

UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

ADDRESSING THE DATA SPARSITY CHALLENGE IN AI-BASED CELLULAR
NETWORKS

A GENERAL EXAMINATION REPORT

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

Ph.D IN ELECTRICAL AND COMPUTER ENGINEERING

BY

HANEYA NAEEM QURESHI

Norman, Oklahoma

2020

ADDRESSING THE DATA SPARSITY CHALLENGE IN AI-BASED CELLULAR
NETWORKS

A GENERAL EXAMINATION REPORT APPROVED FOR THE
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

BY

Dr. Ali Imran, Chair

Dr. Thordur Runolfsson

Dr. Samuel Cheng

Dr. Timothy Ford

Table of Contents

1	Introduction	1
1.1	Motivation	1
1.1.1	Data sparsity challenge	3
1.2	Research objectives	4
1.3	Contributions	5
1.4	Current and planned dissemination and publications	6
1.5	Organization	7
2	Enabling AI-driven cellular network optimization by MDT	8
2.1	Introduction	8
2.2	Relevant work	10
2.3	System model	13
2.4	Quantification of errors in autonomous coverage estimation using MDT .	15
2.4.1	Error due to user positioning uncertainty	15
2.4.2	Quantization error	22
2.4.3	Combined effect of positioning and quantization errors on coverage estimation	27
2.4.4	Error due to scarcity of data	30
2.5	Practical applications	34
2.5.1	Coverage calibration	34
2.5.2	Determining optimal bin width	37
2.6	Conclusion	40
3	A framework towards addressing the training data sparsity challenge in cellular networks	41
3.1	Introduction	41
3.1.1	Related Work	43
3.2	Interpolation methods	45
3.2.1	Matrix completion theory based recovery	45
3.2.2	Inverse distance weighted	48

3.2.3	Gradient plus inverse distance squared	53
3.2.4	Modified Shepard's method	54
3.2.5	Nearest neighbor	55
3.2.6	Natural neighbor	56
3.2.7	Splines	57
3.2.8	Kriging	58
3.3	Methods using contextual information	59
3.3.1	Utilizing geometry of network	60
3.3.2	Through propagation modeling and transmitter parameter estimation	61
3.3.3	Transfer learning	68
3.4	Synthetic data generation	71
3.4.1	Through generative adversarial network	71
3.4.2	Through custom-built system level simulators	75
3.5	Real data generation	77
3.5.1	Through smartphone applications	77
3.5.2	Through testbeds	79
3.6	Conclusion	89
4	Conclusions	92
	References	92

List of Figures

2.1	System model configuration and geographical information.	13
2.2	User distribution	13
2.3	PDF of coverage estimation error due to positioning uncertainty in the absence of bins	16
2.4	Parameter s_1	17
2.5	Probability of misclassification of a user with varying bin width and positioning error radius.	19
2.6	PDF of coverage estimation error due to positioning uncertainty in the presence of bins	20
2.7	Parameter μ_2	21
2.8	Parameter s_2	22
2.9	PDF of coverage estimation error due to quantization error without positioning uncertainty	23
2.10	Variance of error, $E^{Q,P'}$	23
2.11	PDF of coverage estimation error due to quantization in the presence of positioning uncertainty	24
2.12	Parameter μ_4	25
2.13	Parameter s_4	26
2.14	PDF of coverage estimation error due to both positioning uncertainty and quantization	27
2.15	Parameter μ_5	29
2.16	Parameter s_5	29
2.17	Percentage of area with no MDT reports with varying bin width and number of users	30
2.18	Kriging for coverage map reconstruction, $u = 0$ and $w = 10\text{m}$	31
2.19	Comparison of coverage map reconstruction techniques for $u = 0$ and $w = 5\text{m}$	33
2.20	Recovery error with varying bin widths using different reconstruction techniques.	34
2.21	Matrix recovery error with varying bin widths and positioning error radius using Kriging.	35
2.22	Coverage area miscalculated due to different causes.	36

2.23	Different errors in coverage estimation, leading to optimal bin widths. . .	38
3.1	Methods to address data sparsity challenge and chapter organization. . .	42
3.2	Triangular membership function for different adaptive distance-decay parameters (modified from [1]).	54
3.3	Leveraging dense base station deployment to enrich sparse data.	61
3.4	Leveraging cluster geometry to enrich sparse data.	62
3.5	Comparison of Prediction Error after using Transfer Learning on the Source Model.	70
3.6	Comparison of Prediction Performance with Ground Truth after using Transfer Learning on the Source Model.	70
3.7	Leveraging GAN for enriching the sparse training data [2]	72
3.8	Some current and emerging 5G testbeds.	78
3.9	Federated Testbeds.	79
3.10	Framework to address data sparsity challenge.	90

List of Tables

2.1	Network Scenario Settings.	14
3.1	Improvements to IDW interpolation	52
3.2	Comparison of different simulators for solving data sparsity problem . .	75
3.3	Worldwide existing and emerging testbeds for solving data sparsity problem	80

Abstract

The future of cellular networks is contingent on big data enabled artificial intelligence based solutions for various automation tasks. One such enabler for automation is the use of Minimization of drive test (MDT). MDT allows coverage to be estimated at the base station by exploiting measurement reports gathered by the user equipment without the need for drive tests. Thus, MDT is a key enabling feature for data and artificial intelligence driven autonomous operation and optimization in current and emerging 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 the error in MDT-based coverage estimation that stem from scarcity of user reports, while also considering inaccurate user positioning 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 the 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. Moreover, traditional machine learning based techniques for network automation that leverage training and tuning of machine learning based models to determine the behavior of different configuration and optimization parameters, also suffer from training data sparsity challenge. These solutions, though very promising, require vast amounts of training data. Consequently, the success of these solutions is limited by a fundamental challenge faced by research community: sparsity of training data. In this report, this problem is addressed by presenting a comprehensive overview of existing and potential new techniques to enrich sparse data in cellular networks. Along with an overview of existing approaches to address the data sparsity problem, new techniques for enriching sparse training data are also proposed and

demonstrated, such as by matrix completion theory, leveraging different types of network geometries, transfer learning and the use of generative adversarial networks. In addition, an overview of state-of-the art simulators and testbeds is also presented to make readers aware of current and emerging platforms to access real data in order to overcome the data sparsity challenge.

CHAPTER 1

Introduction

1.1 Motivation

Emerging characteristics of wireless cellular networks, such as cell densification, simultaneous operation at multiple frequency bands, control-data plane split, network virtualization, amalgam of new technologies and flexible infrastructure and spectrum allocations are leading to growing complexity, resource inefficiency and shrinking profit margins [3]. To deal with these challenges, zero-touch deep automation in cellular networks is envisioned, not only to provide better quality of experience but also for their technical and commercial viability [4, 5, 6]. This includes functionalities of self-optimization, self-healing and self-configuration [4].

In order to enable these automation capabilities in future cellular networks, one key aspect is designing network and user behavior models. This is of fundamental significance as tuning the network parameters to identify the optimal network configuration, that can maximize the vital key performance indicators, like coverage, capacity, reliability or energy efficiency is necessary for network operators to fulfill the promises made by much anticipated next generation networks.

To date, the design, operation and optimization of wireless networks, has relied either on analytical models or on system level simulations (ray-tracing based simulators). The use of analytical models enables insights into the system behavior. Analytical models also allow optimization problem formulation that can often be solved in a computationally efficient manner. Analytical model based optimization solutions also often lead to system designs with known or partially known optimality properties. In general, however, analytical models are rarely sufficient for accurate system design and optimization of real

wireless networks. For example, cellular networks are often modeled by relying on the abstraction models based on point processes such as Poisson point processes or Matern hard core point process [7, 8]. Such approaches allow analytical tractability, but they are often not accurate enough to capture important details in practical cellular network deployments such as spatio-temporally varying user and Base Station (BS) distributions, complex antenna patterns, and propagation environment, etc.

The use of system-level simulators using ray-tracing, on the other hand, leads to functional system designs. However, while far more accurate than assumptions-based analytical models [9], this approach offers limited insight into system behavior particularly in terms of optimality of its performance for given design parameters. In addition to its complexity, it also requires detailed information about the operating environment and topographical data, such as digital elevation maps, digital terrain models, etc. Hence, design, operation and optimization of large-scale complex networks such as 5G and beyond, using a simulator in lieu of network behavior models may, in addition to being sub-optimal, be inefficient or infeasible altogether due to the huge number of optimization variables and inability to incorporate measured data. It is vital to foresee, that while these two approaches have sufficed to yield functional and economically viable legacy networks, both of the above approaches will particularly become impractical in wake of emerging cellular networks because of their inability to fit into the frame of zero-touch automation.

Due to the limitations of these two existing approaches, a third approach to optimizing networks is becoming increasingly popular: the machine learning (ML)-based data-driven approach [10, 11].

However, traditional ML/AI approaches suffer from many caveats, such as requiring vast amounts of training data, large training time, little interpretability [12] or insights [13] for tuning and optimizing different network parameters and complex forms of input and output parameter relationships [14]. The focus of this report is to address the challenge

of *sparsity of training data* to enhance these AI-enabled approaches for the design, operation and optimization of cellular networks. This challenge is further explained below.

1.1.1 Data sparsity challenge

Machine learning based techniques face a common key challenge that undermine their utility: scarcity/sparsity of the training data. This is because of the several following reasons:

- Conducting independent field trials on a large scale is costly and time-consuming.
- Mathematical models can not be used to generate training data since they are based on too many assumptions and simplifications, that fail to depict real world scenarios.
- Obtaining large amount of pertinent data from network operators is not a trivial task.
- Network operators only try a limited range of COPs in live networks due to high probability of significant network performance impairment of live mobile network during the trial phase. Therefore, only a limited range of COP-KPI data can be obtained.
- Despite sourcing from multiple operators, the real data are expected to be sparse or unevenly distributed.
- In dynamic scenarios, where the number and locations of measurements change, it is infeasible to measure the radio frequency field strength values at every point of interest.
- In ultra-dense deployments, small cells contain far fewer users compared to macro cells. This makes user measurements at the base station of small cells sparse, which

particularly poses a problem for automation solutions that leverage minimization of drive test (MDT) [15, 16, 17]. This problem is further aggravated if smaller bin size is used to reduce quantization error, attributing to the fact that many bins might not be visited by even a single user during the reporting period [17].

Deploying the new 5G and beyond network functionalities in a real world cannot be done arbitrarily. If the training data is poorly distributed or sparse, it might not represent the actual network scenario very well, which could lead to over-fitting during the model training stage. In order to develop accurate models, machine learning algorithms require large amounts of true training data since a model based on sparse data would rely on assumptions and weak correlations [2]. In turn, unscrupulous network design and sub-optimal parameter configuration will hamper not only the capability of future networks that will impact the user experience negatively but will also increment the capital and operational expenditure (CAPEX/OPEX) of mobile operators [18].

1.2 Research objectives

In light of the discussion in section 1.1.1 this report explores the following research questions:

- Minimization of drive test (MDT) could be a promising way to automating cellular networks and address certain aspects of data sparsity challenge. However, although standardized since 3GPP Release 10, why isn't it still implemented?
- How can we overcome some of the issues that are thwarting the practical utility of MDT feature?
- Training data remains sparse even in the age of big data. What framework can be developed towards addressing the training data sparsity challenge in cellular networks for different network scenarios?

- What new techniques can be leveraged to address the data sparsity challenge?

1.3 Contributions

The primary contributions of this report are summarized as follows:

- Minimization of drive test (MDT) allows coverage to be estimated at the base station by exploiting measurement reports gathered by the user equipment without the need for drive tests. Thus, MDT is a key enabling feature for data and artificial intelligence driven autonomous operation and optimization in current and emerging cellular networks. However, to date, the utility of MDT feature remains thwarted by issues such as sparsity of user reports and user positioning inaccuracy. In this report, 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. Utility of the proposed framework is presented by addressing practical applications from network optimization perspective.
- By investigating the interplay between quantization and positioning error to estimate coverage, this report also shows that there exists an optimal bin width for coverage estimation and we determine it as a function of positioning error and user density. This can enable network operators to configure the bin size for given positioning accuracy that results in the most accurate MDT based coverage estimation.
- Several machine learning based techniques are proposed in current literature that leverage training and tuning of machine learning based models to determine the behavior of different configuration and optimization parameters. However, these solutions, though very promising, require vast amounts of training data. Consequently, the success of these solutions is limited by a fundamental challenge faced

by research community: sparsity of training data. In this report, we attempt to solve this problem by presenting a comprehensive overview of existing and potential new techniques to enrich sparse data in cellular networks. Along with an overview of existing approaches to address the data sparsity problem, we also propose and demonstrate new techniques for enriching sparse training data, such as by matrix completion theory, leveraging different types of network geometries, transfer learning and the use of generative adversarial networks. In addition, an overview of state-of-the art simulators and testbeds is also presented to make readers aware of current and emerging platforms to access real data in order to overcome the data sparsity challenge.

1.4 Current and planned dissemination and publications

Journal articles

J1. Qureshi, Haneya Naeem, and Ali Imran. “Optimal bin width for autonomous coverage estimation using MDT reports in the presence of user positioning error.” *IEEE Communications Letters* 23.4 (2019): 716-719.

J2. Qureshi, Haneya Naeem, and Ali Imran. “Addressing the Errors from Positioning Inaccuracy, Quantization and Scarcity in MDT Data to Enable AI Driven Optimization.” (submitted to *IEEE Transactions on Network and Service Management*)

J3. Qureshi, Haneya Naeem, Zaidi, Syed, Masood, Usama and Ali Imran, “Training Data Remains Sparse Even in the Age of Big Data: A Framework Towards Addressing the Training Data Sparsity Challenge in Cellular Networks.” (under review)

J4. Qureshi, Haneya Naeem, and Ali Imran. “Addressing data sparsity challenge in next generation cellular networks.” (planned magazine paper)

Presentations

P1. Qureshi, Haneya Naeem, and Ali Imran, “Artificial Intelligence based approaches to

solving challenges of next generation cellular networks”. 3 Minute Thesis Competition, OU-Tulsa (secured first position)

P2. Qureshi, Haneya Naeem, and Ali Imran, “Planning and Optimizing Networks of The Future: Methods to reduce errors in autonomous coverage estimation using MDT”. OU-Tulsa Research Forum 2019 (won second prize)

P3. Qureshi, Haneya Naeem, and Ali Imran, “Optimal bin width for autonomous coverage estimation using MDT reports in the presence of user positioning error”. OU Graduate Research Meeting 2019 (won the best research presentation award for academic year 2018-2019 in ECE/TCOM department)

1.5 Organization

This report is structured as follows. Chapter 2 addresses different errors in MDT-based coverage estimation to enable AI-driven cellular networks. Chapter 3 presents a framework to address the training data sparsity challenge to enhance ML/AI-based solutions for cellular network design, operation and optimization. Finally, chapter 4 discusses the conclusions and future work, and it thus concludes this report.

CHAPTER 2

Enabling AI-driven cellular network optimization by MDT

2.1 Introduction

The first step to enable such AI based automation is to have abundant telemetric data about network health and performance. This requires continuous gathering of telemetric data about network health and coverage [19]. Currently, however, network performance for the most part is gauged by methods such as drive tests, hardware or software failure alarms at the operation and maintenance center or complaints received from customers [20]. These methods incur inevitable delay and unreliability which stems from human error and low spatio-temporal granularity of reports gathered via drive tests and alarms [21]. In addition, the drive test based measurements are gathered from only a small fraction of the total coverage area, i.e., paved roads and are difficult to obtain in indoor environments. This problem is likely to aggravate with the advent of small cell enabled ultra dense networks, as the probability of cell outages is likely to increase in proportion to the cell density and increasing network complexity [3, 22]. In addition, several use cases for 5G and beyond demand low latency and high reliability requirements, which means that classic methods of drive test based or even alarm based for performance monitoring and outage detection will not suffice [19], [23].

To overcome the aforementioned challenges, 3GPP has standardized a self-organizing network use case, called minimization of drive test (MDT), which exploits the measurement reports gathered by the user equipment [24]. MDT allows the mobile network operators to collect data about network coverage and signal strength from the user equipment (UE) measurements. The UE measurement reports are tagged with their geographical location information, sent to their serving base station (BS) and ultimately used to generate cov-

erage maps [25, 26, 27]. Feeding these MDT reports as input to the network automation and optimization processes, can provide autonomous mechanisms to compensate outages quickly and seamlessly by enabling the network to detect anomalies [28], such as coverage holes, weak coverage spots, sleeping cells [29], or other performance degradation problems and make timely decisions [4], [19]. Therefore, MDT based coverage/performance estimation is a fundamental step to enable and trigger any AI or self-configuration, self-optimization or self-healing routine [3]. However, up till now the utility of MDT feature remains hindered by several issues that include sparsity of user reports and user positioning inaccuracy. These issues cause the following three major types of errors in MDT based autonomous performance estimation solutions, thereby undermining their utility:

1. Positioning error: The reported geographical coordinates of the UE obtained from the positioning techniques, such as assisted global positioning system are susceptible to errors, resulting in the reports being tagged to a wrong location [30]. These locations can also be inaccurate to preserve user privacy.
2. Quantization error: Storing MDT reports from all users is computationally inefficient. Practical implementation demands that, the coverage area is divided into bins. The average received power from each bin is then stored and used to build coverage maps. This results in quantization error due to averaging.
3. Scarcity of user reports: A key challenge in developing MDT based autonomous solutions for self-configuration, self-optimization and self healing 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, attributing to the fact that many bins might not be visited by even a single user during the reporting period.

In order to enable the full realization of MDT-based approach for performance estimation and in turn pave the way to enable AI-based autonomous networks, it is important to

characterize and simultaneously address the aforementioned three errors. Simultaneous characterization of the three errors is essential because of their inter-dependency. This report, not only presents a framework to quantify the three errors and characterize their interplay but also quantifies the overall effect of these errors on coverage estimation concurrently.

2.2 Relevant work

Authors in [20], [21], [30] addressed the reliability of MDT-based coverage estimation in the presence of positioning errors. However, these studies do not take into account the errors resulting from quantization or scarcity of user measurements. Focusing on only the quantization error, authors in [31] estimate the cell radius. Moreover, the work in [31] does not use MDT-based approach for coverage estimation.

User measurements per base station, particularly in the case of emerging small cells can be often sparse. The problem of sparsity of user reports is investigated in [32]-[33].

Authors in [32] use regression clustering for construction of received signal strength maps from a sparse set of MDT measurements. However, this work [32] assumes perfect user locations and a fixed bin width. The authors in [34] analyze the performance of selected spatial interpolation techniques used in the estimation of interference produced by an LTE-Advanced network. The authors in [35] provide a visualization method based on inverse distance weighted interpolation that shows every point of the received data as a heatmap. Another work [36] investigates several classical interpolation methods to reconstruct interference maps in cognitive radio networks. However, the effect of bin width and positioning error on the spatial interpolation techniques investigated in [32]-[36], remains unexplored.

Authors in [37], [38] 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 [39] extend the work in [38] to include a more realistic coverage hole definition, where the coverage of neighboring pixels is also taken into account. Authors in [40] propose a new technique, called Fixed Rank Kriging that is superior in terms of computational complexity as compared to Kriging. Authors in [41] 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 [42] extend this work to a multi-cell scenario. Kriging-based prediction of propagation environment is presented in [43] 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 [44]. However, the works in [37], [38], [39], [40], [41], [42] assume perfect geo-location information.

One work that takes into account the impact of location uncertainty on sparse coverage data is the work in [45], where the authors modify their earlier proposed algorithm [40], [41],[42] 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 combined effect of location uncertainty, quantization and sparsity on coverage estimation.

Studies that consider the recovery of sparse coverage data in an indoor environment include [46], [47], [33]. Using low-cost spectrum sensors in an office indoor environment, authors in [46] 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 [47], analyzes several spatial interpolation techniques based on Inverse Distance Weighting (IDW) and compares them in terms of reliability bounds of interpolation errors. Authors in [33]

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 location inaccuracy in an indoor environment. However, assessing interpolation performance for a wide range of location uncertainties is not the focus of this work [33]. Instead, it considers the interpolation performance for an average location error of 18 meters only.

Most relevant to this study is our earlier work in [17]. In [17], we determined optimal bin width by considering the impact of positioning uncertainty and quantization on coverage estimation using MDT. This work differs from [17] in the following aspects: 1) This study incorporates the effect of sparse user reports in MDT-based coverage estimation and its applications, which was not a focus of the study in [17]. Incorporating this error in coverage estimation is vital because sparsity/scarcity of data is a fundamental challenge that can become bottleneck for MDT-enabled coverage estimation for unleashing its true potential. MDT-enabled network automation require a significant amount of data as the presence of more data results in better and accurate coverage estimation models, instead of relying on assumptions and weak correlations. 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 practical applications. 2) In contrast to the study in [17], which investigates the coverage estimation error by considering its mean value, in this work, we treat the errors as random variables and determine their distributions. 3) We present a solution to solve 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 (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 [17]. 4) The work [17] considers a fixed coverage probability threshold. However, in this

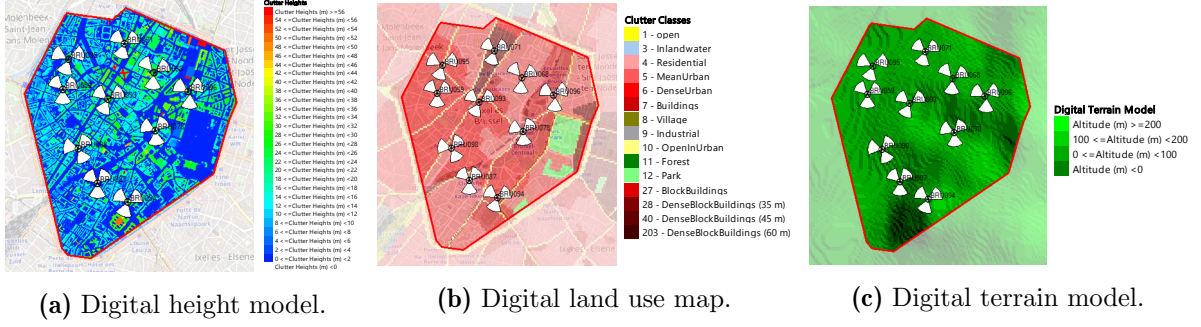


Fig. 2.1: System model configuration and geographical information.

work, 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 work.

2.3 System model

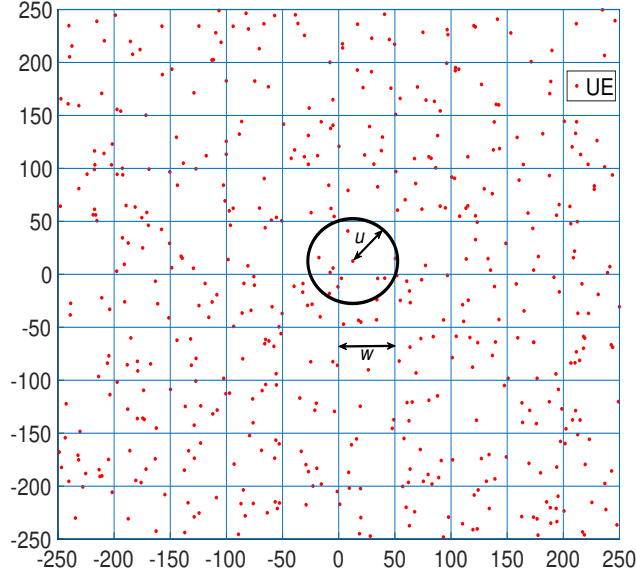


Fig. 2.2: User distribution

We use a ray-tracing based commercial planning tool [48] to create a sophisticated network topology (Fig. 2.1), in order to generate the MDT data in our study. 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

Table 2.1: 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°
Vertical half power beamwidth	4.7°

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. 2.2 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. 2.2 for one user. 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. Other simulation parameters are reported in Table 2.1.

2.4 Quantification of errors in autonomous coverage estimation using MDT

2.4.1 Error due to user positioning uncertainty

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? In order to address this question, we express the error as a random variable due to random distributions of reported and actual positions of users as:

$$E^{P,Q'}(x, y, v, q, u) = r^{P,Q'}(x, y, v, q, u) - r^{P',Q'}(x, y) \quad (2.1)$$

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) and $r^{P',Q'}$ is the received signal strength of the user without positioning uncertainty (in dBm). 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 $w_{min} = 10\text{m}$ to $w_{max} = 50\text{m}$ and u is varied from 0m to 100m. Fig. 2.3 illustrates the PDF of coverage estimation error

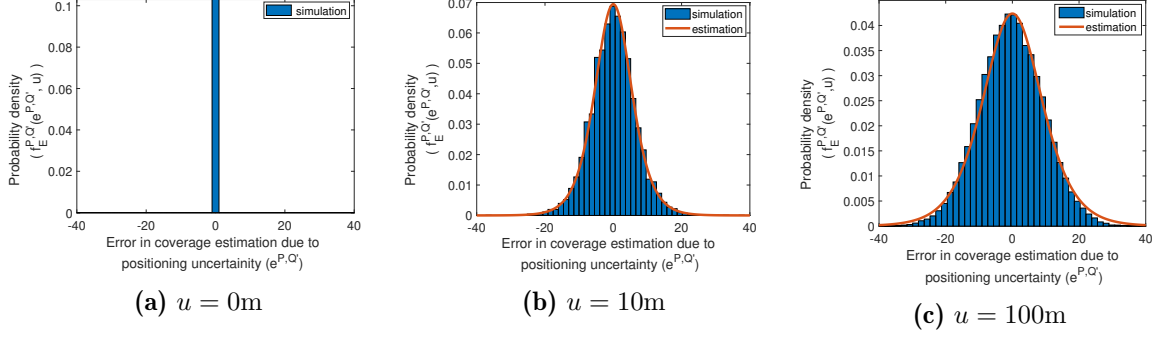


Fig. 2.3: PDF of coverage estimation error due to positioning uncertainty in the absence of bins

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 mean zero and parameter s_1 , that is proportional to the square root of variance. Using multiple terms exponential regression, we determine parameter s_1 as a function of positioning error radius as follows:

$$s_1(u) = a_1 \exp(b_1 u) + c_1 \exp(d_1 u), \quad (2.2)$$

$$\text{where } a_1 = 5.333, b_1 = 0.001, c_1 = -5.325, d_1 = -0.107$$

Fig. 2.4 shows the variation of parameter s_1 with u . The PDF of this error as a function of positioning error radius then becomes:

$$f_E^{P,Q'}(e^{P,Q'}, u) = \frac{\exp\left(-\frac{e^{P,Q'}}{a_1 e^{(b_1 u)} + c_1 e^{(d_1 u)}}\right)}{(a_1 e^{(b_1 u)} + c_1 e^{(d_1 u)}) \left(1 + \exp\left(-\frac{e^{P,Q'}}{a_1 e^{(b_1 u)} + c_1 e^{(d_1 u)}}\right)\right)^2} \quad (2.3)$$

Note that the parameters $\{a_1 \dots d_1\}$ would vary with different path loss and shadowing models. However, this is out of scope of this study and can be part of a future work. Note also that 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.,

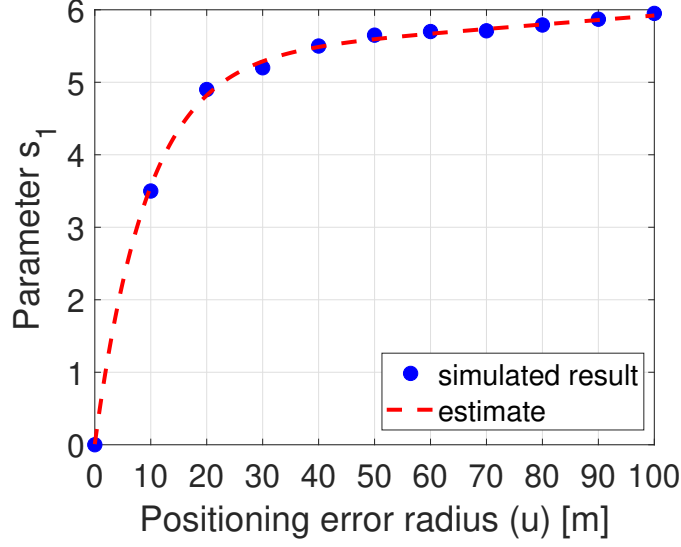


Fig. 2.4: Parameter s_1

by truncating the PDFs according to different operator-defined coverage (RSRP) error thresholds.

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. In order to address this case, we consider the following error measure:

$$E^{P,Q}(x, y, v, q, u, w) = r^{P,Q}(x, y, v, q, u, w) - r^{P',Q}(x, y, w) \quad (2.4)$$

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. 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.

To understand the impact of user positioning error as a function of bin width, consider a user located at the bin center, with coordinates (25,25) as shown in Fig.1. In the presence of no positioning uncertainty, the user is actually present at this location. However, due to positioning uncertainty, the actual location of the user lies within a circular radius u . Depending on the radius u and bin width, w , the probability of user being actually located in adjacent bins would vary, which would impact coverage estimation. We define this probability of misclassification, P_m as the probability that user's actual position lies in bin j , given that its reported position lies in bin i , where $i \neq j$. Using geometry from Fig.1, three cases of P_m can be distinguished depending on u . By expressing $\theta = 2 \cos^{-1}(w/2u)$ and calculating the fraction of area of circle with radius u that lies outside the square with side w , or equivalently, calculating the fraction of user's all possible actual locations that lie outside bin i , P_m when a user is located at the i -th bin center can be derived as follows:

$$P_m(w, u) = \begin{cases} 0, & 0 < u \leq w/2 \\ \frac{4u^2 \cos^{-1}\left(\frac{w}{2u}\right) - 2u^2 \sin\left(2 \cos^{-1}\left(\frac{w}{2u}\right)\right)}{\pi u^2}, & w/2 < u < w/\sqrt{2} \\ \frac{\pi u^2 - w^2}{\pi u^2}, & u \geq w/\sqrt{2} \end{cases} \quad (2.5)$$

P_m as a function of u and w is illustrated in Fig. 2.5. Note that the case when a user is located at the bin center is a lower bound on P_m as P_m will increase as the user moves away from the bin center. Therefore, for any arbitrary user location, the error in coverage estimation due to positioning error in the presence of bins is likely to increase with larger u for the same w or with smaller w for the same u , as the probability of misclassification would increase in these scenarios. It is observed from Fig. 2.5 that a zero probability of user location being misclassified occurs at the combination of large bin width and small positioning error radius. Note that the RSRP perceived by the users is affected by positioning error since the measured RSRP reports are tagged to wrong

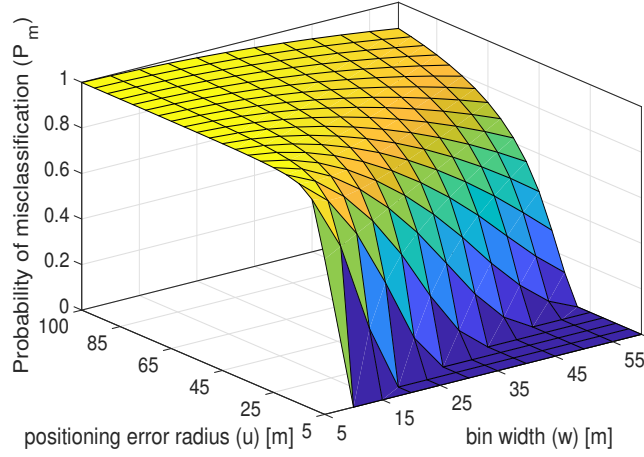


Fig. 2.5: Probability of misclassification of a user with varying bin width and positioning error radius.

locations due to positioning error. This results into error (caused by tagging to wrong location) in the RSRP-location duo reported as part of the MDT reports. This leads to error in the coverage being investigated here. Therefore, the error in coverage estimation due to positioning uncertainty is expected to be the least when bin width is large and positioning error radius is small. In order to capture this effect, we quantify the impact of user positioning error in the presence of bins as follows:

$$E^P = \frac{1}{m^2} \sum_{i=1}^{m^2} |\mathbb{P}[\mathbf{r}_i^{P,Q} > \gamma] - \mathbb{P}[\mathbf{r}_i^{P',Q} > \gamma]| \quad (2.6)$$

where the operator $\mathbb{P}$ represents probability, $\mathbf{r}^{P,Q}$ and $\mathbf{r}^{P',Q}$ are vectorized forms of matrices $\mathbf{R}^{P,Q}$ and $\mathbf{R}^{P',Q}$ respectively. The i -th element of the vector $\mathbf{r}^{P,Q}$, $r_i^{P,Q}$ represents the measured average received power of users in i -th bin in presence of positioning uncertainty and $r_i^{P',Q}$ is the average received power of users in the same bin with no uncertainty.

The effect of positioning error on coverage estimation with varying bin widths is shown in Fig. 2.6. It can be observed from Fig. 2.6 (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

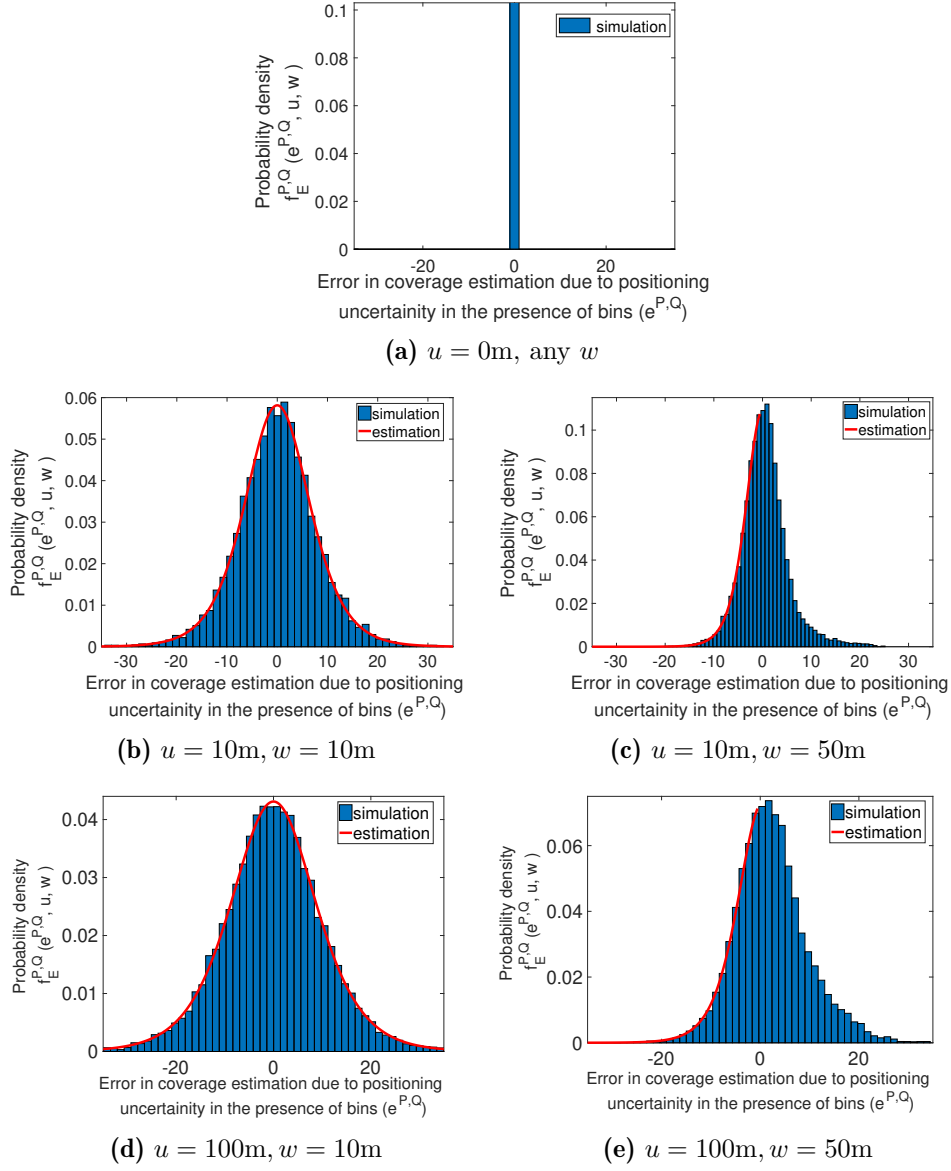


Fig. 2.6: PDF of coverage estimation error due to positioning uncertainty in the presence of bins

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. 2.6 (a). It can also be observed from Fig. 2.6 (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

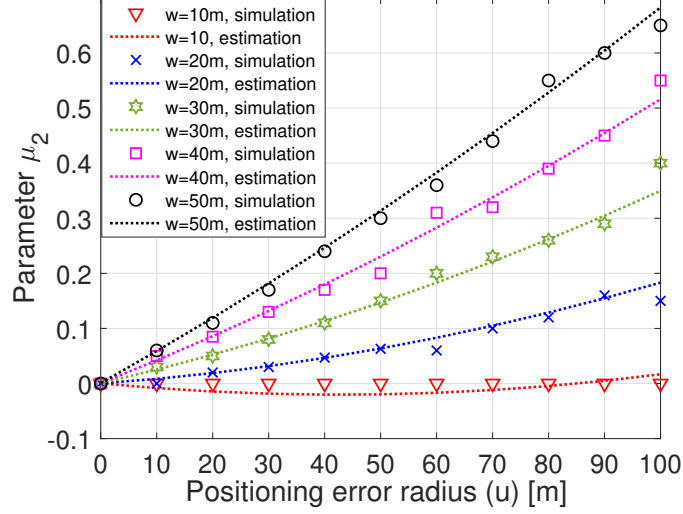


Fig. 2.7: Parameter μ_2

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:

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

$$\forall e^{P,Q} \text{ when } \mu_2 = 0, \text{ for } e^{P,Q} < 0 \text{ when } \mu_2 \geq 0 \quad (2.7)$$

where μ_2 and s_2 are as follows:

$$\mu_2(u, w) = a_2 u + b_2 w + c_2 u^2 + d_2 u w; \quad (2.8)$$

$$s_2(u, w) = (e_2 w^{f_2} + g_2) \exp(h_2 u / (w + i_2)) \\ + (j_2 w^{k_2} + l_2) \exp(m_2 w u + n_2 u) \quad (2.9)$$

where $a_2 = -0.00262, b_2 = 1.587 \times 10^{-5}, c_2 = 1.124 \times 10^{-5}, d_2 = 0.0001663, e_2 = -0.00473, f_2 = 1.538, g_2 = 5.04, h_2 = 0.04628, i_2 = 13.67, j_2 = 0.004732, k_2 = 1.538, l_2 = -5.04, m_2 = 0.001935, n_2 = -0.2277$.

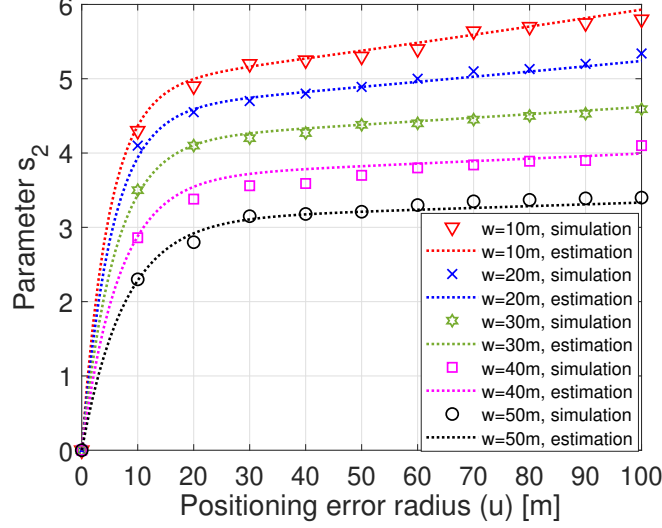


Fig. 2.8: Parameter s_2

The parameters μ_2 and s_2 are shown in Fig. 2.7 and 2.8, respectively and indicate an excellent fit between the simulated results and parameter fitting. For the purpose of determining 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 2.5.1.

2.4.2 Quantization error

Quantization error without positioning uncertainty

The error in coverage estimation incurred due to averaging by dividing the coverage area into bins can be quantified as follows:

$$E^{Q,P'}(x, y, w) = r^{P',Q}(x, y, w) - r^{P',Q'}(x, y) \quad (2.10)$$

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

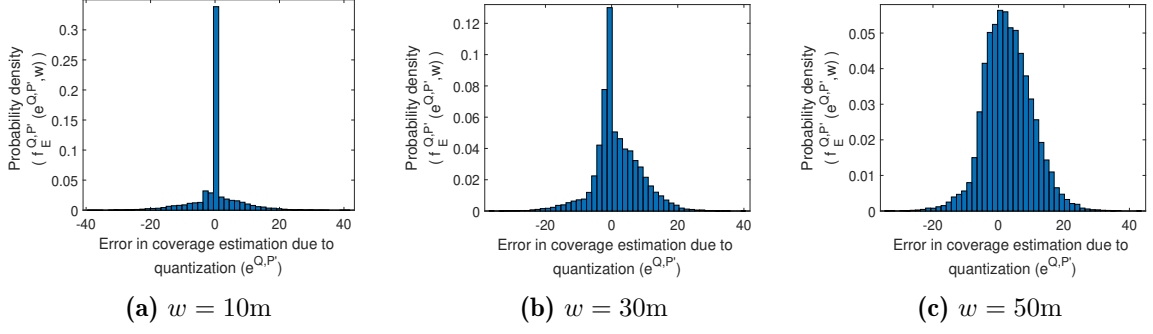


Fig. 2.9: PDF of coverage estimation error due to quantization error without positioning uncertainty

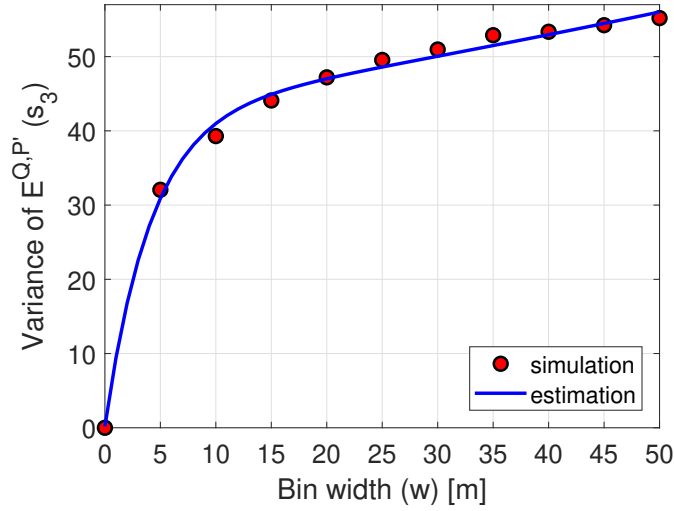


Fig. 2.10: Variance of error, $E^{Q,P'}$

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. 2.9 depicts the PDF of $E^{Q,P'}$ with varying bin widths. It can be observed that this error converges to a delta distribution (no error in coverage estimation due to quantization) as $w \rightarrow 0$. The variance of this error, s_3 is depicted in Fig. 2.10 and it increases with increase in bin width according to:

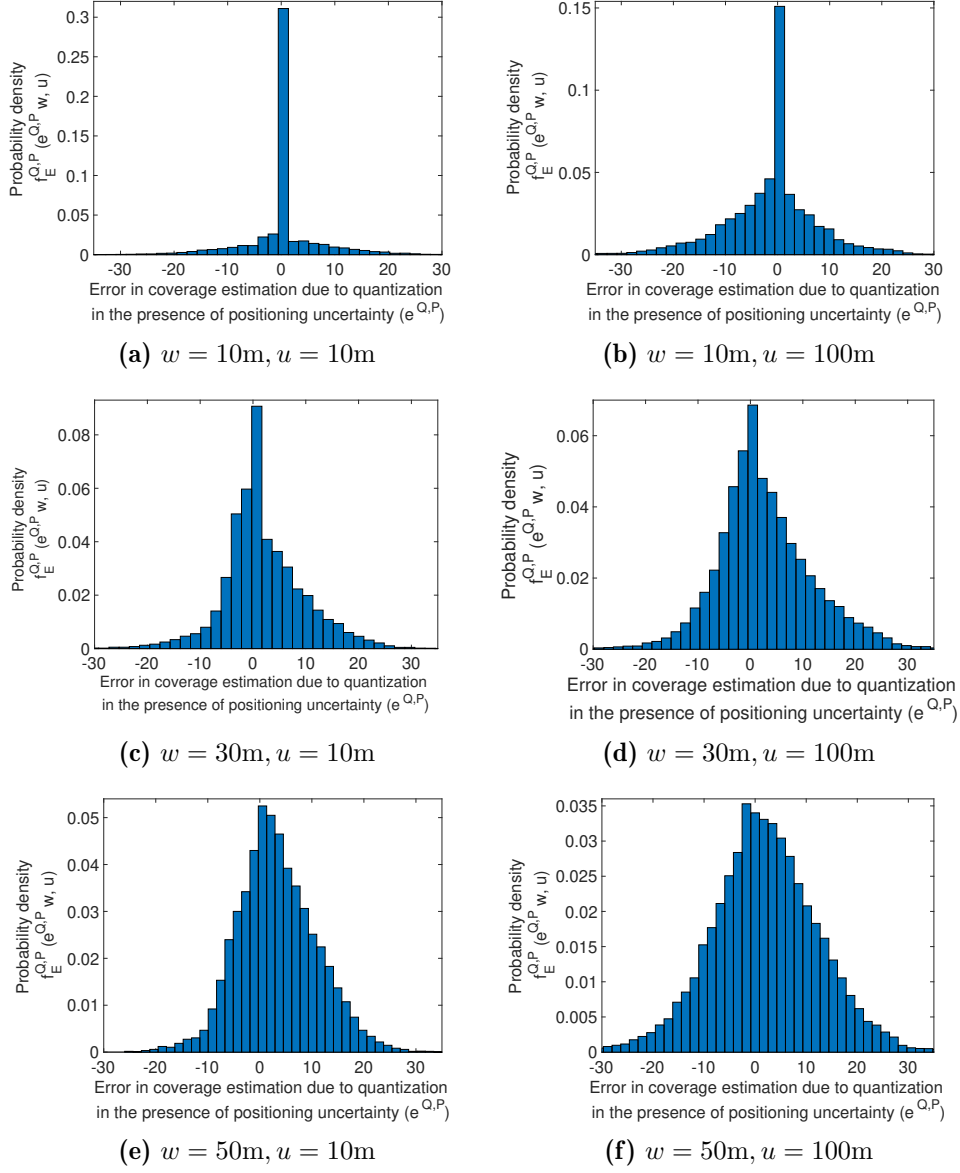


Fig. 2.11: PDF of coverage estimation error due to quantization in the presence of positioning uncertainty

$$s_3(w) = a_3 \exp(b_3 w) + c_3 \exp(d_3 w), \quad (2.11)$$

$$\text{where } a_3 = 42.34, b_3 = 0.005597, c_3 = -42.16, d_3 = -0.2408$$

Quantization error with positioning uncertainty

In this section, we show 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

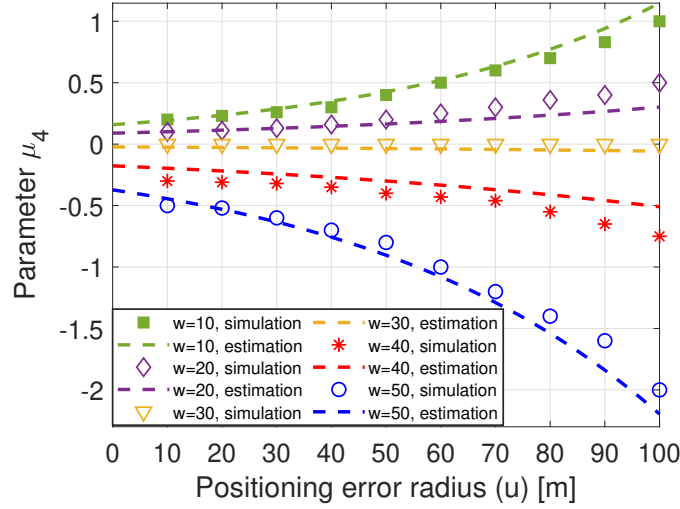


Fig. 2.12: Parameter μ_4

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:

$$E^{Q,P}(x, y, v, q, u, w) = r^{P,Q}(x, y, v, q, u, w) - r^{P,Q'}(x, y, v, q, u) \quad (2.12)$$

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. 2.11 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. 2.11 (a), (c), (e) converge to Fig. 2.9 (a), (b), (c) respectively. However, as u increases, the variance of this error and the percentage of area that is falsely estimated to be covered starts to increase. The PDF of this error can be expressed as follows:

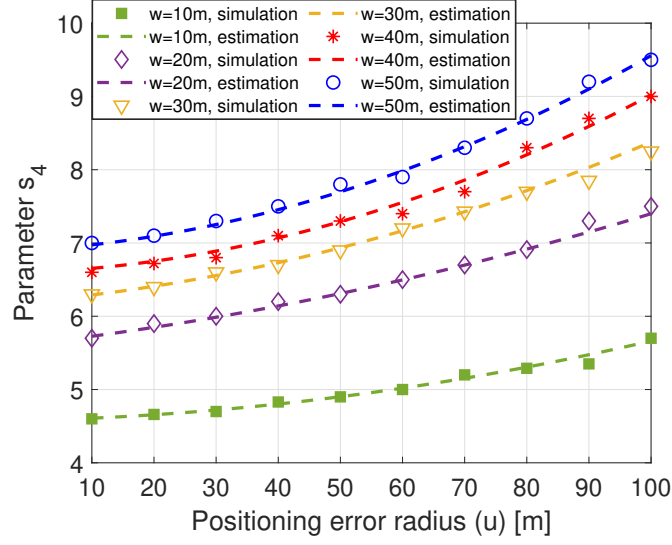


Fig. 2.13: Parameter s_4

$$f_E^{Q,P}(e^{Q,P}, u, w) = \frac{\exp\left(-\frac{e^{Q,P}-\mu_4(u,w)}{s_4(u,w)}\right)}{s_4(u,w) \left(1 + \exp\left(-\frac{e^{Q,P}-\mu_4(u,w)}{s_4(u,w)}\right)\right)^2}, \text{ for } e^{Q,P} > 0 \forall \mu_4 \quad (2.13)$$

where

$$\mu_4(u, w) = (a_4 w^{b_4} + c_4) \exp(d_4 u \exp(e_4 w) + f_4 u \exp(g_4 w)) \quad (2.14)$$

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

$$s_4(u, w) = (h_4 w^3 + i_4 w^2 + j_4 w + k_4) u^2 + (l_4 w^3 + m_4 w^2 + n_4 w + o_4) u + p_4 w^{q_4} + r_4 \quad (2.15)$$

with $h_4 = -1.086e-08, i_4 = 9.973e-07, j_4 = -2.307e-05, k_4 = 0.0002294, l_4 = 1.325e-06, m_4 = -0.0001289, n_4 = 0.00371, o_4 = -0.02346, p_4 = -20.58, q_4 = -0.09358, r_4 = 21.17$. Figures 2.12 and 2.13 illustrate the excellent fit between simulated parameters and (2.14)-(2.15).

Contrary to $E^{P,Q}$, the variance of $E^{Q,P}$ increases with increase in bin width for a fixed

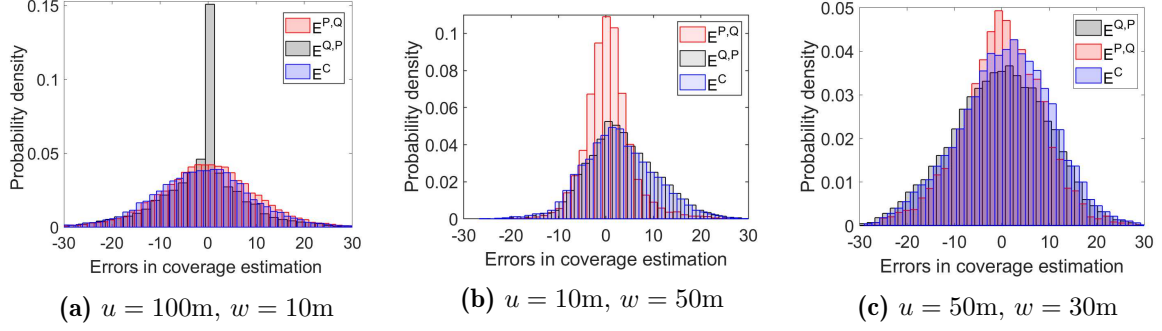


Fig. 2.14: PDF of coverage estimation error due to both positioning uncertainty and quantization

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

2.4.3 Combined effect of positioning and quantization errors on coverage estimation

In Section 2.4.1, we analyzed the effect of positioning error on coverage estimation, both with quantization and without quantization, while in Section 2.4.2, 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 report, 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:

$$E^C(x, y, v, q, u, w) = r^{P,Q}(x, y, v, q, u, w) - r^{P',Q'}(x, y) \quad (2.16)$$

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).

Fig. 2.14 illustrates the PDFs of coverage estimation errors for different bin widths and positioning error radius. In Fig. 2.14 (a), we show the case of large user positioning error ($u = 100\text{m}$) and small quantization error ($w = 10\text{m}$). 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. 2.14). 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. 2.14). 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. 2.14 (b) shows the case of small user positioning error ($u = 10\text{m}$) and large quantization error ($w = 50\text{m}$). 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. 2.14 (c) shows the case for $u = 50\text{m}$ and $w = 30\text{m}$. 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. 2.6 and Fig. 2.11, the distributions of E^C in Fig. 2.14 would converge to a delta distribution only when both $u \rightarrow 0$ and $w \rightarrow 0$ simultaneously.

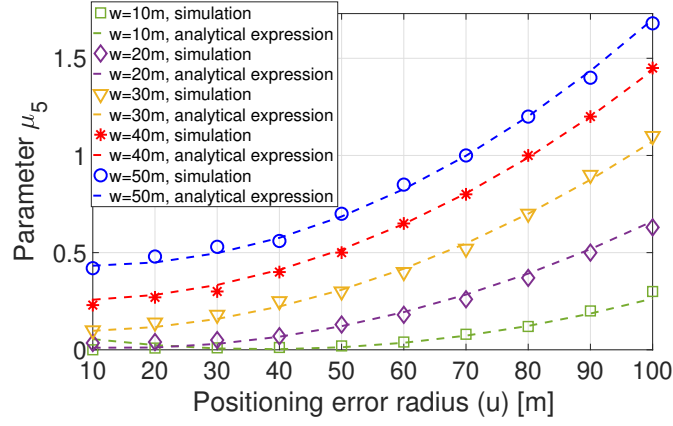


Fig. 2.15: Parameter μ_5

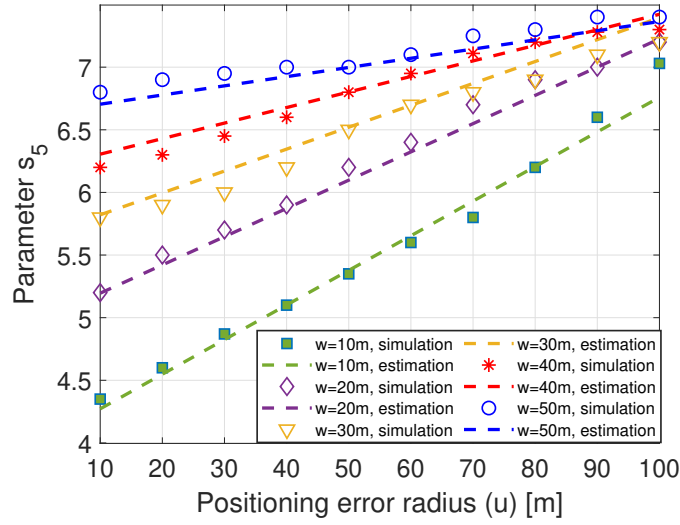


Fig. 2.16: Parameter s_5

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:

$$f_E^C(e^c, u, w) = \frac{\exp\left(-\frac{e^c - \mu_5(u, w)}{s_5(u, w)}\right)}{s_5(u, w) \left(1 + \exp\left(-\frac{e^c - \mu_5(u, w)}{s_5(u, w)}\right)\right)^2}, \text{ for } e^c < 0 \forall \mu_5 \quad (2.17)$$

where

$$\mu_5 = a_5 + b_5 u + c_5 w + d_5 u^2 + e_5 uw + f_5 w^2 + g_5 u^2 w + h_5 uw^2 + i_5 w^3 \quad (2.18)$$

$$s_5 = j_5 wu + k_5 u + l_5 w^{m_5} \quad (2.19)$$

and $a_5 = 0.372, b_5 = -0.008895, c_5 = -0.03936, d_5 = 4.532 \times 10^{-5}, e_5 = 0.0004475, f_5 = 0.0013, g_5 = 2.191 \times 10^{-6}, h_5 = -6.576 \times 10^{-6}, i_5 = -9.658 \times 10^{-6}, j_5 = -0.0005075, k_5 = 0.03271, l_5 = 1.936, m_5 = 0.3147$. Fig. 2.15 and Fig. 2.16 depict parameters u_5 and s_5 respectively.

2.4.4 Error due to scarcity of data

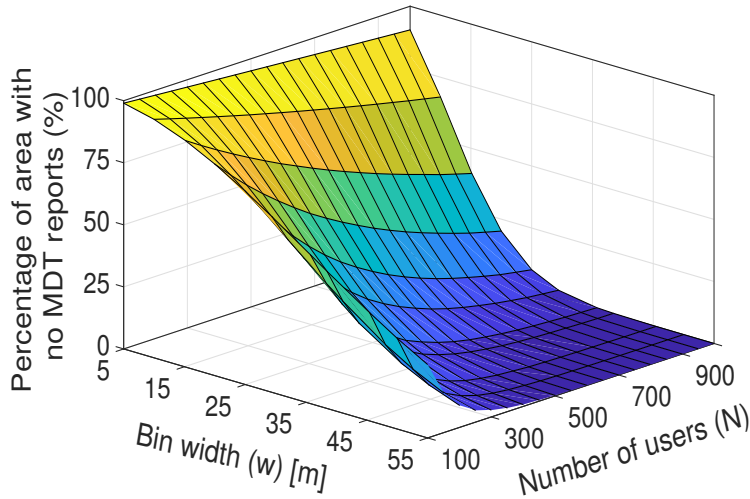


Fig. 2.17: Percentage of area with no MDT reports with varying bin width and number of users

One key challenge in practical implementation of MDT reports based coverage estimation is the sparsity of user reports. The problem of sparse MDT reports is illustrated in Fig. 2.17. From Fig. 2.17, it is observed that the mean percentage of area containing no reported user measurements increases exponentially as the bin width decreases or number of users (user density) decreases. 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.

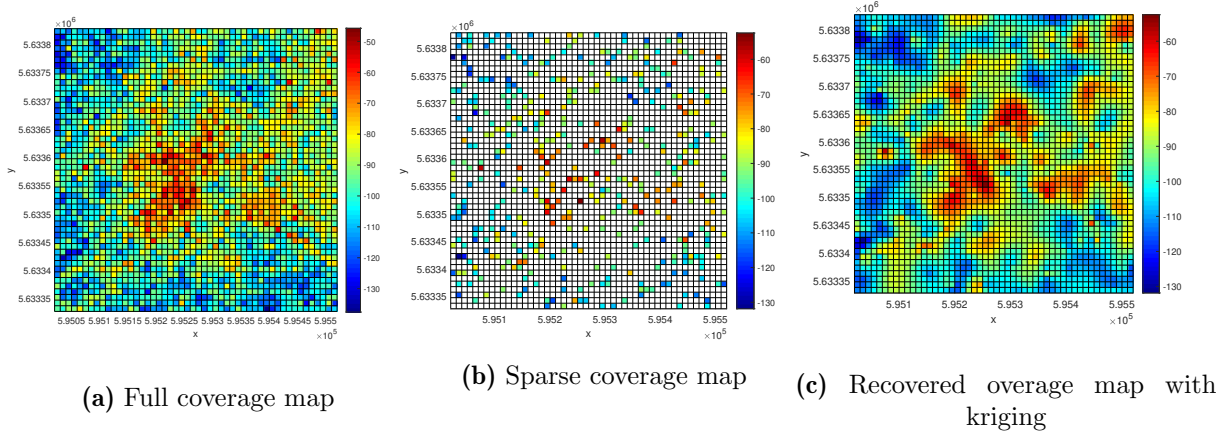


Fig. 2.18: Kriging for coverage map reconstruction, $u = 0$ and $w = 10\text{m}$

Gathered coverage data from different bins can be represented in a matrix $\mathbf{C}$ 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 quantify the accuracy of possible solutions for MDT data sparsity, we use the measure of relative error:

$$E^M = \|\hat{\mathbf{C}} - \mathbf{C}^{full}\|_F / \|\mathbf{C}^{full}\|_F \quad (2.20)$$

where $\mathbf{C}^{full}$ is the matrix with full entries, considering that RSRP measurements are available from all bins and $\hat{\mathbf{C}}$ is the recovered coverage matrix.

Fig. 2.18 illustrates the coverage map for a square area of $500\text{m} \times 500\text{m}$ taken from Fig. 2.1, with 500 users. A bin width of $w = 10\text{m}$ forms a coverage matrix of dimensions 50×50 , leading to 81.1 % missing entries in the matrix, i.e., empty bins as shown in Fig. 2.18 (b).

To predict missing coverage values in empty bins, one approach can be utilizing the average of neighboring bins to approximate the coverage value in an unreported bin.

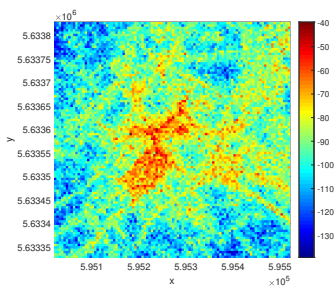
However, this scheme can induce large errors, particularly when two or more adjacent bins are empty. After carrying out a literature review of possible techniques that can be employed in order to recover these missing coverage values, we present a comparison of selected techniques and propose some new approaches to address this issue in the next section. We apply these selected techniques on sparse coverage data and the resulting outputs are shown in the next section. More approaches to addressing this challenge in various cellular networks scenarios will be discussed in Chapter 3.

Comparison of selected techniques to address data sparsity challenge

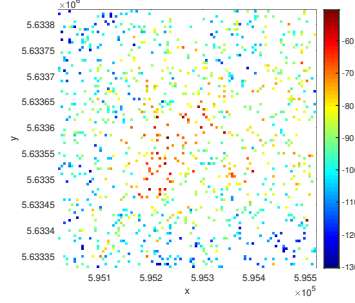
Fig. 2.19 shows a visual comparison of some selected techniques (namely, moving average, matrix completion via different algorithms, inverse distance weighted, nearest neighbor, natural neighbor, spline and kriging techniques) when applied on a sparse coverage map for a bin width of 5m. Details of these techniques are discussed comprehensively in Chapter 3. It can be seen from these figures that kriging interpolation method performs the best. Note that although Fig. 2.19 shows a part of the simulated area, the conclusions remain same for other geographical parts from the simulated area.

Positioning uncertainty is then added to the analysis and the recovery errors for $u = 100\text{m}$ are shown in Fig. 2.20.

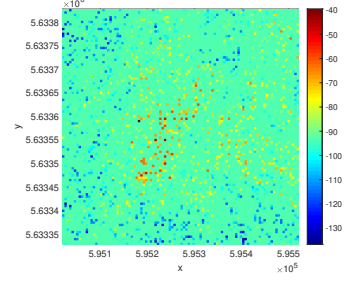
In this chapter we focus only on the accuracy of recovery methods. Other aspects, such as computational complexity are out of scope of this work and can be considered in a future study. From these results, we conclude that kriging works best in extreme scenarios of high positioning uncertainty and low 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 geo-statistical 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 2.5.



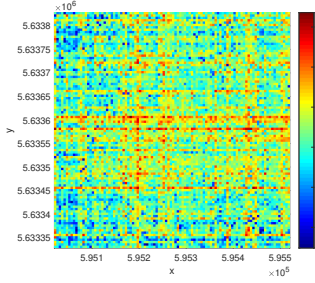
(a) Full coverage map



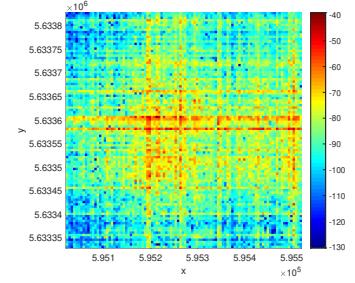
(b) Sparse coverage map



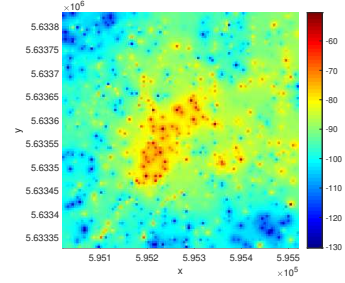
(c) Moving average



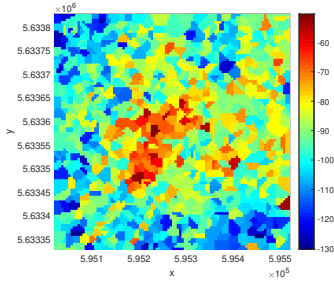
(d) Matrix completion via SVT



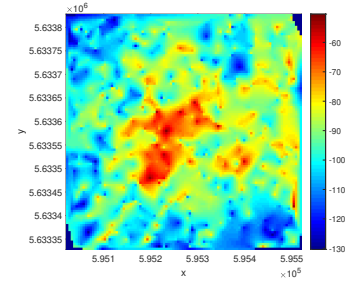
(e) Matrix completion via FPC



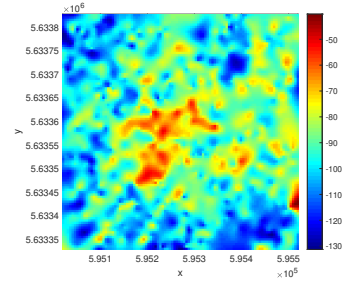
(f) Inverse distance weighted



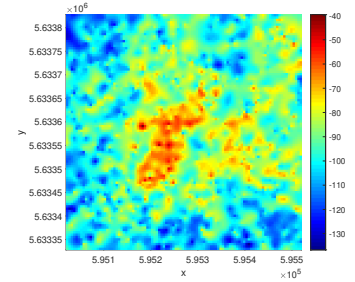
(g) Nearest neighbor



(h) Natural neighbor



(i) Spline



(j) Kriging

Fig. 2.19: Comparison of coverage map reconstruction techniques for $u = 0$ and $w = 5\text{m}$

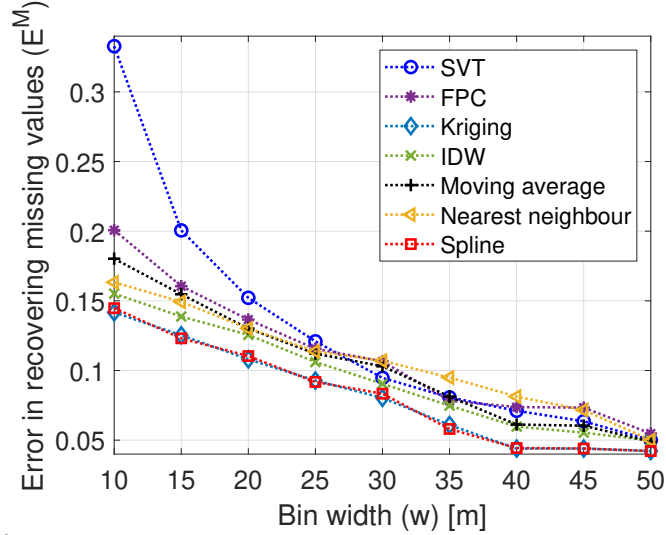


Fig. 2.20: Recovery error with varying bin widths using different reconstruction techniques.

A 3D graph for the matrix recovery error, E^M using Kriging as a function of bin width and positioning error radius is shown in Fig. 2.21. It can be seen from Fig. 2.21 that this error increases with increase in positioning error radius and decrease in bin width.

2.5 Practical applications

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

2.5.1 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.

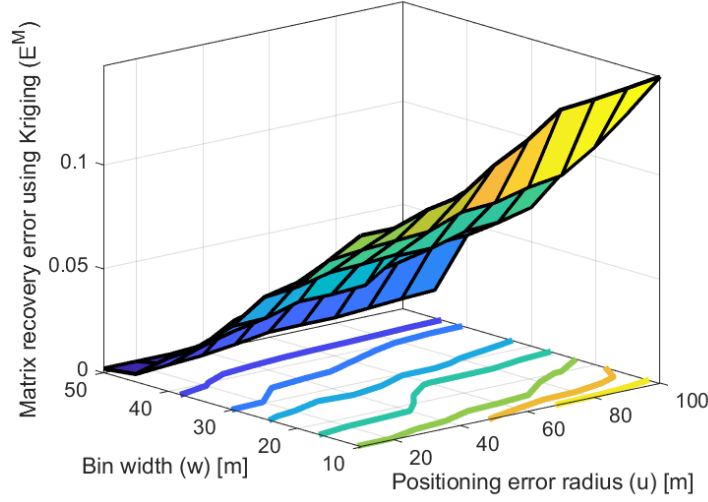


Fig. 2.21: Matrix recovery error with varying bin widths and positioning error radius using Kriging.

The probability of area whose coverage is under-estimated due to given positioning uncertainty, $A_u^P(u, w)$ can be calculated by integrating (2.7) from $-\infty$ to 0 while the probability of area that is over-estimated due to quantization, $A_o^Q(u, w)$ can be determined by integrating (2.13) from 0 to ∞ as follows:

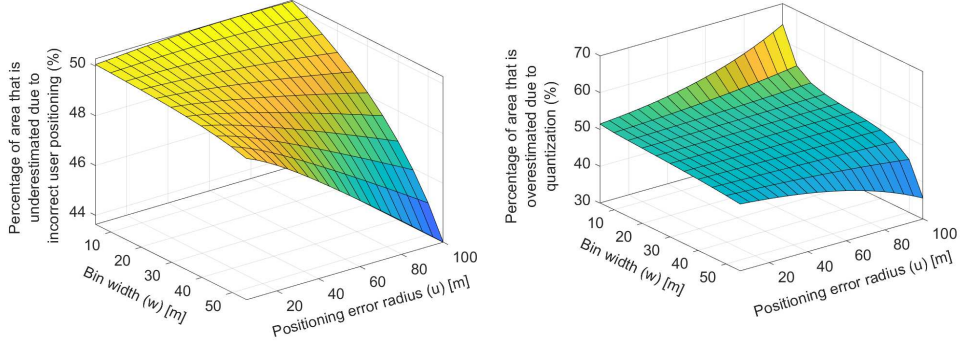
$$A_u^P(u, w) = \int_{-\infty}^0 \frac{\exp\left(-\frac{e^{P,Q}-\mu_2(u,w)}{s_2(u,w)}\right)}{s_2(u, w) \left(1 + \exp\left(-\frac{e^{P,Q}-\mu_2(u,w)}{s_2(u,w)}\right)\right)^2} de^{P,Q}$$

$$A_u^P(u, w) = \frac{1}{e^{\frac{\mu_2(u,w)}{s_2(u,w)}} + 1} \quad (2.21)$$

$$A_o^Q(u, w) = \int_0^{\infty} \frac{\exp\left(-\frac{e^{Q,P}-\mu_4(u,w)}{s_4(u,w)}\right)}{s_4(u, w) \left(1 + \exp\left(-\frac{e^{Q,P}-\mu_4(u,w)}{s_4(u,w)}\right)\right)^2} de^{Q,P}$$

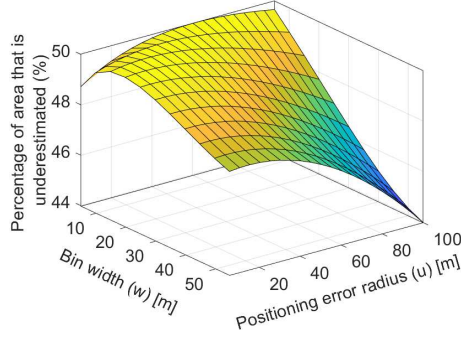
$$A_o^Q(u, w) = 1 - \frac{1}{e^{\frac{\mu_4(u,w)}{s_4(u,w)}} + 1} \quad (2.22)$$

Fig. 2.22a shows the probability of area that is under-estimated due to positioning uncertainty for given bin widths while Fig. 2.22b shows the probability of area that is over-estimated due to quantization for given positioning uncertainties. The probability of area that is under-estimated due to both quantization and user positioning error can be found by integrating (2.17) from $-\infty$ to 0, yielding the expression in (2.23) and illustrated



(a) Percentage of area that is underestimated due to incorrect user positioning as a function of w and u

(b) Percentage of area that is overestimated due to quantization as a function of w and u .



(c) Percentage of area that is underestimated due to both quantization and incorrect user positioning as function of w and u .

Fig. 2.22: Coverage area miscalculated due to different causes.

by Fig. 2.22c.

$$A_u^c(u, w) = \frac{1}{e^{\frac{\mu_5(u, w)}{s_5(u, w)}} + 1} \quad (2.23)$$

The probability of area that is over-estimated is then $1 - A_u^c(u, w)$. Note that the integral limits of (2.21)-(2.23) 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.

2.5.2 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 (2.4), (2.12) and (2.20) 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:

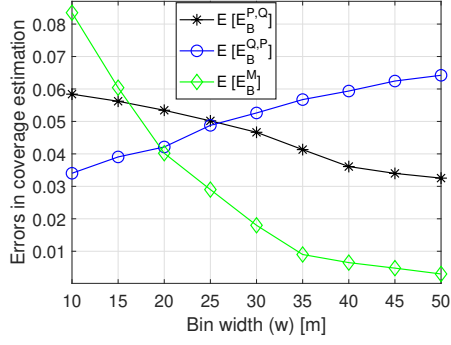
$$E_B^{P,Q} = \frac{1}{n^2} \sum_{i=1}^{n^2} \frac{|\mathbf{r}_i^{P,Q} - \mathbf{r}_i^{P',Q}|}{|\mathbf{r}_i^{P,Q} - \mathbf{r}_i^{P',Q}| + |\mathbf{r}_i^{P',Q}|} \quad (2.24)$$

$$E_B^{Q,P} = \frac{1}{U} \sum_{i=1}^U \frac{|\mathbf{r}_i^{P,Q} - \mathbf{r}_i^{P,Q'}|}{|\mathbf{r}_i^{P,Q} - \mathbf{r}_i^{P,Q'}| + |\mathbf{r}_i^{P,Q'}|} \quad (2.25)$$

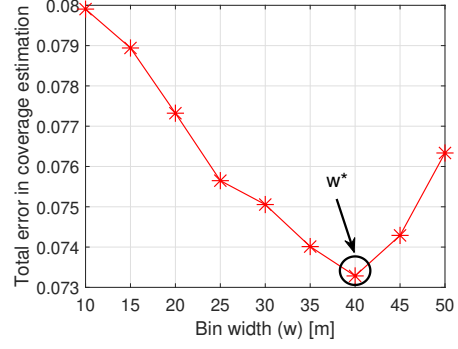
$$E_B^C = \frac{1}{U} \sum_{i=1}^U \frac{|\mathbf{r}_i^{P,Q} - \mathbf{r}_i^{P',Q'}|}{|\mathbf{r}_i^{P,Q} - \mathbf{r}_i^{P',Q'}| + |\mathbf{r}_i^{P',Q'}|} \quad (2.26)$$

where $\mathbf{r}^{P,Q}$ is the measured averaged received power vector of users in bins in the presence of positioning uncertainty and $\mathbf{r}^{P',Q}$ is the averaged received power vector of users in bins without any positioning uncertainty. $\mathbf{r}^{P,Q'}$ is the received power vector at the user level with positioning uncertainty. $\mathbf{r}^{P',Q'}$ 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).

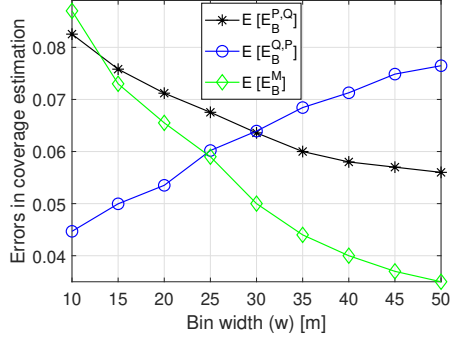
Similarly, a bounded measure for matrix recovery error (this can be considered analogous



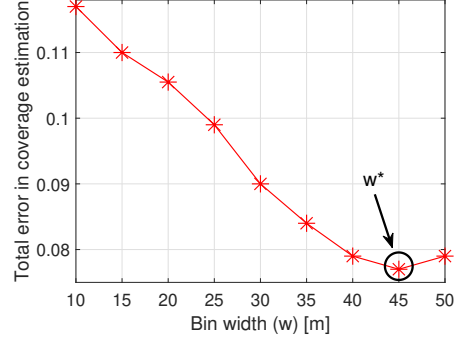
(a) Individual errors, $u = 10\text{m}$



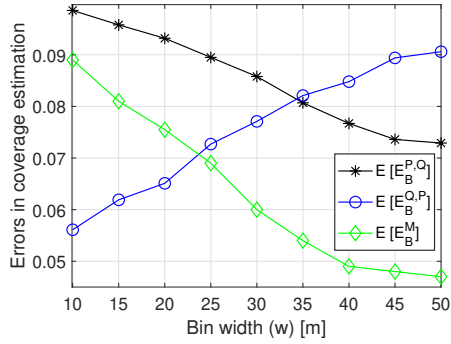
(b) Total error, $u = 10\text{m}$



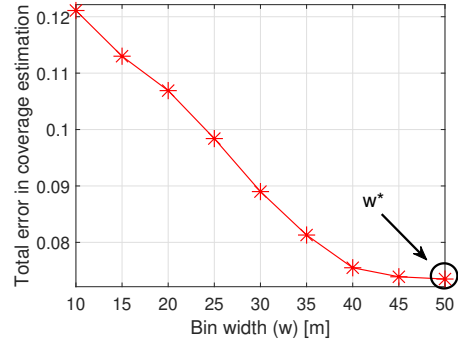
(c) Individual errors, $u = 60\text{m}$



(d) Total error, $u = 60\text{m}$



(e) Individual errors, $u = 100\text{m}$



(f) Total error, $u = 100\text{m}$

Fig. 2.23: Different errors in coverage estimation, leading to optimal bin widths.

to error caused by sparsity of MDT reports) can be expressed as:

$$E_B^M = \frac{1}{n^2} \sum_{i=1}^{n^2} \left(\frac{|\hat{\mathbf{c}}_i - \mathbf{c}_i^{full}|}{|\hat{\mathbf{c}}_i - \mathbf{c}_i^{full}| + |\mathbf{c}_i^{full}|} \right) \quad (2.27)$$

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

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:

$$w^* = \arg \min_w \mathbb{E}(E_B^M + E_B^C) \quad (2.28)$$

$$\text{subject to } w_{min} \leq w \leq w_{max} \quad (2.29)$$

$$\text{positioning error radius} = u \quad (2.30)$$

Owing to the small search space, we can solve (2.28)-(2.30) via brute force.

The quantization error, error due to incorrect user positioning and error due to sparse user reports is shown in Fig. 2.23 (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. 2.17, it becomes difficult to recover the missing coverage values as bin width decreases. Finally, Fig. 2.23 (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.

2.6 Conclusion

In this chapter, we have investigated three key types of errors in MDT-based coverage estimation that stem from sparse user measurements, quantization/binning and inaccurate user positioning. We have determined the distributions of these three types of errors as a function of positioning error radius and bin width. We also present analysis that can not only quantify the error in the estimated coverage but also the direction of the error, i.e., whether the coverage is under estimated or over estimated for given bin width, positioning error and user data sparsity. Our results reveal a very important insight that can play a pivotal role in optimal design of a MDT based coverage algorithm, i.e., there exists an optimal bin width for given user positioning inaccuracy that minimizes the overall error in the MDT based coverage estimation. Thus, for a given positioning accuracy and user density, the findings from this study can be directly used by network operators to configure the bin size that results in most accurate MDT based coverage estimation. Practical applications of this work are then presented such as determining optimal bin widths that minimize the effect of all these errors concurrently and improving the utility of the MDT estimated coverage by its calibration enabled by quantification of the errors. The findings from this work can thus aid in operation and optimization of future cellular networks as MDT based coverage estimation which can not only substantially improve the self-organizing networks based optimization in legacy networks but 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.

CHAPTER 3

A framework towards addressing the training data sparsity challenge in cellular networks

3.1 Introduction

In order to enable these automation capabilities in next generation cellular networks, the process of heterogeneous base station (BS) deployment, implementing existing and newly proposed network features and tuning the associated network parameters has to be meticulous. This is because the process of selecting an optimal network configuration that can maximize the vital key performance indicators, like coverage, capacity, reliability or energy efficiency is a rather challenging task. Identifying the optimal network configuration is necessary for network operators to fulfill the promises made by much anticipated 5G and beyond networks and to realize the efficacy of several new use cases.

Research community heavily rely on mathematical yet tractable analytical models [49]-[54] to propose planning, operation and optimization of different aspects of network. They, however, are based on restrictive assumptions and simplifications with respect to transceiver architecture, base station and user distributions and propagation characteristics, to name a few. Furthermore, stochastic geometry-based models are unable to capture the network dynamics which include mobility management and transmission latency. Therefore, several machine learning (ML) based techniques are proposed in current literature that leverage training and tuning of ML based models to determine the behavior of different configuration and optimization parameters (COPs), such as antenna tilt, transmit power, cell load in relation to different key performance indicators (KPIs), like coverage, capacity or energy efficiency [55]-[57]. These COP-KPI relationships can then be used for COP-KPI optimization. Moreover, in cellular networks context, awareness

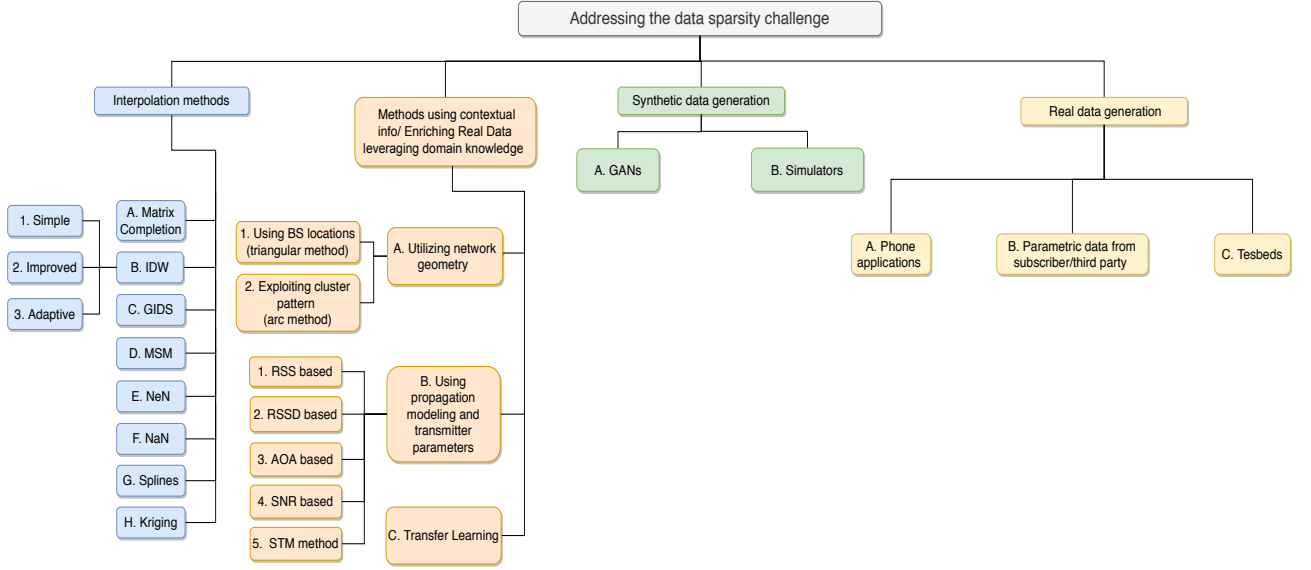


Fig. 3.1: Methods to address data sparsity challenge and chapter organization.

about radio environment in a wireless system is crucial given that the radio spectrum is a limited resource [58]. Ample data is required for constructing radio environment maps which can be used for operations such as spectrum management, to construct interference maps, to make decisions about spectrum availability for enabling dynamic spectrum access, for assessing/monitoring network health, minimizing signalling, interference management, optimization of radio resources allocation, dynamic spectrum allocation, identify bad-signal areas, automatic neighbor relation, minimize drive tests, handovers optimization and coexistence of various technologies [59]-[60]. However, all such techniques face a common key challenge that undermine their utility: scarcity/sparsity of the training data, the reasons for which were outlined in Chapter 1, Section 1.1.1.

It should be noted that measured data can be sparse and still be representative. On the other hand, data can be big but not representative. We begin this chapter by presenting an overview of techniques that will work best in the first case. In the case when data is sparse and representative, but the only information known are the measured data points and their location, interpolation methods in Section 3.2 are likely to perform best.

Moving forward, when some additional information beyond the data points and their locations is known, we can utilize the methods using contextual information or domain

knowledge in Section 3.3 or leverage the synthetic data generation method proposed in Section 3.4. On the contrary, when the available data is big and non-representative or sparse and non-representative, we address this challenge through simulators in Section 3.4.2, smartphone applications in Section 3.5.1 and an overview of existing and emerging testbeds in Section 3.5.2. In addition, for scenarios with no starting real data, for example, for new or anticipated scenarios which are not yet deployed in a real network, simulators, and testbeds to generate real data are going to be the best option for wireless communications community. An overview of this chapter organization is presented in Fig. 3.1.

3.1.1 Related Work

Data scarcity challenge has been addressed in the domain of environment sciences field, such as ecology, marine, agriculture, soil science, elevation, precipitation, and chemical concentrations. [61]-[64]. Authors in [61], [63], [64] survey a wide range of interpolation methods for use in environment sciences field. However, literature on addressing the training data sparsity challenge in cellular networks to enable AI-based solutions remains very constrained.

In cellular networks context, literature remains largely confined to either interference cartography generation techniques in cognitive radio networks [59], [65], [66], [67] or radio environment map reconstruction [68], [47], [69], [70].

In the domain of cognitive radio networks, authors in [59] compare three interpolation methods, namely, natural neighbor, kriging and spline for constructing interference cartographs from a sparse set of data. They conclude that both kriging and natural neighbor interpolations perform similarly when the channel uncertainty is lower and that the average efficiency of all interpolation techniques improves with increased shadowing decorrelation [59]. Authors in [65] conclude that Kriging performs best among nearest neighbor and inverse distance weighted (IDW) methods. Results in [66] again demonstrate the

superior performance of Kriging among nearest neighbors, IDW and triangular irregular network interpolation, but has demonstrated the robustness of IDW method overall. Among the considered techniques in [67], nearest neighbor interpolation is concluded to be the least complex method and natural neighbor, linear, cubic and quadratic interpolation techniques have shown to exhibit comparable performances in terms of primary emitter localization accuracy.

In the context of radio environment maps construction from a sparse set of measurements, the most relevant study is [68]. Advantages, disadvantages and asymptotic complexity comparison of seven different direct (interpolation) techniques and three indirect construction methods is provided. Extending the study in [68] to interference map generation, classification of interpolation methods used for interference map creation based on characteristics such as local/global, exact/inexact, deterministic/stochastic is presented in [71]. Authors in [69] compare Kriging, Modified Shepard's method (MSM) and Gradient plus inverse distance squared (GIDS) and IDW for creating radio environment maps. It concluded that Kriging and IDW are highly flexible and offer trade-off between the computational cost and accuracy, while GIDS is simple from implementation perspective. Authors in [72] use graph signal processing techniques to estimate radio environment map using sparse availability of measured data and compared its performance with Kriging in terms of higher prediction accuracy and reduced time complexity. Focus of the work in [47] is on the reliable estimation of radio interference field with small number of measurements. For this purpose, different variants of IDW spatial interpolation method are employed which have proven robustness when dealing with limited number of observations [47]. A more practical implementation of Kriging based approach using real data from the University of Colorado, Boulder campus has been demonstrated in [70].

However, to the best of authors' knowledge, there is no existing work that presents a consolidated framework to solve the training data sparsity challenge in cellular networks.

3.2 Interpolation methods

When the only information required from cellular network are the measurement values (location-value pair) in order to recover the missing values, we classify such methods as ‘interpolation methods’. Interpolation methods are based on the interpolation approach. In order to address the data sparsity challenge, spatial interpolation assumes that the data are spatially dependent and continuous over space [73].

Consider the scenario in which the predicted coverage area is divided into $n \times n$ bins. Gathered data from different bins can be represented in a matrix $\mathbf{C}$ 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. A comparison of selected reconstruction techniques for a particular scenario, in which coverage map for a square area of $500\text{m} \times 500\text{m}$ with 500 users and bin width of $w = 10\text{m}$ forms a coverage matrix of dimensions 50×50 , leading to 81.1 % missing entries in the matrix, is shown in Fig. 2.18 (b). Different recovery techniques are then applied on this matrix and the resulting outputs are shown in Fig.2.18 (c)-(j). Each technique has its own set of advantages and disadvantages; we elaborate these techniques in this section.

3.2.1 *Matrix completion theory based recovery*

We propose a scheme that jointly exploits matrix factorization theory and convex optimization to recover the missing data in matrix $\mathbf{C}$. We note that this scheme is likely to work well in small cells environments since matrix $\mathbf{C}$ will naturally be low ranked in such scenarios. 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. Moreover, 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 $\mathbf{C}$:

$$\begin{aligned} & \text{minimize} \quad \text{rank}\{\mathbf{P}\} \\ & \text{subject to} \quad P_{ij} = C_{ij} \quad (i, j) \in \Psi \end{aligned} \quad (3.1)$$

However, the problem in (3.1) 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 [74]. However, the analysis presented in [74] proves that the coverage values in vacant bins can be obtained with high accuracy by solving the following alternate convex optimization problem:

$$\begin{aligned} & \text{minimize} \quad \|\mathbf{P}\|_* \\ & \text{subject to} \quad P_{ij} = C_{ij} \quad (i, j) \in \Psi \end{aligned} \quad (3.2)$$

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

$$\|\mathbf{P}\|_* = \sum_{k=1}^n \sigma_k(\mathbf{P}) \quad (3.3)$$

In (3.3), $\sigma_k(\mathbf{P})$ denotes the k th largest singular value of $\mathbf{P}$. (3.2) therefore aims to determine the matrix with minimum nuclear norm that fits the data.

The problem in (3.2) can be solved with the singular value-based threshold (SVT) algorithm presented in [75]. The SVT algorithm solves the following problem:

$$\begin{aligned} & \text{minimize} \quad \eta \|\mathbf{P}\|_* + \frac{1}{2} \|\mathbf{P}\|_F^2 \\ & \text{subject to} \quad \mathcal{O}_\Psi(\mathbf{P}) = \mathcal{O}_\Psi(\mathbf{C}) \end{aligned} \quad (3.4)$$

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 $\mathcal{O}_\Psi(\mathbf{P})$ is equal to P_{ij} if $(i, j) \in \Psi$ and zero otherwise.

It is shown in [75] that the solution of the problem of (3.4) converges to that of (3.2) as $\eta \rightarrow \infty$. The SVT algorithm is iterative and produces a sequence of matrices $\{\mathbf{P}, \mathbf{Q}\}$. At each step, a soft-thresholding operation is performed on the singular values of the matrix $\mathbf{Q}^t$. Thus, by selecting a large value of the parameter, η in (3.4), the sequence of iterates, $\{\mathbf{P}^t\}$ converges to a matrix which nearly minimizes (3.2). Starting with $\mathbf{Q}^0 = \mathbf{0} \in \mathbb{R}^{(n \times n)}$, the algorithm inductively defines

$$\mathbf{P}^t = \text{shrink}(\mathbf{Q}^{t-1}, \eta) \quad (3.5)$$

$$\mathbf{Q}^t = \mathbf{Q}^{t-1} + \Delta_i \mathcal{O}_\Psi(\mathbf{C} - \mathbf{P}^t) \quad (3.6)$$

where $\{\Delta_i\}, i \geq 1$ is a sequence of scalar step sizes, until a stopping criteria is reached. The shrink function in (3.5) applies a soft-thresholding rule at level η to the singular values of the input matrix. It is defined as

$$\text{shrink}(\mathbf{Q}^{t-1}, \eta) = \mathcal{S}_\eta(\mathbf{Q}^{t-1}) := \mathbf{U} \mathcal{S}_\eta(\mathbf{\Sigma}) \mathbf{V}^* \quad (3.7)$$

$$\mathcal{S}_\eta(\mathbf{\Sigma}) = \text{diag}(\{(\sigma_k - \eta)_+\}) \quad (3.8)$$

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 $\mathbf{P}$, shrinking them towards zero. $\mathbf{U}, \mathbf{V}$ are matrices with orthonormal columns and the singular values $\mathbf{\Sigma}$ are positive. $\mathbf{U}, \mathbf{V}$ and $\mathbf{\Sigma}$ are obtained from the singular value decomposition of matrix $\mathbf{P}$ of rank r :

$$\mathbf{P} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^*, \quad \mathbf{\Sigma} = \text{diag}(\{\sigma_k\}), 1 \leq k \leq r \quad (3.9)$$

In case of the presence of random shadowing in the model, the stopping criteria of the algorithm can be modified as follows:

$$\|\mathcal{O}_\Psi(\mathbf{P}^t - \mathbf{C})\|_F^2 \leq (1 + \zeta) m \phi^2 \quad (3.10)$$

where ζ is a fixed tolerance. The SVT algorithm is stopped as soon as $\mathbf{P}^r$ is consistent with the data and obeys (3.10). Therefore, our reconstruction matrix, $\hat{\mathbf{C}}$ is the first $\mathbf{P}^t$

obeying (3.10). This algorithm is outlined in Algorithm 1.

Another similar rank minimization based algorithm used to recover the matrix $\mathbf{C}$ is the fixed point continuation (FPC) algorithm [76]. While SVT is efficient for large matrix completion problems, it only works well for very low rank matrix completion problems. For problems where the matrices are not of very low rank, SVT is slow and not robust and therefore, often fails [76]. To solve this problem, FPC-based algorithm is proposed in [76]. FPC-based algorithm has some similarity with the SVT algorithm in that it makes use of matrix shrinkage as in (3.5)-(3.8). However, it solves (3.4) by leveraging operator splitting technique [77].

Algorithm 1: Singular value thresholding algorithm for finding missing coverage values

Input : sampled set Ψ and sampled entries $\mathcal{O}_\Psi(\mathbf{C})$, tolerance ζ , parameter η , step size Δ , increment α , number of maximum iterations, I_M , shadowing standard deviation ϕ , and cardinality of Ψ , m

Output: $\mathbf{P}^{opt}$

```

1 Set  $\mathbf{Q}^0 = i_0 \Delta \mathcal{O}_\Psi(\mathbf{C})$ 
2 Set  $\tau_0 = 0$ 
3 for  $t = 1$  to  $I_M$ 
4   Set  $h_t = \tau_{t-1} + 1$ 
5   repeat
6     Compute  $[\mathbf{U}^{t-1}, \mathbf{\Sigma}^{t-1}, \mathbf{V}^{t-1}]_{h_t}$ 
7     Set  $t_t = h_t + \alpha$ 
8     until  $\sigma_{h_t-\alpha}^{t-1} \leq \eta$ 
9     Set  $\tau_r = \max\{j : \sigma_j^{t-1} > \eta\}$ 
10    Set  $\mathbf{P}^t = \sum_{j=1}^{\tau_r} (\sigma_j^{t-1} - \tau) \mathbf{u}_j^{t-1} \mathbf{v}_j^{t-1}$ 
11    if  $\|\mathcal{O}_\Psi(\mathbf{P}^t - \mathbf{C})\|_F^2 \leq (1 + \zeta)m\phi^2$  then break
12    Set  $Q_{ij}^t = \begin{cases} 0 & \text{if } (i, j) \notin \Psi \\ Y_{ij}^{t-1} + \Delta(C_{ij} - P_{ij}^t) & \text{if } (i, j) \in \Psi \end{cases}$ 
13 end for  $t$ 
14 Set  $\mathbf{P}^{opt} = \mathbf{P}^t$ 

```

3.2.2 Inverse distance weighted

In this section, we first discuss the simplest form of inverse distance weighted (IDW) method, the simple IDW. Then we highlight several improvements in simple IDW interpolation and finally present an adaptive IDW method from literature.

Simple IDW

The simplest form of IDW method is also known as the Shepard's method. It is based on the assumption that the distribution of signal samples is strongly correlated with distance.

To estimate the missing received signal strength value, $\hat{c}$ (at a particular bin location, D) in the matrix $\mathbf{C}$, weighted average of N known signal strength values, c_k from N adjacent bins are used, where $k = 1 \dots N$. Each known received signal strength value is weighted with a weight that is equal to the inverse of distance, $d_k = d(D, D_k)$ between the location of the bin with missing RSRP value and location of the k -th bin and raised to the power p . Mathematically, the missing received signal strength value is calculated as:

$$\hat{c} = \begin{cases} \frac{\sum_{k=1}^N \frac{1}{d_k^p} c_k}{\sum_{k=1}^N \frac{1}{d_k^p}} & \text{if } d_k \neq 0 \\ c_k & \text{if } d_k = 0 \end{cases} \quad (3.11)$$

The choice of p is an important parameter in this method. For $p < 1$, $\hat{c}$ remains no longer differentiable. Therefore, the exponent has to exceed 1 for the interpolation function to remain differentiable. It is shown by empirical testing that higher exponents tend to make the surface flat near all data points and the gradients over small intervals between data points are very steep. On the other hand, lower exponents tend to produce a relatively flat surface with short blips to achieve appropriate values at data points [78]. When $p = 0$ in (3.11), the missing coverage value is set equal to the weighted arithmetic average of the neighboring coverage values and the recovery method is often termed as the ‘moving average method’.

Simple IDW method's disadvantages are that it leads to the production of the “bull's-eyes” effect, it is sensitive to measurement outliers, it introduces significant errors in case of non-uniform distribution measurements or unevenly distributed data clusters, computational error becomes highly significant in the neighborhood of a data point, the calculation of missing value increases proportionally with the number of data points, leading to inefficiency of the method when the number of data points is large. Also,

there is no way of pre-determining the optimal weighting power factor that will construct the most accurate RF-REM. The appropriate search radius also needs to be optimized. Another drawback is the lack of directionality, i.e., different configurations of co-linear points could yield the same results, attributing to the fact that only the distances from the missing location to the points with known locations are considered and not their direction [68],[78].

However, the advantages of simple IDW method include its efficiency and ease of comprehension since it is intuitive. This interpolation works best with evenly distributed points.

Improved IDW

In order to address the drawbacks of simple IDW method in the preceding subsection, several improvements have been suggested in literature.

Authors in [79], [78] and [80] improve the weighting function by proposing a framework to intelligently select the nearby data points to be used in predicting the missing data point. This approach is developed keeping the overall density of the data points into consideration.

Authors in [78] incorporate a direction factor, in addition to the distance factor in defining the weights. This direction factor is based on the cosine of angle of D_iDD_j , where $i \neq j$ and $i, j = 1 \dots K$. If other data points D_j are in approximately the same direction from D as D_i , then the angles, $l - \cos(D_iDD_j)$ are close to 0. On the other contrary, if other data points are in the opposite D from D_i , then the angles $l - \cos(D_iDD_j)$ are close to 2. The direction factor in the improved weighting function in [78] leverages this fact.

Other improvements to simple IDW involve reduction of computational complexity and errors and making features of the interpolation function desirable, i.e., ensuring non-zero gradients at every location to achieve the desired partial derivatives for the function to

remain differentiable [1], [78].

Since simple IDW assumes that the distance decay is uniform throughout the entire study area, it does not perform well in case of clustered data or data that depicts spatial variability. To address this problem, authors in [81] suggested an improvement based on the weighted median of data in the neighborhood of missing data point. The weighting function in [81] is a function of inverse-distance weights and the de-clustered weights that include the effects of distance and clustering among spatially correlated data in the estimator.

In order to increase the accuracy of predictions through the IDW method, authors in [82] proposed the use of piecewise least-square polynomial regression estimators to increase the accuracy, after evaluating fifteen different estimators using an extensive evaluation data set.

For reducing the “bull-eye” effect in simple IDW method, a distribution-based distance weighting (DDW) technique is used [80]. Weight calculations in DDW method are based on appropriate distributions according to available data, such as Gaussian, Lorentzian and Laplacian distributions. Such a distribution-based calculated ensures that if data variations are very small, then the distribution will have a fairly sharp peak and will cause the weighting to be more sensitive to the distance. On the contrary, if data included in the interpolation are more spread out, a distribution with a larger variance would be a good choice and this would result in the distances having less impact on the weight calculations.

Authors in [80] and [83] propose another improvement to the IDW-based method, that incorporates temporal dimension in addition to spatial dimension. Although these approaches are evaluated in the context of environmental data, such an approach can also be applied to wireless network data. In the approach in [80], time is treated independently from the spatial distance dimension and weights are calculated in two steps: using the inverse of 2D-spatial distance, followed by the inverse of the 1D-temporal distance

Table 3.1: Improvements to IDW interpolation

Improvement	References
Intelligent selection of data in neighborhood	[79], [78], [80]
Addition of directionality	[78]
Reduction of computational complexity	[79],[78],[80], [1], [78]
Reduction of computational errors	[1], [78], [82]
Addition of desirable features	[1], [78]
Extension to clustered/non-uniformly distributed data	[81]
Addition of temporal dimension	[80], [83]
Reduction of “bulls-eyes” effect	[80] [68]

[80]. Authors in [83] assume second-order non-stationarity of both spatial and temporal distributions of the data, based on which they treat the space-time variables in their proposed method as a sum of independent spatial and temporal non-stationarity components. Heterogeneous covariance functions are constructed to obtain the best linear unbiased estimates in spatial and temporal dimensions [83].

Adaptive IDW

The IDW method assumes that the distance-decay structure is uniform throughout the entire study area. However, recognizing the potential of varying distance-decay relationships over area, authors in [1] proposed a variation in the value of weighting parameter, p according to the spatial pattern of sampled points in the neighborhood using information derived from empirical data. Intuitively, when the unsampled location has highly clustered points around its neighborhood, a small p is appropriate so that the nearest sampled values will not have an overwhelming influence on the estimated value. On the contrary, a large p is desirable when data is spatially dispersed since the more reliable source for the estimate will likely be influenced from the closest location, therefore, if a small p value is used in this case, the contributions from local and more reliable sources will be small, resulting in less reliable estimates [1].

In order to adjust p according to the spatial pattern of known data, authors in [1] first quantify the spatial pattern of sample locations in the form of nearest neighbor statistic:

$$R = r_o/r_e, \quad r_e = \frac{1}{2(M/A)^{0.5}} \quad (3.12)$$

where r_e and r_o are the expected and observed average nearest neighbor distances respectively and A is the area under consideration.

After normalizing R to get the normalized local nearest neighbor statistic, μ_R , in the adaptive IDW method, this neighbor statistic carries a fuzzy membership that belongs to certain categories of p . This membership function is depicted in Fig. 3.2. As an example, μ_R corresponding to R of 0.8 will be 0.35, yielding two points in the membership degree (0.3 for category C and 0.7 for category B). The final p would then be a weighted sum of these membership degrees and corresponding p values (0.5 for category B and 1 for category C). Consequently, the final p will be: $0.7 \times 0.5 + 0.3 \times 1 = 0.65$.

Adaptive IDW (AIDW) method can outperform IDW and work well in situations where local variability is relatively large or spatial correlation structure of the data is not strong or data is too limited to support data intensive methods, such as kriging. It is shown to outperform ordinary Kriging, when the spatial structure of data was such that it could not be modeled accurately by a variogram function [1].

However, as compared to IDW, the AIDW method is computational intensive as the distribution of p has to be formulated to find the optimal set of parameter values, which require significant level of heuristics [1].

3.2.3 Gradient plus inverse distance squared

Gradient plus Inverse Distance Squared interpolation (GIDS) combines multiple linear regression and inverse distance based weighted coefficients for the interpolating missing data. By assuming that the data of interest can be represented by a multivariate function,

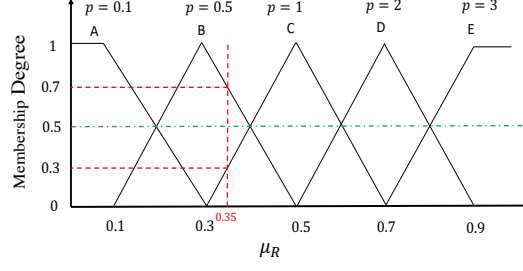


Fig. 3.2: Triangular membership function for different adaptive distance-decay parameters (modified from [1]).

for the unsampled location, D , an ordinary least squared regression is done using N neighboring locations. This yields the coefficients which represent the location gradients. If the measurements are taken at different heights, GIDS method can incorporate the elevation dimension in interpolation too. Assuming $D = (x, y, z)$ with corresponding coefficients C_x, C_y, C_z , representing the x, y, z gradients respectively, the missing data point through GIDS can be estimated as [69]:

$$\hat{c} = \frac{\sum_{k=1}^N (c_k + C_x(x - x_k) + C_y(y - y_k) + C_z(z - z_k)) / d_k^2}{\sum_{o=1}^N 1/d_o^2} \quad (3.13)$$

The advantage of GIDS method is its ability to account for signal level gradients and elevation of the terrain at the interpolated location and at locations of the measurements. However, this method is very sensitive to the selection of neighborhood points as a small neighborhood selection would leave out important measurements and a large neighborhood selection may introduce noise [68].

3.2.4 Modified Shepard's method

The IDW based modified Shepard's method (MSM) is a local interpolation that makes the estimation based on a real multivariate function, f , whose local approximation is referred to as nodal functions. If Q_k is the output of the nodal function of the data point D_k (local approximation to f at x_k, y_k), then the missing value using the MSM method can be written as a weighted average of the nodal functions within some radius influence

(about the missing data point), R_w in the following manner [69], [47]:

$$\hat{c} = \frac{\sum_{k=1}^N W_k Q_k}{\sum_{k=1}^N W_k} \quad (3.14)$$

First, the weights, W_k are calculated by the following formula:

$$W_k = \begin{cases} [R_w - d_k]/R_w d_k]^p & \text{if } d_k < R_w \\ 0 & \text{if } d_k \geq R_w \end{cases} \quad (3.15)$$

Then, another radius, R_v around each known data point is considered and the weights are again calculated using (3.15), this time, replacing R_w with R_v .

This technique can be extended to multivariate case but is dependent upon optimization of R_w , R_v and p . It is also shown to perform poorly if measurements lie in a low-dimensional subspace [68]. However, this method can reduce the “bull’s eye” effect as compared to classical IDW methods.

3.2.5 Nearest neighbor

The nearest neighbor (NeN) method is also known as proximal interpolation or point sampling. Let D_l be the nearest neighbor of the missing point, D and $d(D, D_l)$ denote the distance between D_l and D , then $\min\{d(D, D_k)\} = d(D, D_l)$, $k = 1 \dots N$. In this case, the estimated value will be the same as the value in the nearest sampled location l . Mathematically, the weights, λ_k can be represented as [66]:

$$\lambda_k = \begin{cases} 1 & \text{if } k = l \\ 0 & \text{if } k \neq l \end{cases} \quad (3.16)$$

which leads to the missing point prediction as:

$$\hat{c} = \sum_{k=1}^N \lambda_k c_k = c_l \quad (3.17)$$

Although nearest neighbor approach is of low complexity, it results in sharp transitions between the individual signal level zones and increases noise, especially at the boundary of a given area, since it does not consider the influence of the sample data points apart from the nearest neighboring data point [68], [84].

3.2.6 *Natural neighbor*

The natural neighbor (NaN) interpolation is based on Voronoi decomposition (tessellation) of a set of given points in the plane. The the received signal strength value at a particular location is found from a weighted average of N from all 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 D_k is a natural neighbor of the 2-D point $\mathbf{D}$, the portion of Voronoi region, V_{D_k} stolen away by $\mathbf{D}$ is called the natural region of $\mathbf{D}$ with respect to $\mathbf{D}_k$. 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. If we denote the Lebesgue measure of this natural region by $l_{\mathbf{D}_k}$, the natural coordinate associated to $\mathbf{D}_k$ is used as weights [59]:

$$\lambda_{\mathbf{D}_k}(\mathbf{D}) = \frac{l_{\mathbf{D}_k}(\mathbf{D})}{\sum_k l_{\mathbf{D}_k}(\mathbf{D})} \quad (3.18)$$

The weights are thus the ratio of the area of overlap to the total area of the new polygon. Once the weights are obtained, interpolation to find the missing coverage value can be carried out by a weighted sum of known coverage values.

The natural neighbor interpolation method performs well with non-homogeneous distribution of measurements as well. However, its major drawback is that it can not find missing signal values that lie outside the convex hull of Voronoi polygons since it requires

that the points to be interpolated be in the convex hull of the measurement locations as the Voronoi cells of outer data points are open-ended polygons with an infinite area [68].

3.2.7 *Splines*

The spline method is also referred to as the radius basis function and “rubber sheeting” [68]. It estimates the missing value by a mathematical function or piecewise defined polynomials called splines that minimizes the total surface curvature. This results in a smooth surface that passes exactly through the sampled points.

There are different kinds of splines, such as linear, quadratic, cubic, biharmonic and thin-plate splines. For example, in the case of thin-plate splines, the unknown value is estimated as [59]:

$$\hat{c} = \sum_{k=1}^N w_k ||D - D_k||^2 \ln(||D - D_k||) \quad (3.19)$$

where $||.||$ is the Euclidean norm. w_k can be obtained by solving $\mathbf{O}\mathbf{w} = \mathbf{i}$, where $\mathbf{i}$ and $\mathbf{w}$ are the column vectors of input data points and weights respectively, while $\mathbf{O}$ is the matrix of output of the basis function ($||D - D_k||^2 \ln(||D - D_k||)$ in this case) for all possible input values.

Spline method can generate perform fairly well even with a small number of measurements whose surface can be represented with a gentle variance and perform relatively smooth interpolation, which can be useful to recover coverage maps. However, the spline method is sensitive to outliers due to the inclusion of original data values at the sample points. It also performs poorly when measurements are closely clustered and have large variance [73]. Moreover, solving the system of linear equations in order to determine weights adds to the computational complexity of this method.

3.2.8 Kriging

Kriging, unlike the other methods discussed above, also takes into account the statistical relationships in addition to spatial relationships among the measured data points to estimate the missing values of data.

In Kriging, the weights are based not only on the distance between the measured points and the prediction location but also on the overall spatial arrangement of the measured points [85]. The weight coefficients are calculated by minimizing the variance of the estimation error, σ_e^2 :

$$\sigma_e^2 = \mathbb{V} [\hat{C}_m - C_m] \quad (3.20)$$

where $\mathbb{V}$ is the variance operator and C_m is the missing coverage value located at the 2-D point, $\mathbf{p}$.

The first step in kriging therefore involves creating a prediction surface map in order to uncover the dependency rules to make predictions. To achieve this, kriging first creates a semivariogram and covariance functions to estimate the statistical dependence values that depend on the model of autocorrelation. To solve the optimization problem in (3.20), semivariogram function, γ is used to characterize the spatial correlation.

The next step is to fit a model to the points forming the empirical semivariogram. A mathematical function is used to fit the empirical semivariogram as the theoretical semivariogram model to model spatial autocorrelation. There are many variants of kriging based on advanced and robust semivariogram models, such as simple kriging, block kriging, factorial kriging, kriging with a trend, dual kriging, universal cokriging, kriging with an external drift, indicator kriging, probability kriging, to name a few. A comparison of these variants is presented in [64], [61]. 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 in (3.20), which yield the following solution [66]:

$$\begin{bmatrix} \lambda \\ \delta \end{bmatrix} = \mathbf{X}^{-1} \mathbf{y} \quad (3.21)$$

where $\mathbf{X}$ and $\mathbf{y}$ are defined as:

$$\mathbf{X} = \begin{bmatrix} X_{1,1} & \cdots & X_{1,N} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ X_{N,1} & \cdots & X_{N,N} & \vdots \\ 1 & \cdots & 1 & 0 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \\ 1 \end{bmatrix} \quad (3.22)$$

Each element of matrix, $\mathbf{X}$, $X_{i,j} = \gamma(\|\mathbf{p}_i - \mathbf{p}_j\|)$ and each element of the column vector $\mathbf{y}$, $y_i = \gamma(\|\mathbf{p} - \mathbf{p}_i\|)$. The extra element in the weight vector solution in (3.21), δ , is the result of fitting by assuming a mean trend component in the reconstructed coverage matrix.

The major drawbacks of Kriging are that it requires a large number of measurement points in order to achieve high precision and it involves significant input from the user in order to select the best fit function for the semivariogram. Identifying the most appropriate theoretical variogram for the given data is critical in order for Kriging to perform well. If, for example, the data exhibits large spatial heterogeneity or the number of data points are too less, the theoretical variogram function may not be able to reflect the spatial structure of the data adequately, and therefore, yield poor predictive results when Kriging is performed. Although Kriging has relatively high computational complexity, it is the most commonly applied technique in the literature [85] [73] due to its higher precision. As Kriging is geostatistical method, it also can estimate the variances of predicted values in the unsampled location.

3.3 Methods using contextual information

The preceding section discussed techniques that can be leveraged to address the data sparsity challenge when the only known information are the measured data and their

locations. However, if some additional information other than the observed data is known, taking advantage of that extra information, we can employ other techniques, or use that extra information to enhance the direct methods. We classify those methods as ‘indirect methods’ which require additional information other than the observed data in order to enrich it. This additional information can be knowledge of propagation model, such as path loss and other relevant parameters, transmitter parameters, such as transmit power or antenna patterns, transmitter location estimation, network geometry, or characteristics of the operating environment. This additional information is then combined with observed sparse data to augment it. Based on the availability of known information, different indirect approaches can be employed.

3.3.1 Utilizing geometry of network

Interpolation using locations of data base stations

One way to estimate measurements for bins with no user reports can be using the geometry of the base stations as shown in Fig. 3.3. This is particularly suitable in ultra-dense deployment scenarios [86], where the data base stations (DBSs) are very densely deployed (by virtue of switching OFF DBSs to keep energy consumption and interference low). These additional measurements, after appropriate transformation, can then be used to increase the accuracy of interpolation methods proposed above. However, this approach can complement only simple measurements such as received signal strength.

Exploiting pattern among clusters in polar coordinates

Another way to enrich sparse data in a given network area can be by dividing the area into clusters into polar coordinates as shown in Fig. 3.4. Each cluster has a value that can show a given KPI, such as the average RSRP or SINR of the users in that cluster. To find the missing value in a particular cluster, geometric pattern among other clusters

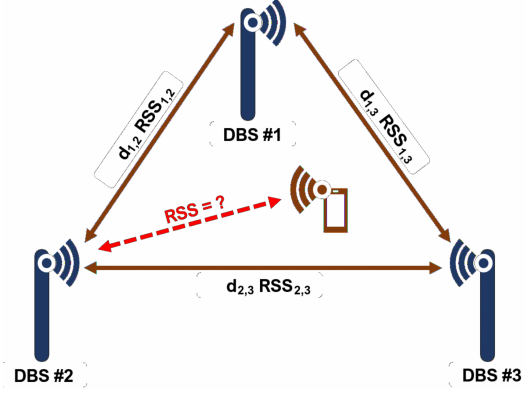


Fig. 3.3: Leveraging dense base station deployment to enrich sparse data.

can be exploited, for example, if we travel along a particular circumference, we observe that the Tx-Rx distance remains constant on that circumference and the only variation is in azimuth angle (θ_1 to θ_4 in Fig. 3.4). Conversely, if we traverse a path radially outwards, we can notice that the azimuth angle remains the same but there is variation in Tx-Rx distance (d_1 to d_3 in Fig. 3.4 assuming base station is located at the center of the sector). If we model the received signal strength as a function of azimuth angle and Tx-Rx distance, this pattern can be exploited to find the unknown signal strength values.

Learning cluster values by exploiting this pattern using a supervised DNN has been proposed in [87]. However, authors in [87] has not used this approach to address the data sparsity challenge. In [87], correlations among their SINRs has been exploited to learn the locations of users at macrocells. However, we propose that such a model based on correlations among SINRs of known clusters can also be used to find the missing SINR in another cluster.

3.3.2 Through propagation modeling and transmitter parameter estimation

RSS based method

The RSS based method to recover sparse data is based on a combination of analytical models with statistical evaluation through measurements [88]. The RSS at a particular

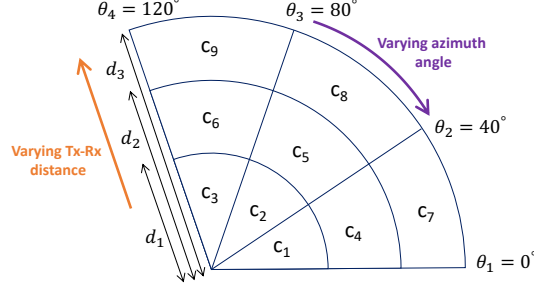


Fig. 3.4: Leveraging cluster geometry to enrich sparse data.

receiver, i located at a distance, d can be represented as:

$$P_i(d) = P_t - L - 10p \log_{10}(d) + \phi \quad (3.23)$$

where P_t is the transmit power, L is the free space path loss and ϕ represents a lognormal random variable for shadowing. L , p and standard deviation of ϕ are environment dependent parameters.

After averaging out RSS measurements (in order to reduce random shadowing effect), and assuming the sample size of RSS measurements is large enough, the average RSS at a particular location can be estimated as:

$$P_i^{av}(d) \approx P_t - L - 10p \log_{10}(d), \text{ where } P_i^{av}(d) = \sum_{k=1}^N P_i^k(d)/N \quad (3.24)$$

After performing some algebraic manipulations, taking the anti-log of (3.24) and representing d in cartesian coordinates, (3.24) can be transformed into a regression problem which can be expressed as a system of linear equation as follows [89]:

$$\begin{bmatrix} 10^{\frac{-L-P_1^{av}(d)}{5p}} & 2x_1 & 2y_1 & -1 \\ 10^{\frac{-L-P_2^{av}(d)}{5p}} & 2x_2 & 2y_2 & -1 \\ \vdots & \vdots & \vdots & \vdots \\ 10^{\frac{-L-P_N^{av}(d)}{5p}} & 2x_N & 2y_N & -1 \end{bmatrix} \begin{bmatrix} 10^{\frac{P_t}{5p}} \\ x_t \\ y_t \\ x_t^2 + y_t^2 \end{bmatrix} = \begin{bmatrix} x_1^2 + y_1^2 \\ x_2^2 + y_2^2 \\ \vdots \\ x_N^2 + y_N^2 \end{bmatrix} \quad (3.25)$$

where x_t, y_t is the transmitter location and (x_i, y_i) is the i -th receiver location. Therefore, by solving (3.25) using least-squares methods, we get estimates for transmit power, P_t and the location of transmitter, (x_t, y_t) . These estimates can then be used to evaluate

estimated received power at the missing location, by first calculating the Tx-Rx distance at the missing location and then using it to find RSS.

Note that since path loss and shadowing parameters in the model are assumed to be known and are highly environment dependent, the quality of estimated is likely to be drastically affected if there is an error in estimation of propagation parameters, caused by, for example, high shadowing fading in the environment. However, this method is likely to show improve if propagation conditions are not too drastic, for example, in rural areas and if the number of receivers with known measurements are large. It is also shown in [89] that unlike IDW and Kriging, RSS-based method is not affected by the minimum distance between receiver and transmitter and therefore, is more robust as compared to direct methods.

RSSD based method

The RSSD method is based on the received signal strength difference (RSSD) between two base stations or transmitters. It is assumed that transmit power is known, transmitter location, (x_t, y_t) is estimated based on the idea that the ratio of the signal powers (or their differences expressed in dB) observed at two different receiver locations is related to the ratios of the transmitter to receiver distances. Specifically, the received power differences between any two receivers, located at (x_a, y_a) and (x_b, y_b) can be represented as [88]:

$$P_{ab} = 5p \log_{10} \left(\frac{(x_t - x_a)^2 + (y_t - y_a)^2}{(x_t - x_b)^2 + (y_t - y_b)^2} \right) \quad (3.26)$$

The transmitter location in (3.26) can then be estimated by solving a linear system of equations of the following form:

$$\begin{bmatrix} 1 - \beta_{12} & -2(x_2 - \beta_{12}x_1) & -2(y_2 - \beta_{12}y_1) \\ 1 - \beta_{13} & -2(x_3 - \beta_{13}x_1) & -2(y_3 - \beta_{13}y_1) \\ \vdots & \vdots & \vdots \\ 1 - \beta_{1N} & -2(x_N - \beta_{1N}x_1) & -2(y_N - \beta_{1N}y_1) \end{bmatrix} \begin{bmatrix} x_t^2 + y_t^2 \\ x_t \\ y_t \end{bmatrix} = \begin{bmatrix} \beta_{12}(x_1^2 + y_1^2) - (x_2^2 + y_2^2) \\ \beta_{13}(x_1^2 + y_1^2) - (x_3^2 + y_3^2) \\ \vdots \\ \beta_{1N}(x_1^2 + y_1^2) - (x_N^2 + y_N^2) \end{bmatrix} \quad (3.27)$$

where $\beta_{ab} = \frac{(x_t - x_a)^2 + (y_t - y_a)^2}{(x_t - x_b)^2 + (y_t - y_b)^2}$. Solution to (3.27) by ordinary least squares using available receiver locations yields estimates for $x_t, y_t, x_t^2 + y_t^2$. Once the transmitter location has been estimated, the received signal level at any location can also be estimated by subtracting the path loss from transmitted signal power. As with RSS based method, this method is also dependent on selection of propagation parameters, such as path-loss exponent and shadowing spread.

AOA based method

Using prior knowledge of transmit power and using measurements from N receivers with known locations, this method first estimates the angles of arrival at the locations of the measurements and combines them with the received signal powers to estimate the location of the transmitter. Once the location of the transmitter and its transmit power is available, any appropriate propagation model can be applied to estimate unknown data at different locations.

The signal model for received signal at i -th receiver is modeled as [90]:

$$\mathbf{R}_i = \sqrt{\alpha}(d_i)\mathbf{h}(\theta_i)s + \mathbf{n}_i \quad (3.28)$$

where s is the complex baseband transmitted signal with known transmit power, d_i is

the unknown distance between the unknown transmitter and receiver, θ_i is the unknown angle by which the signal reached the i -th receiver and $\mathbf{n}_i$ is additive white Gaussian noise vector. The (θ_i, d_i) pair represents a unique position. The directional and attenuation characteristics of the channel $\mathbf{h}$ can be modeled by:

$$\mathbf{h}(\theta_i) = \begin{bmatrix} 1 \\ \exp(j\frac{\pi}{2}\sin(\theta_i)) \end{bmatrix}, \quad \alpha(d_i) = \phi\left(\frac{c}{4\pi f}\right) d_i^{-p} \quad (3.29)$$

For the recovery of missing measurements, first, the angle of arrival based on the received signal strength is estimated at each receiver and then a fusion of these estimates is performed. For angle of arrival estimation, authors in [90] apply the multiple signal classification (MUSIC) algorithm and obtain estimated of the pair (θ_i, d_i) , that translate into a location estimate for the i -th receiver:

$$\begin{bmatrix} \hat{x}_t^i \\ \hat{y}_t^i \end{bmatrix} = \begin{bmatrix} x_i \\ y_i \end{bmatrix} + \begin{bmatrix} \hat{d}_i \cos(\hat{\theta}_i) \\ \hat{d}_i \sin(\hat{\theta}_i) \end{bmatrix} \quad (3.30)$$

Next, these estimated locations are transferred to a central network that combines these estimates. One way to combine these estimates can be through simple averaging. Another fusion method proposed in [91] obtains the following over-conditioned system from the estimates:

$$\begin{bmatrix} -x_1 \sin(\hat{\theta}_1) + y_1 \cos(\hat{\theta}_1) \\ \vdots \\ -x_N \sin(\hat{\theta}_N) + y_N \cos(\hat{\theta}_N) \end{bmatrix} \approx \begin{bmatrix} -\sin(\hat{\theta}_1) & \cos(\hat{\theta}_1) \\ \vdots & \vdots \\ -\sin(\hat{\theta}_N) & \cos(\hat{\theta}_N) \end{bmatrix} \begin{bmatrix} \hat{x}_t \\ \hat{y}_t \end{bmatrix} \quad (3.31)$$

Solving this system of equations through least squares solutions yields the transmitter location, which can then be combined with known transmit power and a suitable propagation model to estimate signal strengths at unknown locations.

SNR based method

The initial steps of this method are similar to AOA based method in which the estimation step at each receiver enables the estimation of the angle of arrival and the received signal power. However, in the later step, combination of the location estimates is done through SNR-aided fusion. The basic idea of this approach is the observation that receivers far away from the transmitter yield worse location estimates. Hence the receiver results are weighted with their respective receiver's SNR, Γ_i as follows [90]:

$$\begin{bmatrix} \hat{x}_t \\ \hat{y}_t \end{bmatrix} = \sum_{i=0}^N \frac{\Gamma_i}{\sum_{k=1}^N \Gamma_k} \begin{bmatrix} \hat{x}_t^i \\ \hat{y}_t^i \end{bmatrix} \quad (3.32)$$

where the received SNR at the i -th receiver is:

$$\Gamma_i(d) = E \left[\frac{\alpha(d_i)P_t}{N_o B} \right] \quad (3.33)$$

with N_o being the the noise power density and B being the bandwidth of the receiver.

Self-tuning method

Another method utilizing propagation parameters but also taking the antenna pattern into account is the self-tuning method (STM) is proposed in [92]. In addition to leveraging characteristics of the operating environment, it performs estimation of the transmitter location, antenna parameters, transmit power and parameters of the propagation model such that the error between available measurements and predicted data is minimized.

Using the sparse data collected, the STM first estimates transmitter parameters and calibrates the propagation model. This is then used to predict missing data, such as signal levels. Among these transmission parameters, the location of transmitter is calculated using localization algorithms based on parameters such as angle of arrival or timing advance, time of arrival or time difference of arrival. Then, based on the transmitter location, distance from transmitter to receiver is calculated. This distance is then used in

an appropriate propagation model. As an example, if the Okumura-Hata model is used, the received power at a particular location can be represented as:

$$\begin{aligned}
P_r = P_t - A_o - A_1 \log_{10}(d) - A_2 \log_{10}(H_e) - \\
A_3 \log_{10}(d) \log_{10}(H) + 3.2(\log_{10}(11.75H_m))^2 - \\
44.49 \log_{10}(f) + 4.78(\log_{10}(f))^2 - L_d - L_c + G
\end{aligned} \tag{3.34}$$

where P_t is the transmit power, d is the transmitter-receiver distance, f is the operating frequency, L_d represents the diffraction loss, L_c is the loss through terrain clutter, H is the height of transmitter and A_o, A_1, A_2, A_3 are the constant coefficients. G represents the antenna gain and can be represented as [92]:

$$G = G_{\max} - F_\theta + F_\theta \left| \cos^{p_1} \left(\frac{\theta_{azi} - \theta_u}{2} \right) \right| - F_\phi + F_\phi \left| \cos^{p_2} \left(\frac{\phi_{tilt} - \phi_u}{2} \right) \right| \tag{3.35}$$

where ϕ_{tilt} is the tilt angle of the antenna, ϕ_u is the vertical angle from the reference axis (for tilt) to the user. θ_{azi} is the angle of orientation of the antenna with respect to horizontal reference axis i.e., positive x-axis, θ_u is the angular distance of the user from the horizontal reference axis. $G_{\max}$ represents the maximum antenna gain and F_θ and F_ϕ are the front to back ratios in both directions, whereas the antenna form is approximated with the cosine functions to the power of p_1 and p_2

We suggest that another option for a more practical directional antenna model defined by 3GPP and utilized in [53] can be as follows:

$$\begin{aligned}
G = \lambda_\phi \left(G_{\max} - \min \left(12 \left(\frac{\phi_u - \phi_{tilt}}{B_\phi} \right)^2, A_{\max} \right) \right) + \\
\lambda_\theta \left(G_{\max} - \min \left(12 \left(\frac{\theta_u - \theta_{azi}}{B_\theta} \right)^2, A_{\max} \right) \right)
\end{aligned} \tag{3.36}$$

The additional antenna parameters in this model are the half power vertical and horizontal beamwidths, B_ϕ and B_θ respectively and the side and back lobe attenuation, $A_{\max}$.

Having defined a suitable propagation and antenna model, the optimal antenna, trans-

mitter and propagation environment parameters can then be obtained by minimizing the mean squared error between the measured and estimated signal strengths. Authors in [92] solved this optimization problem in a non-least squared sense, using prior knowledge of the bounds for the parameters to be optimized.

After solving the optimization problem by a suitable algorithm, the optimized parameters are applied in the calculation of signal levels at unknown location to augment the existing data.

Note that L_d and L_c require knowledge of the propagation environment, such as access to clutter database of a mobile operator or knowledge of the digital elevation model [92]. Also, antenna parameters knowledge through antenna datasheets or antenna diagrams is required in this method.

3.3.3 *Transfer learning*

The interpolation methods in Section II are expected to work well when the sparse data has at least some latent features. Same is true for methods such as GANs (Section IV-A) that are used to enrich sparse data by generating synthetic data. However, for situations where this is not the case, alternatives have to be used. For such data streams where latent features are too little to allow use of GANs, matrix completion or other interpolation techniques identified above, we propose to leverage the transfer-learning paradigm [93].

The key idea here is to utilize similarities among the cells as well as users to address the data scarcity challenge. This method exploits the likely scenarios in networks where some cells have more data (i.e., configuration experiments) than others; e.g., the network operator dedicates a few cells in the network for the experiments for a period of time. Therefore, we propose to leverage the data from those data rich cells for other cells that have limited or no data. Specifically suppose in cell i , we assume that KPIs are generated by a function $f(.)$ of COPs i.e., $[KPI]_{1 \times K} = f_i([COP]_{1 \times C}) + \epsilon$ where ϵ is random noise

and f has marginal distribution $P(f_i)$. Based on historical dataset, this function $f(\cdot)$ can be learned as proposed in [3] due to which we can have all information for the COP-KPI vector $X_i = [COP_1, COP_2, \dots, COP_C, KPI_1, KPI_2, \dots, KPI_K]_i$.

Given that we have no data or limited amount of data in cell j , we can leverage similarities among cells. We propose to transfer knowledge from cell i (source cell) to cell j (target cell) by defining transferability as: $f_i([COP]_{1 \times C}) = f_j([COP]_{1 \times C})$. Using deep transfer learning, knowledge learned from a BS (or cell) can be used to generate data in areas, where little or no data is available. However, the challenging part is to find out that knowledge learned from which cell could be transferred to the target cell. For instance, we can leverage the similarities among cells to find suitable transfer candidates. To quantify similarities among the cells, one approach is to use Wasserstein distance measure [94]. Given two random variables f_i and f_j with marginal distributions $P(f_i)$ and $P(f_j)$ respectively, let ψ denote the set of all possible joint distributions that has marginals of $P(f_i)$ and $P(f_j)$. Then Wasserstein distance between them is defined as:

$$W(f_i, f_j) = \inf_{P_{f_i f_j} \in \psi} \int |f_i - f_j| P_{f_i f_j}(f_i, f_j) d_{f_i} d_{f_j} \quad (3.37)$$

The inf in (3.37) gives joint distribution with f_i and f_j having smallest distance while maintaining the marginals.

As a case study, a commercial planning tool is used to simulate a sophisticated cellular network topology. To capture the similarities among the cells, clustering is done on the clutter (land cover) information available of simulation area, such as percentage of high rise buildings, water, vegetation etc. in each BS (or cell), as it captures the latent environment features which are independent of the BS configuration parameters. Then, we trained a Deep Neural Network from the data rich BS (source cell) to model the behavior of average RSRP across all its serving UEs with respect to different combinations of the transmit power, antenna height and antenna tilt angle. This trained model can be reused in a similar target cell from the same cluster group which has sparse data. This

is where transfer learning jumps in and can be used to fine-tune the source model using the sparse available data in target cell.

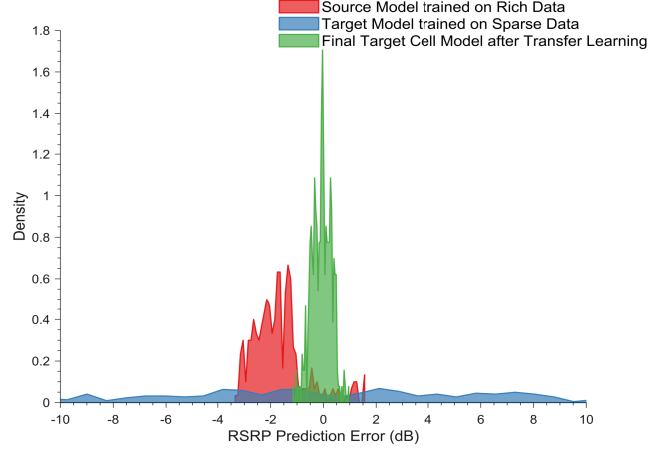


Fig. 3.5: Comparison of Prediction Error after using Transfer Learning on the Source Model.

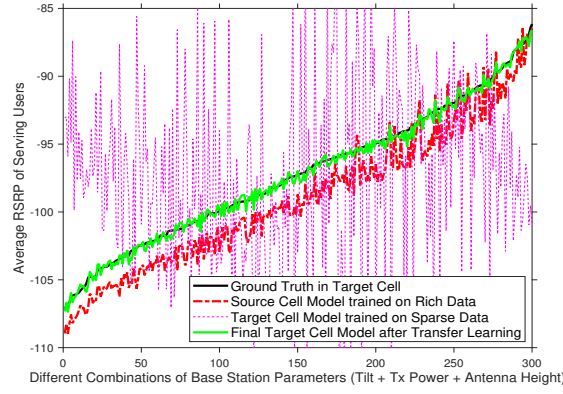


Fig. 3.6: Comparison of Prediction Performance with Ground Truth after using Transfer Learning on the Source Model.

A comparison of the RSRP prediction error after using transfer learning on the source model and a comparison of prediction performance with ground truth after using transfer Learning are shown in Fig. 3.5 and Fig. 3.6. It can be observed from these figures that the final target model using transfer learning depicts an excellent fit to the ground truth in target cell. Our results show that, by using transfer learning the prediction RMSE gets reduced by 80% as compared to only the source model without using transfer learning in the target cell.

3.4 Synthetic data generation

The techniques mentioned in previous sections are likely to work well when the sparse available data is somewhat representative of the whole data or exhibits some degree of correlation. In situations where the available data is sparse and non-representative, the methods presented in preceding sections are likely to perform poorly. Likewise, in other scenarios, the available data can be big, but still not representative. In these cases, the solution lies in either resorting to get real data or generate synthetic data. In this section, we will present two ways to generate synthetic data: through generative adversarial network, which is likely to work well when the sparse data has some latent features and through simulators.

3.4.1 *Through generative adversarial network*

We propose to build on the machine learning concept of Generative Adversarial Networks (GANs) [95] to enrich sparse training data in order to cope with data scarcity issue. The basic idea we propose is to use GAN to generate large amount of synthetic data building on small amounts of real data which will not be distinguishable from real data. The intuition behind GANs is to exploit the potential of deep neural networks (DNNs) to both model nonlinear complex relationships (the generator) as well as classify complex signals (the discriminator).

GAN's success in image processing has been well established [96]-[100]. Although this concept has widely been used in image processing, it can also be used in wireless communications. Here, we propose it to address the data sparsity problem as demonstrated in our work [2]. This paper [2] is the first to use GAN as a tool for synthetic data generation for addressing the data sparsity challenge. In GAN, we begin by setting up a two-player minimax game between the discriminator DNN and generator DNN. In each training epoch, the generator iterates its weights to produce synthetic data trying to fool

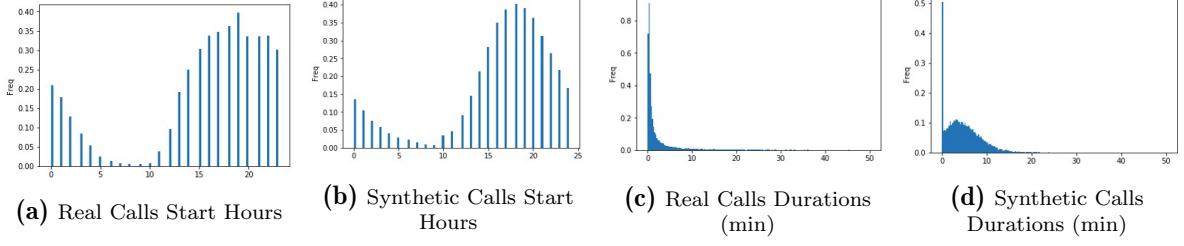


Fig. 3.7: Leveraging GAN for enriching the sparse training data [2]

the discriminator DNN. The discriminator DNN on the other hand, tries to discriminate between real data and generated data i.e., actual COP-KPI or user behavior dataset versus generated synthetic data set. In theory, when Nash equilibrium is reached between the generator DNN and discriminator DNN, the pair of DNNs will provide us a generator that can exactly duplicate or reproduce the distribution of the real data so that the discriminator would be unable to identify whether a sample is synthetic i.e., whether it is generated by the generator DNN or it is from the real data [130]. At this point, the synthetic data generated by the generator DNN are indistinguishable from the real data, and are thus as realistic as possible.

To illustrate this approach further, consider a set of real data for a group of COP-KPI at N sites consisting of C COPs per sites and K network wide KPIs. Let X be the matrix of real data $[X]_{T \times (C \times N + K)} = [COP_{C,N}, KPI_K]_{T \times (C \times N + K)}$ where T is size of historical time window i.e., number of observations. Our objective is to train a generator DNN by utilizing the real data X as the training set. Generated COP-KPI data should be capable of manifesting the same stochastic properties as real data. It should also exhibit the variety of different modes manifesting all possible patterns observed during training. Assume observations x^t for times $t \in T$ are available. Let the actual distribution of the real data be denoted by $\mathbb{P}_X$, which is of course not known at this stage and hard to model.

To begin with, we use a group of noise vector as input z . This input has a known distribution $Z \sim \mathbb{P}_Z$ that is easily drawn from (e.g., jointly Gaussian). The objective here is to transform a sample z sampled from $\mathbb{P}_Z$ such that it follows $\mathbb{P}_X$ (without ever learning $\mathbb{P}_X$ explicitly). We plan to accomplish this by simultaneously training two DNNs:

the generator DNN and the discriminator DNN. Let G denote the generator function parametrized by $\Theta^{(G)}$. We write it as $G(.; \Theta^{(G)})$; Let D represent the discriminator function parameterized by $\Theta^{(D)}$. We write as $D(.; \Theta^{(D)})$. Here, $\Theta^{(G)}$ and $\Theta^{(D)}$ are the weights of two DNNs, respectively.

Generator: During the training process, the generator is trained to intake a set of inputs and by applying a series of up-sampling operations through different function neurons to output realistic data. Let Z be a random variable with distribution $\mathbb{P}_Z$. Then $G(Z; \Theta^{(G)})$ is a new random variable, whose distribution we represent as $\mathbb{P}_Z$.

Discriminator: The discriminator is trained in parallel with the generator. It receives input data either coming from the generator or from real data.

By applying a series of operations of down-sampling using another DNN, it produces a continuous value p_{real} that quantifies to what degree the input samples belong to $\mathbb{P}_X$. The discriminator can be represented as $p_{real} = D(x; \Theta^{(D)})$ where x may come from $\mathbb{P}_{data}$ or $\mathbb{P}_Z$. The discriminator is trained to learn to differentiate between $\mathbb{P}_X$ from $\mathbb{P}_G$. The discriminate thus tries to maximize the difference between $\mathbb{E}[D(G(Z))]$ (generated data) and $\mathbb{E}[D(X)]$ (real data). The Loss functions L_D for discriminator becomes $L_D = -\mathbb{E}_X[D(X)] + \mathbb{E}_Z[D(G(Z))]$ and L_G for generator becomes $L_G = -\mathbb{E}_Z[D(G(Z))]$.

As a large p_{real} means the sample is more realistic, the generator tries to minimize the expectation of $-D(G(.))$ by adjusting G (for a given D), resulting in L_G . On the other hand, for a given G , the discriminator aims to minimize the expectation of $D(G(.))$, and simultaneously increase the real data score. This yields the loss function L_D . Here the functions G and D are parameterized by the weights of the DNN. Finally a minimax two-player game is set up between G and D :

$$\min_{\Theta^{(G)}} \max_{\Theta^{(D)}} V(G, D) = \mathbb{E}_X[D(X)] - \mathbb{E}_Z[D(G(Z))] \quad (3.38)$$

where the value function $V(G, D)$ is negative of L_D .

In the early stages of the training, G just produces COP-KPI data samples $G(z)$ with little match with samples in $\mathbb{P}_X$. The discriminator can reject these data samples with high confidence. Thus, in these early stages of training, L_D is small, and $L_G, V(G, D)$ are both relatively large. As the training progresses, the generator gradually learns to produce data that could let discriminator have high confidence to be true. At the same time, the discriminator is also trained to distinguish these newly fed synthetic data generated by the generator. As training advances and reaches near to the optimal solution, G is able to generate data that mimic real data with a small L_G value. At this stage, the discriminator is not able to differentiate $G(z)$ from $\mathbb{P}_X$ with a large L_D .

To assess the efficacy of approach outlined above, as a preliminary study recently published in [2], we leveraged GAN to generate synthetic call data records (CDRs) data and thus increased training dataset size by enriching the real sparse CDR with realistic synthetic data. CDRs data are selected as preliminary case study because CDR data can be used by a large number of SON solutions such as in [101], [102]. Real network traces with call durations and call start time stamps, provided by one of the leading mobile operators in USA, were used in this study to train the GAN. The discriminator was trained beginning with 20,000 data points (from a record of several hundred thousand). Once the discriminator could reliably differentiate between the real data taken from the record and randomly generated CDR data with two features i.e., call duration and start time, the generator was trained. After the generator was generating data that the discriminator perceived to be real, we used the trained generator to produce another 20,000 CDR data samples. Figs. 3.7a and 3.7c and represent the distribution of the real data used to train the discriminator. Figs. 3.7b and 3.7d show the distribution of the 20,000 synthetic data points produced by the trained generator. These preliminary results show the high similarity between real and synthetic data produced by the proposed GAN based approach.

This synthetic data combined with the sparse real data can then be used for training the

models instead of training the model using sparse real data only. Our preliminary results recently reported in [2] show that this approach can significantly improve the accuracy of the models being trained compared to models trained using sparse real data only. This study thus establishes the feasibility of proposed approach to address training data sparsity/scarcity challenge.

3.4.2 Through custom-built system level simulators

Table 3.2: Comparison of different simulators for solving data sparsity problem

Simulator	Platform	Adaptive Numerology	HO Support	Realistic Mobility	QCI Support	Scheduling Support	mmWave Support	Cloud Based	Free License	Short Description
SyntheticNet [103]	Python	✓	✓	✓	✓	✓	✓	✓*	✓	- Link and System level - Conforms to 3GPP-5G standard - Support of database-aided edge computing
MATLAB/Simulink [104]	MATLAB	✓			✓	✓				- Link level simulator
ns-3 [105]	C++					✓	✓		✓	- Link level simulator - Event Driven: a. Not suitable for large networks b. Takes long time
OMNeT++ [106]	C++		✓			✓	✓		✓	- Link level simulator - Event Driven: a. Not suitable for large networks b. Takes long time
OPNET [107]	C/C++					✓			✓	- Link level simulator - Event Driven: a. Not suitable for large networks b. Takes long time
OpenAir Interface [108]	C++					✓			✓	- Network level - Support of core network: a. Increased complexity b. Limits nodes
5G-K [109]	C++					✓		✓	✓	- Link level - System level - Network level
Vienna-5G [110]	MATLAB	✓				✓	✓		✓	Improved version of Vienna-4G
NYUSIM [111]	MATLAB					✓	✓		✓	mmWave heterogeneous networks
C-RAN [112]	MATLAB					✓	✓	✓	✓	5G cloud-based networks
GTEC [113]	-					✓			✓	- Link level simulator
X. Wang et al. [114]	MATLAB					✓			✓	- Link and System level
V. V. Diaz et al. [115]	N/A						✓		✓	Pathloss model for: a. Rural/Urban b. Macro/Small Cell c. LoS/NLoS
Ke Guan et al. [116]	N/A								✓	Ray tracing with use case of V2V communication

* SyntheticNet built by AI4Networks Research Center simulator supports database-aided edge computing

System level simulators are widely used in both industry and academia due to limitations

of analytical models and field experiments. Analytical models, in the pursuit of ensuring tractability, are often over-simplified. Apart from the limitation of mounting Base Stations (BSs) on predefined locations, the support of antenna height, tilt, transmission power etc. for individual BSs is absent in the analytical model. Furthermore, stochastic geometry-based models are unable to capture the network dynamics which include mobility management and transmission latency. On the other hand, field trials exhibit the most realistic modeling of network performance, evaluation and tuning. However, this approach is impractical owing to the cost and time effort required to conduct field trials on a large scale, and with the high probability of significant network performance impairment of live mobile network during the trial phase.

A list of existing simulators along with a comparison of their features is presented in Table 3.2. As observed from Table 3.2, none of the simulators is based on comprehensive 5G standard incorporating all aspects outlined in the standard. To tackle this problem, we have developed an AI4Networks simulator built on Python platform. The AI4Networks simulator is modular, flexible, microscopic and versatile, built in compliance with the 3GPP Release 15. This simulator supports features like adaptive numerology, actual hand over (HO) criteria and futuristic database-aided edge computing to name a few. Instead of an objected-oriented programming (OOP) based structure like existing simulators, AI4Networks simulator supports commonly used database files (like SQL, Microsoft Access, Microsoft Excel). Site info, user info, configuration parameters, antenna pattern etc. can be directly imported to the simulator. As a result, the simulation environment is more realistic and closer to actual deployment scenarios. For further details of this simulator, the reader is referred to [103].

Python based platform and the flexibility of different input and output data formats in AI4Network simulator can assist in solving the data scarcity challenge by generating ample amounts of synthetic data to enrich the available sparse real data, which can then be used to implement different Self Organizing Networks (SON) related features

or AI based network solutions [3]. Mobile operators can use it for planning, evaluating or even optimization of 5G networks. Research community can also benefit from it by implementing the new ideas on data generated from this 3GPP-based realistic 5G network simulator.

3.5 Real data generation

The above techniques, with the exception of using simulators, are likely to work well when the sparse available data is somewhat representative of the whole data or exhibits some degree of correlation. In situations where the available data is sparse or big but non-representative, the solution lies in obtaining get real data.

One way of getting access to real data can be utilizing historic logs of data gathered by other researchers. However, these logs might become outdated quickly with the emergence of new technologies, heterogeneous deployments or change in traffic patterns, number of users, construction of buildings and other terrain changes. Another way of generating real data can be through the use of mobile phone applications. However, what if researchers require data for scenarios which are not yet deployed in a real network? The techniques presented in previous sections (except simulators), all require some starting real data but with the advent of AI based next generation networks, there exists the potential of new or anticipated scenarios which do not exist in a real network. In such cases, testbeds to generate real data are going to be the best option for wireless communications community.

3.5.1 Through smartphone applications

Many smartphone applications offer the ability to log parameters such as RSRP, RSRQ, SNR, events occurring (handover, cell re-selection), serving time, speed, height, cell ID, along with timestamp and location (latitude, longitude) information). As an example, one of our studies [102], used a novel methodology of utilizing smartphone application,

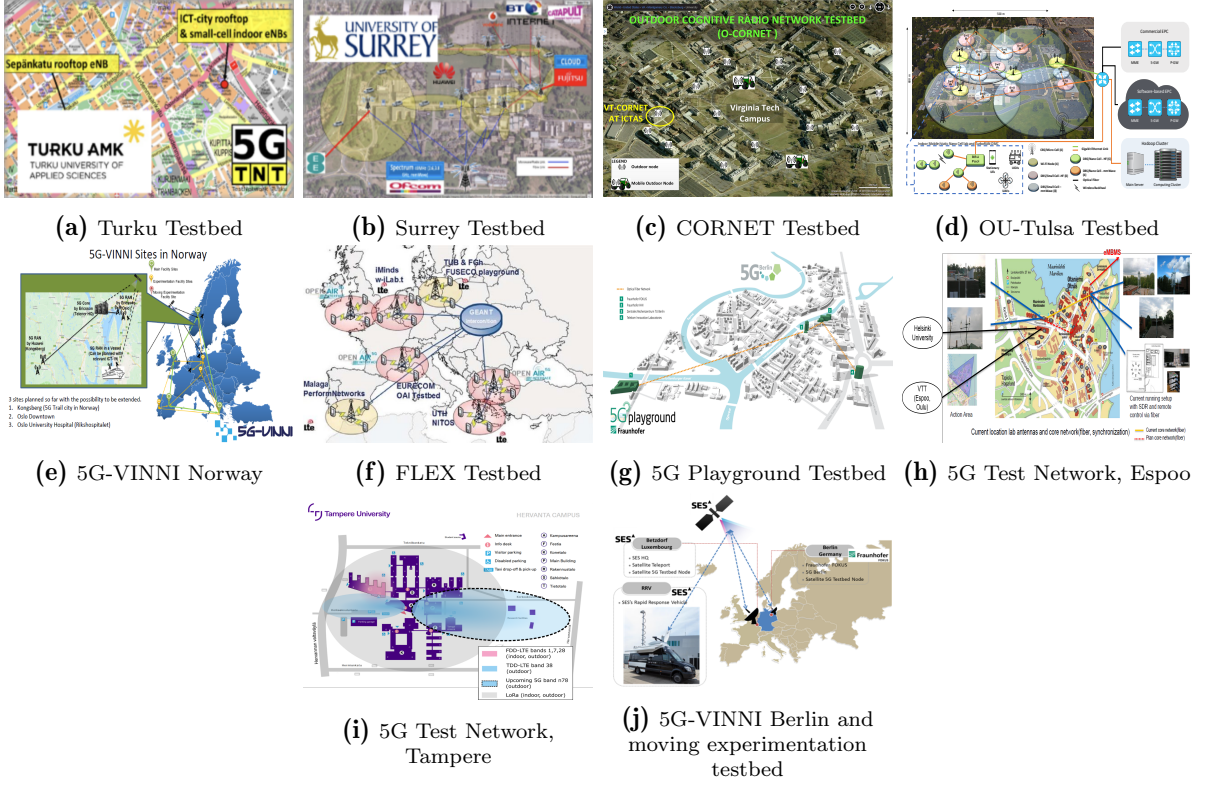


Fig. 3.8: Some current and emerging 5G testbeds.

based on the idea of participatory sensing, to collect real LTE network data for building, training and evaluating the performance of mobility prediction schemes in live network [102]. An android application, “LTE Discovery” was installed on the smartphone to log the timestamp and new cell IDs around the OU-Tulsa campus. This information was then used to build a semi-markov model for mobility prediction.

The quality of data gathered through smartphone applications, however, depends on a number of factors, including measurement capabilities of different smartphones and GPS error inaccuracy for measuring heights and positions. Smartphones equipped with barometers are likely to give a better estimate of heights in scenarios with varying terrains. In addition, transmitter parameters, such as type of antennas and their characteristics remain unknown, unless the network operator is involved. For this reason and for potential new scenarios, the solution is resorting to testbeds.

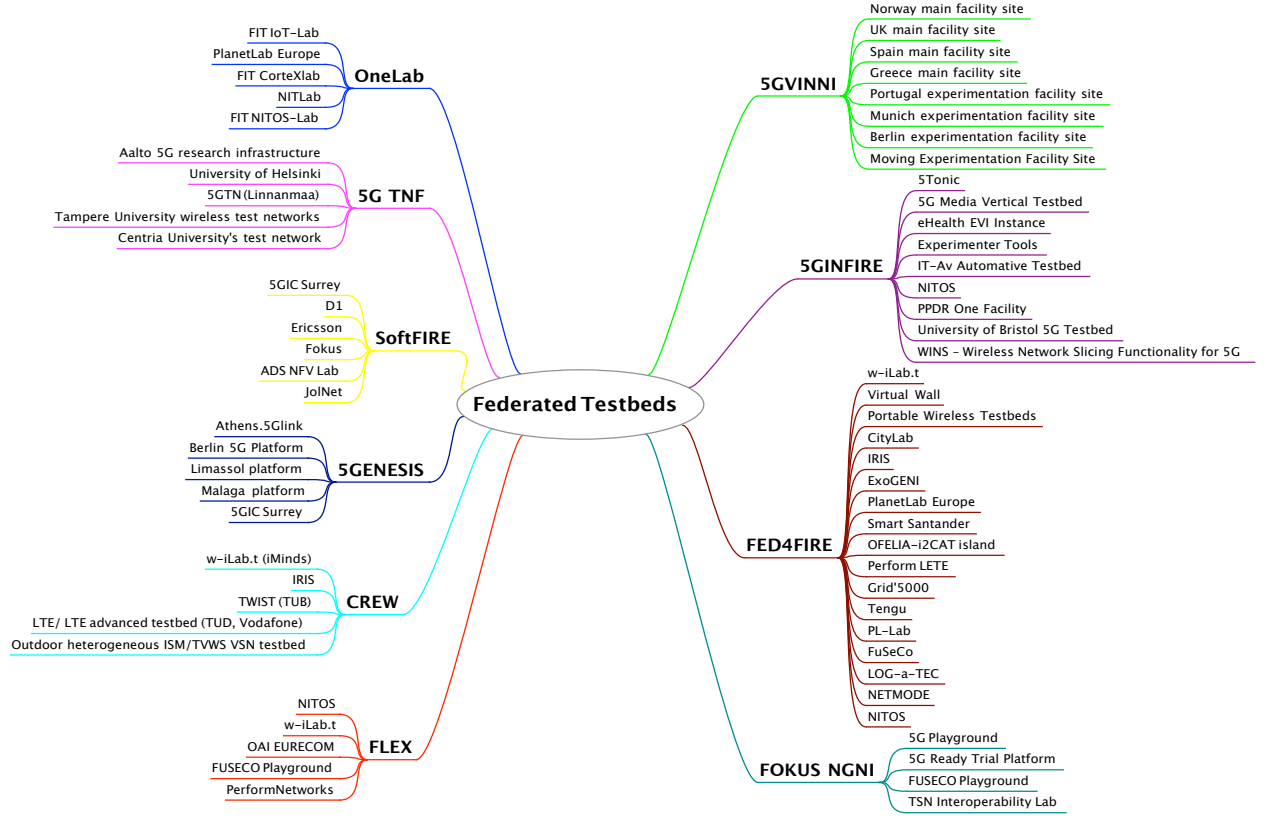


Fig. 3.9: Federated Testbeds.

3.5.2 Through testbeds

Field trials using testbeds generate real training data and provide the most realistic picture of the network. An aerial view of some of these testbeds is presented in Fig. 3.8. However, conducting field trials on a large scale is time-consuming and expensive. For this reason, we have summarized the existing and emerging testbeds in Table 3.3 to make readers aware of current and emerging platforms to access real data in order to overcome data sparsity challenge. Most of these testbeds are open, i.e., available to external experiments. This will foster collaboration among different academic institutions as well as with industry, which will in turn enable the utilization of these existing facilities to the fullest and accelerate quality research in the field.

Apart from individual testbeds, several federations or consortiums of testbeds have been formed around the world. Some key federated testbeds comprising of the testbeds in Table 3.3 are presented in Fig. 3.9.

Table 3.3: Worldwide existing and emerging testbeds for solving data sparsity problem

Testbed	Location	Key Features
NITOS [117] [118]	NITLab, University of Thessaly (UTH), Volos, Greece	- Open (facilities available to external experimenters) - Over 100 wireless indoor and outdoor nodes - 45 nodes equipped with a mixture of Wi-Fi and GNU-radios - One Cloud installation with 200-cores - Multiple wireless sensor network deployments - Cameras, temperature and humidity sensors - Software defined radio testbed with 10 USRP devices - Two programmable robots provide mobility - WiMAX/3G/LTE technologies - 5G virtual infrastructure provisioning by 5GINFIRE [119]
5GIC [120] - [123]	ICS, University of Surrey, Guildford, UK	- 4G LTE and 5G NR - 4km² comprising indoor and outdoor environments - Outdoor: 4G ultra-dense C-RAN comprising 3 macro cells, 39 LTE-A TDD small-cell sites, operating at 2.6 GHz, 1x 4G FDD site operating at 700 MHz, 8x 5G NR TDD sites, operating at 3.5 GHz - Indoor: 6x TDD and 6x FDD cells over 2 floors, and Wi-Fi APs - 28 GHz (PtP), 60GHz (PtMP) mmWave and satellite backhauling also supported - Core Network supports separate 4G and 5G core segments - Supports broadband mobile radio - Fixed core network and service platform based on software defined networking - Supports Internet of Things

ORBIT [124] [125] [126]	WINLAB, Rutgers University, USA	- Open: available for remote or on-site access - Radio grid with 20x20 two-dimensional grid of programmable radio nodes - Outdoor ORBIT network provides a configurable mix of both high-speed cellular (WiMAX, LTE) and 802.11 wireless access - SANDBOX networks used for debugging and controlled experimentation - Software defined networking (SDN) resources - Cloud resources
PhantomNet [127] [128]	Flux Group, University of Utah, USA	- Remotely accessible and sharable - Mobility testbed - Built on top of Emulab - EPC/EPS software (OpenEPC), hardware access points (ip.access eNodeB), PC nodes with mobile radios (Nexus 5 phones and SDR-based) - Provides configuration directives and scripts
LuMaMi [129] [130] [131]	Lund University, Sweden	- Real time 128-antenna MIMO test bed - National Instruments USRP RIO SDRs - LabVIEW system design software and PXI platforms - Mobile base stations - Used for channel sounding, high speed data streaming, evaluation of baseband solutions, assessing circuit design - Demonstrated mobile multi-user tests with University of Bristol [132]
Firecycle [133] [134]	Intrusion Detection Systems Group, Columbia University, USA	- Scalable test bed for large-scale LTE security research - Implement, test, analyze impact of security attacks against LTE mobility network - Prototyping and testing attack mitigation strategies for future cellular networks - Implemented on OPNET

Berlin LTE-A [135] [136] [137]	Center of Berlin, operated from Fraunhofer HHI, Deutsche Telekom, University of Technology, Berlin	- 3 base station sites with 9 sectors - Incorporates LTE key features: frequency dependent scheduling in 20 MHz bandwidth, adaptive MIMO mode selection for 2x2 MIMO utilizing spatial multiplexing, and low round-trip delay on the PHY layer of 8 m
CEWiT LTE and 5G NR [138]	IIT Madras Research Park, Chennai, India	- 2 types of testbeds based on: 1) CEWiT hardware 2) TI's multi-core DSPs - Hardware is made using SDR radio nodes - LTE PHY for UE and eNB has been developed in collaboration with IITM - Basic implementation of LTE L1 downlink and uplink chains - L2 MAC, RLC and a thin layer of PDCP - Both eNodeB and UE implementations - End-to-end IP application flow both in DL, UL - Supports 3GPP Release 8 specifications - Supports up to 20MHz bandwidth - 5G NR for sub 6GHz and mm wave under development
TitanMIMO-6 [139] [140]	Nutaq, Québec, Canada	- Sub 6 GHz wideband Massive MIMO testbed - FDD+TDD capabilities - Up to 56 MHz real-time baseband processing - Radio tumble up to 5 GHz - Nutaq's SDR systems (PicoSDR) can be combined with TitanMIMO system to build up complete HetNet, MUMIMO or CRAN testbed solutions - Enabling evaluation of interoperability behavior for various deployment scenarios

Aalto 5G research infrastructure [141]	Otaniemi, Espoo, Finland	- Network slicing - Support for NB-IOT to be used for IoT hackathon - Mobile, edge computing, VR/AR, Gaming, Industrial Internet - Part of 5G TNF
University of Helsinki Test Network [142]	University of Helsinki, Kumpula campus, (Exactum building), Finland	- 17 Nokia Flexi Zone Indoor Pico BTS (eNBs) - Band: 2600 MHz (E-UTRA 7) FDD - Sync: 1588v2 (PTP) / GPS / Sync-E - 3 connections to cores through VLANs: UH core(s), Aalto core and Nokia core - Part of 5G TNF
VodaPhone Chair [143] [144] [145]	TU Dresden, Germany	- Online Wireless Lab (OWL) testbed - Software Defined Reconfigurable Radio Devices - LabVIEW/LVC in combination with USRPs - Many projects and startups, e.g., 5G Lab Germany, 5GNetMobil, 5G Picture, HPE-5G-Testbed, Airrays GmbH [146]
CORNET [147] [148]	Virginia Tech University, USA	- University-wide testbed - Software-defined, cognitive radios, dynamic spectrum access - 48 indoor SDR nodes, 14 fixed outdoor nodes, 6 mobile units (O-CORNET) - A few LTE-capable nodes (LTE-CORNET) - CORNET nodes are remotely accessible - Awarded grant from DURIP for upgrading to LTE-A - Outdoor network of 15 radio nodes and 2 mobile nodes

5G Playground [149]	Fraunhofer FOKUS and TU Berlin campus, Germany	- Empowers the 5G Berlin testbed - Support for multi-slicing - Ultra reliable, low latency communication in Industria IoT lab of FOKUS - Automotive testbed environment in underground parking of FOKUS building - Coverage of dense urban areas, portable edge nodes in progress - 3 Toolkits: Open5GCore, OpenSDNCore and Open5GMTC
Tampere University Wireless Test Networks [150] [151]	Tampere University, Hervanta, Finland	- Part of 5G TNF - FDD-LTE operating at band 1, 7, and 28 for mostly indoor coverage - TDD-LTE operating at band 38 to provide campus wide outdoor test network - Upcoming outdoor 5G test network in band n78 with 60 MHz channel - LoRa: Digita's LoRaWAN test network in ISM band at 868 MHz
FUSECO Playground [152]	Fraunhofer FOKUS Institute, Berlin, Germany	- Open IMS Core solution - Heterogeneous indoor and outdoor radio access technologies - DSL/WLAN/2G/3G/4G-LTE/LTE-A and soon 5G - M2M communication, IoT, sensor networks - SDN/OpenFlow, NFV cloud environments - Toolkits: Open5GCore, OpenSDNCore and Open5GMTC - OpenMTC, Open Source IMS Core, OpenStack-based Cloud Testbed, OpenXSP

5G Ready Trial Platform [153]	Fraunhofer FOKUS, Berlin, Germany	- Consolidated turn-key solution of the Fraunhofer FOKUS software components - Addresses trial needs of emerging network infrastructures - Edge Instantiation: solution for micro-operators and local networks, provides customized IoT connectivity for x100 devices. - Data Center Instantiation: multi-slice environment, support for multiple parallel instances of IoT and multimedia communication - Technology Elements: Virtual Core network, Network slicing, IoT support, Low delay network, Dynamic spectrum access
Ericsson 5G [154] [155]	Ericsson, Stockholm, Sweden	- Live testing of key capabilities, such as multipoint connectivity with distributed MIMO and 5G-LTE dual connectivity - 5G devices and base stations operate in 15 GHz band - TDD and OFDM - Up to 256 QAM modulation in downlink and up to 64 QAM in the uplink - mm-Wave testbeds 15 GHz and 28 GHz - Bandwidth is 80 MHz, centered at 3.5 GHz - Massive MIMO antenna array of 128 cross-polarized antennas
SK Telecom 5G Playground [156] [157] [158]	SK Telecom R&D Center, Bundang, Korea	- Developing a centimeter-wave 5G radio system with Nokia - 5G 3D system level simulator with Nokia and Ericsson - 3D beamforming techniques with large scale array antennas with Samsung - Developing Anchor-Booster Cell and Massive MIMO with C-RAN with Intel - Achieved 19.1Gbps transmission speed over the air - Futuristic services including 4K live broadcast system, AR/VR

5GTN (Linnanmaa) [159] [160] [161]	University of Oulu and VTT Technical Research Centre of Finland	- Multi-access edge computing - Core network in cloud environment - Cloud systems for applications - Secure connection to other 5G sites worldwide, 10 Gb VPN - Part of 5G TNF
TurboRAN [162]	AI4Networks Research Center, University of Oklahoma, Tulsa, USA	- Developing first end to end programmable cellular test bed for enabling AI based SON research towards 5G and beyond - Complete integrated mobile cellular network over 300,000 m² - Tier 1: 4 outdoors macro cells on 1.2-6 GHz HF band - Tier 2: 16 small cells (programmed to pico or femto cells). 8 small cells can operate on the HF band, other 8 can operate on unlicensed mmWave - Both tier cells are programmable - Both tier cells connected to EPCs and a big data processing Hadoop cluster - Hadoop cluster: 1 high performance master node, 15 slave nodes with high capacity data modems - Support both high mobility and low mobility users
OAI [163]-[166]	EURECOM, France	- Open-source platform - 8-node testbed, equipped OAI compatible RF front-ends - 4 machines that can be used for running OAI as eNodeB - 4 nodes that are equipped with COTS UEs - 2 physical layer emulation modes - 64 antenna Massive MIMO testbed

Munich [167] [168]	TU Munich, Munchen, Germany	- 5G RAN with 2 sectors, carrier frequency: 3.4 GHz, bandwidth: 40 MHz, transmission power: 5 W antennas - 5G Mobile Terminals with vehicular speeds up to 50 km/h, V2X - 5G Core network: HW/SW platform - Hardware: in-house platform of several dozen servers representing a data centre - Software: extended network emulators, controllers, open-source and proprietary switch implementations - Testbed can deploy virtual networks with different topologies - 5G Core network supporting functional split – SDN – NFV Orchestration - Distributed data centres for mobile edge computing use cases
Perform Networks [169] [170] [171]	University of Malaga, Spain	- T2010 conformance testing units by Keysight Technologies - LTE release 8 small cells (Pixies) by Athena Wireless working on band 7 - Polaris Core Network Emulator - Several LTE UEs, working on different bands - ExpressMIMO2 and USRP SDR cards - SIM cards from an Spanish LTE operator to be used on commercial deployments
Centria's Test Network [172]	Centria University of Applied Sciences, Ylivieska, Finland	- TDD-LTE operating at band 40 and 42 for both outdoor and indoor coverage - Upcoming 5G test network in band n78 with 60 MHz channel outdoor network - Implementation plan of first 5G Non-Standalone during 2019 - Later 5G Standalone during 2020 - Part of 5G TNF

w-iLab.t [173] [174] [175]	Ghent and Zwijnaarde, Belgium	- w-iLab.t Office testbed: three 90 m x 18 m floors of iMinds office in Ghent - w-iLab.t Zwijnaarde testbed: 5 km away from w-iLab.t Office in Zwijnaarde - Sensor nodes, Wi-Fi based nodes, sensing platforms, and cognitive radio - Heterogeneous wireless/wired experiments - Virtual Walls: Virtual Wall 1 and 2 containing 206 and 159 nodes respectively - OpenFlow experiments - 20 programmable moving robots
5TONIC [176]	Madrid, Spain	- 9 members: Telefonica, Institute IMDEA Networks, Ericsson Intel, Commscope, Universidad Carlos III de Madrid, Cohere Technologies, Artesyn Embedded Technologies and InterDigital - NFV orchestrator, implemented with Open Source MANO - Dedicated NFVI for 5GINFIRE: 3 server computers, each with 6 cores, 32GB of memory, 2TB NLSAS, network card with 4 GbE ports, DPDK support - Second NFVI: 2 high-profile servers, each equipped with 8 cores in a NUMA architecture, 128GB RDIMM RAM, 4TB SAS and eight 10Gbps Ethernet optical transceivers with SR-IOV capabilities

University of Bristol 5G [177]	University of Bristol, England	- Multi-site network connected through a 10 km fibre - Core network is located at HPN Lab at University of Bristol - Extra edge computing node is available at Watershed - Access technologies are located at Millennium Square (outdoor coverage), “We The Curious” science museum (indoor coverage) - Multi-vendor SDN enabled packet switched network - SDN enabled optical (Fibre) switched network - Nokia 4G and 5G NR - Self-organising multipoint-to-multipoint wireless mesh network - LiFi Access point, Cloud and NFV hosting - 2 different NFV orchestration and management solutions: Open Source MANO , NOKIA CloudBand - 2 cloud/edge computing solutions:Openstack Pike, Nokia MEC - 1 SDN controller: NetOS
--------------------------------------	-----------------------------------	--

3.6 Conclusion

In this chapter, we have presented an overview of key techniques in literature and introduced some emerging new techniques that can be applied to wireless communication domain. It should be noted however, that the success of any technique for solving the data sparsity challenge depends on a number of factors, including type of data under consideration, number of transmitter and receivers, distributions of users and base stations in a given area, distribution of measurement data, level of accuracy required, measurement capability of receivers, dynamics of propagation environment, propagation modeling accuracy, time and computational resources available. Also, highly dynamic spatio-temporal environment would greatly hamper the outputs of techniques covered in this chapter. In that case, using data through simulations and testbeds may provide the best option. Further options on addressing the data sparsity challenge for highly dynamic environments is

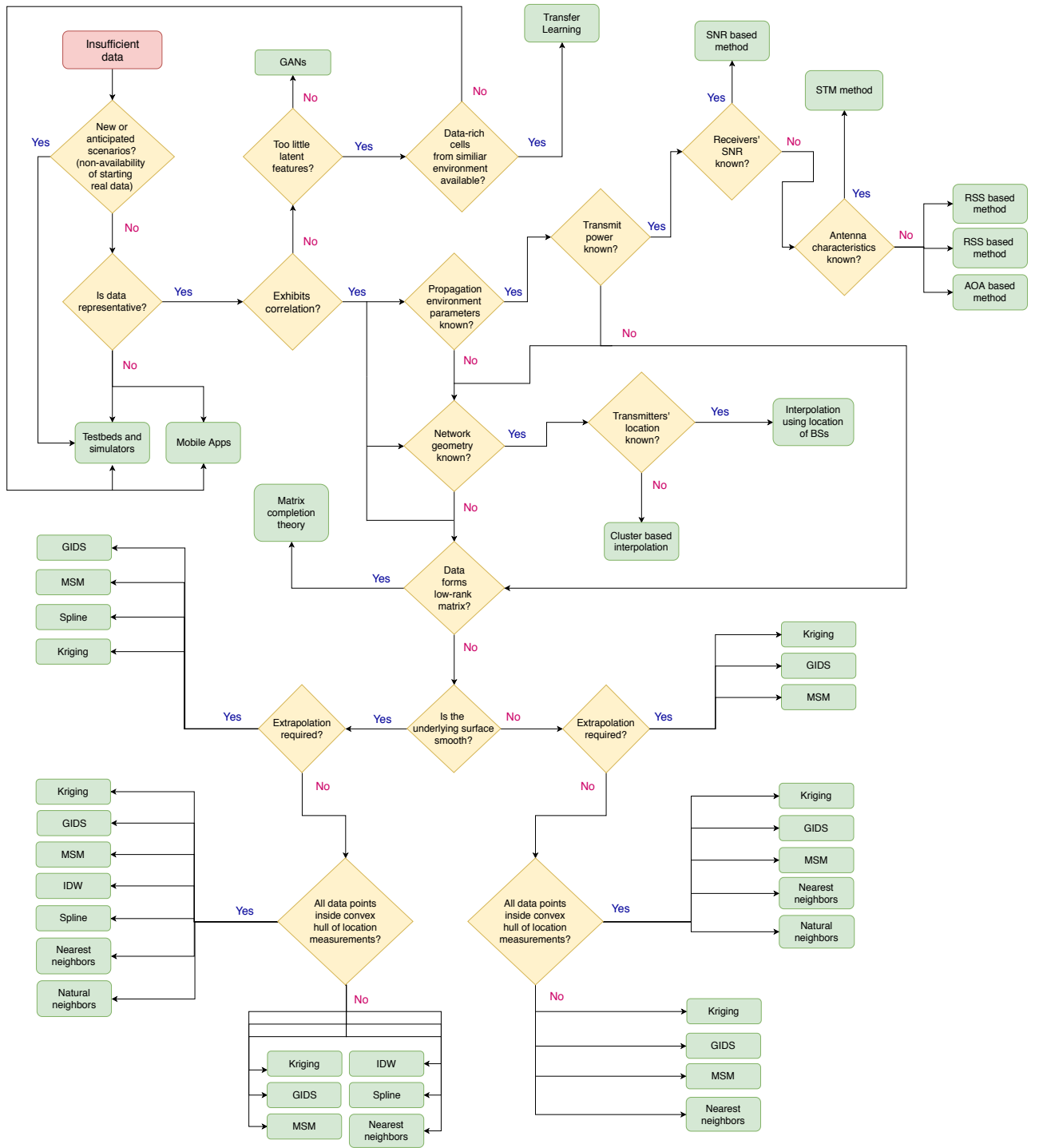


Fig. 3.10: Framework to address data sparsity challenge.

out of the scope of this work and can be considered as part of a future study. Therefore, while a certain technique might work well in a particular scenario, it is likely to perform poorly in other scenarios. It should also be noted that the selection of a performance metric to assess the accuracy of a particular method is important too. As an example, if the metric of mean residual error is used to assess Kriging accuracy, it would always yield zero, since this type of interpolant satisfies the unbiasedness condition, and so some other performance metric, like the average relative error would be more appropriate in this case.

Finally, in order to assess the applicability of a particular method, the tree diagram in Fig. 3.10 is aimed to assist the reader in choosing the possible techniques out of those presented in this chapter, based on available information.

CHAPTER 4

Conclusions

Future cellular networks are envisioned to have big data enabled network automation capabilities, which are envisioned to be enabled by AI. However, traditional ML approaches to enable AI-based automation suffer from various challenges, including data sparsity challenge. This report addresses this challenge. An overview of key techniques in literature is presented and some new techniques are introduced, such as matrix completion theory, leveraging different types of network geometries, transfer learning and the use of generative adversarial networks. Case studies are also demonstrated that show the success of these new techniques in enriching sparse data in cellular networks. In addition, a framework is provided which can assist the reader in choosing the best possible techniques, based on available information. Combining the data sparsity error with other errors in coverage estimation problem, such as quantization and positioning error uncertainty, some practical applications are addressed in a MDT-based case study, which include determining optimal bin widths that minimize the effect of all these errors concurrently and improving the utility of the MDT estimated coverage by its calibration. The findings from this report can enhance AI-based cellular networks and thus aid in their design, operation and optimization.

Bibliography

- [1] G. Y. Lu and D. W. Wong, “An adaptive inverse-distance weighting spatial interpolation technique,” *Computers & geosciences*, vol. 34, no. 9, pp. 1044–1055, 2008.
- [2] B. Hughes, S. Bothe, H. Farooq, and A. Imran, “Generative adversarial learning for machine learning empowered self organizing 5G networks,” in *2019 International Conference on Computing, Networking and Communications (ICNC)*, 2019, pp. 282–286.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] M. Mdini, G. Simon, A. Blanc, and J. Lecoivre, “Introducing an unsupervised automated solution for root cause diagnosis in mobile networks,” *IEEE Transactions on Network and Service Management*, 2019.
- [7] S. Singh, H. S. Dhillon, and J. G. Andrews, “Offloading in heterogeneous networks: Modeling, analysis, and design insights,” *IEEE Transactions on Wireless Communications*, vol. 12, no. 5, pp. 2484–2497, 2013.
- [8] U. S. Hashmi, S. A. R. Zaidi, A. Darbandi, and A. Imran, “On the efficiency tradeoffs in user-centric cloud RAN,” in *2018 IEEE International Conference on Communications (ICC)*, 2018, pp. 1–7.
- [9] P. Mogensen, W. Na, I. Z. Kovács, F. Frederiksen, A. Pokhariyal, K. I. Pedersen, T. Kolding, K. Hugl, and M. Kuusela, “LTE capacity compared to the shannon bound,” in *2007 IEEE 65th vehicular technology conference-VTC2007-Spring*, 2007, pp. 1234–1238.

- [10] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.-C. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Communications*, vol. 24, no. 2, pp. 98–105, 2016.
- [11] Q. Mao, F. Hu, and Q. Hao, "Deep learning for intelligent wireless networks: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 2595–2621, 2018.
- [12] A. Bibal and B. Frénay, "Interpretability of machine learning models and representations: an introduction." in *ESANN*, 2016.
- [13] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.
- [14] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, 2018, pp. 80–89.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] A. Taufique, M. Jaber, A. Imran, Z. Dawy, and E. Yacoub, "Planning wireless cellular networks of future: Outlook, challenges and opportunities," *IEEE Access*, vol. 5, pp. 4821–4845, 2017.
- [19] 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.

- [20] 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.
- [21] 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.
- [22] 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.
- [23] 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.
- [24] 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.
- [25] 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.
- [26] 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.
- [27] 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.
- [28] 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.
- [29] 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.
- [30] 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.
 - [31] 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.
 - [32] 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.
 - [33] 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.
 - [34] 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.
 - [35] 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.
 - [36] 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.
 - [37] B. Sayrac, J. Riihijärvi, P. Mähönen, S. Ben Jemaa, E. Moulines, and S. Gri-moud, “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.

- [38] 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.
- [39] 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.
- [40] 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.
- [41] —, “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.
- [42] 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.
- [43] 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.
- [44] 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.
- [45] 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.
- [46] 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.

- [47] D. Denkovski, V. Atanasovski, L. Gavrilovska, J. Riihijärvi, and P. Mähönen, “Reliability of a radio environment map: Case of spatial interpolation techniques,” in *IEEE 7th international ICST conference on cognitive radio oriented wireless networks and communications (CROWNCOM)*, 2012, pp. 248–253.
- [48] “Atoll, [online] available:<https://www.forsk.com/>.”
- [49] U. S. Hashmi, S. A. R. Zaidi, and A. Imran, “User-centric cloud ran: An analytical framework for optimizing area spectral and energy efficiency,” *IEEE Access*, vol. 6, pp. 19 859–19 875, 2018.
- [50] H. N. Qureshi and A. Imran, “Towards designing systems with large number of antennas for range extension in ground-to-air communications,” in *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2018, pp. 1–5.
- [51] A. AlAmmouri, J. G. Andrews, and F. Baccelli, “Asymptotic analysis of area spectral efficiency in dense cellular networks,” in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 56–60.
- [52] H. N. Qureshi, I. H. Naqvi, and M. Uppal, “Massive MIMO with quasi orthogonal pilots: A flexible solution for TDD systems,” in *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*, 2017, pp. 1–6.
- [53] H. N. Qureshi and A. Imran, “On the tradeoffs between coverage radius, altitude and beamwidth for practical UAV deployments,” *IEEE Transactions on Aerospace and Electronic Systems*, 2019.
- [54] O. Onireti, A. Imran, and M. A. Imran, “Coverage, capacity, and energy efficiency analysis in the uplink of mmwave cellular networks,” *IEEE Transactions on Vehicular Technology*, vol. 67, no. 5, pp. 3982–3997, 2017.
- [55] P. V. Klaine, M. A. Imran, O. Onireti, and R. D. Souza, “A survey of machine learning techniques applied to self-organizing cellular networks,” *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2392–2431, 2017.

- [56] E. Balevi and J. G. Andrews, "Online antenna tuning in heterogeneous cellular networks with deep reinforcement learning," *arXiv preprint arXiv:1903.06787*, 2019.
- [57] A. Zoha, A. Saeed, A. Imran, M. A. Imran, and A. Abu-Dayya, "Data-driven analytics for automated cell outage detection in self-organizing networks," in *2015 11th International Conference on the Design of Reliable Communication Networks (DRCN)*, 2015, pp. 203–210.
- [58] H. B. Yilmaz and T. Tugcu, "Location estimation-based radio environment map construction in fading channels," *Wireless communications and mobile computing*, vol. 15, no. 3, pp. 561–570, 2015.
- [59] S. Üreten, A. Yongaçoğlu, and E. Petriu, "A comparison of interference cartography generation techniques in cognitive radio networks," in *2012 IEEE International Conference on Communications (ICC)*, 2012, pp. 1879–1883.
- [60] C. Phillips, M. Ton, D. Sicker, and D. Grunwald, "Practical radio environment mapping with geostatistics," in *2012 IEEE International Symposium on Dynamic Spectrum Access Networks*, 2012, pp. 422–433.
- [61] J. Li and A. D. Heap, "A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors," *Ecological Informatics*, vol. 6, no. 3-4, pp. 228–241, 2011.
- [62] L. Mitas and H. Mitasova, "Spatial interpolation," *Geographical information systems: principles, techniques, management and applications*, vol. 1, no. 2, 1999.
- [63] F. Susanto, P. de Souza, and J. He, "Spatiotemporal interpolation for environmental modelling," *Sensors*, vol. 16, no. 8, p. 1245, 2016.
- [64] J. Li and A. D. Heap, "A review of spatial interpolation methods for environmental scientists," 2008.
- [65] R. C. Dwarakanath, J. D. Naranjo, and A. Ravanshid, "Modeling of interference maps for licensed shared access in LTE-advanced networks supporting carrier aggregation," in *IEEE IFIP Wireless Days (WD)*, 2013, pp. 1–6.
- [66] 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)*, 2014, pp. 1496–1501.
- [67] S. Üreten, A. Yongaçoğlu, and E. Petriu, “Interference map generation based on delaunay triangulation in cognitive radio networks,” in *2012 IEEE 13th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2012, pp. 134–138.
 - [68] M. Pesko, T. Javornik, A. Košir, M. Štular, and M. Mohorčič, “Radio environment maps: The survey of construction methods.” *KSII Transactions on Internet & Information Systems*, vol. 8, no. 11, 2014.
 - [69] M. Angjelicoski, V. Atanasovski, and L. Gavrilovska, “Comparative analysis of spatial interpolation methods for creating radio environment maps,” in *2011 IEEE 19th Telecommunications Forum (TELFOR) Proceedings of Papers*, 2011, pp. 334–337.
 - [70] C. Phillips, M. Ton, D. Sicker, and D. Grunwald, “Practical radio environment mapping with geostatistics,” in *IEEE International Symposium on Dynamic Spectrum Access Networks*, 2012, pp. 422–433.
 - [71] M. Höyhty, A. Mämmelä, M. Eskola, M. Matinmikko, J. Kalliovaara, J. Ojaniemi, J. Suutala, R. Ekman, R. Bacchus, and D. Roberson, “Spectrum occupancy measurements: A survey and use of interference maps,” *IEEE Communications Surveys & Tutorials*, vol. 18, no. 4, pp. 2386–2414, 2016.
 - [72] F. Yaseen, U. Masood, A. N. Hassan, and I. H. Naqvi, “Graph signal processing-based network health estimation for next generation wireless systems,” *IEEE Communications Letters*, vol. 23, no. 1, pp. 104–107, 2018.
 - [73] M. A. Azpurua and K. D. Ramos, “A comparison of spatial interpolation methods for estimation of average electromagnetic field magnitude,” *Progress in electromagnetics research*, vol. 14, pp. 135–145, 2010.
 - [74] E. J. Candès and B. Recht, “Exact matrix completion via convex optimization,” *Foundations of Computational mathematics*, vol. 9, no. 6, p. 717, 2009.
 - [75] 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.

- [76] 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.
- [77] 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.
- [78] D. Shepard, “A two-dimensional interpolation function for irregularly-spaced data,” in *Proceedings of the 1968 23rd ACM national conference*, 1968, pp. 517–524.
- [79] R. Franke and G. M. Nielson, “Scattered data interpolation and applications: A tutorial and survey,” in *Geometric Modeling*. Springer, 1991, pp. 131–160.
- [80] F. Susanto, P. de Souza, and J. He, “Spatiotemporal interpolation for environmental modelling,” *Sensors*, vol. 16, no. 8, p. 1245, 2016.
- [81] S. Henley, “Nonparametric geostatistics,” pp. 145, year=1981, publisher=Wiley.
- [82] D. Weber and E. Englund, “Evaluation and comparison of spatial interpolators,” *Mathematical Geology*, vol. 24, no. 4, pp. 381–391, 1992.
- [83] M. Deng, Z. Fan, Q. Liu, and J. Gong, “A hybrid method for interpolating missing data in heterogeneous spatio-temporal datasets,” *ISPRS International Journal of Geo-Information*, vol. 5, no. 2, p. 13, 2016.
- [84] R. C. Dwarakanath, J. D. Naranjo, and A. Ravanshid, “Modeling of interference maps for licensed shared access in LTE-advanced networks supporting carrier aggregation,” in *2013 IEEE IFIP Wireless Days (WD)*, 2013, pp. 1–6.
- [85] 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.
- [86] A. Mohamed, O. Onireti, M. A. Imran, A. Imran, and R. Tafazolli, “Control-data separation architecture for cellular radio access networks: A survey and outlook,” *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 446–465, 2015.
- [87] E. Balevi and J. G. Andrews, “Online antenna tuning in heterogeneous cellular

- networks with deep reinforcement learning,” *arXiv preprint arXiv:1903.06787*, 2019.
- [88] S. Alfattani and A. Yonzacoglu, “Indirect methods for constructing radio environment map,” in *2018 IEEE Canadian Conference on Electrical & Computer Engineering (CCECE)*, 2018, pp. 1–5.
 - [89] H. B. Yilmaz and T. Tugcu, “Location estimation-based radio environment map construction in fading channels,” *Wireless communications and mobile computing*, vol. 15, no. 3, pp. 561–570, 2015.
 - [90] G. Sun and J. Van de Beek, “Simple distributed interference source localization for radio environment mapping,” in *IEEE IFIP Wireless Days*, 2010, pp. 1–5.
 - [91] A. Pages-Zamora, J. Vidal, and D. H. Brooks, “Closed-form solution for positioning based on angle of arrival measurements,” in *The 13th IEEE international symposium on personal, indoor and mobile radio communications*, vol. 4, 2002, pp. 1522–1526.
 - [92] M. Pesko, T. Javornik, L. Vidmar, A. Košir, M. Štular, and M. Mohorčič, “The indirect self-tuning method for constructing radio environment map using omnidirectional or directional transmitter antenna,” *EURASIP Journal on wireless communications and networking*, vol. 2015, no. 1, p. 50, 2015.
 - [93] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.
 - [94] Y. Chen, Y. Wang, D. Kirschen, and B. Zhang, “Model-Free Renewable Scenario Generation Using Generative Adversarial Networks,” *IEEE Transactions on Power Systems*, vol. 33, no. 3, pp. 3265–3275, May 2018.
 - [95] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in neural information processing systems*, 2014, pp. 2672–2680.
 - [96] A. Antoniou, A. Storkey, and H. Edwards, “Data augmentation generative adversarial networks,” *arXiv preprint arXiv:1711.04340*, 2017.
 - [97] E. L. Denton, S. Chintala, R. Fergus *et al.*, “Deep generative image models using a laplacian pyramid of adversarial networks,” in *Advances in neural information processing systems*, 2015, pp. 1486–1494.

- [98] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, “Photo-realistic single image super-resolution using a generative adversarial network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.
- [99] H. Huang, P. S. Yu, and C. Wang, “An introduction to image synthesis with generative adversarial nets,” *arXiv preprint arXiv:1803.04469*, 2018.
- [100] C. Bowles, L. Chen, R. Guerrero, P. Bentley, R. Gunn, A. Hammers, D. A. Dickie, M. V. Hernández, J. Wardlaw, and D. Rueckert, “GAN augmentation: augmenting training data using generative adversarial networks,” *arXiv preprint arXiv:1810.10863*, 2018.
- [101] A. Zoha, A. Saeed, H. Farooq, A. Rizwan, A. Imran, and M. A. Imran, “Leveraging intelligence from network cdr data for interference aware energy consumption minimization,” *IEEE Transactions on Mobile Computing*, vol. 17, no. 7, pp. 1569–1582, July 2018.
- [102] H. Farooq and A. Imran, “Spatiotemporal mobility prediction in proactive self-organizing cellular networks,” *IEEE Communications Letters*, vol. 21, no. 2, pp. 370–373, 2016.
- [103] H. F. A. I. Syed Muhammad Asad Zaidi, Marvin Manalastas, “AI4Networks Simulator - A True 3GPP Compliant 5G Network Simulator with Support of AI,” *IEEE Access (Submitted)*, 2019.
- [104] Matlab, “Why Use MATLAB and Simulink for 5G?” Tech. Rep. [Online]. Available: <https://www.mathworks.com/solutions/wireless-communications/5g.html>
- [105] ns 3, “mmWave Cellular Network Simulator,” Tech. Rep. [Online]. Available: <https://omnetpp.org/>
- [106] OMNeT++, “OMNeT++: Discrete Event Simulator,” Tech. Rep. [Online]. Available: <https://apps.nsnam.org/app/mmwave/>
- [107] OPNET, “OPNET: Optimum Network Performance,” Tech. Rep. [Online]. Available: <https://www.openairinterface.org/>
- [108] OpenAirInterface, “OpenAirInterface: 5G Software Alliance for Democratising

Wireless Innovation,” Tech. Rep. [Online]. Available: <http://opnetprojects.com/opnet-simulator/>

- [109] J. Baek, J. Bae, Y. Kim, J. Lim, E. Park, J. Lee, G. Lee, S. I. Han, C. Chu, and Y. Han, “5G K-Simulator of Flexible, Open, Modular (FOM) Structure and Web-based 5G K-SimPlatform,” in *IEEE Annual Consumer Communications Networking Conference (CCNC)*, 2019.
- [110] M. K. Muller, F. Ademaj, T. Dittrich, A. Fastenbauer, B. R. Elbal, A. Nabavi, L. Nagel, S. Schwarz, and M. Rupp, “Flexible multi-node simulation of cellular mobile communications: the Vienna 5G System Level Simulator,” *EURASIP Journal on Wireless Communications and Networking*, vol. 2018, 2018.
- [111] S. Sun, G. R. MacCartney, and T. S. Rappaport, “A novel millimeter-wave channel simulator and applications for 5G wireless communications,” in *2017 IEEE International Conference on Communications (ICC)*, 2017, pp. 1–7.
- [112] N. Mohsen and K. S. Hassan, “C-RAN simulator: A tool for evaluating 5g cloud-based networks system-level performance,” in *2015 IEEE 11th International conference on wireless and mobile computing, networking and communications (WiMob)*, 2015, pp. 302–309.
- [113] T. Domínguez-Bolaño, J. Rodríguez-Piñeiro, J. A. García-Naya, and L. Castedo, “The GTEC 5G link-level simulator,” in *2016 1st International Workshop on Link and System Level Simulations (IWSLS)*. IEEE, 2016, pp. 1–6.
- [114] X. Wang, Y. Chen, and Z. Mai, “A novel design of system level simulator for heterogeneous networks,” in *2017 IEEE Globecom Workshops (GC Wkshps)*, 2017, pp. 1–6.
- [115] V. V. Díaz and D. M. Aviles, “A path loss simulator for the 3GPP 5G channel models,” in *2018 IEEE XXV International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, 2018, pp. 1–4.
- [116] K. Guan, B. Ai, T. Kürner, R. He, A. Möller, and Z. Zhong, “Integrating composite urban furniture into ray-tracing simulator for 5G small cells and outdoor device-to-device communications,” in *2016 10th European Conference on Antennas and Propagation (EuCAP)*, 2016, pp. 1–5.
- [117] D. Giatsios, “FLEX - FIRE LTE testbeds for open experimentation: Flex overview,”

in *3RD INTERNATIONAL NORNET USERS WORKSHOP, OSLO*, 2015.

- [118] “Nitos - network implementation testbed using open source platforms, [online] available: <http://nitlab.inf.uth.gr>.”
- [119] “5g virtual infrastructure provisioning over nitos testbed, [online] available: <https://5ginfire.eu/nitos/>.”
- [120] K. Kondepu, F. Giannone, S. Vural, B. Riemer, P. Castoldi, and L. Valcarengi, “Experimental demonstration of 5G virtual EPC recovery in federated testbeds,” in *2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*, 2019, pp. 712–713.
- [121] “5G innovation centre, university of surrey, [online] available: <https://www.surrey.ac.uk/5gic>.”
- [122] J. Costa-Requena, A. Poutanen, S. Vural, G. Kamel, C. Clark, and S. K. Roy, “Sdn-based upf for mobile backhaul network slicing,” in *2018 European Conference on Networks and Communications (EuCNC)*, 2018, pp. 48–53.
- [123] *Surrey Platform*. [Online]. Available: <https://5genesis.eu/surrey-platform/>
- [124] M. Ott, I. Seskar, R. Siraccusa, and M. Singh, “Orbit testbed software architecture: Supporting experiments as a service,” in *First International Conference on Testbeds and Research Infrastructures for the Development of NeTworks and COMmunities*, 2005, pp. 136–145.
- [125] T. Chen, M. B. Dastjerdi, G. Farkash, J. Zhou, H. Krishnaswamy, and G. Zussman, “Open-access full-duplex wireless in the orbit testbed,” *arXiv preprint arXiv:1801.03069*, 2018.
- [126] “Open-access research testbed for next-generation wireless networks (orbit) , [online] available: <http://www.orbit-lab.org/>.”
- [127] “Open-access research testbed for next-generation wireless networks (orbit) , [online] available: <https://www.phantomnet.org/>.”
- [128] A. Banerjee, J. Cho, E. Eide, J. Duerig, B. Nguyen, R. Ricci, J. Van der Merwe, K. Webb, and G. Wong, “Phantomnet: Research infrastructure for mobile net-

working, cloud computing and software-defined networking,” *GetMobile: Mobile Computing and Communications*, vol. 19, no. 2, pp. 28–33, 2015.

- [129] E. Luther, “5G massive MIMO testbed: From theory to reality,” *white paper*, 2014.
- [130] S. Malkowsky, J. Vieira, L. Liu, P. Harris, K. Nieman, N. Kundargi, I. C. Wong, F. Tufvesson, V. Öwall, and O. Edfors, “The world’s first real-time testbed for massive MIMO: Design, implementation, and validation,” *IEEE Access*, vol. 5, pp. 9073–9088, 2017.
- [131] J. Vieira, S. Malkowsky, K. Nieman, Z. Miers, N. Kundargi, L. Liu, I. Wong, V. Öwall, O. Edfors, and F. Tufvesson, “A flexible 100-antenna testbed for massive MIMO,” in *2014 IEEE Globecom Workshops (GC Wkshps)*, 2014, pp. 287–293.
- [132] S. Mattisson, “Overview of 5G requirements and future wireless networks,” in *ES-SCIRC 2017-43rd IEEE European Solid State Circuits Conference*, 2017, pp. 1–6.
- [133] J. Jermyn, R. P. Jover, M. Istomin, and I. Murynets, “Firecycle: A scalable test bed for large-scale LTE security research,” in *2014 IEEE International Conference on Communications (ICC)*, 2014, pp. 907–913.
- [134] J. L. Jermyn, “Discovering network control vulnerabilities and policies in evolving networks,” Ph.D. dissertation, Columbia University, 2017.
- [135] T. Wirth, L. Thiele, T. Haustein, O. Braz, and J. Stefanik, “LTE amplify and forward relaying for indoor coverage extension,” in *2010 IEEE 72nd Vehicular Technology Conference-Fall*, 2010, pp. 1–5.
- [136] “Berlin LTE-advanced testbed , [online] available: <https://www.hhi.fraunhofer.de/en/departments/wn/research-groups/software-defined-radio/research-topics/berlin-lte-advanced-testbed.html>.”
- [137] T. Wirth, V. Venkatkumar, T. Haustein, E. Schulz, and R. Halfmann, “LTE-advanced relaying for outdoor range extension,” in *2009 IEEE 70th Vehicular Technology Conference Fall*, 2009, pp. 1–4.
- [138] “5g testbed, [online] available: <https://cewit.org.in/testbed/>.”
- [139] *5G Ready Trial Platform*. [Online]. Available: <https://www.nutaq.com/products/>

- [140] Nutaq, *TitanMIMO-6 Sub 6 GHz Massive MIMO Testbed PRODUCT SHEET*. [Online]. Available: https://www.nutaq.com/wp-content/uploads/2015/07/TitanMIMO6_09_16_2014_Final.pdf
- [141] *ESPOO Aalto 5G research infrastructure*. [Online]. Available: <http://5gtnf.fi/sites/espoo/>
- [142] *HELSINKI*. [Online]. Available: <http://5gtnf.fi/sites/helsinki/>
- [143] *VODAFONE CHAIR MOBILE COMMUNICATION SYSTEMS*. [Online]. Available: <https://www.vodafone-chair.org/>
- [144] W. Anwar, S. Dev, K. Kulkarni, N. Franchi, and G. Fettweis, “On PHY abstraction modeling for IEEE 802.11 ax based multi-connectivity networks,” in *IEEE Wireless Communications and Networking Conference (WCNC 2019)*, 2019.
- [145] N. F. A.H. Mahdi, K. Kulkarni and G. Fettweis, “On network deployment for ultra-reliable communication using multi-connectivity,” in *IEEE Vehicular Technology Conference (VTC Fall 2019)*, 2019.
- [146] *VODAFONE CHAIR & RESEARCH*. [Online]. Available: <https://www.vodafone-chair.org/chair+research#projects>
- [147] T. R. Newman, A. He, J. Gaeddert, B. Hilburn, T. Bose, and J. H. Reed, “Virginia tech cognitive radio network testbed and open source cognitive radio framework,” in *2009 5th International Conference on Testbeds and Research Infrastructures for the Development of Networks & Communities and Workshops*, 2009, pp. 1–3.
- [148] T. R. Newman, S. S. Hasan, D. DePoy, T. Bose, and J. H. Reed, “Designing and deploying a building-wide cognitive radio network testbed,” *IEEE Communications Magazine*, vol. 48, no. 9, pp. 106–112, 2010.
- [149] *5G Playground*. [Online]. Available: https://www.fokus.fraunhofer.de/go/en/fokus_testbeds/5g-playground
- [150] *TAMPERE Tampere University wireless test networks (Hervanta)*. [Online]. Available: <http://5gtnf.fi/sites/tampere/>

- [151] R. Yasmin, J. Petäjäjärvi, K. Mikhaylov, and A. Pouttu, “On the integration of LoRaWAN with the 5G test network,” in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 2017, pp. 1–6.
- [152] *FUSECO Playground*. [Online]. Available: https://www.fokus.fraunhofer.de/go/en/fokus_testbeds/fuseco_playground
- [153] *5G Ready Trial Platform*. [Online]. Available: <https://www.fokus.fraunhofer.de/go/en/5GRTP>
- [154] *Ericsson 5G radio test bed biggest contribution to 5G development in Asia*. [Online]. Available: <https://www.ericsson.com/en/news/2015/10/ericsson-5g-radio-test-bed-biggest-contribution-to-5g-development-in-asia>
- [155] A. E. R. C. P. M. Bjorn Halvarsson, Arne Simonsson and H. Asplund, “5g nr testbed 3.5 ghz coverage results,” in *IEEE 87th Vehicular Technology Conference: VTC2018-Spring*, 2018.
- [156] *SK Telecom*. [Online]. Available: <https://www.sktelecom.com/index.html>
- [157] *Korean ICT News, SK Telecom opens 5G Playground to lead Innovation towards 5G Commercialization*. [Online]. Available: https://www.netmanias.com/en/post/korea_ict_news/8251
- [158] “SK telecom’s 5G architecture design and implementation guidelines (Version 1.35),” *5G Tech Lab Corporate R&D Center, SK telecom*, 2015.
- [159] *OULU 5GTN (Linnanmaa)*. [Online]. Available: <http://5gtnf.fi/sites/oulu/>
- [160] E. Piri, P. Ruuska, T. Kanstrén, J. Mäkelä, J. Korva, A. Hekkala, A. Pouttu, O. Liinamaa, M. Latva-Aho, K. Vierimaa *et al.*, “5GTN: A test network for 5G application development and testing,” in *IEEE European Conference on Networks and Communications (EuCNC)*, 2016, pp. 313–318.
- [161] M. Latva-aho, A. Pouttu, A. Hekkala, I. Harjula, and J. Mäkelä, “Small cell based 5G test network (5GTN),” in *IEEE International Symposium on Wireless Communication Systems (ISWCS)*, 2015, pp. 231–235.

- [162] *TurboRAN*. [Online]. Available: <http://bsonlab.com/TurboRAN/>
- [163] “Openairinterface testbed, [online] available: <https://oailab.eurecom.fr/oai-testbed>.”
- [164] C. Y. Yeoh, M. H. Mokhtar, A. A. A. Rahman, and A. K. Samingan, “Performance study of lte experimental testbed using openairinterface,” in *IEEE 18th International Conference on Advanced Communication Technology (ICACT)*, 2016, pp. 617–622.
- [165] “Openairinterface massive mimo testbed : A 5G innovation platform, [online] available: <https://www.openairinterface.org/>.”
- [166] N. Nikaein, M. K. Marina, S. Manickam, A. Dawson, R. Knopp, and C. Bonnet, “Openairinterface: A flexible platform for 5G research,” *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 5, pp. 33–38, 2014.
- [167] *Munich Experimentation Facility Site*. [Online]. Available: <https://www.5g-vinni.eu/munich-experimentation-facility-site/>
- [168] T. Heyn, J. Morgade, S. Petersen, K. Pfaffinger, E. Lang, M. Hertlein, and G. Fischer, “Integration of broadcast and broadband in LTE/5G (IMB5)-experimental results from the embms testbeds,” in *IEEE European Conference on Networks and Communications (EuCNC)*, 2016, pp. 319–324.
- [169] “Performnetworks testbed, [online] available: <http://morse.uma.es/performnetworks>.”
- [170] A. Díaz-Zayas, C. A. García-Pérez, Á. M. Recio-Pérez, and P. Merino-Gómez, “PerformLTE: A testbed for LTE testing in the future internet,” in *International Conference on Wired/Wireless Internet Communication*. Springer, 2015, pp. 46–59.
- [171] “Performnetworks: a testbed for exhaustive interoperability and performance analysis for mobile networks.”
- [172] *YLIVIESKA Centria University of Applied Sciences test network*. [Online]. Available: <http://5gtnf.fi/sites/ylivieska/>
- [173] S. Verstichel, E. De Poorter, T. De Pauw, P. Becue, B. Volckaert, F. De Turck, I. Moerman, and P. Demeester, “Distributed ontology-based monitoring on the

IBBT wilab. t infrastructure,” in *International Conference on Testbeds and Research Infrastructures*. Springer, 2010, pp. 509–525.

- [174] “w-ilab.t (iminds), [online] available: <http://www.crew-project.eu/wilabt.html>.”
- [175] W. V. B. J. I. M. Bouckaert, Stefan and P. Demeester, “The wilab. t testbed,” in *International Conference on Testbeds and Research Infrastructures*, Springer Berlin Heidelberg, 2010, pp. 145–154.
- [176] *5TONIC: an open research and innovation laboratory focusing on 5G technologies*. [Online]. Available: <https://www.5tonic.org>
- [177] *UNIVERSITY OF BRISTOL 5G TESTBED*. [Online]. Available: <https://5ginfire.eu/university-of-bristol-5g-testbed/>