2017 IEEE 18th International Conference on Mobile Data Management

# Experimental Study of Telco Localization Methods

Yukun Huang[§] Weixiong Rao[§] Fangzhou Zhu[†] Ning Liu[†] Mingxuan Yuan[†] Jia Zeng[†] Hua Yang[‡]
[§]School of Software Engineering, Tongji University, China [†]Huawei Noah's Ark Lab, Hong Kong [‡]MSC Software Corporation, CA, USA
**Corresponding Authors**: wxrao@tongji.edu.cn, {yuan.mingxuan, zeng.jia}@huawei.com

***Abstract*—Telecommunication (Telco) localization is a technique to accurately locate mobile devices (MDs) using measurement report (MR) data, and has been widely used in Telco industry. Many techniques have been proposed, including measurement-based statistical algorithms, fingerprinting algorithms and different machine learning-based algorithms. However, it has not been well studied yet on how these algorithms perform on various Telco MR data sets. In this paper, we conduct a comprehensive experimental study of five state-of-art algorithms for Telco localization. Based on real data sets from two Telco networks, we study the localization performance of such algorithms. We find that a Random Forest-based machine learning algorithm performs best in most experiments due to high localization accuracy and insensitivity to data volume. The experimental result and observation in this paper may inspire and enhance future research in Telco localization.**

## I. INTRODUCTION

With the widespread use of location-based service (LBS) in the past decade, it has become an essential requirement to accurately locate a mobile device (MD) outdoor. Telco localization is a technique which uses measurement report (MR) from Telco networks (GSM, CDMA and LTE) to calculate the position of an MD. Compared with popular Global Positioning System (GPS), Telco localization has the following advantages [30]: *1*) energy-efficient, *2*) available in most mobile phones or devices, *3*) better network coverage, *4*) active when making calls or mobile broadband (MBB) services, and *5*) crowd spatiotemporal behavior analysis without bringing any burden to MDs. Thus Telco localization has attracted intensive research interests in Telco industry [2], [6], [14], [17], [18], [24], [25]. Many real applications using Telco localization have been deployed, such as coarse localization in Google mobile map, enterprise vehicle management, and Disney passenger flow monitoring.

In literature there are three main categories of Telco localization techniques. The *1st* category is called *the measurement-based statistical method* [20]. The basic idea is to use absolute point-to-point distance estimates or angles estimates from Telco signals to calculate location. Classic methods include as Angle of arrival (AOA) technique [24], Time of Arrival (TOA) technique [18] and Received signal strength (RSS)-based single source localization [25]. These methods predict an MD's location by estimating the distance from the MD to base stations (BSs). They do not require complex calculation and usually need extra equipments to the Telco network. Nevertheless, these methods are weak in term of localization precision. The *2nd* category is called *the fingerprinting method* [14]. Fingerprinting method divides the urban area into small grids and represents each grid by an associated fingerprint. For example in Cellsense [13], the fingerprint in a grid is the Radio Signal Strength Index (RSSI) distribution of the MDs within this grid. The *3rd* category is called *the machine learning-based method* [27], which builds models, such as Random Forest (RF) and artificial neural network (ANN), from sample inputs. These methods use well-trained models to estimate the location of an MD from MR data. A recent work [30] proposed a two-layer RF regression model and achieved high precision and performance. Beyond machine learning-based models, in this paper, we further extend the models for Telco localization in two ways: regression and classification. A regression model can be directly used to calculate the GPS numeric coordinates, while a classification model can be used to find one target grid which an input MR record belongs to (we need first to cut the urban area to small grids beforehand).

Although many techniques have been proposed for Telco localization, there is still lack of a comprehensive performance comparison. It should be deeply studied on how different algorithms perform on the data from various Telco networks and how we should select appropriate models. In this work, we extensively study the performance of five typical algorithms, including RSS-based algorithm [25], state-of-the-art fingerprinting algorithm-Cellsense [13], and three machine learning-based algorithms (i.e. Random Forest, Multilayer perceptron neural network and XGBoost [7]) in the following aspects:

- Accuracy comparison: Which algorithm can achieve the best localization accuracy.
- Data Source, Data Volume and BS neighbors:
  1) How would these algorithms perform with data from different Telco networks (including GSM and LTE in this work)[1]?
  2) How would these algorithms perform with different volumes of data?
  3) How would the count of BSs in MR data impact localization accuracy?
- Gridding mechanism: How would the classification model with a gridding mechanism improves localization accuracy when compared with regression model without gridding?
- Time Efficiency: How would these algorithms perform in terms of computation time?

To summarize, we make the following contributions:

1) We conduct a comprehensive performance study on Telco localization algorithms. We identify Cellsense and

[1]Note that in remote rural areas, calling quality is still very embarrassing, signal coverage is not perfect, and base station construction is lagging. Therefore, GSM is still need to be researched as well as LTE.

2375-0324/17 $31.00 © 2017 IEEE
DOI 10.1109/MDM.2017.48

Random Forest (RF) as the best localization algorithms in 2G and 4G data sets respectively. In particular, RF wins in most experiments for its high accuracy and insensitivity to data volume or neighbors of MR data.

2) We conclude that different data sources have significant impact on localization precision. In detail, BS density is more important than the number of neighbors in MR data. Because of higher BS density, the algorithms perform better on 4G MR data, although 4G MR data is with a smaller amount of neighbors.
3) We verify that gridding mechanism can help improving most algorithms for better localization accuracy. Gridding mechanism transforms the prediction result from numeric GPS coordinates to classified grids. This mechanism can improve the accuracy by using a smaller grid size but at the cost of calculation complexity.

The rest of this paper is organized as follows. Section II introduces the MR data and five Telco localization algorithms. Section III describes our experiment setting, and Section IV reports the detail of experimental results. Finally Section V concludes the work. Figure I summarizes the mainly used terms (or short names) and associated meanings in the paper.

| Term | Meaning |
|---|---|
| BS | Base Station |
| MD | Mobile Device |
| ML estimator | Maximum Likelihood estimator |
| MLP | Multilayer Perceptron |
| MR | Measurement Report |
| RF | Random Forest |
| RSSI | Radio Signal Strength Index |
| Telco | Telecommunication |

TABLE I: Used Terms and associated meanings

## II. Background

In this section, we first briefly introduce Telco localization problem and MR data (Section II-A), and next review literature work (Section II-B).

### A. Telco Localization and MR data

MR records the connection information of an MD when the MD connects with a Telco network. Such information contains signal measurements and information of base stations (BSs). These BSs are called *neighbors*. Normally there are $1 \sim 6$ neighbors in each MR record. One of these BSs, called *service BS*, provides communication services to the MD. In the rest of this paper, the first neighbor and the service BS represent the same BS. As shown in Figure 1, a typical MR record contains a user id (IMSI: International Mobile Subscriber Identification Number), connecting time, BS IDs and the corresponding RSSIs (Radio Signal Strength Index). Telco localization is a technique to calculate the position of an MD by making use of the signal measurement information in MR records.

Figure 1 describes how Telco localization works. A Telco localization algorithm takes a set of MR records as input to calculate or predict the position of an MD. It can be regarded as a function between signal measurement and real position. Machine learning-based Telco localization algorithms build prediction model based on training data sets. A *regression* model [30] can directly predict numeric GPS longitudes and latitudes as MD locations.

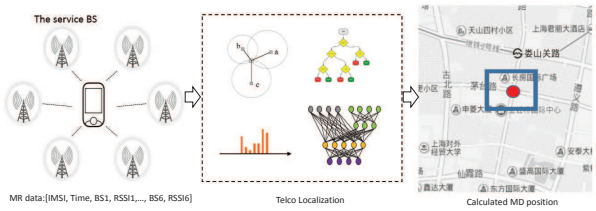


Fig. 1: MR data and Telco Localization

TABLE II: An example of the 2G MR record.

| Field | Example | Field | Example |
|---|---|---|---|
| Time | 2016/5/3 13:34:36 | MS | MS1 |
| Frame Number | 80121 | Direction | UL |
| Message Type | Measurement Report | Event | NULL |
| Event Info | NULL | **LAC** | 6243 |
| **Cell Id** | 61954 | BSIC(Num) | 33 |
| **ARFCN BCCH** | 579 | **RxLev Full (dBm)** | -71 |
| RxLev Sub (dBm) | -64 | RxQual Full | 6 |
| RxQual Sub | 0 | MS TxPower (dBm) | 30 |
| **LAC [1]** | 6205 | **Cell Id [1]** | 61954 |
| BSIC (Num) [1] | 34 | **ARFCN [1]** | 571 |
| **RxLev [1]** | -53 | **LAC [2]** | 6243 |
| ... | ... | ... | ... |
| **Cell Id [6]** | 53395 | BSIC (Num) [6] | 64 |
| **ARFCN [6]** | 575 | **RxLev (dBm) [6]** | -80 |

TABLE III: An example of the 4G MR record.

| Field | Example | Field | Example |
|---|---|---|---|
| MRTime | 2016/5/3 19:16:34 | IMSI | ****058 |
| **Serving eNodeBID** | 99129 | **Serving CellID** | 1 |
| **eNodeBID_1** | 99130 | **CellID_1** | 1 |
| **RSRP_1** | –96.63 | RSSI_1 | -57.13 |
| **RSRQ_1** | -19.63 | **eNodeBID_2** | 99130 |
| **CellID_2** | 3 | **RSRP_2** | –98.75 |
| RSSI_2 | -44.88 | **RSRQ_2** | -20 |
| ... | ... | ... | ... |
| ... | ... | **eNodeBID_6** | 99167 |
| **CellID_6** | 3 | **RSRP_6** | -88.5 |
| RSSI_6 | -44.25 | **RSRQ_6** | -19.75 |

In this paper, we divide the urban areas into small grids beforehand and next predict the grid containing the MD (then the centroid of the grid can be as the MD's position) by a *classification*-based machine learning algorithm.

We have two available MR data sets provided by one of the largest Chinese Telco operators. One is for 2G (GSM) and another for 4G (LTE). GSM is regarded as the second generation (2G) of mobile phone system, whose capabilities are achieved by allowing multiple users on a single channel via multiplexing. 2G enabled mobile phones can be used for data along with voice communication. LTE, actually standing for "long term evolution", is often marketed as 4G technology by Telco companies that package it as part of their wireless or mobile service. LTE is based on GSM/EDGE and UMTS/HSPA network technologies, and provides an increase to both capacity and speed using new techniques for modulation. Therefore, 4G can support various services by quick data access.

Table II gives an example of 2G GSM MR data. In the 2G MR data, a BS is uniquely identified by (LAC, Cell ID). ARFCN and RxLev (ARFCN is absolute radio-frequency channel number and RxLev is received signal level) indicate a code for transmission and signal strength (RSSI = RxLev). As shown in Table II, the MR record contains values from 6 nearby sectors. But more than 50% MR records in the 2G data set have missing values with no more than two sectors [29].

Table III gives an example of 4G LTE MR data. In the 4G data, a BS is uniquely identified by eNodeBID. RSSI = RSCP − EcNo (RSCP is Received Signal Code Power and EcNo is Energy per Bit to Noise Power Density). Similar to the MR records in 2G data set, the MR records in 4G data set theoretically contains items from 6 nearby sectors and actually almost all MR records in 4G data set contain only one sector. Besides RSSI, there are some other signal measurements such as RSRP and etc. These parameters are related with RSSI. For example, RSRP (Reference Signal Receive Power) is the average power of Resource Elements (RE) that carry cell specific Reference Signals (RS) over the entire bandwidth. Thus, RSRP is only measured in the symbols carrying RS [22]. While RSSI is a parameter which provides information about total received wide-band power (measure in all symbols) including all interference and thermal noise [22]. Therefore, RSRP provides information about signal strength and RSSI helps determining interference and noise information. Since the logarithmic ratio of 100 subcarriers to one subcarrier is 20 dB (e.g. $10 \times log100 = 20$), RSSI tends to measure about 20 dB higher than RSRP does. Thus RSRP represents the same information in the localization point of view. To be consistent with previous works [30], we use BS ID, BS locations and RSSIs to do localization.

We mainly use time stamp, identification of base station (BS ID), and RSSI to build the MR-based positioning system. The columns we do not mentioned above (Frame Number, Direction and etc.) are useless for this work.

In this work, we experimentally study how different Telco localization algorithms perform on 2G and 4G MR data. Section III will give two real data sets of 2G and 4G networks, and Section IV will report a comprehensive experimental study of different recent algorithms for Telco localization on the data from 2G and 4G networks.

### *B. Telco Localization Algorithms*

**Alg#1: ML Estimator** [25] is a recent measurement-based statistical method [20]. It outperforms other methods such as linear estimators. The basic idea comes from the RSS-based single source localization in sensor networks: the positions of unknown-location sensors are estimated from the known-location sensors [25]. The position of an MD can be calculated from RSSIs under a function of signal propagation power.

**Alg#2: Cellsense** [13] is the state-of-the-art fingerprinting method for Telco localization. It divides the urban areas into small grids. Each grid is represented by a fingerprint consisting of the RSSI distributions of each BS from the MDs within this grid (from the training data). For each BS in a grid, the RSSI distribution is set to be normal distribution. When the training data is not enough, the missing values are filled by normal distribution.

**Alg#3: Random Forest** (RF) [30] is a recently implemented system. It makes use of the signal measurements in MR data as features and the position point or position grid as label to train a machine learning model. The position point (i.e. longitude and latitude) can be predicted by a regression model. Figure 3 shows the basic work procedure of a machine learning algorithm for Telco localization. Random forest (RF) is one of the most widely used ensemble learning methods for classification, regression and other tasks [12]. It has demonstrated good effects in many applications [9], [23], [5]. Beyond the regression-based localization model in [30], in this paper, we mainly use the classification-based RF model and also compare a RF-based regression model against a RF-based classification model.

**Alg#4: Multilayer perceptron** (MLP) is one of feed-forward artificial networks. MLP contains multiple layers of nodes in a directed graph with a layer fully connected to the next one. Each node is a neuron with a nonlinear activation function except for the input nodes. MLP trains the network by back-propagation to calculate the weights between all couples of neurons [21]. In this paper, we adopt the MLP with a three-layer network: *1*) input layer, *2*) hidden layer and *3*) output layer. A vector of variable values of MR data is presented to the input layer. The input layer standardizes these values so that the range of each variable is -1 to 1. The input layer distributes the values to each neuron in the hidden layer. The value from each input neuron is multiplied by a weight, and the resulting weighted values are added together producing a combined value. The weighted sum is fed into an activation function, which outputs a value. The outputs from the hidden layer are distributed to the output layer. The process of computation in output layer is same to hidden layer. In hidden layer we use Relu (Rectified Linear Units) activation function [11]. In order to implement grid mechanism, the values from output layer are put in a transfer function called *Softmax* function [3], so that the final values are the probabilities of each grid. Though the three layers network is almost the simplest neural network, it can fit nearly all linear or nonlinear function. By using Relu activation function, we can use MLP to solve the classification problem effectively [15].

**Alg#5: XGBoost** is a very recent work to provide an optimized distributed gradient boosting library [7]. It implements multiple machine learning algorithms under the parallel Gradient Boosting framework (also known as GBDT, GBM) by automatically taking advantage of multithreaded parallel CPU. XGBoost is used widely by data scientists to achieve state-of-the-art results on many machine learning challenges, and XGBoost achieves good performance in KDD cups [26].

TABLE IV: Implementation Tools

| | ML estimator | Cellsense | RF | MLP | XGBoost |
|---|---|---|---|---|---|
| Numpy | √ | √ | √ | √ | √ |
| Scipy | √ | | | | |
| Pandas | √ | √ | √ | √ | √ |
| sk-learn | | | √ | | |
| Keras | | | | √ | |
| XGBoost | | | | | √ |

TABLE V: Statistics of 2G and 4G data

| Dataset | record cnt | Distance to srv BS | | Avg. RSSI of srv BS |
|---|---|---|---|---|
| | | Avg. | Median | |
| GSM | 46133 | 286.7 m | 182.3 m | -64.3 dB |
| LTE | 46641 | 187.0 m | 157.0 m | -71.2 dB |

| | BS cnt | RSSIs of all BSs | | Avg. dist. btw BS |
|---|---|---|---|---|
| | | Avg. | Median | |
| GSM | 618 | -66 dBm | -66 dBm | 512.9 m |
| LTE | 973 | -84.8 dBm | -74.6 dBm | 277.5 m |

## III. EXPERIMENTAL SETUP

### A. Implementation and Platform

We implement all algorithms with *Python 2.7* and several open source tools such as *Pandas* [19], *sk-learn* [10], *Theano* [4] and etc. Table IV summarizes the tools we used for each algorithm. All algorithms use *Pandas* to read data and store them into format of *numpy.array*. For the ML estimator, we utilize *scipy.minimize* to find out the optimal solution. *sk-learn* provides a RF API. For artificial network models such as MLP, we adopt Theano-based Keras packages. Keras is a minimalist, highly modular neural network library, written in Python. In terms of XGBoost, we download its library from the official implementation [1].

We run the ML estimator algorithm, Cellsense, RF, and XGBoost on a Ubuntu Linux machine with an Intel(R) Core(TM) i7-4930k CPU 3.40GHz, 32GB RAM and 12 cores. Each core has 32KB of L1 cache and 256KB L2 cache and all these cores have a sharing 12MB L3 cache. Because of the high computation cost, the training of MLP model are executed by a GeForce GTX 780 GPU.

### B. Datasets

We use two real MR datasets covering a 2 Km × 4 Km urban area from 2G (GSM) and 4G (LTE) networks respectively. The MR data is generated every 1 second frequency when users equipped with Android phones are driving cars. The cars are moving on each street in the testing area by 3 times forward run and 3 times backward run (except those one-way streets). Table V gives the detailed statistics of our datasets. From Table V, we can see that 4G BSs are much denser than 2G BSs. Since each 4G BS only needs to support a much smaller area, the signal strength of 4G BS is weaker than 2G signals (the smaller the RSSI value, the better the signal quality). Also, an MD does not need to measure many neighbors in a LTE network because of the dense deployment of BSs. Figure 2 gives the distribution of BS neighbors in 2G and 4G data respectively. We can find that most 4G MR records only have one neighbor. Our experiment will show that the localization accuracy of 4G data (though with fewer neighbors and weaker RSSIs) is better than the one of 2G data. It indicates that the density of BSs plays an important role in Telco localization.

TABLE VI: Varying volumes of training data

| Source data | 2G | | 4G | |
|---|---|---|---|---|
| volume of data | train | test | train | test |
| large | 41580 | 4553 | 41966 | 4675 |
| middle | 24386 | 4553 | 24511 | 4675 |
| small | 15518 | 4553 | 15289 | 4675 |

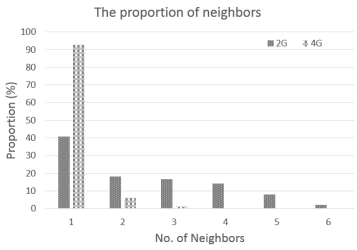


Fig. 2: Neighbor proportion of 2G and 4G data

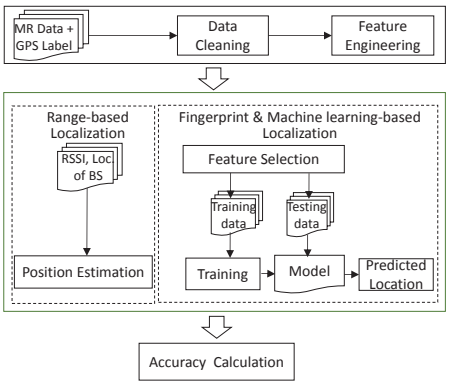


Fig. 3: Telco localization experiment flow

### C. Telco Localization Experiment Flow

We next highlight the experiment flow in Figure 3. The MR source data with GPS label (the target location to recover by Telco localization) is cleaned before used as the input of localization algorithms. For example, we need to remove those MR records containing the GPS locations which are out of the testing area. For fingerprinting and machine learning-based algorithms, we represent the signal measurements in MR records as features and divide the data into training data and testing data. The training data is used to train a fingerprint database or machine learning model. Then we locate the MR data in the testing data and calculate the accuracy by comparing the predicted results with the real GPS. The range-based algorithm does not need a training process. It extracts the RSSIs and locations of BSs and estimates the distances between an MD

and nearby BSs with signal propagation function. Then the location of an MD is calculated from the estimated distances. The range-based algorithm directly locates the MR records in testing data and calculate the accuracy.

## IV. EXPERIMENTAL RESULTS

Table VII shows the default settings of each algorithm, and Table VIII lists the parameter values specially set for Section IV-A and Section IV-B1. It is mainly because the two sections are to find the best localization accuracy and we thus tune best parameter values for the associated algorithms and data sets. By default, all algorithms are tested by using a classification model with the grid size equal to 30 meters.

TABLE VII: default parameters

| | Parameter | values |
|---|---|---|
| Cellsense | std | 5 |
| RF | number of trees | 100 |
| MLP | number of perceptions | 500 |
| XGBoost | boosting iterations | 100 |

TABLE VIII: Parameters set for two different data sets

| | 2G | | | 4G | | |
|---|---|---|---|---|---|---|
| | Parameter | | Val. | Parameter | | Val. |
| ML estimator | P0 | $\beta$ | 0 3 | P0 | $\beta$ | 0 3 |
| Cellsense | std | | 6 | std | | 5 |
| RF | # of trees | | 120 | # of trees | | 60 |
| MLP | # of perceptions | | 800 | # of perceptions | | 600 |
| XGBoost | boosting iterats. | | 180 | boosting iterats | | 180 |

### A. Accuracy comparison

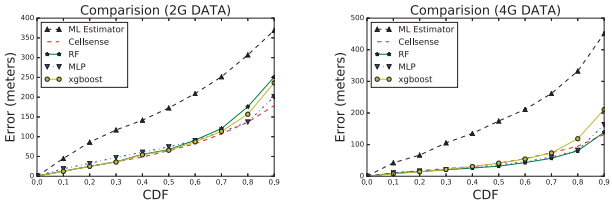


Fig. 4: Comparison of all algorithms. left: 2G; right: 4G

Firstly, we give a baseline experiment to compare the localization accuracy of all five algorithms in Figure 4. The *x*-axis in this figure indicates the cumulative distribution function (CDF) [16], [30] of errors with respect to the proportion of the test data, and *y*-axis is the associated localization error (meters). For example the localization error of 0.5 means the median error. According to this figure, Cellsense performs the best in 2G data, and RF achieves the minimum error in 4G data. The behind rationale is as follows: Cellsense requires more neighbors to construct histogram, and 2G MR data can provide enough neighbors for the construction. Meanwhile, due to a small amount of neighbors in 4G data, Cellsense is with only the third best localization accuracy. The accuracy of MLP and XGBoost is very close. For example, XGBoost performs slightly better in 2G data, while MLP has higher accuracy than XGBoost in 4G data. The performance of ML estimator is the worst no matter the data sets, with the median error lager than 100 meters.

TABLE IX: Median error (m) under different data sources

| | 2G | 4G |
|---|---|---|
| ML | 172.8 | 174.6 |
| Cellsense | 64.8 | 41.3 |
| RF | 67.3 | 31.8 |
| MLP | 74.3 | 35.7 |
| XGBoost | 65.7 | 41.3 |

### B. Data Source, Volume and Quality

*1) Data Source:* In this experiment, we study how the data source, either 2G or 4G, affects the localization performance. Figure 5 shows the CDF errors of ML estimator, Cellsense, RF, MLP and XGBoost, respectively. We can see, except the ML estimator, the four other algorithms perform better on 4G data than on 2G data. That is because the deployment of LTE BSs is much denser than the one of GSM BSs. The performance of the ranged-based method (i.e., ML estimator) does not vary too much from data sources (either 2G or 4G data). This is because ML estimator uses the same signal propagation model to calculate the distances from an MD to nearby BSs.

We summarize the localization results of five algorithms in Table IX. For example, the localization accuracy has increased 36.3%, 52.7%, 52.0%, and 37.1% respectively in Cellsense, RF, MLP and XGBoost. Based on two experimental result above, it is very clear that ML estimator performs the worst no matter the data sets and sources, with the median error lager than 100 meters. Thus, in the rest of this section, we focus on the study of other four algorithms except ML estimator.

*2) Data Volume:* We now would like to study how the volume of training data can affect the localization. We randomly divide the data into training set and testing set. We use the testing data to predict a MD's location. As described in Table VI, we have three training sets: *large*, *middle* and *small*. Figure 6 gives the localization error of four algorithms on 2G and 4G data. All four algorithms are sensitive to the volume of training data. More training data benefits better performance of the algorithms. It is because more training data means more information to represent each small urban area. Among these algorithms, Cellsense is much more sensitive than others: Cellsense requires more training data to construct the histogram of each cell. Remarkably, XGBoost shows bad performance on the small training data for both 2G and 4G data sets. The result indicates the high demand of training data by XGBoost. In addition, we note that the change of localization errors on 4G data is larger than the one 2G data with the growth of training data volume. All four algorithms are more sensitive to data volumes on 4G data than 2G data. One reason is as follows. When the volume of training data in 4G data set is decreased, more BS neighbors are lost. Consequently, the average feature information becomes smaller sharply. Based on this experiment, we will use the *large* training data due to the best performance for localization algorithms if not specially mentioned.

*3) Neighbors:* In the two data sets, many records contain the information of $< 6$ neighbor BSs. Instead, the majority of MR records contain 1 to 3 neighbors. In this experiment, we

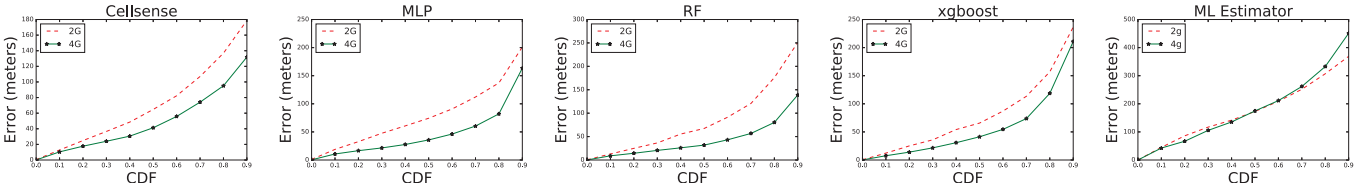


Fig. 5: Comparison between 2G and 4G data sources

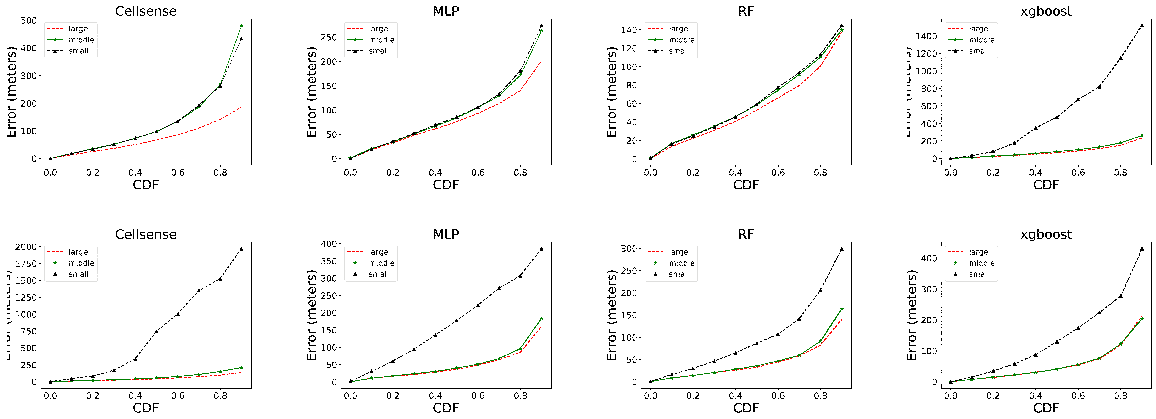


Fig. 6: Performance with different volumes of training data. $1^{st}$ row: 2G; $2^{nd}$ row: 4G

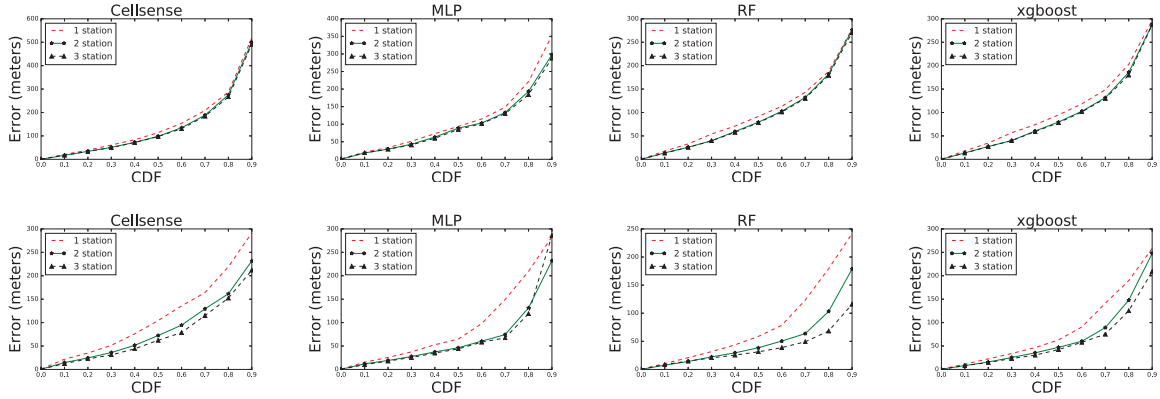


Fig. 7: Influence of neighbors. $1^{st}$ row: 2G; $2^{nd}$ row: 4G

TABLE X: Median error (m) with various neighbors

| | 2G | | | 4G | | |
|---|---|---|---|---|---|---|
| Number of BSs | 1 | 2 | 3 | 1 | 2 | 3 |
| Cellsense | 114.1 | 98.2 | 97.4 | 104.4 | 72.3 | 62.2 |
| RF | 91.6 | 79.3 | 78.3 | 58.8 | 38.5 | 31.3 |
| MLP | 93.7 | 88.9 | 85 | 64.7 | 46.2 | 43.8 |
| XGBoost | 94.7 | 79.6 | 78.1 | 62.9 | 47.1 | 42.4 |

study how the amount of of neighbor BSs affects localization accuracy. In order to study the performance of $k$ neighbors, we sort MR records by the count of neighbor BSs in each MR record. In this way, we only remain the MR records at least $k$ neighbors.

Figure 7 gives the CDF of various neighbors. More neighbors can help improving the accuracy of Cellsense, RF and MLP except XGBoost. We summarize the statistics of results in Table X. More neighbors in MR do improve the localization accuracy. For example, when the count of neighbors varies from 1 to 2, the localization accuracy of all four algorithms has increased by 13.9%, 13.4%, 5%, and 15.9% in 2G data, and 30.7%, 34.5%, 28.6%, and 25.1% in 4G, respectively. We conclude that the better localization performance on 4G data mainly benefits from more neighbor BSs. The reason is due to the higher BS deployment density on 4G data. In addition, we find that the growth tendency of the localization algorithms becomes significantly small when both 2G and 4G data sets use 3 neighbor BSs.

### C. Gridding

We adapt gridding mechanism to ML estimator, MLP and RF and next compare the performance between regression and classification. We do not plot the result of Cellsense and

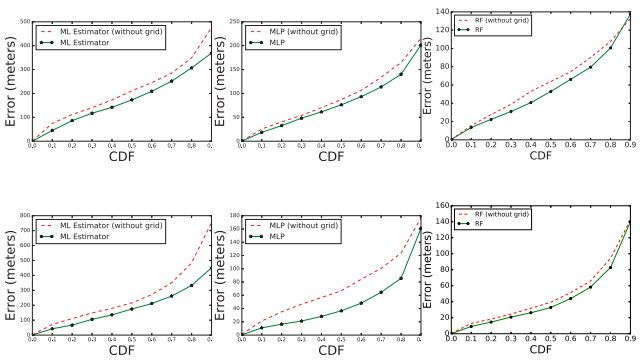


Fig. 8: Regression vs. Classification. $1^{st}$ row:2G, $2^{nd}$ row:4G

XGBoost. It is mainly because Cellsense is based on griding and XGBoost can only predict one-dimensional variables and thus the both algorithms cannot solve the regression problem. Figure 8 indicates that gridding mechanism leads to much better localization precision. The reason is that in the two data sets, the MDs are moving on the roads by car. Thus, by splitting the roads into smaller grids, the classification model can help achieving higher localization accuracy. Remarkably, among the three algorithms, RF benefits the most significantly from the griding mechanism. In particular, RF classification outperforms RF regression on both 2G and 4G data sets.

In addition, Figure 9 plots the CDF of all algorithms with various grid size. From the grid size varies from 50 meters to 150 meters, the localization errors of almost all algorithms become larger, indicating worse accuracy. It is obvious because the centroid of each grid, when given a larger grid size, more coarsely represent the real location.

### D. Efficiency

Finally, we measure the running time of five localization algorithms. To better understand the running time of localization algorithms, we break down the associated running time into two parts: training time and prediction time. All results in this experiment are measured on the *half* training data in Table VI. Table XI shows the time of all methods. Rangebase (i.e., ML estimator) does not need the training phase and thus all time is used to predict the localization. For the four other algorithms, the training time is larger than prediction time, and particularly the training time of MLP and XGBoost dominates the overall time. For example, MLP is time consuming due to many iterations of computation among neurons to tune a large amount of parameters before convergence. Similar situation occurs on XGBoost which also requires many iterations of computation. Moreover, The highest time of XGBoost among all algorithms is mainly due to the sequential boosting used in the experiment: each tree in XGBoost is built only after the previous one is finished. To sum up, Cellsense and Random Forest have a significant advantage of time efficiency.

In addition, we are interested in how all algorithms perform with various grize size. As shown in Table XI, a larger grid size leads to faster running time, because a smaller amount of grids are used. Note that Figure 9 indicates a larger grid size leads to worse localization accuracy, involving the tradeoff between computation efficiency and localization accuracy.

TABLE XI: Running Time (s) (T: Training, P: Prediction): Classification with various grid size vs. Regression

| Grid Size | 50m | | 100m | | 150m | | Regression | |
|---|---|---|---|---|---|---|---|---|
| | T | P | T | P | T | P | T | P |
| ML Estimator | | 619.6 | | 279 | | 160 | | 204 |
| Cellsense | 17.34 | 11.35 | 17.27 | 5.94 | 17.33 | 4.3 | | |
| RF | 10.26 | 1.35 | 4.81 | 0.54 | 3.29 | 0.36 | 1.56 | 0.15 |
| MLP | 891.62 | 0.07 | 710.9 | 0.04 | 647.71 | 0.02 | 501.76 | 0.02 |
| XGBoost | 9804 | 0.15 | 2717 | 0.07 | 1323 | 0.03 | | |

Table XI also lists the used running time for regression. Depending upon the grid size, the regression is slower than the classification approach with a large grid size (e.g., 150 meters) and instead faster than the one with a small grid size (e.g., 50 meters).

### E. Summary of Experimental Result

We summarize the above experiment result as follows. 1) Among the five Telco localization algorithms, we find that Random Forest achieves the best localization accuracy in most experiments and offers unique advantages including insensitivity to data volume or neighbors of MR data.

2) Our experiment shows that data sources have significant impact on localization. In detail, BS density is more important than the number of neighbors in MR data. Due to higher BS density, all algorithms perform better on 4G MR data, although 4G MR data detects fewer neighbors.

3) We verify that gridding mechanism can help improving most algorithms for better localization accuracy. Gridding mechanism transforms the prediction result from numeric coordinates to classified grids. This mechanism can improve the accuracy by using a smaller grid size but at the cost of calculation complexity.

## V. Conclusions

In this paper, we present a thorough performance evaluation of five Telco localization algorithms ML estimator, Cellsense, RF, MLP, and XGBoost on two MR data sets. We vary data source, volumes, localization mechanism (regression or classification) and grid size and measure localization accuracy and computation efficiency of such algorithms. The performance evaluation of the algorithms indicates that Cellsense performs best in 2G MR data and RF best on 4G MR data. Moreover, RF outperforms other algorithms due to unique advantages such as insensitivity to data volume or neighbors of MR data. Therefore, we would like to recommend the RF-based classification approach for Telco localization.

As future work, we will continue the performance study of localization algorithms in the following directions. For example, we would like to incorporate other data such as magnetic field signals [8] into our evaluation framework. Also, with other signal data such as Wifi, we are interested in the evaluation of indoor localization algorithms [28].

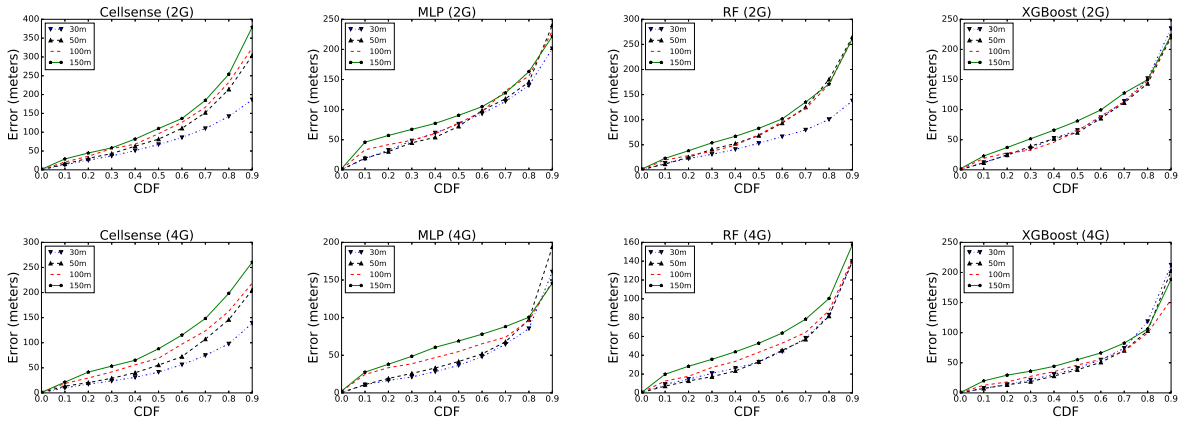


Fig. 9: Grid Size. $1^{st}$ row: 2G; $2^{nd}$ row: 4G

**Acknowledgment**: This work is partially sponsored by National Natural Science Foundation of China (Grant No. 61572365, 61503286), Science and Technology Commission of Shanghai Municipality (Grant No. 14DZ1118700, 15ZR1443000, 15YF1412600) and Huawei Innovation Research Program (HIRP).

## References

[1] XGBoost: eXtreme Gradient Boosting. https://github.com/dmlc/xgboost/.

[2] H. Aly and M. Youssef. Dejavu: an accurate energy-efficient outdoor localization system. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 154–163. ACM, 2013.

[3] Y. Anzai. *Pattern Recognition & Machine Learning*. Elsevier, 2012.

[4] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio. Theano: A cpu and gpu math compiler in python. In *Proc. 9th Python in Science Conf*, pages 1–7, 2010.

[5] L. Calderoni, M. Ferrara, A. Franco, and D. Maio. Indoor localization in a hospital environment using random forest classifiers. *Expert Systems with Applications*, 42(1):125–134, 2015.

[6] M. Y. Chen, T. Sohn, D. Chmelev, D. Haehnel, J. Hightower, J. Hughes, A. LaMarca, F. Potter, I. Smith, and A. Varshavsky. Practical metropolitan-scale positioning for gsm phones. In *International Conference on Ubiquitous Computing*, pages 225–242. Springer, 2006.

[7] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. *arXiv preprint arXiv:1603.02754*, 2016.

[8] H.-H. Chiang. *Magnetic Field Feature Analysis of Smartphone Application Activities Using Android MI Sensors*. Master Thesis of Science in Electrical Engineering University of California, Los Angeles, 2013.

[9] R. Díaz-Uriarte and S. A. De Andres. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1):1, 2006.

[10] R. Garreta and G. Moncecchi. *Learning scikit-learn: machine learning in python*. Packt Publishing Ltd, 2013.

[11] R. H. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, and H. S. Seung. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 405(6789):947–951, 2000.

[12] T. K. Ho. Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 1, pages 278–282. IEEE, 1995.

[13] M. Ibrahim and M. Youssef. Cellsense: An accurate energy-efficient gsm positioning system. *IEEE Transactions on Vehicular Technology*, 61(1):286–296, 2012.

[14] H. Koshima and J. Hoshen. Personal locator services emerge. *IEEE spectrum*, 37(2):41–48, 2000.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[16] I. Leontiadis, A. Lima, H. Kwak, R. Stanojevic, D. Wetherall, and K. Papagiannaki. From cells to streets: Estimating mobile paths with cellular-side data. In *Proceedings of the 10th ACM International on Conference on emerging Networking Experiments and Technologies, CoNEXT 2014, Sydney, Australia, December 2-5, 2014*, pages 121–132, 2014.

[17] D.-B. Lin and R.-T. Juang. Mobile location estimation based on differences of signal attenuations for gsm systems. *IEEE transactions on vehicular technology*, 54(4):1447–1454, 2005.

[18] L. Lopes, E. Viller, and B. Ludden. Gsm standards activity on location. In *Novel Methods of Location and Tracking of Cellular Mobiles and Their System Applications (Ref. No. 1999/046), IEE Colloquium on*, pages 7–1. IET, 1999.

[19] W. McKinney. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc.", 2012.

[20] N. Patwari, J. N. Ash, S. Kyperountas, A. O. Hero, R. L. Moses, and N. S. Correal. Locating the nodes: cooperative localization in wireless sensor networks. *IEEE Signal processing magazine*, 22(4):54–69, 2005.

[21] F. Rosenblatt. *Perceptions and the theory of brain mechanisms*. Spartan books, 1962.

[22] S. Sesia, I. Toufik, and M. Baker. *LTE-the UMTS long term evolution*. Wiley Online Library, 2015.

[23] V. Svetnik, A. Liaw, C. Tong, and T. Wang. Application of breimans random forest to modeling structure-activity relationships of pharmaceutical molecules. In *International Workshop on Multiple Classifier Systems*, pages 334–343. Springer, 2004.

[24] S. Swales, J. Maloney, and J. Stevenson. Locating mobile phones and the us wireless e-911 mandate. In *Novel Methods of Location and Tracking of Cellular Mobiles and Their System Applications (Ref. No. 1999/046), IEE Colloquium on*, pages 2–1. IET, 1999.

[25] R. M. Vaghefi, M. R. Gholami, and E. G. Ström. Rss-based sensor localization with unknown transmit power. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2480–2483. IEEE, 2011.

[26] H. Wei, B. Shi, and J. Chen. Location based services recommendation with budget constraints.

[27] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[28] Z. Yang, C. Wu, Z. Zhou, X. Zhang, X. Wang, and Y. Liu. Mobility increases localizability: A survey on wireless indoor localization using inertial sensors. *ACM Comput. Surv.*, 47(3):54:1–54:34, 2015.

[29] H. Zang, F. Baccelli, and J. Bolot. Bayesian inference for localization in cellular networks. In *INFORCOM*, pages 1–9, 2010.

[30] F. Zhu, C. Luo, M. Yuan, Y. Zhu, Z. Zhang, T. Gu, K. Deng, W. Rao, and J. Zeng. City-scale localization with telco big data. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 439–448. ACM, 2016.