = 25$ dBm

## 6.7 Conclusion and discussion

In this chapter, a systematic statistical method is developed to deal with missing data in fingerprinting systems. A specific model is proposed to describe the missing mechanisms in the context of cellular RSS measurements. Next, statistical techniques tailored for such missing data are proposed. At a first level, the missing mechanism is assumed to be present exclusively during the localization phase; a maximum likelihood-based positioning procedure is then proposed, which takes into account the missing mechanism explicitly (Full ML method). At the second level of modeling, the missing mechanism is assumed to be present during both training and localization phases. Here a Multiple Imputation (MI) data completion algorithm is developed specific to the training phase. Then, handling the missing data in localization phase send us back to the first level of the modeling.

The efficiency of the promoted statistical methodology was supported by simulation results, in the context of a GSM fingerprinting system. Concerning the first level of the problem, the positioning performance of the proposed full ML method was compared with the Fixed Imputation and the Observed ML methods, where the Full ML technique notably outperformed the others. In our simulated scenario, the improvement was about 15% w.r.t the Observed ML method, and varied between 15% and 45% w.r.t the Fixed Imputation method.

Concerning the second level of the problem, we observe that in most cases, the proposed MI- Full ML method outperforms the naive Fixed Imp. method. Precisely, for a low transmission power, no improvement was provided by applying the MI- Full ML method. But the performance improvement goes up to 50 %, as the transmission power increases. In other words, the performance of proposed method improves as the transmission power increases.

Recall that the sampling step of the proposed MI method in this work was implemented by using a non-optimal accept-reject method. This sampling could possibly be performed by means of Markov chain Monte Carlo (MCMC) techniques or by using sequential importance sampling methods. However, the very high complexity of the distribution of the unobserved variables renders the design of such sampling schemes very complicated.

As a further discussion, we note that the proposed method in this work can deal with the missing data which are generated according to the missing mechanism defined in section 6.3.2. Other types of missingness (e.g. issued from temporary switch-off of base stations) can not be treated by the proposed framework. In future works, one may envisage an extended modeling for the missing mechanism, such that it includes more missingness factors.

# Bibliography

[1] M.M. El-Said, A. Kumar, A.S. Elmaghraby, "Pilot pollution interference reduction using multi-carrier interferometry," in *Proceedings of the 8th IEEE International Symposium on Computers and Communication*, vol. 2, 2003, pp. 919 – 924.

[2] J. Niemela, J. Lempiainen, "Mitigation of pilot pollution through base station antenna configuration in WCDMA," in *Proceedings of the IEEE 60th Vehicular Technology Conference*, vol. 6, 2004, pp. 4270 – 4274.

[3] J. L. Schafer, J. W. Graham , "Missing data: our view of the state of the art," in *Psychological Methods*, vol. 7, no. 2, June 2002, pp. 147–177.

[4] D. B. Rubin, "Inference and missing data," *Journal of Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[5] R. J. A. Little, and D. B. Rubin, *Statistical Analysis with Missing Data.* Wiley, New York, 1987.

[6] Z. Ghahramani and M.I. Jordan, "Learning from incomplete data," Lab Memo No. 1509, CBCL Paper No. 108, MIT AI Lab, Tech. Rep., 1995.

[7] C. Dessiniotis, "Motive project deliverable 2.1," European Project FP6-IST 27659, Tech. Rep., September 2006.

[8] Z. Wu, C. Li, J. Kee-Yin Ng, and K. R.P.H. Leung, "Location estimation via support vector regression," *IEEE Transactions on Mobile Computing*, vol. 6, no. 3, pp. 311 – 321, march 2007.

[9] A. Arya, P. Godlewski, P. Mellé, "A hierarchical clustering technique for radio map compression in location fingerprinting systems," in *Proceedings of the International Conference on Vehicular Technology*, May 2010, pp. 1 – 5.

[10] D. Zimmermann, J. Baumann, M. Layh, F.M. Landstorfer, R. Hoppe, "Database correlation for positioning of mobile terminals in cellular networks using wave propagation models," in *Proceedings of the Conference on Vehicular Technology*, vol. 7, September 2004, pp. 4682 – 4686.

[11] A. Arya, P. Godlewski, Marine Campedel, Ghislain du Chéné, "Radio database compression for accurate energy-efficient localization in fingerprinting systems," *To be appeared in IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 2011.

[12] K. Kaemarungsi, P. Krishnamurthy, "Modeling of indoor positioning systems based on location fingerprinting," in *Proceedings of the twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 2, March 2004, pp. 1012–1022.

[13] N. Swangmuang and P. Krishnamurthy, "Location fingerprint analyses toward efficient indoor positioning," in *Proceedings of the IEEE International Conference on Pervasive Computing and Communications*, March 2008, pp. 100 – 109.

[14] M. Khalaf-Allah, K. Kyamakya, "Database correlation using bayes filter for mobile terminal localization in GSM suburban environments," in *Proceedings of the Conference on Vehicular Technology*, vol. 2, May 2006, pp. 798–802.

[15] M. A. Youssef, A. Agrawala, A. Udaya Shankar, "WLAN location determination via clustering and probability distributions," in *Proceedings of the Conference on Pervasive Computing and Communications*, March 2003, pp. 143– 150.

[16] S. Fang, T. Lin, P. Lin, "Location fingerprinting in a decorrelated space," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 20, no. 5, pp. 685 – 691, May 2008.

[17] G. Wei and M. Tanner, "A Monte Carlo implementation of the EM algorithm and the Poor Man's Data Augmentation algorithms," *Journal of the American Statistical Association*, vol. 85, no. 411, pp. 699–704, 1990.

[18] R. Levine and G. Casella, "Implementations of the Monte Carlo EM Algorithm," *Journal of Computational and Graphical Statistics*, vol. 10, no. 3, pp. 422–439, 2001.

[19] G. Seber and A. Lee, *Linear regression analysis*. Wiley, 2003.

[20] O. Cappé and E. Moulines and T. Ryden, *Inference in Hidden MArkov Models*. Springer, 2005.

# Chapter 7

# Conclusions and perspectives

In regards to the emerging interest for Location Based Services (LBS), this thesis was initiated with the goal of providing "low-cost" and "continuous" LBS to the end users. The *Location Fingerprinting* (LFP) method has been adopted as the main axis of our studies, where it is investigated in a *machine learning* perspective.

## Database compression by using cluster analysis

As the first major contribution of this dissertation, we tackled the problem of radio database compression in LFP systems, by reducing the number of records. We proposed to perform the compression by applying a "clustering" process during the training phase. At a first step, standard clustering algorithms such as k-means and the minimum variance-based hierarchical method where examined.

The algorithms are applied in a concatenated "location-radio" signal space. In other words, the input data points for clustering algorithms consist of "location" parts and "radio" parts. In a basic scenario, the radio parts concern RSS measurements. In more elaborated cases, besides RSS, some other parameters (like time related measurements) are also included in the radio parts. The radio parameters could be measured from various radio access technologies (heterogeneous networks). However, one concern here to develop the cluster analysis is to apply an appropriate distance metric, since the data points consist of components belonging to different natures. In this work, the Euclidian distance has been adopted since it is a common choice in many studies of the literature. Precisely, we adopted a generalized form of Euclidian distance with heuristic fixed weights, which allows to attribute different weight factors to location and radio parts in the concatenated vectors.

It is noteworthy that for any adopted distance metrics, the RSS vectors may be expressed in dBm, or in their natural unit (milliwatt). Everywhere in this work, we adopted the dB format, since it showed a better performance during some preliminary tests. The dB representation is the format adopted in most of works in the fingerprinting literature.

The performance of the standard clustering algorithms based on the Euclidian distance, was examined by computer simulations. Two different environments were examined: a first area with large mask objects (average masks length about 500 m), and a second area with small mask objects (average masks length about 100 m). The equivalent propagation exponent for both environments was equal to 3.8 (typical value in urban environments). In both simulated scenarios, we observed a superior positioning performance of clustering techniques with respect to the naive griding method. The average positioning error obtained by clustering algorithms was stabilized at about 50-60 m for the first area, and about 90-100 m for the second area. The cluster analysis was more effective in the area with large mask objects than it was in the area with small mask objects, since the radio propagation was less homogeneous in the former case. In fact, the more the masks in the area are important, the more one moves away from a homogenous propagation situation; this leads to a better performance of clustering algorithms.

Once the standard clustering methods were examined, as a next step, a clustering algorithm well-tailored to the structure of the radio database was proposed. The main idea was to consider the weight factors of the distance metric as variables that would be optimized during the clustering process. We defined the concept of *feature types* in association with the records; a feature type is defined as all the stored parameters in a record that belong to the same nature. Based on this attribution, a *Block-based Weighted Clustering* (BWC) scheme was proposed; this scheme imposes equal weight factors to blocks of components belonging to the same feature type. The weight factors associated to feature types are optimized during the clustering process. This might be considered as providing a refined distance metric, adapted to the specific structure of records in the database. It is noteworthy that in this work, minimization of the total (or average) positioning error is not taken into account as an independent criterion for the optimization process. Enhanced clustering methods may be envisaged by considering this criterion.

Computer simulations and real experiments were conducted to evaluate the performance of the proposed BWC technique in the context of a cellular fingerprinting system. The results of both simulated and real experiments show that the proposed BWC technique outperforms the standard clustering algorithms and also other compression methods, like PCA and KCCA, taking the positioning accuracy as the performance criterion. In the simulated experiments all the clustering techniques provided a performance superior to that of the naive griding method. Although this superiority was incontestable in the simulated

scenario, in the real experiments the griding method showed a performance competitive enough, and it outperformed the standard clustering methods. This phenomenon needs further investigations, to interpret the reasons of the degraded performance of the cluster analysis, in the context of the real experiments. One possible reasoning is that the real measurements are picked up on close points (with a distance of about 10 m), and along pavements on street level. One may say that the measurements are performed over some homogenous sub-areas, and hence the clustering techniques are less effective in this context.

The cluster analysis in this work was proposed with the goal of reducing the computation and transmission loads, in order to decrease the terminal power consumption in mobile-based LFP systems. A complexity analysis was performed to evaluate the computation and transmission loads issued from clustering techniques, and to provide a comparison with other compression methods in the literature, such as Principle Component Analysis (PCA) and Kernel Canonical Correlation Analysis (KCCA). Based on the performed analysis, the clustering techniques outperform the other compression methods in the complexity viewpoint.

## Missing data handling procedures

In the next part of the thesis, we tackled the problem of *missing data* in the RSS-based fingerprinting systems. A specific missing mechanism was proposed to describe the missingness occurring in RSS measurements, issued from the 3GPP-defined scanning process (as in 2G and 3G). Our modeled missing mechanism proceeds based on two parameters: the receiver minimum sensitivity for signal detection $\lambda$, and the maximum number of base stations to be measured in the radio measurements $B_{max}$. Next, statistical methods were developed at two different levels, to deal with missing data. At the first level, the missing mechanism was assumed to be present exclusively during the localization phase; a maximum likelihood-based positioning procedure was then proposed, which takes into account the missing mechanism explicitly (Full ML method). At the second level of modeling, the missing mechanism was assumed to be present during both training and localization phases. Here a Multiple Imputation (MI) data completion algorithm was developed specific to the training phase. Then, handling the missing data in localization phase would lead to the first level of the modeling.

The efficiency of the proposed statistical methodology was examined by computer simulations, in the context of a GSM fingerprinting system. To be consistent with 3GPP standards, the receiver minimum sensitivity and the maximum number of base stations to be measured were set at -110 dBm and 7, respectively. Concerning the first level of the problem, the positioning performance of the proposed full ML method was compared

with the Fixed Imputation and the Observed ML methods, where the Full ML technique notably outperformed the others. In our simulated scenario, the improvement in average positioning error was about 15% w.r.t the Observed ML method, and varied between 15% and 45% w.r.t the Fixed Imputation method.

In the second level of the problem, the missing mechanism was supposed to be present during both training and localization phases. According to simulation results, for low transmission powers, the proposed MI-Full ML did not bring further improvement w.r.t the Fixed Imputation technique. But by increasing the transmission power, the MI-Full ML method notably outperformed the other technique. The average positioning error provided by the Fixed Imputation technique improved up to 30 % by applying the MI-Full ML method. The reason of this observation is that for a higher transmission power, a higher percentage of missing data is due to $B_{max}$. The information of this portion of missing data may be retrieved by using the proposed treating methods. On the other hand, the other portion of missing data due to $\lambda$ is not retrievable. Therefore, the performance of proposed method improves as the transmission power increases.

We notice that the sampling step of the proposed MI method in this work was implemented by using a non-optimal accept-reject method. This sampling could possibly be performed by means of Markov chain Monte Carlo (MCMC) techniques or by using sequential importance sampling methods. However, the very high complexity of the distribution of the unobserved variables renders the design of such sampling schemes very complicated. As a further discussion, we note that the proposed method in this work can deal with the missing data which are generated according to the missing mechanism defined in section 6.3.2. Other types of missingness (e.g. issued from temporary switch-off of base stations) can not be treated by the proposed framework. In future works, one may envisage an extended modeling for the missing mechanism, such that it includes more missingness factors.

# Glossary

| | |
|---|---|
| **3GPP** | 3rd Generation Partnership Project |
| **ANN** | Artificial Neural Networks |
| **AOA** | Angle of Arrival |
| **BCCH** | Broadcast Control Channel |
| **BWC** | Block-based Weighted Clustering |
| **CDMA** | Code Division Multiple Access |
| **CGI** | Cell Global Identity |
| **CPICH** | Common Pilot Channel |
| **CRLB** | Cramér-Rao lower bound |
| **E-OTD** | Enhanced Observed Time Difference |
| **FCC** | Federal Communications Commission |
| **GIS** | Geographical Information System |
| **GNSS** | Global Navigation Satellite Systems |
| **GPS** | Global Positioning System |
| **GSM** | Global System for Mobile Communications |
| **ICA** | Independent Component Analysis |
| **IPDL** | Idle Period DownLink |
| **KCCA** | Kernel Canonical Correlation Analysis |
| **KNN** | K-Nearest Neighbors |
| **LA** | Location Area |
| **LAN** | Local Area Network |
| **LBS** | Location Based Services |
| **LCS** | Location Service |
| **LFP** | Location Fingerprinting |
| **LMU** | Location Measurement Units |
| **LTE** | Long Term Evolution |
| **MAP** | maximum a posteriori probability |

**MCEM** Monte Carlo EM algorithm
**MCMC** Markov chain Monte Carlo
**MI** Multiple Imputation
**MIMO** Multiple Inputs Multiple outputs
**ML** Maximum Likelihood
**NLoS** Non Line of Sight
**OMA** Open Mobile Alliance
**OTD** Observed Time Difference
**OTDoA** Observed Time Difference of Arrival
**PCA** Principal Component Analysis
**RFID** Radio Frequency Identification
**RIT** Radio Interface Timing
**RRM** Radio Resource Management
**RSCP** Received Signal Code Power
**RSS** Received Signal Strength
**RTD** Real Time Difference
**RTT** Round Trip Time
**SMLC** Serving Mobile Location Center
**SNR** Signal to Noise Ratio
**SVM** Support Vector Machines
**TA** Timing Advance
**TDoA** Time Difference of arrival
**ToA** Time of Arrival
**TTFF** Time To First Fix
**U-TDOA** Uplink Time Difference Of Arrival
**UMTS** Universal Mobile Telecommunications System
**UWB** Ultra Wide Band
**WLAN** Wireless Local Area Networks

# List of figures

# List of tables