Localization of LTE Measurement Records with Missing Information

Paper Id: xxxx

***Abstract*—As cellular networks like 4G LTE networks get more and more sophisticated, mobiles also measure and send enormous amount of mobile measurement data (in TBs/week/metropoliton) during every call and session. The mobile measurement records are saved in data center for further analysis and mining, however, these measurement records are not geo-tagged because the measurement procedures are implemented in mobile LTE stack. Geo-tagging (or localizing) the stored measurement record is a fundamental building block towards network analytics and troubleshooting since the measurement records contain rich information on call quality, latency, throughput, signal quality, error codes etc. In this work, our goal is to localize these mobile measurement records. Precisely, we answer the following question: *what was the location of the mobile when it sent a given measurement record?* We design and implement novel machine learning based algorithms to infer whether a mobile was outdoor and if so, it infers the latitude-longitude associated with the measurement record. The key technical challenge comes from the fact that measurement records do not contain sufficient information required for triangulation or RF fingerprinting based techniques to work by themselves. Experiments performed with real data sets from an operational 4G network in a major metropolitan show that, the median accuracy of our proposed solution is around 20 m for outdoor mobiles and outdoor classification accuracy is more than 98%.**

I. Introduction

As cellular technologies evolve from 4G LTE to 5G, these networks are becoming increasingly difficult to manage and troubleshoot. This is due to heterogeneity of cells and unprecedented network complexity and scale. Indeed, in a large metropolitan city like New York City, the cellular network of an operator can easily have thousands of cells combining macro and diverse small cell technologies like metro cells, distributed antenna systems etc; also, these cells are configured through tens of thousands of critical network wide parameters. While management, analytics, and diagnostics of such a complex network is of paramount importance, but it is also a significant challenge. Towards this end, network providers are moving towards network measurement data driven network analytics and measurements.

In modern cellular systems, enormous amount of mobile measurement data is collected during each call/session. The measurements are typically performed by mobiles and sent back to the network and eventually saved in the data center (see Fugure 1 in Section III). Since this data is named differently by different telecom vendors, in this paper we will refer to this measurement data as LUMD (*LTE UE Measurement Data*). LUMD has rich information on mobile's performance metrics like throughput, latency, call drop during previous session etc. and also on RF metrics like signal strength (RSRP) and SINR. The measurement records are sent on a per-procedure basis (e.g., service request, session beginning, attach etc.) and also on a network defined event basis (e.g., relative signal strength of neighboring cell etc.). See Section III for more details. However, since the mobile measurement records are sent from mobile stack that do not have access to application layer, the mobile measurement records are not geo-tagged. Absence of latitude and longitude in LUMD poses challenges in LUMD based use-cases that rely on the location of the mobiles. These applications could range from identify exact locations of poor coverage to troubleshooting temporary call failures. Thus geo-tagging the measurement records is a fundamental building block towards location dependent network analytics and diagnostics.

The goal of this work is achieve the following objectives: *(i)* infer whether any measurement record was generated from an indoor mobile or from an outdoor mobile, and *(ii)* if the measurement record was generated from an outdoor mobile, estimate the latitude-longitude of the mobile when the measurement record was generated. In this work, we develop machine learning algorithms and present experiment results from a real data set from an operation 4G LTE deployment.

In principle, localizing LUMD could be done using *triangulation* based localization principles used in GPS devices or in mobile applications so long as measurement records contain signal strength related information from three or more cells. However, in LUMD records, many of the signal strengths of neighboring cells are missing. LUMD typically has signal strength (RSRP) from at most two cells: the serving cell and the strongest neighboring cell. In fact, many LUMD records are equipped with RSRP from only the serving cell. Since triangulation based approaches require signal strength from at least three and ideally four or more cells, the problem of LUMD localization is essentially a problem of localization with missing cell-strength information. Another approach would be to first create an RF figer-print at different locations using training data and then use the figer-prints to infer the location of measurement records. However, this approach too suffers from a similar problem of having not sufficient cell signals per location.

A. Approach and Contributions

Our approach: Every location in a wireless coverage area is characterized by unique RF fingerprint due to the presence of signals from multiple transmitters. Triangulation and RF fingerprinting uses different techniques to infer the location for a given RF signature. In our problem, since measurement

records contain one or two cell signal information, this can be viewed as projection of RF signature into a lower dimensional plane and this makes inferring locations challenging. To mitigate this problem, we stitch together multiple measurement records from a mobile to create a time-series of these RF signatures. Our approach essentially combines localization principles based on RF fingerprinting and probabilistic path-tracking used for robot localization. At a high level, our approach has two steps for localizing measurement records from outdoor mobiles:

1) Instead of viewing each LUMD record in isolation, for each mobile, we *stitch* together LUMD records from that mobile over a "session duration" and model it as a suitable Markovian time series. The problem now reduces to identifying locations (states) of the entire path of the mobile.
2) The above solution method assumes that the probabilities characterizing the underlying Markovian structure can be learned. We perform supervised learning to estimate these probabilities. The training data for supervised learning comes from drive test carried out by network providers.

The details of the above two steps are provided in Section V. The rationale behind localizing the path taken by a mobile is two-fold: first, localization accuracy of the individual points can be improved if there is a nearby point that is more accurately localized; and second, we also make use the road network to constrain points to lie on the road whenever the mobile is moving.

Our Contributions: We outline our main contributions as follows:

1) *Novel framework:* To the best of our knowledge, ours is the first work to develop a systematic study of geo-tagging mobile measurement records in modern cellular systems while tackling the challenges posed by insufficient cell signal information that would be required by triangulation and vanilla RF figer-printing based approach.
2) *Algorithms:* We develop novel machine learning algorithms by combining elements from supervised learning based RF figer-printing and particle-filter based Hidden Markov Model learning used for robot path-tracking. We also present how standard machine learning algorithms can be adapted for indoo-outdoor classification of mobile measurement records.
3) *Evaluation:* We present experimental results using real data-set from an operational 4G LTE network in a major metropolitan to show the efficacy of our design. Our results show a median location accuracy of around 20 m whereas indoor-outdoor classification accuracy is more than 98%.

The rest of the paper is organized as follows. Section II provides an overview of related work abd Section III provides some background and introduces relevant terminologies. Section IV presents the problem setting and states the precise localization problem. Section V presents the main localization algorithm and Section VI describes how measurement records can be classified as indoor or outdoor. We present experimental validation in Section VII and finally we conclude in Section VIII.

II. Related Work

Though our work is on localizing measurement records (i.e., estimating mobile location when measurement was generated) and most of the localization work in the literature is on localizing devices, the objectives are similar. In the following, we highlight some of the work on localization in wireless systems and point out the main difference in our work.

Much of the work on localization in wireless networks is for indoor localization. Two of the popular techniques are triangulation and RF fingerprinting. In triangulation based localization, geometric principles are used to localize a device based on signals from multiple (more than 3 typicall) access points. Some of works on triangulation based localization using wireless signals are [8], [16], [17]. On the other hand, fingerprinting based localization techniques uses training data to create RF fingerprint of the area and this fingerprint is used to localize devices. See [2], [6], [9] for some of works on RF fingerprinting based localization. In [21], the authors propose unsupervised learning method for highly accurate indoor localization without the need for training data; also see [7], [14]. Among other works, [13] proposes RFId based indoor localization and [3] describes Bluetooth based indoor localization. However, the common thread in all these works is that, signals from multiple (more than two mostly) wireless access points are available and thus creating a unique signature at different locations. In our problem, since measurement records have no more than two signal strength information at, the records can be viewed as carrying a lower dimensional projection of the unique multi-dimensional RF signature.

To overcome the above problem, our work creates a time-series of measurement records from the same mobile which increases lifts the RF signals of a mobile into a higher dimension. Once that is performed, principles from robot localization can be used. Localizing robot paths is an extensively studied research area. In such a problem, at different locations, some noisy version of robot's state (location, velocity, accelerometer reading et.) is observed, and the goal is to estimate the correct state from sequence of noisy state observations. In [18], the authors provide an excellent survey of research in this rich area. A more detailed treatment of different techniques for robot localization is there in the book [19]. [10] provides an excellent survey of another closely related problem in the field: simultaneous localization and mapping. However, the fundamental difference between localizing robot path and localizing sequence of measurement records from a mobile is that, unlike robot measurements, cellular measurement records do not contain a noisy version of the state (location and velocity), rather contains RF information which has some unknown dependence of location.

Work on outdoor localization using cellular networks has been very few primarily because most smart phones carry GPS capability. The only work that attempts to localize measurement records in 3G context is! [11]. The use of mobile measurement records have also been looked at [22] to improve paging efficiency and recently in [12] for fast measurement analytics.

III. Background and Terminologies

A. Relevant 4G LTE Terminologies

Though our techniques could apply to any future cellular system, we use LTE terminologies for convenience. The terminologies [15] relevant for our purpose are described below.

UE (user equipment): UE refers to the mobile end-device.

Cell: In LTE networks, a cell refers to coverage footprint of a base station transmitter typically ensuring a cell coverage radius around 0.5 km-5 km. In LTE macro cells, each cell typically has a directional base-station transmitter with 120° sectorized antennas.

eNodeB (eNB): The eNB is the network element that interfaces with the UE and hosts critical protocol layers like PHY, MAC, and Radio Link Control (RLC) etc. Each eNB typically has 3 base station transmitters with 120° antennas.

Reference Signal Received Power (RSRP): In LTE networks, UEs make certain measurements of received signal strength for each nearby cell transmitter. RSRP is the total measured time-average received power at UE of all downlink reference signals across the entire bandwidth from a *given cell transmitter*. RSRP is a measure of the received signal strength of a cell transmitter at a UE.

RSSI and RSRQ: RSSI (Received Signal Strength Indicator) is the total measured received power at the UE over the entire band of operation from *all cell transmitters.* RSRQ of a given cell transmitter at a UE is RSSI scaled by average RSRP (of that cell) per reference symbol.

B. LTE UE Measurement Data (LUMD)

In 4G LTE networks, during each session and call, mobile related measurement data is collected by the network [15]. This measurement is referred to by different names by different network vendors, for e.g., Alcatel-Lucent based systems refer to this as *Per Call Measurement Data* (PCMD), Ericsson systems refer to this as *General Performance Event Handling* (GPEH), Nokia systems refer to this as Megamon. In this paper, we will refer to these measurements as LUMD which is an abbreviation for LTE UE Measurement Data.

LUMD provided call/session measurement data is eseentially a view of the user experience within LTE system. The measurements are either procedure based and event based. Procedure based measurements are sent from mobiles when certain pre-defined procedures take place, for e.g., attach, detach, hand-off, session end, session initiation etc. Event based measurements are sent from UE, when certain standards defined pre-defined events occur. Few useful events (defined by standards [1]) for our purpose are as follows:

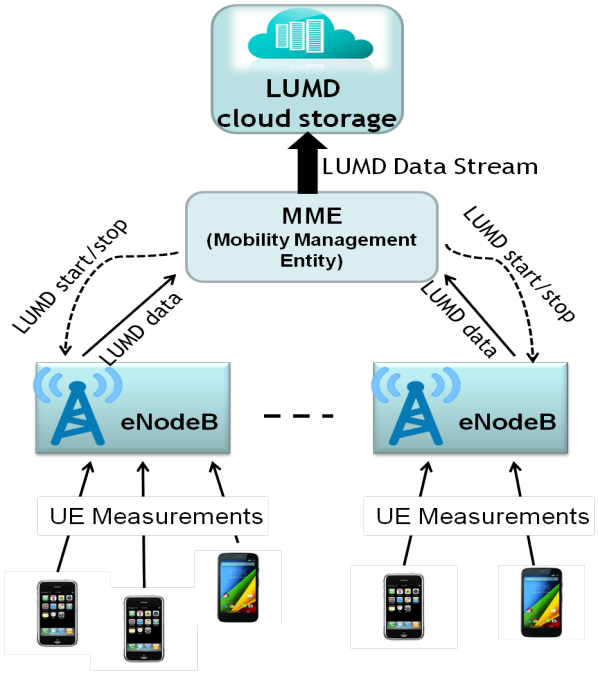


Fig. 1. LUMD Data Collection.

1) *A1 and A2:* A1 (resp. A2) event are triggered at UE when serving cell RSRP becomes better (resp. worse) than a network defined threshold (dBm).
2) *A3:* A3 event is triggered at UE when a neighbor cell RSRP becomes better than serving cell RSRP by an offset (in dB) specified by the network.
3) *A4:* A4 event is triggered at UE when a neighbor cell RSRP becomes better than a network defined threshold (dBm).

Measurement data collection architecture: The LUMD data collection architecture is shown in Figure 1. LUMD is collected at both the eNodeB and MME (Mobility Management Entity). The MME serves as the coordinator of the LUMD data. After LUMD collection is turned on at the eNodeB, it collects the records and sends the data to the MME. MME aggregates and temporarily saves LUMD from multiple eNodeBs and sends it periodically (typically in minutes time-scale) to the data center where LUMD is saved and analyzed. Scalable storage of LUMD, which can easily run into TB in a week per metropolitan, in the data center is an important design problem and beyond the scope of this paper.

Contents of LUMD: LUMD record contains data related to signaling performance on per UE, per bearer level for different procedures, user experience such as data throughput and procedure duration, eNodeB internal UE related data such as MIMO decision, SINR, buffer size, and normalized power headroom etc. What information is present depends on procedure/event that led to the measurement record. For our purpose, we are interested in RF information contained in measurement records. These are RSRP and RSRQ information. A LUMD

record contains the following RF information:

- *RSRP:* Most LUMD contain RSRP of the serving cell that a UE is associated with. In addition, only when LUMD is generated due to an A3 or A4 event as described earlier in this section, it might also contain the RSRP of *one* neighboring cell (typically the strongest one).
- *RSRQ:* LUMD also contains RSRQ of the serving cell. Note that once RSRP and RSRQ are known, the corresponding RSSI can be uniquely computed since RSRQ is defined as RSSI scaled by RSRP per reference symbol.

The important thing to note is that RSRP and RSRQ information is available from no more than two cells in an LUMD record.

IV. Problem Statement

Consider an LTE network with K cells (we refer to each cell transmitter as a base station) located at positions $\{y_k\}_{k=1}^K$. In general, the locations of the cells, $\{y_k\}_{k=1}^K$, may or may not be known.

Mobiles travel along a road network represented by a graph $G_r = (V, E)$ where V denotes graph nodes represented by a latitude-longitude tuple and E denotes valid direct path between two nodes.

There are two two types of data relevant to our discussion:

1) **Training data in the form of *drive test data*:** This is essentially geo-tagged data sent from a set of locations in the road graph nodes V. Precisely, we are given n locations $\{x_i\}_{i=1}^n$ and for each location, RSRP of multiple cells. We denote by $\{R_{i,k}\}_{i=1}^n$ the signal strength of cell-k sent from training location x_i. Note that, for a location x_i the data $R_{i,k}$ is only available for a small subset of cells near location x_i. We will also denote the set of training data by $\mathcal{D}_{tr}$.
2) **LUMD data or observed data:** This data is not geo-tagged but comes with time stamp. Precisely, for every mobile, we are given time instants $t_i, i = 1, 2, \ldots, T$ for each t_i we are also given RSRP $\tilde{R}_k(t_i)$ where $k \in K(t_i)$; $K(t_i)$ denotes the set of cells reported by the mobile at time t. Typically $|K(t_i)|$ takes value one or two. Though we have LUMD for each mobile-m, we drop the dependence of m on $R_k(t)$ and $K(t)$ as we are essentially perform the same algorithm for each mobile separately. The locations of mobiles $\tilde{x}(t_i)$ at different times t_i are unknown.

Thus the problem can be succinctly stated as follows:

Problem of localization with missing RSRPs: We are given training data consisting of locations $\{x_i\}_{i=1}^n$ and associated RSRPs $\{R_{i,k}\}_{i=1}^n$ of cell-i at location x_i. Estimate the unknown location of a sequence of measurements $\tilde{R}_k(t_i)$ where $i = 1, 2, \ldots, m,\ k \in K(t_i)$. Assume that the locations are drawn from locations in a road network given by $G_r = (V, E)$.

Remark 1. *We make three important remarks:*

1) **Pre-processing LUMD:** *In our problem statement, we have assumed that LUMD from a UE forms a time-series of measurements. In practice, the available LUMD records are available as discrete records that carry mobile identifiers (IMSI) and time-stamps of measurements. Thus, a pre-processing step is required to stitch together multiple LUMD records from a UE to create time-series LUMD sequence from each UE during a session.*
2) **Randomness of RSRP:** *The RSRPs denoted by $R_{i,k}$s are random variables. To this end, in the drive test data there could be more than one x_i representing the same location but each reporting different samples from the underlying random variable. Similarly, the RSRPs in LUMD is also a realization of the random variable.*
3) **Accounting for RSRQ:** *For ease of exposition, we describe our model and algorithm based on observation of RSRPs of different cells. In LTE, in addition to RSRP, mobiles could also report another quantity known as RSRQ (see Section III). This additional information cane be easily incorporated into our model and algorithm by treating $R_{i,k}$'s and $\tilde{R}_k(t_i)$'s as vector of RSRP and RSRQ measurements. Note that, our software implementation does take RSRP and RSRQ both into account and so do the presented results.*

V. Learning Algorithms

A. Hidden Markov Model

We now describe our model for learning. First we represent the motion of the mobile and the observed LUMD data using suitable Hidden Markov Model (HMM). In HMM, the system (mobile) moves from one hidden state (location etc.) to another and in each state certain observations (RSRP etc) are made. The goal is to infer the hidden state from the observations based on prior knowledge about the transition probabilities between hidden states and observations in the states.

Denote by $\tilde{x}(t)$ the unknown locations of the mobile. Then the sequence of locations $\{\tilde{x}(t),\ t = 1, 2, \ldots$ of in the graph G_r form a Markov chain. Given the outputs $\{\tilde{R}(t)\}_t = 1, 2, \ldots,$ of the HMM we infer the hidden states $\{\tilde{x}_i\}_{i=1}^m$ using different filtering algorithms.

We now describe our HMM in more detail:

- *Hidden States:* The hidden states of the HMM are the location and the velocity of the mobile. At time-t, the HMM has state described by $(\tilde{x}(t), v(t))$.
- *State transition probabilities and mobility model:* These transition probabilities model how transition happens from one hidden state to another. We assume that that, given the previous location and velocity, the current location and the velocity are statistically independent, i.e.,

$$\begin{aligned} &p(\tilde{x}(t_i), v(t_i) \mid \tilde{x}(t_{i-1}), v(t_{i-1})) \\ &= p(\tilde{x}(t_i) \mid \tilde{x}(t_{i-1}), v(t_{i-1})) \\ &\quad \times p(v(t_i) \mid \tilde{x}(t_{i-1}), v(t_{i-1}))\ . \end{aligned}$$

We assume a suitable mobility model of the mobile which determines how it moves along the graph G_r and

also helps us to calculate the above probabilities. We assume that the mobile updates its speed according to the following equation.

$$v(t_i) = e^{-\beta\tau} v(t_{i-1}) + (1 - e^{-\beta\tau})\mathcal{N}(\mu, \sigma^2) \quad (1)$$

where $\tau = t_i - t_{i-1}$, $\mathcal{N}(\mu, \sigma^2)$ is initial velocity distribution and β is a scaling constant. Let $d_i = v(t_i) \times \tau$ and x_i be a point on the graph G such that $d_G(x_i, x_{i-1}) = d_i$ along a path (if there is no such point we round d_i to the nearest such point). Under this mobility model the likelihood $p(\hat{x}_i|\hat{x}_{i-1}, \hat{v}_{i-1}, G)$ is given by

$$\begin{aligned} &p(\hat{x}(t_i)|\hat{x}(t_{i-1}), v(t_{i-1})) \\ &= \frac{1}{C\sigma_\tau(1 - e^{-\beta\tau})\tau\sqrt{2\pi}} \\ &\times \exp\left(-\frac{(d_G(\tilde{x}(t_i), \tilde{x}(t_{i-1})) - \tau v(t_{i-1})e^{-\beta\tau})^2}{2\sigma_\tau^2\tau^2(1 - e^{-\beta\tau})^2}\right), \end{aligned} \quad (2)$$

where C is a normalizing constant. The expressions in (1) and (2) completely describe the transition probabilities between the hidden states. Note that the mean μ and variance σ^2 of the initial speed is ideally based on city road under consideration and the associated speed limits.

- *Observations in states:* The LUMD records at different states (locations) represent the observations of HMM model. The probability distribution (also called the likelihood function) of an observation (LUMD record) conditioned on a location is denoted by $p(\tilde{R}_i|\hat{x}_i)$. In our approach, these probabilities can be learnt from the drive test data using regression on drive test data to estimate $p(\tilde{R}_i|\hat{x}_i)$. This is outlined in Section V-C.

B. Particle Filter Based Localization

The optimal MAP solution to this localization problem is given as follows.

$$\begin{aligned} &\{\hat{x}(t_i)\}_{i=1}^m \\ &= \arg \max_{\{x(t_i)\}_{i=1}^m \in V^m} \Pr(\{x(t_i)\}_{i=1}^m | \{\tilde{R}(t_i)\}_{i=1}^m, G_r, \mathcal{D}_{tr}) \end{aligned} \quad (3)$$

Solving (3) exactly requires a complexity of $O(|V|^m)$ which is computationally infeasible since the size of the graph G is very large. Hence we need to design more efficient algorithms.

In particle filter based localization algorithm $LocalizeUEpf$ we maintain set of N particles and their corresponding weights or likelihoods where each particle represents a sequence of possible location of the UE. Recall that t_i denotes the time at which the UE sends the LUMD record $\tilde{R}_i$ and v_i is the speed of the UE at time t_i. Let $d_G(x, y)$ be the shortest distance between points $x, y \in V$ calculated along the edges of the graph G. The pseudocode is presented in Algorithm 1. N_{th} is a non-degeneracy parameter input which determines when less probable particles are to be discarded.

Algorithm 1 $LocalizeUEpf(\mathcal{D}_{tr}, \mathcal{C}, G, N_{th})$

1: Sample N particles $\mathcal{P}_j = \{\hat{x}_1^{(j)}, \hat{v}_1^{(j)}\}$, $j = 1, \ldots, N$ from prior distribution $p(\tilde{x}_1, \tilde{v}_1|G)$
2: Initialize importance weights $\hat{w}_1^{(j)} \leftarrow p(\tilde{R}_1|\hat{x}_1^{(j)}, \mathcal{C})$, $j = 1, \ldots, N$
3: Normalize $w_1^{(j)} \leftarrow \hat{w}_1^{(j)} / \sum_{l=1}^N \hat{w}_1^{(l)}$, $j = 1, \ldots, N$
4: **for** $i = 2$ to m **do**
5: **for** $j = 1$ to N **do**
6: Sample $\hat{x}_i^{(j)}$ from distribution $p(\hat{x}_i^{(j)}|\hat{x}_{i-1}^{(j)}, \hat{v}_{i-1}^{(j)}, G)$
7: Update weight $\hat{w}_i^{(j)} \leftarrow \hat{w}_{i-1}^{(j)} \times p(\tilde{R}_i|\hat{x}_i^{(j)}, \mathcal{C})$
8: Update speed $\hat{v}_i^{(j)} = d_G(\hat{x}_i^{(j)}, \hat{x}_{i-1}^{(j)})/(T_i - T_{i-1})$
9: $\mathcal{P}_j \leftarrow \mathcal{P}_j \cup \{\hat{x}_i^{(j)}, \hat{v}_i^{(j)}\}$
10: **end for**
11: Normalize $w_i^{(j)} \leftarrow \frac{\hat{w}_i^{(j)}}{\sum_{l=1}^N \hat{w}_i^{(l)}}$
12: $\hat{N}_{eff} \leftarrow \frac{1}{\sum_{l=1}^N (w_i^{(l)})^2}$
13: **if** $\hat{N}_{eff} < N_{th}$ **then**
14: Sample N particles with replacement from current particle set $\{\mathcal{P}_j\}_{j=1}^N$ with probabilities $\{\hat{w}_i^{(j)}\}_{j=1}^N$. Update particle set with the new sampled set
15: $w_i^{(j)} \leftarrow \frac{1}{N}$ for $j = 1, \ldots, N$
16: **end if**
17: **end for**
18: $j^* = \arg\max_{j=1,\ldots,N} w_m^{(j)}$
19: Output location estimate $\{\hat{x}_i^{(j^*)}\}_{i=1}^m$
20: Output distribution $p(\{\hat{x}^{(j)}\}_{i=1}^m | \{\tilde{R}_i\}_{i=1}^m, \mathcal{C}, G) = w_m^{(j)}$ for $j = 1$ to N

There are a couple of important considerations in the actual implementation of our algorithm as we note below.

- *Choice of prior distribution:* The prior distribution $p(\tilde{x}_1, \tilde{v}_1|G)$ can be critical to the performance of the algorithm. There are two options. In the first option, the prior can be uniform over the cell coverage area but constrained to lie on the road graph. The cell coverage area can be obtained from the drive test area or from estimate of cell radius of the cell assuming the cell site location is known apriori. In the second option, location likelihood from some other algorithm (typically based on geometric principles) can be used as prior.
- *Robust PF based algorithm:* We also make our PF algorithm robust by doing several runs, deleting the outliers, and averaging the rest. For detecting the outliers, the output of the K runs are *clustered* using a unsupervised clustering algorithm like *affinity propagation*, then each cluster is assigned a probability based on the likelihood of observations, and the cluster with highest probability is taken as representative of accurate estimates.

Runtime: The runtime of the $LocalizeUEpf$ is bounded

as $O(mN)$, where m is the length of the LUMD sequence and N is the number of particles.

C. Regression based Observation Likelihood

Our goal is to estimate the probability of a observing an RSRP given a location. To achieve this, we resort to Random Forest based regression [5] on the drive test data. The rationale behind choosing Random Forest is as follows: first, the drive test data is spread over a non-contiguous location because coverage areas in a cell are not necessarily connected. Secondly, wireless RSRP manifests quite different properties in different locations and Random Forest is ideal for automatically segmenting an area into locations where the RSRPs exhibit strong spatial correlation.

The regressions steps are as follows:

1) For each location and each base station that can be heard at that location, we take the empirical mean and standard deviation of all corresponding drive test data RSRP.
2) For each cell, model the spatial variation of RSRP-statistics (i.e., mean and standard deviation) using *Random Forest* where the latitude and the longitude are taken as features of the model and the RSRP-statistic of the cell is the output. Each such *Random Forest* is trained using data aggregated in previous step. Also, compute the mean square error (or *cross validation error*) for each random forest.
3) Denote by $RndFrst_m(x,c)$ ($RndFrst_s(x,c)$) the random forest predictor of mean (standard deviation) of RSRP for cell-c at location x. Let $(\sigma_{RF}(c))^2$ be the corresponding mean square error of the predictor. Then we model

$$p(\tilde{R}_i|\hat{x}_i) = \mathcal{N}(RndFrst(\hat{x}_i,c), \sigma_c^2(\hat{x}_i)) \ , \qquad (4)$$

where

$$\sigma_c^2(x) = RndFrst_s(x,c) + \sigma_{RF}^2(c) \ ,$$

and the serving cell-c can be obtained from the LUMD record. In general, we can choose any spatial regressor instead of random forest. However, choosing random forest makes the model robust to cell propogation properties and to the fact that the coverage area of the cell could be disjoint.

This regression procedure can be repeated for other RF measurements like RSRQ as well.

VI. Indoor Outdoor Classification

So far we have provided a scheme for localizing measurement records under the assumption that the measurement records are generated from an outdoor mobile. In practice, mobiles can be outdoor as well as indoor. In the following, we adapt standard machine learning techniques to infer whether a measurement record was generated from an indoor mobile or an outdoor mobile. Our main contribution is in showing that the combination of RSRP and RSSI provide excellent feature for indoor-outdoor classification of measurement records. Here we show how SVM based classifier is easily appicable to our problem, the results are presented in Section **??**.

We set-up the indoor-outdoor classification as a supervised learning problem. The training data for this problem consists of two sets of data that is usually collected by mobile operators: *(i) walk test* of the cellular network that is carried out at different indoor locations within the network, and *(ii) drive test* of the cellular network using a vehicle in different city streets. In the following, we precisely describe the training data set that can be easily created out the above walk and drive test data.

Training data for classification: The training data $\mathcal{T}$ is a set of 3-tuple $(R_i, S_i, z_i),\ i = 1, 2, \ldots n$ where R_i denotes a RSRP measurement, S_i denotes RSSI measurement, and z_i is a $0-1$ variable that takes value 1 iff the RSRP-RSSI pair has come from indoor (walk test) else it takes value 0 (i.e., the record has come from a drive test). Note that, we consider RSSI values instead of RSRQ measurements because RSRQ is simply a scaled version of RSSI which is uniquely retrievable from RSRQ and RSRP.

Indoor-Outdoor classification problem: Given above training data set $\mathcal{T}$ and m LUMD records containing RSRP and RSSI information $(R'_i, S'_i), i = 1, 2, \ldots m$, infer whether z'_i is one or zero (indoor or outdoor) for each LUMD record.

This is a classical supervised learning based classification problem where the drive/walk test data is training data and LUMD contents are test data set. This problem can be solved using the following steps:

1) Use SVM[1] (Suppoer Vector Machine) [4] based classifier to fit a function f_{svm} $f_{svm} : (R_i, S_i) \rightarrow z_i \in \{0, 1\}$ where f is trained using the training data set $\mathcal{T}$.
2) For each LUMD record, predict $\hat{z}_i = f_{svm}(R'_i, S'_i)$.
3) (Optional) Using $\hat{z}_i$, concatenate $(R'_i, S'_i, \hat{z}_i)$ with training data set $\mathcal{T}$ to form a new training data set $\mathcal{T}'$. Re-train SVM based classifier f_{svm} using the new training set $\mathcal{T}'$.
4) (Optional) Repeat the last two steps till distance between $(\hat{z}_1, \hat{z}_2, \ldots \hat{z}_m)$ in two successive iteration is smaller than a pre-defined threshold ϵ, say 0.001.

The last two optional steps improve the accuracy of classification especially when LUMD records can be classified in a batch. This technique is also known an semi-supervised learning [4].

Putting it all together: The classification along with the localization scheme can be put together as follows:

1) Use the drive test data to train the condition probabilities of underlying HMM as per regression based scheme in Section V-C.
2) Use drive and walk test data to train the SVM based indoor-outdoor classification engine as described in this section.
3) For each LUMD record, use the classification scheme in this section to infer which records came from indoor and which came from outdoor locations.

[1]

4) For the outdoor LUMD records, first create LUMD sequence for each UE and then use Algorithm 1 to estimate the latitude-longitude of the mobile when the record was generated.

VII. Evaluation

In this section, we present evaluation of our proposed technique. The objective of our evaluation is three folds: to understand the accuracy of our localization scheme, to evaluate how much the accuracy depends of fraction of network coverage area that is drive tested, and to evaluate the extent to which RSRP-RSSI pair serves as good feature set for indoor-outdoor classification.

A. Methodology

We validate of our solution with measurement data collected from the LTE deployment of a top-3 service provider in US in the area of Chelsea in New York City (see map in Figure 2). In this area, we use the drive test data (from outdoor locations) to validate our results; we also make use of walk test data from the same location for evaluating indoor-outdoor classification. Note that, the drive test data is used for training the Hidden Markov Model in Algorithm 1. In practice, once the HMM is trained, LUMD records from UE can be localized using our techniques, however, we will not be able to validate the accuracy as we do not have access to ground truths, i.e., actual mobile location from which the LUMD records were sent. Thus, to validate our approach, we divided the drive test data locations into two sets as follows:

- *Training locations:* A random chosen subset of drive test data locations were chosen as *training locations* and all drive test data from these training locations were chosen to train our HMM probabilities.
- *Test locations:* The drive test data locations that were not part of training locations were chosen as *test locations*. In addition, we also allow for some randomly selected locations (a small fraction) to be part of test locations.

Synthesizing test LUMD records: The drive test data at test locations include much more information than LUMD record that would be generated at those locations. To exactly mimic LUMD record that would be generated at the test locations, we perform the following steps for each test location to synthesize LUMD record (we only synthesize the contents relevant for our purpose):

1) Find the serving cell in the drive test data and include the RSRP and RSRQ of the serving cell in and RSRQ.
2) Verify if the strongest RSRP of any non-serving cell satisfies A3 or A4 event condition (see Section III) and if so, include the RSRP and RSRQ of that neighbor cell in the LUMD record. Note that, we strip of any location information from the LUMD record, however we separately maintain it simply to compare with estimated location from the thus created LUMD data.

Once LUMD records are generated at each location, we synthesize LUMD records for a moving user using the following steps: *(i)* start at a random location in the street map, select LUMD record at this location based on LUMDs created at this location, *(ii)* a new next location of user is generated by sampling the nearby test locations probabilistically where the probabilities are that of user moving to the new location based on velocity distribution given by (1) and a fixed travel time of 10s, *(iii)* generate an LUMD record at this new location based on the LUMD records generated at each location, *(iv)* repeat previous three steps till no new location can be found in map (due to absence of nearby test location) or number of LUMD records reaches a threshold (chosen as 6 since we have rarely observed more than $6-8$ LUMD records from one user during a session in real LUMD data).

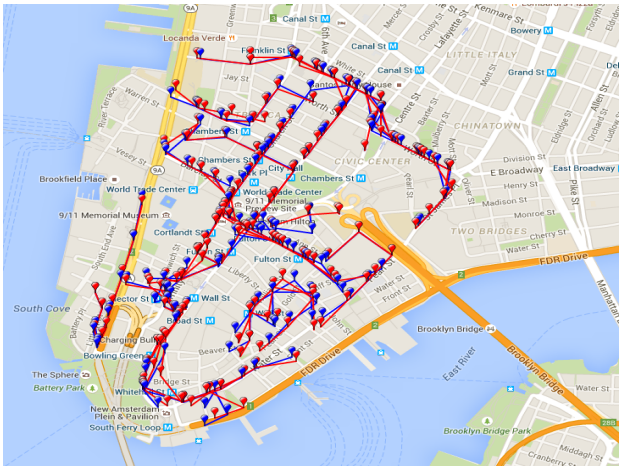


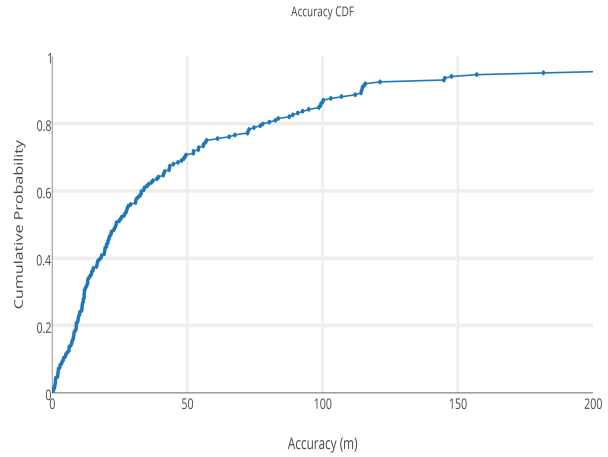
Fig. 2. Comparison of actual (red) and predicted (blue).

Size of data set: Our data set consisted of around 129000 drive test data points at 19000 distinct locations. We present results with two different splits between training and test locations: one with percentage of locations unique to training locations, unique to test locations, common to both respectively $50\%, 40\%, 10\%$, and another with the corresponding fractions $70\%, 20\%, 10\%$. All our results are averages over more than 100 user generated LUMD sequences.

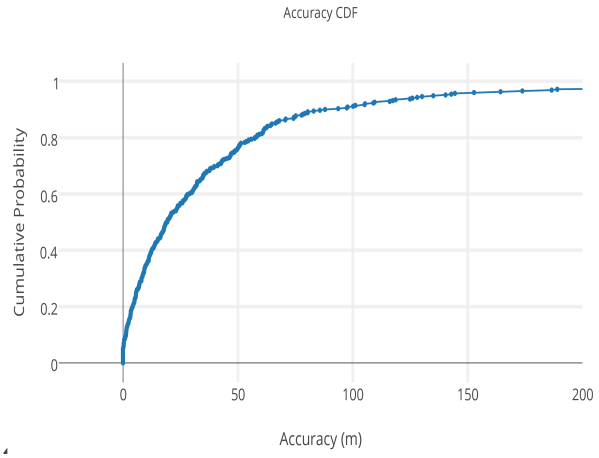
B. LUMD Localization Results

In Figure 2, we show the predicted and actual locations of all mobiles for which we generated LUMD records. As it can be seen, the actual locations and the estimated locations are quite close. In the following, we present more detailed analysis of the results.

Accuracy CDF: In Figure 3(a) and Figure 3(b), we show the accuracy distribution for two different cases of fraction of locations used for training. When the training locations are 50% of locations, the median accuracy is around $25m$ and when the training locations are 70% of locations, the median accuracy is around $20m$. At a higher percentile, the accuracy is around $50m$ with 70% training locations and around $75m$ for 50% training locations. This implies that, when smaller fraction of network coverage area is drive tested, the median

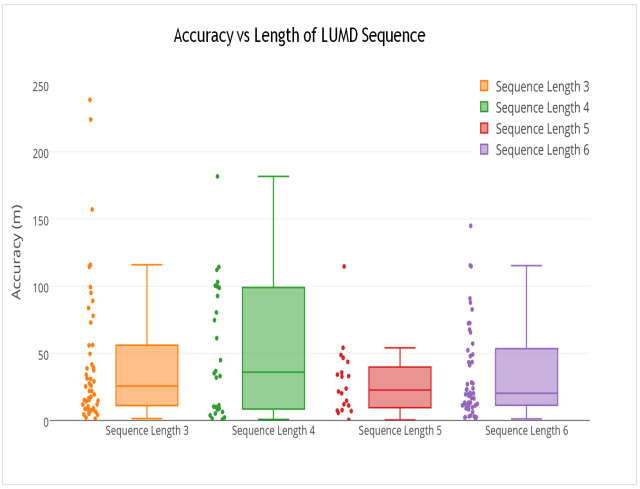


(a)

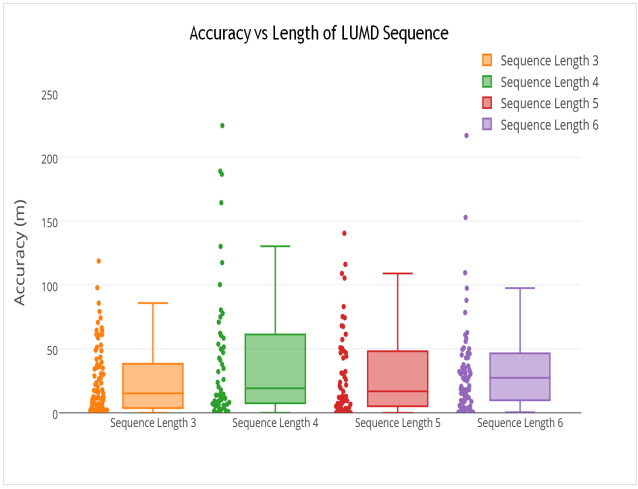


(b)

Fig. 3. CDF of accuracy. when training locations, test locations, common locations are respectively $0.5, 0.4, 0.1$ and $0.7, 0.2, 0.1$.



(a)



(b)

Fig. 4. LUMD length v/s accuracy when training locations, test locations, common locations are respectively $0.5, 0.4, 0.1$ and $0.7, 0.2, 0.1$.

accuracy does not get affected much but the probability of large inaccuracy increases. However, a median accuracy is the range of $20m - 30m$ range is significant improvement over previous non-machine learning based techniques in literature that reported median accuracies of more than $100m$ [11].

Accuracy vs. length of LUMD sequences: Another relevant question is whether the performance of scheme depends critically on length of LUMD sequences because our technique relies on stitching together multiple LUMD records from the same user. In Figure 4(a) and Figure 4(b), we show the accuracies in the form a box plots. For different LUMD sequences, we show boxes that represent IQR or inter quantile range ($25 - 75$-th percentile) and the middle line in each box represents the median. As it can be seen that the median accuracy of our scheme does not change much with length of LUMD sequences. For example, the median accuracy with 70% locations with training data, has all median accuracies within $30m$ for any LUMD sequence of length less than 6.

Remark 2. *An important question is related to whether a median accuracy of 20m is good enough. One measure of this to identify the extent to which the location uncertainty area is reduced. To illustrate this, since measurement records contain serving cell information and typical serving cell radius is around 500m in LTE in urban areas, a rough figure of initial location uncertainty area is* $\pi \times 500^2$. *With a median error of* $20m$, *the location uncertainty area is reduced to a factor of* $20^2/500^2 \approx 0.16\%$. *For indoor localization, if a Wi-Fi coverage area is taken as* $200m$, *a similar reduction would require the localization error to be around* $3.2m$.

C. Indoor Outdoor Classification Results

In Figure VII-C, we show the how well we can classify whether a record comes from an indoor mobile or not.The classification score we use is the popular F1 score [20] defined by

$$F1\ Score\ = \frac{2PR}{P+R},$$

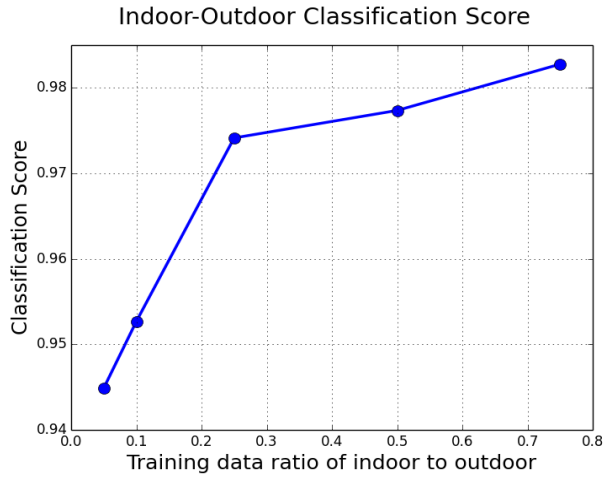


Fig. 5. Indoor-outdoor classification accuracy for different proportion of drive and walk test data.

where P is precision defined as the number of true positives (i.e., indoor) divided by number of predicted positives and R is recall defined as the number of true positives divided by the number of actual positives. The reason behind showing the classification score for different proportion of walk to drive test data is the following. In practice, collecting drive test data is much more prevalent and easier and thus we wanted to understand the performance as we have smaller amount of walk test data as compared to drive test data. As we can see in Fire VII-C, the classification score is in the range 0.95 even when walk test data set is just 10% of drive test data set and the outdoor classification increases up to 0.98 when walk test data set is around 80% of drive test data set. This is a string evidence of fact that the RSRP, RSSI tuple can serve as very good feature set for indoor outdoor classification.

VIII. Concluding Remarks

In this paper, we have developed localization algorithms of measurement records in LTE networks and we have also shown that measurement records can classified as indoor or outdoor with appropriate training. We have shown median accuracy of 20m in urban settings which is a significant improvement over more than 100m accuracy reported with non machine learning based techniques. A more challenging problem is to identify indoor locations at least in terms of buildings. This could require more training or combining LUMD with Wi-Fi signatures available from mobiles.

References

[1] 3GPP Technical Specification 36.331. Available from www.3gpp.org.

[2] AL., A. V. Accurate gsm indoor localization. In *UbiComp* (2005).

[3] ALTINI, M., BRUNELLI, D., FARELLA, E., AND BENINI, L. Bluetooth Indoor Localization with Multiple Neural Networks. In *IEEE Proceedings of the 5th IEEE International Conference on Wireless Pervasive Computing* (2010), pp. 295–300.

[4] BISHOP, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[5] BREIMAN, L. Random forests. *Mach. Learn. 45*, 1 (Oct 2001), 5–32.

[6] CHEN, Y., LYMBEROPOULOS, D., LIU, J., AND PRIYANTHA, B. Fm-based indoor localization. In *ACM MobiSys* (2012).

[7] CHINTALAPUDI, K., IYER, A. P., AND PADMANABHAN, V. N. Indoor localization without the pain. In *In Proc. ACM MobiCom Conference* (2010), pp. 173–184.

[8] CRUZ, C. C., COSTA, J. R., AND FERNANDES, C. A. Hybrid UHF/UWB antenna for passive indoor identification and localization systems. *IEEE Transactions on Systems, Antennas and Propagation 61*, 1 (Jan 2013), 1354–361.

[9] EL-KAFRAWY, K., YOUSSEF, M., EL-KEYI, A., AND NAGUIB, A. F. Propagation modeling for accurate indoor wlan rss-based localization. In *IEEE VTC Fall* (2010), pp. 1–5.

[10] ET. AL., J. A. The slam problem: A survey. In *Proceedings of the 2008 Conference on Artificial Intelligence Research and Development* (2008).

[11] ET. AL., M. J. F. Wireless network analysis using per call measurement data. *Bell Labs Technical Journal 11*, 4 (2007), 307–313.

[12] KUMAR, S., HAMED, E., KATABI, D., AND LI, L. E. Lte radio analytics made easy and accessible. *SIGCOMM Comput. Commun. Rev. 44*, 4 (2014), 211–222.

[13] NI, L. M., LIU, Y., LAU, Y. C., AND PATIL, A. P. Landmarc: Indoor location sensing using active rfid. *Wireless Networks 10*, 6 (Nov 2004), 701–710.

[14] RAI, A., CHINTALAPUDI, K. K., PADMANABHAN, V. N., AND SEN, R. Zee: zero-effort crowdsourcing for indoor localization. In *MOBICOM'12* (2012), pp. 293–304.

[15] SESIA, S., TOUFIK, I., AND BAKER, M. *LTE, The UMTS Long Term Evolution: From Theory to Practice*. Wiley, 2009.

[16] STELZER, A., POURVOYEUR, K., AND FISCHER, A. Concept and application of LPM - a novel 3-D local position measurement system. *IEEE Transactions on Microwave Theory and Techniques 52*, 12 (Dec 2004).

[17] TARZIA, S. P., AND ET AL. Indoor localization without infrastructure using the acoustic background spectrum. In *ACM MobiSys* (2011).

[18] THRUN, S. Robotic mapping: A survey. In *Exploring Artificial Intelligence in the New Millenium*, Morgan Kaufmann, p. 2002.

[19] THRUN, S., BURGARD, W., AND FOX, D. *Probabilistic robotics*. Intelligent robotics and autonomous agents. the MIT Press, Cambridge (Mass.) (London), 2005.

[20] W, D. M. Evaluation: From precision, recall and f-measure to roc, informedness, markedness correlation, 2011.

[21] WANG, H., FARID, M., SEN, S., YOUSSEF, M., ELGOHARY, A., AND CHOUDHURY, R. R. No need to war-drive: Unsupervised indoor localization. In *ACM MobiCom* (2012), pp. 197–210.

[22] ZANG, H., AND BOLOT, J. C. Mining call and mobility data to improve paging efficiency in cellular networks, 2007.