Unsupervised Labelling of Sequential Data for Location Identification in Indoor Environments

Iker Perez[a,*], James Pinchin[a], Michael Brown[a], Jesse Blum[a], Sarah Sharples[b]

[a] *Horizon Digital Economy Research Institute, The University of Nottingham, UK*
[b] *Faculty of Engineering, The University of Nottingham, UK*

Abstract

In this paper we present indoor positioning within unknown environments as an unsupervised labelling task on sequential data. We explore a probabilistic framework relying on wireless network radio signals and contextual information, which is increasingly available in large environments. Thus, we form an informative spatial classifier without resorting to a pre-determined map, and show the potential of the approach using both simulated and real data sets.

Results demonstrate the ability of the procedure to segregate structures of radio signal observations and form clustered regions in association to areas of interest to the user; thus, we show it is possible to differentiate location between closely spaced zones of variable size and shape.

Keywords: Unsupervised Labelling, Sequential Data, Indoor Positioning, Ubiquitous Computing, Graphical Models

1. Introduction

The practical determination of a pedestrian's position indoors remains an open problem; signals from satellite based positioning systems such as the *Global Positioning System* (GPS) are attenuated and delayed by building structures, to the extent that they are undetectable or yield very poor position estimates. A variety of alternative technologies have been developed; largely turning to *signals of opportunity* such as WiFi or Bluetooth (e.g. Mok & Retscher (2007); Yang et al. (2015)), or relying on inertial technologies and detailed building information (e.g. Pinchin et al. (2012)). An overview on the topic and available options for the deployment of *indoor positioning systems* can be found in Harle (2013) and Gu et al. (2009).

[*] Corresponding author. Iker Perez, Nottingham Geospatial Building, Jubilee Campus, Wollaton Road, Nottingham NG8 1BB, UK. +44 (0) 754 728 4617.

Email addresses: iker.perez@nottingham.ac.uk (Iker Perez), james.pinchin@nottingham.ac.uk (James Pinchin), michael.brown@nottingham.ac.uk (Michael Brown), jesseblum@gmail.com (Jesse Blum), sarah.sharples@nottingham.ac.uk (Sarah Sharples)

Positioning methodologies associated to the use of signals of opportunity commonly resort to supervised learning approaches; these dissociate statistical structures of signal observations from various *fingerprinted* locations within a training set, and then construct a predictive model mapping observed signals to vector locations (Madigan et al. (2005)); however, the collection of enough data for useful applications is labour intensive (cf. Chen et al. (2006)). Within indoor environments, the need for extensive dedicated hardware and repeated physical surveys of the space also deters real-world applications; buildings may not necessarily be accessible and can be subject to constant alterations on their layout. The requirement for constant map maintenances usually yields implemented solutions unusable over short time-periods (Bolliger et al. (2009)); thus, there is a need to study alternative approaches that allow drawing meaningful inference regarding the whereabouts of a device (cf. Sharples et al. (2015)), while ignoring the internal arrangement of infrastructure.

In this paper, we address the problem of retrospectively labelling sequences of locations of a mobile device, within some unknown indoor environment. Labels must be assigned in an spatially informative manner; and in this particular context, we are restricted to the use of observations on WiFi access points (APs) and their received signal strength indicators (RSSIs), along with meta-data linked to the activities undertaken by device carriers. In standard wireless networks, especially within large indoor environments, there commonly are multiple APs serving end-users, uniquely identifiable through their MAC address; such signals can be received by hardware in many mobile devices and there exists literature on their use for (mostly supervised) location estimation (see Roos et al. (2002); Youssef et al. (2003); Ladd et al. (2004) and references therein). In addition, the rise of ubiquitous computing permits the feed of substantial contextual information for use within an unsupervised learning scheme.

A motivational example to this work is given by the task of locating medical doctors within a hospital, while lacking any knowledge regarding the internal arrangement of the buildings or the placement of APs. We note that hospitals, as well as universities, airports or museums, are generally vast and complex spaces encompassing several buildings. While it is possible to fingerprint the environment in order to infer precise room-level positioning, information on doctors' duties over their shifts can be of sufficient discriminative power, in order to segregate structures of observations of WiFi signals over differing locations. Thus, this can enable the construction of an informative spatial classifier that segregates physical points over clusters associated to specific labels.

To address this issue, we divert from traditional coordinate-system positioning and resort to an unsupervised labelling approach on sequential data. Here, labels respond to unknown physical locations, identifiable from meta-data information; for instance, hospital wards, university faculties or museum galleries. We propose a scalable probabilistic model formalizing a joint distribution over signal observations and position-label sequences, and reduce the problem to an inference and learning task on a *dynamic Bayesian Network* (dBN). The method relies on an intuitive *Input-Output* network where transitions are assumed under the influence of activity information and thus defined through a *maximum-entropy classifier*. Hence, we resort to belief-propagation and numerical optimization routines to learn device location-label sequences, AP observation probabilities and density of RSSIs.

Within an unsupervised learning scheme and to the best of our knowledge, the proposed approach first allows to draw meaningful inference on sequential indoor location, while overcoming the need for inertial measurements on a device along with assumptions regarding physical models for transition dynamics (cf. Shin et al. (2012); Wang et al. (2012)). In addition, the underlying learning task is tractable without resorting to variational or simulation-based approximate inference methods. Its practical benefits are varied; on one hand, indoors spatial classifications do not rely on a pre-determined map and can respond to environmental or AP placement alterations. Also, physical clusters associated to labels may be influenced through outdoors structural observations of buildings gained from satellite images, or from a conceptual understanding of meta-data. Tests on simulated data and a small-scale study within a major university hospital in the United Kingdom show that, given a sufficiently rich data set regarding the activity of mobile device carriers, it is not only possible to dissociate positioning within large and physically separated building-blocks, but also between areas of interest relatively close to each other and of significantly differing forms.

The rest of the paper is organized as follows. In Section 2 we review recent work on indoor positioning and probabilistic approaches to the matter, and we offer an overview of literature discussing Bayesian Networks and Graphical Models. In Section 3 we formulate the labelling problem, while Section 4 discusses parameter learning within the model, allowing the formation of a positioning classifier. Section 5 presents results on simulated data and a case study within a university hospital and, finally, Section 6 offers a discussion on the advantages, key contributions and possible extensions of the work.

2. Related Work

A probabilistic indoor positioning approach commonly requires mapping WiFi signal strength distributions to different locations; these distribution *fingerprints* can later be used for classification, inferring the most likely posterior location given a set of signal observations. These differentiated stages are usually referred to as *offline* and *online* phases and some examples of this work include Kaemarungsi & Krishnamurthy (2004); Brunato & Battiti (2005); Roos et al. (2002); Youssef et al. (2003); Yim (2008). Alternative approaches based on *triangulation* and *trilateration* (Bahl & Padmanabhan (2000); Krishnan et al. (2004); Peterson et al. (1998); Li et al. (2000)) study physical signal propagations. Their aim is to locate a device by measuring angles or distances to signal sources. Such an approach is usual in global positioning problems and it requires detailed maps along with physical changes over time. We refer the reader to Liu et al. (2007) for a technical review on existing wireless indoor positioning solutions.

Indoor positioning problems have seldom been posed as purely unsupervised tasks; working with labelled data is standard and the use of clustering or latent variable models has been limited to the task of improving computational scalability (cf. Swangmuang & Krishnamurthy (2008); Ma et al. (2008)). Yet, recent work in Shin et al. (2012); Wang et al. (2012); Pinchin et al. (2014) has begun studying varied unsupervised schemes aiming to bypass the need for fingerprinting. The premise in Shin et al. (2012); Wang et al. (2012) is to combine identifiable signatures present in common locations along with odometry tracings. On the other hand, work in Pinchin et al. (2014) integrated contextual information regarding reception devices within a positioning classifier, with aims of improving the behavioural understanding of the device carriers.

Given the inherent sequential structure in the positioning of a reception device, *probabilistic graphical models* provide a suitable framework for labelling purposes in this work. These models encode complete distributions over random variables using graph-based representations, and some special cases such as the state-space *Kalman Filter* have long found vast applications in global positioning systems (e.g. Krakiwsky et al. (1988); Gustafsson et al. (2002); Fung & Grimble (1983); Sasiadek et al. (2000)). Additionally, *Hidden Markov Models* (HMMs), *Maximum-Entropy Markov Models* (MEMMs) and *Conditional Random Fields* (CRFs) have also been studied for diverse supervised labelling problems with sequential data (Dietterich (2002); Lafferty et al. (2001); McCallum et al. (2000)), and recent work has used properties of HMMs and *Markov Random Fields* for developing indoor positioning systems (Seitz et al. (2010b,a); Hoang et al. (2013); Shen et al. (2012)). For an introduction on learning dBNs we refer the reader to Ghahramani (2001); Murphy (2002); also, general theory on graphical models can be found in Koller & Friedman (2009); Jordan (1998).

3. Formulation and Reduction of the Labelling Problem

Let $x_t \in \mathcal{X}$ denote the location-label of a mobile device at time $t \in \{1, \cdots, T\}$, for some $T > 1$. Furthermore, let $\boldsymbol{a}_t \in \mathcal{A}$ define a binary vector representing observed and unobserved access points within a wireless network, over a single scan of the device. In addition, denote by $\boldsymbol{r}_t \in \mathcal{R}$ the set of RSSIs (if any) during the scan. Here, $\mathcal{X}$ defines a family of labels associated with a set of unknown location-clusters $\{X_1, \cdots, X_n\}$, rather than a finite family of coordinates on a discretized Cartesian system (cf. Seitz et al. (2010b,a); Hoang et al. (2013)). Further, each cluster is a collection of coordinates within some undefined Euclidean Space. The set $\mathcal{A}$ is formed by the combinatorial space of m binary outputs, with cardinality $|A| = 2^m$, and

$$\mathcal{R} = \big\{(R_1, \cdots, R_m) \ : \ R_i \in \mathbb{R}^- \cup \{\emptyset\} \text{ for } i \in \{1, \cdots, m\}\big\}.$$

Note that observable signal strength measurements from WiFi networks are often given in units of *decibel-milliwatts* (dBm) and are strictly negative; physical models (Harley (1989); Hashemi (1993)) suggest a signal propagation according to a log-normal function within indoor environments. However, such an assumption breaks under the presence of walls or furniture (cf. Campos et al. (2014)) and different experiments have arrived at different distributions (Ladd et al. (2004)). A general model remains unavailable and discrete Gaussian approximations are common (cf. Seitz et al. (2010b,a)), sometimes incorporating a signal strength decay factor based on euclidean log-distance (Madigan et al. (2005)). In addition, given a large location-cluster in association to a label, its global signal-reception strength will be susceptible to noise from the changing distances to APs on the reception device (device carriers altering their position within the physical clustered area). Thus, the resulting observations in our work reflect weighted averaged distributions over the positioning of the devices within clusters in $\{X_1, \cdots, X_n\}$, and normal approximations perform well.

Signal observations are independent given the location-label of a reception device; also, in order to ease notation we assume reception capability is equivalent across devices. Hence, the conditional dependence structure between the above variables is such that

$$\begin{aligned}\mathbb{P}(\boldsymbol{a}_t, \boldsymbol{r}_t | x_t = x) &= \mathbb{P}(\boldsymbol{r}_t | \boldsymbol{a}_t, x_t = x)\mathbb{P}(\boldsymbol{a}_t | x_t = x) \\ &= \prod_{i=1}^{m} \Big[\mathbb{I}_{\{a_t(i)=0\}}(1 - p_{x,i}) + \mathbb{I}_{\{a_t(i)=1\}} \frac{p_{x,i}}{\sigma_{x,i}} \psi\Big(\frac{r_t(i) - \mu_{x,i}}{\sigma_{x,i}}\Big)\Big],\end{aligned} \tag{1}$$

where $\psi(\cdot)$ denotes the standard normal probability density function. Here, $p_{x,i}$ is the probability of observing AP i from cluster x over a single scan; $\mu_{x,i}$ and $\sigma_{x,i}$ represent the mean and standard deviation of its RSSI. Note than an extension allowing for varying reception strengths of devices is straightforward.

3.1. Meta-data and Graphical Model

In addition to radio signal scans, this paper assumes the presence of contextual information regarding the activities of device carriers. Such information is becoming increasingly accessible with the rise of ubiquitous computing, and can take the form of a log with tasks, duties or any indications suggesting a carrier *likely* visited a uniquely named location over a bounded time-span. Thus, it is possible to determine an amount of identifiable spatial clusters $\{X_1, \cdots, X_n\}$ in association to labels named after meta-data locations. In addition, it is possible to define a probability distribution over labels $\mathcal{X}$, for any time $t \in \{1, \cdots, T\}$.

Let $\boldsymbol{d}_t \in \mathcal{D}$ be a vector of *counts* associating activity-log data to the family $\{X_1, \cdots, X_n\}$ at time t, with $\mathcal{D} = \mathbb{N}_0 \times \overset{n}{\cdots} \times \mathbb{N}_0$; for instance, $\boldsymbol{d}_t$ could reflect active duties that a doctor has in association to each location of interest in a hospital. Then, the assigned label of a reception device forms a categorical distribution whose prior membership probabilities can be defined in proportion to counts in $\boldsymbol{d}_t$, so that

$$\mathbb{P}(x_t = x | \boldsymbol{d}_t = \boldsymbol{d}) \propto \delta^{\boldsymbol{d}(x)} \quad \text{for} \quad \delta > 1 \ . \tag{2}$$

We refer to δ as the *weighting* parameter; it reflects the relevance of counts in d. An extension allowing for combinations of weights is straightforward; however, the focus here is on a single value for simplicity. We notice that if $\boldsymbol{d} = 0_{1\times n}$

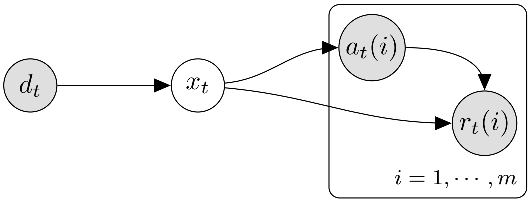


Figure 1: Graphical model for variables x_t, a_t, r_t and d_t, we note that x is a latent node.

then the underlying distribution is uniform. Figure 1 presents the relation within location-label, signal observation and count variables in plate notation.

Latent location-label nodes and signal observation structures can be inferred from the model in Figure 1 through iterative *expectation-maximization* (Moon (1996)). However, two radio scans will be relatively close in time and sequential patterns should be exploited to improve the prediction accuracy of a classifier. The complete dependence structure within variables is represented with a dBN as in Figure 2, a type of network in relation to *tree structured HMMs* or *hidden Markov decision trees* (Jordan et al. (1997); Ghahramani (2001)).

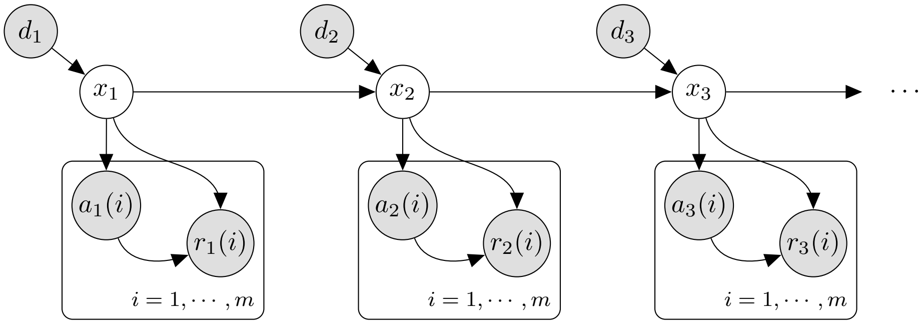


Figure 2: Dynamic Bayesian Network representing a partial structure of dependencies within activity data, latent location-labels and signal observation variables.

3.2. Exponential Transitions and Likelihood Function

In this network (Figure 2), we require label transitions to be under the influence of activity information. We define a maximum-entropy classifier such that

$$\mathbb{P}(x_t = x' | x_{t-1} = x, \boldsymbol{d}_t = \boldsymbol{d}) \propto \begin{cases} e^{\nu_x \boldsymbol{d}(x')} & \text{if } x' = x, \\ e^{\lambda_{x,x'} + \nu_x \boldsymbol{d}(x')} & \text{if } x' \neq x, \end{cases} \tag{3}$$

where counts in $\boldsymbol{d}$ are *features* and $\{\lambda_{x,x'}, \nu_x \ : \ x, x' \in \mathcal{X} \text{ and } x \neq x'\}$ are independent *weights* for each transition origin. Note that it is possible to extend the framework and incorporate additional information, such as odometry data or precise location information (for instance, card swipes or fingerprint scans). Additionally,

$$e^{\lambda_{x,x'}+\nu_x \boldsymbol{d}(x')} = e^{\lambda_{x,x'}} \delta_x^{\boldsymbol{d}(x')}$$

with $\delta_x = e^{\nu_x}$, so that λ values define an uninfluenced transition matrix for each origin x; and ν_x incorporates count data proportionally as in (2). The prior initial state distribution is assumed uninformative, so that $\mathbb{P}(x_1 = x' | \boldsymbol{d}_1 = \boldsymbol{d}) \propto 1$.

The log-likelihood function conditional on activity data is given by

$$\begin{aligned} l(\boldsymbol{a}_{1:T}, \boldsymbol{r}_{1:T}, x_{1:T} | \boldsymbol{d}_{1:T}) &= \log \mathbb{P}(x_1 | \boldsymbol{d}_1) \\ &\quad + \sum_{t=1}^{T} \log \mathbb{P}(\boldsymbol{a}_t, \boldsymbol{r}_t | x_t) + \sum_{t=2}^{T} \log \mathbb{P}(x_t | x_{t-1}, \boldsymbol{d}_t), \end{aligned}$$

with probability functions as described in (1) and (3).

4. Parameter and Location Estimation

Given the model structure, the learning task is reduced to estimating parameters and states. Here, a training procedure offering a good trade-off between computational cost and quality of fit requires an inner *forward-backward* loop. However, traditional Baum-Welch equations do not accommodate for conditional label transitions in a parameter update stage (cf. Ghahramani (2001)), and exact Bayesian inference through MCMC simulations is impractical in problems with high dimensionality and long data sets (Rydén et al. (2008)). Thus, we iterate between belief propagation, analytical Maximum Likelihood updates and *Limited-memory Broyden-Fletcher-Goldfarb-Shanno* (L-BFGS) in an expectation maximization routine, helping scalability and handling computational restrictions. We note that L-BFGS is, to date, the preferred algorithm for fitting maximum-entropy models and CRFs, and has largely surpassed generalized iterative scaling algorithms for use with MEMMs (Malouf (2002)).

Starting values $\{p_{x,i}^{(0)}, \mu_{x,i}^{(0)}, \sigma_{x,i}^{(0)} : x \in \mathcal{X}, i \in \{1, \cdots, m\}\}$ are averaged-out over the signal observation data set. Additionally,

$$\lambda_{x,x'}^{(0)} < 0 \quad \text{and} \quad \nu_x^{(0)} > 0 \quad \text{for all} \quad x, x' \in \mathcal{X} \text{ with } x \neq x'.$$

These strict inequalities impose a starting positive correlation between activity information and location-labels, and allow for identifiability in the problem. As per usual, forward-backward equations compute

$$\gamma_x(t) = \mathbb{P}(x_t = x | \boldsymbol{a}_{1:T}, \boldsymbol{r}_{1:T}, \boldsymbol{d}_{1:T}) \tag{4}$$

and

$$\xi_{x,x'}(t-1) = \mathbb{P}(x_{t-1} = x, x_t = x' | \boldsymbol{a}_{1:T}, \boldsymbol{r}_{1:T}, \boldsymbol{d}_{1:T}),$$

under the current estimate for model parameters. At each iteration k, parameter updates are given by

$$p_{x,i}^{(k)} \leftarrow \frac{\sum_t \gamma_x(t)\mathbb{I}_{\{a_t(i)=1\}}}{\sum_t \gamma_x(t)}, \quad \mu_{x,i}^{(k)} \leftarrow \frac{\sum_t \gamma_x(t)\mathbb{I}_{\{a_t(i)=1\}} r_t(i)}{\sum_t \gamma_x(t)\mathbb{I}_{\{a_t(i)=1\}}}$$

and

$$\sigma_{x,i}^{(k)} \leftarrow \frac{\sum_t \gamma_x(t)\mathbb{I}_{\{a_t(i)=1\}}(r_t(i) - \mu_{x,i}^{(k)})^2}{\sum_t \gamma_x(t)\mathbb{I}_{\{a_t(i)=1\}}},$$

for all $x \in \mathcal{X}$ and $i \in \{1, \cdots, n\}$. In addition, transition parameters in (3) are numerically updated so that

$$(\nu_x^{(k)}, \{\lambda_{x,x'}^{(k)} \ : \ x' \in \mathcal{X} \text{ with } x \neq x'\}) \leftarrow \arg\max_{\nu, \boldsymbol{\lambda}} \mathcal{S}(x, \nu, \boldsymbol{\lambda})$$

for all $x \in \mathcal{X}$, with

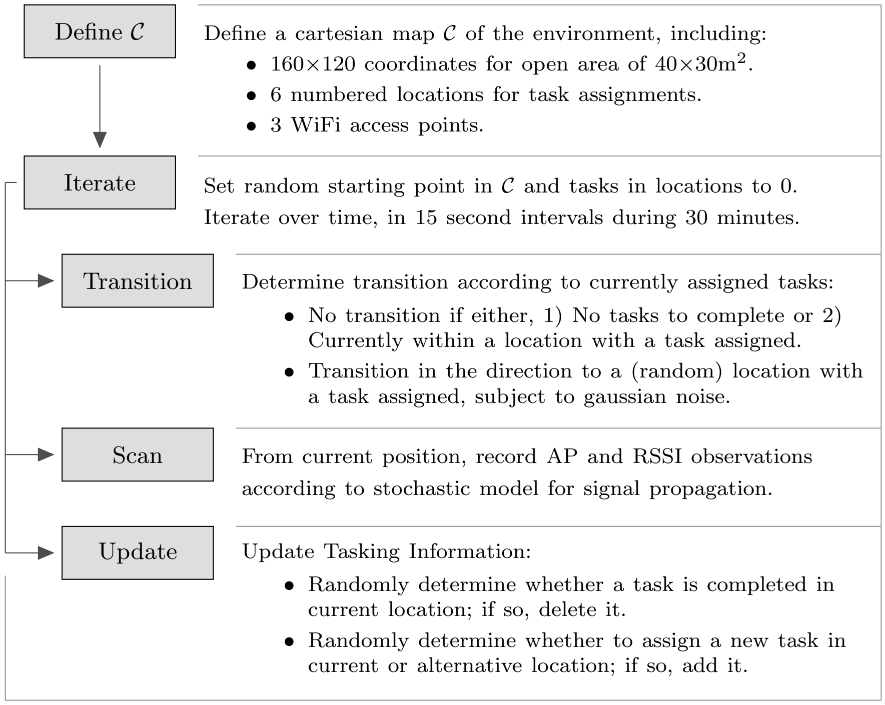
$$\begin{aligned}\mathcal{S}(x, \nu, \boldsymbol{\lambda}) = &\sum_{t=2}^{T} \xi_{x,x}(t-1)\nu \boldsymbol{d}_t(x) + \sum_{t=2}^{T} \sum_{\substack{x' \in \mathcal{X} \\ x' \neq x}} \xi_{x,x'}(t-1)(\nu \boldsymbol{d}_t(x') + \lambda_{x'}) \\ &- \sum_{t=2}^{T} \gamma_x(t-1) \log(e^{\nu \boldsymbol{d}_t(x)} + \sum_{\substack{x' \in \mathcal{X} \\ x' \neq x}} e^{\nu \boldsymbol{d}_t(x') + \lambda_{x'}}).\end{aligned}$$

The above steps are repeated until convergence, and sequential location-labels are assigned according to resulting distributions in (4). Alternatively, it is possible to use the Viterbi algorithm to extract the most likely sequence of labels.

5. Coordinate Classifier and Results

In order to test the validity of the labelling approach, we resort to both simulated and real data sets. Simulations present human transitions within an indoor environment; reception device carriers move through a set of numbered locations as they must complete assigned tasks, aiming to recreate the working environment within a hospital. Also, WiFi network scans are produced and recorded over small, equally-spaced intervals. Figure 3 offers a step-wise procedure diagram for each 30 minute simulation in our experiment.

We notice that task assignments and the choice to complete them in one or another order is randomized; thus, it is not possible to assume a device is in any given location at any point in time. After each simulation, positioning data is ignored and only sequences of tasks and WiFi observations are stored; in Figure 4 we observe a sample task-flow obtained from a single simulation, including creation times, completion times and locations assigned. For instance, during



Define a cartesian map $\mathcal{C}$ of the environment, including:

- 160×120 coordinates for open area of $40 \times 30\text{m}^2$.
- 6 numbered locations for task assignments.
- 3 WiFi access points.

Set random starting point in $\mathcal{C}$ and tasks in locations to 0. Iterate over time, in 15 second intervals during 30 minutes.

Determine transition according to currently assigned tasks:

- No transition if either, 1) No tasks to complete or 2) Currently within a location with a task assigned.
- Transition in the direction to a (random) location with a task assigned, subject to gaussian noise.

From current position, record AP and RSSI observations according to stochastic model for signal propagation.

Update Tasking Information:

- Randomly determine whether a task is completed in current location; if so, delete it.
- Randomly determine whether to assign a new task in current or alternative location; if so, add it.

Figure 3: Step-wise procedure diagram for a 30 minute simulation of transitions.

the 12th minute, we observe two active tasks in association to location 6, along with a single task linked to location 3; this enables the construction of counts $\boldsymbol{d}$ within the model, and offers prior knowledge on where a device is *likely* to be. A small set of real *tasking-data* within a university hospital, along with radio scans, is also available and used for model-testing purposes within a wider environment.

5.1. Cartesian Coordinate Classifier

We recall that the spatial sequential-labelling approach introduced in this paper bypasses the need to know the layout of a environment; its aim is to cluster signal observations under the influence of a maximum-entropy model for transitions, and assign location-labels accordingly. Hence, a specific point in a room could be associated with different labels depending on the signal scan obtained, specially in areas distant from any point of interest identified in the meta-data. However, for model-testing purposes within a known Cartesian

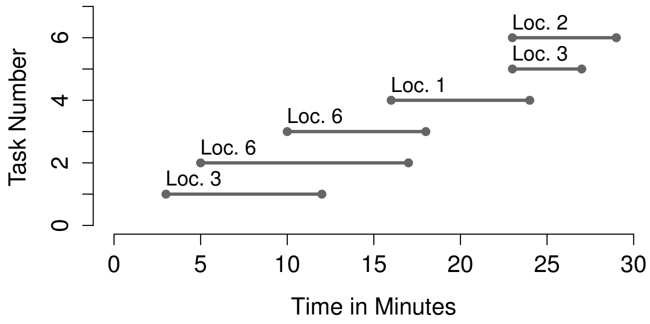


Figure 4: Task-flow with assigned locations for a simulated transition interval. Creation and completion times are represented by circled dots.

coordinate system $\mathcal{C}$, a classifier $h(\cdot) : \mathcal{C} \to \mathcal{X}$ can be constructed. Note that

$$\begin{aligned}\mathbb{P}(x_{\boldsymbol{y}} = x) &= \sum_{\mathcal{A}} \int_{\mathcal{R}} \mathbb{P}(x_{\boldsymbol{y}} = x | \boldsymbol{a}_y, \boldsymbol{r}_y) \mathbb{P}(\boldsymbol{a}_y, \boldsymbol{r}_y) \mathrm{d}\boldsymbol{r}_y \\ &= \mathbb{E}_{\boldsymbol{a}_y, \boldsymbol{r}_y} \Big[\frac{\mathbb{P}(\boldsymbol{a}_y, \boldsymbol{r}_y | x_{\boldsymbol{y}} = x)}{\sum_{x \in \mathcal{X}} \mathbb{P}(\boldsymbol{a}_y, \boldsymbol{r}_y | x_{\boldsymbol{y}} = x)} \Big], \end{aligned} \tag{5}$$

where $\boldsymbol{y} = (y_1, y_2) \in \mathcal{C}$ denotes a position in a 2-dimensional Cartesian map. Here, $\mathbb{P}(\boldsymbol{a}_y, \boldsymbol{r}_y)$ denotes the distribution of signal observations as seen from $\boldsymbol{y}$ and can be approximated with a collection of *fingerprints* sufficiently large. Then, h is defined so that

$$h(\boldsymbol{y}) = \arg\max_{x} \mathbb{P}(x_{\boldsymbol{y}} = x) \ ,$$

and $\mathbb{P}(x_{\boldsymbol{y}} = x)$ in (5) is approximated averaging the inner part of the expectation across fingerprints.

5.2. Results on Simulated Data

Here, we discuss results obtained from feeding meta-data and radio scans from multiple simulated transitions into the model. We recall that the learning procedure ignores physical positioning data; yet, we aim to evaluate its ability to form segregated physical clusters in association to labels in the presented experiment (see Figure 3). Note that each label is associated to one out of six locations where simulated tasks are carried out.

First, the top plot in Figure 5 presents prior membership probabilities over time in relation to 6 different locations, as introduced in (2). This information relates to a single simulation for a mobile device carrier within the recreated indoor environment; it offers an insight on the location a device is *likely* to be spatially associated with. Probabilities are displayed on a gray-scale and higher intensities reflect higher likelihoods. On the other hand, the bottom plot shows labelled locations as inferred from radio scans and estimated transitions; these represent the sequence of inferred labels over time for the device carrier.

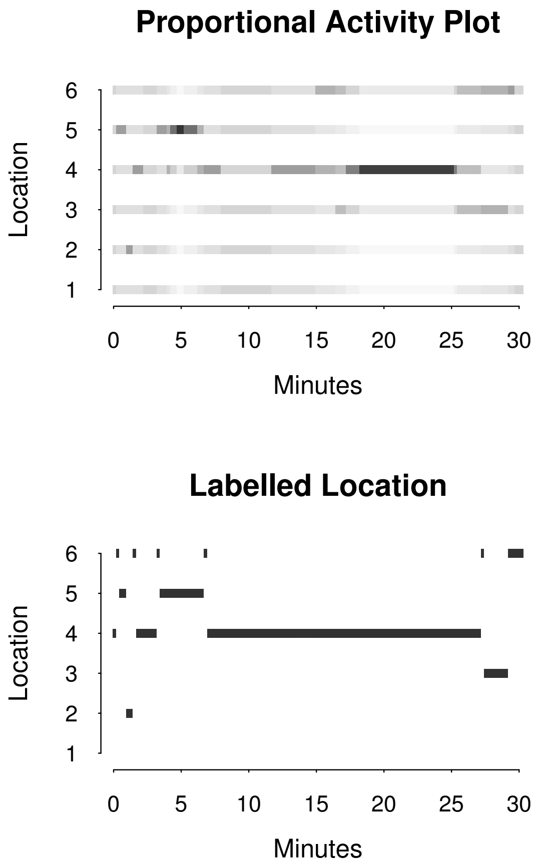


Figure 5: Prior activity likelihood and label sequence plots in relation to locations within the indoor environment. Probabilities displayed on a gray-scale.

In addition, Figure 6 displays the Cartesian map representation $\mathcal{C}$ of the simulated environment, along with color-coded label membership intensities over coordinates $\boldsymbol{y} \in \mathcal{C}$, as deduced from (5). There, we observe how segregated clusters of coordinates are formed in association to each activity-point of interest; at any given time, it is safe to assume no task is being undertaken in one location provided no corresponding label was assigned. Also, we observe that the location of APs (represented by red dots) is likely to influence the shape of coordinate groups.

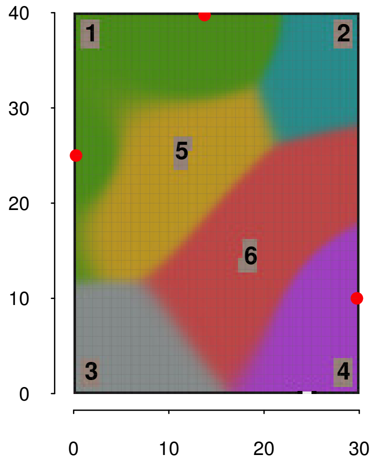


Figure 6: Color-coded map of the environment with learnt clusters, distance shown in meters. Red dots represent the location of APs.

5.3. Results on Real Data

The data set used includes information regarding duties over night shifts for several doctors. This tasking-data is for a 1-week time-window and has been retrieved from a hospital management system in place within a university hospital. This information has been merged with WiFi radio scans obtained from devices carried by doctors; and the joint structure of the data resembles that discussed in the simulated experiment.

Due to the size of the indoor environment and the small amount of observations available, we do not aim to achieve ward-level labelling. The focus rather lies in the creation of *super-clusters*; aiming to segregate physical areas that include several locations of interest. For that matter, tasking information on different wards in the hospital is grouped together whenever there exists some certainty regarding ward proximity; such knowledge may be gained both from meta-data information and outdoors observations of building structures. In practice, should a doctor have a task associated to a (named or numbered) ward, then he has a task associated to the greater (named or numbered) region that ward falls in. Hence, we study the ability of the proposed methodology in order to form segregated regions in association to groups of locations (wards).

In this trial, over 500 APs where observed and a subset of most-relevant ones was isolated, using characteristic associations to different wards as noticed in tasking information. In Figure 7 we find the probability of observing different APs from formed super-clusters within the hospital (each in association to one area label); we recall that signal reception capabilities are assumed uniform across devices, and we notice significant associations between certain APs and specific clusters. Also, high probability values are rare within large regions, given the limited range of radio signals; on the contrary, the opposite is true in small isolated areas such as the *Nurse Coordinator* room, where reception devices were stored in-between doctor shifts. In addition, Figure 8 presents the distribution of signal reception strength for various APs, as observed from a single clustered region in association to the *South* label; we recall that these represent weighted distributions over different points within the clustered area. The weighting is associated to the time device carriers spend at specific physical points. Normal approximations to signal densities are shown in dashed red.

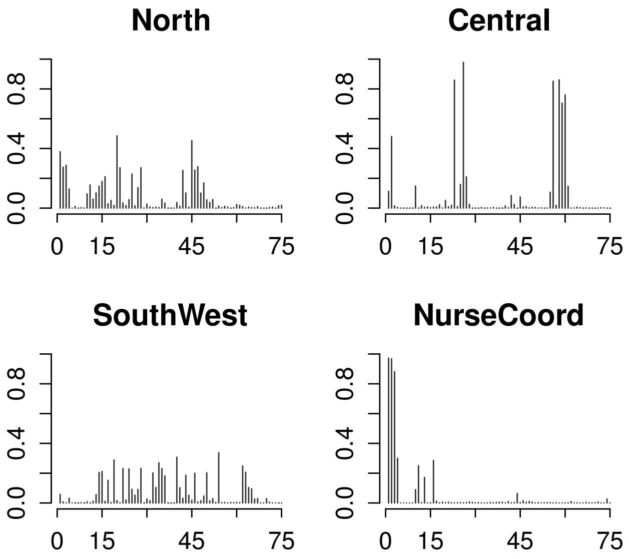


Figure 7: Probability of observing APs from 4 different super-clusters within the hospital. The horizontal axis denotes the APs.

Finally, Figure 9 presents a 2-dimensional scaled map of the hospital, along with color-labelled fingerprints distributed over different regions and floors; the choice of fingerprint placements is made in relation to observed wards within the tasking data set. Information relating doctor duties is strongly linked to southern areas of the hospital, while their presence elsewhere is rare. Thus, we notice the ability to segregate radio scans from small parts within southern buildings, whereas the northern block is grouped as a whole.

6. Discussion

The work in this paper has addressed an indoor positioning problem as an unsupervised labelling task on sequential data; exploiting the use of contemporary techniques in pattern recognition and natural language processing, in order

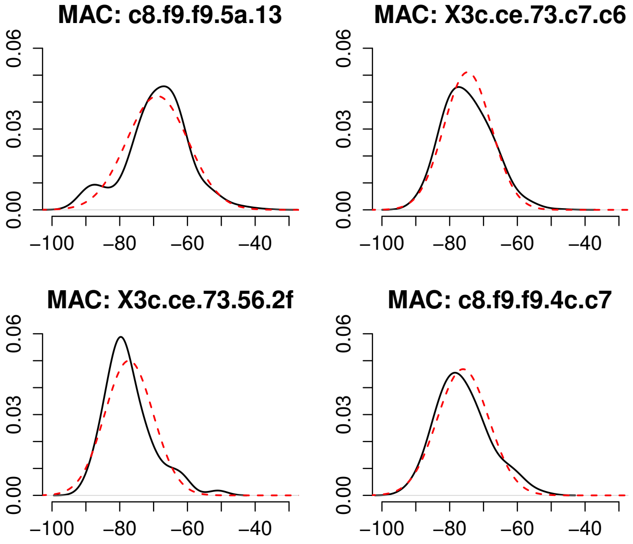


Figure 8: RSSIs for various APs, as observed from *south* supercluster. Normal approximations to signal densities are shown in a dotted red line.

to formalize informative spatial classifiers within unknown ubiquitous environments. Tests on simulated data have shown the ability of the framework in order to dissociate structures of radio signal observations within small places; forming clustered regions in association to areas of interest to the user. Furthermore, a small-scale study within a large hospital environment has confirmed the potential of such a probabilistic approach in order to extend on current positioning literature in relation to the extraction of *indoor location landmarks* (cf. Wang et al. (2012)).

Some restrictions in the present approach relate to the process of AP selection in large environments, prior to a labelling study. Lacking on knowledge regarding the placement of APs presents a challenge in order to ensure a choice with proper signal covering of the entire environment. In addition, an extensive amount can increase the complexity, become computationally demanding and raise traditional concerns on identifiability and starting parameter choices; also, in combination with poor quality or not sufficiently large contextual information can lead to over-fitting. Super-clusters in Figure 9 are biased by patterns of few task allocations within the contextual information available; thus, further work is required in order to determine an optimal training data-set size for practical applications and achieving ward-level labelling.

In conclusion, the work first proposes an efficient method and valuable working framework that provides evidence regarding the power of labelling as an inference solution for real-world unsupervised indoor location; resorting to increasingly available information within the context of hospitals, museums or

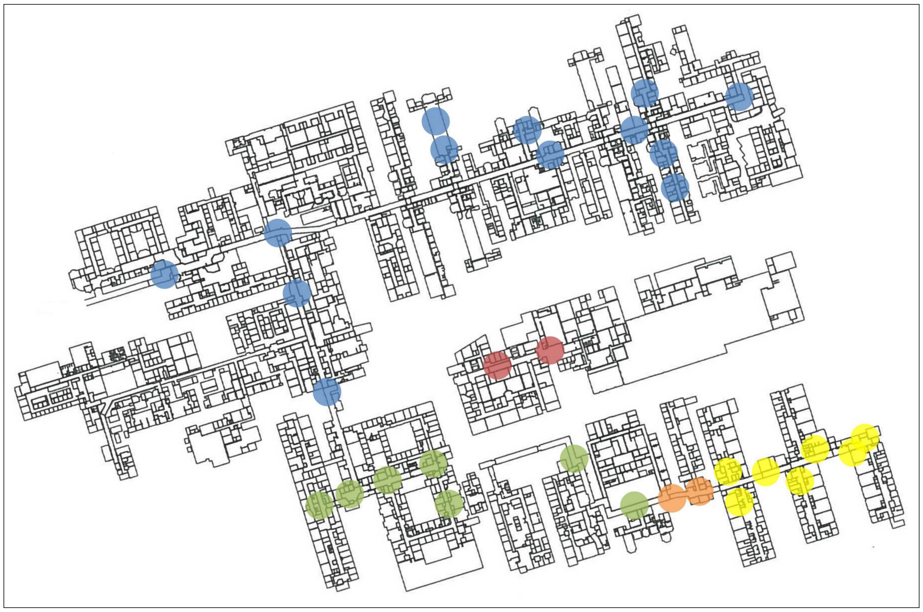


Figure 9: Hospital map along with color-labelled fingerprints. North, central, south west, south and southeast areas represented by blue, red, green, orange and yellow colors respectively.

airports. Additionally, the paper contributes to the study of indoor positioning without relying on pre-determined maps (cf. Shin et al. (2012); Wang et al. (2012)), and extends on literature analysing applications of graphical models generally focused on the fields of biology, language processing and computer vision. Finally, the work presents practical applications in the managerial and behavioural study within a workplace ((Pinchin et al., 2014)).

Acknowledgements

We would like to thank all the medical and administrative staff; visitors and patients; medical students, junior doctors and hospital volunteers that took part in the various studies. We thank Sarah Martindale, Gavin Smith, John Blakey and Dominick Shaw for assistance and comments that greatly improved the manuscript.

This work was supported by The Health Foundation through the Insight 2014 project "Informatics to identify and inform best practice in out of hours secondary care" (7382) and RCUK thought the Horizon Digital Economy Research grants (EP/G065802/1, EP/M000877/1).

References

Bahl, P., & Padmanabhan, V. N. (2000). Radar: An in-building rf-based user location and tracking system. *INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, *2*, 775–784.

Bolliger, P., Partridge, K., Chu, M., & Langheinrich, M. (2009). Improving location fingerprinting through motion detection and asynchronous interval labeling. In *Location and Context Awareness* (pp. 37–51). Springer.

Brunato, M., & Battiti, R. (2005). Statistical learning theory for location fingerprinting in wireless lans. *Computer Networks*, *47*, 825–845.

Campos, R. S., Lovisolo, L., & de Campos, M. L. R. (2014). Wi-fi multi-floor indoor positioning considering architectural aspects and controlled computational complexity. *Expert systems with applications*, *41*, 6211–6223.

Chen, M. Y., Sohn, T., Chmelev, D., Haehnel, D., Hightower, J., Hughes, J., LaMarca, A., Potter, F., Smith, I., & Varshavsky, A. (2006). Practical metropolitan-scale positioning for gsm phones. In *UbiComp 2006: Ubiquitous Computing* (pp. 225–242). Springer.

Dietterich, T. G. (2002). Machine learning for sequential data: A review. In *Structural, syntactic, and statistical pattern recognition* (pp. 15–30). Springer.

Fung, P. T.-k., & Grimble, M. J. (1983). Dynamic ship positioning using a self-tuning kalman filter. *Automatic Control, IEEE Transactions on*, *28*, 339–350.

Ghahramani, Z. (2001). An introduction to hidden markov models and bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, *15*, 9–42.

Gu, Y., Lo, A., & Niemegeers, I. (2009). A survey of indoor positioning systems for wireless personal networks. *Communications Surveys & Tutorials, IEEE*, *11*, 13–32.

Gustafsson, F., Gunnarsson, F., Bergman, N., Forssell, U., Jansson, J., Karlsson, R., & Nordlund, P.-J. (2002). Particle filters for positioning, navigation, and tracking. *Signal Processing, IEEE Transactions on*, *50*, 425–437.

Harle, R. (2013). A survey of indoor inertial positioning systems for pedestrians. *Communications Surveys & Tutorials, IEEE*, *15*, 1281–1293.

Harley, P. (1989). Short distance attenuation measurements at 900 mhz and 1.8 ghz using low antenna heights for microcells. *Selected Areas in Communications, IEEE Journal on*, *7*, 5–11.

Hashemi, H. (1993). Impulse response modeling of indoor radio propagation channels. *Selected Areas in Communications, IEEE Journal on*, *11*, 967–978.

Hoang, M. K., Schmalenstroeer, J., Drueke, C., Tran Vu, D., & Haeb-Umbach, R. (2013). A hidden markov model for indoor user tracking based on wifi fingerprinting and step detection. *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European*, (pp. 1–5).

Jordan, M. I. (1998). *Learning in graphical models* volume 89. Springer Science & Business Media.

Jordan, M. I., Ghahramani, Z., & Saul, L. K. (1997). Hidden markov decision trees. *Advances in neural information processing systems*, (pp. 501–507).

Kaemarungsi, K., & Krishnamurthy, P. (2004). Modeling of indoor positioning systems based on location fingerprinting. *INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies*, *2*, 1012–1022.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Krakiwsky, E. J., Harris, C. B., & Wong, R. V. (1988). A kalman filter for integrating dead reckoning, map matching and gps positioning. In *Position Location and Navigation Symposium, 1988. Record. Navigation into the 21st Century. IEEE PLANS'88., IEEE* (pp. 39–46). IEEE.

Krishnan, P., Krishnakumar, A., Ju, W.-H., Mallows, C., & Gamt, S. (2004). A system for lease: Location estimation assisted by stationary emitters for indoor rf wireless networks. *INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies*, *2*, 1001–1011.

Ladd, A. M., Bekris, K. E., Rudys, A. P., Wallach, D. S., & Kavraki, L. E. (2004). On the feasibility of using wireless ethernet for indoor localization. *IEEE Transactions on Robotics and Automation*, *20*, 555–559.

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data, .

Li, X., Pahlavan, K., Latva-aho, M., & Ylianttila, M. (2000). Comparison of indoor geolocation methods in dsss and ofdm wireless lan systems. In *Vehicular Technology Conference, 2000. IEEE-VTS Fall VTC 2000. 52nd* (pp. 3015–3020). IEEE volume 6.

Liu, H., Darabi, H., Banerjee, P., & Liu, J. (2007). Survey of wireless indoor positioning techniques and systems. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, *37*, 1067–1080.

Ma, J., Li, X., Tao, X., & Lu, J. (2008). Cluster filtered knn: A wlan-based indoor positioning scheme. In *World of Wireless, Mobile and Multimedia Networks, 2008. WoWMoM 2008. 2008 International Symposium on a* (pp. 1–8). IEEE.

Madigan, D., Einahrawy, E., Martin, R. P., Ju, W.-H., Krishnan, P., & Krishnakumar, A. (2005). Bayesian indoor positioning systems. *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE*, *2*, 1217–1227.

Malouf, R. (2002). A comparison of algorithms for maximum entropy parameter estimation. In *proceedings of the 6th conference on Natural language learning-Volume 20* (pp. 1–7). Association for Computational Linguistics.

McCallum, A., Freitag, D., & Pereira, F. C. (2000). Maximum entropy markov models for information extraction and segmentation. In *ICML* (pp. 591–598). volume 17.

Mok, E., & Retscher, G. (2007). Location determination using wifi fingerprinting versus wifi trilateration. *Journal of Location Based Services*, *1*, 145–159.

Moon, T. K. (1996). The expectation-maximization algorithm. *Signal processing magazine, IEEE*, *13*, 47–60.

Murphy, K. P. (2002). Dynamic bayesian networks. *Probabilistic Graphical Models, M. Jordan*, *7*.

Peterson, B. B., Kmiecik, C., Hartnett, R., Thompson, P. M., Mendoza, J., & Nguyen, H. (1998). Spread spectrum indoor geolocation. *Navigation*, *45*, 97–102.

Pinchin, J., Brown, M., Blum, J., Shaw, D., & Blakey, J. (2014). Integrating wifi based positioning with a job management system to study task management behaviour. In *Position, Location and Navigation Symposium-PLANS 2014, 2014 IEEE/ION* (pp. 193–200). IEEE.

Pinchin, J., Hide, C., & Moore, T. (2012). A particle filter approach to indoor navigation using a foot mounted inertial navigation system and heuristic heading information. In *Indoor Positioning and Indoor Navigation (IPIN), 2012 International Conference on* (pp. 1–10). IEEE.

Roos, T., Myllymäki, P., & Tirri, H. (2002). A statistical modeling approach to location estimation. *Mobile Computing, IEEE Transactions on*, *1*, 59–69.

Rydén, T. et al. (2008). Em versus markov chain monte carlo for estimation of hidden markov models: A computational perspective. *Bayesian Analysis*, *3*, 659–688.

Sasiadek, J., Wang, Q., & Zeremba, M. (2000). Fuzzy adaptive kalman filtering for ins/gps data fusion. In *Intelligent Control, 2000. Proceedings of the 2000 IEEE International Symposium on* (pp. 181–186). IEEE.

Seitz, J., Vaupel, T., Jahn, J., Meyer, S., Boronat, J. G., & Thielecke, J. (2010a). A hidden markov model for urban navigation based on fingerprinting and pedestrian dead reckoning. *Information Fusion (FUSION), 2010 13th Conference on*, (pp. 1–8).

Seitz, J., Vaupel, T., Meyer, S., Boronat, J. G., & Thielecke, J. (2010b). A hidden markov model for pedestrian navigation. *Positioning Navigation and Communication (WPNC), 2010 7th Workshop on*, (pp. 120–127).

Sharples, S., Brown, M., Pinchin, J., Blum, J., Nagiyev, A., Ryan, B., Shaw, D., & Blakey, J. (2015). Ubiquitous technologies for capture of real-world performance. In *Proceedings 19th Triennial Congress of the IEA* (p. 14). volume 9.

Shen, G., Yu, J., & Tan, L. (2012). Hierarchical rss-based indoor positioning using a markov random field model. In *Wireless Communications, Networking and Mobile Computing (WiCOM), 2012 8th International Conference on* (pp. 1–4). IEEE.

Shin, H., Chon, Y., & Cha, H. (2012). Unsupervised construction of an indoor floor plan using a smartphone. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, *42*, 889–898.

Swangmuang, N., & Krishnamurthy, P. V. (2008). On clustering rss fingerprints for improving scalability of performance prediction of indoor positioning systems. In *Proceedings of the first ACM international workshop on Mobile entity localization and tracking in GPS-less environments* (pp. 61–66). ACM.

Wang, H., Sen, S., Elgohary, A., Farid, M., Youssef, M., & Choudhury, R. R. (2012). No need to war-drive: unsupervised indoor localization. In *Proceedings of the 10th international conference on Mobile systems, applications, and services* (pp. 197–210). ACM.

Yang, J., Wang, Z., & Zhang, X. (2015). An ibeacon-based indoor positioning systems for hospitals. *International Journal of Smart Home*, *9*, 161–168.

Yim, J. (2008). Introducing a decision tree-based indoor positioning technique. *Expert Systems with Applications*, *34*, 1296–1302.

Youssef, M. A., Agrawala, A., & Udaya Shankar, A. (2003). Wlan location determination via clustering and probability distributions. *Pervasive Computing and Communications, 2003.(PerCom 2003). Proceedings of the First IEEE International Conference on*, (pp. 143–150).