Localization of 5G Measurement Records

Paper Id: xxxx

***Abstract*—**

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 [4], 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 changes have been made to support lower latency and higher throughput. The most profound advanced technology adopted in 5G network is the utilization of the massive multiple input multiple output (MIMO) antennas. With the usage of massive MIMO, the mobile will report beam related metric to the base station. In this paper, we propose a new localization algorithm to estiamte the location of the measruement report utilizing the new reported metric introduced in 5G network.

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 paper, we combine the localization principles based on RF fingerprinting and probablitistic path-tracking used for robot localization, which is similar to [4]. However, different from [4], the unique property of 5G network, such as the beamrelated information is used in this paper. The focus of this paper is on estimating location for outdoor mobiles. 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 commein the field. However, given 5G network is not available in the field now, in this paper, we generate the drive training data and testing data from a 5G simulator.

The details of the above two steps are provided in Section IV.

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 VI describes the extension of the joint localizationa and channel model estimation algorithm. We present experimental validation in Section VII and finally we conclude in Section VIII.

II. Relevant 5G Terminologies

NOTE: This section needs to be updated. Now, it is copied from previous paper, thus, it is about LTE.

Though our techniques could apply to any future cellular system, we use LTE terminologies for convenience. The terminologies [5] 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.

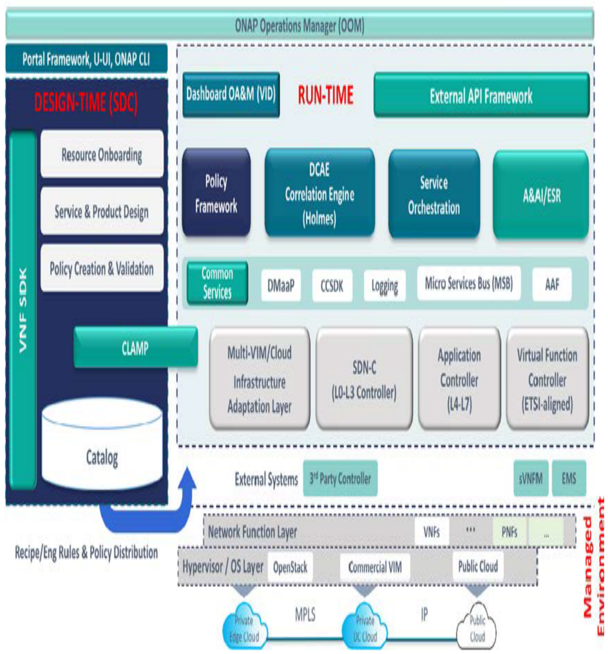


Fig. 1. LUMD Data Collection.

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 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).
- *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

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 $\{B_{i,k}^j\}_{i=1}^n$, the jth best beam indices sent from training location x_i with cell k and $\{R_{i,k}^j\}_{i=1}^n$ are the corresponding signal strength associated with those reported beam indices. Note that, for a location x_i, the data $R_{i,k}^j$ and $B_{i,k}^j$ are 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 $\{\tilde{R}_k^j(t_i)\}_{j=1}^4$ and beam indices $\{\tilde{B}_k^j(t_i)\}_{j=1}^4$ 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 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 $\{R_{i,k}^j\}_{i=1}^n$ and beam indices $\{B_{i,k}^j\}_{i=1}^n$of cell-i at location x_i. Estimate the unknown location of a sequence of measurements $\tilde{R}_k^j(t_i)$ and $\tilde{B}_k^j(t_i)$ where $i = 1, 2, \ldots, m,\ k \in K(t_i), j = 1, 2, 3, 4$. 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 new 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 (HMM) 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 this paper, we use particle filter to solve the localization problem presented in HMM. In this approach, the following two probablistic models are needed:

- state transition model: this desribes how the hidden state moves from one state to another state as illustrated in section IV-A.
- channel model: this describes the probability of observing the

In the proposed algorithm $5GLocalizeAlgo$ 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 $\tilde{R}_i^j$ 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.

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 Gauss-Markov Mobility Model [3].

$$S_t = \alpha S_{t-1} + (1-\alpha)\bar{S} + \sqrt{(1-\alpha^2)} S_{x_{t-1}} \tag{1}$$

$$d_t = \alpha d_{t-1} + (1-\alpha)\bar{d} + \sqrt{(1-\alpha^2)} d_{x_{t-1}} \tag{2}$$

where S_t and d_t are the new speed and direction of the mobile at time interval t. S_{t-1} and d_{t-1} are random variables from a Gaussian distribution with mean $\bar{S}$ and $\bar{d}$.

At each time interval the next location is calculated based on the current location, speed, and direction of movement. Specifically, at time interval t, a mobile's position is given by the equations.

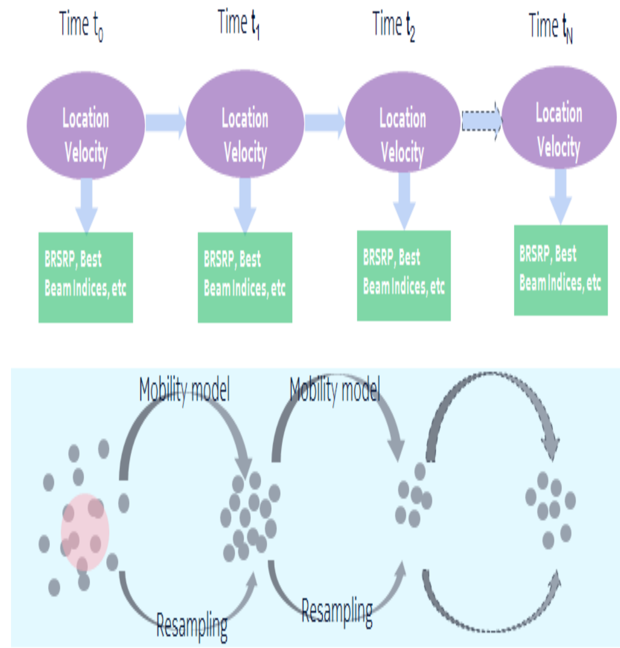


Fig. 2. LUMD Data Collection.

$$x_t = x_{t-1} + S_{t-1} cos(d_{t-1}) \tag{3}$$

$$y_t = y_{t-1} + S_{t-1} sin(d_{t-1}) \tag{4}$$

where (x_t,y_t) and (x_{t-1},y_{t-1}) are the x and y coordinates of the mobiles position at the tth and $(t-1)$ st time intervals, respectively.

V. Machine learning based Channel modeling

A. Regression based Observation Likelihood

The SSSRP reported at different states (locations) is part of the observations of HMM model. The probability distribution (also called the likelihood function) of an observed SSSRP on a location is denoted by $p(\tilde{R}_i|\hat{x}_i)$. In our approach, these probabilities can be learnt from the training data using regression on training (or say drive test) data to estimate $p(\tilde{R}_i|\hat{x}_i)$. There are different machine learning regression algorithm in the literature. However, SSSPR has the following two special properties:

- The drive test data is spread over a non-contiguous location because coverage areas in a cell are not necessarily connected.
- Wireless SSSRP manifests quite different properties in different locations

Thus, we use random forest regressor to segment the area into locations where the SSSPRs exhibit strong spatial correlation.

The regressions steps are as follows:

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

Offline inference based on training data. This does not need real time processing.
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 $\hat{w}_1^{(n)} \leftarrow p(\tilde{R}_1^1, \tilde{R}_1^2, \tilde{R}_1^3, \tilde{R}_1^4, \tilde{B}_1^1, \tilde{B}_1^2, \tilde{B}_1^3, \tilde{B}_1^4|\hat{x}_1^{(n)}, \mathcal{C})$, $n = 1, \ldots, N$
Normalize $w_1^{(n)} \leftarrow \hat{w}_1^{(n)} / \sum_{l=1}^{N} \hat{w}_1^{(l)}$, $n = 1, \ldots, N$
for $i = 2$ to m **do**
    for $n = 1$ to N **do**
        Sample $\hat{x}_i^{(n)}$ based on Gauss-Markov model in (3) and (4).
        Update weight $\hat{w}_i^{(n)} \leftarrow \hat{w}_{i-1}^{(n)} \times p(\tilde{R}_i^1, \tilde{R}_i^2, \tilde{R}_i^3, \tilde{R}_i^4, \tilde{B}_i^1, \tilde{B}_i^2, \tilde{B}_i^3, \tilde{B}_i^4|\hat{x}_i^{(n)}, \mathcal{C})$
    end for
    Normalize $w_i^{(n)} \leftarrow \frac{\hat{w}_i^{(n)}}{\sum_{l=1}^{N} \hat{w}_i^{(l)}}$
    $\hat{N}_{eff} \leftarrow \frac{1}{\sum_{l=1}^{N} (w_i^{(l)})^2}$
    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
        $w_i^{(n)} \leftarrow \frac{1}{N}$ for $n = 1, \ldots, N$
    end if
end for
$n^* = \arg\max_{n=1,\ldots,N} w_m^{(n)}$
Output location estimate $\{\hat{x}_i^{(n^*)}\}_{i=1}^{m}$
Output distribution
$p(\{\hat{x}^{(n)}\}_{i=1}^{m} | \{\tilde{R}_i^{1,2,3,4}\}_{i=1}^{m}, \{\tilde{B}_i^{1,2,3,4}\}_{i=1}^{m}, \mathcal{C}, G) = w_m^{(j)}$ for $n = 1$ to N

1) For each location, each base station and each reported beam indices that can be heard at that location, we take the empirical mean and standard deviation of all corresponding drive test data SSSRP.
2) For each cell and each beam indices, model the spatial variation of SSSRP-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 SSSRP-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 $RndFrst_m(x, b, c)$ ($RndFrst_s(x, c)$) the random forest predictor of mean (standard deviation) of SSSRP for cell-c, beam-b at location x. Let $(\sigma_{RF}(b, c))^2$ be the corresponding mean square error of the predictor. Then we model

$$p(\tilde{R}_i|\hat{x}_i) = \mathcal{N}(RndFrst(\hat{x}_i, b, c), \sigma_c^2(\hat{x}_i)) \ , \quad (5)$$

where

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

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.

B. Classification based Observation Likelihood

As part of the observations of HMM model, UE reports up to four observed best beam indices. The probability distribution (also called the likelihood function) of a particular observed beam indices as its j-th best reported beam indices conditioned on a location is denoted by $p(\tilde{B}_i^j|\hat{x}_i)$. In our approach, these probabilities can be learnt from the drive test data using classification on training (or say drive test) data to estimate $p(\tilde{B}_i^j|\hat{x}_i)$. There are different machine learning classifier in the literature.

Thus, we use random forest classifier to segment the area into locations where the coverage of beam indices exhibit strong spatial correlation.

The classification steps are as follows:

1) For each location, each base station that can be heard at that location, we take the corresponding drive test data beam indices.
2) For each cell, model the likelihood of a certain classification (beam indices) using *Random Forest* where the latitude and the longitude are taken as features of the model and the likelihood of the beam indices is the output. Each such *Random Forest* is trained using data aggregated in previous step.
3) In this paper, we use the random forest classifier in the machine learning library (sklearn) to predict the probability of reporting a specific beam indices. The function $predict\_proba(X)$ can be used to get the probability of class X.

C. Inference based Observation Likelihood

In the above two sections V-A and V-B, we have descrbed how to do the inference based on SSSRP and beam indices seperately. In this subsection, we will combine them to construct the combined inference based on the observation.

$$\begin{aligned} &p(\tilde{R}_i^1, \tilde{R}_i^2, \tilde{R}_i^3, \tilde{R}_i^4, \tilde{B}_i^1, \tilde{B}_i^2, \tilde{B}_i^3, \tilde{B}_i^4|\hat{x}_i) \\ &\qquad = \prod_{j=1}^{4} p(\tilde{R}_i^j|\hat{x}_i, \tilde{B}_i^j) p(\tilde{B}_i^j|\hat{x}_i), \end{aligned} \quad (6)$$

VI. Extension: Joint Localization and Channel Modling

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.

$$\mathcal{C}^{t+1} = \arg\max_{\mathcal{C}} E_{\{\hat{x}_i\}_{i=1}^m|\mathcal{D}_1,\mathcal{D}_2,\mathcal{C}^t} \log P(\mathcal{D}_1, \mathcal{D}_2, \{\hat{x}_i\}_{i=1}^m|\mathcal{C}) \tag{7}$$

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

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 $\mathcal{C}$

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.

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.

A. Methodology

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. As there is no commercial deployed 5G network yet, we dont have the drive test data from the field. Instead, we use our 5G system simulator to generate both training data and test data.

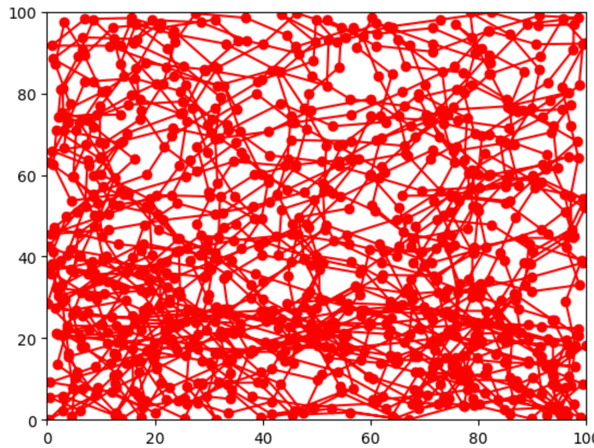


Fig. 3. Training data points

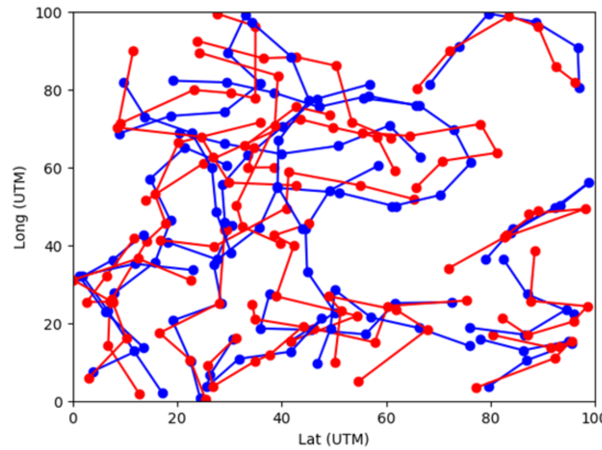


Fig. 4. Comparison of actual (red) and predicted (blue).

The 5G system we use in this evaluation is operating at mmwave band, and the cell ISD is 100m. The testing data points are illustrated in Figure 3. Instead of diving the training data into training and validation, we use the gridsearchCV and cross validation functionaility in sklearn [1] to avoid the overfitting since we have limited number of training data.

B. Localization Results

In Figure 4, we show the predicted and actual locations of all mobiles. 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.

Inference accuarcy Figure 5(a) and Figure 5(b) illustrate the beam indices at different locations when RF channel conditons are LoS and mixed LoSNLoS, respectively. Figure 5(c) and Figure 5(d) shows the predicted beam indices using random forest classification. As we can see from these figures, when the training data is more spreaded as shown in Figure 5(b), the random forest classifier will divided the whole area into smaller region. The accuracy sore for LoS is 0.78 and for mixed LoSNLoS case is 0.57.

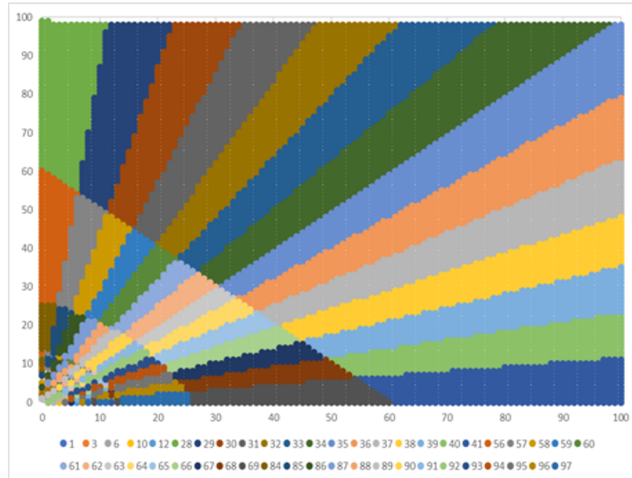
Accuracy CDF: In Figure 6(a) and Figure 6(b), we show the accuracy distribution under two different type of RF channel condition, line of sight (LoS) vs. Mixed LoS NLoS. The RF channel scenarios are illustrated in [2]. With LoS channel, the median accuracy is around $4m$ and with Mixed LoS and NLoS channel, the median accuracy is around $12m$.These results have not included additional constraints such as that users move on prescribed paths such was walkways etc. as well as additional columns in the data matrix such as TA, etc. With the additional information, the localization accuracy is expected to improve further.

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.

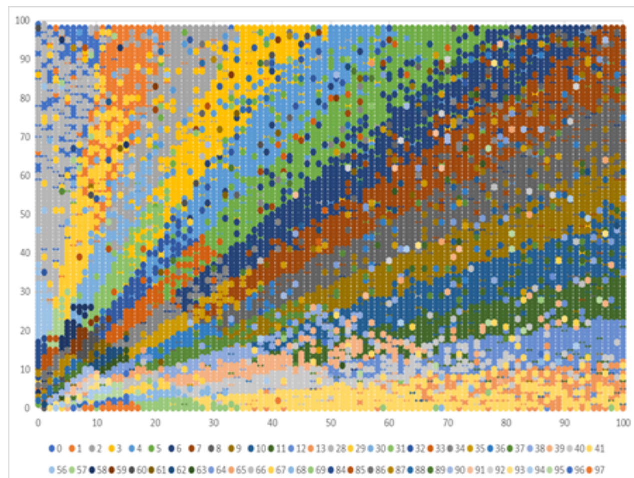
VIII. Concluding Remarks

References

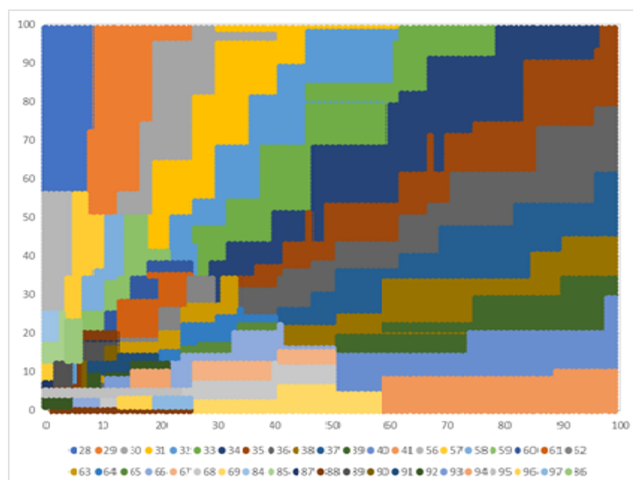
[1] *http://scikit-learn.org/*.
[2] *http://www.3gpp.org/ftp//Specs/archive/38*$_series$/38.901/38901 $-$ $e20.zip$.
[3] Camp, T., Boleng, J., and Davies, V. A survey of mobility models for ad hoc network research. *Wireless Communications and Mobile Computing, Special Issue: Mobile Ad Hoc Networking Research, Trends and Applications 2*, 5 (2002), 439–547.
[4] Ray, A., Deb, S., and Mnogioudis, P. Localization of lte measurement records with missing information. In *IEEE INFOCOM 2016* (2016), pp. 1–9.
[5] Sesia, S., Toufik, I., and Baker, M. *LTE, The UMTS Long Term Evolution: From Theory to Practice*. Wiley, 2009.



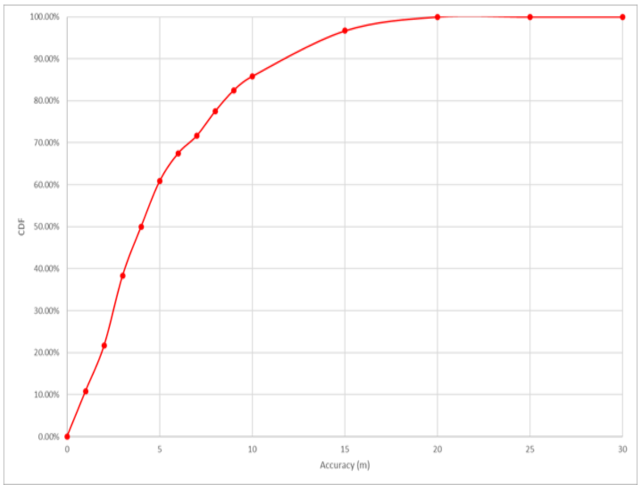
(a)



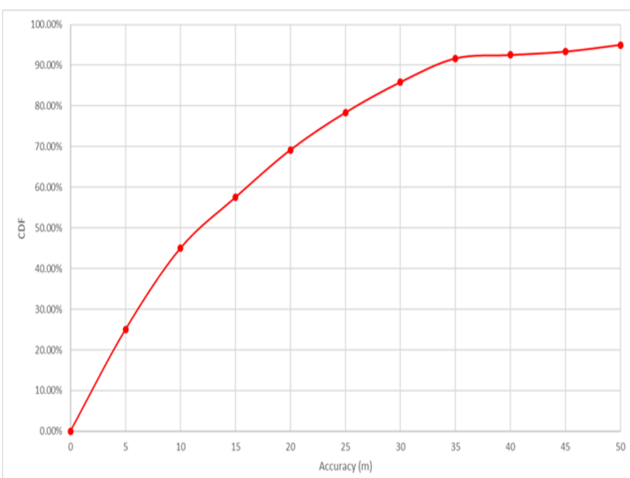
(b)



(c)



(a)



(b)

Fig. 6. CDF of accuracy. with different RF channel condition (LoS vs. MixedLoSNLoS)