Wireless Pers Commun (2016) 89:365–384
DOI 10.1007/s11277-016-3270-4

Outdoor Location Estimation Using Received Signal Strength-Based Fingerprinting

Chao Ning[1,2] · Rui Li[1] · Kejiong Li[3]

Published online: 31 March 2016

Abstract A cluster-based intersection fingerprinting technique for outdoor location estimation using received signal strength (RSS) is proposed. The performance of the proposed scheme is demonstrated by making comparisons using RSS data from a simulated grid-based urban propagation model, RSS data generated by a network planning tool applied to a rural environment, and RSS data from real city environment. The proposed scheme first uses an optimal clustering scheme to portion the large outside area into different small regions based on the use of RSS deviations from the path loss model. For each region, a fine RSS distribution model is built to provide a good support for further positioning. An improved intersection method is then presented to find the most likely geographical area to further estimate a mobile user's location. A comparison between cluster-based and grid-based environment partitioning is made. The experimental results show that the proposed clustering scheme gives good support for location estimation and the positioning accuracy is significantly improved.

Keywords Clustering · Fingerprinting · Intersection · Received signal strength

1 Introduction

Localisation has become more and more popular in pervasive computing environments, for example, positioning a mobile user in an emergency environment. Although the global positioning system (GPS) is overwhelmingly popular for mobile devices, it is not always

✉ Chao Ning
c.ning12@imperial.ac.uk

[1] Department of Geophysics, Chengdu University of Technology, Chengdu, China

[2] Department of Earth Science and Engineering, Centre for Reservoir Geophysics, Imperial College London, London, UK

[3] School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK

the best option as: (a) GPS relies on special hardware support and has high complexity, high battery consumption and latency, which impacts on its widespread commercial use; (b) the access to GPS signals is often limited in some environments such as urban areas with many high buildings, mountainous terrain and indoor areas.

Consequently, a variety of wireless localisation techniques have been proposed in the literature, including Time of Arrival (ToA), Time Difference of Arrival (TDoA), Angle of Arrival (AoA) and Received Signal Strength (RSS)-based methods. Among these techniques, RSS-based method is seen as economic for wireless networks because it does not require additional hardware such as high precision clocks (ToA and TDoA) and antenna arrays in transmitter or receiver (AoA). Moreover, RSS data can be readily collected indoors or outdoors for most wireless systems and the data can be used to obtain either range estimates or connectivity information [1].

Many RSS-based localisation approach make use of fingerprinting [2] to overcome the limitations of traditional triangulation approaches and performs well for non-line-of-sight circumstances especially in a complex environment. However, previous fingerprint studies [3–19] mainly focus on indoor localisation due to the difficulty in acquiring the large amount of data that need to be processed for larger outdoor areas.

In this paper, we propose a outdoor location estimation scheme exhibiting a high accuracy in localising mobile stations (MSs) even with a relatively low density of reference RSS data. The proposed scheme consists of two phases: the offline training phase and the online localisation phase. In the first phase the outdoor area is partitioned into small clusters by analysing the RSS collected from historical data using an improved clustering scheme. Then in the online localisation phase the mobile location can be estimated by further analysing these clusters with the help of a refined intersection approach. The novel features that contribute to the greater accuracy include: (a) clusters are created using RSS deviations resulting from the observed path loss model which capture better the wireless topography in a complex environment, rather than the raw RSS, in each RSS component analysis. As a result, the clusters remain invariant to the variation in base station (BS) transmitter power; (b) the relationship between the accurate estimation of the cluster membership probability and the optimal number of clusters is applied to manage the trade-off between number of cluster and accuracy of clustering; (c) the application of an intersection approach in the online phase improves the accuracy of location estimation.

The rest of this paper is organised as follows. Section 2 reviews related work on location fingerprinting using RSS and analyses its two main characteristics, followed by the detailed description of the proposed location estimation scheme in Sect. 3. Performance evaluations of the proposed method and the existing localisation methods are presented and compared in Sect. 4 using both numerically simulated and real measured data. Finally, Sect. 5 discusses the results and outlines open issues for future research.

2 Location Fingerprinting Based on Received Signal Strength

The idea of RSS-based fingerprinting is simple and effective: the measurements of RSS are collected from a number of known locations to generate a database of location fingerprints (a.k.a. radio map) based on a variety of partitioning models in the training phase. Then in the online phase new RSS observations measured at unknown locations are compared with all the fingerprints in the radio map to estimate the locations based on preferred algorithms. Two popular partitioning models and typical location fingerprinting techniques are outlined in the following parts.

2.1 Partitioning Models

Grid-based partitioning and *cluster-based partitioning* are the two most popular partitioning models. The grid-based methods [8, 10] generally divide the simulation environment into uniform regular grids and then attempt to map a MS location to a point on a grid element. The spacing of the grid influences the accuracy of the position estimation [4]. Coarse grids result in loss of the accuracy dramatically. On the other hand, using fine grids can increase the accuracy but also requires a more laborious site-survey. Moreover, another key issue is that uniform grids can not reflect the topography. Some location-aware applications [8] (mainly indoor ones), do not use uniform grids but use a topographical model. In this case the environment, e.g. a office building, is divided into cells where a cell corresponds to a specific office room or hallway segment.

Recently, many cluster-based location estimation methods have been proposed with positive results. Most of these method focused on indoor environments and the IEEE 802.11b wireless LAN networks (WLAN) [9, 11, 12]. The clustering algorithms partition the environment into regions that are more homogeneously covered by the radio signal. In [9, 11], the Joint Clustering approach covers access points (APs) during the offline phase, and then applies a Maximum Likelihood estimator to determine the most probable location within the cluster in the online phase. Yiqiang et al. [12] proposed an algorithm known as CaDet for power-efficient location estimation by selecting the APs in an indoor wireless environment. The environment is modelled as a space of 99 locations, each representing a 1.5-m grid cell. In the offline phase, it uses K-means clustering method based on the similarity of RSS from APs. A decision tree over the grids in each cluster is built in the online phase for location estimation with high accuracy.

However, there are two points worth noting: firstly, previous cluster-based partitioning research [9, 11, 12] did not pay attention to the cluster stability and scalability issues of handling a large amount of data without loss of important correlation information. Secondly, the clustering results are dominantly affected by the path loss effect rather than by the correlations between RSS values associated with the topographical effects. To cope with these issues, the proposed clustering scheme makes an approximate adjustment for the distance effect and then works with the residual. The approach also has the benefit that the clusters are still invariant to the power of the BSs. The improved accuracy of our clustering approach has been demonstrated by the experimental results.

2.2 Fingerprinting-Based Techniques

Fingerprinting-based approaches can generally be categorized into two categories: *deterministic approaches* and *probabilistic approaches*. The former use deterministic inference algorithms to estimate a MS location. This essentially involves calculations of the similarity between new RSS observations at unknown locations and the trained RSS data with known location information. For example, the RADAR system [3, 13], a RF based system for locating and tracking users inside buildings, represents the first 802.11 fingerprinting structure for localisation developed by Microsoft Research. The system carries out K-Nearest-Neighbour algorithm (KNN) based on Euclidean distance function to find the K nearest neighbours of a user. Then the average of the coordinates of these K locations is used as the estimate of the user's location. Ni et al. [14] and Li and Salter [15] improved the accuracy by using a weighted average of the coordinates of the K nearest neighbours. The weight values are taken as the inverse of the Euclidean distances. This method is referred to as Weighted K-Nearest Neighbours (WKNN). The experimental results in [15] indicate that the KNN and the WKNN can provide a relatively higher accuracy than the simple Nearest Neighbour (NN) method, particularly when K = 3 and K = 4. However, when a high density radio map is available, the NN method can perform as well as other complicated methods [16]. There are several variants of the KNN method, e.g. the Database Correlation Method (DCM) [17, 18]. Our proposed approach also belongs to deterministic category.

Probabilistic approaches are often used as a tool for cope with the incompleteness in RSS models, e.g. [19]. These methods use the training RSS samples to construct a probability distribution of RSS at desired locations as the content of a radio map, and then calculate the likelihood or posterior probabilities. Although probabilistic methods are reported to provide higher positioning accuracy than deterministic approaches [19], they are quite computationally complex when the observed data is high-dimensional.

3 Outdoor Location Estimation with Clustering

The proposed location estimation scheme consists of two phases: a offline training phase and an online localisation phase, as illustrated in Fig. 1.

In the training phase, the real time RSS samples collected during the network planning stage are analysed. Based on this analysis, the large target area can be partitioned into small ones using a clustering method. The relatively homogeneous RSS distribution within each small region is then accurately modelled. When a new MS is collected, these models are involved in the K-Nearest Neighbour Venn Probability Machine (KNN-VPM) algorithm in order to estimate which small region the new MS is most probably located in. Within this region, further location estimation can be made with acceptable precision. As such, our proposed location estimation scheme is tolerant to the measurement errors.

3.1 Training Phase

The two main objectives in the training phase of the proposed scheme are to find the optimal clustering result and to create the RSS distribution model. Previous clustering studies [9, 11, 12] did not pay attention to the cluster stability and scalability issues of handling a large amount of data without loss of important correlation information. In the proposed clustering scheme, the Affinity Propagation [20] method is applied to produce clusters, while the Venn Probability Machine (VPM) [21] is utilized to determine the probability of cluster membership.

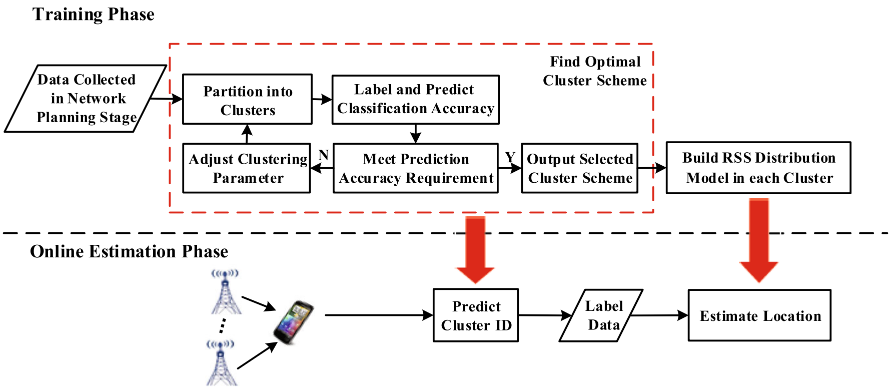

Training Phase
Find Optimal Cluster Scheme
Data Collected in Network Planning Stage
Partition into Clusters
Label and Predict Classification Accuracy
Adjust Clustering Parameter
N
Meet Prediction Accuracy Requirement
Y
Output Selected Cluster Scheme
Build RSS Distribution Model in each Cluster
Online Estimation Phase
Predict Cluster ID
Label Data
Estimate Location

Fig. 1 The overview of the proposed location estimation scheme

3.1.1 Clustering Mobile Stations' RSS Data

In the context of wireless networks, there are two benefits of Affinity Propagation (AP) clustering technique for this research work: (a) the clusters emerge naturally and the number of clusters is related to a pre-set "preference" value, rather than by setting the number of clusters in advance; (b) it allows great flexibility in the face of dynamic environments, since all clustering parameters can be changed across iterations.

In this work, the similarity calculation in the AP clustering process is based on the Mahalanobis distance rather than the Euclidean distance in the signal space to create distinct and stable clusters. This is because Mahalanobis distance function can avoid giving too much weight to correlated RSS values in the distance function and enables both non-linear and linear decision boundaries. Comparisons between the Euclidean distance and Mahalanobis distance on real data set are presented in Sect. 4.

Let $\mathbf{r_i} = (\mathbf{r_{i,1}}, \mathbf{r_{i,2}}, \ldots, \mathbf{r_{i,q}})$ represent the RSS tuple of MS i received from q neighbouring antennas in the area of interest. The RSS deviations that can be obtained based on the log-distance path loss models can be given as $\boldsymbol{\rho}_i = (\rho_{i,1}, \rho_{i,2}, \ldots, \rho_{i,q})$. For any two MSs, such as MS i and MS k, the similarity between them can be expressed as

$$s(i,k) = -\sqrt{(\boldsymbol{\rho}_i - \boldsymbol{\rho}_k)\Sigma^{-1}(\boldsymbol{\rho}_i - \boldsymbol{\rho}_k)^T}. \tag{1}$$

Because the signal strength received by MSs from different BSs can be correlated, the covariance matrix Σ in the signal space is used in (1) to describe the mutual dependence of the signal strength received by the two MSs from different BSs. If there are q adjacent BSs, Σ can be estimated as

$$\Sigma_{q\times q} = \begin{bmatrix} \Sigma_{1,1} & \cdots & \Sigma_{1,q} \\ \vdots & \ddots & \vdots \\ \Sigma_{q,1} & \cdots & \Sigma_{q,q} \end{bmatrix}, \tag{2}$$

where $\Sigma_{j,p} = \frac{1}{(n-1)}\sum_{i=1}^{n}(\rho_{i,j} - \bar{\rho}_j)(\rho_{i,p} - \bar{\rho}_j)^T$, $1 \le j,p \le q$, n is the number of MSs, the superscript T represents transpose and $\bar{\rho}_j$ is the average RSS deviation of all MSs from BS j, $\bar{\rho}_j = \frac{1}{n}\sum_{i=1}^{n}\rho_{i,j}$. If Σ is replaced by a identity matrix, Eq. (1) will become the Euclidean distance function.

Moreover, the use of RSS deviation data for calculating the similarity can eliminate the effects of distance dependent path loss attenuation to some extent. Therefore, the effects of multipath and shadowing associated mainly with the topography can be better captured. On the contrary, similarity calculation using raw RSS will be dominated by the path loss. Comparisons between using RSS deviation and raw RSS are made in Sect. 4.

According to the log-distance path model [22], the RSS measurement P_{rss} (in dBm) at the distance d from a transmitter can be calculated as

$$P_{rss} = PTR + \kappa + \gamma\log(d/d_0), \tag{3}$$

where PTR represents the transmit power (in dBm) of the transmitter, d_0 is reference distance for the antenna area and its value is set to 100 m in this research. The values of parameter γ and κ are heavily dependent on the environment and can be estimated by a least squares linear regression based on all MSs' RSS from the transmitter in the training phase.

3.1.2 Estimation of the Accuracy of Cluster Identification

The Venn Probability Machine (VPM) [21] is a classification system usually applied on top of an existing learning algorithm, e.g. KNN, to augment predictions with probability estimates. In this research, the VPM is used for determining the probability of cluster membership. Based on this, we can manage the trade-off between the accuracy of cluster identification and the number of clusters.

Specifically, the RSS training data set are randomly split into two portions as the cluster training and cluster testing sets, respectively. The cluster training set is used as the representatives of the clusters that have been produced, while the cluster testing set is allocated to these clusters based on KNN. As such, we can calculate the probability of one MS in the cluster testing set belonging to one cluster, which means the most probable cluster ID for each testing MS can be estimated and verified. According to this resultant accuracy of cluster identification and the number of clusters produced, the preference value of the Affinity Propagation method can be optimised iteratively.

The process of cluster estimation can be formulated as follows and described in detail in Algorithm (1). Let R represent the space of RSS tuples of MSs from the neighbouring BSs, and C be the space of cluster IDs, and $Z = R \times C$, which denotes the pair [RSS tuple, cluster ID] for every MS in the area of interest. The clustering set $C = \{C_1, C_2, C_3, \ldots, C_T\}$ and T is the number of clusters. The training data set, TR, can be represented as $TR = \{z_1, z_2, z_3, \ldots, z_N\}$, where $z_n = [r_n, c_n]$, $c_n \in C$. The testing data set, TS, can be denoted as $TS = \{z_{N+1}, z_{N+2}, z_{N+3}, \ldots, z_{N+S}\}$. Note that the cluster ID of each test data tuple is assumed unknown in the cluster identification process and is only used for cluster verification.

Algorithm 1 K-Nearest Neighbours Venn Probability Machine

```
Requried:
    $k_{max}$: the maximum value of nearest neighbours used
    Cluster ID: $\{C_1, C_2, C_3, ..., C_T\}$
    Training data set $TR = \{z_1, z_2, z_3, ..., z_N\}$, $(z_n = [r_n, c_n], c_n \in C)$
    Test data set $TS = \{z_{N+1}, z_{N+2}, z_{N+3}, ..., z_{N+S}\}$
Steps:
for $s = 1$ to $S$ do
    $TM = \{z_1, z_2, z_3, ..., z_N, z_{N+s}\}$.
    Using RSS to calculate the distances and get each $z_i$ its neighbours $Neighbour(z_i)$
    in ascending order of respective distance.
    for $t = 1$ to $T$ do
        Assign $z_{N+s} \in C_t$
        for $k = k_{max}$ to $1$ do
            if $\exists z_p \in TR$ such that $Neighbour(z_p)(1:k) == Neighbour(z_{N+s})(1:k)$ then
                $k_{eff} = k$
                Put $z_p$ into $\mathcal{Z}$
                Fill $\mathcal{Z}$ with all other $z_q$ in $TR$ that satisfy
                    $Neighbour(z_q)(1:k_{eff}) = Neighbour(z_{N+s})(1:k_{eff})$
                Break
            end if
        end for
        for $\tau = 1$ to $T$ do
            Calculate the frequency of each cluster
                $P_{t,\tau} = \frac{sizeof(\{z_\tau \in \mathcal{Z}, c_\tau \in C_t\})}{sizeof(\mathcal{Z})}$
        end for
    end for
    $c_{N+s} = \arg\max_{c_{N+s} \leq T}(\min\{P_{c_{N+s},1}, ..., P_{c_{N+s},T}\} + \max\{P_{c_{N+s},1}, ..., P_{c_{N+s},T}\})$
end for
```

First, choose one test MS from the test data set (step 1) and combine it with the training set TR to form a new data set (step 2). Use KNN algorithm to obtain a list of neighbours for each MS (step 3). Then the process works recursively on different cluster IDs (step 4). Specifically, the test MS is assigned with current cluster ID (step 5), so each MS gets a list of cluster IDs which can be converted from its list of neighbours. *Compare the first k (initially k_{max}) cluster ID in the respective list between the test MS and each training MS* (step 6). If there exists no training MSs that has the same sequence of the first k cluster ID with the test MS (step 7), decrease k by one (step 6) and repeat. Once a training MS satisfying this condition is found, the effective value of k is set as k_{eff} (step 8). Then, put this training MS and all other eligible training MSs into collection $\mathcal{Z}$ (steps 9 and 10). The normalised frequency of each cluster can be obtained by counting the number of MSs in $\mathcal{Z}$ (steps 14–16). These probabilities also compose the corresponding column of the frequency matrix. Repeat step 4 until all the cluster IDs are analysed. As a result, all the columns of the matrix can also be filled. Finally, the mean of the maximum and minimum values of each row is regarded as the probability that the selected test MS belongs to each corresponding cluster. Therefore, the cluster ID of the test MS can be estimated as the one with the largest probability (step 18). The cluster ID of the other test MSs can be estimated in the same way. Further cluster verification will compare the estimated cluster IDs with $\{c_{N+1}, c_{N+2}, c_{N+3}, \ldots, c_{N+S}\}$, and hence yields the accuracy of cluster identification.

3.1.3 Selecting the Number of Clusters

As demonstrated in the previous sections, the AP clustering and VPM method are combined in the training phase. However, changes of the preference value of the AP clustering can result in quite difference clustering results and impact on the accuracy of cluster identification. In fact, a stable clustering result can be very useful for the purpose of monitoring a dynamic MS environment and predicting users' locations. Therefore, it should be taken into consideration of the stable clustering and coherent partitioning in the RSS space.

An example is shown in Fig. 2 to illustrate the selection of the optimal number of clusters. The thick blue line represents the relationship between the AP preference

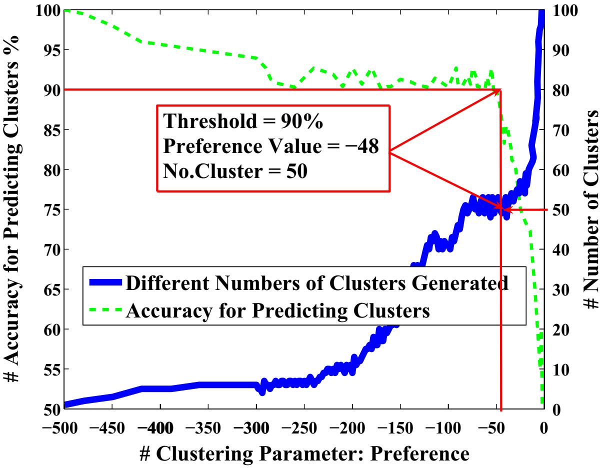


Fig. 2 An example of the selection of cluster number

parameter value and the number of clusters produced. The green dashed line depicts the dependence of the cluster prediction accuracy on the number of clusters produced. It can be seen that there is a trade-off between the number of clusters and the accuracy of location estimation. A greater number of clusters generated in the training phase comes at the cost of reduced cluster identification accuracy. For example, if there is only one cluster, the cluster identification accuracy must be 100 % but the accuracy of location estimation will be poor. On the contrary, if there are so many clusters that each MS has an exclusive cluster, partitioning scheme will eventually useless for the location estimation. Therefore, an optimal number of cluster is a vital factor for accurate location estimation. In this example, the objective is to find the right balance between the accuracy of cluster identification and the number of clusters. The number of clusters is the maximum number of clusters that can still satisfy the accuracy requirements for cluster identification. As can be seen in Fig. 2, if the threshold accuracy of cluster identification is taken as 90 %, the corresponding maximum number of clusters is 50.

3.1.4 Building RSS Distribution Model in Each Cluster

After the optimal number of clusters is determined, another regression models for each combination of cluster and BS needs to be carried out to find the optimal parameter for the RSS distribution model.

For a MS j in a cluster C_i, we use the signal strength received by MS j from BS b to calculate the RSS distance between MS j and BS b following

$$\hat{d}_{j,b} = 10^{(PTR - r_{j,b})/10\alpha_{i,b}}, \tag{4}$$

where $r_{j,b}$ is the RSS of MS j from BS b, PTR is the value of the transmission power which is given a default value of 48 dBm for outdoor GSM environment in this research. The objective is to determine the optimal parameter $\alpha_{i,b}$ for RSS distribution model of cluster C_i and BS b. In this research, the optimal $\alpha_{i,b}$ is calculated by minimizing the sum of the squared errors $\sum_{j=1}^{n_i}(d_{j,b} - \hat{d}_{j,b})^2$ where $d_{j,b}$ is the geographical distance between the locations of BS b and MS j. n_i is the number of training MSs in cluster C_i.

3.2 Online Estimation Phase

In this phase, the location of a new MS is estimated. Given a new MS m, its the observed RSS tuple from q neighbouring BSs is $r_m = (r_{m,1}, r_{m,2}, \ldots, r_{m,q})$. The detailed process of its location estimation is described below:

Step 1 The cluster which MS m belonging to can be estimated following Algorithm (1). Call this cluster C_i.

Step 2 Based on function (4) and the optimised parameter $\alpha_{i,b}$, the RSS distance from MS m to each BS can be calculated as $\hat{d}_{m,b}(1 \le b \le q)$, as shown in Fig. 3. For the simplicity, only three BSs are plotted.

Step 3 By applying KNN algorithm, MS m's K nearest neighbours in terms of RSS distance can be found among the MSs in the cluster C_i of the training data set. These neighbours are sorted as $\{p_1, \ldots, p_k, \ldots, p_K\}$ in ascending order of RSS distance. Then a deviation value of each of MS m's K neighbours is calculated using

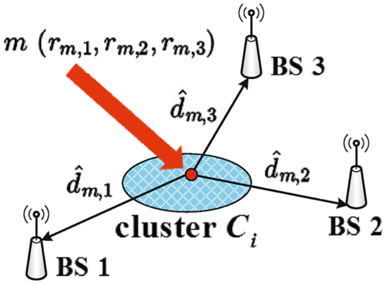


Fig. 3 Application of RSS distribution models for new MS *m*

$$\delta_{p_k,b} = \sqrt{\left|d_{p_k,b} - \overline{d_{m,b}}\right|}, \tag{5}$$

where $d_{p_{k,b}}$ is the geographical distance between BS *b* and MS p_k, $\overline{d_{m,b}}$ is the geographical distance from the centroid of the *K* nearest neighbours to BS *b*:

$$\overline{d_{p,b}} = \sum_{k=1}^{K} d_{p_k,b}/K. \tag{6}$$

Let μ and σ be the mean value and the standard deviation of these deviations, respectively. Since the RSS distribution is skew in the real environment, the confidence interval derived from the normal distribution cannot be used in this case. Instead, a two-sided confidence interval can be estimated by applying Chebyshev's inequality [23] as

$$P\left(\left|\delta_{m,b} - \mu\right| \geq \lambda\sigma\right) \leq \frac{1}{\lambda^2}, \tag{7}$$

which can provide a lower bound for how much probability mass lies outside a the chosen confidence range. For example, if the value of $\frac{1}{\lambda^2}$ on the right hand side is set as 0.01, there is always at least 99 % of the probability of being inside the distance band interval no matter what type of the distribution is. The value of the upper bound of the band width as the uncertainly band, $\delta_{m,b}$ is now chosen. Thus, the estimated geographical distance, $d_{m,b}$, between the new MS *m* and BS *b* falls in the range of

$$d_{m,b} \in \left[\hat{d}_{m,b} - \delta_{m,b}, \hat{d}_{m,b} + \delta_{m,b}\right]. \tag{8}$$

Step 4 At this step, the aim is to narrow the scope of where MS *m* is most likely to be located. Here we use an intersection scheme as described as follows. Due to the uncertainty of estimated geographical distance of MS *m*, several intersection areas are generated. This is illustrated in Fig. 4, where the intersection areas that contain at least one MS are ABCDEA and AEFGA. As inspired by [24], we design a search strategy to select the optimal intersection as below.

1. The intersection area that has the most nearest neighbours of MS *m* is selected. In Fig. 4, the intersection area ABCDEA has three neighbours and the area AEFGA has two. Therefore intersection area ABCDEA is selected as MS *m*'s estimate area.
2. If more than one intersection areas are qualified, select the one where the sum of RSS distance between each nearest neighbour and MS *m* is the smallest.

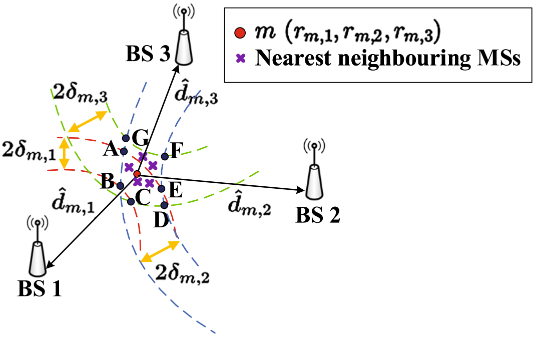


Fig. 4 Uncertainty area of location estimation

Step 5 Collect all the training MSs in cluster C_i that are located in the selected intersection area, and use the WKNN algorithm to determine MS m's location. Assuming there are $K^{'}$ qualified training MSs, and $(r_1, \ldots, r_i, \ldots, r_{K'})$ and $(l_1, \ldots, l_i, \ldots, l_{K'})$ denote their RSS tuples and location vectors, respectively. The location of MS m is finally estimated by

$$\hat{l}_m = \sum_{i=1}^{K'} w_i l_i, \tag{9}$$

where the weighting factor w_i is a normalized weight for each training MS and can be calculated as

$$w_i = \frac{1}{\left\| \mathbf{r_i} - \mathbf{r_m} \right\| \cdot \sum_{\mathbf{i=1}}^{\mathbf{K'}} \frac{\mathbf{1}}{\|\mathbf{r_i} - \mathbf{r_m}\|}} \tag{10}$$

Figure 5 briefly illustrates how the above six steps are processed using a real training data set.

4 Performance Evaluation

In this section the proposed location estimation scheme is tested with three different data sets. In each experiment, we randomly divide the data into two sets of equal size. The first one is processed in the training phase, while the other half is used only for the online phase.

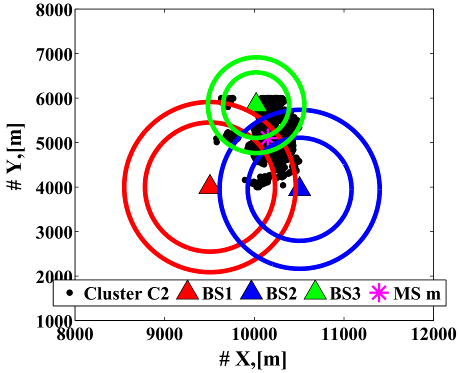


Fig. 5 An example of the location estimation for a new MS m

The performance of the proposed intersection strategy is compared with KNN algorithms using different nearest neighbours based on two different partitioning models, viz. grid and clustering. For fair comparison the number of grid elements is equal to the number of clusters produced by AP clustering method.

4.1 Simulated Data and Real Data

4.1.1 Scenario 1: A Numerically Simulated Urban Propagation Model

A 2 km × 2 km square area with four BSs at each corner is built as shown in Fig. 6. The propagation model used in the simulation is based on the reference propagation model of COST-231 urban that is the combination of typical logarithmic path loss model and Rayleigh fading model. For simplicity, reflection, diffraction and scattering effects are not taken into account. The area is divided into 20 × 20 elements with rectangular grids. For each grid element, there is a propagation feature that represents the shadowing variation in the urban environment. The shadowing feature in each grid is given by the mean of the shadowing variation deviation using the uniform distribution of (0, 1). The mean of the shadowing variation is -5, 0, 7, 15 and 25 dBm, respectively, as shown in Fig. 6.

Additionally, the MSs are uniformly distributed over the whole area and an equal number of sample MSs are selected from each grid element. Every MS can receive signal strength from the four BSs. The parameters of simulation configuration are given in Table 1. To verify whether the clusters represent the features of the topography, transmitter power is emitted from each BS to accommodate different physical situations. Two settings for BS transmit powers are applied: high power (48 dBm) and low power (40 dBm). Therefore, there are in total 2^4 possible combinations of power settings and the BSs. If the

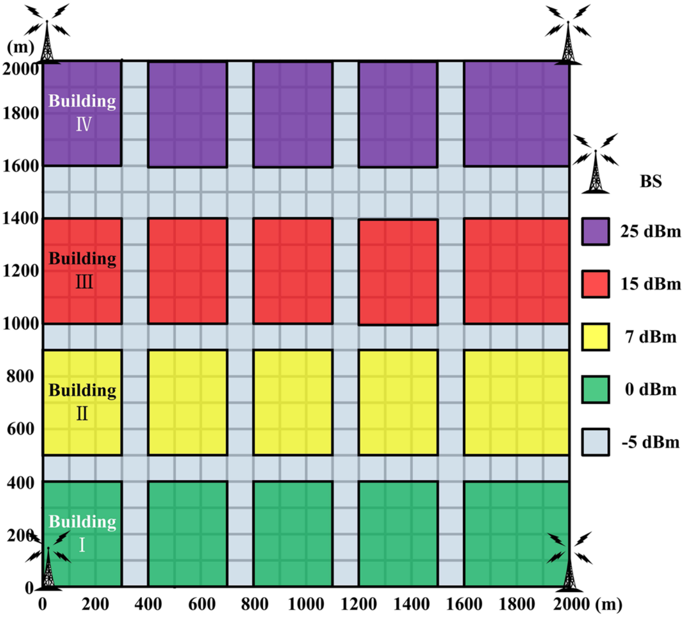


Fig. 6 Topography of the urban environment simulation

Table 1 Configuration parameters used in the simulation

System setting	
Distance between BS to BS	2.0 km
Minimum mobile-to-BS distance	20 m
Total number of MSs	3200 (8 MSs in each grid element)
Propagation Environment	
Minimum transmit power	40 dBm
Maximum transmit power	48 dBm
Shadowing deviation	
Building I	0 dBm
Building II	7 dBm
Building III	15 dBm
Building IV	25 dBm
Street	−5 dBm

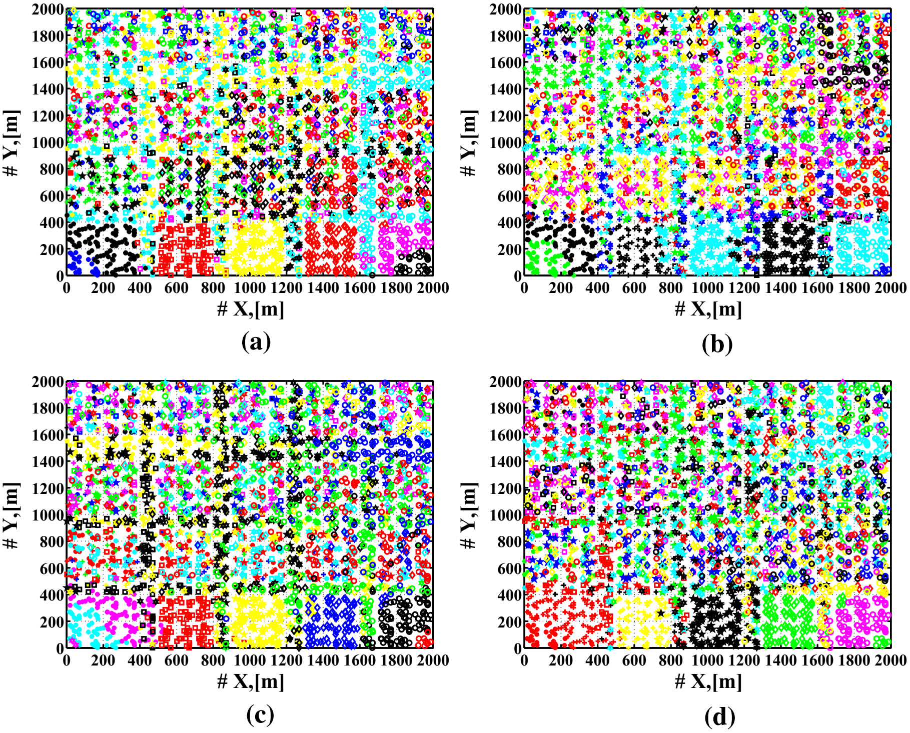


Fig. 7 Clustering results in scenario 1. **a** 0000, **b** 0110, **c** 1101, **d** 1111

high power setting is denoted as “1” and the low power setting as “0”, these 2^4 factorial experiments can be simply expressed as: 0000, 0001, ..., 1110, 1111.

In order to better analyse and compare the results of 16 settings, the locations of MSs are unchanged in all experiments. Figure 7 shows four examples of the results of clustering MSs based on deviation RSS over different powers at the four BSs. In the figure, different

Table 2 Comparison of estimation error between KNN and intersection methods based on cluster and grid Models in scenario 1 (in meters)

	Scenario 1: a simple simulated urban propagation model			
	KNN		Intersection	
	Cluster	Grid	Cluster	Grid
Mean error	145.7	186.7	144.8	161.2
Variance	127.9	174.2	142.5	152.1
50 percentile	107.8	119.0	97.3	116.0
75 percentile	196.4	279.6	199.2	224.4
90 percentile	319.1	446.1	335.4	361.0

colours represent different clusters and the cluster distribution can be seen to reflect the topological feature of the simulation area to some extent, especially for the places with a relatively small shadow variation. For the area with relatively large shadow variation, such as block IV in Fig. 6, the cluster distribution is scattered with respect to the geographical locations of the MSs, though in the four dimensional RSS space they are compact.

The number of clusters produced in every test is not exactly the same but quite close about 50 clusters. All the 16 results show that the distributions of produced clusters exhibit roughly the same structure as expected mathematically as a result of using the deviation RSS. Moreover, with unchanged parameters, each experiment has been tested many times and the results showed good stability in the clustering results. Although the simulated urban model used is simple, it can be further improved by analysing the azimuth and elevation power distribution of the transmission antenna to make sure whether this approach can be used in various scenarios in wireless networks.

Table 2 summarizes the information in terms of the mean, variance, 50th, 75th and 90th percentile values of the error distance for the intersection and KNN approaches based on cluster and grid partitions. Although the propagation model in scenario 1 is based on grid elements, the results show that the cluster-based positioning methods provide better accuracy than the grid-based partitioning. For example, 50 % of distance errors using the Cluster-Intersection scheme are within 97.3 m, whereas the Grid-Intersection scheme reports 116.0 m, i.e. a 19.7 m improvement.

4.1.2 Scenario 2: The Island of Jersey

The pilot signal strength data for the island of Jersey from network planning tool ASSET 3G is processed in this scenario. Figure 8 shows the topographic map of the centre of the island. There are six BSs covering an area of 8 km × 6 km and the clustering result is depicted in Fig. 9.

As can be seen from the results in Fig. 9, the clusters can generally represent the features of the current geographical patterns to a certain extent, particularly the contour of highways and roads. Combining with the results in scenario 1, it can be found that less shadowing variability in an area can result in more topographical features in clustering results.

This experiment has been tested several times as well and the clustering result is quite stable. Note that as many as 160 clusters are produced in the central area due to of the complex terrain. Despite this, considering the large number of test points and the

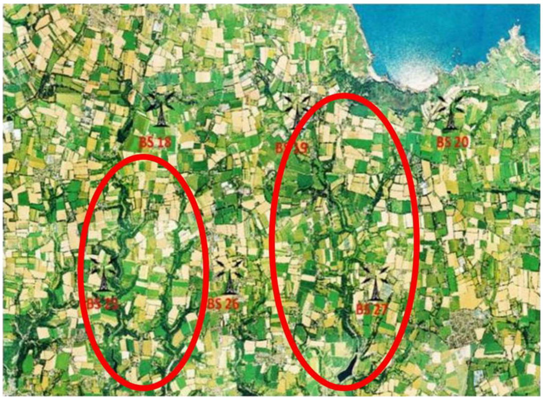


Fig. 8 The topography map in scenario 2

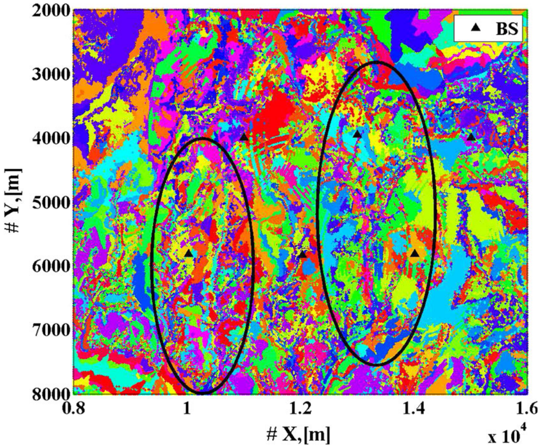


Fig. 9 Clustering result in scenario 2

complexity of the model, the result is quite inspiring since the topography features are well exhibited.

Comparisons of estimation errors between KNN and intersections methods based on cluster and grid models are given in Table 3. For the KNN methods, the results of cluster-base models show a significant improvement comparing with the grid-based models. The localisation accuracies of intersection methods are quite similar and much higher than those of KNN method. For example, the 90 % of distance errors of Cluster-KNN method are within 42.5 m, while for Cluster-Intersection method it reads 23.3 m. Overall, the proposed intersection method presents much higher accuracy in this rural scenario.

4.1.3 Scenario 3: Queen Mary Campus with Real Data

To test the proposed location estimation scheme in a real environment, the data from a GSM network is collected around the Queen Mary campus. The campus is in a city area with some high buildings nearby. The data was acquired from mobile application developed for an Android smartphone. Every second the application records the latitude and longitude of the current location from GPS, and collects the signal strengths from the

Table 3 Comparison of estimation error between KNN and intersection methods based on cluster and grid models in scenario 2 (in meters)

	Scenario 2: the Island of Jesery			
	KNN		Intersection	
	Cluster	Grid	Cluster	Grid
Mean error	17.8	26.0	10.5	10.5
Variance	31.3	18.0	15.6	15.7
50 percentile	7.8	22.3	5.3	5.2
75 percentile	17.3	30.6	11.1	11.0
90 percentile	42.5	44.0	23.3	23.4

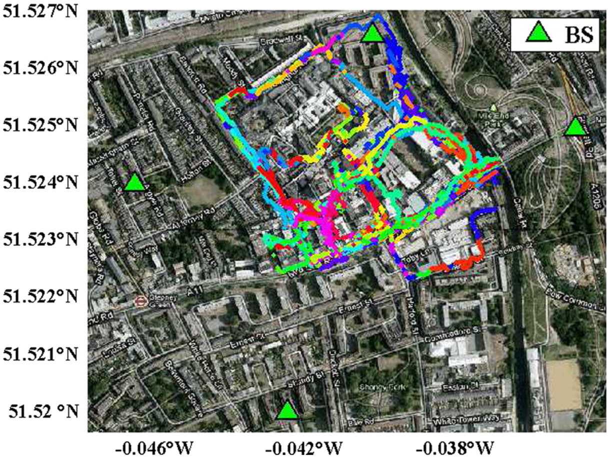


Fig. 10 Clustering results in scenario 3

Table 4 Comparison of estimation error between KNN and intersection methods based on cluster and grid models in scenario 3 (in meters)

	Scenario 3: Queen Mary campus			
	KNN		Intersection	
	Cluster	Grid	Cluster	Grid
Mean error	27.7	57.4	22.8	57.1
Variance	37.7	76.9	26.0	77.6
50 percentile	14.3	20.3	13.4	20.4
75 percentile	32.5	76.3	30.6	73.8
90 percentile	58.2	174.5	50.2	176.7

surrounding BSs. The locations of all the nearby BSs were obtained from the server of Sony Ericsson lab. Here we focus on the four nearest BSs. Figure 10 shows a map of Queen Mary campus that covers 475 m × 365 m. The colour-lines represents the result of clustering 9277 test points on different paths. The optimal number of clusters is 70.

Table 4 shows comparison of the results using the intersection and KNN methods based on grid and cluster model, respectively. As can be clearly seen that the intersection and KNN methods based on the cluster model outperforms these two methods based on the grid model. Particularly, the proposed Cluster-Intersection scheme also shows a better performance than the Cluster-KNN scheme.

4.2 RSS Deviations and Raw RSS

Comparing with using raw RSS in the calculation of RSS distance, using RSS deviations instead improves the estimation precision. This is because the similarity calculation using the raw RSS are dominated by the distance path loss rather than the topography. By comparison, clusters produced from the deviation data are a better reflection of the topography.

Figure 11 depicts the comparisons of clustering distribution between using the raw RSS and deviation RSS when the same number of clusters is created using the island of Jersey data. It is obvious that significant improved results is obtained in Fig. 11b by using the RSS deviations where the contour of the highway and road is well depicted.

Remarkable difference between the two results can be noticed in the vicinity of the BSs. In fact, if the topography were uniform, there would be only ring segments generated around the BSs in Fig. 11a. This would simply reflect the RSS attenuation with distance rather than topography. On the other hand, Fig. 11b illustrates that clusters produced using RSS deviations resulting from the observed path loss model which capture better the wireless topography especially in a complex environment than in Fig. 11a.

Another comparison is made using the Queen Mary data. The cumulative distribution function of the error distance for intersection method based on both RSS deviation and raw RSS data sets as shown in Fig. 12, under the premise that the same number of clusters is

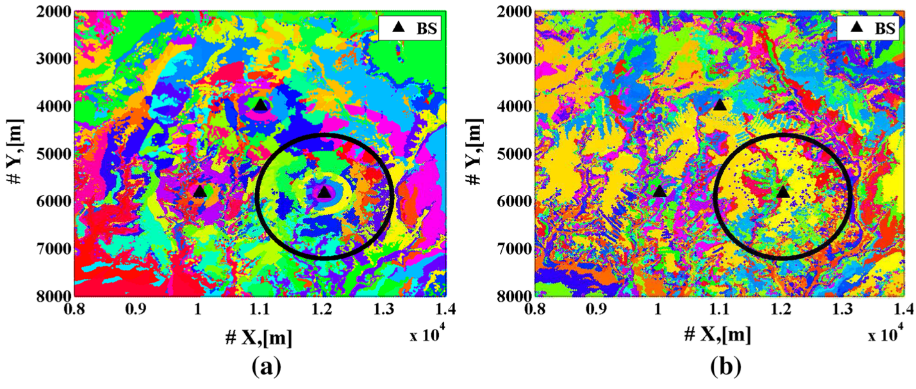
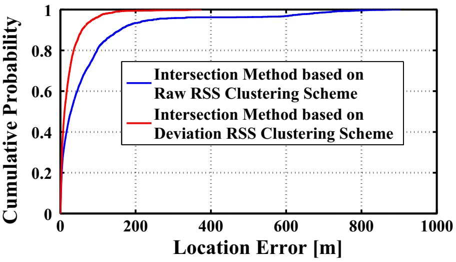


Fig. 11 The comparisons of clustering results between using the raw RSS and deviation RSS in the island of Jersey data. **a** Clustering based on raw RSS. **b** Clustering based on deviation RSS

Fig. 12 Location estimation results based on raw RSS and deviation RSS clustering scheme in Queen Mary data by using intersection approach



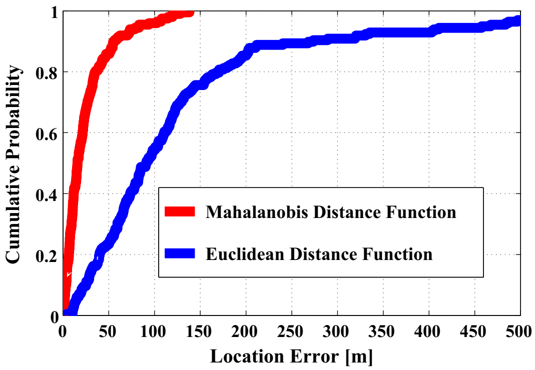


Fig. 13 Cumulative percentage of error for cluster-based intersection approach using Mahalanobis distance and Euclidean distance in Queen Mary scenario

produced. For the intersection method the mean values of the distance error using RSS deviations is 22.8 m, which is much more accurate than the result of using raw RSS data which is 72.6 m. Moreover, To achieve the same cumulative probability approaching 100 %, the location error of using RSS deviation is less than 200 m while the location error of using raw RSS data is about 700 m. This also indicates that the clustering using RSS deviation data can produce more accurate results than using raw RSS data.

4.3 Mahalanobis Distance and Euclidean Distance

The correlation between signal strength is very pervasive in the real environment, which impacts on the estimation accuracy. Therefore, eliminating these correlation information in the clustering needs to be considered. For this purpose, the Mahalanobis distance function can be applied to calculate the RSS similarity between any two MSs from different transmitters. As a result the high correlation between signal strength from different transmitters can be mitigated. Another advantage of using Mahalanobis is that it can automatically account for the scaling of the coordinate axes.

Figure 13 compares the cumulative probability with respect to the localisation error of the two distance functions. The results are generated by applying the Cluster-intersection method to the Queen Mary data, under the premise that the same number of clusters is created. It can be clearly observed that the proposed method using Mahalanobis distance function significantly outperforms that based on clustering using Euclidean distance. More specifically, the cumulative probability curve of using Mahalanobis distance function rises dramatically at small location errors and reaches 100 % at location error about 140 m. By comparison, the curve of Euclidean distance function grows relatively slow from beginning and becomes even slower after location error of 200 m. This indicates that the accuracy of the clustering identification of using Mahalanobis distance function is much higher than using Euclidean distance function. Thus, the Mahalanobis distance function could be used as an appropriate alternative to the Euclidean distance for location estimation in positioning systems.

5 Conclusion

In this paper, an outdoor location estimation scheme based on a cluster-based intersection fingerprinting technique using Received Signal Strength (RSS) has been proposed. The improved performance of the proposed scheme is demonstrated by making comparisons of results using three different outdoor data sets including numerically simulated data and read data. Several improvements have been made to the similarity calculation involved in the clustering. First the RSS deviations resulting from the observed path loss model which capture better the wireless topography in a complex environment is used in similarity calculation instead of raw RSS data. Second the Mahalanobis distance function is applied instead of Euclidean distance function to eliminate the correlation information in the clustering results. Combining with these improvements, our proposed clustering and intersection schemes provide good support for outdoor location estimation and, as a result, the accuracy of localisation is significantly improved.

References

1. Giorgetti, G., Gupta, S. K. S., & Manes, G. (2008). Localization using signal strength: To range or not to range?. In *Proceedings of the first ACM international workshop on Mobile entity localisation and tracking in GPS-less environments*, San Francisco, California, USA, pp. 91–96.
2. Dil, B. J., & Havinga, P. J. M. (2010). RSS-based localisation with different antenna orientations. In *Telecommunication networks and applications conference (ATNAC), Australasian*, Vol. 2010, pp. 13–18.
3. Bahl, P., & Padmanabhan, V. N. (2000). RADAR: an in-building RF-based user location andtracking system. In *INFOCOM 2000. Nineteenth annual joint conference of the IEEE computer and communications societies. Proceedings on IEEE*, pp. 775–784.
4. Kaemarungsi, K., & Krishnamurthy, P. (2004). Modeling of indoor positioning systems based on location fingerprinting, in INFOCOM. In *Twenty-third annual joint conference of the IEEE computer and communications societies*, Vol. 2004, pp. 1012–1022.
5. Kaemarungsi, K., & Krishnamurthy, P. (2004). Properties of indoor received signal strength for WLAN location fingerprinting. In *Mobile and ubiquitous systems: Networking and services*. MOBIQUITOUS 2004. The first annual international conference on, 2004, pp. 14–23.
6. Leppakoski, H., Tikkinen, S., & Takala, J. (2010). Optimizing radio map for WLAN fingerprinting. *Ubiquitous Positioning Indoor Navigation and Location Based Service (UPINLBS)*, 1–8.
7. Bshara, M., & Van Biesen, L. (2009). Localization in WiMAX networks depending on the available RSS-based measurements. *International Journal on Advances in Systems and Measurements*, *2*, 214–223.
8. Haeberlen, A., Flannery, E., Ladd, A. M., Rudys, A., Wallach, D. S., & Kavraki, L. E. (2004). Practical robust localisation over large-scale 802.11 wireless networks. In *Proceedings of the 10th annual international conference on Mobile computing and networking*, Philadelphia, PA, USA, pp. 70–84.
9. Youssef, M. A., Agrawala, A., & Udaya Shankar, A. (2003). WLAN location determination via clustering and probability distributions. In *Pervasive computing and communications, 2003. (PerCom 2003). Proceedings of the first IEEE international conference on*, pp. 143–150.
10. Ibrahim, M., & Youssef, M. (2010). Cell sense: A probabilistic RSSI-based GSM positioning system. In *Global telecommunications conference (GLOBECOM 2010), IEEE*, Vol. 2010, pp. 1–5.
11. Youssef, M., & Agrawala, A. (2005). The Horus WLAN location determination system. In *Proceedings of the 3rd international conference on mobile systems, applications, and services*, Seattle, Washington, pp. 205–218.
12. Yiqiang, C., Qiang, Y., Jie, Y., & Xiaoyong, C. (2006). Power-efficient access-point selection for indoor location estimation. *IEEE Transactions on Knowledge and Data Engineering*, *18*, 877–888.
13. Bahl, P., Padmanabhan, V. N., & Balachandran, A. (2000). Enhancements to the RADAR user location and tracking system, technical report, Microsoft Research.
14. Ni, L., Liu, Y., Lau, Y., & Patil, A. (2004). LANDMARC: Indoor location sensing using active RFID. *Wireless Networks*, *10*, 701–710.

15. Li, B., Salter, J., Dempster, A. G., & Rizos, C. (2006). Indoor positioning techniques based on wireless LAN. In *LAN, first IEEE international conference on wireless broadband and ultra wideband communications*.
16. Honkavirta, V., Perala, T., Ali-Loytty, S., & Piche, R. (2009). A comparative survey of WLAN location fingerprinting methods. In *Positioning, navigation and communication, 2009. 6th workshop on WPNC 2009*, pp. 243–251.
17. Lakmali, B. D. S., & Dias, D. (2008). Database correlation for GSM location in outdoor and indoor environments. In *Information and automation for sustainability, 2008. 4th international conference on ICIAFS 2008*, pp. 42–47.
18. Laitinen, H., Lahteenmaki, J., & Nordstrom, T. (2001). Database correlation method for GSM location. In *Vehicular technology conference, 2001. VTC 2001 Spring. IEEE VTS 53rd, 4*, pp. 2504–2508.
19. Roos, T., Myllymaki, P., Tirri, H., Misikangas, P., & Sievanen, J. (2002). A probabilistic approach to WLAN user location estimation. *International Journal of Wireless Information Networks*, *9*, 155–164.
20. Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science*, *315*, 972–976.
21. Vovk, V., Shafer, G., & Nouretdinov, I. (2004). Self-calibrating probability forecasting. *Advances in Neural Information Processing Systems, 16*, 1133–1140.
22. Goldsmith, A. (2005). *Wireless communication*. Cambridge: Cambridge university press.
23. Papoulis, A., & Pillai, S. U. (2002). *Probability, random variables, and stochastic processes*. Noida: Tata McGraw-Hill Education.
24. Yiming, J. (2010). Practical precision bound for indoor location determination. In *Computer and Information Application (ICCIA), International Conference on*, Vol. 2010, pp. 410–413.

Chao Ning received his B.S and M.Sc degrees in University of Electronic Science and Technology of China, in 2008 and 2011. He is currently a Ph.D. candidate in Imperial College London. His current research interests are primarily in data analysis and inverse problem.

Rui Li received his B.S, M.Sc and Ph.D. degrees in Chengdu University of Technology in 1982, 1985 and 1995, respectively. Since 1999 he has been a professor of earth science at Chengdu University of Technology.

Kejiong Li received her BS degree in University of Electronic Science and Technology of China in 2008, and Ph.D. degree in Queen Mary University of London in 2013. Her current research interests are primarily in indoor and outdoor localisation, data analysis and machine learning.