Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2020.DOI

Enhanced MDT-based Performance Estimation for AI Driven Optimization in Future Cellular Networks

HANEYA NAEEM QURESHI *, **(Student Member, IEEE)**, **ALI IMRAN** *, **(Senior Member, IEEE), AND ADNAN ABU-DAYYA** † **(Senior Member, IEEE)**

*Electrical and Computer Engineering Department, University of Oklahoma, Tulsa, OK 74135, USA
† Department of Electrical Engineering, Qatar University, PO Box 210531, Doha, Qatar

Corresponding author: Haneya Naeem Qureshi (e-mail: haneya@ou.edu).

This material is based upon work supported by the National Science Foundation under Grant Numbers 1559483, 1619346, 1923669 and Qatar National Research Fund (QNRF) under Grant No. NPRP12-S 0311-190302. The statements made herein are solely the responsibility of the authors. For more details about these projects, please visit:http://www.ai4networks.com.

ABSTRACT

Minimization of drive test (MDT) allows coverage estimation at a base station by leveraging measurement reports gathered at the user equipment (UE) without the need for drive tests. Therefore, MDT is a key enabling feature for data and artificial intelligence driven autonomous operation and optimization in future cellular networks. However, to date, the utility of MDT feature remains thwarted by issues such as sparsity of user reports and user positioning inaccuracy. We characterize three key types of errors in MDT-based coverage estimation that stem from inaccurate user positioning, scarcity of user reports and quantization. For the first time, the presented analysis shows existence of joint interplay between these errors on coverage estimation that result from inter-dependency between positioning error and bin width. The analysis also shows that there exists an optimal bin width for given user positioning inaccuracy and user density that minimizes the overall error in MDT-based estimated coverage. Utility of our framework is presented by addressing two applications from network optimization perspective: determining optimal bin width to maximize accuracy of MDT-based coverage estimation and its calibration to further improve its accuracy.

INDEX TERMS Minimization of drive test, autonomous coverage estimation, optimal bin width, sparse data, coverage calibration, data driven optimization

I. INTRODUCTION

FUTURE cellular networks will require artificial intelligence (AI) enabled self-configuration, self-optimization and self-healing capabilities, not only to provide better quality of experience but also for their technical and commercial viability [1], [2]. Such AI driven automation will necessitate continuous gathering of telemetric data about network performance and coverage [3]. Currently, cellular network operators rely on drive tests, hardware or software failure alarms, and complaints received from their customers to measure the performance of their networks. However, these methods incur inevitable delays and unreliability that stems from human error and low spatio-temporal granularity of the gathered information [4]. These issues will most likely aggravate with the advent of small cell enabled ultra dense and complex networks, where the probability of cell outages will increase [5]. In addition, several use cases for 5G and beyond demand low latency and high reliability, which means that classic drive tests or alarm based approaches to performance monitoring and outage detection will not suffice [3], [6].

To overcome the aforementioned challenges, 3GPP has standardized minimization of drive test (MDT) that allows network performance estimation at a base station by leveraging measurement reports gathered at the user equipment (UE) without the need for drive tests [7]. The MDT reports contain network coverage related key performance indicators (such as received signal strength) measured at the UE. These reports are tagged with the UE's geographical location information and then sent to their serving base stations (BS) to generate coverage maps [8]- [10]. Using this MDT based coverage information, network operators can develop AI enabled autonomous mechanisms to compensate for outages quickly and seamlessly by enabling the network to detect anomalies such as coverage holes, weak coverage spots, sleeping cells, electromagnetic interference issues or diagnose the root causes of network issues [11]–[15]. Therefore, MDT based coverage and performance estimation is a fundamental step to enable AI based advanced self-configuration, self-optimization or self-healing routines [16]. However, utility of the MDT feature still remains hindered by the following three major types of errors:

1) Positioning error: The geographical location reported by the UE, determined using positioning techniques such as global positioning system (GPS), are susceptible to errors. Moreover, these locations may also be imprecise to preserve user privacy. This results in the reports being tagged to a wrong location [17].
2) Quantization error: Storing MDT reports from all users is computationally inefficient. Practical implementation demands that the coverage area be divided into bins. The reports from multiple UEs in each bin are averaged while building coverage maps. This results in quantization error due to averaging.
3) Scarcity of user reports: A key challenge in developing MDT based autonomous routines for ultra-dense deployments is that small cells contain far fewer users compared to macro cells. This makes the MDT reports from small cells sparse. This problem is further aggravated if smaller bin size is used to reduce quantization error due to the fact that many bins might not be visited by even a single user during the reporting period.

To enable accurate MDT-based performance estimation and in turn pave the way to fully autonomous network management, it is crucial to characterize and simultaneously address the aforementioned errors. Existing studies in literature either address the challenges of positioning inaccuracy, data sparsity and quantization individually or study only two of these errors jointly. None of the existing studies propose MDT-based performance estimation while taking into account all three errors jointly. There is a need to jointly characterize these errors because they are inter-dependent, as we later show in this paper. Only when these errors are jointly analyzed, interesting trade-offs are revealed and new insights are discovered, that enables the optimization of parameters for enhanced MDT-based performance estimation in a realistic scenario. The main advantages/utilities of the proposed method and the insights drawn therein are the following:

- We present a framework to determine the optimal bin width that will minimize the overall error in coverage estimation, even in the presence of sparse MDT reports. Given a certain user density and positioning error, network operators can therefore configure the bin width that will minimize the overall error in coverage estimation using the framework presented in this study.
- The findings and insights from this study can help network operators to calibrate the observed coverage in order to estimate the true coverage.
- Network operators can also determine the directionality of coverage estimation error (i.e., whether the coverage is over-estimated or under-estimated and by what amount) in a given area.
- Accurate coverage estimation can then enable network operators to solve many issues, such as detecting and compensating outages and anomalies including coverage holes, weak coverage spots, detecting sleeping cells, solving electromagnetic interference issues or other performance degradation problems.

To the best of authors' knowledge, this study is the first to not only present a framework to quantify these errors and characterize their interplay, but also to quantify the joint effect of these errors on overall coverage estimation.

A. RELATED WORK

We categorize previous studies related to our work into six groups based on whether they take into account the errors resulting from positioning, quantization, and data sparsity individually or jointly.

1) Positioning uncertainty only:

Authors in [18], [4], [17] addressed the reliability of MDT-based coverage estimation in the presence of positioning errors. A signal reliability expression and the cell coverage expressions that take the error in position estimation into consideration under shadowing and non-shadowing channel models is derived in [18]. The work in [18] is extended to incorporate base station location uncertainty in addition to user positioning error in [4], [17], where analytical model that allow the quantification of error in MDT-based coverage estimation as a function of user and base station positioning errors is developed. However, the impact of quantization or sparse MDT measurements is not the focus of these works. Acknowledging the limitations of position estimation methods such as GPS positioning or metrics such as observed time difference of arrival in combination with angles of reception, authors in [19] propose to use big data processing to obtain network performance, such as coverage evaluation as a function of location. To this end, the authors in this work leverage deep neural network and Bayesian probability theory-based techniques to reduce the number of required drive test measurements for LTE networks. They predict LTE signal quality metrics using drive test measurements using these techniques. However, deep learning-based training requires abundant data and is not likely to work in scenarios with sparse user data. Moreover, the joint impact of quantization and positioning inaccuracy is not the focus of the work in [19]. In contrast, presented work aims to address this challenge of sparse user data along with exploring the trade-off of bin size and positioning accuracy for enhanced MDT-based performance estimation.

The authors in [20] highlight the limitations of positioning estimation techniques, such as GPS, time of arrival, time difference of arrival and angle of arrival. They also point out that some of these techniques, like the time and angle of arrival-based methods are significantly impacted by the multi-path effect, and so will perform poorly if the UE is in non-line-of-sight environment. To address these challenges, they evaluate the estimation accuracy of radio frequency signatures with UE location in a dense urban area based on machine learning techniques of regression tree, random forest, neural networks, support vector machines and linear regression. The inputs to these models are combinations of RSRP, reference signal received quality, or physical cell ID of the serving and neighboring cell and the outputs of the estimation model are the latitude and longitude of the UE. However, it should be noted that a large amount of training data is required for machine learning techniques, especially neural networks. Moreover, the impact of quantization or sparse measurements was not the focus of this work. Authors in [21] propose a system that utilizes a collaborative filtering algorithm for improving robustness and accuracy of the MDT reports when both base station position and GPS MDT data are abnormal. After first estimating the position of base station, the abnormal GPS MDT data is detected and eliminated with the help of estimated base station position. They also propose a fast matching k-nearest neighbor algorithm to improve the location efficiency and reduce the location cost under the constraint of ensuring high location accuracy. However, the impact of varying grid sizes and sparse MDT reports and their joint interplay with each other and with positioning error is not investigated in this study.

Another attempt using a data-mining approach to enhance the GPS signal in order to overcome the degradation of the positioning accuracy due to noise of satellite-based positioning system is made in [22]. Authors in [22] build a large-scale precision GPS receiver grid system to collect real-time GPS signals for training. Gaussian process regression is chosen to model the vertical total electron content distribution of the ionosphere of the Earth. The experiments show that the noise in the real-time GPS signals often exceeds the breakdown point of conventional robust regression methods. To address this challenge, authors propose a Filter-Reweight-Retrain Algorithm for GPS signal enhancement. This consists of separating the signals into clean and noisy groups. Then an initial regression model on the clean signals is trained and the noisy signals are re-weighted based on the residual error. A final model is retrained on both the clean signals and the re-weighted noisy signals. However, this study does not consider MDT-based performance estimation nor does it consider the impact of bin width or sparse measurements. By using MDT reports, authors in [23] focus on characterizing multipath time, coherence bandwidth and doppler shift of propagation channels in mobile networks. In this study, the GPS position of UE is considered reliable if its uncertainty shape has a radius below 39 m. The bin size is taken to be 10 × 10 meters, because it was the highest resolution currently available with the visualization tool authors used. It is also assumed that a UE under excellent GPS coverage provides its location to the network with an accuracy of less than 10 meters. Therefore, this study is limited to a fixed bin size and location inaccuracy. In contrast, in this study, we comprehensively analyse MDT-based performance by varying positioning uncertainty radius and bin size and reveal new insights and solutions for performance enhancement that stem from analyzing the trade-off between these factors.

In summary, though positioning error is well investigated in literature as represented by the aforementioned studies i.e., [4], [17]–[23],], none of these existing studies take into account the errors resulting from quantization or scarcity of user measurements. This study aims to fill this gap by proposing enhanced MDT-based performance estimation methods while taking into account other errors such as sparse user measurements and quantization errors that were not taken into account by [4], [17]–[23].

2) Quantization only:

By considering only the quantization error, authors in [24] estimate cell radius. Using the assumption of uncorrelated lognormal shadowing, the authors in [24] analyze the quantization noise requirements of radio frequency prediction and coverage. The quantization error was represented as an equivalent error in the cell radius estimate. It is found that the minimum resolution bin size required to mitigate spatial quantization noise effects is roughly one-fortieth of the cell radius. However, this work does not consider the effect of sparse measurements or location uncertainty in the analysis. Moreover, the work in [24] does not use MDT-based approach for coverage estimation.

3) Data sparsity only:

The problem of sparsity of UE reports is investigated in [25]–[37]. Authors in [25] use regression clustering for construction of received signal strength maps from a sparse set of MDT measurements. The authors in [26] analyze the performance of selected spatial interpolation techniques used in the estimation of interference produced by an LTE-Advanced network. The authors in [27] provide a visualization method based on inverse distance weighted interpolation that shows every point of the received data as a heatmap. Another work [28] investigates several classical interpolation methods to reconstruct interference maps in cognitive radio networks. Authors in [29] propose a new technique, called Fixed Rank Kriging that is superior in terms of computational complexity as compared to Kriging. Authors in [30] use this technique to study the tradeoff between computational complexity and prediction accuracy when using Kriging to predict coverage, using real measurement data. The authors in [31] extend the work in [30] to a multi-cell scenario. Kriging-based prediction of propagation environment is presented in [32] for two different frequencies and environments. This work is extended to study how Kriging behaves in the presence of propagation model uncertainties, that stem from shadowing [33]. Studies that consider the recovery of sparse coverage data in an indoor environment include [34], [35], [36]. Using low-cost spectrum sensors in an office indoor environment, authors in [34] present an accuracy comparison between the spatial interpolation methods of Kriging, Gradient Plus Inverse Distance Squared and Inverse Distance Weighted methods. The results show that there is no significant difference in the accuracy for the considered interpolation methods, relative to the variability in the measurements reported by different low-cost devices. Another study in an indoor environment [35], analyzes several spatial interpolation techniques based on Inverse Distance Weighting (IDW) and compares them in terms of reliability bounds of interpolation errors. Authors in [36] compare various interpolation techniques, including Kriging, splines, weighted moving average, theissen polygons, trend surfaces, classification, in terms of accuracy, spatial distribution of measurements, measurement density and impact of a fixed fixed location inaccuracy in an indoor environment. However, assessing interpolation performance for a wide range of location uncertainties is not the focus of this work [36]. Instead, it considers the interpolation performance for an average location error of 18 meters only. Another methodology to build complete path loss grids for a given site from sparse user measurements is proposed in [37], where the path loss is estimated for missing locations by tuning a propagation model and extrapolating the path loss for neighboring pixels, using the available drive test measurements for certain pixels. For each of these pixels, a parameter from the standard propagation model (SPM) is tuned so that the resulting path loss equals the measured value. The remaining missing parameters of the SPM model are then estimated by weighting these tuned factors of the same clutter. Once all parameters are obtained, the path loss is estimated using SPM model, resulting in a completed path loss grid. However, this approach relies on a fixed bin width of 50 m and does not take into account user positioning errors. Moreover, it relies on an empirical path loss model, which is based on simplifications and does not capture the real world scenario features, such as detailed geographical information. In contrast, our study aims to evaluate, propose and optimize parameters for enhanced MDT-based performance estimation in a realistic scenario, while taking into account variable bin widths and positioning error uncertainties.

In summary, in contrast to present study, the effect of bin width and positioning error on the spatial interpolation techniques investigated in [25]–[37], remains unexplored.

4) Data sparsity and quantization:

Authors in [38], [39] use Bayesian kriging technique on cellular network data to build radio environment maps for the purpose of coverage hole detection. They show that the accuracy of such a technique is directly impacted by the bin size. Authors in [40] extend the work in [39] to include a more realistic coverage hole definition, where the coverage of neighboring pixels is also taken into account. A framework to establish a relationship between geographical data and user data using crowdsourced measurements in three region types: downtown, single-family residential, and multi-family residential is proposed in [41]. A neural network is trained to predict the key performance indicators in terms of RSRP and path loss estimation. The framework uses geographical features to predict the received signal level in different sub-regions. The impact of choosing different sizes of sub-regions (called tile size) is evaluated for three different sizes: 100-m, 200-m and 300-m square. Authors in [41] observed that too large or too small tile sizes hinders the ability of the model to capture the correlation between geographical characteristic changes and the resulting channel propagation. Since the 200-m square tile size showed the best performance for dataset used, this size is used for the remainder of the paper to study other issues related to key performance indicator prediction. This work does not consider the impact of positioning error in the proposed framework. Moreover, a large number of measurements are needed to train the neural network, before prediction of unknown measurements can take place.

In summary, the works in [38]–[41] investigate interpolation techniques i.e., quantization error while assuming perfect geolocation information. In contrast to presented study, these works do not take positioning uncertainty into account.

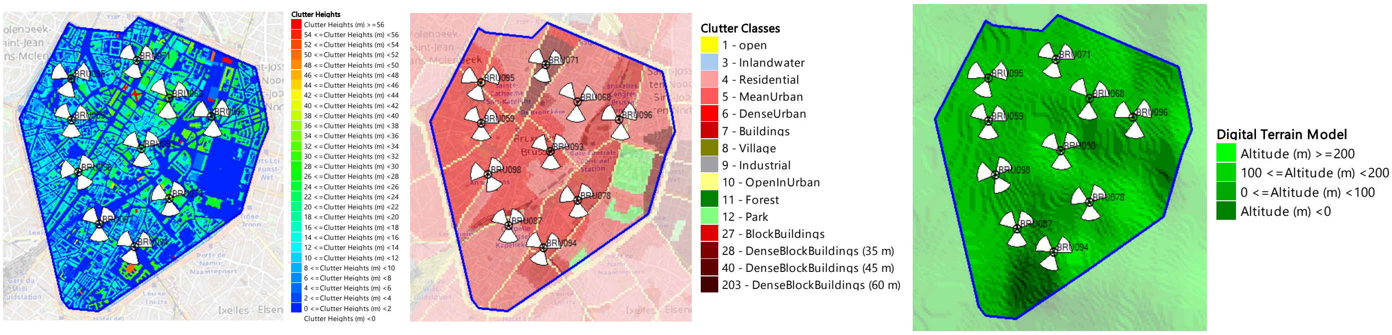
5) Positioning uncertainty and data sparsity:

One work that takes into account the impact of location uncertainty on sparse coverage data is [42], where the authors modify their earlier proposed algorithm [29], [30], [31] to incorporate the location uncertainty in the measurements. However, this work is limited to studying the impact of location uncertainty on the prediction algorithm and does not focus on the joint effect and inter-dependency of location uncertainty, quantization and sparsity on coverage estimation.

6) Positioning uncertainty and quantization:

In order to eliminate the RSRP difference due to inaccurate user positioning, authors in [43] use the calculation of RSRP difference of each UE based on multi-dimensional scaling algorithm. Based on the measured data, they achieve to concentrate the RSRP variation into a small range, which can effectively improve the positioning performance. In addition, based on the RSRP feature vector, authors propose a UE motion state classification method based on Adaboost method. Simulations are carried out for varying user positioning errors and four bin sizes with bin widths of 1 m, 5 m, 10 m and 15 m to test the positioning performance. Results using RSRP data show that the proposed positioning method can achieve almost 40m (at 85% CDF) positioning accuracy. However, this work requires data pre-processing techniques to prevent the effect of outliers, which is not measurement error. It is also concluded in [43] that the bin width monotonically decreases as the positioning accuracy increases. In contrast, analysis in this study considers the joint interplay of bin with and positioning accuracy, which leads to the existence of an optimal bin width, that we later show in this work. Moreover, sparsity of user reports is also not the focus of this work.

The most relevant study to our work is the study in [44], where authors determine optimal bin width by considering the impact of positioning uncertainty and quantization on coverage estimation using MDT. Our work differs from [44] in the following aspects: (1) In this work, we incorporate the effect of sparse user reports in MDT-based coverage estimation and its applications, which was not a focus of the


BRU095
BRU071
BRU068
BRU096
BRU059
BRU093
BRU098
BRU078
BRU087
BRU094
Clutter Heights
Clutter Heights (m) >=56
54 <=Clutter Heights (m) <56
52 <=Clutter Heights (m) <54
50 <=Clutter Heights (m) <52
48 <=Clutter Heights (m) <50
46 <=Clutter Heights (m) <48
44 <=Clutter Heights (m) <46
42 <=Clutter Heights (m) <44
40 <=Clutter Heights (m) <42
38 <=Clutter Heights (m) <40
36 <=Clutter Heights (m) <38
34 <=Clutter Heights (m) <36
32 <=Clutter Heights (m) <34
30 <=Clutter Heights (m) <32
28 <=Clutter Heights (m) <30
26 <=Clutter Heights (m) <28
24 <=Clutter Heights (m) <26
22 <=Clutter Heights (m) <24
20 <=Clutter Heights (m) <22
18 <=Clutter Heights (m) <20
16 <=Clutter Heights (m) <18
14 <=Clutter Heights (m) <16
12 <=Clutter Heights (m) <14
10 <=Clutter Heights (m) <12
8 <=Clutter Heights (m) <10
6 <=Clutter Heights (m) <8
4 <=Clutter Heights (m) <6
2 <=Clutter Heights (m) <4
0 <=Clutter Heights (m) <2
Clutter Heights (m) <0
Clutter Classes
1 - open
3 - Inlandwater
4 - Residential
5 - MeanUrban
6 - DenseUrban
7 - Buildings
8 - Village
9 - Industrial
10 - OpenInUrban
11 - Forest
12 - Park
27 - BlockBuildings
28 - DenseBlockBuildings (35 m)
40 - DenseBlockBuildings (45 m)
203 - DenseBlockBuildings (60 m)
Digital Terrain Model
Altitude (m) >=200
100 <=Altitude (m) <200
0 <=Altitude (m) <100
Altitude (m) <0

(a) Digital height model.

(b) Digital land use map.

(c) Digital terrain model.

FIGURE 1: System model configuration and geographical information.

TABLE 1: Network Scenario Settings.

System Parameters	Values
Carrier Frequency	2100 MHz
Maximum transmit power	43 dBm
Cell sectors	3 sectors per BS
Path loss model	Aster propagation (ray-tracing)
Propagation matrix resolution	5 m
BS height	30 m
Geographical information	Ground heights, building heights, land use map
User distribution	Poisson Distribution
Antenna gain	18 dBi
Horizontal half power beamwidth	63^o
Vertical half power beamwidth	4.7^o

study in [44]. Incorporating this error in coverage estimation is vital because data sparsity is a fundamental challenge that can become bottleneck for MDT-based coverage estimation. To this end, this work analyzes the effect of quantization, positioning uncertainty and sparsity of MDT-data independently as well as studies their combined effect on coverage estimation and its potential applications. (2) In contrast to the study in [44] that investigates the coverage estimation error by considering its mean value, we treat the errors as random variables and determine their distributions. (3) We present a solution to minimize the errors incurred in coverage estimation due to positioning uncertainty and quantization by presenting results and analysis that can enable network operators to calibrate the observed coverage in order to estimate the true coverage. We do so by not only quantifying the coverage estimation errors due to different factors, but also determining the directionality of coverage estimation error (i.e., whether the coverage is over-estimated or under-estimated and by what amount). Such coverage calibration and directionality of coverage estimation error is not considered in [44]. (4) The study in [44] considers a fixed coverage probability threshold. In contrast, we present a generic analysis by considering the difference between the actual and perceived RSRPs. Specific operator defined threshold-based coverage estimation errors can be easily derived from the results and analysis presented in this paper.

B. CONTRIBUTIONS AND ORGANIZATION

The organization and key contributions of this paper can be summarized as follows:

- We quantify the effect of positioning error without bins (Section III-A-1) and its joint effect with bins (Section III-A-2) on coverage estimation. By analyzing the error distributions, we analytically express the probability density function (PDF) of these errors as a function of positioning error radius and bin width.
- We determine the distributions of coverage estimation error caused by quantization without positioning error (Section III-B-1) and concurrently with positioning error (Section III-B-2).
- By analyzing combined effect of quantization and incorrect user positioning on coverage estimation, we show that the effect of quantization is not independent of positioning error radius and vice versa (Section III-C).
- We investigate the error in MDT based coverage estimation caused by scarcity of user measurements by leveraging existing and potential new techniques. We also study the joint effect of this error in tandem with other sources of errors, i.e, characterize the error in coverage estimation due to sparsity as a function of positioning error radius and bin width. (Sections III-D and III-E).
- Utility of this study is discussed in Section IV by considering two practical applications:
  - Coverage calibration: From a network design and optimization perspective, it is necessary not only to know how much coverage area is misclassified, but also to know the directionality of misclassified coverage (i.e., whether the coverage area is over-estimated or under-estimated and by what amount). We address this by utilizing the expressions derived for various types of errors in the preceding sections (Section IV-A).
  - Determining optimal bin width: While on one hand, decreasing bin size reduces the quantization error, on the other hand, it increases the error in coverage estimation due to incorrect user positioning as well as the error stemming from sparsity. This calls for an optimization of bin width that would minimize the

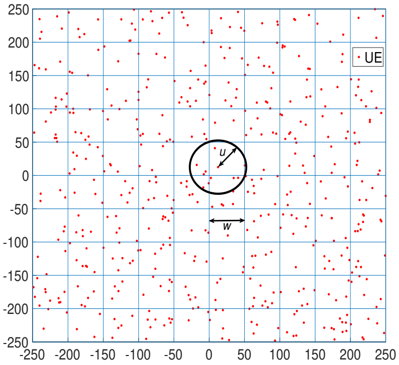


FIGURE 2: User distribution

TABLE 2: List of acronyms.

Acronym	Meaning
MDT	Minimization of Drive Test
AI	Artificial Intelligence
UE	User Equipment
LTE	Long-Term Evolution
GPS	Global Positioning System
RSRP	Reference Signal Receive Power
PDF	Probability density function
RV	Random variable
SVT	Singular Value Thresholding
FPC	Fixed Point Continuation
IDW	Inverse Distance Weighted
DTM	Digital Terrain Model
3GPP	3rd Generation Partnership Project

overall error under positioning error and user sparsity constraints. To the best of authors' knowledge, this paper is the first to show that there exists an optimal bin width which minimizes all three errors concurrently. Leveraging this finding, we present a framework to determine the optimal bin width that minimizes these errors simultaneously (Section IV-B).

- This work is concluded in Section V.

II. SYSTEM MODEL

We use a ray-tracing based commercial planning tool to create a sophisticated network topology (Fig. 1), in order to generate the MDT data in our study [45]. For the calibration of propagation model, environmental conditions, terrain profile and buildings were considered and also validated through drive tests in the simulator. Therefore, it can be assumed that coverage data obtained from this simulator represents the ground truth very closely in the area under consideration.

Users are distributed according to Poisson distribution. The area of interest is divided into n^2 bins of width, w as shown in Fig. 2 for 500 users and bin width of 50m. Given a reported UE position, we assume that its actual location is within a circular disc with radius u which is centered at the reported UE position, as illustrated in Fig. 2 for one UE. Hence, the actual position of the i^{th} UE with coordinates (x_i, y_i) is generated as $(x_i + u\sqrt{v_i}\cos(2\pi q_i), y_i + u\sqrt{v_i}\sin(2\pi q_i))$, where v_i and q_i are pseudo random, pseudo independent numbers uniformly distributed in [0, 1]. The shadowing effect is modeled by a random variable, which follows a zero mean Gaussian distribution with standard deviation ϕ in dB, based on clutter type.

The network scenario settings are reported in Table 1. The geographical data comprising of geo raster data and geo vector data was used to depict a realistic scenario. The raster data gives a grid-based representation of the terrain with a defined resolution. The raster files we used are DTM (Digital Terrain Model) representing the elevation of the ground over sea level, clutter classes representing the type of terrain (land cover or land use) and clutter heights (also called a digital height model) representing individual heights (altitude of clutter over the DTM, for example, building heights). Each pixel of a clutter class file contains a code which corresponds to a certain type of ground use or cover. In the clutter height file, a height is given for each point on the map. The geo vector data models the buildings and their height, in the form of one or several ArcView SHP files. All of these geo files were incorporated into our model to represent a realistic scenario. The path loss model was chosen to be aster propagation model, rather than empirical or semi-empirical path loss models [46]–[49], that are based on measurements in a specific environment and limited in their ability to capture idiosyncrasies of various propagation environments. In contrast, the aster propagation model is based on advanced ray-tracing propagation techniques and incorporates vertical diffraction over roof-tops, horizontal diffraction/reflection based on ray-launching and ray-tracing calculation on raster data as well as on vector building data. It also has the support of automatic calibration using continuous wave to further calibrate the model. All these features of aster propagation model enabled the modeling of a realistic network scenario. Considering the benefits of spectrum reuse and decreased the co-channel interference, we chose 3 sectors per cell over the conventional omni-cells [50]. Since cell sectoring improves the signal-to-interference ratio using a directional antenna, we chose a practical directional antenna model. Instead of modeling the antenna pattern through an analytical equation, which is often based on assumptions and simplifications, we used a practical 3-D antenna model which is composed of both horizontal and vertical antenna patterns. The antenna datasheets are incorporated into the simulator for more realistic modeling. A carrier frequency of 2100 MHz (also called Band 1 by the 3GPP [51]) was used because it is one of the 2100MHz is the most widely used band in the world [52]. Base station height of 30 m was chosen, because based on information from our industry experience and collaborators, 30m to 45m is common macro cell height for successful communication. A very high base station height will overshoot effective communication and a low height will undershoot it [53], [54]. The maximum power is regulated to be 43 dBm by the FCC [55]. That is why we chose a maximum transmit for 43 dBm for our study. It is the commonly used maximum transmit power for macro cells by various operators and recommended by

TABLE 3: List of symbols.

Symbol	Description	Symbol	Description
n	Number of bins/grids	w	Bin width in meters
w_{max}	Maximum bin width, w in meters	w_{min}	Minimum bin width, w in meters
x, y	Actual coordinates of UE. x_i, y_i are coordinates of i-th UE	u	Positioning error radius in meters
v, q	Pseudo random, pseudo independent numbers uniformly distributed in [0, 1] used in modeling the reported position of UE. v_i, q_i is i-th realization of these RVs	$E^{P,Q'}$	Coverage estimation error due to positioning uncertainty in the absence of bins. $e^{P,Q'}$ is a realization of this RV. $f_E^{P,Q'}(e^{P,Q'})$ is its PDF
$r^{P',Q'}$	Received signal strength in dBm of UE without positioning uncertainty	$r^{P,Q'}$	Measured/perceived received signal strength in dBm of UE in the presence of positioning uncertainty
$\boldsymbol{C}$	$n \times n$ sparse matrix of coverage data. C_{ij} is the entry located at the i-th row and j-th column of $\boldsymbol{C}$	s_1	Scale parameter of logistic distribution, $E^{P,Q'}$. It is proportional to the standard deviation of $E^{P,Q'}$
$E^{Q,P'}$	Coverage estimation error due to quantization error without positioning uncertainty. $e^{Q,P'}$ is a realization of this RV. $f_E^{Q,P'}(e^{Q,P'})$ is its PDF	$E^{P,Q}$	Coverage estimation error due to positioning uncertainty in the presence of bins. $e^{P,Q}$ is a realization of this RV. $f_E^{P,Q}(e^{P,Q})$ is its PDF
$E^{Q,P}$	Coverage estimation error due to quantization error due to quantization in the presence of positioning uncertainty. $e^{Q,P}$ is a realization of this RV. $f_E^{Q,P}(e^{Q,P})$ is its PDF	E^C	Coverage estimation error due to both positioning uncertainty and quantization. e^C is a realization of this RV. $f_E^C(e^C)$ is its PDF
$r^{P,Q}$	Measured averaged received power of UEs in a bin in the presence of positioning uncertainty	$r^{P',Q}$	Averaged received power of UEs in a bin with no positioning uncertainty
μ_2	Location parameter/mean of logistic distribution, $E^{P,Q}$	μ_4	Location parameter/mean of logistic distribution, $E^{Q,P}$
s_2	Scale parameter of logistic distribution, $E^{P,Q}$. It is proportional to the standard deviation of $E^{P,Q}$	s_4	Scale parameter of logistic distribution, $E^{Q,P}$. It is proportional to the standard deviation of $E^{Q,P}$
s_3	Variance of $E^{Q,P'}$	μ_5	Location parameter/mean of logistic distribution, E^C
$\{a_1 \dots d_1\}$	Parameters of s_1 obtained through distribution fitting	$\{e_2 \dots n_2\}$	Parameters of s_2 obtained through distribution fitting
$\{a_2 \dots d_2\}$	Parameters of μ_2 obtained through distribution fitting	$\{a_3 \dots d_3\}$	Parameters of s_3 obtained through distribution fitting
$\{h_4 \dots r_4\}$	Parameters of s_4 obtained through distribution fitting	$\{a_4 \dots g_4\}$	Parameters of μ_4 obtained through distribution fitting
s_5	Scale parameter of logistic distribution, E^C. It is proportional to the standard deviation of E^C	$\{\boldsymbol{P}, \boldsymbol{Q}\}$	Sequence of matrices produced by the iterative SVT algorithm. $\{\boldsymbol{P}^t\}$ and $\{\boldsymbol{Q}^t\}$ are the matrices at t-th step
$\{a_5 \dots i_5\}$	Parameters of μ_5 obtained through distribution fitting	$\{j_5 \dots l_5\}$	Parameters of s_5 obtained through distribution fitting
ϕ	Standard deviation of Gaussian shadowing in dB	$\hat{\boldsymbol{C}}$	Reconstructed/estimated matrix $\boldsymbol{C}$
Ψ	A set of cardinality m sampled at random	σ_k	k^{th} Largest singular value of a matrix
$\mathcal{O}_\Psi$	Orthogonal projector onto the span of matrices vanishing outside of Ψ	η	Regularization parameter in objective function of SVT algorithm
$\{\Delta_i\}$	Sequence of scalar step sizes	$\mathcal{S}_\eta$	Shrink function that applies soft-thresholding rule at level η
$\boldsymbol{U}, \boldsymbol{V}$	Matrices with orthonormal columns, obtained by singular value decomposition of $\boldsymbol{P}$	C_m	Missing coverage value at a particular bin location in matrix $\boldsymbol{C}$, $\hat{C_m}$ is its estimate
ζ	A fixed tolerance in the stopping criteria of SVT algorithm	m	Total number of coverage entries in coverage matrix $\boldsymbol{C}$
r	Rank of a matrix	$\boldsymbol{\Sigma}$	Singular values of $\boldsymbol{P}$
$A_o^Q(u, w)$	Probability of area that is over-estimated due to quantization	p	Distance decay parameter in IDW algorithm
V_{n_k}	Voronoi region of 2-D point n_k	E^M	Error in covering missing coverage values
$A_u^P(u, w)$	Probability of area whose coverage is under-estimated due to given positioning uncertainty	d_k	Distance between the location of the bin with missing coverage value and location of the k-th bin
$A_u^c(u, w)$	Probability of area that is under-estimated due to both quantization and user positioning error	$\boldsymbol{C}^{full}$	Matrix with full entries, considering that RSRP measurements are available from all bins
$E_B^{Q,P}$	Bounded error $E^{Q,P}$ between 0 and 1	E_B^C	Bounded error E^C between 0 and 1
E_B^M	Bounded error E^M between 0 and 1	$E_B^{P,Q}$	Bounded error $E^{P,Q}$ between 0 and 1
$\hat{\boldsymbol{c}}, \boldsymbol{c}^{full}$	Vectorized forms of matrix $\hat{\boldsymbol{C}}, \boldsymbol{C}^{full}$	w^*	Optimal bin width for coverage estimation

standards [56]. A 5m resolution was sufficient to capture the propagation characteristics for our study. A lesser resolution was providing redundant information and a using a greater resolution lead to missing information. The distribution of users is modeled according to a Poisson distribution. Rather than using a grid-based or uniformly distributed users, that is is too idealized, we have obtained a realistic user distribution by generating the user distribution using a Monte Carlo algorithm. This user distribution is based on the traffic database and traffic maps and is weighted by a Poisson distribution between simulations of the same group. With this modeling, there could be some users arbitrarily close to each other, thus providing a more realistic depiction [57].

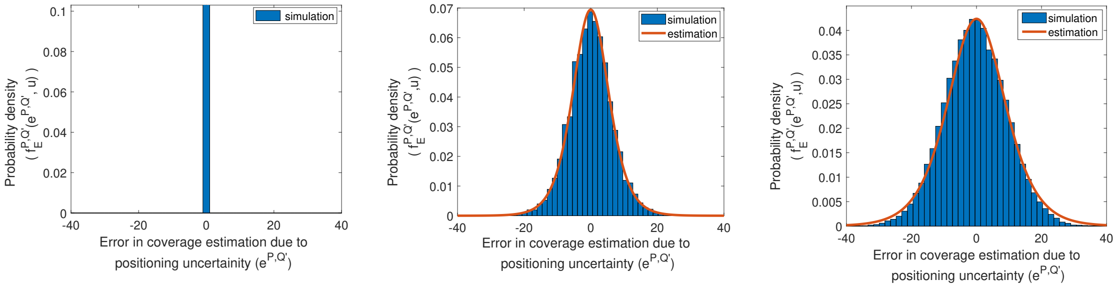
The list of acronyms and symbols used in this paper are given in Table 2 and 3 respectively.

III. QUANTIFICATION OF ERRORS IN AUTONOMOUS COVERAGE ESTIMATION USING MDT

A. ERROR DUE TO USER POSITIONING UNCERTAINTY

1) Error due to user positioning uncertainty without bins

In this section, we address the following challenge: When the coverage area is not divided into bins, how much coverage is misclassified due to positioning uncertainty as a function of positioning error radius? To address this, we express the error



(a) $u = 0m$ **(b)** $u = 10m$ **(c)** $u = 100m$

FIGURE 3: PDF of coverage estimation error due to positioning uncertainty in the absence of bins

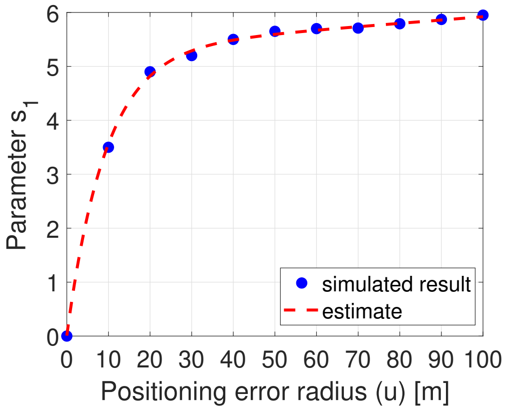


FIGURE 4: Parameter s_1

as a random variable due to random distributions of reported and actual positions of users as:

$$E^{P,Q'}(x,y,v,q,u) = r^{P,Q'}(x,y,v,q,u) - r^{P',Q'}(x,y) \tag{1}$$

where the superscripts P and P' indicate the presence and absence of user positioning error respectively. The superscript Q' indicates no quantization. $r^{P,Q'}$ is the measured/perceived received signal strength of the user in the presence of positioning uncertainty (in dBm). This is a function of UE actual coordinates, x and y, positioning error radius, u and random variables v and q, which are used in the modeling the reported UE position. Since $r^{P',Q'}$ is the received signal strength of the user without positioning uncertainty (in dBm), it is not a function of positioning error radius, u.

Note that the RSRP of the users would not be affected by positioning error, however, the measured RSRP reports would be tagged to wrong locations due to positioning error since the received signal estimation is based on the measurement report, which is tagged to a wrong position. Thus, the PDF of coverage estimation error due to positioning uncertainty at the user-level, $f_E^{P,Q'}(e^{P,Q'})$ represents the probability of users that are misclassified by a certain amount due to positioning uncertainty. Note also that since the random variable, $E^{P,Q'}$ is a difference in dB, the probability that this random variable takes on a value greater than 0 ($E^{P,Q'} = e^{P,Q'} > 0$) represents the probability of users whose coverage is over-estimated by $e^{P,Q'}$ and $f_E^{P,Q'}(e^{P,Q'})$ corresponding to $E^{P,Q'} = e^{P,Q'} < 0$ represents the probability of users whose coverage is under-estimated by $e^{P,Q'}$.

In our simulations, the bin width is varied from $w_{min} =$ 10m to $w_{max} =$ 50m and u is varied from 0m to 100m. Fig. 3 illustrates the PDF of coverage estimation error due to positioning uncertainty in the absence of bins for $u = 0, 10$ and 100m. It can be observed that the variance of this error increases with increase in positioning error radius. Using distribution-fitting tools, we determine that this error distribution follows a Logistic Distribution with zero mean and parameter s_1, that is proportional to the square root of variance. This parameter can be interpreted as a scaled measure of how much variation or dispersion exists from the mean. It is a function of positioning uncertainty, u. Using multiple terms exponential regression, we determine parameter s_1 as a function of positioning error radius as follows:

$$s_1(u) = a_1\exp(b_1 u) + c_1\exp(d_1 u), \tag{2}$$

where $a_1 = 5.333, b_1 = 0.001, c_1 = -5.325, d_1 = -0.107$

Fig. 4 shows the variation of parameter s_1 with u. This parameter increases continuously with increase in positioning uncertainty because as the positioning error radius gets bigger, the probability of reported MDT geographical coordinates being farther away from its actual coordinates increases. In addition, the number of possibilities of discrepancies between actual and reported locations increase as positioning uncertainty increases, hence leading to an increased variation from the mean of the coverage error estimation due to positioning uncertainty. The PDF of this error as a function of positioning error radius then becomes:

$$f_E^{P,Q'}(e^{P,Q'},u) = \frac{\exp\left(-\frac{e^{P,Q'}}{a_1 e^{(b_1 u)} + c_1 e^{(d_1 u)}}\right)}{\left(a_1 e^{(b_1 u)} + c_1 e^{(d_1 u)}\right)\left(1+\exp\left(-\frac{e^{P,Q'}}{a_1 e^{(b_1 u)} + c_1 e^{(d_1 u)}}\right)\right)^2} \tag{3}$$

where $e^{P,Q'}$ is one realization of the random variable, $E^{P,Q'}$ and u is the positioning error radius. The parameters $\{a_1 \ldots d_1\}$ would vary with different path loss and shadowing models. However, this is out of scope of this paper and can be part of a future work. Moreover, the errors in coverage estimation are quantified as errors between the actual and perceived RSRP measurements in this work. For generality, we do not consider a specific RSRP threshold-based coverage definition. However, the coverage estimation error based on different RSRP thresholds can be easily inferred from our results and analysis, i.e., by truncating the PDFs according to different operator-defined coverage (RSRP) error thresholds.

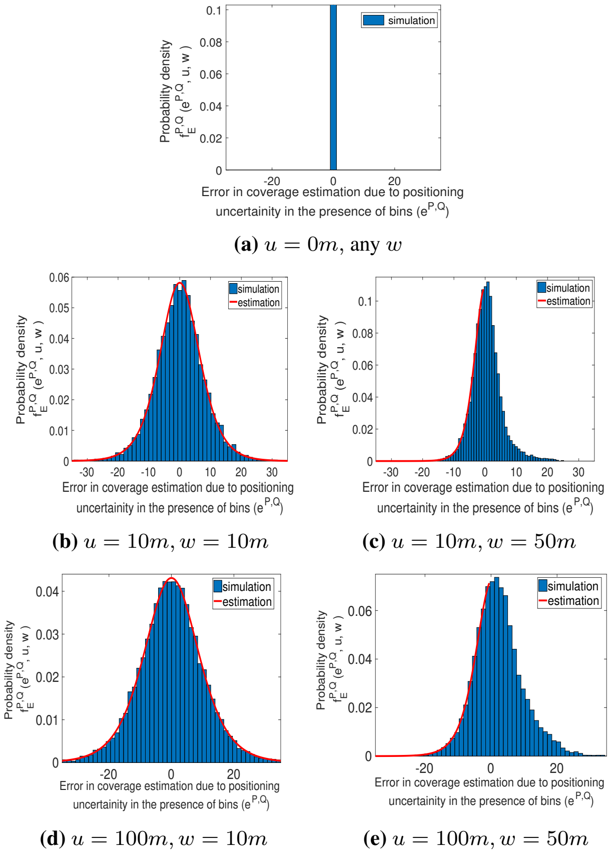
2) Error due to user positioning uncertainty with bins

The preceding section quantified the impact of user positioning error on coverage estimation without binning. In scenarios where the coverage area is divided into bins, the coverage estimation would be impacted by both positioning error as well as the bin width. To address this case, we consider the following error measure:

$$E^{P,Q}(x,y,v,q,u,w) = r^{P,Q}(x,y,v,q,u,w) - r^{P',Q}(x,y,w) \tag{4}$$

where $r^{P,Q}$ is the measured averaged received power of users in a bin of width w in presence of positioning uncertainty and $r^{P',Q}$ is the averaged received power of users in the same bin with no positioning uncertainty. Since this error in coverage estimation characterizes the effect of positioning uncertainty when the coverage area is divided into bins, it is a function of bin width, w, in addition to the UE location parameters. The integral of PDF of this error from $0 < E^{P,Q} < \infty$ thus represents the percentage of misclassified area that is over-estimated on average and the integral of PDF from $-\infty < E^{P,Q} < 0$ represents the percentage of misclassified area that is under-estimated.

The effect of positioning error on coverage estimation with varying bin widths is shown in Fig. 5. It can be observed from Fig. 5 (b)-(e) that for the same positioning error radius, the variance of this error decreases as the bin width increases, attributing to the fact that for the same positioning error radius, the effect of positioning error on coverage estimation will be greater when bin width is small as the probability of a user being actually located in adjacent bins instead of the reported bin is likely to increase with decreasing bin width. Similarly, for the same bin width, the error variance increases with increasing positioning error radius. Note that since the plotted error in coverage estimation captures the effect of positioning error only, it approaches the delta function as the positioning error radius reduces to 0m, as shown in Fig. 5 (a). It can also be observed from Fig. 5 (d) and (e), that for the same user positioning uncertainty, the percentage of area that is falsely estimated to be covered (i.e., over-estimated coverage) increases with increase in bin width. These findings can be used to calibrate the coverage estimated through MDT, for given values of positioning error radius and bin width. In order to facilitate this goal, we determine an analytical expression by performing distribution fitting for part of the PDF of $E^{P,Q}$ for a range of bin widths and positioning error radii, yielding the following expression:



(a) $u = 0m$, any w

(b) $u = 10m, w = 10m$

(c) $u = 10m, w = 50m$

(d) $u = 100m, w = 10m$

(e) $u = 100m, w = 50m$

FIGURE 5: PDF of coverage estimation error due to positioning uncertainty in the presence of bins

$$f_E^{P,Q}(e^{P,Q},u,w) = \frac{\exp\left(-\frac{e^{P,Q}-\mu_2(u,w)}{s_2(u,w)}\right)}{s_2(u,w)\left(1+\exp\left(-\frac{e^{P,Q}-\mu_2(u,w)}{s_2(u,w)}\right)\right)^2},$$

$$\forall\ e^{P,Q} \text{ when } \mu_2 = 0, \text{ for } e^{P,Q} < 0 \text{ when } \mu_2 \geq 0 \tag{5}$$

where μ_2 and s_2 are as follows:

$$\mu_2(u,w) = a_2u + b_2w + c_2u^2 + d_2uw; \tag{6}$$

$$\begin{aligned} s_2(u,w) = (e_2w^{f_2} + g_2)\exp(h_2u/(w+i_2)) \\ +(j_2w^{k_2} + l_2)\exp(m_2wu + n_2u) \end{aligned} \tag{7}$$

where $a_2 = -0.00262, b_2 = 1.587 \times 10^{-5}, c_2 = 1.124 \times 10^{-5}, d_2 = 0.0001663, e_2 = -0.00473, f_2 = 1.538, g_2 = 5.04, h_2 = 0.04628, i_2 = 13.67, j_2 = 0.004732, k_2 = 1.538, l_2 = -5.04, m_2 = 0.001935, n_2 = -0.2277.$ $e^{P,Q}$ represents one realization from the distribution of the error variable, $E^{P,Q}$ and u and w are the positioning error radius and bin widths, respectively.

The parameters μ_2 and s_2 are shown in Fig. 6 and 7, respectively and indicate an excellent fit between the simulated results and parameter fitting. These parameters are both functions of the bin width, w and the positioning error radius, u and represent the mean and measure of variability in the error distribution, $E^{P,Q}$. For the purpose of determining

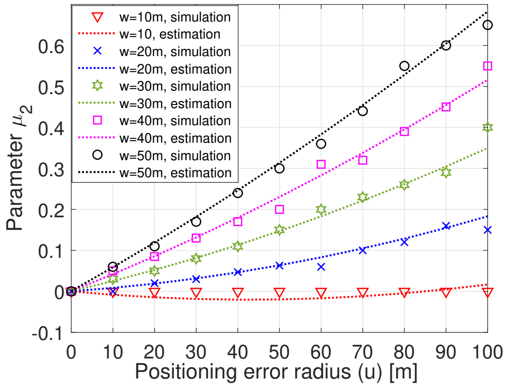


FIGURE 6: Parameter μ_2

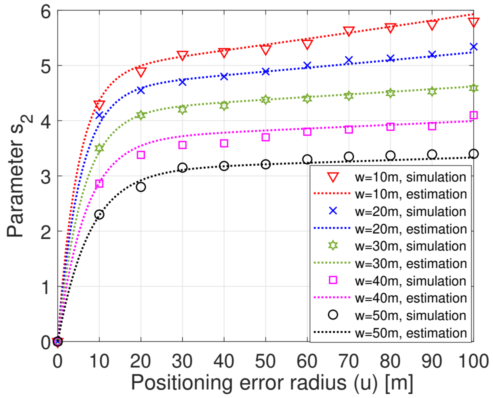


FIGURE 7: Parameter s_2

the directionality of misclassified coverage, and ultimately calibrating for correct coverage estimation, the percentage of area that is under-estimated is sufficient since the remaining fraction would be the percentage of area that is over-estimated. Further discussion on utility of these results from coverage calibration perspective is presented in Section IV.

B. QUANTIZATION ERROR

1) Quantization error without positioning uncertainty

We quantify the error in coverage estimation incurred due to averaging by dividing the coverage area into bins as:

$$E^{Q,P'}(x,y,w) = r^{P',Q}(x,y,w) - r^{P',Q'}(x,y) \qquad (8)$$

where $r^{P',Q'}$ is the received signal strength of a user without positioning inaccuracy and $r^{P',Q}$ is the averaged received signal strength being reported from the bin in which the same user resides, in absence of positioning inaccuracy. Alternatively, $r^{P',Q}$ is the averaged received signal strength that is being reported from a bin, where a user resides with an individual received signal strength equal to $r^{P',Q'}$. Assuming a constant user density, $r^{P',Q}$ is a function of bin width in addition to user locations since a larger bin width would mean more spatially spread users with more widely different received powers in that bin, resulting in greater averaging error, whereas a smaller bin width would mean lesser averaging error.

Fig. 8 depicts the PDF of $E^{Q,P'}$ with varying bin widths. This error PDF captures the effect of quantization error on coverage estimation, assuming perfect geo-location information. It can be observed from Fig. 8 that the spread of this error increases as the bin with increases, attributing to the fact that more and more users lie in a bin as the bin width increases and hence the error stemming from the averaging of more users reflect the coverage estimation. This spread of the error distribution is captured quantitatively by its variance, s_3 in Fig. 9, which shows how the variance of coverage estimation error due to quantization increases as the bin with increases. This error converges to the delta distribution as $w \rightarrow 0$ and we can observe that the higher frequency of zero coverage estimation error as $w \rightarrow 0$. This is because as $w \rightarrow 0$, effectively, each user converges to a single bin. The variance of this error, which is a function of bin width, w is depicted in Fig. 9. It increases with increase in bin width according to:

$$s_3(w) = a_3 \exp(b_3 w) + c_3 \exp(d_3 w), \qquad (9)$$

where $a_3 = 42.34, b_3 = 0.005597, c_3 = -42.16, d_3 = -0.2408$

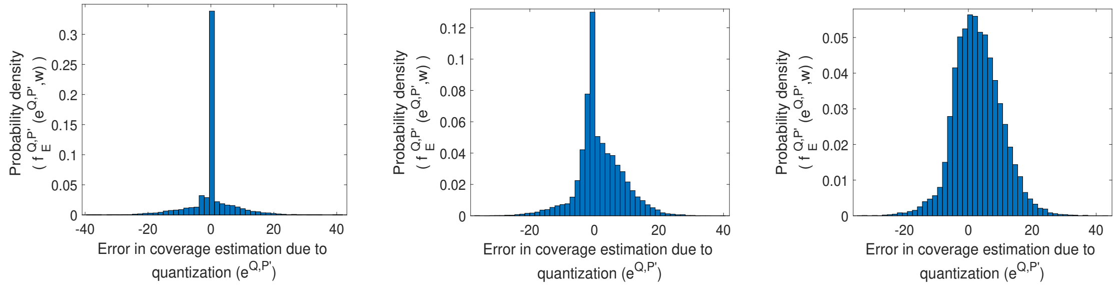
2) Quantization error with positioning uncertainty

We now investigate how the presence of user positioning uncertainty changes the distribution of coverage estimation error due to quantization that has been illustrated in the previous section. More specifically, for a given positioning uncertainty, in order to correctly calibrate coverage, it is important for the network operator not only to know how much coverage is misclassified, but also know the directionality of misclassified coverage (i.e., whether the coverage is over-estimated or under-estimated and by what amount). To aid this goal, we define the error in coverage estimation due to quantization in the presence of positioning uncertainty as:

$$E^{Q,P}(x,y,v,q,u,w) = r^{P,Q}(x,y,v,q,u,w) - r^{P,Q'}(x,y,v,q,u) \qquad (10)$$

where $r^{P,Q}$ is the measured averaged received signal strength that is being reported from a bin, where a user is reported to reside in the presence of positioning uncertainty with its measured received power (at user-level) equal to $r^{P,Q'}$ in the presence of the same positioning uncertainty.

Fig. 10 illustrates the PDF of this error with varying bin widths and positioning error radius. It is observed that $E^{Q,P} \rightarrow E^{Q,P'}$ as $u \rightarrow 0$ as Fig. 10 (a), (c), (e) converge to Fig. 8 (a), (b), (c) respectively. This is because when u approaches 0, the effect of positioning uncertainty diminishes and so the coverage estimation error in the presence of positioning uncertainty, $E^{Q,P}$ converges to that without any positioning uncertainty, $E^{Q,P'}$. However, as u increases, the variance of this error and the percentage of area that is falsely



(a) $w = 10m$ **(b)** $w = 30m$ **(c)** $w = 50m$

FIGURE 8: PDF of coverage estimation error due to quantization error without positioning uncertainty

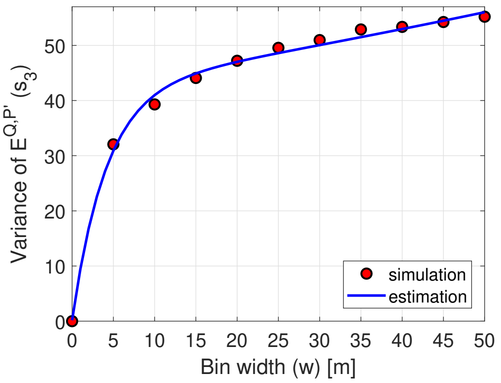


FIGURE 9: Variance of error, $E^{Q,P'}$

estimated to be covered starts to increase. The PDF of this error can be expressed as follows:

$$f_E^{Q,P}(e^{Q,P},u,w) = \frac{\exp\left(-\frac{e^{Q,P}-\mu_4(u,w)}{s_4(u,w)}\right)}{s_4(u,w)\left(1+\exp\left(-\frac{e^{Q,P}-\mu_4(u,w)}{s_4(u,w)}\right)\right)^2}, \text{for } e^{Q,P} > 0\ \forall\ \mu_4 \qquad (11)$$

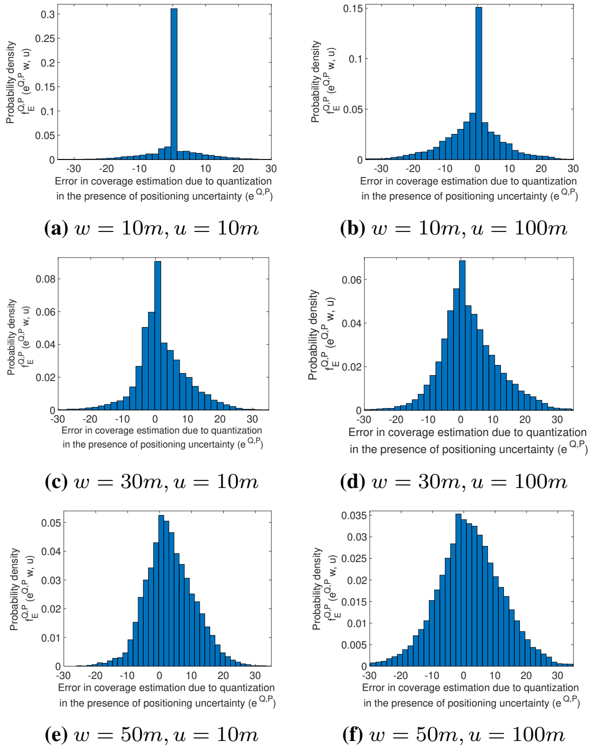
where

$$\mu_4(u,w) = (a_4 w^{b_4} + c_4)\exp\left(d_4 u \exp(e_4 w) + f_4 u \exp(g_4 w)\right) \qquad (12)$$

where $a_4 = -0.0002718, b_4 = 1.948, c_4 = 0.1824, d_4 = 0.03437, e_4 = -0.05875, f_4 = 0.0003263, g_4 = 0.07777$ and

$$s_4(u,w) = (h_4 w^3 + i_4 w^2 + j_4 w + k_4)u^2 + (l_4 w^3 + m_4 w^2 + n_4 w + o_4)u + p_4 w^{q_4} + r_4 \qquad (13)$$

with $h_4 = -1.086e-08, i_4 = 9.973e-07, j_4 = -2.307e-05, k_4 = 0.0002294, l_4 = 1.325e-06, m_4 = -0.0001289, n_4 = 0.00371, o_4 = -0.02346, p_4 = -20.58, q_4 = -0.09358, r_4 = 21.17$. The variable $e^{Q,P}$ represents an instance of the random variable that captures the effect of quantization with positioning uncertainty on coverage estimation. Fig. 11 shows the mean of this error, given by (12) and Fig. 13 illustrates the scale parameter s_4 of this error distribution, that is proportional to the standard deviation of this error and can be interpreted as a scaled measure of how much variation or dispersion exists from the mean. Fig. 11 and 12 illustrate the excellent fit between simulated parameters and (12)-(13).



(a) $w = 10m, u = 10m$ **(b)** $w = 10m, u = 100m$

(c) $w = 30m, u = 10m$ **(d)** $w = 30m, u = 100m$

(e) $w = 50m, u = 10m$ **(f)** $w = 50m, u = 100m$

FIGURE 10: PDF of coverage estimation error due to quantization in the presence of positioning uncertainty

By comparing Fig. 5 and Fig. 10, we observe contrary trends in the error variance with varying bin widths and positioning error radius: contrary to $E^{P,Q}$, the variance of $E^{Q,P}$ increases with increase in bin width for a fixed positioning error radius. This is because $E^{Q,P}$ characterizes the effect of quantization for a fixed positioning error radius, which increases with increase in bin width, owing to greater averaging error of received signal strength measurements with increase in bin width. On the other hand, $E^{P,Q}$, captures the effect of positioning error radius on coverage estimation. The effect of positioning error becomes more profound with decrease in bin width as the probability of a user being actually located in adjacent bins and not the reported bin

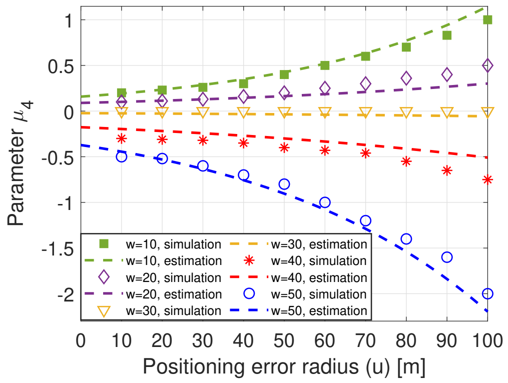


FIGURE 11: Parameter μ_4

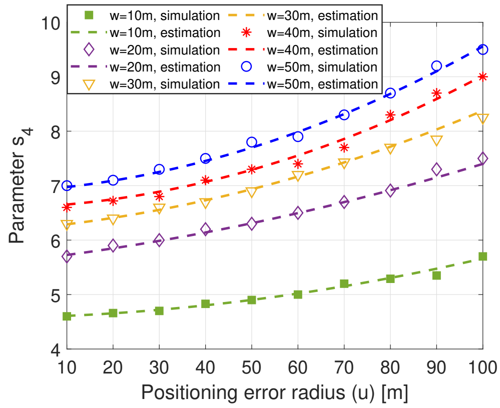


FIGURE 12: Parameter s_4

increases with decrease in bin width, for a fixed positioning error radius.

C. COMBINED EFFECT OF POSITIONING AND QUANTIZATION ERRORS ON COVERAGE ESTIMATION

In Section III-A, we analyzed the effect of positioning error on coverage estimation, both with quantization and without quantization, while in Section III-B, we investigated the effect of quantization error on coverage estimation, both with and without positioning uncertainty. The results and analysis from the preceding section serve as a basis for coverage calibration for a given bin width or a given user positioning error. However, in applications where the goal is to minimize effect of both errors simultaneously, following questions arise:

- Is the impact of user positioning error on coverage estimation independent of quantization error?
- If the two errors are dependent, how do they affect coverage estimation using MDT?

We will begin by addressing the first question in this section. This paper, for the first time investigates the concurrent effect of user positioning uncertainty and quantization on coverage estimation. In order to reveal this interplay, we characterize the error in coverage estimation due to both positioning and quantization errors as follows:

$$E^C(x,y,v,q,u,w) = r^{P,Q}(x,y,v,q,u,w) - r^{P',Q'}(x,y) \tag{14}$$

where $r^{P,Q}$ is the measured averaged received signal strength that is being reported from a bin, where a user resides in the presence of positioning uncertainty (both quantization and positioning inaccuracy) with its received signal strength equal to $r^{P',Q'}$, in the absence of positioning uncertainty (no positioning inaccuracy and no quantization). This error is a function of bin width and location parameters defined in Table 3.

Fig. 13 illustrates the PDFs of coverage estimation errors for different bin widths and positioning error radius. In Fig. 13 (a), we show the case of large user positioning error ($u = 100m$) and small quantization error ($w = 10m$). It can be seen from the figure that the effect of quantization alone on coverage estimation leads to almost no error in coverage estimation (see gray histogram for $E^{Q,P}$ in Fig. 13). However, a large positioning error causes a large error in coverage estimation (shown by large variance of red histogram of $E^{P,Q}$ in Fig. 13). The combined effect of the two errors is shown by E^C and it is dominated by the error in user positioning since a large user positioning error overshadows the small quantization error. On the contrary, Fig. 13 (b) shows the case of small user positioning error ($u = 10m$) and large quantization error ($w = 50m$). Here, the combined error in coverage estimation follows the distribution of $E^{Q,P}$, since the large error due to quantization is much more significant than the small error due to user positioning. Fig. 13 (c) shows the case for $u = 50m$ and $w = 30m$. Over here, the variance of error in coverage estimation due to quantization alone, user positioning error alone, and due to both quantization and user positioning uncertainty is large. Note that unlike Fig. 5 and Fig. 10, the distributions of E^C in Fig. 13 would converge to a delta distribution only when both $u \to 0$ and $w \to 0$ simultaneously.

The mathematical expression to characterize the distribution of under-estimating coverage due to both quantization and user positioning error in this scenario is found to be as follows:

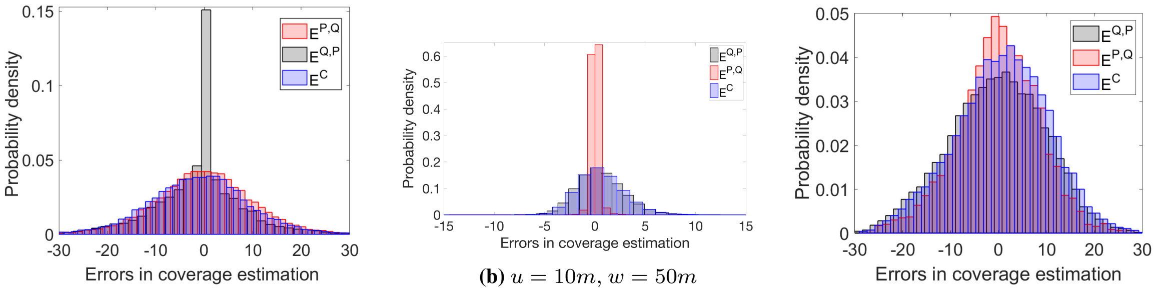
$$f_E^C(e^c,u,w) = \frac{\exp\left(-\frac{e^c-\mu_5(u,w)}{s_5(u,w)}\right)}{s_5(u,w)\left(1+\exp\left(-\frac{e^C-\mu_5(u,w)}{s_5(u,w)}\right)\right)^2}, \text{for } e^C < 0 \;\forall\; \mu_5 \tag{15}$$

where

$$\mu_5 = a_5 + b_5u + c_5w + d_5u^2 + e_5uw + f_5w^2 + g_5u^2w + h_5uw^2 + i_5w^3 \tag{16}$$

$$s_5 = j_5wu + k_5u + l_5w^{m_5} \tag{17}$$

and $a_5 = 0.372, b_5 = -0.008895, c_5 = -0.03936, d_5 = 4.532 \times 10^{-5}, e_5 = 0.0004475, f_5 = 0.0013, g_5 = 2.191 \times 10^{-6}, h_5 = -6.576 \times 10^{-6}, i_5 = -9.658 \times 10^{-6}, j_5 = -0.0005075, k_5 = 0.03271, l_5 = 1.936, m_5 = 0.3147$. e^c represents a realization of the error in coverage estimation



(a) $u = 100m, w = 10m$

(b) $u = 10m, w = 50m$

(c) $u = 50m, w = 30m$

FIGURE 13: PDF of coverage estimation error due to both positioning uncertainty and quantization

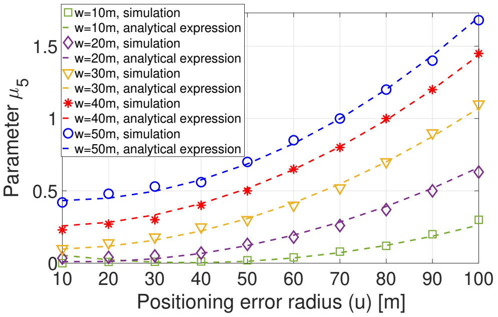


FIGURE 14: Parameter μ_5

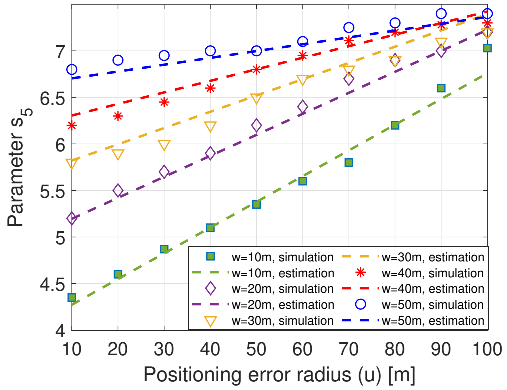


FIGURE 15: Parameter s_5

due to both quantization and user positioning errors, u is the positioning error radius and w is the bin width. Fig. 14 and Fig. 15 depict parameters μ_5 and s_5 respectively.μ_5 is the mean of E^C represented by (16) and s_5, given by (17) captures the variability in E^C.

D. ERROR DUE TO SCARCITY OF DATA

An additional challenge, that is jointly related to quantization and positioning uncertainty is the sparsity of user reports. The problem of sparse MDT reports is illustrated in Fig. 16.

From Fig. 16, it is observed that the mean percentage of area containing no reported user measurements increases exponentially as the number of users (user density) decreases. Moreover, it also increases as the bin width decreases, since for the same number of users in a given area, decreasing bin width leads to a greater number of empty bins, that translates to a higher percentage of the total area with no MDT reports.

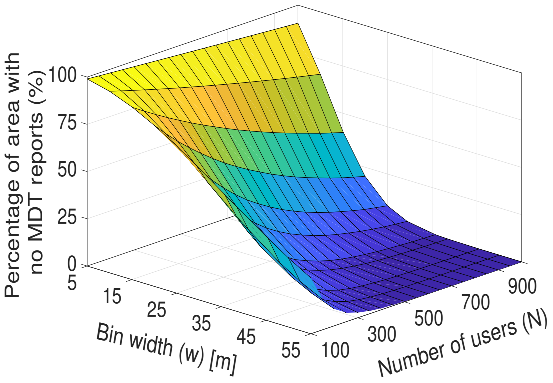


FIGURE 16: Percentage of area with no MDT reports with varying bin width and number of users

Therefore, it is important to find a robust method to predict the coverage status of empty bins.

Consider the scenario in which the predicted coverage area is divided into $n \times n$ bins. Gathered coverage data from different bins can be represented in a matrix $\boldsymbol{C}$ of dimensions $n \times n$. Thus, the coverage area forms a square matrix $\in \mathbb{R}^{(n \times n)}$, where each entry is located at the i-th row and j-th column. Following the time window for gathering measurements and updating the coverage map, it is possible that values are available in only m random bins where $m < n \times n$ such that $\{C_{ij} : (i,j) \in \Psi\}$ and Ψ is a set of cardinality m sampled at random.

In order to recover these missing coverage values, we apply and compare selected popular techniques from literature to our data and also propose some new approaches to address this issue in this section.

1) Rank minimization based matrix completion

We propose a scheme that jointly exploits matrix factorization theory and convex optimization. We note that matrix $\boldsymbol{C}$ will naturally be low ranked. This observation stems from the fact that propagation conditions are mostly dominated by line of sight in small cells and the standard deviation of shadowing is generally small. Also, it is shown in the study in [58] that the shadowing phenomenon that heavily determines coverage values, particularly in a small cell environment, remains correlated over small distances that separate users in the same small cell. This leads to the following optimization problem in order to find the missing values in matrix $\boldsymbol{C}$:

$$\begin{aligned} &\text{minimize} \quad \text{rank}\{\boldsymbol{P}\} \\ &\text{subject to} \quad P_{ij} = C_{ij} \quad (i,j) \in \Psi \end{aligned} \tag{18}$$

The problem in (18) is known to be not only NP-hard, but also all known algorithms that provide exact solutions require time doubly exponential in the dimension n in both theory and practice [59]. However, the analysis presented in [59] proves that (18) can be approximately solved and thus coverage values in vacant bins can be obtained by solving the following alternate convex optimization problem [60]:

$$\begin{aligned} &\text{minimize} \quad ||\boldsymbol{P}||_* \\ &\text{subject to} \quad P_{ij} = C_{ij} \quad (i,j) \in \Psi \end{aligned} \tag{19}$$

where $||\boldsymbol{P}||_*$ is the nuclear norm and is given as:

$$||\boldsymbol{P}||_* = \sum_{k=1}^{n} \sigma_k(\boldsymbol{P}) \tag{20}$$

In (20), $\sigma_k(\boldsymbol{P})$ denotes the k^{th} largest singular value of $\boldsymbol{P}$ and n is the number of bins. (19) therefore aims to determine the matrix with minimum nuclear norm that fits the data.

The problem in (19) can be solved with the singular value-based threshold (SVT) algorithm presented in [61]. The SVT algorithm solves the following problem:

$$\begin{aligned} &\text{minimize} \quad \eta||\boldsymbol{P}||_* + \frac{1}{2}||\boldsymbol{P}||_F^2 \\ &\text{subject to} \quad \mathcal{O}_\Psi(\boldsymbol{P}) = \mathcal{O}_\Psi(\boldsymbol{C}) \end{aligned} \tag{21}$$

where $\mathcal{O}_\Psi$ is the orthogonal projector onto the span of matrices vanishing outside of Ψ so that the $(i,j)^{th}$ component of $\mathcal{O}_\Psi(\boldsymbol{P})$ is equal to P_{ij} if $(i,j) \in \Psi$ and zero otherwise. $||.||_F$ denotes the Frobenius norm. η is a regularization parameter in the objective function and is shown in [61] that the solution of the problem in (21) converges to that of (19) as $\eta \to \infty$.

The SVT algorithm is iterative and produces a sequence of matrices $\{\boldsymbol{P}, \boldsymbol{Q}\}$. At each step, a soft-thresholding operation is performed on the singular values of the matrix $\boldsymbol{Q}^t$. By selecting a large value of the parameter, η in (21), the sequence of iterates, $\{\boldsymbol{P}^t\}$ converges to a matrix which nearly minimizes (19). Starting with $\boldsymbol{Q}^0 = \boldsymbol{0} \in \mathbb{R}^{(n \times n)}$, the algorithm inductively defines

$$\boldsymbol{P}^t = \text{shrink}(\boldsymbol{Q}^{t-1}, \eta) \tag{22}$$

$$\boldsymbol{Q}^t = \boldsymbol{Q}^{t-1} + \Delta_i \mathcal{O}_\Psi(\boldsymbol{C} - \boldsymbol{P}^t) \tag{23}$$

where $\{\Delta_i\}, i \geq 1$ is a sequence of scalar step sizes, until a stopping criteria is reached. The shrink function in (22) applies a soft-thresholding rule at level η to the singular values of the input matrix. It is defined as

$$\text{shrink}(\boldsymbol{Q}^{t-1}, \eta) = \mathcal{S}_\eta(\boldsymbol{Q}^{t-1}) := \boldsymbol{U}\mathcal{S}_\eta(\boldsymbol{\Sigma})\boldsymbol{V}^* \tag{24}$$

$$\mathcal{S}_\eta(\boldsymbol{\Sigma}) = \text{diag}(\{(\sigma_k - \eta)_+\}) \tag{25}$$

where $f_+ = \max(0, f)$. Equivalently, this operator is the positive part of f and simply applies a soft-thresholding rule to the singular values of $\boldsymbol{P}$, shrinking them towards zero. $\boldsymbol{U}, \boldsymbol{V}$ are matrices with orthonormal columns and the singular values $\boldsymbol{\Sigma}$ are positive. $\boldsymbol{U}, \boldsymbol{V}$ and $\boldsymbol{\Sigma}$ are obtained from the singular value decomposition of matrix $\boldsymbol{P}$ of rank r:

$$\boldsymbol{P} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*, \quad \boldsymbol{\Sigma} = \text{diag}(\{\sigma_k\}), 1 \leq k \leq r \tag{26}$$

To cope with the presence of random shadowing in our model, we modify the stopping criteria of the algorithm as follows:

$$||\mathcal{O}_\Psi(\boldsymbol{P}^t - \boldsymbol{C})||_F^2 \leq (1+\zeta)m\phi^2 \tag{27}$$

where ζ is a fixed tolerance, m is the total number of entries in the sparse coverage matrix and ϕ is the standard deviation of shadowing that is modelled as a Gaussian random variable. Therefore, our reconstruction matrix, $\hat{\boldsymbol{C}}$ is the first $\boldsymbol{P}^t$ obeying (27).

Another similar rank minimization based algorithm used to recover the matrix $\boldsymbol{C}$ is the fixed point continuation (FPC) algorithm in [62]. While SVT is efficient for large matrix completion problems, it only works well for very low rank matrix completion problems. It is shown in [62] that for problems where the matrices are not of very low rank, SVT is slow and not robust and therefore, often fails. To solve this problem, FPC-based algorithm is proposed in [62]. FPC-based algorithm has some similarity with the SVT algorithm in that it makes use of matrix shrinkage as in (22)-(25). However, it solves (21) by leveraging operator splitting technique [63]. This technique computes the solution numerically by first separating the original equation into parts over a time step, calculating the solution to each part separately and then combining the solutions to the form the final solution.

2) Inverse distance weighted interpolation method

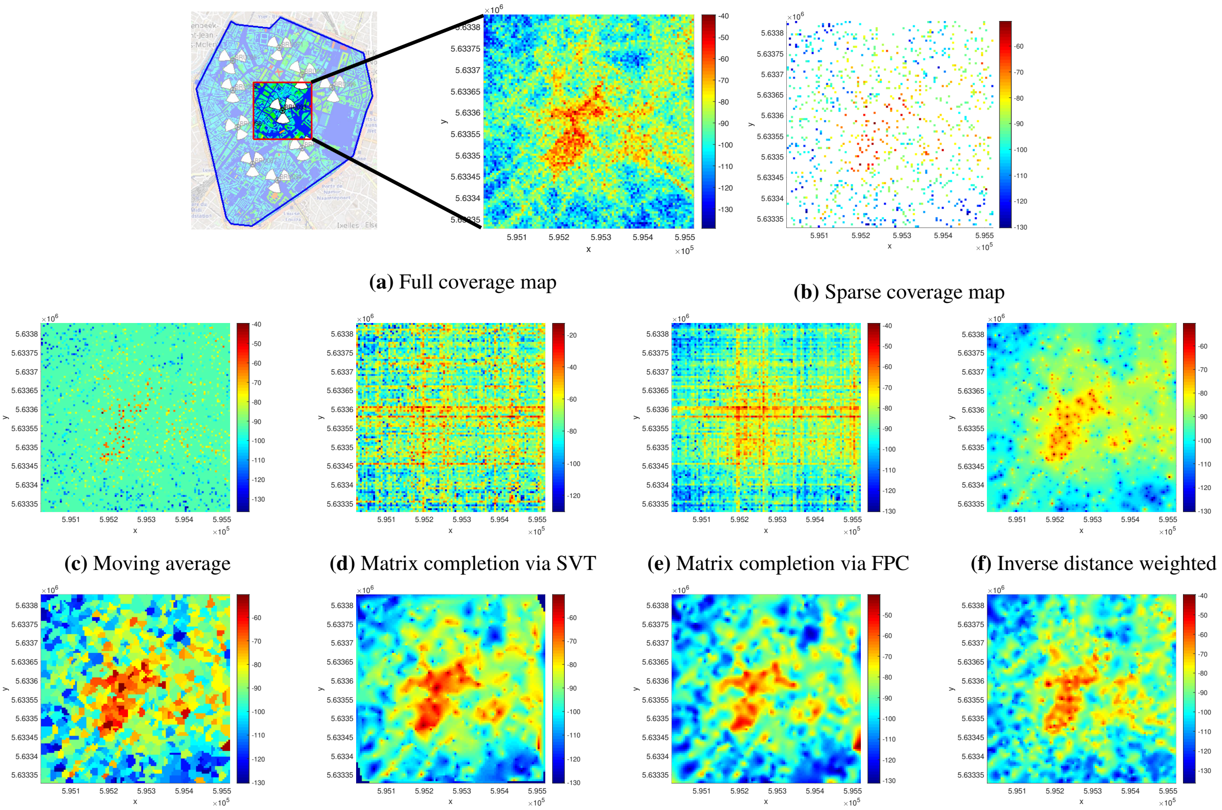
To calculate the missing received signal strength value, $\hat{C_m}$ (at a particular bin location) in matrix $\boldsymbol{C}$, weighted average of N known signal strength values C_k from N adjacent bins are used, where $k = 1 \ldots N$ [64]. Each known received signal strength value is weighted with a weight that is equal to the inverse of distance, d_k between the location of the bin with missing RSRP value and location of the k-th bin and raised to the power p. We take $p = 2$.

3) Moving average method

The missing coverage value by the moving average method is set equal to the weighted arithmetic average of the neighboring coverage values. Mathematically, $p = 0$ in inverse distance weighted method.

4) Nearest neighbor method

The measure of Euclidean distance is used to calculate the distances between the interpolating location and locations of the known measurements and the measurement with the minimum Euclidean distance is selected as the missing signal strength value.



(a) Full coverage map

(b) Sparse coverage map

(c) Moving average

(d) Matrix completion via SVT

(e) Matrix completion via FPC

(f) Inverse distance weighted

(g) Nearest neighbor

(h) Natural neighbor

(i) Spline

(j) Kriging

FIGURE 17: Comparison of coverage map reconstruction techniques for $u = 0$ and $w = 5$m

5) Natural neighbor method

The natural neighbor method finds the received signal strength value at a particular location as a weighted average of those available available measurements which fall within its 'natural neighborhood'. The natural neighbors of any point are those associated with neighboring Voronoi polygons. If the 2-D point n_k is a natural neighbor of the 2-D point p, the portion of Voronoi region, V_{n_k} stolen away by p is called the natural region of p with respect to n_k. Initially, a Voronoi diagram is constructed of all the available coverage values. Then, a new Voronoi polygon is created around the interpolation point (missing coverage value). The proportion of overlap between this new polygon and the initial polygons is then used as weights.

6) Spline method

This method estimates the signal strength value at a particular location by using piecewise defined polynomials called splines. We use a biharmonic spline interpolation, where the interpolating surface is a linear combination of Green functions centered at each data point. The amplitudes of the Green functions are found by solving a linear system of equations [65]- [66].

7) Kriging method

Kriging is based on statistical models that include autocorrelation among the measured points. The weights in kriging are based on the overall spatial arrangement of the measured points, in addition to the distance between measured points and the prediction location [67].

The first step in kriging is creating a prediction surface map in order to uncover the dependency rules to make predictions. This is done by creating semivariogram and developing covariance functions. The next step is to fit a model to the points forming the empirical semivariogram. For our data, the stable model semivariogram in [68] yielded best results. Kriging weights then come from the semivariogram that was developed by analyzing the spatial nature of the data. These weights are a result of minimizing the variance of the estimation error, $\mathbb{V}\left[\hat{C_m} - C_m\right]$, where $\mathbb{V}$ is the variance operator and C_m is the missing coverage value.

E. COMPARISON OF SELECTED TECHNIQUES TO ADDRESS THE DATA SPARSITY CHALLENGE

For the comparison of selected proposed and existing techniques to address data sparsity in MDT, we select a square area of 500m × 500m shown in Fig. 17a as a case study example. We apply the techniques on sparse coverage map shown in Fig. 17b. The resulting visual outputs for a bin width of 5m are shown in Fig. 17 (c)-(j). It can be seen from these figures that kriging interpolation method performs the best. This is because in contrast to other interpolation methods where the weights are dependent solely on the distance to the prediction location, the weights in kriging are based on the overall spatial arrangement of the measured points as well. Note that although Fig. 17 shows a part of the simulated area, the conclusions remain same for other geographical parts from the simulated area.

In order to quantify the accuracy of possible solutions for MDT data sparsity, we use the measure of relative error:

$$E^M = ||\hat{\boldsymbol{C}} - \boldsymbol{C}^{full}||_F / ||\boldsymbol{C}^{full}||_F \tag{28}$$

where $\boldsymbol{C}^{full}$ is the matrix with full entries, considering that RSRP measurements are available from all bins and $\hat{\boldsymbol{C}}$ is the recovered coverage matrix. $||.||_F$ represents the Frobenius norm operator.

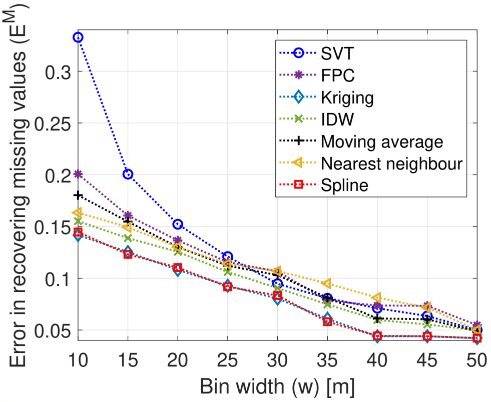


FIGURE 18: Recovery error with varying bin widths using different reconstruction techniques for $u = 100m$.

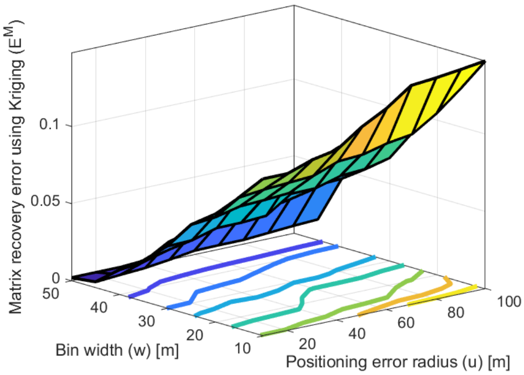


FIGURE 19: Matrix recovery error with varying bin widths and positioning error radius using Kriging.

Positioning uncertainty is then added to the analysis and the recovery errors for $u = 100$m are shown in Fig. 18. From this figure, we can observe that the trends in coverage estimation error with jointly varying bin width and positioning uncertainty remain consistent for other interpolation methods too. In this work we focus only on the accuracy of recovery methods. Other aspects, such as computational complexity are out of scope of this work and can be considered in a future study. From these results, we conclude that kriging works best in extreme scenarios of high positioning uncertainty and low bin widths. This is because other methods are directly based on the surrounding measured values or on specified mathematical formulas that determine the smoothness of the resulting surface, whereas kriging is based on geostatistical methods. Therefore, it performs better even in conditions such as large user positioning uncertainty. Hence, we select this technique for the case studies presented in Section IV.

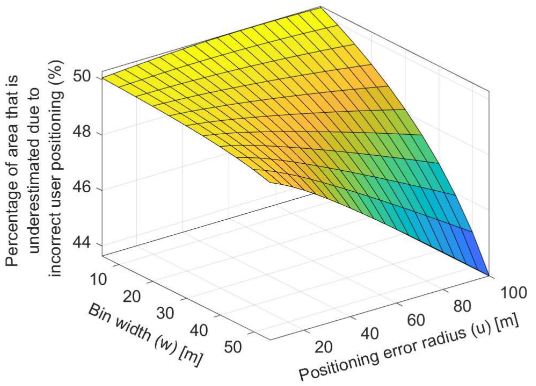


FIGURE 20: Percentage of area that is underestimated due to incorrect user positioning as a function of w and u.

The joint characterization of the matrix recovery error, E^M using Kriging, with the bin width and positioning error radius is shown in Fig. 19. It can be seen from Fig. 19 that this error increases with increase in positioning error radius for different fixed bin widths, since the disparity in actual and reported locations increases with increase in positioning error, making it difficult to recover the actual coverage values. This error also increases with decrease in bin width for different positioning errors, owing to the smaller percentage of available MDT reports as the bin width decreases. Fig. 19 therefore, quantifies the inter-dependencies between these factors.

IV. POTENTIAL APPLICATIONS

From a cellular network design perspective, the analysis and insights obtained from results of this study can be used for many potential practical applications. In this section, we present two fundamental applications of our work related to network planning and optimization, i.e., coverage calibration and determining optimal bin width.

A. COVERAGE CALIBRATION

Having investigated and characterized the various types of errors in MDT-based autonomous coverage estimation, we can now 1) quantify coverage estimation error and 2) determine the direction of coverage estimation error, i.e., is the coverage over-estimated or under-estimated and by what amount? This information can be used by network operators to correctly calibrate the coverage for different geographical areas.

The probability of area whose coverage is under-estimated due to given positioning uncertainty, $A_u^P(u, w)$ can be calculated by integrating (5) from $-\infty$ to 0 while the probability of

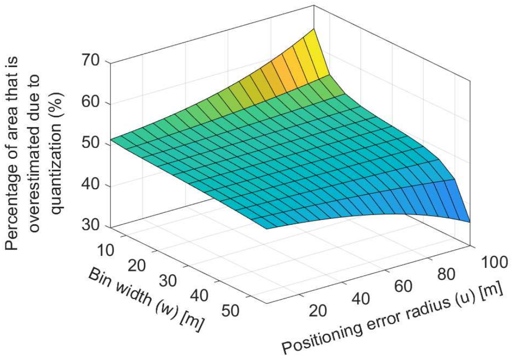


FIGURE 21: Percentage of area that is overestimated due to quantization as a function of w and u.

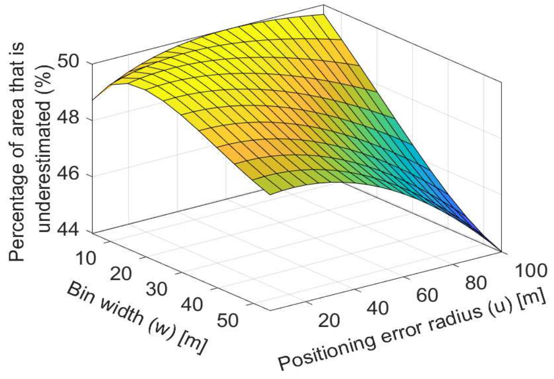


FIGURE 22: Percentage of area that is underestimated due to both quantization and incorrect user positioning as function of w and u.

area that is over-estimated due to quantization, $A_o^Q(u,w)$ can be determined by integrating (11) from 0 to ∞ as follows:

$$A_u^P(u,w) = \int_{-\infty}^{0} \frac{\exp\left(-\frac{e^{P,Q}-\mu_2(u,w)}{s_2(u,w)}\right)}{s_2(u,w)\left(1+\exp\left(-\frac{e^{P,Q}-\mu_2(u,w)}{s_2(u,w)}\right)\right)^2}\,\mathrm{d}e^{P,Q}$$

$$A_u^P(u,w) = \frac{1}{\mathrm{e}^{\frac{\mu_2(u,w)}{s_2(u,w)}}+1} \tag{29}$$

$$A_o^Q(u,w) = \int_{0}^{\infty} \frac{\exp\left(-\frac{e^{Q,P}-\mu_4(u,w)}{s_4(u,w)}\right)}{s_4(u,w)\left(1+\exp\left(-\frac{e^{Q,P}-\mu_4(u,w)}{s_4(u,w)}\right)\right)^2}\,\mathrm{d}e^{Q,P}$$

$$A_o^Q(u,w) = 1 - \frac{1}{\mathrm{e}^{\frac{\mu_4(u,w)}{s_4(u,w)}}+1} \tag{30}$$

Fig. 20 shows the probability of area that is under-estimated due to positioning uncertainty for given bin widths while Fig. 21 shows the probability of area that is over-estimated due to quantization for given positioning uncertainties. Using such figures, given a specific bin with and positioning error radius, network operators can estimate what percentage of the total area under consideration is being falsely estimated due to which error source. The probability of area that is under-estimated due to both quantization and user positioning error can be found by integrating (15) from $-\infty$ to 0, yielding the expression in (31) and illustrated by Fig. 22.

$$A_u^c(u,w) = \frac{1}{\mathrm{e}^{\frac{\mu_5(u,w)}{s_5(u,w)}}+1} \tag{31}$$

The probability of area that is over-estimated is then $1 - A_u^c(u,w)$ Note that the integral limits of (29)-(31) can also be modified based on minimum coverage thresholds determined by the network operator. Given the bin width and positioning error radius, Fig. 19-21 can be used to calibrate observed coverage in order to estimate true coverage in a specified area.

B. DETERMINING OPTIMAL BIN WIDTH

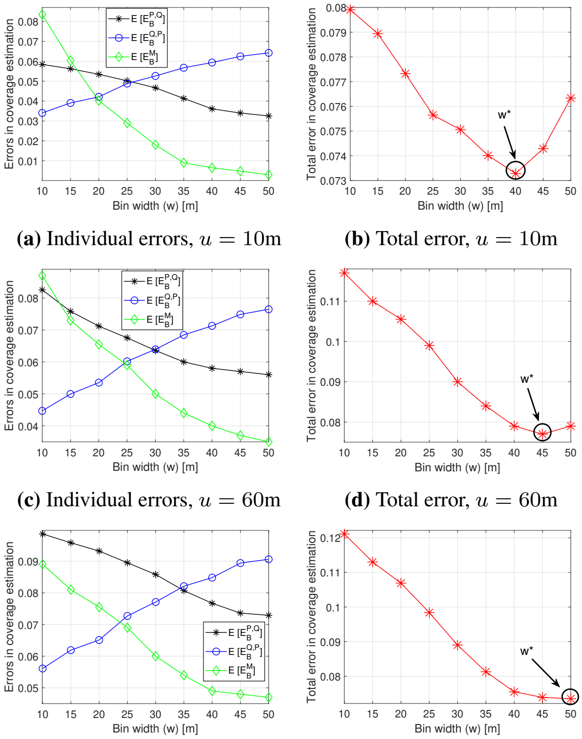
While on one hand, decreasing bin size reduces the quantization error, on the other hand, it increases the error in coverage estimation due to incorrect user positioning and sparsity of user reports. This study is the first to show that there exists an optimal bin width for given user positioning error that can minimize the overall error in the MDT based coverage error, i.e., the combined error caused by quantization (dictated by bin size), user positioning inaccuracy and error due to sparse MDT reports. This calls for an optimization of bin width that would minimize the overall error under positioning error constraints. The errors in (4), (10) and (28) can have an upper bound of greater than 1. In order to get a bounded measure between 0 and 1 of these errors and to enable comparison of combined quantization and user positioning error with matrix recovery error, we define new bounded error measures based on the relative error measures as follows:

$$E_B^{P,Q} = \frac{1}{n^2}\sum_{i=1}^{n^2} \frac{|\boldsymbol{r}_i^{P,Q} - \boldsymbol{r}_i^{P',Q}|}{|\boldsymbol{r}_i^{P,Q} - \boldsymbol{r}_i^{P',Q}| + |\boldsymbol{r}_i^{P',Q}|} \tag{32}$$

$$E_B^{Q,P} = \frac{1}{U}\sum_{i=1}^{U} \frac{|\boldsymbol{r}_i^{P,Q} - \boldsymbol{r}_i^{P,Q'}|}{|\boldsymbol{r}_i^{P,Q} - \boldsymbol{r}_i^{P,Q'}| + |\boldsymbol{r}_i^{P,Q'}|} \tag{33}$$

$$E_B^{C} = \frac{1}{U}\sum_{i=1}^{U} \frac{|\boldsymbol{r}_i^{P,Q} - \boldsymbol{r}_i^{P',Q'}|}{|\boldsymbol{r}_i^{P,Q} - \boldsymbol{r}_i^{P',Q'}| + |\boldsymbol{r}_i^{P',Q'}|} \tag{34}$$

where $\boldsymbol{r}^{P,Q}$ is the measured averaged received power vector of users in bins in the presence of positioning uncertainty and $\boldsymbol{r}^{P',Q}$ is the averaged received power vector of users in bins without any positioning uncertainty. $\boldsymbol{r}^{P,Q'}$ is the received power vector at the user level with positioning uncertainty. $\boldsymbol{r}^{P',Q'}$ is the received power vector at the user level without positioning uncertainty (i.e., the user reporting RSRP value from a particular location is actually present at that exact location).



(a) Individual errors, $u = 10$m **(b)** Total error, $u = 10$m

(c) Individual errors, $u = 60$m **(d)** Total error, $u = 60$m

(e) Individual errors, $u = 100$m **(f)** Total error, $u = 100$m

FIGURE 23: Different errors in coverage estimation, leading to optimal bin widths.

Similarly, a bounded measure for matrix recovery error (this can be considered analogous to error caused by sparsity of MDT reports) can be expressed as:

$$E_B^M = \frac{1}{n^2} \sum_{i=1}^{n^2} \left(\frac{|\hat{\boldsymbol{c}}_i - \boldsymbol{c}_i^{full}|}{|\hat{\boldsymbol{c}}_i - \boldsymbol{c}_i^{full}| + |\boldsymbol{c}_i^{full}|} \right) \tag{35}$$

where $\hat{\boldsymbol{c}} = \text{vect}(\hat{\boldsymbol{C}})$ and $\boldsymbol{c}^{full} = \text{vect}(\boldsymbol{C}^{full})$ are vectorized forms of matrices $\hat{\boldsymbol{C}}$ and $\boldsymbol{C}^{full}$.

For the percentage of area from where MDT reports are unavailable, we want to minimize the matrix recovery error and for the remaining fraction of the total geographical area, we want to minimize total quantization and averaging error. The optimization problem can then be formulated as:

$$w^* = \underset{w}{\text{arg min}} \;\; \mathbb{E}\left(E_B^M + E_B^C\right) \tag{36}$$

$$\text{subject to} \quad w_{min} \le w \le w_{max} \tag{37}$$

$$\text{positioning error radius} = u \tag{38}$$

Owing to the small search space, we can solve (36)-(38) via brute force.

The quantization error, error due to incorrect user positioning and error due to sparse user reports is shown in Fig. 23 (a), (c) and (e) for $u = 10, 60$ and 100 m respectively. Quantization error increases with increase in bin width owing to greater spatial gap among users in a given bin as bin width increases. On the contrary, error due to incorrect user positioning decreases with increase in bin attributing to the fact that for a given positioning error radius, a larger bin width would mean a lesser probability that a particular user reporting MDT data from a given bin is in fact present in any of the adjacent bins. This error is then combined with the matrix recovery error. Since the number of vacant entries in the coverage matrix increases as the bin width decreases as previously illustrated by Fig. 16, it becomes difficult to recover the missing coverage values as bin width decreases. Finally, Fig. 23 (b), (d) and (f) show the effect of all errors simultaneously. We note that the optimal bin width increases as positioning error radius increases. This work therefore presents a framework to determine the optimal bin width that minimizes overall error in MDT-based coverage estimation and can be extended for different UE densities and environmental conditions, that can be focus of a future work.

V. CONCLUSION

In this paper, we investigate the joint effects of the errors due to sparse user measurements, quantization/binning and inaccurate user positioning on MDT-based coverage estimation. We show that there is a need to jointly characterize these errors as they are interdependent and present a framework to quantify these errors and characterize their interplay by determining the error distributions in coverage estimation as a function of user density, bin width, and positioning error radius. In our analysis, we quantify both the error in estimated coverage as well as its direction, i.e., whether the coverage is over estimated or under estimated for any given scenario. Our results reveal that there exists an optimal bin width for a given user positioning inaccuracy and user density that minimizes the overall coverage estimation error. This insight is fundamental to optimal design of MDT based coverage estimation algorithms. Finally, we present two fundamental applications of our work related to network planning and optimization, i.e., coverage calibration and determining optimal bin width. Our findings can not only help substantially improve the self-organizing networks based optimization in legacy networks, but can also act as key enabler for most of the AI based automation use cases envisioned for the operation and optimization of future cellular networks such as 5G and beyond. These use cases include the automatic detection of coverage holes, detection of weak coverage spots or identification of sleeping cells. Another important direction on this issue is the consideration of mobility dynamics in the network (e.g., high-speed train scenario). Speed of the mobile users in the network could affect the user density both spatially and temporally. Hence, the frequency of MDT reports available in a certain area would be dependent on the mobility dynamics of the users in that area. This in turn would impact the sparsity issue and the interplay of positioning error and quantization with user density and is an important direction that is worthy of investigation in the future. Another interesting direction of this work can be its extension to 5G and beyond deployment, for example, in millimeter wave transmission. Accurate coverage estimation problem is crucial in next generation millimeter wave (mmWave) networks, since propagation conditions at millimeter wave bands differ significantly as compared to sub-6GHz bands. Specifically, mmWave spectrum does not offer the broad coverage that sub-6 GHz spectrum supports and is sensitive to external factors. In contrast to the rich multi-path propagation considered in this work, mmWave networks are likely to have only a few propagation paths, that would impact the coverage. Moreover, in a mmWave network scenario, there would be small cells with relatively low user mobility and few simultaneous users due to small coverage area. Therefore, MDT-based accurate estimation of the limited coverage in millimeter wave networks is crucial and this work can be extended to such scenarios.

REFERENCES

[1] O. G. Aliu, A. Imran, M. A. Imran, and B. Evans, "A survey of self organisation in future cellular networks," IEEE Communications Surveys Tutorials, vol. 15, no. 1, pp. 336–361, 2013.

[2] M. G. Kibria, K. Nguyen, G. P. Villardi, O. Zhao, K. Ishizu, and F. Kojima, "Big data analytics, machine learning, and artificial intelligence in next-generation wireless networks," IEEE access, vol. 6, pp. 32 328–32 338, 2018.

[3] A. Asghar, H. Farooq, and A. Imran, "Self-healing in emerging cellular networks: Review, challenges, and research directions," IEEE Communications Surveys & Tutorials, vol. 20, no. 3, pp. 1682–1709, 2018.

[4] I. Akbari, O. Onireti, A. Imran, M. A. Imran, and R. Tafazolli, "Impact of inaccurate user and base station positioning on autonomous coverage estimation," in IEEE 20th International Workshop on Computer Aided Modelling and Design of Communication Links and Networks (CAMAD), 2015, pp. 114–118.

[5] A. Taufique, M. Jaber, A. Imran, Z. Dawy, and E. Yaacoub, "Planning wireless cellular networks of future: Outlook, challenges and opportunities." IEEE Access, vol. 5, pp. 4821–4845, 2017.

[6] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," IEEE Communications Surveys & Tutorials, vol. 18, no. 3, pp. 1617–1655, 2016.

[7] 3rd Generation Partnership Project, "Universal Terrestrial Radio Access (UTRA) and Evolved Universal Terrestrial Radio Access (E-UTRA); Radio measurement collection for Minimization of Drive Tests (MDT); Overall description; Stage 2 (Release 10), 3GPP Standard TS 37.320, Version 10.2.0," Tech. Rep., June 2011.

[8] A. Galindo-Serrano, B. Sayrac, S. B. Jemaa, J. Riihijärvi, and P. Mähönen, "Harvesting MDT data: Radio environment maps for coverage analysis in cellular networks," in 8th International Conference on Cognitive Radio Oriented Wireless Networks. IEEE, 2013, pp. 37–42.

[9] P.-C. Lin, "Minimization of drive tests using measurement reports from user equipment," in 2014 IEEE Global Conference on Consumer Electronics (GCCE), Oct 2014, pp. 84–85.

[10] A. Gómez-Andrades, R. Barco, P. Muñoz, and I. Serrano, "Data analytics for diagnosing the RF condition in self-organizing networks," IEEE Transactions on Mobile Computing, vol. 16, no. 6, pp. 1587–1600, 2017.

[11] N. Samaan and A. Karmouch, "Network anomaly diagnosis via statistical analysis and evidential reasoning," IEEE transactions on network and service management, vol. 5, no. 2, pp. 65–77, 2008.

[12] A. Zoha, A. Saeed, A. Imran, M. A. Imran, and A. Abu-Dayya, "A SON solution for sleeping cell detection using low-dimensional embedding of MDT measurements," in IEEE International Symposium on Personal, Indoor, and Mobile Radio Communication, 2014, pp. 1626–1630.

[13] D. Micheli and R. Diamanti, "Statistical Analysis of Interference in a Real LTE Access Network by Massive Collection of MDT Radio Measurement Data from Smartphones," in 2019 Photonics & Electromagnetics Research Symposium-Spring (PIERS-Spring), 2019, pp. 1906–1916.

[14] M. Mdini, G. Simon, A. Blanc, and J. Lecoeuvre, "Introducing an unsupervised automated solution for root cause diagnosis in mobile networks," IEEE Transactions on Network and Service Management, 2019.

[15] P. Szilágyi and S. Nováczki, "An automatic detection and diagnosis framework for mobile communication systems," IEEE transactions on Network and Service Management, vol. 9, no. 2, pp. 184–197, 2012.

[16] A. Imran, A. Zoha, and A. Abu-Dayya, "Challenges in 5G: how to empower SON with big data for enabling 5G," IEEE network, vol. 28, no. 6, pp. 27–33, 2014.

[17] I. Akbari, O. Onireti, A. Imran, M. A. Imran, and R. Tafazolli, "How reliable is MDT-based autonomous coverage estimation in the presence of user and BS positioning error?" IEEE Wireless Communications Letters, vol. 5, no. 2, pp. 196–199, 2016.

[18] I. Akbari, O. Onireti, M. A. Imran, A. Imran, and R. Tafazolli, "Effect of inaccurate position estimation on self-organising coverage estimation in cellular networks," in Proceedings of European 20th European Wireless Conference, 2014, pp. 1–5.

[19] J. Thrane, M. Artuso, D. Zibar, and H. L. Christiansen, "Drive test minimization using deep learning with bayesian approximation," in 2018 IEEE 88th Vehicular Technology Conference (VTC-Fall), 2018, pp. 1–5.

[20] N. Kanazawa, A. Nagate, and A. Yamamoto, "Field experiment of localization based on machine learning in LTE network," in 2018 IEEE 88th Vehicular Technology Conference (VTC-Fall), 2018, pp. 1–6.

[21] W. Fang and B. Ran, "An Accuracy and Real-Time Commercial Localization System in LTE Networks," IEEE Access, 2020.

[22] M. Lin, X. Song, Q. Qian, H. Li, L. Sun, S. Zhu, and R. Jin, "Robust gaussian process regression for real-time high precision GPS signal enhancement," in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 2838–2847.

[23] A. Scaloni, P. Cirella, M. Sgheiz, R. Diamanti, and D. Micheli, "Multipath and Doppler characterization of an electromagnetic environment by massive MDT measurements from 3G and 4G mobile terminals," IEEE Access, vol. 7, pp. 13 024–13 034, 2019.

[24] P. Bernardin and K. Manoj, "The postprocessing resolution required for accurate RF coverage validation and prediction," IEEE transactions on vehicular technology, vol. 49, no. 5, pp. 1516–1521, 2000.

[25] F. Sohrabi and E. Kuehn, "Construction of the RSRP map using sparse MDT measurements by regression clustering," in IEEE International Conference on Communications (ICC), 2017, pp. 1–6.

[26] J. D. Naranjo, A. Ravanshid, I. Viering, R. Halfmann, and G. Bauch, "Interference map estimation using spatial interpolation of MDT reports in cognitive radio networks," in Wireless Communications and Networking Conference (WCNC), 2014 IEEE. IEEE, 2014, pp. 1496–1501.

[27] R. V. Akhpashev and V. G. Drozdova, "Spatial interpolation of LTE measurements for minimization of drive tests," in 2018 19th International Conference of Young Specialists on Micro/Nanotechnologies and Electron Devices (EDM). IEEE, 2018, pp. 6403–6405.

[28] J. D. Naranjo, A. Ravanshid, I. Viering, R. Halfmann, and G. Bauch, "Interference map estimation using spatial interpolation of MDT reports in cognitive radio networks," in 2014 IEEE Wireless Communications and Networking Conference (WCNC). IEEE, 2014, pp. 1496–1501.

[29] H. Braham, S. B. Jemaa, B. Sayrac, G. Fort, and E. Moulines, "Low complexity spatial interpolation for cellular coverage analysis," in 2014 12th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), 2014, pp. 188–195.

[30] ——, "Coverage mapping using spatial interpolation with field measurements," in 2014 IEEE 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC), 2014, pp. 1743–1747.

[31] H. Braham, S. B. Jemaa, G. Fort, E. Moulines, and B. Sayrac, "Fixed rank kriging for cellular coverage analysis," IEEE Transactions on Vehicular Technology, vol. 66, no. 5, pp. 4212–4222, 2016.

[32] N. Perpinias, A. Palaios, J. Riihijärvi, and P. Mähönen, "A measurement-based study on the use of spatial interpolation for propagation estimation," in 2015 IEEE International Conference on Communications (ICC), 2015, pp. 2715–2720.

[33] N. Perpinias, J. Riihijarvi, and P. Mahonen, "Impact of model uncertainties on the accuracy of spatial interpolation based coverage estimation," in 2017 IEEE Wireless Communications and Networking Conference (WCNC), 2017, pp. 1–6.

[34] Z. El-friakh, A. M. Voicu, S. Shabani, L. Simić, and P. Mähönen, "Crowdsourced indoor Wi-Fi REMs: Does the spatial interpolation method matter?" in 2018 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN), 2018, pp. 1–10.

[35] D. Denkovski, V. Atanasovski, L. Gavrilovska, J. Riihijärvi, and P. Mähönen, "Reliability of a radio environment map: Case of spatial interpolation techniques," in 2012 7th international ICST conference on cognitive radio oriented wireless networks and communications (CROWNCOM), 2012, pp. 248–253.

[36] M. Molinari, M.-R. Fida, M. K. Marina, and A. Pescape, “Spatial interpolation based cellular coverage prediction with crowdsourced measurements,” in Proceedings of the 2015 ACM SIGCOMM Workshop on Crowdsourcing and Crowdsharing of Big (Internet) Data, 2015, pp. 33–38.

[37] D. Fernandes, L. S. Ferreira, M. Nozari, P. Sebastião, F. Cercas, and R. Dinis, “Combining drive tests and automatically tuned propagation models in the construction of path loss grids,” in 2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), 2018, pp. 1–2.

[38] B. Sayrac, J. Riihijärvi, P. Mähönen, S. Ben Jemaa, E. Moulines, and S. Grimoud, “Improving coverage estimation for cellular networks with spatial bayesian prediction based on measurements,” in Proceedings of the 2012 ACM SIGCOMM workshop on Cellular networks: operations, challenges, and future design, 2012, pp. 43–48.

[39] B. Sayrac, A. Galindo-Serrano, S. B. Jemaa, J. Riihijärvi, and P. Mähönen, “Bayesian spatial interpolation as an emerging cognitive radio application for coverage analysis in cellular networks,” Transactions on Emerging Telecommunications Technologies, vol. 24, no. 7-8, pp. 636–648, 2013.

[40] A. Galindo-Serrano, B. Sayrac, S. B. Jemaa, J. Riihijärvi, and P. Mähönen, “Automated coverage hole detection for cellular networks using radio environment maps,” in 2013 11th International Symposium and Workshops on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), 2013, pp. 35–40.

[41] R. Enami, D. Rajan, and J. Camp, “RAIK: Regional analysis with geodata and crowdsourcing to infer key performance indicators,” in 2018 IEEE Wireless Communications and Networking Conference (WCNC), 2018, pp. 1–6.

[42] H. Braham, S. B. Jemaa, G. Fort, E. Moulines, and B. Sayrac, “Spatial prediction under location uncertainty in cellular networks,” IEEE Transactions on Wireless Communications, vol. 15, no. 11, pp. 7633–7643, 2016.

[43] L. Ma, N. Jin, Y. Zhang, and Y. Xu, “RSRP difference elimination and motion state classification for fingerprint-based cellular network positioning system,” Telecommunication Systems, vol. 71, no. 2, pp. 191–203, 2019.

[44] H. N. Qureshi and A. Imran, “Optimal bin width for autonomous coverage estimation using MDT reports in the presence of user positioning error,” IEEE Communications Letters, 2019.

[45] “Atoll, [online] available:https://www.forsk.com/.”

[46] M. Hata, “Empirical formula for propagation loss in land mobile radio services,” IEEE transactions on Vehicular Technology, vol. 29, no. 3, pp. 317–325, 1980.

[47] M. S. Rani, S. Behara, and K. Suresh, “Comparison of standard propagation model (SPM) and Stanford university interim (SUI) radio propagation models for long term evolution (LTE),” IJAIR, vol. 3, pp. 221–228, 2012.

[48] I. Mohamed, “Path-loss estimation for wireless cellular networks using okumura/hata model,” Science Journal of Circuits, Systems and Signal Processing, vol. 7, no. 1, pp. 20–27, 2018.

[49] “ITU-R Recommendation P.452-15. Prediction procedure for the evaluation of interference between stations on the surface of the Earth at frequencies above about 0.1 GHz,” 2013.

[50] G. K. Chan, “Effects of sectorization on the spectrum efficiency of cellular radio systems,” IEEE transactions on vehicular technology, vol. 41, no. 3, pp. 217–225, 1992.

[51] 3rd Generation Partnership Project, “GPP TS 36.101 V16.2.0 (2019-06). Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA);User Equipment (UE) radio transmission and reception (Release 16),” 2019.

[52] Spectrum Monitoring. Frequencies. Accessed on: 19 July 20, 2020. [Online]. Available: https://www.spectrummonitoring.com/frequencies/

[53] R. D. Straw and G. Hall, “Antenna height and communications effectiveness,” in Newington CT 06111, 225 Main Street. The American Radio Relay League, 1999.

[54] M. Deruyck, W. Joseph, and L. Martens, “Power consumption model for macrocell and microcell base stations,” Transactions on Emerging Telecommunications Technologies, vol. 25, no. 3, pp. 320–333, 2014.

[55] Federal Communications Commission. Report and Order and Future Notice of Proposed Rulemaking [Online]. Accessed on: 19 July 20, 2020. [Online]. Available: https://docs.fcc.gov/public/attachments/FCC-16-89A1.pdf

[56] “Technical Report. ETSI TR 136 942 V13.0.0 (2016-01). Evolved Universal Terrestrial Radio Access (E-UTRA), Radio Frequency (RF) system scenarios (3GPP TR 36.942 version 13.0.0 Release 13 16-01),” 2016.

[57] M. Haenggi, Stochastic geometry for wireless networks. Cambridge University Press, 2012.

[58] X. Yang and R. Tafazolli, “A method of generating cross-correlated shadowing for dynamic system-level simulators,” in 14th IEEE Proceedings on Personal, Indoor and Mobile Radio Communications, 2003. PIMRC 2003., vol. 1. IEEE, 2003, pp. 638–642.

[59] E. J. Candès and B. Recht, “Exact matrix completion via convex optimization,” Foundations of Computational mathematics, vol. 9, no. 6, p. 717, 2009.

[60] E. J. Candès and T. Tao, “The power of convex relaxation: Near-optimal matrix completion,” IEEE Trans. Inf. Theor., vol. 56, no. 5, pp. 2053–2080, May 2010. [Online]. Available: http://dx.doi.org/10.1109/TIT.2010.2044061

[61] J.-F. Cai, E. J. Candès, and Z. Shen, “A singular value thresholding algorithm for matrix completion,” SIAM Journal on Optimization, vol. 20, no. 4, pp. 1956–1982, 2010.

[62] S. Ma, D. Goldfarb, and L. Chen, “Fixed point and bregman iterative methods for matrix rank minimization,” Mathematical Programming, vol. 128, no. 1-2, pp. 321–353, 2011.

[63] E. T. Hale, W. Yin, and Y. Zhang, “Fixed-point continuation for l1-minimization: Methodology and convergence,” SIAM Journal on Optimization, vol. 19, no. 3, pp. 1107–1130, 2008.

[64] P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, Geographic information systems and science. John Wiley & Sons, 2005.

[65] X. Deng and Z.-a. Tang, “Moving surface spline interpolation based on Green?s function,” Mathematical Geosciences, vol. 43, no. 6, pp. 663–680, 2011.

[66] P. Wessel and D. Bercovici, “Interpolation with splines in tension: a Green’s function approach,” Mathematical Geology, vol. 30, no. 1, pp. 77–93, 1998.

[67] A. Konak, “A kriging approach to predicting coverage in wireless networks,” International Journal of Mobile Network Design and Innovation, vol. 3, no. 2, pp. 65–71, 2009.

[68] H. Wackernagel, V. D. Oliveira, and B. Kedem, “Multivariate geostatistics,” SIAM Review, vol. 39, no. 2, pp. 340–340, 1997.