Localization of 5G Measurement Records

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

Fuled with the emerging cloud technologies (cloud computing, cloud storage, etc), software defined network (SDN) together with network functions virtualization (NFV) is transforming the techonology, and business models in telcommunication industry. Several open source platforms (ONAP, XRAN, etc) have emerged to provide a platform to collect data and share technogies among different vendors and operators. As the wireless network is now evolving into 5G, tremendous data will be shared in these open source platform and thus create a lot of opportunity to provide better service using these data. These data can include a wide variety of measurements, such as the service throughput of the mobile device, the serving cell, the signal strength, etc. In [3], a novel localization algorithm has been proposed to estimate the location when measurement reports are made in LTE systems. The paper shows that the medium accuracy of 20m can be achieved for outdoor mobibles. When celluar network evloves from 4G LTE to 5G, a lot of things have been changes to support lower latency and higher throughput. The most profound difference in LTE and 5G network is the utilization of the massive MIMO. With the usage of massive MIMO, the mobile will report beam related metric to the base station. Thus in this this paper, we propose a new localization algorithm to estiamte the location of the measruement report utilizing the new report metric introduced in 5G.

As the open source platform is a new platform under development, it provivdes an opportunity to request different measurement metric. The goal of this work is two folds: 1). identify the measuremnets unique to 5G network for localization. 2). to estimate the latitude-longitude of the mobile when the measurement record was generated for 5G network. In this work, we develop machine learning algorithms and present experiment results from a 5G NS-3 simulator. As far as we know, this is the first paper discussing localization issues in 5G networks.

Similar to [3], in this paper, we combine the localization principles based on RF fingerprinting and probablitistic path-tracking used for robot localization. However, different from [3], the unique property of 5G network, such as the beam-related information is used in this paper. The focus of this paper is on estimating location for outdoor mobiles. [3] has illustrated how to classify a mobile as an indoor or outdoor mobile.

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. This is done by performing supervised learning using the unique property of 5G networks. The training data for supervised learning may come from drive test carried out by network providers once 5G is deployed commercially. However, given 5G network is not available in the field now, in this paper, we first generate a set of data as drive training data and then generate another set of data as test data. This will be illustrated more in the later section of simulation setup and results.

The details of the above two steps are provided in Section IV. 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 main contribution in this paper is that we have proposed a new localization algorithm appliable to 5G network. Based on our best knowledge, this is the first localization algorithm designed for 5G network.

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

II. Relevant 5G Terminologies

Though our techniques could apply to any future cellular system, we use LTE terminologies for convenience. The terminologies [4] 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.

gNodeB (gNB): 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 (SSSRP): 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.

beam indices: 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.

We will need a figure to show the connection of 5G network to ONAP/xRAN, etc where the measured data is stored.

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 timescale) 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

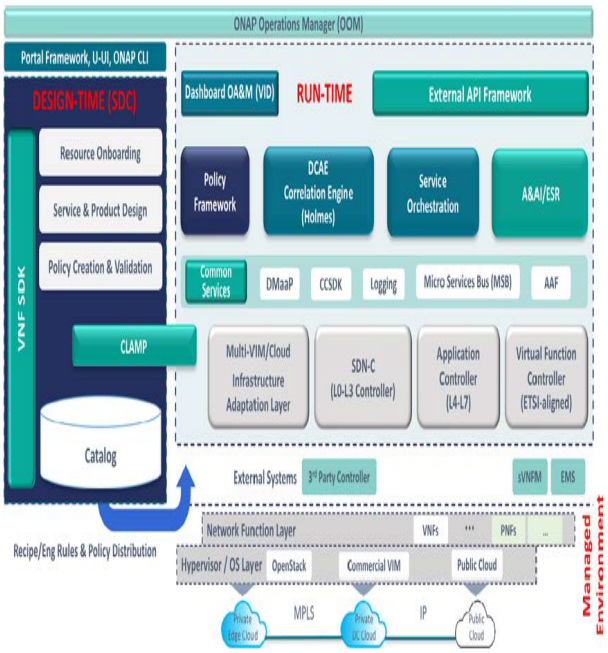


Fig. 1. LUMD Data Collection.

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).
- *beam indices:* 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.

III. Problem Statement

In this paper, we focus on one isolated cell in 5G network, which is the measurement report only has the information of the serving cell. In the field, when we have multiple cells, the additional information on the measruement report from neigboring cells can be easily incoorbated into the proposed algorithm and improve the localization accuary.

Consider in an 5G network, 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:** 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, up to four SSSRPs and beam indices of the serving cell. Note that standard allows each UE to report up to four best beam indices and corresponding SSSRPs. We denote by $\{R_{i,k}\}_{i=1}^n$ the signal strength of cell-k and $\{B_{i,k}^j\}_{i=1}^n$, the jth best beam indices sent from training location x_i. Note that, for a location x_i, the data $R_{i,k}$ and $B_{i,k}^j$ 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 SSSRP $\tilde{R}_k(t_i)$ and beam indices $\tilde{B}_k(t_i)$ 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 in 5G network: We are given training data consisting of locations $\{x_i\}_{i=1}^n$ and associated SSSRPs $\{R_{i,k}\}_{i=1}^n$ and beam indices $\{B_{i,k}\}_{i=1}^n$of cell-i at location x_i. Estimate the unknown location of a sequence of measurements $\tilde{R}_k(t_i)$ and $\tilde{B}_k(t_i)$ 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 $G_r = (V, E)$. Note that SSSRP and beam indices are the unique metric reported in 5G network.

Note that, in the algorithm illustration in this paper, we only focus on SSSRP and beam indices. However, in the field, there may be other measurement reports that can be used to help to improve the localization accuary. For example, timing alignment, etc. These addtional measurement reports can be easily incoorbated into our proposed algorithm.

IV. Localization Algorithms

The framework we use for tackling the localization problem is hidden markov model as illustrated in Figure 2. The hidden states in HMM in our case are the locations and the velocity. The observations corresponding to each hidden state are the reported measurement reports, such as SSSRP and the beam indices. The system moves from one hidden state to another hidden state with some underlying mobility model. 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.

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

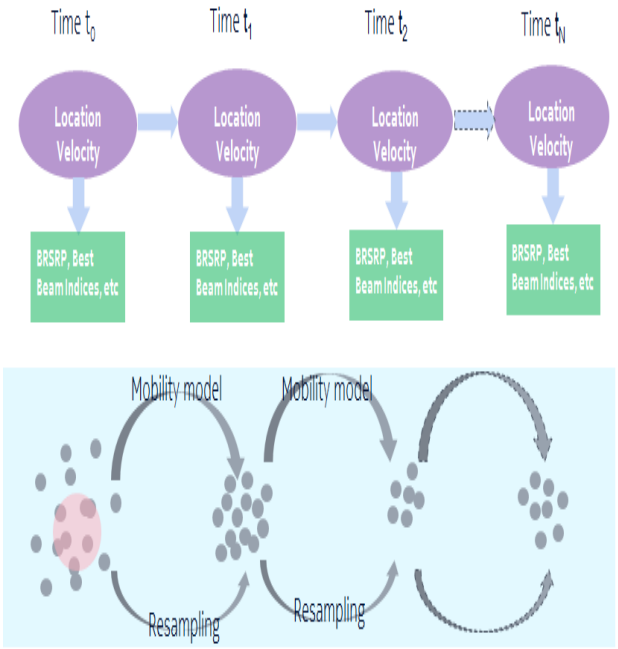


Fig. 2. LUMD Data Collection.

presented in Algorithm 1. N_{th} is a non-degeneracy parameter input which determines when less probable particles are to be discarded.

A. State transition probabilities and mobility model

These transition probabilities model how transition happens from one hidden state to another. 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.

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

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).

B. Regression based Observation Likelihood

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 $p(\tilde{R}_i, \tilde{B}_i^j | \hat{x}_i)$. In our approach, these probabilities can be learnt from the drive test data using regression on training (or say drive test) data to estimate $p(\tilde{R}_i, \tilde{B}_i^j | \hat{x}_i)$. 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 [1] on the drive test data. The rationale behind choosing Random

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

Sample N particles $\mathcal{P}_j = \{\hat{x}_1^{(j)}, \hat{v}_1^{(j)}\}$, $j = 1, \ldots, N$ from prior distribution $p(\tilde{x}_1, \tilde{v}_1|G)$
Initialize importance weights $\hat{w}_1^{(j)} \leftarrow p(\tilde{R}_1|\hat{x}_1^{(j)}, \mathcal{C})$, $j = 1, \ldots, N$
Normalize $w_1^{(j)} \leftarrow \hat{w}_1^{(j)} / \sum_{l=1}^{N} \hat{w}_1^{(l)}$, $j = 1, \ldots, N$
for $i = 2$ to m **do**
  for $j = 1$ to N **do**
    Sample $\hat{x}_i^{(j)}$ from distribution $p(\hat{x}_i^{(j)}|\hat{x}_{i-1}^{(j)}, \hat{v}_{i-1}^{(j)}, G)$
    Update weight $\hat{w}_i^{(j)} \leftarrow \hat{w}_{i-1}^{(j)} \times p(\tilde{R}_i|\hat{x}_i^{(j)}, \mathcal{C})$
    Update speed $\hat{v}_i^{(j)} = d_G(\hat{x}_i^{(j)}, \hat{x}_{i-1}^{(j)})/(T_i - T_{i-1})$
    $\mathcal{P}_j \leftarrow \mathcal{P}_j \cup \{\hat{x}_i^{(j)}, \hat{v}_i^{(j)}\}$
  end for
  Normalize $w_i^{(j)} \leftarrow \frac{\hat{w}_i^{(j)}}{\sum_{l=1}^{N} \hat{w}_i^{(l)}}$
  $\hat{N}_{eff} \leftarrow \frac{1}{\sum_{l=1}^{N} (w_i^{(l)})^2}$
  if $\hat{N}_{eff} < N_{th}$ **then**
    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
    $w_i^{(j)} \leftarrow \frac{1}{N}$ for $j = 1, \ldots, N$
  end if
end for
$j^* = \arg\max_{j=1,\ldots,N} w_m^{(j)}$
Output location estimate $\{\hat{x}_i^{(j^*)}\}_{i=1}^{m}$
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

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

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

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.

C. Classification based Observation Likelihood

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 $p(\tilde{R}_i, \tilde{B}_i^j|\hat{x}_i)$. In our approach, these probabilities can be learnt from the drive test data using regression on training (or say drive test) data to estimate $p(\tilde{R}_i, \tilde{B}_i^j|\hat{x}_i)$. 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 [1] 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

$$p(\tilde{R}_i|\hat{x}_i) = \mathcal{N}(RndFrst(\hat{x}_i, c), \sigma_c^2(\hat{x}_i)) \ , \qquad (3)$$

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.

V. Extension: Joint Localization and Channel Modling

Each LUMD record R_t may be viewed as a $K+2$ length vector containing K RSRP measurements (one for each base station), the measured RSRQ, and the timing advance (TA) from the serving base station. However for any particular record R_t several of these $K+2$ fields may be missing. The power received (in dBm) or RSRP from base station k by an UE at position x be $P_k(x)$. Since the transmit power from each base station is known, the variable $P_k(x)$ captures the channel from base station k to an UE at position x.

Problem: Let $\mathcal{D}_1 = \{X_t, R_t\}_{t=1}^n$ represent the labeled drive test dataset. $\mathcal{D}_2 = \{\tilde{R}_i, T_i\}_{i=1}^m$ is the unlabeled PCDM data, where T_i is the time when record $\tilde{R}_i$ is sent. The problem of joint channel modeling and localization is to use the complete data $\mathcal{D}_1, \mathcal{D}_2$, graph G and base station locations $\{y_k\}_{k=1}^K$ to do the following.

1) Estimate the unknown user locations $\tilde{x}_i$ for $i = 1$ to m.
2) For every location $x \in V$ estimate the RSRP $P_k(x)$ from each base station $k \in [K]$ to an UE at location x.

A. EM Algorithm

In this section we describe our main algorithm for joint channel modeling and UE localization. The Joint Channel Modeling and Localization (JCML) algorithm takes as input the labeled and unlabeled datasets $\mathcal{D}_1, \mathcal{D}_2$, the graph G and a parameter N_{em} specifying the number of expectation-maximization (EM) iterations to be performed. It outputs the UE location estimates $\{\hat{x}_i\}_{i=1}^m$ and the channel model $\mathcal{C}$ in the region of interest. The main idea is to use expectation-maximization procedure to iteratively improve the estimates of both the channel model $\mathcal{C}$ and the location estimates $\{\hat{x}_i\}_{i=1}^m$. The basic EM algorithm improves the channel in each iteration from the previous according to the following equation.

$$\mathcal{C}^{t+1} = \arg\max_{\mathcal{C}} E_{\{\hat{x}_i\}_{i=1}^m | \mathcal{D}_1, \mathcal{D}_2, \mathcal{C}^t} \log P(\mathcal{D}_1, \mathcal{D}_2, \{\hat{x}_i\}_{i=1}^m | \mathcal{C}) \tag{4}$$

The high level pseudo-code of the algorithm is shown in Algorithm 2.

Note that in line 4 $\{\hat{x}_i\}_{i=1}^m$ can be the mode or samples from the distribution $p(\{\hat{x}_i, \tilde{R}_i\}_{i=1}^m, |\mathcal{C}, G)$ as a stochastic approximation for the EM algorithm. The JCML algorithm uses two main subroutines. The first subroutine $ChannelModel$ computes the channel function $\mathcal{C}$ from the labeled dataset $\mathcal{D}$. $\mathcal{C}$ can be viewed as a function which can output the distribution of RSRP from any base station k to any UE location $x \in V$. The second subroutine $LocalizeUE$ use the channel model function $\mathcal{C}$, the LUMD data $\{\tilde{R}_i, T_i\}$ and the graph G to come up with location estimates of the UE $\{\hat{x}_i\}_{i=1}^m$ and its corresponding distribution. Each time a location estimate is computed it is used with the corresponding record $\tilde{R}_i$ and the dataset $\mathcal{D}_1$ to further improve the channel model using an expectation-maximization procedure. Therefore in the next iteration we can obtain a better location estimate. We now describe each subroutine. 3, we can just estimate the current location $\hat{x}_i$ instead of re-estimating all previous locations.

Algorithm 2 $JCML(\mathcal{D}_1, \mathcal{D}_2, G, N_{em})$

1: $\mathcal{C} \leftarrow ChannelModel(\mathcal{D}_1)$
2: **for** $j = 1$ to N_{em} **do**
3: $\quad \{\hat{x}_i\}_{i=1}^m, p(\{\hat{x}_i, \tilde{R}_i\}_{i=1}^m, |\mathcal{C}, G) \leftarrow LocalizeUE(\mathcal{D}_2, \mathcal{C}, G)$
4: $\quad \mathcal{D} \leftarrow \mathcal{D}_1 \cup \{\tilde{R}_i, \hat{x}_i\}_{i=1}^m$
5: $\quad \mathcal{C} \leftarrow ChannelModel(\mathcal{D})$
6: **end for**
7: Output $\{\hat{x}_i\}_{i=1}^m = \arg\max_{\{\hat{x}_i\}_{i=1}^m} p(\{\hat{x}_i, \tilde{R}_i\}_{i=1}^m, |\mathcal{C}, G)$
8: Output $\mathcal{C}$

VI. 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 3). 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 II) 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).

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 3, 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 4(a) and Figure 4(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 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 $100m$ [2].

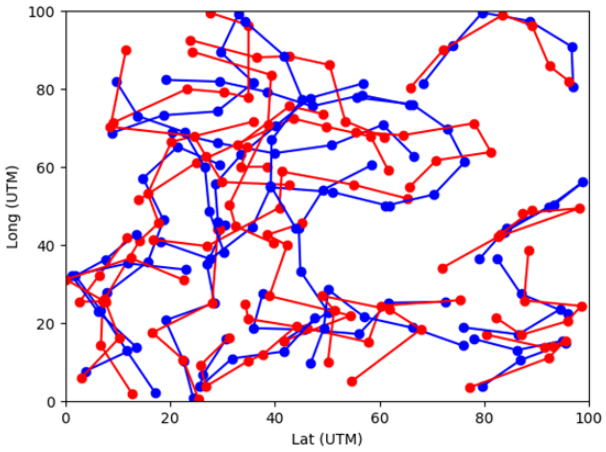


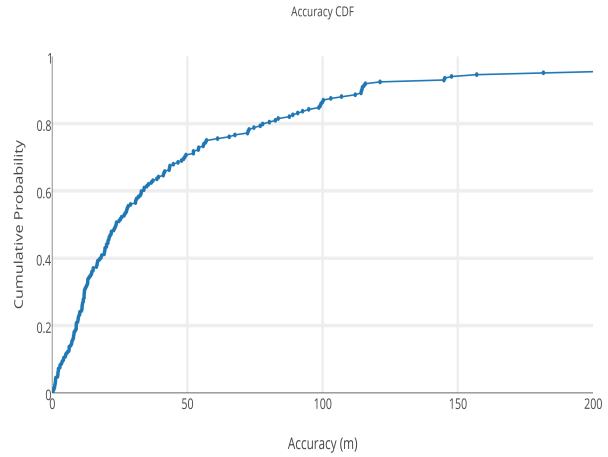
Fig. 3. Comparison of actual (red) and predicted (blue).

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 5(a) and Figure 5(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.

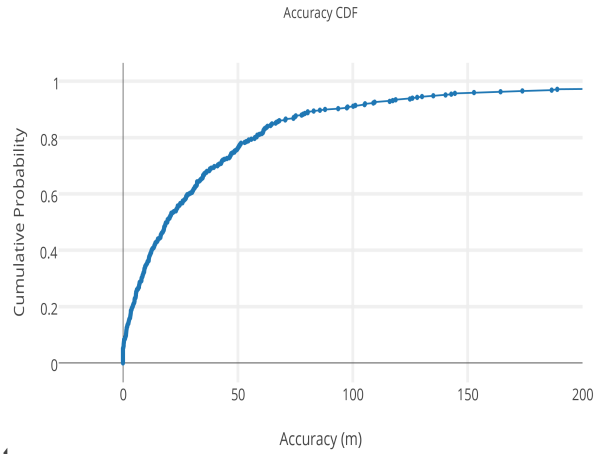
Remark 1. *An important question is related to whether a median accuracy of 20m is good enough. One measure of this to identify the extent to which the location uncertainty area is reduced. To illustrate this, since measurement records contain serving cell information and typical serving cell radius 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* $3.2m$.

VII. 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



(a)



(b)

Fig. 4. CDF of accuracy. when training locations, test locations, common locations are respectively $0.5, 0.4, 0.1$ and $0.7, 0.2, 0.1$.

could require more training or combining LUMD with Wi-Fi signatures available from mobiles.

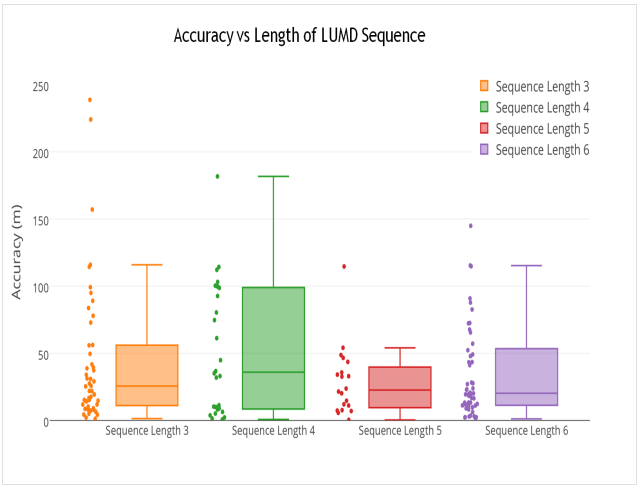
REFERENCES

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

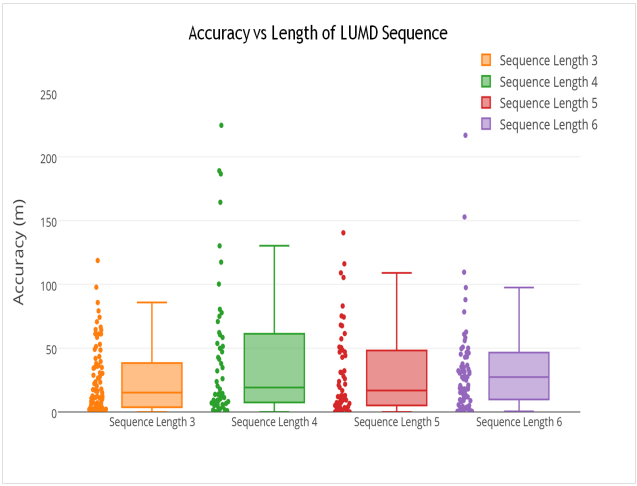
[2] ET. AL., M. J. F. Wireless network analysis using per call measurement data. *Bell Labs Technical Journal 11*, 4 (2007), 307–313.

[3] RAY, A., DEB, S., AND MNOGIOUDIS, P. Localization of lte measurement records with missing information. In *IEEE INFOCOM 2016* (2016), pp. 1–9.

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



(a)



(b)

Fig. 5. LUMD length v/s accuracy when training locations, test locations, common locations are respectively $0.5, 0.4, 0.1$ and $0.7, 0.2, 0.1$.