Thèse de statistiques dans une PME : La localisation intra-muros WiFi

Thierry Dumont [1]

[1] *Bureau E 20 Bât.G, Université Paris Ouest, 200 Av. République Nanterre France*

Résumé. Cet exposé revient sur la collaboration entre une PME et l'université Paris-sud autour d'une thèse CIFRE. Nous décrivons et analysons la méthode mise en œuvre pour répondre au besoin de localisation à l'intérieur des bâtiments. Cette méthode a permis le développement d'un logiciel de positionnement de terminaux WiFi dont la mise en place est simplifiée par rapport aux autres systèmes existant sur le marché.

Mots-clés. Localisation, Modèles de Markov cachés, Inférence, algorithme EM séquentiel.

Abstract. This presentation retraces the collaboration between a medium sized compagny and Paris-sud university through a CIFRE PhD contract. We describe and analyse the method developped to answer the indoor localization problem. This method allowed the developpement of a WiFi devices positioning software which, unlike other systems existing on the market possesses a simplified installation procedure.

Keywords. Localization, Hidden Markov models, Inference, online EM algorithm

1 Introduction

L'entreprise ID Services basée à Orsay (91) est une filiale du groupe européen Autotech-ID créée en 2002 qui compte une trentaine de salariés en France. L'entreprise est spécialiste de l'identification automatique dans la logistique, l'industrie ou la grande distribution. Elle propose à ses clients des solutions matérielles, informatiques et techniques pour le suivi des marchandises à l'aide notamment de terminaux mobiles connectés. Une des compétences de l'entreprise est la mise en place et la maintenance d'architectures WiFi offrant une connectivité optimale pour le parc de terminaux mobiles de ses clients.

Fort de son expérience dans ce domaine, la direction d'ID Services entreprend en 2008 de proposer à ses clients un outil de localisation en temps réel de leur matériel informatique exploitant ces signaux WiFi. Une telle technologie pouvant permettre le développement de solutions innovantes en matière de traçabilité et d'identification automatique. Un premier prototype est développé par une équipe d'ingénieurs et de développeurs permettant de confirmer la pertinence de l'approche envisagée. Néanmoins la précision atteinte par le prototype et son mode de fonctionnement complexe ne permettent pas de commencer le processus d'industrialisation. Un travail de fond est nécessaire pour comprendre les

phénomènes aléatoires impactant l'efficacité du système et améliorer l'algorithme de localisation. Débute alors un partenariat entre ID Services et la faculté de mathématiques d'Orsay concrétisé par une thèse CIFRE qui durera de 2009 à 2012.

2 Présentation du problème et première analyse

2.1 Modélisation

Dans le domaine des télécommunications la puissance d'un signal radio est donnée par le RSS (*Receive signal strength* exprimée en *dBm*). Cette information sur la puissance des signaux est utilisée par les terminaux connectés en WiFi pour s'appairer au point d'accès dont le signal possède le meilleur ratio signal/bruit. C'est cette information sur la puissance de l'ensemble des signaux reçus par un terminal que l'on souhaite exploiter pour le localiser. Cette donnée fluctue beaucoup même lorsque le terminal ne bouge pas rendant la localisation difficile. Une modélisation Gaussienne de ces fluctuations est communément utilisée. Malgré ces fluctuations le RSS moyen dépend fortement de la position du terminal par rapport au point d'accès. On appellera la fonction associant à chaque position de l'espace le RSS moyen d'un signal provenant d'un point d'accès donné la *carte de propagation associée au point d'accès*. Une telle carte de propagation construite à l'aide de mesures prises dans les locaux de l'entreprise est représentée sur la figure 1.

Le terminal mobile (porté par une personne) se déplace à l'intérieur d'un bâtiment et mesure à intervalles de temps réguliers la puissance des signaux WiFi qui l'entourent. Notons $\{X_t\}_{t\in\mathbb{N}}$ la suite de positions (non observées) prises par le mobile à chaque relevé de puissances. X_t est donc une position sur une carte $\mathbb{X}$ représentant le bâtiment. Nous supposerons $\mathbb{X} \subset \mathbb{R}^2$. Y_t désigne les puissances mesurées à l'instant t par le terminal. $Y_t \in \mathbb{R}^\ell$ où ℓ représente le nombre total de points d'accès dans le bâtiment. Une modélisation du processus bivarié $\{(X_t, Y_t)\}_{t\geq 0}$ par modèles de de Markov cachés (HMM pour *hidden Markov models*) est particulièrement bien adapté dans notre cas. L'équation (1) précise le modèle HMM considéré :

$$\begin{aligned} X_t|X_{t-1} = x &\sim Q(x, \mathrm{d}x') = q(x, x')\mathrm{d}x' \\ Y_t &= F^\star(X_t) + \epsilon_t \ , \end{aligned} \tag{1}$$

avec $q(x, x') \propto \mathbb{1}_{\mathbb{X}}(x') \exp\left(-\frac{||x-x'||^2}{a}\right)$ a étant un paramètre (supposé connu) dépendant de la vitesse moyenne du mobile, $F^\star : \mathbb{X} \to \mathbb{R}^\ell$ où pour $j = 1, \dots \ell$, $F_j^\star$ est la carte de propagation du signal associée au point d'accès WiFi j et $\{\epsilon_t\}_{t\geq 0}$ est un bruit Gaussien de densité $\varphi_{\sigma^\star,2}(z) \propto e^{-||z||^2/2\sigma^{\star,2}}$.

2.2 Loi *a priori* sur $F^\star$

En champ libre (sans obstacles), Friis [5] établi une relation entre la puissance $\mu(x)$ (en dBm) d'un signal radio reçu en une position x et la distance séparant x à l'émetteur du signal O :

$$\mu(x) = c + d\log(||x - O||) ,$$

où c et d dépendent de la fréquence du signal radio, de sa puissance d'émission et des gains des antennes de réception et d'émission. Cependant, comme illustré sur la figure 1, à l'intérieur des bâtiments la propagation du signal est perturbée par les nombreux obstacles présents et n'est donc isotrope comme en champ libre. Nous modéliserons ces perturbations par un champ gaussien noté δ de sorte que pour $j = 1, \ldots, \ell$ et $x \in \mathbb{X}$,

$$F_j^\star(x) = c_j^\star + d_j^\star \log(||x - O_j||) + \delta_j^\star(x) .$$

où O_j est la position dans $\mathbb{X}$ du point d'accès j.

2.3 Filtrage

Les informations de puissances $Y_0, \ldots, Y_n, \ldots$ sont donc générées de manière séquentielle par le terminal. À chaque nouvelle mesure Y_n nous souhaitons construire un estimateur $\widehat{X}_n$ de la position X_n qu'occupait le terminal lors de cette prise de mesure. En conséquence nous souhaitons exploiter la mesure dite de filtrage : $\phi_{n|n}$ qui est la distribution de X_n conditionnelle aux observations passées : $Y_0, \ldots, Y_n$. Notre HMM étant complètement dominé, $\phi_{n|n}$ est à densité par rapport à la mesure de Lebesgue sur $\mathbb{X}$. $\widehat{X}_n$ peut être construit sur la base de cette mesure de filtrage comme étant, par exemple, le prédicteur de maximum *a posteriori* : $\widehat{X}_n := \text{argmax}_{x\in\mathbb{X}}\phi_{n|n}(x)$ ou la moyenne *a posteriori* : $\widehat{X}_n := \int_{x\in\mathbb{X}} x\phi_{n|n}(x)\mathrm{d}x$. Les mesures de filtrage peuvent être construites récursivement :

$$\phi_{n|n}(x') \propto \int_{x\in\mathbb{X}} \phi_{n-1|n-1}(x) q(x, x') \varphi_{\sigma^\star,2}(F^\star(x') - Y_n)\mathrm{d}x ,$$

et approximées grâce à des méthodes d'échantillonage séquentielles (*c.f.* Chapitre 7 de [2] pour une description de ces méthodes de filtrage séquentielles).

Dès lors le défi consiste à estimer les quantités permettant le calcul (ou l'approximation) des mesures de filtrage à savoir $F^\star$ et $\sigma^{\star,2}$.

2.4 Algorithme EM

L'algorithme EM (*Expectation Maximisation*) est un algorithme itératif très répandu pour approcher l'estimateur du maximum de vraisemblance $\widehat{\theta} = \text{argmax}_{\theta\in\Theta} L_\theta(Y_{0:n})$ où

$$L_\theta(Y_{0:n}) = \int_{x_{0:n}} p_\theta(x_{0:n}, Y_{0:n})\mathrm{d}_{x_{0:n}}$$

dans les modèles de Markov cachés paramétriques. Notons

$$Q(\theta;\theta') = \mathbb{E}_{\theta'}\left[\log p_\theta(X_{0:n}, Y_{0:n})|Y_{0:n}\right]$$

Partant d'une valeur initiale $\theta^{(0)}$ l'algorithme EM construit récursivement une suite d'estimateurs $\{\theta^{i)}\}_{i\geq 0}$ en posant pour tout $i \geq 1$, $\theta^{(i)} = \text{argmax}_\theta\ Q(\theta;\theta^{(i-1)})$. Ainsi construit la vraisemblance des observations augmente à chaque itération de l'algorithme pour converger vers un maximum local.

3 EM par blocs et prise en compte de la loi a priori

L'EM classique présenté dans la section 2.4 nécessite de parcourir l'ensemble des données pour construire la fonction $Q(\cdot,\theta^{(i-1)})$ à chaque itération de l'algorithme. Dans [4] nous utilisons une version séquentielle de l'EM dans l'esprit de [1] permettant d'actualiser l'estimation de nos paramètres à mesure que les données de puissances sont produites par le terminal. Nous faisons de plus intervenir la loi *a priori* sur $F^\star$ présentée dans la section 2.2. Avant toute chose nous procédons à une discrétisons de l'environnement $\mathbb{X}$ pour rendre notre modèle paramétrique : considérons $\mathcal{G} \subset \mathbb{X}$ un ensemble fini de points de $\mathbb{X}$ (une grille fine). Nous cherchons à estimer $\theta^\star = \left(\{c_j^\star, d_j^\star, \{\delta_j^\star(x)\}_{x\in\mathcal{G}}\}_{j=1}^\ell, \sigma^2\right)$ sur la base des observations $\{Y_t\}_{t\geq 0}$. Notons π la loi a priori imposée sur le processus $\delta^\star|_\mathcal{G} = \{\{\delta_j^\star(x)\}_{x\in\mathcal{G}}\}_{j=1}^\ell$:

$$\pi(\delta) \propto \exp\left(-\frac{1}{2}\sum_{j=1}^{\ell}\delta_j^T\Sigma^{-1}\delta_j\right) ,$$

où Σ est la matrice de covariance commune (supposée connue ici) des processus gaussiens $\delta_j^\star$ restreints à $\mathcal{G}$. La vraisemblance des observations $y_{0:n}$ sous $\theta = (c, d, \delta, \sigma^2)$ est donnée par $L_\theta(y_{0:n}) = \sum_{x_{0:n}\in\mathcal{G}^{n+1}} \nu(x_0)\varphi_{\sigma^2}(F_\theta(x_0) - y_0)\prod_{t=1}^n q(x_{t-1}, x_t)\varphi_{\sigma^2}(F_\theta(x_t) - y_t)$, où ν est une distribution initiale pour X_0 et $F_{\theta,j}(x) = c_j + d_j\log(||x - O_j||) + \delta_j(x)$. Dans [4] nous montrons que la quantité $Q(\theta;\theta')$ peut se mettre sous la forme d'un produit scalaire :

$$Q(\theta;\theta') = < S_{\theta',0:n}, f(\theta) > ,$$

où S est de la forme $S_{\theta',0:n} = \frac{1}{n+1}\sum_{t=0}^n \mathbb{E}'_\theta(s(X_t, Y_t)|Y_{0:n})$. [4] propose d'exploiter le caractère additif de $S_{\theta',0:n}$ pour construite une version séquentielle par bloc de l'EM : considérant une valeur initiale $\theta^{(0)} = (c^{(0)}, d^{(0)}, \overrightarrow{0}, \sigma^{2,(0)})$ ($\delta^{(0)} = \overrightarrow{0}$ correspondant à une propagation isotrope du signal) et une suite croissante $0 = T_0 < T_1 < \cdots$ d'indices temporels, $\theta^{(i+1)}$ est défini comme maximisant de la fonction :

$$\theta \mapsto < S_{i+1}, f(\theta) > + \frac{1}{T_{i+1} - T_i}\log(\pi(\delta)) ,$$

où $S_{i+1} := S_{\theta^i,T_i:T_{i+1}-1} = \frac{1}{T_{i+1}-T_i}\sum_{t=T_i}^{T_{i+1}-1}\mathbb{E}_{\theta^{(i)}}(s(X_t, Y_t)|Y_{T_i:T_{i+1}-1})$. Ainsi définis $\theta^{(i+1)}$ dépend de l'estimateur précédent $\theta^{(i)}$ et des mesures générées dans le $i+1$-ème bloc de mesures : $Y_{T_i}, \cdots, Y_{T_{i+1}-1}$. Parallèlement, une version moyennée de ces estimateurs $\{\widehat{\theta^{(i)}}\}_{i\geq 1}$

est construite sur la base des statistiques moyennées : $\widehat{S}_i = \frac{1}{T_i}\sum_{k=1}^{i}(T_k - T_{k-1})S_k$. Finalement, une étape dite de "stabilisation" est introduite dans [4] pour contenir le phénomène d'instabilité apparaissant sur l'estimateur non moyenné : régulièrement (tous les K blocs de mesures), $\widehat{\theta}^{(i)}$ est substitué à $\theta^{(i)}$ de sorte que $\theta^{(i+1)}$ dépend de $\widehat{\theta}^{(i)}$ et des mesures du $i+1$-ème bloc de mesures. Des résultats de simulations sont présentés dans la figure 2. La qualité de l'estimation est mesurée par la précision en localisation offerte par les différents estimateurs. La figure 2(a) indique la précision obtenue par le prédicteur du maximum *a posteriori* de la section 2.3 en utilisant pour chaque bloc de mesures i l'estimateur par bloc $\theta^{(i)}$ et l'estimateur moyenné $\widehat{\theta}^{(i)}$. Ces résultats sont comparés à la précision que l'on qualifiera d'optimale obtenue en appliquant notre algorithme de filtrage avec le vrai paramètre $\theta^\star$. Il apparaît clairement que nos deux estimateurs font diverger notre précision en localisation mais que la précision obtenue sous l'estimateur moyenné est beaucoup plus stable. Ce phénomène a pour cause le fait que notre processus caché $\{X_t\}_{t\geq 0}$ est une marche aléatoire. Si dans un bloc de mesures donné i, X_t n'explore qu'une petite partie de la carte, l'estimateur $\theta^{(i)}$ sera impacté par la non représentativité de ces mesures. Cette mauvaise estimation impacte alors le calcul de l'estimateur suivant $\theta^{(i+1)}$ et démarre ainsi un réaction en chaîne. Les résultats de la méthode de substitution sont illustrés sur la figure 2(b). La substitution est effectuée tous les 5 blocs de mesures et semble enrayer cette réaction en chaîne. La précision offerte par l'estimateur moyenné semble converger vers la précision optimale. Nous avons soumis cette méthode d'estimation séquentielle avec substitution à des données réelles (figure 3). La précision est calculée grâce à quatre phases de mesures durant lesquelles la vraie position du terminal est connue et comparée à la position prédite par l'algorithme avec l'estimateur moyenné. L'expérience 1 démarre la localisation avec un paramètre initial $\theta^{(0)}$ calibré grâce à une campagne de mesures effectuée préalablement offrant une bonne précision initiale (quantile à 80% $\sim$ 5 mètres). Dans l'expérience 2, $\theta^{(0)}$ est choisi pour offrir une précision initiale très mauvaise (quantile à 80% $\sim$ 15 mètres). Pour ces deux expériences la précision en localisation s'améliore au cours du temps et, bien que les paramètres initiaux utilisés soient très différents, la précision obtenue après 20000 mesures est comparable pour les deux expériences (quantile à 80% $\sim$ 4 mètres).

4 Conclusion

Suite à cette étude nous avons développé un système de localisation autonome dont la mise en place ne nécessite que la connaissance de la position des points d'accès. Cette étude à aussi été le point de départ d'une analyse plus théorique autour du modèle décrit par (1). [3] étudie l'identifiabilité de tels modèles et une méthode d'estimation non paramétrique qui, contrairement à la méthode utilisée ici ne repose sur aucune discrétisation ou distribution *a priori* sur $F^\star$.

Références

[1] O. Cappé. Online EM algorithm for Hidden Markov Models. *J. Comput. Graph. Statist.*, 20(3) :728–749, 2011.

[2] O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer, 2005.

[3] T. Dumont and S. Le Corff. Nonparametric regression on hidden phi-mixing variables : identifiability and consistency of a pseudo-likelihood based estimation procedure. *arXiv preprint arXiv :1209.0633*, 2012.

[4] T. Dumont and S. Le Corff. Simultaneous localization and mapping in wireless sensor networks. *Signal Processing*, 101(0) :192 – 203, 2014.

[5] H. T. Friis. A note on a simple transmission formula. *Proceedings of the IRE*, 34(5) :254–256, 1946.

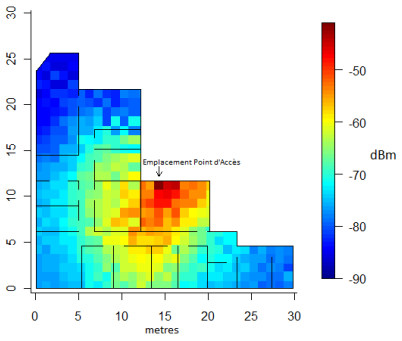
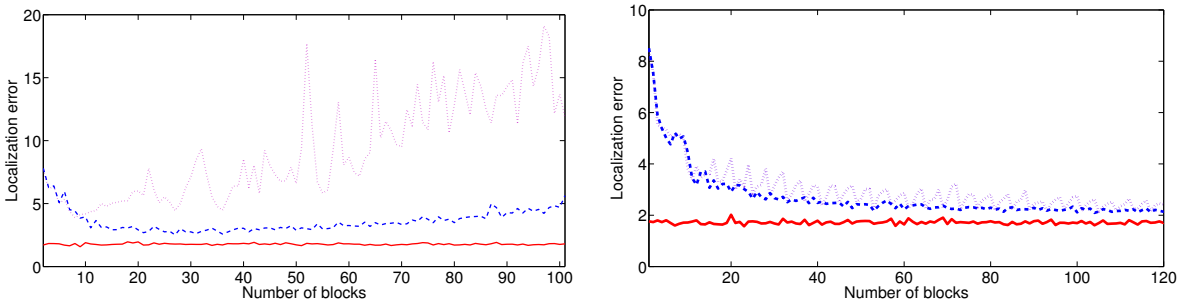


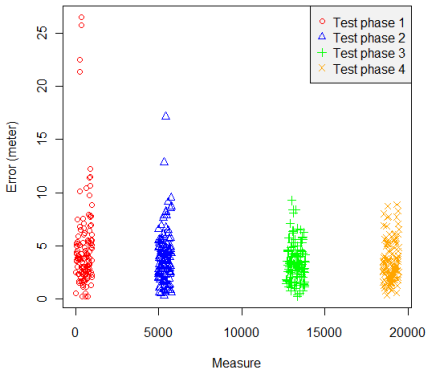
FIGURE 1 – Carte de propagation



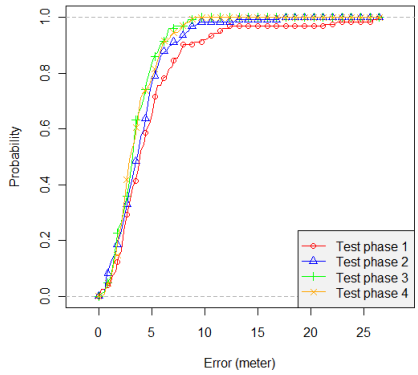
(a) méthode sans substitution

(b) méthode avec substitution

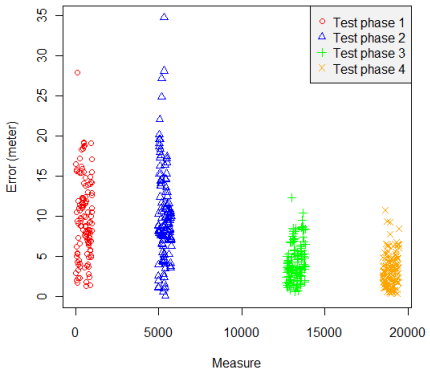
FIGURE 2 – Simulation : Quantiles à 80% de l'erreur en localisation (en mètres) calculé sur chaque bloc de mesure en utilisant les paramètres $\theta^\star$ (traits pleins), $\theta^{(k)}$ (pointillés) et $\widehat{\theta}^{(k)}$ (tirés)



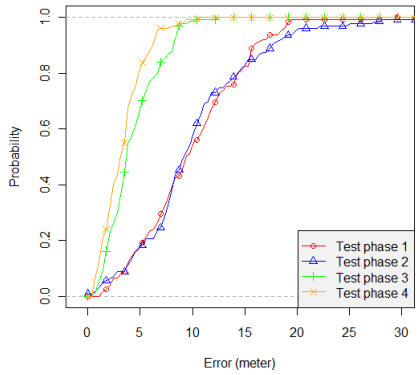
(a) Expérience 1 : Erreurs en localisation sur les quatre phases de test.



(b) Expérience 1 : Répartition des erreurs réparties par phase de test.



(c) Expérience 2 : Erreurs en localisation sur les quatre phases de test.



(d) Expérience 2 : Répartition des erreurs réparties par phase de test.

FIGURE 3 – Données réelles