UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

ADDRESSING TRAINING DATA SPARSITY AND INTERPRETABILITY CHALLENGES IN AI BASED CELLULAR NETWORKS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY IN ELECTRICAL AND COMPUTER ENGINEERING

BY

HANEYA NAEEM QURESHI
Norman, Oklahoma
2021

ADDRESSING TRAINING DATA SPARSITY AND INTERPRETABILITY CHALLENGES IN AI BASED CELLULAR NETWORKS

A DISSERTATION APPROVED FOR THE
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

BY THE COMMITTEE CONSISTING OF

Dr. Ali Imran, Chair

Dr. James Sluss

Dr. Samuel Cheng

Dr. Timothy Ford

Acknowledgments

I would like to express my sincere gratitude to my advisor, Dr. Ali Imran, for his continuous support during my PhD. His expertise, guidance, support, unwavering encouragement and patience has made this possible.

I would also like to thank the rest of my dissertation committee, Dr. Samuel Cheng, Dr. Timothy Ford, Dr. James Sluss and, Dr. Thordur Runolfsson who is no longer with us. Their encouragement and insightful comments during my preliminary and general exams were very valuable.

I would also like to thank our industry collaborators, especially Dr. Omar Al-Kalaa (FDA) and Dr. Hasan Farooq (Ericsson). I really enjoyed working with them during our collaborations.

Throughput my Ph.D., I was supported by various NSF research grants secured by Dr. Ali Imran. So thanks NSF!

I would also like to thank the University of Oklahoma, specifically the Tulsa Graduate College and its dean, Dr. Sluss, and Student Affairs for providing me financial support to attend various international conferences and international visiting researcher opportunities during my Ph.D.

I would also like to thank my fellow colleagues in the AI4Networks Research Center for their support and providing a research conducive work environment.

Finally, I can't thank my family enough: my parents, my sister and my husband, Kamran, for their continuous support, prayers, love and encouragement.

Table of Contents

List of Figures

List of Tables

Abstract

To meet the diverse and stringent communication requirements for emerging networks use cases, zero-touch artificial intelligence (AI) based deep automation in cellular networks is envisioned. However, the full potential of AI in cellular networks remains hindered by two key challenges: (i) training data is not as freely available in cellular networks as in other fields where AI has made a profound impact and (ii) current AI models tend to have black box behavior making operators reluctant to entrust the operation of multi-billion mission critical networks to a black box AI engine, which allow little insights and discovery of relationships between the configuration and optimization parameters and key performance indicators. This dissertation systematically addresses and proposes solutions to these two key problems faced by emerging networks.

A framework towards addressing the training data sparsity challenge in cellular networks is developed, that can assist network operators and researchers in choosing the optimal data enrichment technique for different network scenarios, based on the available information. The framework encompasses classical interpolation techniques, like inverse distance weighted and kriging to more advanced ML-based methods, like transfer learning and generative adversarial networks, several new techniques, such as matrix completion theory and leveraging different types of network geometries, and simulators and testbeds, among others. The proposed framework will lead to more accurate ML models, that rely on sufficient amount of representative training data. Moreover, solutions are proposed to address the data sparsity challenge specifically in Minimization of drive test (MDT) based automation approaches. 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. For the first time, this dissertation reveals

the existence of an optimal bin width for coverage estimation in the presence of inaccurate user positioning, scarcity of user reports and quantization error. The presented framework can enable network operators to configure the bin size for given positioning accuracy and user density that results in the most accurate MDT based coverage estimation.

The lack of interpretability in AI-enabled networks is addressed by proposing a first of its kind novel neural network architecture leveraging analytical modeling, domain knowledge, big data and machine learning to turn black box machine learning models into more interpretable models. The proposed approach combines analytical modeling and domain knowledge to custom design machine learning models with the aim of moving towards interpretable machine learning models, that not only require a lesser training time, but can also deal with issues such as sparsity of training data and determination of model hyper-parameters. The approach is tested using both simulated data and real data and results show that the proposed approach outperforms existing mathematical models, while also remaining interpretable when compared with black-box ML models. Thus, the proposed approach can be used to derive better mathematical models of complex systems. The findings from this dissertation can help solve the challenges in emerging AI-based cellular networks and thus aid in their design, operation and optimization.

CHAPTER 1

Introduction

1.1 Motivation

The key features of emerging cellular networks, such as high multi-Gbps peak data speeds, ultra-low latency, massive device connectivity, high reliability, increased network capacity, and increased availability are set to revolutionize many industries and enable new use cases and applications [3]. Different emerging network use cases have diverse communication requirements. In some cases, such as time-critical applications in healthcare sector, these requirements can be very stringent. To meet these requirements, 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] -[3]. This includes artificial intelligence (AI) enabled self-configuration, self-optimization and self-healing capabilities [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 sys-

tem 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, the full potential of AI in cellular networks remains hindered by two key challenges: (i) training data is not as freely available in cellular networks as in other fields

where AI has made a profound impact and (ii) current AI models tend to have black box behavior making operators reluctant to entrust the operation of multi-billion mission critical networks to a black box AI engine, which allow little insights and discovery of relationships between the configuration and optimization parameters and key performance indicators. This leads to little interpretability [12] or insights [13] for tuning and optimizing different network parameters and complex forms of input and output parameter relationships [14].

This dissertation systematically addresses and proposes solutions to these two key problems of ***sparsity of training data*** and ***lack of interpretability*** faced by emerging networks These two challenges are 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]-[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 [18]. 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 [19].

1.1.2 Lack of interpretability challenge

In additional to requiring vast and representative amounts of training data, a key limitation of traditional ML-based approaches is that they allows little interpretability [12] or insights [13], and remain predominately as black-box models [14]. However, from cellular networks design and optimization perspective, interpretability is a paramount quality

that machine learning methods should aim to achieve if they are to be applied in practice. Instead of using machine learning as a black box, having an interpretable machine learning based model (such as an explicit analytical relationship between different key performance indicator (KPIs) and configuration and optimization parameters(COPs)) will allow insights and discovery of relationships between different KPIs and COPs, which can then to be used to optimize networks autonomously.

1.2 Research objectives

In light of the discussion in sections 1.1.1 and 1.1.2, following research questions are explored in this dissertation:

- Different 5G and beyond (5G&B) enabled use cases have diverse and stringent communication technical requirements. What are some examples of these requirements and how do they compare with the current state of 5G capabilities for different use cases? Why are meeting these requirements a challenge? How can future cellular networks meet these stringent requirements?

- Emerging cellular networks will require artificial intelligence (AI) enabled automation capabilities to meet the requirements of future use cases. Minimization of drive test (MDT) can 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?
- How can we create interpretable models of complex systems, such as cellular networks, using big data and machine learning? Is there a way to leverage analytical modeling and domain knowledge to turn black box machine learning models into grey or white box models?
- Can traditional machine learning models, like neural networks be used to derive COP-KPI equations for cellular network?

1.3 Contributions

The primary contributions of this dissertation are summarized as follows:

- Diverse quantitative requirements for emerging cellular networks use cases are investigated and current gaps and challenges in meeting those requirements are identified.
- 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. Quantification of three key types of errors in MDT-based coverage estimation that stem from inaccurate user positioning, scarcity of user reports and quantization is done. 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, the findings in this dissertation show 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. To solve this problem, a framework to address the data sparsity challenge in cellular networks is proposed based on available information and combination of techniques including interpolation, domain-knowledge based, generative adversarial neural networks, transfer learning, simulators and testbeds. Potential new techniques to enrich sparse data in cellular networks are also proposed, such as by matrix completion theory, and domain knowledge-based techniques leveraging different types of network geometries and network parameters.

- Traditional machine learning based techniques to model different network aspects also suffer from the challenge of lack of interpretability. A combination analytical modeling and domain knowledge to custom design machine learning models with the aim of moving towards interpretable machine learning models is proposed. The proposed approach is tested using both simulated data and real data and results show that the proposed approach outperforms existing mathematical models, while also remaining interpretable. Thus, it can be used to derive better mathematical models of complex systems.

1.4 Dissemination and publications

Throughout the course of preparation for this dissertation, several dissemination activities were carried out. These activities have resulted in the following presentations and (accepted or pending) peer reviewed articles.

Awards

A1. Awarded Gallogly College of Engineering Dissertation Excellence Award by University of Oklahoma, 2021.

A2. Secured first position in 3 Minute Thesis Competition, OU-Tulsa on presenting "Artificial Intelligence based approaches to solving challenges of next generation cellular networks", 2020.

A3. Won the FDA Fellowship Award for "work to develop an evaluation framework for 5G technology in medical devices", 2020.

A4. Won the best research presentation award for academic year 2018-2019 in ECE/TCOM department on presenting the work titled:"Optimal bin width for autonomous coverage estimation using MDT reports in the presence of user positioning error" at OU Graduate Research Meeting 2019.

A5. Won second prize at OU-Tulsa Research Forum 2019 on presenting poster titled: "Planning and Optimizing Networks of The Future: Methods to reduce errors in autonomous coverage estimation using MDT" and first prize at OU-Tulsa Research Forum 2018.

A6. Won second position in IEEE Tulsa section poster competition, 2018.

Journals

J1. Haneya Naeem Qureshi, Marvin Manalastas, Asad Zaidi, Ali Imran and Mohamad Omar Al Kalaa, "Service Level Agreements for 5G and Beyond: Overview, Challenges

and Enablers of 5G-Healthcare Systems," in IEEE Access, vol. 9, pp. 1044-1061, 2021, DOI: 10.1109/access.2020.3046927.

J2. Haneya Naeem Qureshi and Ali Imran. "Optimal bin width for autonomous coverage estimation using MDT reports in the presence of user positioning error." IEEE Communications Letters, vol. 23, issue 4, pp.716-719, 2019, DOI:10.1109/lcomm.2019.2899094.

J3. Haneya Naeem Qureshi, Ali Imran, Adnan Abu-Dayya. "Enhanced MDT-Based Performance Estimation for AI Driven Optimization in Future Cellular Networks." IEEE Access, vol.8, pp.161406-161426, 2020, DOI:10.1109/access.2020.3021030.

J4. Haneya Naeem Qureshi, Marvin Manalastas, Ali Imran and Mohamad Omar Al Kalaa, "Service Level Agreements for 5G-Enabled Healthcare Systems: Challenges and Considerations", IEEE Networks (accepted), 2021.

J5. Ahmad Asghar, Hasan Farooq, Haneya Naeem Qureshi, Adnan Abu-Dayya, and Ali Imran. "Entropy Field Decomposition Based Outage Detection for Ultra-Dense Networks", IEEE Access, 2021, DOI: 10.1109/access.2021.3056551.

J6. Shruti Bothe, Haneya Naeem Qureshi, and Ali Imran, "Which statistical distribution best characterizes modern cellular traffic and what factors could predict its spatiotemporal variability?", IEEE Communications Letters, vol. 23, issue 5, pp.810-813, 2019, DOI: 10.1109/lcomm.2019.2908370.

J7. Haneya Naeem Qureshi, Marvin Manalastas, Aneeqa Ijaz, Ali Imran, Yongkang Liu and Mohamad Omar Al Kalaa, "Communication Requirements in 5G-Enabled Healthcare Applications: Overview and Considerations" (submitted to MDPI Healthcare), 2021.

J8. Haneya Naeem Qureshi, Usama Masood, Ali Imran. "Outage Detection for Emerging Networks: Key Challenges and Solutions" (submitted to IEEE Computational Intelligence), 2021.

J9. Ahmad Asghar, Usama Masood, Haneya Qureshi and Ali Imran, "An AI based coordination framework for conflict avoidance in automated network functions: A key

step towards zero-touch automation" (submitted to IEEE Computational Intelligence), 2021.

J10. Haneya Naeem Qureshi, Umar Farooq, Marvin Manalastas, Ali Imran, Yongkang Liu and Mohamad Omar Al Kalaa, "5G-Healthcare in Mobile Scenarios: Challenges and Considerations" (under review with FDA for submission in IEEE Communications Magazine), 2021.

J11. Haneya Naeem Qureshi, Asad Zaidi, Usama Masood 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 with Ericsson for submission in IEEE Communication Surveys and Tutorials), 2021

J12. Haneya Naeem Qureshi and Ali Imran. "Towards Using Big Data and Machine Learning to Derive Mathematical Models of Complex Systems" (under review), 2021

Conferences

C1. Joel Shodamola, Haneya Naeem Qureshi, Usama Masood, and Ali Imran. "Towards Addressing the Spatial Sparsity of MDT Reports to Enable Zero Touch Network Automation", in IEEE Global Communications Conference (GLOBECOM), Madrid, Spain, 2021.

C2. Haneya Naeem Qureshi and Ali Imran. "Towards designing systems with large number of antennas for range extension in ground-to-air communications." in IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), pp. 1-5, Bologna, Italy, 2018, DOI: 10.1109/pimrc.2018.8580713.

1.5 Organization

This dissertation is structured as follows. Chapter 2 identifies key communication requirements for advanced use cases in emerging networks through examples from healthcare industry and the importance of meeting those requirements through service level

agreements. Several challenges and gaps in existing approaches to meet the requirements of emerging networks use cases are also identified in this chapter. Chapter 3, 4, and 5 of this dissertation aim to propose ways to solve these challenges towards meeting the stringent requirements for use cases leveraging emerging networks. Chapter 3 addresses different errors in MDT-based coverage estimation to enable AI-driven cellular networks. Chapter 4 presents a framework to address the training data sparsity challenge to enhance ML/AI-based solutions for cellular network design, operation and optimization. Chapter 5 addresses the lack of interpretability challenge in traditional ML-based solutions by proposing to combine classical machine learning, analytical modeling and domain knowledge. Finally, chapter 6 discusses the conclusions and future works, and it thus concludes this dissertation.

CHAPTER 2

Identification of key requirements for advanced use cases in emerging cellular networks and challenges in their compliance

2.1 Introduction

This chapter identifies key communication requirements for advanced use cases in emerging networks. In some cases, such as time-critical applications in healthcare sector, these requirements can be very stringent. The importance of meeting these stringent requirements is discussed in context of meeting service level agreements for emerging cellular networks. Several challenges and gaps in existing approaches to meet the requirements and service level agreements in emerging networks use cases are also identified.

2.2 Service level agreements for emerging network use cases

Emerging networks are set to revolutionize many industries and enable advanced applications with estimates of 1.2 billion 5G connections by 2025 [20]. One of the industries where 5G and beyond networks are expected to create a significant impact is healthcare [21, 22, 23]. In this section, identification of the communication requirements and challenges in their compliance are identified using specific examples from advanced emerging networks use cases from the healthcare sector.

Different use cases require different quality of service (QoS) guarantees as shown in Fig. 2.1, which illustrates some 5G-enabled healthcare use cases and qualitatively highlights several communication key performance indicators (KPIs) requirements for each. Inspired by several sources including [24], combined with domain knowledge, we identify commonly used KPIs to highlight the uniqueness of different healthcare use cases in Fig. 2.1. For

example, while energy efficiency might be important for wearable and implantable devices, it might not be so significant in other use cases like remote robotic-assisted surgery. Similarly, mobility related KPIs are more important for use cases such as connected ambulance on the move compared to their values to the static use case of telesurgery. Therefore, it is important to properly document the communication requirements of each application and ensure that those requirements are met to successfully enable the different 5G-enabled healthcare applications. Such documentation can be done through service level agreements (SLAs) between a service provider and the consumer. An SLA details the various aspects of services that the service provider will provide to the consumer, which include but are not limited to performance metrics and guarantees, service level failure and indemnification clauses, service level monitoring process, security and privacy management frameworks, and costs, among others. Thus, SLAs can provide assurance of the guaranteed level of services for facilitating 5G-healthcare use cases.

However, the unique technical characteristics and peculiarities of emerging 5G & beyond technologies make the traditional practices and procedures across SLA stages inadequate for 5G & beyond enabled healthcare Compared with wireless technologies that are currently common in medical devices like Wi-Fi and Bluetooth, 5G is a centrally-managed network that expands the set of stakeholders participating to deliver the medical device functionality. Assessing and managing the risks of communication loss, delay, or disruption is complicated by the rich set of 5G features that are necessary to enable some medical device applications like network slicing where maintaining the performance of several network slices at the same time is challenging compared to the existing service assurances in legacy networks [25].

Moreover, 5G and beyond networks will operate in a multi-domain, multi-operator environment with increasing number of users and varying applications with diverse requirements. Accordingly, these networks resemble an assembly of different autonomous networks, each having their own role in the service provision, their own technology and

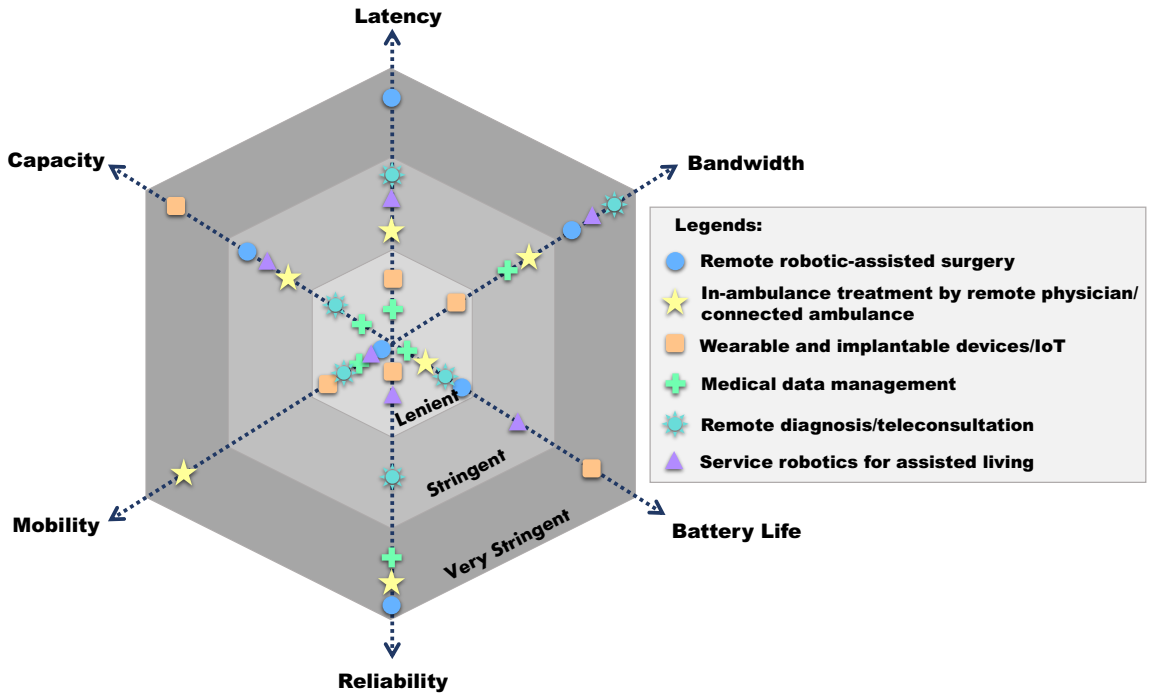


Fig. 2.1: Some 5G-enabled advanced applications and their different degrees of expected requirements. The three tiers of the hexagon indicate the level of expected requirements. The innermost tier corresponds to lenient KPI requirement, the middle tier corresponds to stringent KPI requirement and the outermost tier corresponds to highly stringent KPI requirement. For example, remote robotic-assisted surgery needs very stringent latency, data rate and reliability requirements, stringent capacity requirement and lenient battery life and mobility requirements.

operated by separated entities [26]. Therefore, ensuring that various 5G-enabled medical devices receive the communication services needed per their unique requirements is important, especially for applications that perform critical functions (e.g., life-supporting, life-sustaining).

There are gaps in the literature regarding 5G and beyond SLAs. SLAs in literature are discussed in various technical domains, such as IT data centers [27, 28, 29, 30, 31], web services [32, 33, 34, 35, 36], optical communication systems [37, 38], cloud computing and IoT [39, 40]. Literature reports on SLAs for cloud computing and IoT are numerous. The studies in [41] and [42] identified over 300 existing works related to SLAs in the domain of cloud services in IoT. To present a systematic and comprehensive literature review on the

topic, authors in [41] did a systematic mapping study on management of SLAs for cloud computing and IoT and categorized their findings into various SLA stages and aspects and analyzed select reports in [42]. However, the focus of [42] was not to compare the technical details of the existing literature, but to to analyze the existing literature and categorize the relevant reports with respect to their research contribution areas, maturity level of the evaluated contributions, tool support and application domains within cloud computing and IoT. Notably, the authors concluded that there are few studies focusing on concrete metrics for qualitative or quantitative assessment of quality of service (QoS) in SLAs, which highlights a need for in-depth research on metric specification and measurement methods for SLAs.

There is scarce literature addressing SLAs for 5G and beyond networks. To the best of our knowledge, the only papers that discuss SLAs in this context are [43, 44, 45, 46, 25, 47, 48, 49, 50]. There are gaps in literature regarding 5G-healthcare SLAs that should be addressed to facilitate the implementation of 5G-enabled use cases. SLAs for 5G and beyond networks are addressed in a limited number of articles that primarily aim to propose specific technical solutions and the evaluation of those solutions. No previous work has comprehensively investigated whether traditional SLAs are adequate for 5G and beyond networks or detailed the challenges and limitations that can render them insufficient, which are gaps that will be filled in this chapter. Moreover, there is no existing work that addresses any aspect of SLAs in advanced use cases of 5G-healthcare systems.

2.2.1 Why are traditional SLA approaches insufficient in 5G and beyond networks?

A comparison of SLAs in 5G and beyond environment with legacy SLAs is provided in this section. Specifically, it is identified why traditional SLA approaches will not suffice for 5G and beyond enabled use cases and applications and how to overcome those challenges in

Section 2.2.1. All stakeholders in 5G SLAs can benefit from this information to facilitate 5G-enabled applications.

5G and beyond networks have new and evolved technical characteristics that are not considered in existing practices of SLA generation and management. Hereafter, these aspects are described and grouped based on the section of the 5G network architecture where they appear and discuss how they can be addressed in evolved 5G SLAs.

RAN side/PHY layer aspects

5G and beyond networks are highly heterogeneous, including multi-vendor equipment, multi-operator, multi-modal environments, and multi-frequency spectrum allocations (e.g., sub-6 GHz, millimeter wave spectrum [mmWave]). Accordingly, there are new SLA considerations to the 5G radio access network (RAN) and physical layer (PHY).

Given the plethora of existing network carriers (i.e., spectrum physical resources or bearers) in the sub-6 GHz bands, the user equipments (UEs) should be camped on the optimal carrier for a given SLA service type. For example, in the case of SLAs leveraging ultra-reliable low-latency communication (URLLC), voice users should camp on larger coverage bands with commonly limited bandwidth and UEs with low latency requirements should be camped on medium bands with larger bandwidth. Accordingly, an evolved SLA should include the mechanism and guarantees for carrier association, i.e., assurance that UEs will be camped on the desired band identified for the specific use-case. In massive machine type communications (mMTC) based SLAs, searching for multiple bands can have negative implications on the energy efficiency of power-constrained IoT devices, which also can be addressed by a band selection clause in evolved 5G SLAs.

Notably, the use of mmWave spectrum contributes to enhanced 5G network capabilities compared to legacy networks. Using mmWave alleviates the capacity crunch in existing networks because of the limited spectrum available in sub-6 GHz bands. However, cell discovery in mmWave bands is challenging due to pencil-like beams, which might delay or

prevent the UE from associating with a nearby large bandwidth mmWave cell. Therefore, SLAs in the 5G context, should also consider the probability of miss-association and the related impact to maintaining high download and upload speeds in SLAs leveraging enhanced mobile broadband (eMBB) use case.

Also relevant in the mmWave spectrum is the UE hand over (HO) process, especially in high mobility use cases. A successful mmWave HO completes the cell discovery process of the HO target cell including the challenging beam alignment that can be complicated by the user mobility or environmental changes like obstructions and nearby objects. Therefore, new metrics addressing cell discovery and beam alignment issues as a function of the user speed can be incorporated in evolved 5G SLAs for high-mobility scenarios.

Moreover, 3GPP specifies adaptive 5G numerology (i.e., frame structure) in order to accommodate diverse services like eMBB, mMTC, URLLC and the associated user requirements [75]. Compared to 4G networks, where the transmission time interval (TTI) is fixed to 1 ms, 5G networks can adapt the transmission by varying the TTI or symbol duration to address the desired KPI constraints, while considering the impact of UE mobility and varying channel conditions. For example, an adaptive numerology to meet the latency requirements for URLLC applications might be a subcarrier spacing of 120 kHz and slot duration (i.e., equivalent to TTI) of 0.125 ms. When TTI becomes smaller, the signals will be transmitted in a larger bandwidth since frequency is inversely proportional to time scale. Due to larger signal bandwidth, the channel will be more susceptible to frequency selective fading, which occurs when the signal bandwidth becomes larger than the coherence bandwidth of the channel. A consequence of frequency selective fading is that different frequency components in the signal get attenuated by different amounts, which limits the range of communication or cell radius. Therefore, larger TTI is suited for eMBB/mMTC use cases or use cases that require a larger radius, but with smaller TTI, lower latency can be achieved at the cost of reduced cell size. Another factor to consider is the subcarrier spacing where a small value leads to a short TTI, which might be desirable

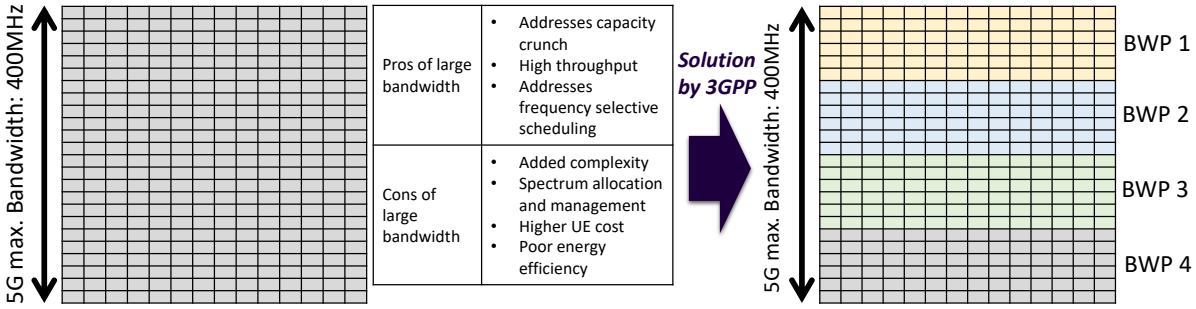


Fig. 2.2: The concept of 5G bandwidth adaptation.

for quick transmissions and hybrid automatic repeat request (HARQ) feedback. Hence, in contrast to legacy SLAs, SLAs for 5G and beyond should consider the TTI constraints to ensure the harmony between the application requirements and network capabilities (e.g., a conflict arising when the SLA specifies 0.125 ms TTI but the network is configured to support 1 ms TTI).

Another 5G physical (PHY) layer aspect is the division of spectrum into the bandwidth parts specified in 5G new radio (NR) as illustrated in Fig. 2.2. A static bandwidth allocation close to the upper end of possible values (i.e., 400 MHz) is challenging for IoT devices and sensors having low power and low processing capabilities that are typical in mMTC applications. Therefore, the introduction of bandwidth adaptation in 5G can provide flexibility and facilitate power saving. This highlights the importance of considering energy efficiency in 5G SLAs and how it relates to the bandwidth allocated to the user by the 5G network to ensure a desired application receives adequate network resources and avoid being under-scheduled.

4G LTE networks perform resource allocation as multiples of one time slot, where 1 slot = 1 ms = 14 orthogonal frequency-division multiplexing (OFDM) symbols. 5G introduces the concept of mini slots where a UE can be allocated resources on the symbol level (e.g., 2, 4 or 7 symbols in a minislot). The concepts of minislots and adaptive numerology are illustrated in Fig. 2.3. Also possible in 5G is aggregating slots to reduce the signaling overhead during resource allocation. Instead of acknowledging every physical resource block (PRB) separately, ACK/NACK are sent for a group of PRBs due to slot

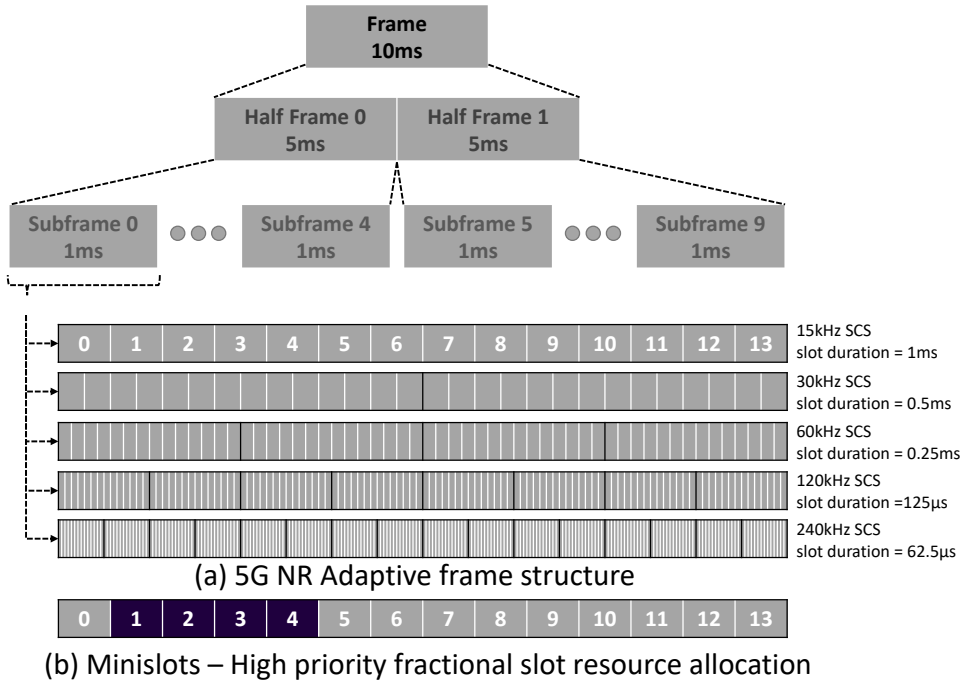


Fig. 2.3: 5G adaptive numerology and minislots.

aggregation. Moreover, minislots can pre-empt normal transmissions, which can be useful for URLLC services and time-critical communication. Accordingly, 5G SLAs can be augmented to consider limits on the variable allocated resources, i.e., how many symbols in a mini slot are needed and would be provisioned for a specific service, whether slot aggregation is allowed, and whether and how frequently minislot pre-emption is allowed.

Core side/network layer aspects

5G network slicing is an innovative flexibility in the network architecture to facilitate the provision of 5G network resources according to specific SLAs. Network slicing permits the partitioning of network architecture into virtual elements, such that each virtual element is suited for a specific use-case or SLA. However, to enable SLA assurance and verification, the network performance data collected to establish SLA KPIs should address the network slice which can be different from the data collected for the overall network.

Unlike SLAs in legacy telecommunication networks that share many similarities resulting in similar SLA metrics, slice-based 5G networks can offer unique services that can be

addressed in a per-slice SLA approach, where individual SLAs have unique elements, metrics and structure. Notably, the business model, SLA structure, QoS specifications, cost model, and the level of service can differ between slices [44]. Accordingly, new scheduling and resource allocation mechanisms (e.g., via weighted slice distribution strategy) and network admission control policies can be considered in the per-slice SLA. Other types of SLAs that can be applicable in a 5G network slicing environment include shared SLAs (i.e., shared between specific number of customers that use the same slice) and hybrid SLAs (i.e., expected to serve certain customers first and then serve the authorized customers of the same slice [44]).

5G and beyond networks are dynamic and can adapt the provided service according to the customer demand for specific KPIs. Accordingly, dynamic SLAs should be considered to capture the limits within which the service provider and customer will operate. An example of dynamic service provisions is those of cloud services where the provider offers cloud facilities in various modes that are capable of scaling up or down in real time to meet the customer demand for resources. This flexibility is coupled with a dynamic change in the SLA QoS parameters [51]. Another example is a telesurgery platform requiring low-latency communication for the duration of the procedure, i.e., the customer can be charged for a network slice to meet their demand for latency and bandwidth for the duration of the surgery. However, when the surgery is complete, the customer would invoke the mechanism specified in the dynamic SLA with the service provider to change their demand for network resources [44].

2.2.2 Challenges in compliance of SLAs in 5G and beyond systems

Evolved SLAs for dynamic 5G and beyond networks are more complex than existing ones in terms of agile network management to accommodate novel applications and dynamic QoS requirements. In this section, we identify and describe 5G SLA challenges during the various stages of the SLA lifecycle, including challenges in the stages of SLA development,

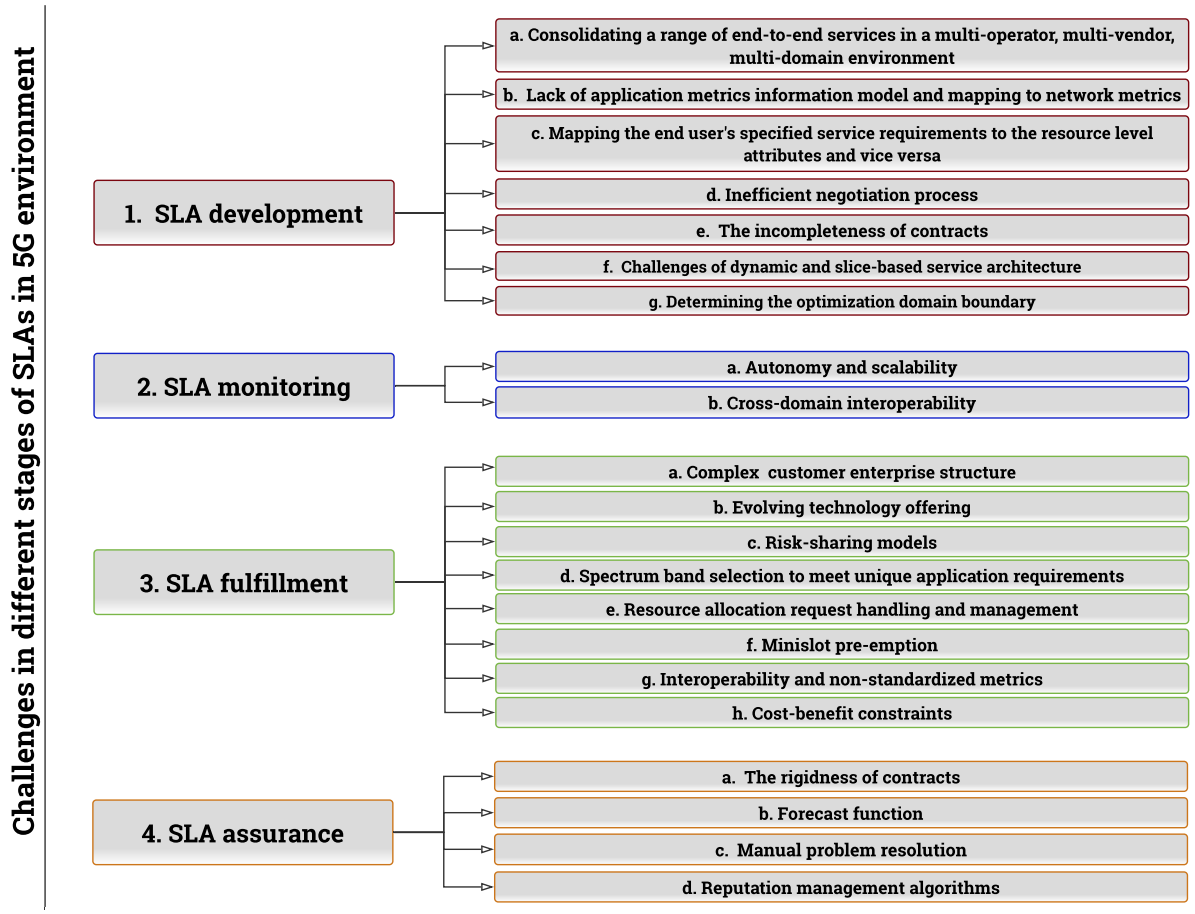


Fig. 2.4: Challenges of 5G SLAs categorized according to the development, monitoring, fulfillment, and assurance SLA parts.

monitoring, fulfillment and assurance. These challenges are illustrated in Fig. 2.4 and are described below. Notably, there is a correlation between the challenges identified in this section given the common theme of dynamic and heterogeneous 5G and beyond networks.

Challenges in SLA development

Specifying the customer communication needs and mapping those needs to the 5G network technical capabilities establish the theme of challenges during the SLA development that include the following:

Consolidating a range of end-to-end services in a multi-operator, multi-vendor, multi-domain environment In 5G and beyond networks, the service is provided as a result of

a multi-stakeholder collaboration that involve multiple network technologies. Ownership of the entire ecosystem is commonly not held by a single entity. Outsourcing of service functions is expected to increase in the 5G business model to save costs, reduce risk, or to benefit from specialized service providers [52]. In this case, networks providers lease parts of their networks, which can be managed through agreements with the lessees and between providers and end customers. Accordingly, delivering a desired service to the end customer involves processes for alignment and coordination between the various involved service providers. This highlights the opportunity to establish methods for developing SLAs where multiple parties are involved in the service delivery. One work in this direction is proposed in [53]. Other propositions in this context are given in [26], where two scenarios are identified to provide an end-to-end service for an end-user: (i) the end-user must manage different SLAs and is the only one who manages their interactions from end-to-end; (ii) the end-user manages only one SLA with a service provider and all necessary information for service management is propagated into the network from end to end, including out-sourced components. Furthermore, it is not straightforward to implement an end-to-end service level management system that can accurately and granularly measure network performance in a 5G environment with varying logical architectures, functional splits, and QoS needs across network layers [26].

Lack of application metrics information model and mapping to network metrics Considering the application side, information models or templates might not exist to identify the communication performance metrics and other technical details that are needed to fulfill the intended functionality of the plethora of 5G service types and applications. Such templates help the stakeholders to cooperate and negotiate tradeoffs to facilitate service delivery. On the network side, choosing a configuration of network parameters to meet the desired application performance can benefit from a mapping between the SLA metrics and 5G network parameters that highlights the sensitivity of desired performance to the change in network configuration. This can be accomplished by leveraging domain knowl-

edge in both the service application area and 5G network management, which exceeds in complexity compared to the legacy networks because of the increase in the number of network parameters and their complex interdependencies. The work in [26] attempts to map ten services to ten network technology independent parameters by considering four performance classes: 1) very high performance, 2) high performance, 3) default performance, and 4) indifferent. However, this work does not consider 5G applications and metrics.

Mapping the end-user's specified service requirements to the resource level attributes and vice versa The exchange of information between SLA stakeholders becomes challenging with the increase in number and business interests of the stakeholders. Accordingly, reaching a compromise that satisfies the SLA requirements can benefit from a precise mapping of the customer high-level communication requirements (e.g., achieving a specific latency value for a telesurgery platform) to the low-level network KPIs and network policy resource-level attributes [45]. This helps bridge the gap between the expectations of customers and service providers and facilitate negotiation clarity between stakeholders in the SLA development phase. The studies in [46] and [45] aim to address this challenge using data analytics and artificial neural networks to automatically identify the interdependencies between different parameters. A framework that implements the reverse process is proposed in [54], where the authors address the translation of low-level metrics to high-level SLA terms that are used in cloud service level agreements.

Inefficient negotiation process Manual negotiations of SLA metrics and service assurances can be inefficient. This is especially true in 5G and beyond networks due to the increased complexities highlighted earlier in this section. Accordingly, it is likely that automated inter-domain negotiation processes will be developed and used to determine the importance of different KPIs by analyzing the predefined service parameters while leveraging historic data documenting the service provider's negotiations [45]. This ap-

proach also helps focus on the most relevant KPIs for a certain application for inclusion in the SLA.

The incompleteness of contracts SLAs are inherently limited by the technical scenarios envisioned upon SLA creation. Hence, changing requirements might lead to situations that are not covered by the SLA terms. Furthermore, verifiable data can be challenging to obtain for service level specification. Accordingly, it is not uncommon to find qualitative statements such as "as soon as possible" in the SLA [55]. These gaps in SLA coverage could result in conflict, which highlights the importance of transparency, ongoing communication, and cooperation between the SLA stakeholders.

Challenges of dynamic and slice-based service architecture While network slicing contributes to maintaining cost-effective network operations, it is challenging for the network operator to allocate portions of the network on-demand. The trade-off between static and dynamic network slicing, which is also applicable to static and dynamic SLAs, involves network efficiency, complexity, and cost. In a static slicing scenario, simplicity is achieved by configuring the network once to allow users continuous access to the allocated network resources without impacting other slices. However, cost and network efficiency are sub-optimal considering that users allocated to a busy slice cannot benefit from the resources available in an idle slice. Dynamic network slicing on-demand can alleviate this inefficiency. However, the challenge is to decide when and which slices to pre-empt to provide the users in the slice covered by the SLA with the agreed services. Moreover, accurate SLA assurance verification in a slice-based environment relies on per-slice KPI monitoring, which should be clearly captured in the SLA.

Determining the optimization domain boundary The SLA stakeholders should consider the limits of their influence on the network optimization strategies and the impact of those strategies on the services promised to the customer and the services provided by

the network operators to other customers.

Challenges in SLA monitoring

Revolving around the task of capturing useful data streams in a heterogenous network to facilitate adequate SLA monitoring, we describe the following challenges of SLA monitoring in 5G and beyond environments.

Autonomy and scalability Manual monitoring of SLA parameters can be expensive, time-consuming, and unscalable. Although the automated monitoring tools used by network operators could be leveraged to support SLA monitoring, access to these tools is commonly reserved to the internal use of the service provider. Using common signaling (e.g., generalized multi-protocol label switching) with a generic policy manager or a third party can help automate the SLA monitoring tasks. However, this will include the added burden of mapping the SLA requirements of each SLA to the technical configurations of network equipment used by the service provider and the specification of tools to generate SLA performance metrics [26].

Another challenge is the data volume resulting from data collection for SLA monitoring. Service quality metrics are specified based on detailed infrastructure-based measurements that can generate large volumes of data, which is challenging for customers to analyze and determine the service consistency with the SLA terms. To alleviate the burden of analyzing large data volumes, the stakeholders can identify the most important and relevant data stream and only gather the associated technical reports for assessment. Although this approach can reduce the administrative burden on the SLA stakeholders, there can be cases where the customer requires detailed data collection for traceability and compliance with external reporting commitments. The importance of SLA monitoring automation is further highlighted by the large number of technical counters in heterogenous 5G and beyond networks, the use of vendor-specific monitoring tools by network operators, and the

lack of unified data format for collected data. Accordingly, a gap in the existing methods is the lack of automated SLA monitoring methods that are capable of efficiently addressing the SLA monitoring tasks of 5G and beyond networks. Automated, scalable, and transparent data collection and aggregation helps build trust between SLA stakeholders and promotes efficient use of resources to achieve the customer desired application.

Cross-domain interoperability SLA monitoring methods for 5G and beyond networks should account for cross-domain monitoring involving multiple organizations (e.g., network operators, connectivity outsourcing companies), systems, and entities (e.g., network performance monitor, service and application monitor, virtualization manager or storage manager). Therefore, methods should be considered to permit management information flow across administrative domain boundaries and facilitate an end-to-end view of the service provision in a common platform that promotes cooperation between multiple organizations and integrates multiple domain monitoring modules. However, the lack of standardized performance metrics for use in data collection and aggregation hinders the automation and interoperability of such platform across multiple domains for 5G SLA monitoring.

Challenges in SLA fulfillment

SLA fulfillment is closely related to SLA monitoring. However, the impact of business needs and expectations of the SLA stakeholders highlight the challenges listed below.

Complex customer enterprise structure In complex company structures, it is challenging to correlate the quality of services in terms of business value creation. With growing enterprise complexity, the number of internal customer entities increases along with their inter-dependencies and potentially conflicting requirements. When a value model for the procured services is absent, the sensitivity of the business value of a desired application to service changes is not easily predictable.

Evolving technology offering Customers might attempt to improve their connected applications to remain competitive (e.g., serve more subscribers, increase access speed to medical imaging data). However, there is no financial incentive for the service providers to offer technical capabilities beyond what is needed to meet the established SLA terms. Accordingly, evolving the technology offered by the service providers can be regarded as a challenge since such investment in service quality can impact the provider's cost structures. Customers wishing to expand their access to improved technology should initiate a new negotiation process with the service provider [55]. Therefore, the customer should maintain up-to-date technology landscaping efforts in the evolving 5G and beyond networks to be aware of what can be done with improved communication capabilities. On the other hand, the service providers can benefit from the targeted marketing of their communication service offerings to industry verticals.

Risk-sharing models Business costs and success can be perceived differently by the SLA stakeholders, which extends to the associated risk to that success. Accordingly, the SLA stakeholders should determine if and how to consider risk-sharing of the end-to-end service provided to the end-user. Unique industry verticals can approach this topic according to their unique needs.

Spectrum band selection to meet unique application requirements Due to an increasing number of sub-6 GHz carriers in 5G and beyond networks, a challenge for service providers is to ensure that users are camped on the optimal carrier in 5G according to the service type. Spectrum bands in 5G networks are divided into low, medium, and high bands corresponding to less than 1 GHz, 1 GHz to 6 GHz, and 24 GHz to 40 GHz, respectively. Band selection is important because it ensures minimum inter-frequency hand overs by avoiding measurement gaps, which is the key contributor to voice muting occasions (i.e., due to cell radio shifting to another carrier during measurement gaps). In 5G voice services, the biggest problem is call muting, rather than call dropping or call

quality. Muting is a gap in voice packets or real-time transport protocol (RTP) packets, which is perceived by human ears as silence. Call dropping means that a call ends unexpectedly. However, in 5G packet-based voice service, with VoLTE, users are left on the receiving end of silence (i.e., go mute during the call) due to loss of voice packets. Packet loss has a pronounced impact on time-critical applications with low bandwidth requirements whose users would expect to be camped on a low spectrum band with relatively small bandwidth. However, low bands are congested with 2G, 3G, 4G and other services. Accordingly, medium-band with larger bandwidths compared to low-band can be considered for time-critical applications (i.e., SLAs for URLLC use cases).

Resource allocation request handling and management In 5G and beyond use cases, provisions like spectrum sharing and infrastructure sharing complicate the resource allocation in SLA management. For example, short-term services requested through signaling can be challenging to meet by the service provider because of the complexity of managing the network resource reservation while balancing the overall services offered to all customers and maximizing resource utilization [45]. Bandwidth adaptation in 5G and beyond networks and how it can impact the desired application should also be considered and documented in the SLA.

Managing spectrum sharing scenarios would be a challenge as well. If used, spectrum sharing practices should be addressed in the SLA, whereby some service providers might consider the temporary transfer of some or all their spectrum access rights. Furthermore, the optional use of unlicensed spectrum bands is commonly best-effort and lacks service guarantees due to the lack of interference protection in unlicensed spectrum, which raises concerns for wireless coexistence. For example, the coexistence impact of LTE-Licensed Assisted Access (LAA) on users of unlicensed spectrum including wireless medical devices was investigated in [56]. Authors in [57] address the problem of modeling and evaluating the coexistence of LTE LAA in the unlicensed band. Accordingly, considerations of wireless coexistence should be addressed in the SLA if applicable to the offered service.

Another SLA consideration is the network physical resource sharing and its impact on the offered service. Often, a customer does not need a high QoS at all times. For example, in the case of connected ambulance facilitating patient treatment by a remote physician while in transport, the service level needed to operate the associated connectivity would only be needed while the patient is on the way to hospital. Once the patient reaches the hospital, that communication service is no longer needed. For such applications, customers can request on-demand services that are charged on a pay-as-you-use basis, which might be an incentive for the provider to share the network resources between users to achieve profitability [55].

Minislot pre-emption In URLLC use-cases, 5G minislots can pre-empt normal transmissions, which can be useful when there is a need for time-critical communication. However, pre-emption can negatively impact other network users, e.g., a user will be affected if its transmission is pre-empted because of another higher priority user. Therefore, the SLA should consider the trade-offs of using minislot pre-emption that are application specific and lack established best practices.

Interoperability and non-standardized metrics Interoperability should be considered between the various components of the 5G-enabled medical device application [58] in addition to the interoperability between various network equipment vendors to facilitate SLA service delivery. Interoperability challenges for SLA fulfillment are further highlighted by the fact that network performance metrics are commonly vendor-specific, where each network equipment vendor defines metrics using its own set of counters and naming conventions. In addition to managing non-standardized network performance metrics, SLA fulfillment includes the challenge of translating the customer requirements to technical specifications [55], which can be presented as customer business goals. In this case, the SLA stakeholders develop a mapping between the technical and business metrics to align the SLA with their business goals and document the expected business value contribution

of the measurable network performance metrics. Business metrics indicate the progress of a stakeholder's goals and can include metrics for marketing (e.g., incremental sales), sales (e.g., average profit margin), financial value (e.g., debt-to-equity ratio), software as a service (SaaS, e.g., customer retention rate), or social media (e.g., number of twitter followers) [59], [60].

Cost-benefit constraints The customer budget might limit the level of service obtained from the network service provider. Accordingly, the challenge is to maintain a tolerable customer cost-benefit ratio including the cost assessment of possible technical solutions that can meet the customer expectations and the associated trade-offs.

Challenges in SLA assurance

This part of SLA management assures that the provided service achieves the performance set in the SLA.

The rigidness of contracts While foreseeable future requirements are considered during SLA development, the unpredictable change in customer requirements is challenging to address for SLA assurance. Unpredictable requirements encountered during the lifecycle of SLAs complicate the SLA applicability to evolving customer needs where the established correlations might become outdated between business needs, network performance metrics, and cost. Notably, the incentive to adapt an SLA to new situations decreases as the contract period nears its end [55]. Accordingly, considering dynamic SLAs in 5G and beyond networks can help prepare the stakeholders to address evolving technical and business situations during for the SLA duration.

Forecast function An open research question is the development of continuous network forecasting and optimization techniques to optimize a set of desired network aspects (e.g., coverage, energy efficiency, spectral efficiency) based on variable inputs (e.g., traffic,

environmental factors). Although there are reports on advancements in this area, it is unclear what the optimal mapping is between the proposed forecasting techniques and network parameters [26]. However, in dynamic 5G and beyond networks, forecast functions are central to the deployment of features like network slicing, where the network resources are dynamically optimized between slices to improve utilization while meeting the SLA service levels [46]. Hence, the challenge is to develop, deploy, and document a forecast function that meets the optimization objectives and constraints for every network slice with the available input streams.

Manual problem resolution With increasing complexity and heterogeneity of 5G and beyond networks, the lack of automatic problem resolution is challenging. To facilitate efficient service problem resolution, automated tools can be useful in root cause analysis, trouble ticketing, and traffic forecasting.

Reputation management algorithms SLA penalties can negatively impact the service provider reputation [61]. This is augmented in cases where the service performance metrics include client reviews. In 5G and beyond networks, a challenge in the review-based evaluation can be the ease of generation of large volumes of dummy clients by a service provider to build their reputation or damage the competition. For example, due to posting fraudulent reviews against its competitor HTC, Samsung was fined $340,000 in 2013 by the Taiwan Federal Trade Commission [62]. The interested reader is referred to the comprehensive study in [63] for more information on this topic.

Notably, the challenges in meeting requirements for emerging networks identified in this section can be solved using AI-based automation tools.

2.2.3 Considerations and open research questions for enabling advanced emerging networks use cases

This section illustrates open research questions in Fig. 2.5 based on the SLAs challenges described in section 2.2.2. These include trade-offs and practical implementation considerations in 5G network resource allocation. Other considerations include optimizing device performance when using bandwidth adaptation, network slice sharing modes, and dynamic network resource optimization. With the mmWave spectrum enabling 5G performance, research is also needed to understand the integration of user mis-association probability to mmWave cells in the medical device risk evaluation and strategies to address it in the SLA. This is relevant in scenarios like connected ambulance where mobility is important to deliver the intended functionality. With 5G and beyond networks becoming increasingly complex in terms of the number of configuration and optimization parameters and counters, adaptive algorithms to reduce the large set of observable network counters and metrics are needed for helping in efficient network monitoring, especially during the SLA service monitoring and assurance phases. Moreover, algorithms are also needed to flexibly map and optimize the network configuration parameters to meet a desired healthcare application requirement while maintaining business objectives for all stakeholders. This can extend to the development of continuous forecasting and optimization techniques, where AI can play a key role. Given the evolving nature of 5G and beyond networks, and the diverse communication requirements for different healthcare use cases, mechanisms for dynamic SLA negotiation are also needed. Additionally, the heterogenous and multi-domain nature of 5G and beyond networks indicates that collaboration frameworks are needed between the SLA stakeholders to promote interoperability and service delivery.

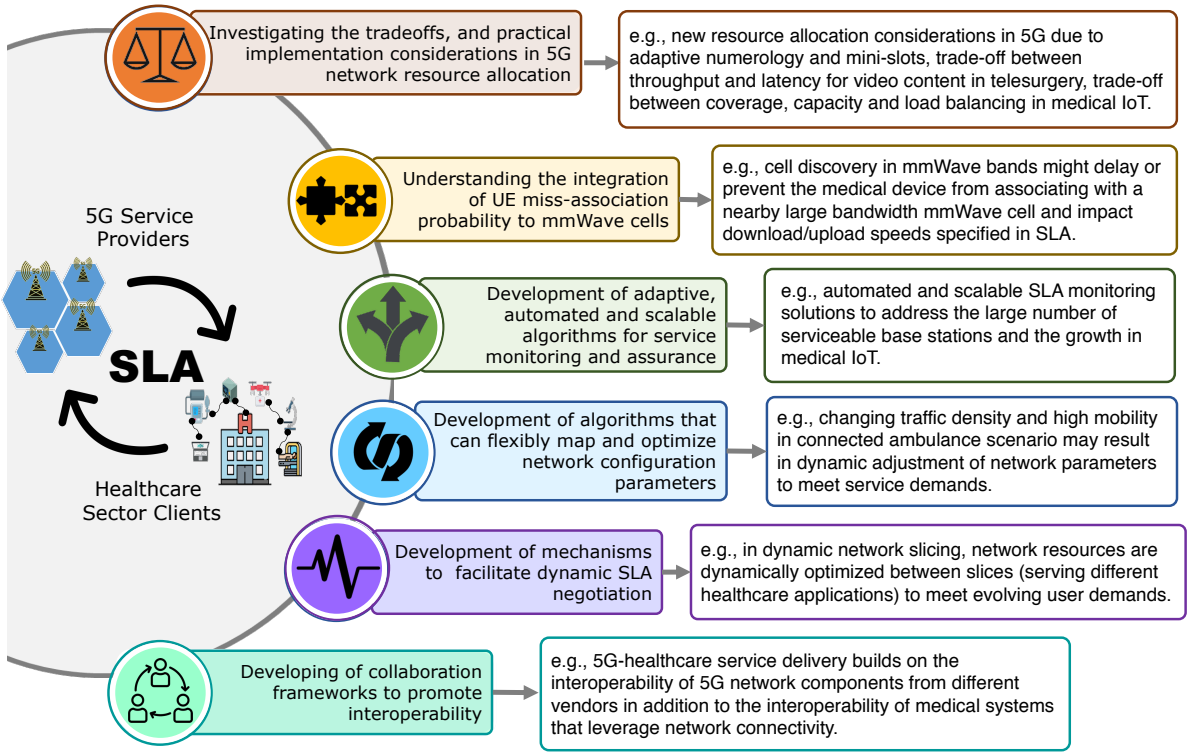


Fig. 2.5: Future research directions to facilitate the safe use of communication in emerging network applications through SLAs.

2.3 Communication requirements for advanced use cases in emerging networks

Knowledge of requirements for emerging networks use cases is important for all the stakeholders including developers, network providers and regulatory authorities in the healthcare sector to facilitate safe and effective healthcare [64]. The expansive set of 5G configuration and optimization parameters offer network operators flexible options in setting up their networks and dynamically optimizing network performance to achieve a desired objective. Accordingly, a large set of parameters can impact the needed performance for a 5G-healthcare use case. 5G network providers can use specific quantitative key performance indicators (KPIs) to assess the feasibility of a given 5G-enabled healthcare use case, provide the level of service required for the safe and effective functioning of 5G-enabled healthcare applications, and for drafting service level agreements with their customers [64].

Clearly specified KPIs can also be helpful to the regulatory authorities like the U.S. Food

and Drug Administration (FDA) to help assess whether communication service levels and quality of service requirements are met to support the safe and effective use of a 5G-enabled medical device.

In this section, we identify, compare, and summarize the communication requirements for several advanced use cases that can be enabled by 5G. Although, we focus on quantitative requirements, qualitative requirements are also highlighted. Furthermore, we identify gaps in the existing literature and highlight considerations in this area. Specifically, we have surveyed the technical requirements for remote robotic-assisted surgery, mobile connected ambulance (i.e., in-ambulance treatment by remote physicians), wearable and implantable devices, and service robotics for assisted living. To the best of our knowledge, the closest work to this section on the similar topic is the recent magazine article by Cisotto et al. [65], which highlights select quantitative requirements for the use cases of telepresence and robotic-assisted telesurgery, remote pervasive monitoring, healthcare in rural areas, and mobile health (m-Health). Compared to the related work, this section includes references specific to the use of 5G in healthcare, in addition to those addressing the communication requirements of the healthcare applications regardless of the enabling communication technology, which can inform how applications use 5G.

2.3.1 Key performance indicators for emerging networks applications

While KPIs such as data rate, accessibility, reliability, and mobility have been widely used in the performance evaluation of 4G cellular networks, the diversity and heterogeneity of 5G applications are calling for further expansion to incorporating novel sets of KPIs for measuring adequacy and efficacy of 5G-enabled services. The taxonomy shown in Fig. 2.6 highlights the vastness of 5G network KPIs. Inspired by [66, 67, 68] and combined with domain knowledge, this taxonomy classifies 5G KPIs into four categories: network, service, application, and user. Each category also includes high-level and low-level KPIs. High-level ones measure the overall performance of the network based on metrics defined

by the standardization bodies such as 3rd Generation Partnership Project (3GPP). However, most of the time, these high-level KPIs are focused on characterizing general features of the cellular system/service. With this regard, we also introduce low-level KPIs under each high-level one to further instantiate specific requirements. A certain 5G-enabled healthcare application might depend on a given set of KPIs to deliver its function while having low sensitivity to others.

The service level KPIs often discussed in 5G-enabled healthcare literature to address several aspects of the communication network include **availability**, **accessibility**, **reliability**, **data rate** and **retainability**. Availability is the fraction of time the network is available to provide the services users demand [69]. Accessibility is discussed in the context of connectivity time, which measures the time to establish a network connection, starting at the user request and ending at the beginning of the data transmission. Reliability is addressed through several low-level KPIs shown in Fig. 1: *throughput*, *latency*, *jitter*, and *packet loss rate (PLR)* or *bit error rate (BER)*. User throughput during active time is the size of a burst divided by the time between the arrival of the first packet and the reception of the last packet of the burst. Latency corresponds to the travel time of data packets from the source to the destination (i.e., one-way, or end-to-end latency) [70]. The round-trip latency is the time it takes a signal to be sent plus the time spent to receive an acknowledgement of that signal. Jitter is a measure of the variation in the time of arrival between packets. If uncontrolled, jitter impacts the audio and video quality which can negatively impact applications where this type of communication is used (e.g., telesurgery, remote diagnosis, service robotics for assisted living). PLR is the fraction of packets that failed to reach the receiver out of total number of transmitted packets. BER is the total number of bits received in error over the total number of bits sent. Like jitter, high BER/PLR negatively impacts the audio and video quality. Also relevant to the service level is the data rate, which is a measure of the volume of successfully received application data, expressed in bits, within a period expressed in seconds. A high data rate is relevant in applications that transport large volumes of data. Service retainability refers to the

count of radio link interruptions following the activation of that link between the user and the network. A related measure of service retainability is the number of reconnections, i.e., the count of attempts a user performs to re-establish network connection following a link failure.

The overall network characteristics are addressed in the literature with several network level KPIs like network **bandwidth**, **utilization** and **spectral efficiency**. Bandwidth refers to the network maximum aggregated data transmission rate. *Connection density* and *traffic density* are measures of utilization. Connection density refers to the number of connected devices per unit area. This is relevant in connected IoT application, where the number of connected devices is large. Traffic density (or area traffic capacity) is a measure of the volume of catered data in a unit area. Spectral efficiency is the maximum number of bits the network can provide to users every second using a given bandwidth.

On the user level, KPIs of **battery or power consumption**, **range** and **payload size** are commonly reported in literature covered in this paper. User battery consumption and the its associated low-level KPI, *duty cycle*, which is the ratio between an application active (ON) and idle (OFF) times, are relevant in IoT devices where transmissions are intermittent and battery lifetime is limited. Range is the distance at which the signal transmitted is sufficient for the transmitter and receiver to communicate effectively. Another relevant KPI discussed in literature is the user payload size, which can be controlled to balance the transferred data volume with the incurred transmission overhead. This promotes efficient network resource usage while helping to meet specific application needs.

On the application level, **security** and **position accuracy** are the most commonly discussed KPIs in literature reviewed in this paper. Security refers to the network ability to identify, isolate, and eliminate threats to its infrastructure, users, and their data. Position accuracy is a measure of the difference between the estimated and actual user locations. The 3GPP (the entity that develops 5G specifications) has set different position accuracy targets for different scenarios ranging from several meters for emergency calls to few decimeters for

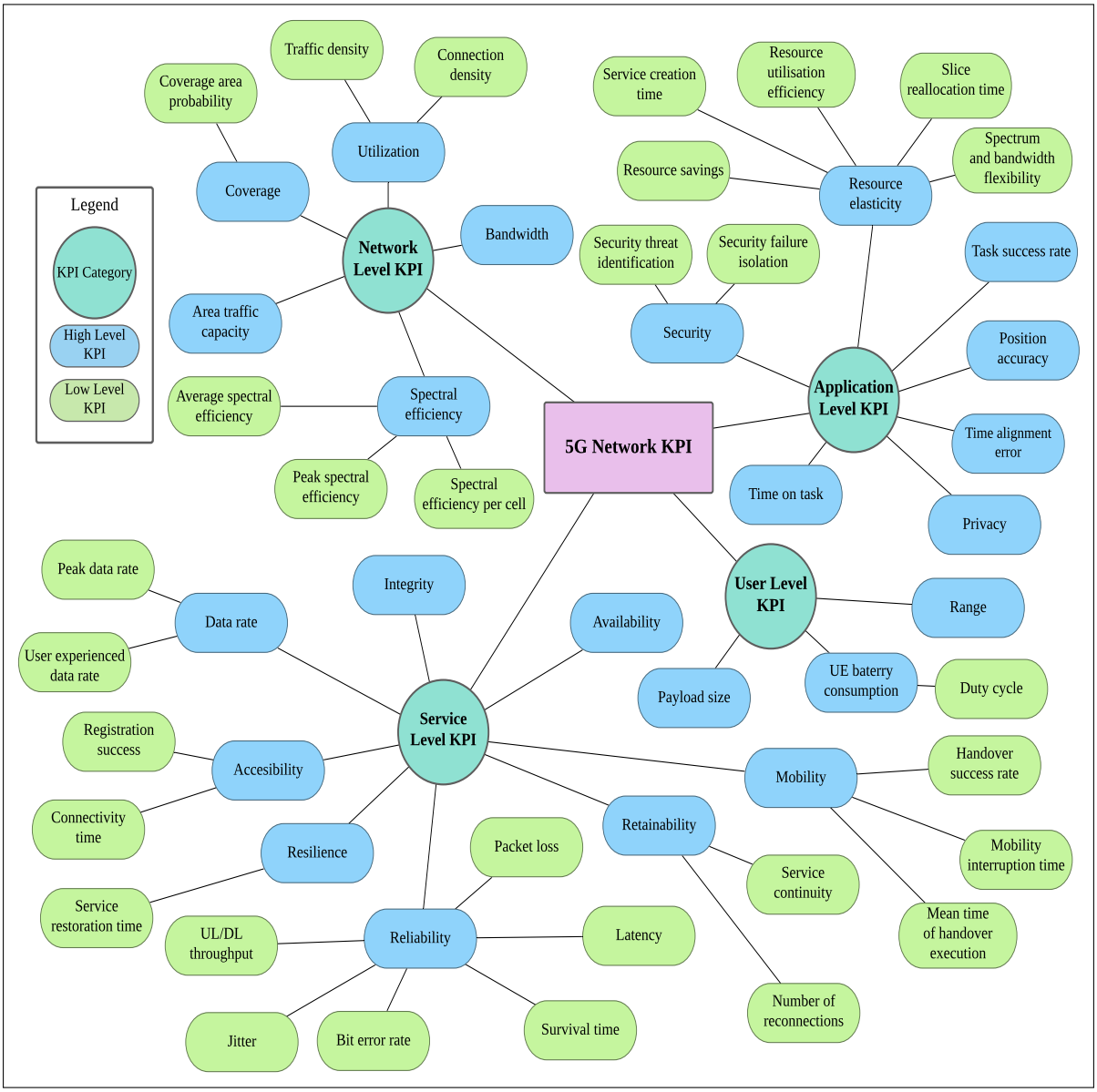


Fig. 2.6: Taxonomy of 5G network KPIs.

indoor plant operations and vehicle-to-everything (V2X) [71].

Although relevant to enabling 5G healthcare functions, some KPIs are seldom addressed in the articles reviewed in this paper. For example, the **network-coverage** is relevant to all applications using its services. While network *coverage area probability* is related to user activity range, it refers to the percentage of service area where users can receive a desired service. On the application level, **privacy** is relevant to healthcare applications because it refers to the ability of the network to keep the data which passes through

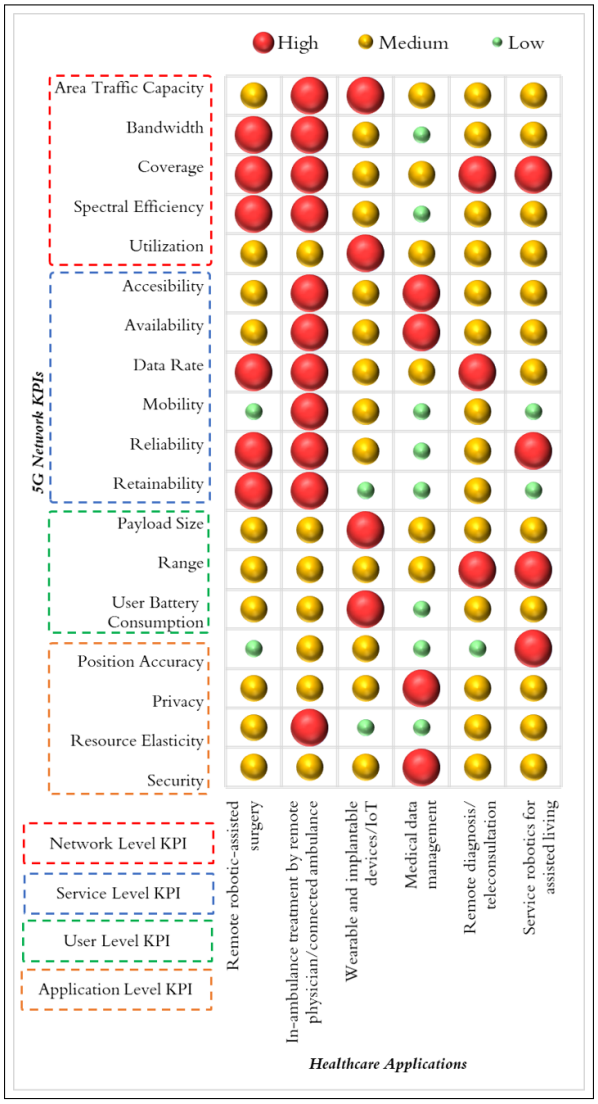


Fig. 2.7: Examples of 5G-enabled applications concepts and their projected needs for some communication KPIs.

or stored in it private. Also on the application level, network **resource elasticity** is relevant in applications with temporary need for elevated connection capabilities like in-ambulance treatment and other emergency related applications. Resource elasticity describes the network ability of responding to temporal and spatial fluctuations in traffic demand by redistributing available resources to seamlessly meet the demand of critical applications [72]. On the service level, **mobility** is relevant to applications that are mobile like connected ambulance. Mobility is the maximum user speed that a network can support. It also refers to the ability of a network to support mobile users. A measure of mobility can be the *rate of successful handovers* between the coverage sites. Additional

examples of KPIs related to the service level include the *service restoration time* under resilience and *survival time* under reliability. The former refers to the period in which the services are restored to normal operating status after experiencing a downtime. The latter is the tolerable packet delay in which an application can still function effectively.

Fig. 2.7 illustrates a subjective summary of the general relevance of the high-level 5G network KPIs we investigated in Fig. 2.6 to the following applications: remote robotic-assisted surgery, connected ambulance or in-ambulance treatment by remote physician, healthcare IoT applications, medical data management, teleconsultation and remote diagnosis, and service robotics for assisted living. These applications are only considered as generic concepts, which recognizes that realistic medical devices implementing one or more of these application concepts have unique KPI needs. Furthermore, the FDA guidance document on radio frequency wireless technology in medical devices recommends that the medical device wireless quality of service (QoS) is specific to the medical device [73]. Accordingly, this summary can help inform the KPI value specifications that should be determined for the specific intended use of a medical device and its design. Relevance is qualitatively described as high, medium, or low. Notably, remote robotic-assisted surgery needs careful provisioning of several KPIs including reliability where low-level KPIs such as latency, jitter, and packet loss fall under. However, when the scenario is implemented in an operating room, mobility is not as relevant as other KPIs since the connection will not move across multiple network cells. On the contrary, in-ambulance treatment by remote physician or connected ambulance needs exceptional mobility support since the data exchange occurs while the ambulance is mobile. Support for mobility in this case complements other relevant KPIs like reliability, data rate, availability, coverage, and resource elasticity to enable the exchange of diverse data streams (e.g., video, audio, file transfer, control commands). The number of connected wearable devices is expected to grow globally from 720 million in 2019 to more than 1 billion in 2022 [74]. Accordingly, the KPIs of utilization and UE battery consumption are highly relevant for enabling the network connectivity for such devices given their energy constraints. In the case of med-

ical data management, security and privacy are more relevant compared to other KPIs, such as reliability. Like other services that use audio and video, remote diagnosis or teleconsultation are negatively impacted with degraded reliability. Other relevant KPIs for this use case include coverage, range, and utilization to facilitate the service access by many users. Finally, we note that reliability, range, and position accuracy are relevant in the service robotics for assisted living use case where the robot is mobile in a limited area. The following sections will identify the requirements for each of these use cases.

Remote robotic-assisted surgery

Several studies have addressed quantitative KPI requirements for remote robotic-assisted surgery, which we also refer to as telesurgery for the remainder of this chapter. This use case involves the use of a robotic-assisted surgery platform by a surgeon located in a remote geographic location. The most commonly reported KPIs include latency, data rate and packet loss [75, 76, 77, 65, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102]. Few studies have also reported quantitative requirements for reliability, communication service availability, payload size, traffic density, connection density, service area dimension, survival time, range, and duty cycle [85, 89, 65, 99]. Table 2.1 presents the reported latency requirements for several communication streams that can be used during telesurgery like camera flows, vital signs, and feedback for force and vibration. Latency in this context is considered end-to-end. Compared to latency, quantitative requirements for jitter are less investigated in literature. The reported jitter requirements are detailed in Table 2.2. Similarly, requirements for data rate are detailed in Table 2.3. These requirements can be influenced by different compression techniques used. Reported packet loss and BER requirements are presented in Table 2.4. Reports of other KPIs, such as reliability, availability, survival time, etc. are listed in Table 2.5. The ability of current 5G networks to meet these KPIs will be discussed in Section 2.3.2.

Notably, the reported KPI values are inconsistent across literature reports, which could

be attributed to the varying type of tasks considered by the researchers during telesurgery. Additionally, the equipment used to perform telesurgery and the simulation environment also varies across studies. To detail the context of the telesurgery KPI specification, we also labeled the original source of the reported KPIs in each study as detailed in Tables 2.1-2.5. Most KPI values were found in experiment and simulation settings of the individual studies with exceptions where the values are a consensus view of the achievable performance by wireless stakeholders [77], [88], and [85]. [77] is a white paper by the 5G Infrastructure Public Private Partnership (5GPPP) that highlighted use cases for 5G in healthcare and suggested quantitative requirements. A technical requirements document was compiled by the IEEE 802.15 Task Group 6 for Body Area Networks (BAN), formed in 2007 to help develop a communication standard optimized for the low power devices and operation, in or around the human body to serve a variety of applications including medical applications. The report in [85] outlined findings from the National Science Foundation (NSF) funded workshop on ultra-low latency wireless networks. The report addressed healthcare application requirements of the emerging applications, including telesurgery, in terms of throughput, latency, and reliability. In the following, the relevant experimental and simulation studies are summarized.

Experiment based In [78], the Aesop 1000TS robot (Computer Motion, Goleta, CA) was adapted to hold a metal pin in addition to a laparoscope and camera (Stryker Instruments, San Jose, CA). Programmed incremental time delays were introduced in the audiovisual acquisition and the number of errors made while performing tasks at various time delay intervals was noted. A remote surgeon in Baltimore, MD performed tasks 9000 miles away in Singapore and determined that a delay of < 700 ms is acceptable.

A teleoperation capable ZEUSTM robotic minimally invasive surgery system was used in [79], with a dedicated communication link by Bell Canada and Telesat Canada. This link included a wired link with a roundtrip delay of 64 ms, a satellite link with a roundtrip delay of 580 ms, and a software simulated delay link through a local switch. Different tasks

were performed from London, Ontario to Halifax, Nova Scotia, Canada. These included dry (typical surgical maneuvers at latencies from 0 to 1 s, in increments of 100 ms) and wet (internal mammary artery takedown on a pig) experiments. A heuristic mathematical model accompanied the task completion times and error rate results, showing acceptable delays of up to 300 ms and 800 ms for simple tasks with training. It was concluded that the effect of delay is not pronounced until the round-trip time exceeds 400 ms and the maximum tolerable delay is approximately 600 ms.

Another work using the ZEUS system is [80], which attempted a remote robot-assisted laparoscopic cholecystectomy on a 68-year-old woman with a history of abdominal pain and cholelithiasis. The surgeon's subsystem (Equant's point of presence, New York) and patient's subsystem (operating room in European Institute of Telesurgery, Strasbourg) were connected via a high-speed terrestrial network (i.e., ATM service), with a round-trip distance of over 14,000 km. Robot motion data had a high priority and a rate guarantee of 512 Kbps within the 10 Mbps virtual path. The operation was carried out successfully in 54 min, with a 155 ms mean time lag for transmission. The study estimated that 300 ms was the maximum time tolerable delay.

Authors in [87] attempted a robot-assisted laparoscopic gall bladder removal for six pigs with the surgeon located in Strasbourg, France and animals located in Paris, France using the ZEUS system. The time lag was artificially increased from 20 ms up to 551.5 ms. It was concluded that no packet was lost during any surgical procedure, the round-trip delay was 78–80 ms, with additional 70 ms for video coding and decoding and a few milliseconds for rate adaptation, summing to 155 ms [87].

To study the impact of haptic feedback in virtual environments, two experimental platforms were implemented in [95]. Platform 1 consisted of two sites, the University of Belfast separated by a few hundred meters and linked by Gigabit Ethernet connection. The configuration of the experimental platform consisted of four 100 Mbps Ethernet segments, two 1000 Mbps fiber optic segments and four PCs. One PC was connected to a

Table 2.1: Latency requirements for telesurgery.

Data type	Reported Latency	Source
2D camera flow	< 150 ms [75] [76]	Experiment [87]
	< 200 ms [77] [65]	Other [77]
	< 700 ms [78]	Experiment [78]
	< 600 ms [79]	Experiment [79]
	< 300 ms [80]	Experiment [80]
3D camera flow	< 150 ms [75] [76]	Experiment [87]
	< 300 ms [81]	Experiment [81]
	< 500 ms [82]	Experiment [82]
	< 400 ms [83] [84]	Simulation[93]
	280 ms [103]	Experiment [103]
	20-50 ms [85]	Other [85]
	2-60 ms [104]	Experiment [104]
	146-202 ms [105]	Experiment [105]
	28 ms [106]	Experiment [106]
	258-278 ms [107]	Experiment [107]
	0.25-5 ms [108]	Simulation [108]
Audio flow	< 150 ms [75] [76] [86] [83]	Experiment [87]
	100 ms [85]	Other [85]
Temperature	< 250 ms [75] [65] [76] [88] [89]	Other [88]
Blood pressure	< 250 ms [75] [89] [65] [76] [88]	Other [88]
Heart rate	< 250 ms [75] [65] [76] [88] [89]	Other [88]
Respiration rate	< 250 ms [75] [65] [76] [88] [89]	Other [88]
ECG	< 250 ms [75] [65] [76] [88] [89]	Other [88]
EEG	< 250 ms [75] [65] [76] [88] [89]	Other [88]
EMG	< 250 ms [75] [65] [76] [88] [89]	Other [88]
Force	3-10 ms [75] [76]	Experiment [92]
	1-10 ms [85]	Other [85]
	3-60 ms [83]	Experiment [94]
	< 50 ms [84] [90]	Experiment [95] & Simulation [90]
	40 ms [84]	Experiment & Simulation [84]
	< 100 ms [91]	Experiment [91]
Vibration	< 5.5 ms [75] [76] [83] [86]	Experiment [92]
	< 50 ms [84]	Experiment [95]
	1-10 ms [85]	Other [85]

PHANToM Desktop, two generated background traffic, and one ran the remote virtual environment. In Platform 2, one of the computers is used to emulate network impairments. Haptic data, network congestion, and network-impairments were analyzed using these two platforms by introducing controlled delay (0 ms to 50 ms), jitter (1 ms to 15 ms), and packet loss (0.1% to 50%). Study participants self-scored the sense of force

Table 2.2: Jitter requirements for telesurgery.

Data type	Reported Jitter	Source
2D camera flow	3-30 ms [75] [65]	Simulation [96] Simulation [93]
3D camera flow	3-30 ms [75] [65] 3-55 ms [108] < 30 ms [85][89] [83] [84] [93] [96]	Simulation [96] Simulation [93] Simulation [108] Other [85]
Audio flow	< 30 ms [89] [75] [83] [84] [65] 50 ms [85] 3-55 ms [108]	Simulation [96] Simulation [93] Other [85] Simulation [108]
Force	< 2 ms [75] [89] [65] [84] 10 ms [85] 1-10 ms [83]	Experiment [95] Simulation [96] Other [85] Experiment [97]
Vibration	< 2 ms [84] [89] [75] [65] 10 ms [85] 1-10 ms [83]	Experiment [95] Simulation [96] Other [85] Experiment [97]

feedback. The haptic QoS requirements were summarized by less than 10 ms delay, less than 3 ms jitter, 1% to 5% for packet loss rate, and haptic data transmission rate of approximately 1 kHz.

The study in [84] involved both simulation and practical experiments where multimodal data is transmitted over a QoS-enabled Internet Protocol (IP) network. The force feedback device was the PHANToM desktop from SensAble Technologies Inc., which could provide force up to 3.3 N in 3 axis directions and generate 1000 packets/s of position and force data during the haptic collaboration actions. In the experiments, the force feedback device was used to manipulate moving virtual objects and to provide the user with feedback from the virtual environment. The end-to-end delay experienced by the haptic traffic was found to decrease from 200 ms (best effort) to 40 ms by running the haptic application in a Differentiated services (DiffServ) network.

To understand the impact of vibration feedback latency, authors in [92] built a system consisting of a liquid crystal display (LCD), touch sensor, rod device with a vibrator, microcontroller, and a host computer. The microcontroller (NXP semiconductors, mbed

Table 2.3: Data rate requirements for telesurgery.

Data type	Reported data rate	Source
2D camera flow	$<$ 10 Mbps [75] [76]	Simulation [96] Experiment [95]
3D camera flow	137 Mbps-1.6 Gbps [75] [76]	Simulation [83]
	$\approx$ 8 Mbps [104]	Experiment [104]
	95-106 Mbps [105]	Experiment [105]
	2.5–5 Mbps [84] [83]	Simulation [96] Experiment [95]
	1 Gbps [85]	Other [85]
	$>$ 1 Gbps [65]	Simulation [83]
Audio flow	22-200 Kbps [75] [76] [84] [83]	Experiment [86]
Temperature	$<$ 10 Kbps [75] [76] [89]	Other [88]
Blood pressure	$<$ 10 Kbps [75] [76] [89]	Other [88]
Heart rate	$<$ 10 Kbps [75] [76] [89]	Other [88]
Respiration rate	$<$ 10 Kbps [75] [76] [89]	Other [88]
ECG	72 Kbps [75] [76] [89]	Other [88]
EEG	84.6 Kbps [75] [76] [89]	Other [88]
EMG	1.536 Mbps [75][76] [89]	Other [88]
Force	128-400 Kbps [75] [76]	Experiment [83] [86]
	500 Kbps–1 Mbps [84]	Simulation [96]
	128 Kbps [83]	Experiment [98]
Vibration	128-400 Kbps [75]	Experiment [83] [86]
	500 Kbps–1 Mbps [84]	Simulation [96]
	128 Kbps [83]	Experiment [98]

NXP LPC1768) controlled the feedback latency from 0.1 to 25.6 ms, according to an adaptive staircase algorithm. 24 participants first sat in front of the touchscreen and were instructed to tap the touchscreen by raising the rod as quickly as possible after the rod head made contact with the touchscreen with an approach velocity of 0.1-0.5 m/s. After the practice, they experienced a 25.6 ms delayed vibration. The participants then conducted eight staircases for further experiments involving two surface conditions (wood or metal). The results showed a 5.5 ms detection threshold of the vibration feedback latency.

Another experimental study proposed a multiplexing scheme that is evaluated using a

teleoperation system consisting of a KUKA light weight robot arm (KUKA Robotics), a JR3 force/torque sensor, a force dimension Omega 6 haptic device [86], and real-time Linux-based Xenomai development software. Using the robot arm, the human operator could move toys and peg them in corresponding holes, which was considered as a representative task for the teleoperation applications. Haptic teleoperation experiments were performed and KPIs considered were varying end-to-end signal latencies (force delay, video delay, audio delay), packet rates, peak delay, convergence time, and peak signal-to-noise ratio (PSNR) for visual quality.

In [94], authors demonstrated an experiment on haptic interaction between two users over a network with 2.4 Gbps connection. Authors used two PHANToM force-feedback devices at both sites; one was located at UCL VECG Lab, London, UK and the second was in MIT Touch Lab, Massachusetts, USA. The experimental subjects were to cooperate in lifting a virtual box together under different conditions.

Authors in [91] proposed a mutual tele-environment system named "HaptoClone", which mutually copies adjacent 3D environments optically and physically using micro-mirror array plates technology. Haptic feedback was also given by using an airborne ultrasound tactile display. Different objects were touched by users and the perceived delay of tactile feedback was measured. Simulations showed that a 100 ms delay was allowable to achieve the real-time interaction.

Other experimental studies using robot systems of SoloAssist (AKTORmed) in Germany, Panda robot (Franka Emika) in Italy, 3D-microscope (Karl Storz) and TiRobot system (Tinavi), and MicroHand (WEGO Group) in China are surveyed in [101].

Simulation based The surgical simulator dV-Trainer from Mimic technologies Inc., Seattle, USA was used in [81, 82]. In [81], sixteen medical students performed an energy dissection and a needle-driving exercise on the dV-Trainer, with latencies varying between 0 and 1,000 ms with a 100 ms interval. These latencies were communication latencies

Table 2.4: Packet loss or bit error rate for telesurgery.

Data type	Reported loss	Source
2D camera flow	$< 10^{-3}$ [75] [76]	Experiment [95] [96]
3D camera flow	$< 10^{-3}$ [75] [76]	Experiments [95] [96]
	$< 1\%$ [84] [83]	Experiments [95] [96] & Simulation [93]
	0.01-0.06% [108]	Simulations [108]
Audio flow	$< 10^{-2}$ [75] [76]	Experiments [95] [96]
	0.01-0.06% [108]	Simulations [108]
	$< 1\%$ [84] [83]	Experiments [95] [96], Simulation [93]
	10^{-5} [85]	Other [85]
Temperature	$< 10^{-3}$ [75] [76]	Other [88]
	$< 10^{-10}$ [89] (BER)	Other [88]
Blood pressure	$< 10^{-3}$ [75] [76]	Other [88]
	$< 10^{-10}$ [89] (BER)	Other [88]
Heart rate	$< 10^{-3}$ [75] [76]	Other [88]
	$< 10^{-10}$ [89] (BER)	Other [88]
Respiration rate	$< 10^{-3}$ [75] [76]	Other [88]
	$< 10^{-10}$ [89] (BER)	Other [88]
ECG	$< 10^{-3}$ [75] [76]	Other [88]
	$< 10^{-10}$ [89] (BER)	Other [88]
EEG	$< 10^{-3}$ [75] [76]	Other [88]
	$< 10^{-10}$ [89] (BER)	Other [88]
EMG	$< 10^{-3}$ [75] [76]	Other [88]
	$< 10^{-10}$ [89] (BER)	Other [88]
Force	$< 10\%$ [84]	Experiments [95] [96]
	$< 10^{-4}$ [75] [76]	Experiments [95] [96]
	0.01-10% [83]	Experiments [95] [96]
	< 0.1 [90]	Experiments [90]
Vibration	$< 10\%$ [84]	Experiments [98] [95] [96]
	$< 10^{-4}$ [75] [76]	Experiments [98] [95] [96]
	0.01-10% [83]	Experiments [98] [95] [96]

from the time that a movement was initiated by the surgeon until the image of the movement is visible on the surgeon's monitor. The difficulty, security, precision, and fluidity of manipulation were self-scored by subjects. It was concluded that the surgical performance deteriorates in an exponential way as the latency increases. This study further concluded that latencies less than 200 ms were ideal for telesurgery; 300 ms was

Table 2.5: Other requirements for telesurgery.

KPI	Reported requirement	Source
Reliability	$1 - 10^{-7}$	[65] [99]
Availability	$1 - 10^{-5}$	[65]
Payload size	Big [1]	[65]
Traffic density	Low [Gbps/km^2]	[65]
Connection density	Low [$/km^2$]	[65]
Service area dimension	10 m x 10 m x 5 m	[65]
Survival time	0 ms	[65]
Range	Up to 200 km 300 km	[85] [65]
Duty cycle for vital signal monitoring	$<$ 1-10%	[89]

also suitable; 400–500 ms may be acceptable; and 600–700 ms was only acceptable for low risk and simple procedures. Surgery was quite difficult at 800–1,000 ms. The same simulator was utilized in [82]. However, in this study, instead of students, 37 surgeons were involved and performed different exercises in an easy-to-difficult order. The dV-Trainer simulator was permitted to introduce fixed latencies into the exercises between the gesture on the grips and the visual feedback on the console. Instead of a self-scoring system as in [81], the dV-trainer in [82] included a built-in scoring system, capturing instrument collisions, drops, etc. This study concluded that although the impact of delay is related to the difficulty of the procedures, but overall, delays of 100 to 200 ms caused no significant impact, delays higher than 500 ms caused a noticeable increase in surgical risk, and surgery became extremely difficult and should be avoided at delays higher than 700 ms.

In [84], following experiments on a testbed, a probability density function (PDF) model of the haptic traffic from a distributed haptic virtual environments (DHVE) application was created for the use in a simulated DiffServ network using OPNET simulation tool. Subsequently, the effect of running the haptic traffic over a DiffServ IP network was obtained. Results indicated that the haptic throughput increases with the increase in the

[1]By big payload, we mean when the packet exceeds 10 Kb [65].

queue scheduling weight.

Another work leveraging a similar testbed, used a force-feedback haptic device in the PHANToM experimental testbed [96]. The set-up involved two computers that were connected through a gigabit Ethernet fiber optic link running on the best effort IP service. The collected network traces from the test network were used to generate statistical models of each type of DVHE traffic that can be used in the standard network simulation packages such as OPNET. The measured network parameters included throughput, packet lost, delay, and jitter. Results from this simulation model showed a close match of simulation network throughputs with experimental throughputs of 850 Kbps and 630 Kbps in asynchronous and synchronous modes, respectively. DHVE effective throughput deteriorated sharply above 90% background load. End-to-end delays of more than 5 ms occurred at above 90% background load. The impact of jitter, latency, and packet loss was studied in [93] using the analytical models, OPNETWORK, and OPNET simulators. For audio, the simulated traffic behavior model was based on two-state (ON-OFF) Markov modulated rate process (MMRP) with the exponentially distributed time at each state. For video, the model was based on K-state MMRP. The QoS requirements for the audio were reported as: delay $<$ 150 ms, jitter $<$ 30 ms, and packet loss $<$ 1%. For video, these requirements were concluded as: delay $<$ 400 ms, jitter $<$ 30 ms, and packet loss $<$ 1%.

Another simulation-based study to investigate the haptic- audio-visual data communication used an interpersonal communication system, HugMe, which consisted of a haptic jacket for a remote person to simulate nurture touching. A haptic device for a local person to communicate his feelings with the remote person, and a depth camera to capture the image and depth information of the remote person and send it back [83].

Several studies citing jitter requirements for telesurgery have referred to the work in [98] that used Image Server and Haptic Handshake applications. The network emulation in [98] consisted of two endpoint computers and a third intervening computer that simulates the network using NISTNet software. The Handshake application is intended to train

students remotely in surgical procedures by placing a haptic device at each endpoint and having the instructor guide the movements of the student remotely. The performance was evaluated under varying packet loss, delay, and jitter. Minimum end-to-end performance requirements for throughput was 128 Kbps, packet loss was less than 10%, delay was less than 20 ms with abrupt movement and less than 80 ms with gentle movement, and jitter was less than 1 ms.

In [90], authors investigated the effect of packet loss and latency in multimodal telepresence systems. The packet loss caused the impression of time delay and influenced the perception of the subsequent events. The simulated haptic feedback force was generated via PHANToM haptic device. The visual 3D environment was presented on a monitor, which was fixed above the haptic device and tilted 80° toward the observer. The visual space was collocated with (i.e., projected into) the haptic space by means of a mirror and participants viewed the mirrored image through a pair of shutter glasses for the stereo image presentation. Visual-haptic event judgment was investigated under packet loss rates of 0, 0.1, 0.2, and 0.3, respectively. The minimum required latency for visual-haptic events was concluded to be 50 ms. Finally, telesurgery reports using software-defined networking (SDN), fog, and cloud infrastructures are described and compared in [108].

Connected ambulance

Table VI summarizes the literature relevant to the connected ambulance use case in terms of the investigated communication KPIs. The literature covers a wide range of applications termed as connected ambulance. In essence, this involves providing medical care enroute to a healthcare facility while exchanging relevant data (e.g., imaging, vital signs, audio, video) with healthcare providers. Requirements for 5G-enabled mobile healthcare in general are discussed in [109], where the authors propose to implement two-way connectivity between ambulances and hospitals across the UK. The KPIs discussed in the paper include the maximum allowed end-to-end latency for different data types (i.e.,

150 ms for camera and audio flow, 250 ms for vital signs, and less than 10 ms for force and vibration). Data rate requirements for different data types were also specified, with the highest data rate requirement being 10 Mbps for two-way visual multimedia streaming, followed by haptic feedback including force and vibration data types with 400 Kbps each, and then audio multimedia stream with a requirement of 200 Kbps. Depending on the required quality and bandwidth constraints the data rate requirements for audio data can vary between 22-200 Kbps. Moreover, different types of vital signs were assigned different data rates, with EEG having the highest requirement of up to 86.4 Kbps [109].

The studies in [77, 65] also highlighted some general requirements for this use case, including 10 ms latency, 2 ms jitter, < 2 ms survival time, $1 - 10^{-5}$ service availability, $1 - 10^{-7}$ reliability, and 0.05 Mbps data rate.

The project "improving treatment with rapid evaluation of acute stroke via mobile telemedicine" (iTREAT) in [110] reported that 93% of connected ambulance cases achieved a minimum 9 min of continuous, live video transmission with a mean mobile connectivity time of 18 min and 87.5% of tests achieved bidirectional audio video quality with ratings of 4 out of 5 or higher, excluding one route with poor transmission quality. The transport routes were 20 min to the University of Virginia Medical Center and 30 test runs were performed. Limitations of this study include manual ratings of the service quality, not explicitly incorporating patient while testing, exclusion of one route with poor coverage conditions, small size of study, and being limited to one region.

Another e-ambulance study used biosensor emulators in a laboratory to mimic biosensor communication behavior and studied KPIs with the varying number of biosensors and payload sizes [111, 112, 113]. Reported outcomes include an upper bound of 250 ms on latency, 0.4 Mbps for average overall throughput, and the success ratio of transmitted samples varied between 97.7% and 99.9%.

A connected ambulance use case was investigated in [114] in the context of proposing a video encoding configuration that jointly optimizes the clinical video quality, time-

varying bandwidth availability, and heterogeneous device's performance capabilities. The proposed model estimated structural similarity quality with a median accuracy error of less than 1%, bitrate demands with the deviation error of 10% or less, and encoding frame rate within a 6% margin.

The study in [115] proposed measurement-based requirements for high-definition ultrasound images (uplink rate > 20 Mbps, downlink rate > 5 Mbps, network delay < 80 ms, jitter < 30 ms), 4K video (uplink rate > 20 Mbps, downlink rate > 20 Mbps, network delay < 50 ms, jitter < 20 ms). Reliability was set to 99.99%, and mobility was of 0-120 km/h. The measured download rate inside the ambulance, which is a user of a 5G private network, reached 1361.21 Mbps, and upload rate reached 257.52 Mbps.

Handling specific patient conditions was also addressed in the context of connected ambulance e.g., prehospital stroke evaluation and treatment [116]. A Prehospital Stroke Study at the Universitair Ziekenhuis Brussel investigated the safety, technical feasibility, and reliability of in-ambulance telemedicine [117]. A total of 43 attempts were made to perform a prehospital teleconsultation of neurological and non-neurological conditions (e.g., strokes, trauma, respiratory, gastro-intestinal, acute pain, intoxication, labor, dysglycemia, vascular disease). The authors concluded that 30 teleconsultations were performed with success rate of 73.2%. Transient signal loss occurred during 6 teleconsultation sessions (14.6%). The time before the connection was re-established varied from 38 seconds to 5 minutes and 47 seconds. Permanent signal losses occurred in 5 teleconsultations (12.2%). The success rates for the communication of blood pressure, heart rate, blood oxygen saturation, glycemia, and electronic patient identification were 78.7%, 84.8%, 80.6%, 64.0%, and 84.2%, respectively. Communication of a prehospital report to the in-hospital team had a 94.7% success rate and prenotification of the in-hospital team 90.2%. Most problems were caused by unstable bandwidth of the 3G/4G mobile network, limited high speed broadband access, and software, hardware, or human error. The study's main limitations include the small sample size, short study duration, and complex observational design. A

TABLE VI: Summary of literature for relevant connected ambulance KPIs.

Use case	KPIs	Data type	Tools	Study
A mobile small cell-based ambulance in the uplink direction in a heterogeneous network.	Latency, data rate, PLR, retainability, spectral efficiency	Ultrasound video	LTE Sim system level simulator	[60]
Project proposal aiming to capture more than 6000 ambulances across the UK provided by 200 different vendors	Latency, data rate, PLR	Ultrasound video, in-ambulance video vital signs, EEG, ECG, force, vibration	Sonography and vital-signs-measuring equipment in ambulances	[53]
Ambulance transporting cardiac patients to hospital	Latency, PLR	12-Lead ECG	Philips standard (basic device model without advanced features such as computer-assisted ECG interpretations), embedded, integrated ECG device in ambulance	[61]
Ambulance transporting stroke patients to hospital	Retainability, bandwidth (mean and maximal upload and download speeds for data transfer), accessibility	Audio-video, blood pressure, heart rate, blood oxygen saturation, glycemia, and electronic patient identification	PreSSUB 3.0 system in ambulance	[62]
Ambulance transporting stroke patients to hospital	Bandwidth (median maximal and average upload download speed)	Audio-video, blood pressure, heart rate, blood oxygen saturation, glycemia, temperature, cardiac rhythm, Glasgow Coma Scale (GCS), and electronic patient identification	PreSSUB 3.0 system in ambulance	[63]
Mobile stroke treatment units for patients with acute onset of stroke-like symptoms	Service restoration time, PLR, latency	CT, audio-video, vital signs	MSTUs with CT system, camera (RP-Xpress; InTouch Health)	[64]
Ambulance transporting cardiac patients to hospital	Retainability, PLR	12-lead ECGs	Rhythm-surveillance and defibrillation equipment	[65]
Ambulance transporting stroke patients to hospital	Throughput, number of reconnections	Audio, video, vital signs	TeleBAT system in ambulance	[66]
Ambulance transporting stroke patients in rural area to hospital	Retainability, reliability	Audio, video,	iPad, Jabber video app, University of Virginia Health System firewall, COR IBR600 LE-VZ; CradlePoint router, 4G Verizon Wireless sim, AP-CW-M-S22-RP2-BL and AP-CG-S22-BL antennas	[54], [67]
Stroke patients in mobile stroke units en route to hospital	Reliability, retainability	Audio, video, ECG, vital signs	MEYTEC GmbH telemedicine systems of Vimed car and Vimed Doc for videoconferencing and teleradiology	[68], [69]
Simulation of mobile ambulance using emulated biosensor data	Latency, average throughput, PLR	Body temperature, blood pressure, heart rate	Data Distribution Service (DDS) middleware, biosensor emulator	[55]–[57]
Connected ambulance prototype study with QoS control in network slicing environment	Uplink/downlink throughput, latency (average per-hop)	Video slices (eHealth, conferencing, surveillance and entertainment)	MEC-based TeleStroke service by SliceNet, NetFPGA cards, SimpleSumeSwitch architecture, LTE eNodeBs, OpenFlow-enabled switches, Software Development Kit (SDK), Dell Edge Gateway, P4 NetFPGA	[70]
Connected Ambulance prototype study in network slicing environment	Average packet loss, latency (round trip time), throughput (frames per second)	Audio, video	eHealth infrastructure at Dell, Ireland, pfSense security, OpenVPN, Dell Edge Gateway series 3003, LTE SIMS, OpenMANO OSM, MEC by SliceNET	[71]
Testing of video encoding framework on ultrasound videos of carotid artery in connected ambulance scenario.	Bitrate, data rate, time-varying bandwidth availability	Ultrasound videos of the common carotid artery	Multi-objective optimization, Philips ATL 5000 ultrasound machine, x265 open source software, Ubuntu 14.04.4 LTS/Linux 64-bit platform	[58]
Prediction of ambulances' future locations to overcome mobility-based challanges	Position accuracy	GPS data	Apache Spark, Spark SQL, algorithms	[72]
Proposition of an architecture for connected ambulance	Uplink/downlink rate, number of device connections, latency, speed, reliability, jitter	Ultrasound image, vital signs, video	Vital signs monitor, ultrasound equipment, video cameras	[59]
Connected ambulance evaluation in network slicing environment using a test platform	Downlink/uplink data rate, uplink latency	Video, CT image, vital signals, medical record	5G customer-premises equipment (CPE) signal transceiver and a 5G user plane function (UPF) gateway service flow forwarding device, medical data acquisition device, MEC cloud computing node	[73]
Ambulance transporting stroke patients to hospital	Reliability, retainability	Audio, video	In-Touch RP-Xpress telemedicine device, Verizon Jetpack 4G LTE mobile hotspot (4620LE) for 4G LTE	[74]
Ambulance transporting stroke patients to hospital	Retainability	Audio, video	VIMED CAR, head and body cameras, specialized microphones	[75]
Report compiled by industry experts and academic researchers based on their studies	Latency, jitter, survival time, communication service availability, reliability, data rate	4K video, audio	Reference given to [21]	[10]

continuation of this study was carried out in [118], that addressed patients with suspected acute stroke and reported median maximal and average upload speeds as 196 Kbps and 40 Kbps, respectively. The download median maximal speed is reported as 407 Kbps and average speed is reported 12 Kbps, using 4G. An experimental study evaluated the use of mobile stroke treatment units (MSTUs) to diagnose and treat 100 residents of Cleveland who had an acute onset of stroke-like symptoms [119]. It was concluded that there were 6 instances of video disconnection, of which 5 were because of an area of poor wireless reception and one was due to the compatibility issue of the devices. No video disconnections lasted longer than 60 s. One limitation pointed out by the authors is the small sample size of this study.

TeleBAT system in [120] used an integrated mobile telecommunications system while transporting patients to the University of Maryland hospital via an ambulance. Results showed feasibility of the case, with number of disconnections resulting from coverage holes, or network switching.

Another case study [121], consisted of a combination of two studies, PrioLTE2 (Reliability of Telemedically Guided Pre-hospital Acute Stroke Care With Prioritized 4G Mobile Network Long-Term Evolution) study and TeDir (TeleDiagnostics in Prehospital Emergency Medicine [Tele-Diagnostik im Rettungsdienst]) study. Remote neurologist in this study rated audiovisual quality. The study in [121] reported high inter-rater reliabilities between the onboard and remote neurologist and 16 out of 18 treatment decisions agreed. Limitations of this study included 12.6% of the teleconsultations not being completed due to the failure of video connection, higher rate of aborted attempts than the previous studies (1% in [119] and 2% in [122]), small number of patients and inclusion of the data from 2 separate studies with different assessment metrics.

A prehospital utility of rapid stroke evaluation using in-ambulance telemedicine (PURSUIT) pilot feasibility study was conducted in [123]. Actors performing pre-scripted stroke scenarios of varying stroke severity were used in live acute stroke assessments. It

is concluded that 80% of the sessions were conducted without major technological limitations. Reliability of video interpretation was defined by a 90% concordance between the data derived during the real-time sessions and those from the scripted scenarios. A previous pilot study, StrokeNET in Berlin, could not conclude assessments because the audio video was lost in 18 out of 30 scenarios [124].

As for cardiac patients, a study published in 2010 [125] demonstrated the transmission of 12-lead electrocardiography (ECG) in an ambulance driving at 50-100 km/h to the cell phone of the attendant emergency medical technician and then to the hospital and to the cell phones of off-site cardiologists using a 3G network, after going through the hospital ECG-processing server. It was concluded that the ECG can be transmitted successfully at the first attempt in all five trials, except in one remote, mountainous ambulance service area. The average transmission time of an ECG report ranged from 91 to 165 s. Interruption of ambulance ECG transmission occurred in up to 27% of transmissions. [126] reported a 1 year study included data from 17 ambulances enroute to Silkeborg Central Hospital (distance ranging from 20-75 km) transmitting 12-lead ECGs and involving 250 patients with the suspected diagnosis of acute myocardial infarction using a GSM network. Results indicated that 86% of prehospital diagnoses were successful. Geographically related transmission problems were the primary reason for failure. Limitations of this study included patient history taking by direct communication between the physician and patient and the lack of a randomized setup.

Mobility is one of the unique features of the connected ambulance use cases and this raises the connectivity issues that can be observed in high-speed moving vehicles (e.g., poor signal quality, multiple handovers, greater occurrences of connection drops, and penetration loss from metallic walls of vehicle). To address these challenges, authors in [127] evaluated data streaming between one ambulance and hospital nodes on the uplink with a small cell inside the ambulance traveling at a speed of 120 km/h. In the simulation scenario, a transceiver was installed on the roof of the ambulance to transmit/receive data

to/from the backhaul macrocell network. The small cell installed inside the ambulance made a wireless connection between the paramedics and the small cell access point (SAP). The SAP and the transceiver were connected through a wired network. The PLR value when using the small cell was reduced to 4.8% compared to 14% in case of 10 users trying to connect to the outside macrocell base station. All 10 users were located in the same ambulance. Throughput also improved by a small amount with the small cell. Authors concluded that using small cell inside the ambulance could be particularly useful in high bandwidth congestion scenarios. Another way to help address mobility challenges can be to predict the future location of the ambulance based on its previous locations as reported in [128]. The authors proposed an algorithm, NextSTMove, which is 300% faster than traditional algorithms and achieved accuracies of 75% to 100%.

Among the 5G features that can enable connected ambulances is network slicing, where logical network resources can be provisioned to accommodate specific application demands. A study conducted in network slicing environment using facilities at the 5G Prototyping Lab at Dell EMC facilities Ireland and SliceNet reported an average round trip latency of 296.91 ms from client to core, an average round trip time of 50.68 ms from client to edge, and an average packet loss of 7.2% for the core and 0.1% at the edge [129]. Another study was carried out in [130] using the same experimental tools with the added features like QoS control based on the data plane programmability and low-latency cloud-based mobile edge computing (MEC) platform. Throughput was evaluated for the coordinated and uncoordinated network slicing strategies and ranged from 0 to 18 Mbps. In QoS-aware slicing, average delay of less than 0.05 ms was observed. However, in non-QoS aware slicing, no guarantee of low latency was given for any network transmission.

Another network slicing system architecture for 5G-enabled ambulance service was tested in the experimental settings with ambulance speed of 30 km/h. Two types of data were considered in this study, video data for remote consultation and uploading of 4.5 GB

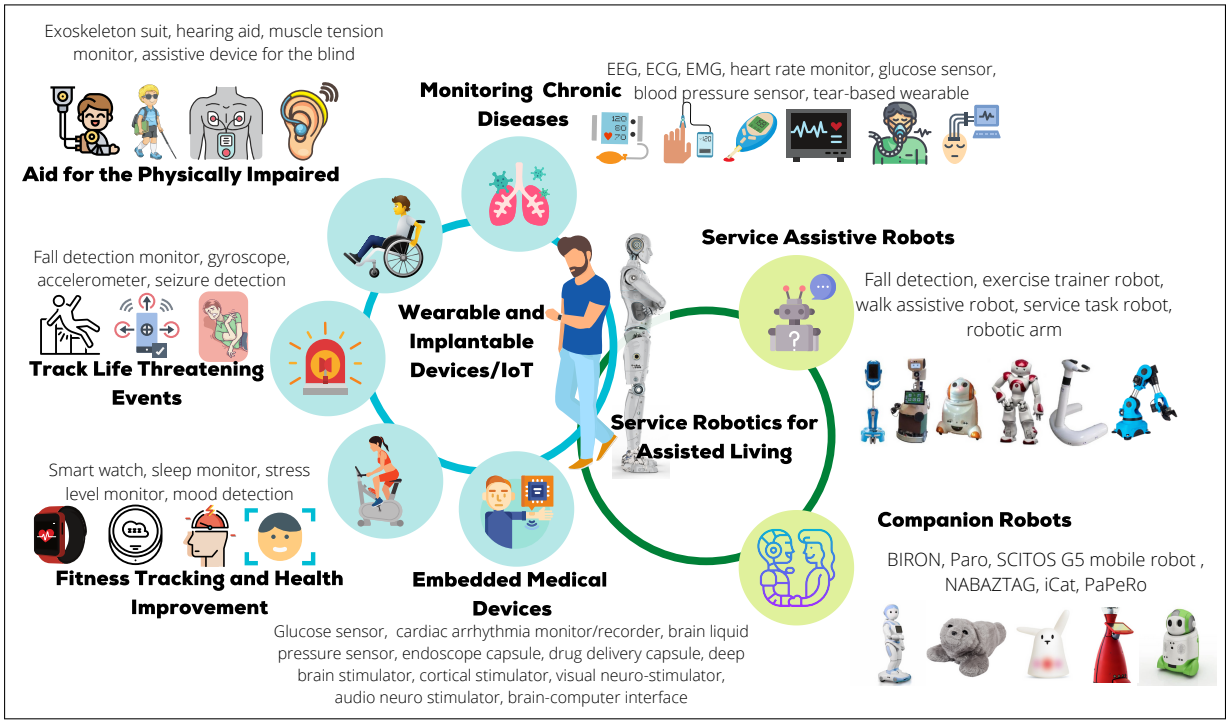


Fig. 2.8: Types of healthcare IoT devices and service assistive robots.

of computed tomography (CT) image data from an ambulance to a destination hospital affiliated with the Zhengzhou University [131]. For video data, the average downlink speed of 1080p 30 Hz HD video in the 5G network environment was 4.6 Mbps, compared to 3.5 Mbps with unstable network and packet loss in 4G. For CT data, the upload time was shortened by 33 percent in 5G as compared to 4G and the average latency for 5G was 12.88 ms, compared to 76.85 ms for 4G which was 6 times that of 5G.

Other relevant studies are ongoing by the groups such as PRE-hospital Stroke Treatment Organization's (PRESTO) [132, 133] and EU 5G PPP Trials working group by SliceNET [134, 135].

Healthcare IoT

Based on the American Society of Engineers, medical internet of things refers to the amalgamation of the medical devices and applications that connect to healthcare information technology systems by leveraging the networking technologies[136]. Healthcare IoT sys-

tems encompass diverse applications and computational capabilities and target diverse populations. Notably, many healthcare IoT systems predate 5G and are being used with 4G and local area wireless technologies like Wi-Fi and Bluetooth. However, 5G can enable an expanded use of healthcare IoT and facilitate the development of novel applications [154]. Accordingly, we dedicate this section to highlighting the wide range of healthcare IoT applications and summarizing their reported communication KPIs. We broadly categorize healthcare IoT systems, which include, medical and non-medical devices, into five types as shown in Fig. 3: 1) fitness tracking and health improvement, 2) chronic disease monitoring, 3) aid for the physically impaired, 4) tracking of life threatening events, and 5) embedded/implantable medical devices.

Applications targeted for the healthy individuals can be used for a wide range of purposes including routine monitoring, lifestyle improvement, or disease prevention, where they act as early warning systems [137]. Examples include smart watches [138, 139] that can monitor heart rate, blood glucose level, blood pressure, and breathing rate. Other fitness and health improvement wearables include temperature sensors [140, 141], pulse oximeter SpO2 [142, 143, 144], sleep trackers [145], fertility and pregnancy trackers [146], and monitors for respiration [147], blood pressure [148, 149, 150, 151], pH [152, 153], stress [154], mood [155], and sleep [156].

Patients with underlying conditions or those who need assisted living in chronic scenarios can benefit from the applications for measuring and reporting electroencephalogram (EEG) [157, 158], ECG [[159, 148, 160], electromyography (EMG) [161, 162] heart rate [163, 164, 165] for cardiac patients, glucose [166, 167], insulin for diabetic patients [168, 169, 170], continuous respiratory rate for chronic respiratory patient [171]. For assisting the physically impaired, there are numerous wearable devices to help improve the quality of life such as hearing aid (ear to ear communication) [172, 173], devices for disability assistance, e.g., muscle tension monitor [174], muscle tension stimulation [175], wearable assistive devices for the blind [176, 177, 178, 179], devices for speech impairment

[180, 181], artificial/wearable limbs [182, 183, 184], and exoskeleton suits [185]. Other examples that can be used by the elderly and Alzheimer or epilepsy patients include wearables for fall detection [186, 187, 188], seizure detection [189, 190], and gyroscopes [191] and accelerometers [192] for localization monitoring. Examples of implantable devices include pacemakers [193] and implantable cardioverter defibrillators (ICD) [194], and implanted actuator [195, 196].

Despite the diversity of healthcare IoT applications, the underlying KPIs requirements are shared by most. However, KPI levels vary for different applications. Following are some of the KPI requirements for this category.

Energy efficiency is vital for the battery-operated devices, where the needed battery lifetime can range from a few days to a few years. Accordingly, battery lifetime can be $>$ 1 week[2] for non-implantable devices, for monitoring ECG, EEG, EMG, glucose, etc. [197]. For implantable devices, this figure can grow to several years (e.g., $>$ 3 years for deep brain stimulator) or remain within the range of hours for some applications like $>$ 24 hr for capsule endoscopes[89]. The importance of battery lifetime increases in implanted devices given the risks associated with the device replacement because of depleted battery. In an attempt to overcome constraints on the battery form factor to accommodate specific implant application, solutions for energy harvesting were considered in the literature that can benefit from the energy present in the environment, human body, and wireless signals [198]. Duty cycle is also relevant in this context, where a lower duty cycle contributes to longer battery lifetime. It captures the tradeoff between the need to timely communicate data and the cost of battery power to do so. The work in [89] reports on duty cycle requirements ranging from $<$ 1% (e.g., temperature sensors, fall detection devices, respiration monitors) to $<$ 50% (e.g., implantable endoscope capsules).

The efficiency of data transmission during the device ON time is described by the data rate with varying requirements according to the application and the used transmission

[2]The life-time numbers are expected/calculated based on normal use conditions for continuous monitoring.

protocol. Literature reports offer a wide array of data rate requirements. For example, [89] reports that monitoring devices for temperature, heart rate, breathing, blood pressure, blood sugar, and oxygenation require $<$ 10 Kbps data rate, 72 Kbps for ECG, 86.4 Kbps for EEG, 1 Mbps for deep brain stimulation and capsule endoscopy, and 1-1.5 Mbps for EMG and location tracking devices [89, 199]. Other references, [200, 201, 197, 202], listed different values including 128-320 Kbps for deep brain stimulators, 3 Kbps per ECG channel per link, and 16 bps for the wearable temperature sensors. Data rate can be influenced by the device processing capabilities, the data use model (i.e., real-time processing by an external processor is associated with demand for a high data rate while applications suitable for post-processing can use a low data rate), and the capabilities of the wireless technology being considered. With the advancement of 5G, literature reports now point to a higher data rate to be supported by wearables (e.g., 10 Mbps [203], 0.1-5 Mbps [65].) Requirements for BER also varied by application and were reported in [204] that generally range from 10^{-10} to 10^{-5}. Specific examples included an ultrasonic wearable device prototype designed to be used as heart rate monitor, ECG respiratory rate monitor, and step counter reported a BER requirement of lower than 10^{-5} using a transmission power of 13 dBm [205]. BER for vital sign monitoring devices such as ECG, pulse oximeters, and implantable devices like hearing aid is reported as $< 10^{-10}$ [89]. To facilitate the diverse healthcare IoT applications, the overall reliability and service availability should be $1 - 10^{-3}$ [65].

Latency requirements also varied across the applications and by the source. [199] reports $<$ 50 ms latency for monitors of chronic disease and emergency event detection. Vital signs monitors were assigned a latency of $<$ 1 s while fitness tracking devices increased latency tolerance to a few seconds. A blanket latency requirement for wearables was set at 250 ms in [65, 89] while survival time was set at 10 ms in [65, 77] and jitter $<$ 25 ms in [65]. Other reported latency values include $<$ 50 ms for deep brain stimulators and $<$ 100 ms for hearing aids [197]. In [206], LTE based data transmission experiments using a real time video wearable device (i.e., BlueEye) under impaired channel loss and

propagation loss were performed. The purpose of the study was to test whether mHealth services could be used in the locations with poor coverage conditions. For different mobility scenarios, the jitter values obtained were 0.473 ms for the static users, 2.05 ms for the pedestrian users, and 3.54 ms for the vehicular users. In an attempt to reduce latency in healthcare IoT applications, significant research was dedicated to data processing and analytics at the edge side of the system to circumvent delays caused by the processing lag and cross network data transfer [207, 208]. In this context, latency of transmitting various raw ECG captures from a gateway to a remote cloud was compared with the total latency of processing on fog computing service and transmitting preprocessed ECG data in [209]. At the data rate of 9 Mbps there was 48.5% latency reduction by leveraging fog computing in this case. This comes at the cost of addressing data security and privacy while in transport between the device and the cloud. To help manage medical device risks including security, a risk management process is specified in the international organization for standardization (ISO) 14971 standard for the application of risk management to the medical devices [210]. Moreover, the FDA published a draft guidance on the content of premarket submissions for the management of cybersecurity in medical devices [211], which provides recommendations to industry regarding cybersecurity aspects of the medical device cybersecurity management, such as risk assessment. Security KPIs in the context of 5G-enabled healthcare applications are summarized in [64], including authenticity, confidentiality, integrity, agility, vulnerability, resilience, mitigation/recovery time, and proactiveness.

Network-level KPIs were addressed in the context of healthcare IoT including a connection density of 20,000 devices/km^2 in remote pervasive monitoring settings such as in smart home wearables and $10,000$ devices/km^2 for general mHealth wearables [77, 65]. Other reported KPIs include 50 Gbps/km^2 traffic density and 50 km user activity range [65]. Given that the healthcare IoT includes diverse applications that can be used in the diverse environments, their enabling KPIs can be influenced by practical deployment factors such as number of nodes, topology, operating frequencies, transmit power restrictions height of

device [212], interference and co-existence [212, 213], and others. Finally, we note that one of the emerging 5G-enabled healthcare applications is medical augmented reality/virtual reality (AR/VR). According to a study by Qualcomm, [214], the requirements for AR/VR can go to as high as 10-50 Mbps for 360° 4K video, 50-200 Mbps for 360° 8K video, and up to 5000 Mbps (or 5 Gbps) for 6 degree-of-freedom (DoF) video. Moreover, a study by Facebook indicates a real-time playback rate of 4 Gbps (or 32 Gbps) for 6 DoF video, indicating there might be some use cases where individual sustained per-user rates of > 1 Gbps might be needed [215].

Robots for assisted living

Robots in assisted living environments have been widely studied in literature [75, 76, 77, 65, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102]. An assistive robot can be defined as an aiding device that has the ability to process the sensory information for helping the physically/mentally impaired or elderly persons to perform tasks of daily living without the need of attendants, in hospital or at home [227]. Assistive robots can be broadly classified into two categories, i.e., services assistive robots and companion robots as shown in Fig. 3. In this section our focus is on the communication KPIs for this application with a summary provided in Table VII of the reported cellular network KPIs.

Position accuracy is pertinent to robots used for fall detection and real-time assistance. The authors in [228] demonstrated that by exploiting the information from the reflected multipath components, increased accuracy and robustness in localization can be achieved. Moreover, they proposed 5G mmWave as one of the promising solutions for indoor accurate localization for assistive living.

According to the EU Horizon 2020 project "Robots in Assisted Living Environments" [229], assisted living considerations include reliability, connectivity, low battery discharge profile, low latency, high communication success rate, and minimum localization error

Table VII: Summary of literature for relevant assistive robots KPIs.

KPI	Service Robot	Assigned Tasks	Target Population	Study
UE battery	Mobile robot BENDER with telepresence capabilities	Assistance in routine tasks and user localization	Elderly	[216]
Latency, PLR	Companion robot	User finding and medication reminder	Elderly	[217]
Latency, data rate	Cloud robot	Monitoring of vital signs	Elderly	[218]
Accessibility, position accuracy	Domestic health assistant Max	Assistance in routine tasks,user searching and following	Healthy elderly	[219]
Throughput (packets per seconds)	Domestic robot DoRo	Video streaming through robot cameras	Elderly and children	[220]
Latency, PLR, position accuracy (mean localization error)	Service robot	Recognition and localization of users	Healthy elderly	[221]
Latency (round trip time), retainability (total service time)	Mobile robot DoRo	Personalized medical support and pre-set reminder event	Elderly people with chronic diseases (multi-morbidity)	[222]
Latency, reliability	Nao, Qbo and Hanson robots	Streaming of teleoperation website	Elderly and children	[223]
Position accuracy	ASTRO robot	Assistance in routine tasks, health related reminders	Healthy elderly	[224]
Position accuracy	Assistive robotic arm	Tablet placement infront of patient	Patients with limited or no mobility	[225]
Position accuracy	Mobile humanoid robot GARMI	Support for household tasks and emergency assistance	Elderly and patients	[226]

with appropriate feedback to support people with limited mobility, who require assistance and companionship.

To provide personalized medical support to the elderly in the presence of several chronic diseases, the authors in [222] designed a hybrid robot-cloud approach. The robot au-

tonomously reached the user with the pre-set reminder events acting as a physical reminder. This case study in DomoCasa Lab (Italy) evaluated the robot (DoRo) based on KPIs such as latency (i.e., round trip time), retainability (i.e, in terms of total service time), along with robot processing time (RPT), average travel time, and mean velocity. Latency over the 20 experimental trials was reported as 56 ms and RPT as 0.012 ms. For the use case where DoRo had to travel 12.6 m to deliver the services with a mean velocity of 0.31 m/s, the total service time was 40.08 s.

The ASTROMOBILE system was evaluated in [224]. The mean path length for the simplest use case (moving in the kitchen) was 9.6 m with a mean velocity of 0.51 m/s, path jerk of 0.023×10^6, and a mean position accuracy error is 0.98 m.

Under the German research project SERROGA, which lasted from 2012 to mid 2015, a companion robot for domestic health assistance was developed [219]. Its services include communication, emergency assistant, physical activity motivator, navigation services, pulse rate monitoring, fall detection, and others. The robot was evaluated in different apartments and labs for a minimum of 29 min and a maximum duration of 255 min, with a velocity range of 0.25-0.27 m/s for distance covered of 355-2600 m. The robot was able to complete the user following tasks with a positioning accuracy of 95%.

A cloud-robotic system for the provisioning of assistive services for the promotion of active and healthy ageing in Italy and Sweden was assessed in [221] on the basis of latency (i.e, round trip time), PLR (i.e, data loss rate), position accuracy (i.e, mean localization error), and localization root mean square error (RMSE) KPIs. The reliability and responsiveness of the cloud Database Management Service (DBMS) was evaluated based on latency, as the time a robot waits for the user position, after a request to the server. The study took place in two sites: smart home in Italy (Domocasa lab) and residential condominium in Sweden (Angen). The mean latency in Domocasa lab was 40 ms, while for the Swedish site it was 134.57 ms. The local host latency acquired during the experimentation was 7.46 ms, and was used as a benchmark. The rate of service failures was less than 0.5%

in Italy, and 0.002% for the Angen site. In Domocasa and Angen, the mean absolute localization error was 0.98 m and 0.79 m, respectively, while the RMSE were 1.22 m and 0.89 m, respectively. On average, the absolute localization error considering the two setups was 0.89 m, and the RMSE was 1.1 m. The use of the presence sensors increased the localization accuracy in the selected positions by an average of 35%.

Assistive living robots domain can suffer from errors caused by the communication connection issues, latency, and spatiotemporal dynamic environment changes. To improve the autonomy and efficiency of robots in smart environment, the authors in [230] proposed a framework for the improvement of the assistive robots performance through a context acquisition method, an activity recognition process, and a dynamic hierarchical task planner. Also, authors in [231] proposed to use full duplex 5G communication for reliable and low latency robot based assistive living.

2.3.2 Requirements for use cases in emerging networks vs current status of network capabilities

5G technology was developed to meet the use cases specified by the International Telecommunication Union (ITU) International Mobile Telecommunications-2020 (IMT-2020). These are enhanced mobile broadband (eMBB), ultra-reliable and low latency communications (URLLC), and massive machine type communications (mMTC). As detailed in the previous sections, many applications can benefit from the communication capabilities of these 5G use cases. A study based on simulation confirmed that the 3GPP 5G system complies with the ITU IMT-2020 performance requirements [232]. 5G trials and commercial deployments are accelerating throughout the world [233, 234, 235]. These show varying levels of performance toward the theoretical goals. For example, 2 Gbps throughput and 3 ms latency were achieved in Austria using spectrum in the 3.7 GHz band [233]. In another 5G trial in Belgium, 2.94 Gbps throughput and 1.81 ms latency were achieved. The peak throughput of 15 Gbps, 5 Gbps, and 4.3 Gbps in 5G trials were also reported

by European network operators Telia, Elisa, and Tele2 Lithuania, respectively [233]. In the U.S., AT&T reported on 5G use cases like video streaming, downloading, and conferencing and achieved upload and download speeds around 1 Gbps [233]. Sprint tested streaming 5G virtual reality systems and 4K video and achieved peak download speeds of more than 2 Gbps using the 73 GHz mmWave spectrum [234]. Verizon achieved 4.3 Gbps speeds by aggregating C-band spectrum with mmWave spectrum in a lab trial [235].

Although, commercial 5G coverage is still limited [236, 237, 238], 5G tests by OpenSignal in 2020 compared services offered by Verizon (mmWave), T-Mobile (mmWave, 600 MHz), Sprint (2.5 GHz), and AT&T (850 MHz) [239]. The report concluded that users should not automatically expect speeds of several hundred Mbps on 5G, as in the tests, they observed an average 5G download speeds ranging from 47.5 Mbps to 722.9 Mbps. They also noted that the U.S. carrier's 5G services are held back by 5G spectrum availability and some services are fast, however, they are limited by the coverage. Those with greater coverage offer slow speeds due to the limited spectrum. They also highlighted the need for the U.S. carriers to repurpose large portions of the mid-band spectrum for 5G in the U.S. to facilitate the 5G performance goals.

Comparing the realistic performance reports with the most stringent data rate requirement for telesurgery (i.e., 1.6 Gbps for 3D camera flow as listed in Table III), we note that the throughput requirements of many healthcare use cases might be possible to meet with current 5G. However, use cases requiring 6 DoF content like AR/VR might be challenging with current 5G capabilities. Furthermore, our review highlights that the latency for the haptic feedback can go as low as 1 ms and for connected ambulance, the lower limit is 10 ms. However, realistic latency figures are expected to remain in the 10-12 ms range [240, 241], rather than 1-2 ms. 5G mmWave frequencies—also known as frequency range 2 (FR2)—can support large subcarrier spacing resulting in smaller transmission time interval and thus improving latency. This indicates a favorable latency requirement support for use cases when using the mmWave spectrum. However, this comes at

the expense of limited coverage due to the wave propagation properties in the mmWave spectrum, which can impact applications that need mobility support like the connected ambulance. Moreover, the realistic deployments and trials are limited by the specific used configurations and the small set of reported KPIs like downlink throughput and latency. Accordingly, enabling a specific healthcare application using 5G requires a collaboration between the application developer, 5G network service provider, and the application user to ensure that the service meets the application requirements for communication and that the application can be used safely.

2.3.3 Gaps in literature and future considerations

Although some reports describe individual KPIs in detail, the trade-offs between multiple KPIs and their interactions with configuration and optimization parameters (COPs) in advanced use cases are often omitted. For example, one trade-off between throughput and latency for next generation video content is described in [214], which states that achieving 5-20 ms latency requires 400-600 Mbps throughput, while achieving 1-5 ms latency requires 100-200 Mbps throughput. Another example of trade-offs is between coverage, capacity, and load balancing [242] or the trade-off between coverage, height of BS, and antenna parameters [243]. Such trade-offs are rarely considered in the literature on 5G-enabled healthcare use cases, which can complicate applications with conflicting requirements like achieving high throughput with high mobility or low battery consumption. One way to study these trade-offs might be to combine several KPIs into a new one. For example, Samsung developed representative KPIs to describe the performance of multi-objective optimization involving more than two KPIs, such as sum of log of data rate, considering both throughput and fairness. It can be used as a joint KPI of wearable devices applications to represent both energy efficiency and throughput, energy efficiency and delay, or energy efficiency and reliability [244].

Another gap in the literature is the limited 5G network scenarios that are assessed.

Limitations include the small number of network trials, small number of infrastructure configurations, small coverage area, and the lack of spatiotemporal variability for trials being conducted in the laboratory settings. A critical analysis of 5G network failure modes that can impact 5G-enabled healthcare use cases is an open question not addressed in the literature. For example, only the success of the connected ambulance use case is discussed in the literature. However, this use case might be negatively impacted in situations with extremely high mobility, high user density, a disaster scenario where a large number of ambulances rush to the same point, a cell outage, or the presence of multiple critical traffic flows in the network.

Moreover, network KPIs are commonly vendor-specific, where each network equipment vendor specifies the performance metrics using its own set of counters and naming conventions. This may give rise to the challenge of managing non-standardized KPIs. The large number of technical counters in the heterogeneous 5G deployments, the use of vendor-specific monitoring tools by the network operators, and the lack of unified data format for collecting and reporting the performance data also pose a challenge for managing the service level agreements between the 5G network operators and the end users of the 5G-enabled healthcare systems [64].

AI can help enable bridge the above highlighted gaps by automating several tasks, such as determination of optimal COP-KPI relationships that can be used for network design, operation and optimization.

2.4 Conclusion

5G and beyond networks will transform many industries by enabling novel use cases and applications, such as telesurgery, remote patient diagnosis, smart medication, and healthcare big data management in the healthcare sector. However, facilitating user access to novel 5G and beyond enabled applications requires that integration of technology takes place safely and effectively to deliver the intended application function. SLAs are a frame-

work for documenting the communication requirements for different use cases. However, current practices for SLAs will not suffice for 5G and beyond networks. In this chapter, tradeoffs, challenges and practical implementation considerations in 5G network resource allocation like provisioning minislots for a specific service, optimal triggering of minislots pre-emption, optimizing device performance when using bandwidth adaptation, network slice sharing modes, and dynamic network resource optimization, UE miss-association probability to mmWave cells are discussed. With increasing network complexity, the need arises for adaptive and automated algorithms to reduce the large set of observable network counters and metrics and facilitate efficient network monitoring for service assurance. Additionally, algorithms are also needed to flexibly map and optimize network configuration parameters to meet desired application objectives while maintaining business objectives for all stakeholders. The heterogenous and multi-domain nature of 5G and beyond network illustrate the opportunity to develop collaboration frameworks to promote interoperability and robust service delivery. To this end, robust artificial intelligence based solutions for designing, operating, optimizing and managing emerging networks can greatly help in the full realization of the emerging networks based advanced use cases.

Understanding the communication KPI requirements for 5G-enabled use cases can help the application developers, 5G network providers, and regulatory authorities in the healthcare sector to promote safe and effective healthcare. Knowledge of requirements for 5G-enabled use cases highlighted in this chapter can also help network service providers, users, and regulatory authorities in developing, managing, monitoring, and evaluating service level agreements in advanced emerging networks use cases. We have also identified gaps in the existing literature and highlight considerations in this space, including the lack of focus on quantitative requirements, omitting relevant KPIs, overlooking the trade-offs between multiple KPIs and COPs, the lack of unified KPI specifications across different network operators and equipment vendors and lastly, the limitations 5G scenarios conducted in the existing trials. The gaps in this space and considerations highlighted in this

chapter can help direct future 5G-enabled application studies towards addressing these gaps to facilitate the efficient implementation of 5G technology in various applications. Moreover, quantitative and qualitative KPI requirements for different use cases, including remote robotic-assisted surgery, mobile connected ambulance, wearable, and implantable devices in the healthcare IoT, and service robotics for assisted living are identified. A comparison of 5G-healthcare requirements with the status of 5G capabilities reveals that some healthcare applications can be supported by the existing 5G services while others might be challenging, especially those with stringent latency requirement. This calls for zero-touch artificial intelligence (AI) based deep automation in cellular networks to meet the diverse and stringent communication requirements for emerging networks use cases.

CHAPTER 3

Enabling AI-driven cellular network optimization by solving the challenges in MDT-based coverage estimation

3.1 Introduction

The first step to enable 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 [245]. 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 [246]. 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 [247]. 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, 248]. 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 [245], [249].

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 [250]. 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 coverage maps [251]-[253]. 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 [254], such as coverage holes, weak coverage spots, sleeping cells [255], or other performance degradation problems and make timely decisions [4], [245]. 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 [256]. 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 dissertation, 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.

3.2 Relevant work

Authors in [246], [247], [256] 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 [257] estimate the cell radius. Moreover, the work in [257] 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 [258]-[259].

Authors in [258] use regression clustering for construction of received signal strength maps from a sparse set of MDT measurements. However, this work [258] assumes perfect user locations and a fixed bin width. The authors in [260] analyze the performance of selected spatial interpolation techniques used in the estimation of interference produced by an LTE-Advanced network. The authors in [261] provide a visualization method based on inverse distance weighted interpolation that shows every point of the received data as a heatmap. Another work [262] 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 [258]-[262], remains unexplored.

Authors in [263], [264] use Bayesian kriging technique on cellular network data to build

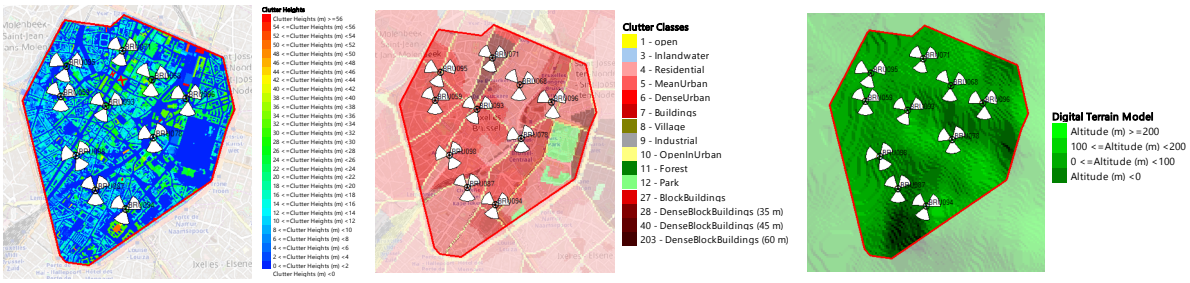
radio environment maps for the purpose of coverage hole detection. They show that the accuracy of such a technique is directly impacted by the bin size. Authors in [265] extend the work in [264] to include a more realistic coverage hole definition, where the coverage of neighboring pixels is also taken into account. Authors in [266] propose a new technique, called Fixed Rank Kriging that is superior in terms of computational complexity as compared to Kriging. Authors in [267] 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 [268] extend this work to a multi-cell scenario. Kriging-based prediction of propagation environment is presented in [269] 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 [270]. However, the works in [263], [264], [265], [266], [267], [268] assume perfect geo-location information.

One work that takes into account the impact of location uncertainty on sparse coverage data is the work in [271], where the authors modify their earlier proposed algorithm [266], [267],[268] 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 [272], [273], [259]. Using low-cost spectrum sensors in an office indoor environment, authors in [272] 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 [273], analyzes several spatial interpolation techniques based on Inverse Distance Weighting (IDW) and

compares them in terms of reliability bounds of interpolation errors. Authors in [259] 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 [259]. 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 uncertainity 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



(a) Digital height model. **(b)** Digital land use map. **(c)** Digital terrain model.

Fig. 3.1: System model configuration and geographical information.

[17]. 4) The work [17] considers a fixed coverage probability threshold. However, in this 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.

3.3 System model

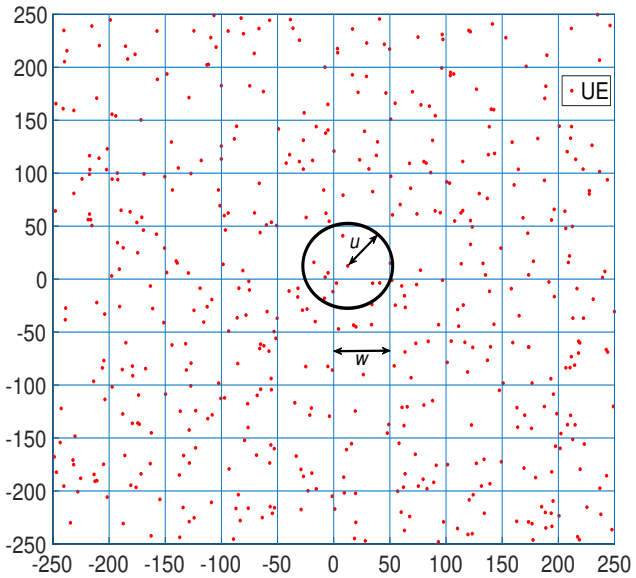


Fig. 3.2: User distribution.

We use a ray-tracing based commercial planning tool [274] to create a sophisticated network topology (Fig. 3.1), 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

Table 3.1: Network Scenario Settings.

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

be assumed that coverage data obtained from this simulator represents the ground truth very closely in the area under consideration.

Users are distributed according to Poisson distribution. The area of interest is divided into n^2 bins of width, w as shown in Fig. 3.2 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. 3.2 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 3.1.

3.4 Quantification of errors in autonomous coverage estimation using MDT

3.4.1 Error due to user positioning uncertainty

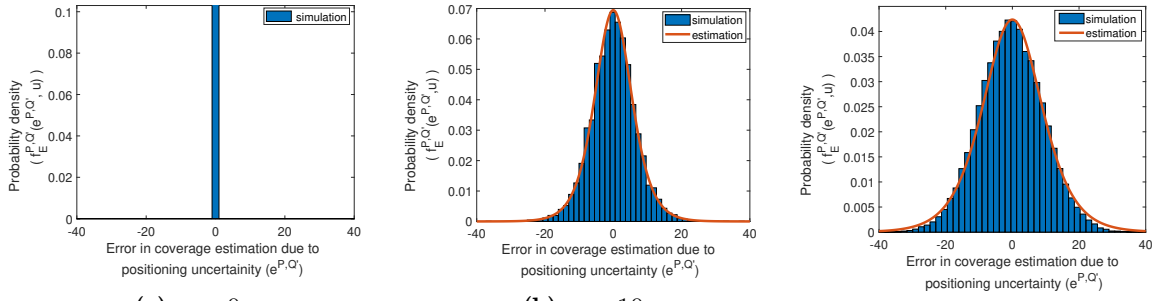
Error due to user positioning uncertainty without bins

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:

$$E^{P,Q'}(x, y, v, q, u) = r^{P,Q'}(x, y, v, q, u) - r^{P',Q'}(x, y) \tag{3.1}$$

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 $w_{min} = 10$m to $w_{max} = 50$m and u is varied from 0m to 100m. Fig. 3.3 illustrates the PDF of coverage estimation error



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

Fig. 3.3: 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:

$$s_1(u) = a_1 \exp(b_1 u) + c_1 \exp(d_1 u), \tag{3.2}$$

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

Fig. 3.4 shows the variation of parameter s_1 with u. The PDF of this error as a function of positioning error radius then becomes:

$$f_E^{P,Q'}(e^{P,Q'}, u) = \frac{\exp\left(-\frac{e^{P,Q'}}{a_1 e^{(b_1 u)} + c_1 e^{(d_1 u)}}\right)}{(a_1 e^{(b_1 u)} + c_1 e^{(d_1 u)})\left(1 + \exp\left(-\frac{e^{P,Q'}}{a_1 e^{(b_1 u)} + c_1 e^{(d_1 u)}}\right)\right)^2} \tag{3.3}$$

Note that the parameters $\{a_1 \ldots d_1\}$ 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.,

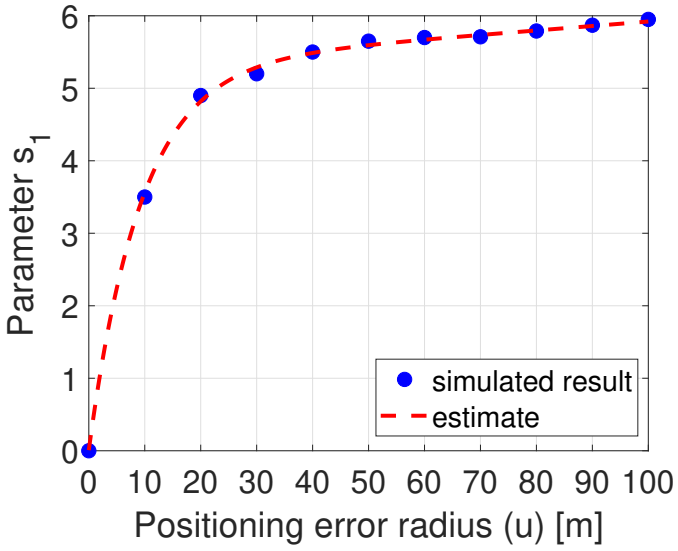


Fig. 3.4: Parameter s_1

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:

$$E^{P,Q}(x, y, v, q, u, w) = r^{P,Q}(x, y, v, q, u, w) - r^{P',Q}(x, y, w) \tag{3.4}$$

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. 3.2. 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:

$$\mathrm{P}_m(w,u) = \begin{cases} 0, & 0 < u \leq w/2 \\ \frac{4u^2\cos^{-1}\left(\frac{w}{2u}\right) - 2u^2\sin\left(2\cos^{-1}\left(\frac{w}{2u}\right)\right)}{\pi u^2}, & w/2 < u < w/\sqrt{2} \\ \frac{\pi u^2 - w^2}{\pi u^2}, & u \geq w/\sqrt{2} \end{cases} \tag{3.5}$$

P_m as a function of u and w is illustrated in Fig. 3.5. 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. 3.5 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

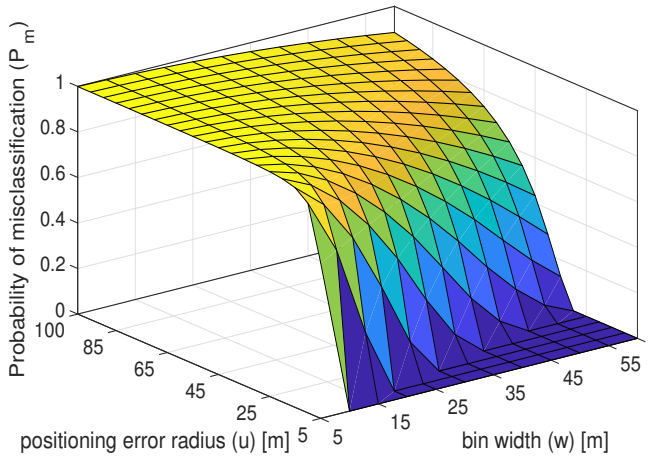


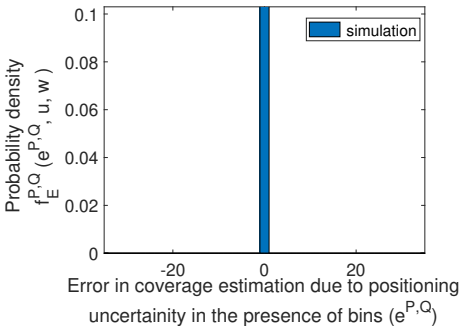
Fig. 3.5: Probability of misclassification of a user with varying bin width and positioning error radius.

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:

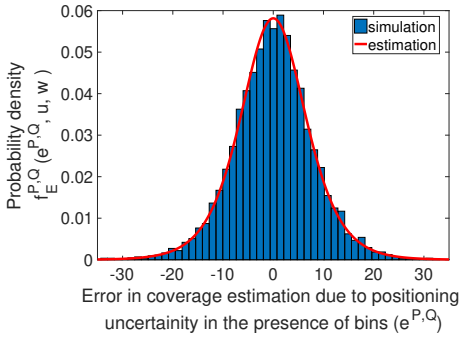
$$E^P = \frac{1}{m^2} \sum_{i=1}^{m^2} |\mathbb{P}[\boldsymbol{r}_i^{P,Q} > \gamma] - \mathbb{P}[\boldsymbol{r}_i^{P',Q} > \gamma]| \tag{3.6}$$

where the operator $\mathbb{P}$ represents probability, $\boldsymbol{r}^{P,Q}$ and $\boldsymbol{r}^{P',Q}$ are vectorized forms of matrices $\boldsymbol{R}^{P,Q}$ and $\boldsymbol{R}^{P',Q}$ respectively. The i-th element of the vector $\boldsymbol{r}^{P,Q}$, $r_i^{P,Q}$ 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.

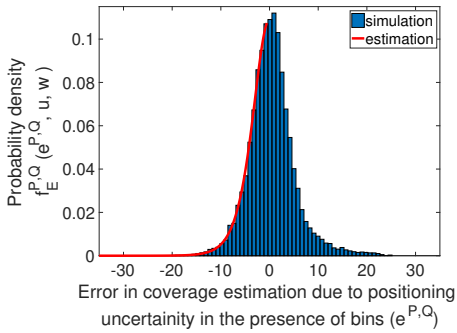
The effect of positioning error on coverage estimation with varying bin widths is shown in Fig. 3.6. It can be observed from Fig. 3.6 (b)-(e) that for the same positioning error radius, the variance of this error decreases as the bin width increases, attributing to the fact that for the same positioning 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



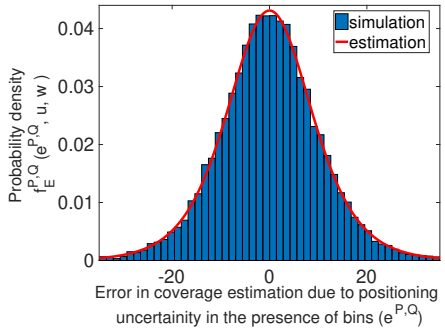
(a) $u = 0$m, any w



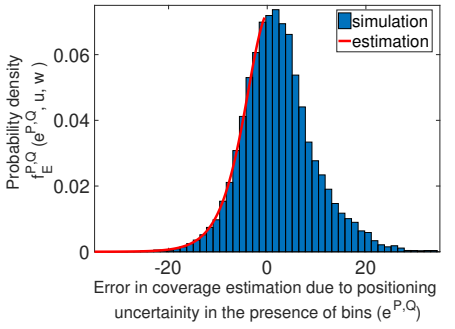
(b) $u = 10$m, $w = 10$m



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



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



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

Fig. 3.6: 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. 3.6 (a). It can also be observed from Fig. 3.6 (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

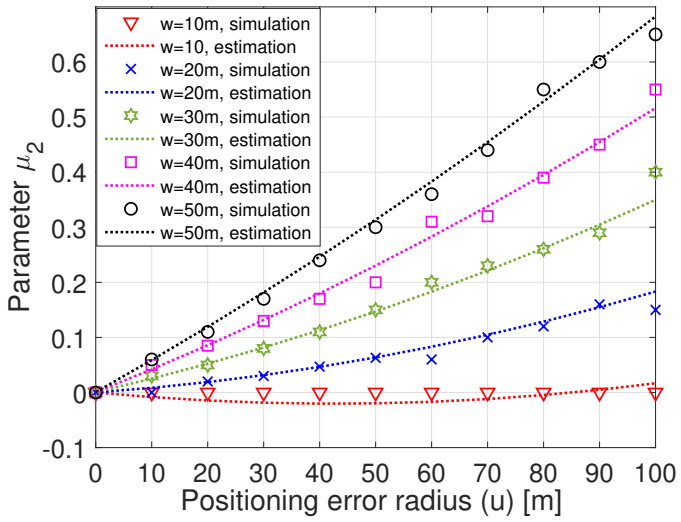


Fig. 3.7: Parameter μ_2

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:

$$f_E^{P,Q}(e^{P,Q},u,w) = \frac{\exp\left(-\frac{e^{P,Q}-\mu_2(u,w)}{s_2(u,w)}\right)}{s_2(u,w)\left(1+\exp\left(-\frac{e^{P,Q}-\mu_2(u,w)}{s_2(u,w)}\right)\right)^2},$$
$$\forall\, e^{P,Q} \text{ when } \mu_2 = 0, \text{ for } e^{P,Q} < 0 \text{ when } \mu_2 \geq 0 \tag{3.7}$$

where μ_2 and s_2 are as follows:

$$\mu_2(u,w) = a_2u + b_2w + c_2u^2 + d_2uw; \tag{3.8}$$

$$s_2(u,w) = (e_2w^{f_2} + g_2)\exp(h_2u/(w+i_2)) + (j_2w^{k_2} + l_2)\exp(m_2wu + n_2u) \tag{3.9}$$

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

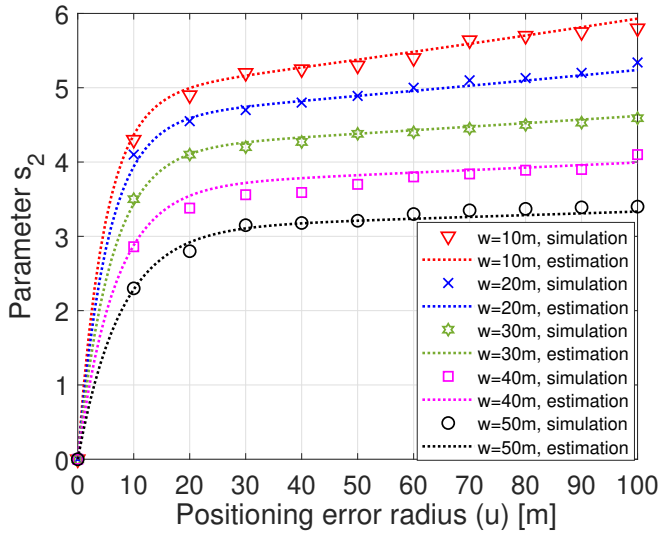


Fig. 3.8: Parameter s_2

The parameters μ_2 and s_2 are shown in Fig. 3.7 and 3.8, 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 3.5.1.

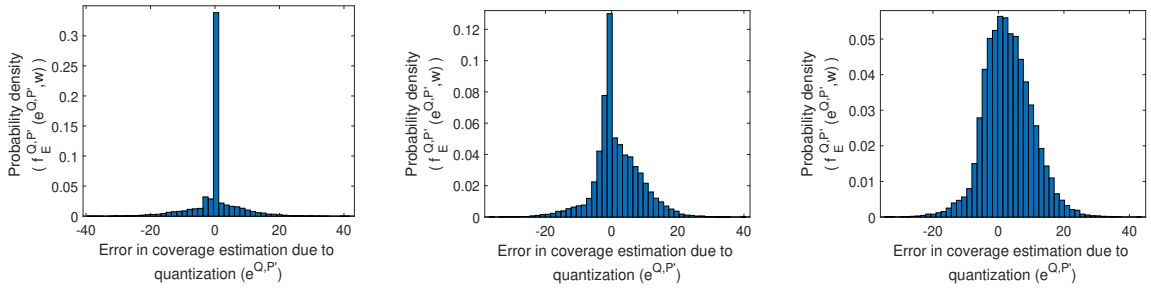
3.4.2 Quantization error

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:

$$E^{Q,P'}(x,y,w) = r^{P',Q}(x,y,w) - r^{P',Q'}(x,y) \tag{3.10}$$

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



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

Fig. 3.9: PDF of coverage estimation error due to quantization error without positioning uncertainty

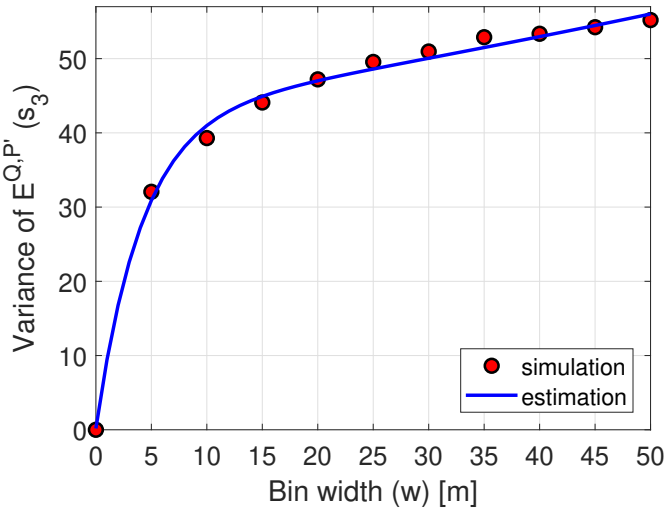
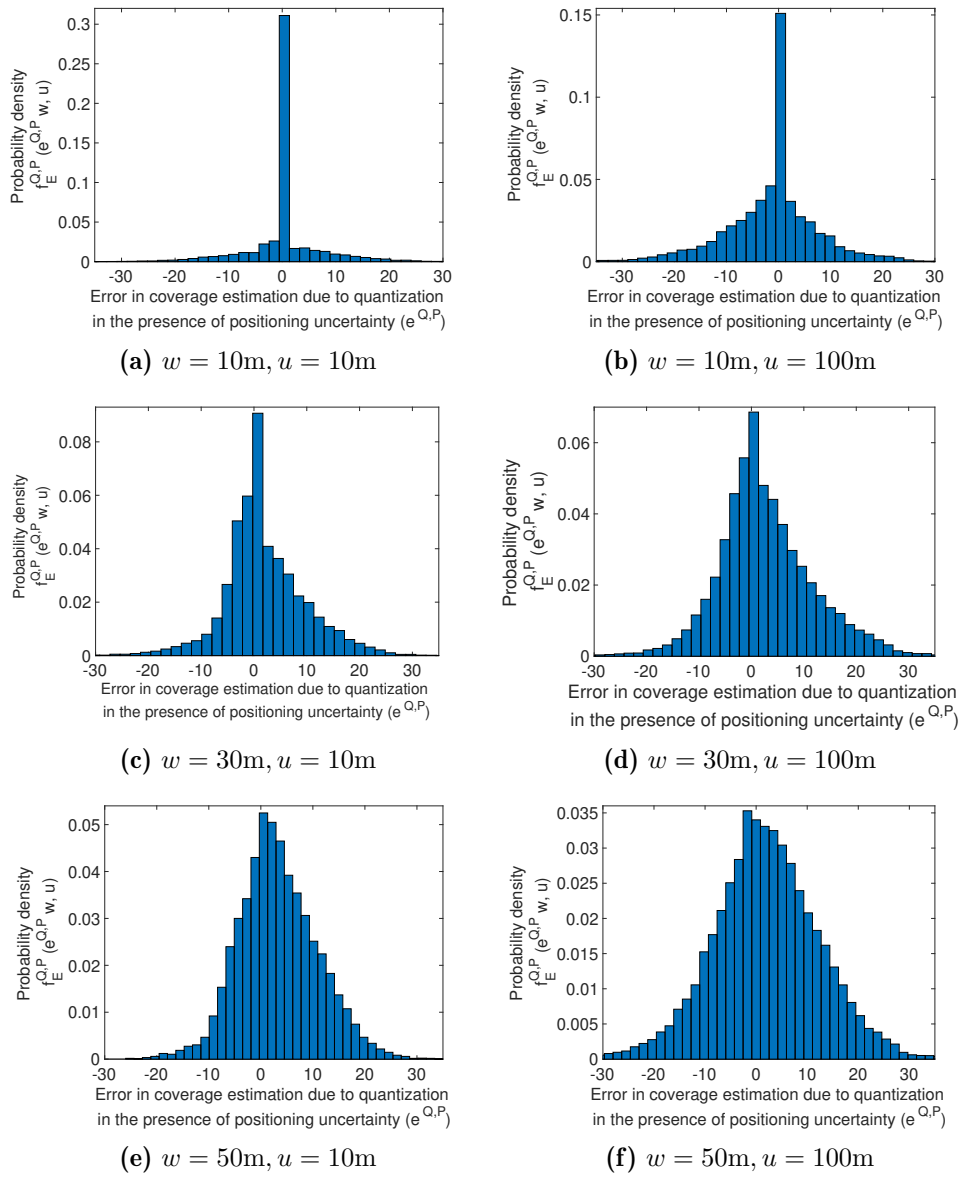


Fig. 3.10: Variance of error, $E^{Q,P'}$

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. 3.9 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 $w \to 0$. The variance of this error, s_3 is depicted in Fig. 3.10 and it increases with increase in bin width according to:



(a) $w = 10\text{m}, u = 10\text{m}$ **(b)** $w = 10\text{m}, u = 100\text{m}$

(c) $w = 30\text{m}, u = 10\text{m}$ **(d)** $w = 30\text{m}, u = 100\text{m}$

(e) $w = 50\text{m}, u = 10\text{m}$ **(f)** $w = 50\text{m}, u = 100\text{m}$

Fig. 3.11: PDF of coverage estimation error due to quantization in the presence of positioning uncertainty

$$s_3(w) = a_3 \exp(b_3 w) + c_3 \exp(d_3 w), \tag{3.11}$$

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

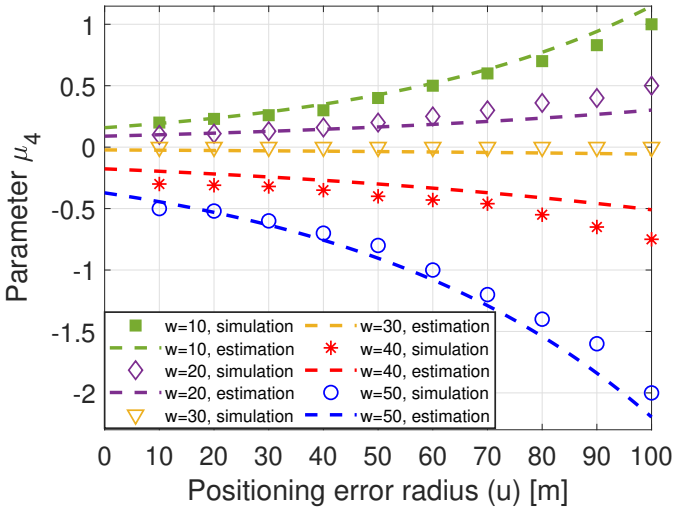


Fig. 3.12: Parameter μ_4

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:

$$E^{Q,P}(x, y, v, q, u, w) = r^{P,Q}(x, y, v, q, u, w) - r^{P,Q'}(x, y, v, q, u) \tag{3.12}$$

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. 3.11 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. 3.11 (a), (c), (e) converge to Fig. 3.9 (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:

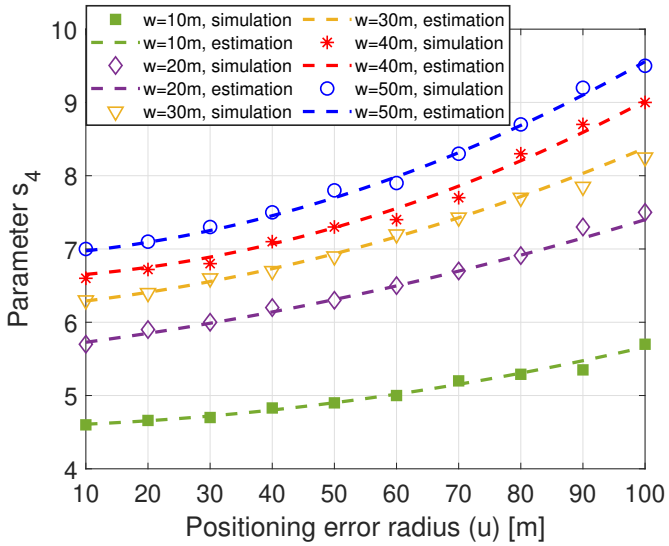


Fig. 3.13: Parameter s_4

$$f_E^{Q,P}(e^{Q,P},u,w)=\frac{\exp\left(-\frac{e^{Q,P}-\mu_4(u,w)}{s_4(u,w)}\right)}{s_4(u,w)\left(1+\exp\left(-\frac{e^{Q,P}-\mu_4(u,w)}{s_4(u,w)}\right)\right)^2}\text{, for } e^{Q,P}>0\;\forall\;\mu_4 \tag{3.13}$$

where

$$\mu_4(u,w)=(a_4w^{b_4}+c_4)\exp\left(d_4u\exp(e_4w)+f_4u\exp(g_4w)\right) \tag{3.14}$$

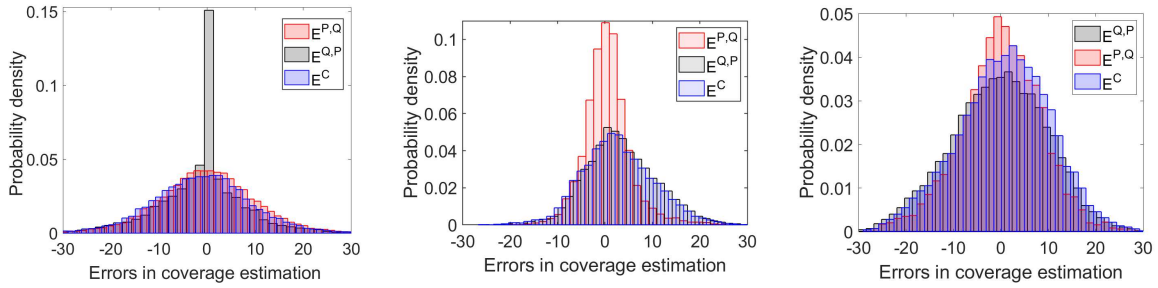
where $a_4=-0.0002718, b_4=1.948, c_4=0.1824, d_4=0.03437, e_4=-0.05875, f_4=0.0003263, g_4=0.07777$

and

$$\begin{aligned}s_4(u,w)=(h_4w^3+i_4w^2+j_4w+k_4)u^2+\\(l_4w^3+m_4w^2+n_4w+o_4)u+p_4w^{q_4}+r_4\end{aligned} \tag{3.15}$$

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 3.12 and 3.13 illustrate the excellent fit between simulated parameters and (3.14)-(3.15).

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



(a) $u = 100$m, $w = 10$m **(b)** $u = 10$m, $w = 50$m **(c)** $u = 50$m, $w = 30$m

Fig. 3.14: 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.

3.4.3 Combined effect of positioning and quantization errors on coverage estimation

In Section 3.4.1, we analyzed the effect of positioning error on coverage estimation, both with quantization and without quantization, while in Section 3.4.2, 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 dissertation, for the first time investigates the concurrent effect of user positioning uncertainty and quantization on coverage estimation. In order to reveal this interplay, we characterize the error in coverage estimation due to both positioning and quantization errors as follows:

$$E^C(x, y, v, q, u, w) = r^{P,Q}(x, y, v, q, u, w) - r^{P',Q'}(x, y) \tag{3.16}$$

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. 3.14 illustrates the PDFs of coverage estimation errors for different bin widths and positioning error radius. In Fig. 3.14 (a), we show the case of large user positioning error ($u = 100$m) and small quantization error ($w = 10$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. 3.14). 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. 3.14). 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. 3.14 (b) shows the case of small user positioning error ($u = 10$m) and large quantization error ($w = 50$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. 3.14 (c) shows the case for $u = 50$m and $w = 30$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. 3.6 and Fig. 3.11, the distributions of E^C in Fig. 3.14 would converge to a delta distribution only when both $u \to 0$ and $w \to 0$ simultaneously.

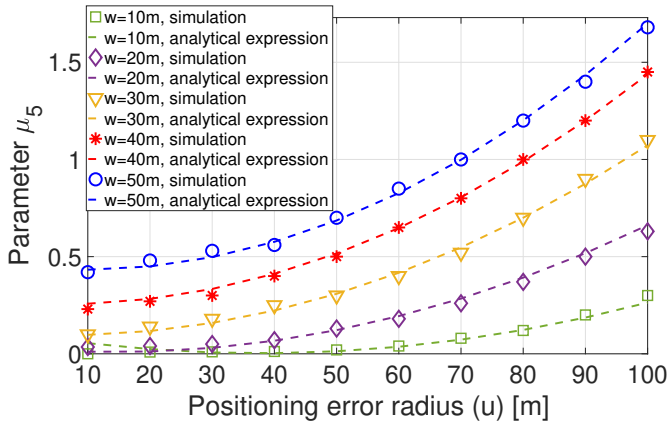


Fig. 3.15: Parameter μ_5

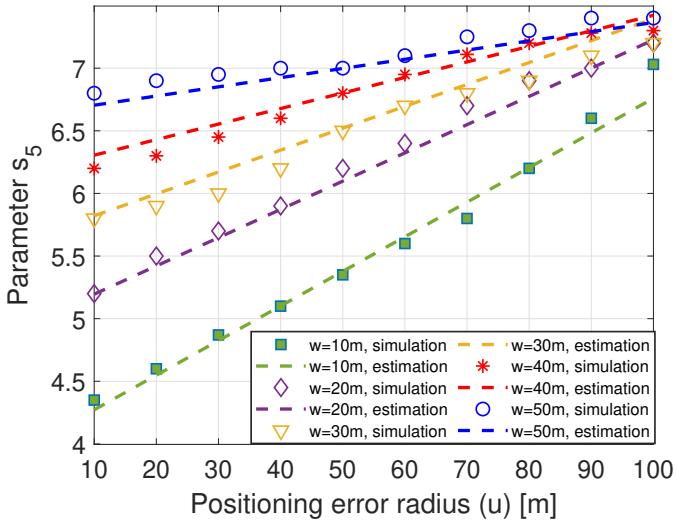


Fig. 3.16: Parameter s_5

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:

$$f_E^C(e^c, u, w) = \frac{\exp\left(-\frac{e^c-\mu_5(u,w)}{s_5(u,w)}\right)}{s_5(u,w)\left(1+\exp\left(-\frac{e^C-\mu_5(u,w)}{s_5(u,w)}\right)\right)^2}, \text{for } e^C < 0 \,\forall\, \mu_5 \tag{3.17}$$

where

$$\mu_5 = a_5 + b_5 u + c_5 w + d_5 u^2 + e_5 uw + f_5 w^2 + g_5 u^2 w + h_5 u w^2 + i_5 w^3 \tag{3.18}$$

$$s_5 = j_5 wu + k_5 u + l_5 w^{m_5} \tag{3.19}$$

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. 3.15 and Fig. 3.16 depict parameters u_5 and s_5 respectively.

3.4.4 Error due to scarcity of data

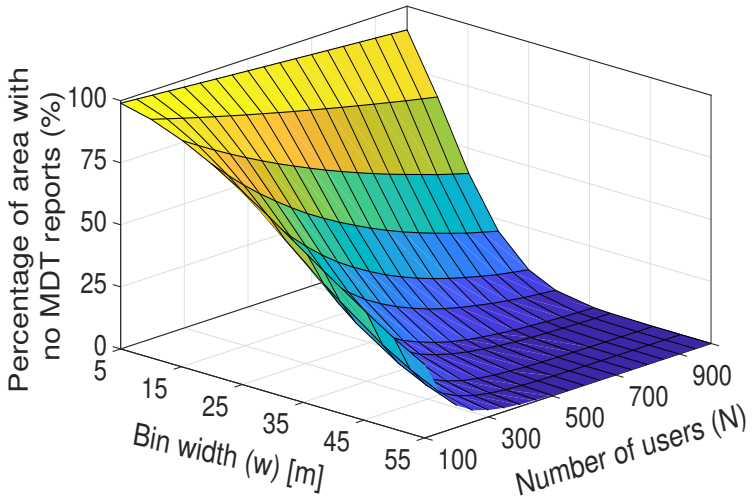
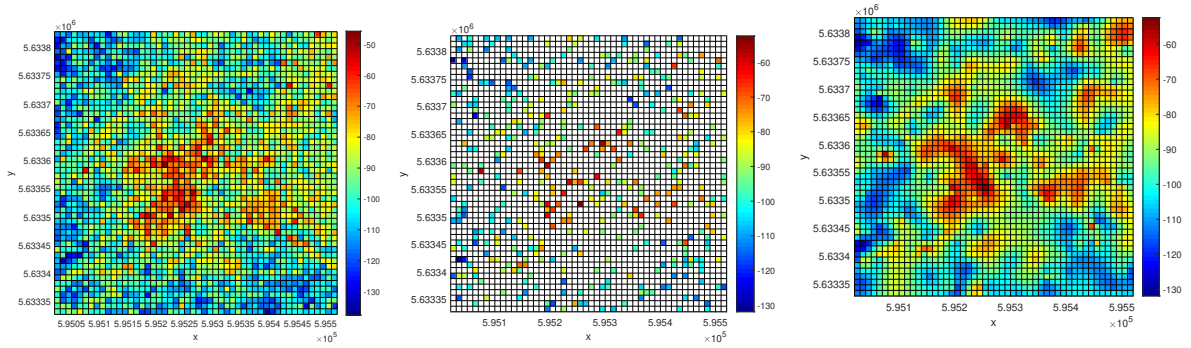


Fig. 3.17: Percentage of area with no MDT reports with varying bin width and number of users

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. 3.17. From Fig. 3.17, 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.



(a) Full coverage map

(b) Sparse coverage map

(c) Recovered overage map with kriging

Fig. 3.18: Kriging for coverage map reconstruction, $u = 0$ and $w = 10$m

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

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

$$E^M = ||\hat{\boldsymbol{C}} - \boldsymbol{C}^{full}||_F / ||\boldsymbol{C}^{full}||_F \tag{3.20}$$

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

Fig. 3.18 illustrates the coverage map for a square area of 500m $\times$ 500m taken from Fig. 3.1, with 500 users. A bin width of $w = 10$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. 3.18 (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.

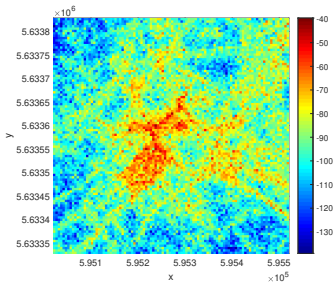
However, this scheme can induce large errors, particularly when two or more adjacent bins are empty. After carrying out a literature review of possible techniques that can be employed in order to recover these missing coverage values, we present a comparison of selected techniques and propose some new 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 4.

Comparison of selected techniques to address data sparsity challenge

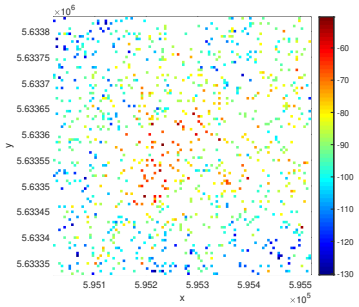
Fig. 3.19 shows a visual comparison of some selected techniques (namely, moving average, matrix completion via different algorithms, inverse distance weighted, nearest neighbor, natural neighbor, spline and kriging techniques) when applied on a sparse coverage map for a bin width of 5m. Details of these techniques are discussed comprehensively in Chapter 4. It can be seen from these figures that kriging interpolation method performs the best. Note that although Fig. 3.19 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 $u = 100$m are shown in Fig. 3.20. 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 geostatistical methods. Therefore, it performs better even in conditions such as large user positioning uncertainty. Hence, we select this technique for the case studies presented in Section 3.5.

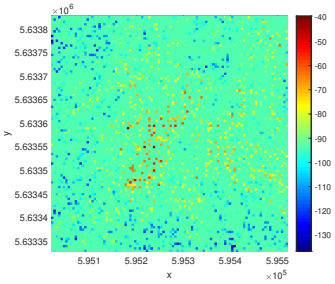
A 3D graph for the matrix recovery error, E^M using Kriging as a function of bin width



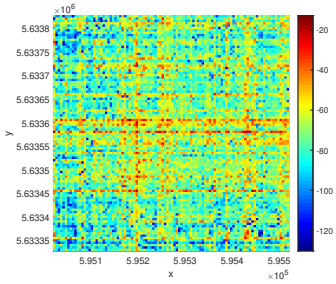
(a) Full coverage map



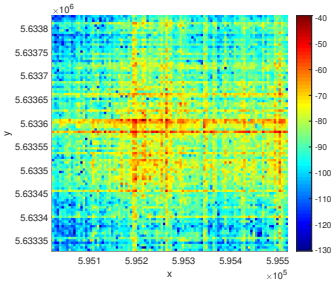
(b) Sparse coverage map



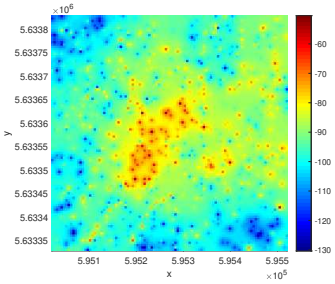
(c) Moving average



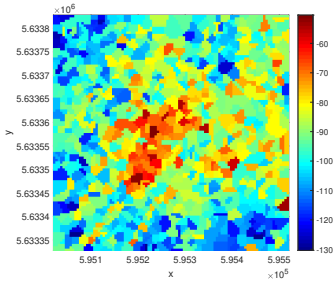
(d) Matrix completion via SVT



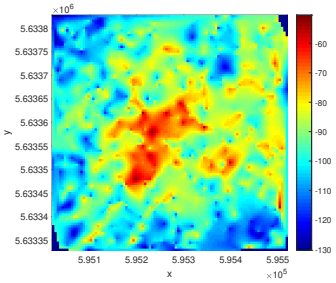
(e) Matrix completion via FPC



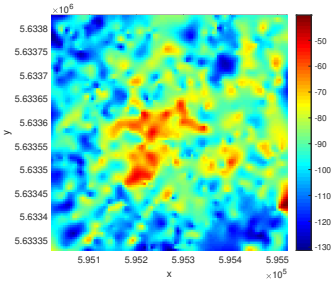
(f) Inverse distance weighted



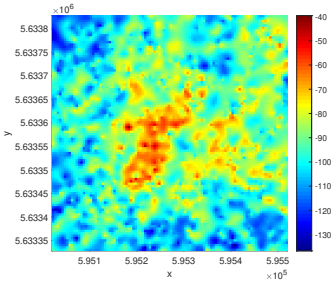
(g) Nearest neighbor



(h) Natural neighbor



(i) Spline



(j) Kriging

Fig. 3.19: Comparison of coverage map reconstruction techniques for $u = 0$ and $w = 5$m

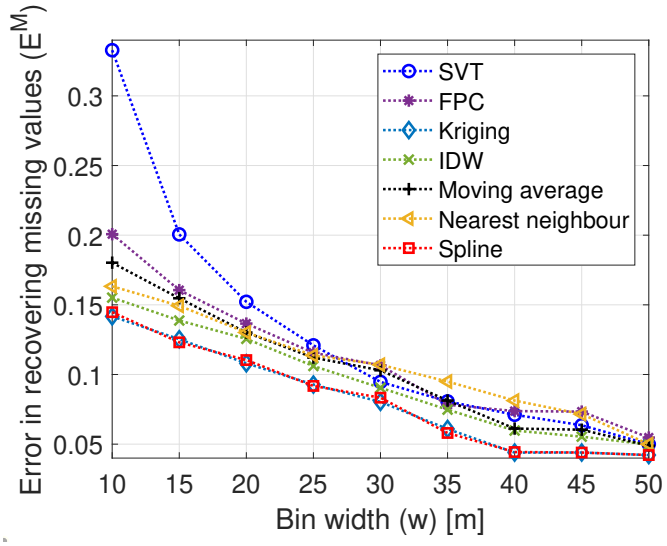


Fig. 3.20: Recovery error with varying bin widths using different reconstruction techniques.

and positioning error radius is shown in Fig. 3.21. It can be seen from Fig. 3.21 that this error increases with increase in positioning error radius and decrease in bin width.

3.5 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.

3.5.1 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.

The probability of area whose coverage is under-estimated due to given positioning uncer-

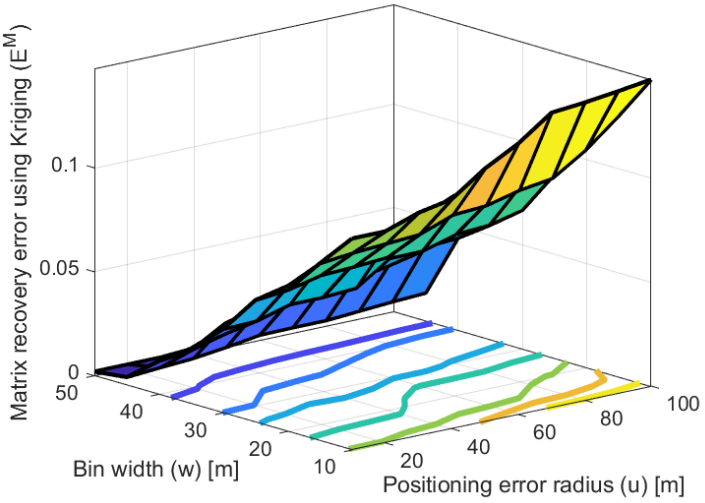


Fig. 3.21: Matrix recovery error with varying bin widths and positioning error radius using Kriging.

tainty, $A_u^P(u,w)$ can be calculated by integrating (3.7) 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 (3.13) from 0 to ∞ as follows:

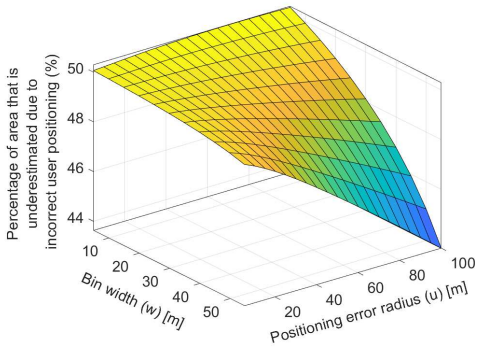
$$A_u^P(u,w) = \int_{-\infty}^{0} \frac{\exp\left(-\frac{e^{P,Q}-\mu_2(u,w)}{s_2(u,w)}\right)}{s_2(u,w)\left(1+\exp\left(-\frac{e^{P,Q}-\mu_2(u,w)}{s_2(u,w)}\right)\right)^2}\, de^{P,Q}$$

$$A_u^P(u,w) = \frac{1}{e^{\frac{\mu_2(u,w)}{s_2(u,w)}}+1} \tag{3.21}$$

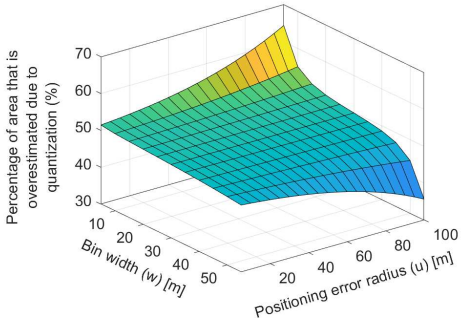
$$A_o^Q(u,w) = \int_{0}^{\infty} \frac{\exp\left(-\frac{e^{Q,P}-\mu_4(u,w)}{s_4(u,w)}\right)}{s_4(u,w)\left(1+\exp\left(-\frac{e^{Q,P}-\mu_4(u,w)}{s_4(u,w)}\right)\right)^2}\, de^{Q,P}$$

$$A_o^Q(u,w) = 1 - \frac{1}{e^{\frac{\mu_4(u,w)}{s_4(u,w)}}+1} \tag{3.22}$$

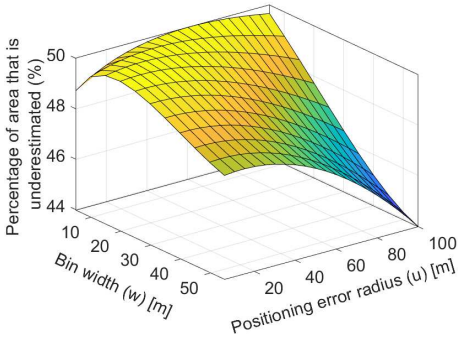
Fig. 3.22a shows the probability of area that is under-estimated due to positioning uncertainty for given bin widths while Fig. 3.22b shows the probability of area that is over-estimated due to quantization for given positioning uncertainties. The probability of area that is under-estimated due to both quantization and user positioning error can be found by integrating (3.17) from $-\infty$ to 0, yielding the expression in (3.23) and illustrated by Fig. 3.22c.



(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. 3.22: Coverage area miscalculated due to different causes.

$$A_u^c(u,w) = \frac{1}{\mathrm{e}^{\frac{\mu_5(u,w)}{s_5(u,w)}} + 1} \tag{3.23}$$

The probability of area that is over-estimated is then $1 - A_u^c(u,w)$ Note that the integral limits of (3.21)-(3.23) 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.

3.5.2 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 (3.4), (3.12) and (3.20) 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:

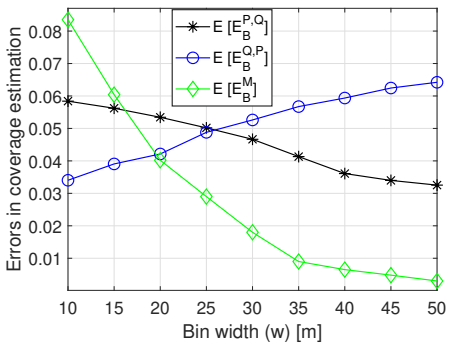
$$E_B^{P,Q} = \frac{1}{n^2} \sum_{i=1}^{n^2} \frac{|\boldsymbol{r}_i^{P,Q} - \boldsymbol{r}_i^{P',Q}|}{|\boldsymbol{r}_i^{P,Q} - \boldsymbol{r}_i^{P',Q}| + |\boldsymbol{r}_i^{P',Q}|} \tag{3.24}$$

$$E_B^{Q,P} = \frac{1}{U} \sum_{i=1}^{U} \frac{|\boldsymbol{r}_i^{P,Q} - \boldsymbol{r}_i^{P,Q'}|}{|\boldsymbol{r}_i^{P,Q} - \boldsymbol{r}_i^{P,Q'}| + |\boldsymbol{r}_i^{P,Q'}|} \tag{3.25}$$

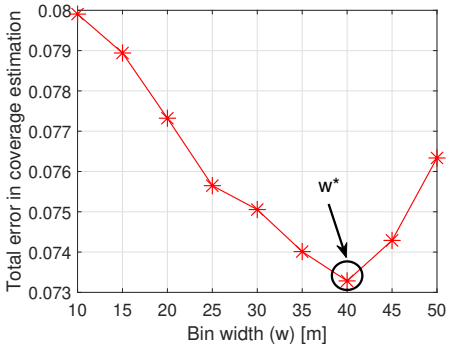
$$E_B^{C} = \frac{1}{U} \sum_{i=1}^{U} \frac{|\boldsymbol{r}_i^{P,Q} - \boldsymbol{r}_i^{P',Q'}|}{|\boldsymbol{r}_i^{P,Q} - \boldsymbol{r}_i^{P',Q'}| + |\boldsymbol{r}_i^{P',Q'}|} \tag{3.26}$$

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

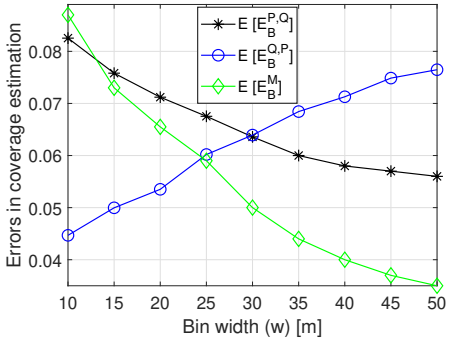
Similarly, a bounded measure for matrix recovery error (this can be considered analogous



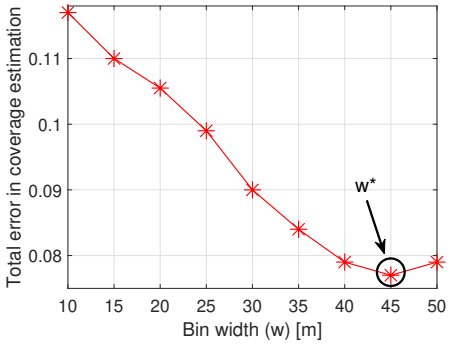
(a) Individual errors, $u = 10$m



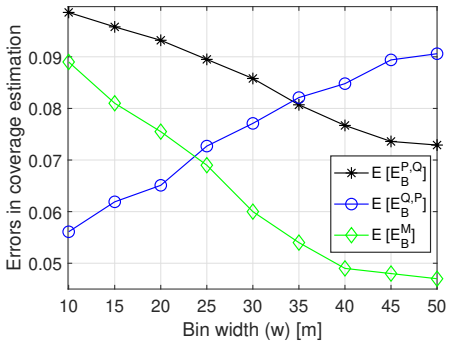
(b) Total error, $u = 10$m



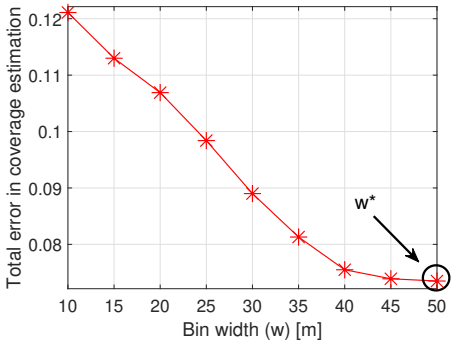
(c) Individual errors, $u = 60$m



(d) Total error, $u = 60$m



(e) Individual errors, $u = 100$m



(f) Total error, $u = 100$m

Fig. 3.23: Different errors in coverage estimation, leading to optimal bin widths.

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

$$E_B^M = \frac{1}{n^2} \sum_{i=1}^{n^2} \left(\frac{|\hat{\boldsymbol{c}}_i - \boldsymbol{c}_i^{full}|}{|\hat{\boldsymbol{c}}_i - \boldsymbol{c}_i^{full}| + |\boldsymbol{c}_i^{full}|} \right) \tag{3.27}$$

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

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

$$w^* = \arg\min_{w} \ \mathbb{E}\left(E_B^M + E_B^C\right) \tag{3.28}$$

$$\text{subject to} \quad w_{min} \leq w \leq w_{max} \tag{3.29}$$

$$\text{positioning error radius} = u \tag{3.30}$$

Owing to the small search space, we can solve (3.28)-(3.30) via brute force.

The quantization error, error due to incorrect user positioning and error due to sparse user reports is shown in Fig. 3.23 (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. 3.17, it becomes difficult to recover the missing coverage values as bin width decreases. Finally, Fig. 3.23 (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.

3.6 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 4

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

4.1 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 [275]-[280] 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 [281]-[283]. These COP-KPI relationships can then be used for COP-KPI optimization. Moreover, in cellular networks context, awareness

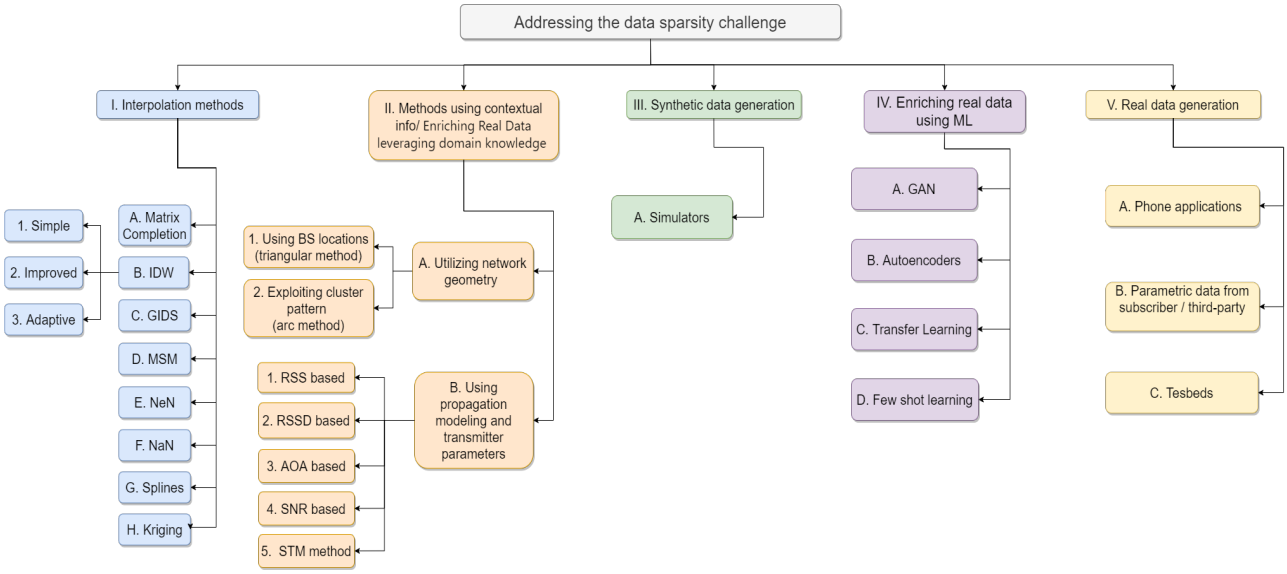


Fig. 4.1: Methods to address data sparsity challenge and chapter organization.

about radio environment in a wireless system is crucial given that the radio spectrum is a limited resource [284]. 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 [285]-[286]. 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 4.2 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 contexual information or domain knowl-

edge in Section 4.3. Several machine learning techniques can also be leveraged to address the data sparsity challenge. These include generative adversarial networks, autoencoders, transfer learning and few shot learning techniques (Chaper 4.5). On the contrary, when the available data is big and non-representative or sparse and non-representative, the solution lies in either resorting to generate synthetic data (Chapter 4.4) or get real data (Chapter 4.6). 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 most likely going to be the best option for wireless communications community. An overview of this chapter organization is presented in Fig. 4.1.

4.1.1 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. [287]-[290] . Authors in [287], [289], [290] 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 [285], [291], [292], [293] or radio environment map reconstruction [294], [273], [295], [296].

In the domain of cognitive radio networks, authors in [285] 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 [285]. Authors in [291] conclude that Kriging performs best among nearest

neighbor and inverse distance weighted (IDW) methods. Results in [292] 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 [293], 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 [294]. Advantages, disadvantages and asymptotic complexity comparison of seven different direct (interpolation) techniques and three indirect construction methods is provided. Extending the study in [294] 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 [297]. Authors in [295] compare Kriging, Modified Shepard's method (MSM) and Gradient plus inverse distance squared (GIDS) and IDW for creating radio enviromment 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 [298] 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 [273] 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 [273]. A more practical implementation of Kriging based approach using real data from the University of Colorado, Boulder campus has been demonstrated in [296].

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.

4.2 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 [299].

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 $\boldsymbol{C}$ of dimensions $n \times n$. Thus, the coverage area forms a square matrix $\in \mathbb{R}^{(n \times n)}$, where each entry is located at the i-th row and j-th column. Following the time window for gathering measurements and updating the coverage map, it is possible that values are available in only m random bins where $m < n \times n$ such that $\{C_{ij} : (i, j) \in \Psi\}$ and Ψ is a set of cardinality m sampled at random. A comparison of selected reconstruction techniques for a particular scenario, in which coverage map for a square area of 500m $\times$ 500m with 500 users and bin width of $w = 10$m forms a coverage matrix of dimensions 50×50, leading to 81.1 % missing entries in the matrix, is shown in Fig. 3.18 (b). Different recovery techniques are then applied on this matrix and the resulting outputs are shown in Fig.3.18 (c)-(j). Each technique has its own set of advantages and disadvantages; we elaborate these techniques in this section.

4.2.1 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 $\boldsymbol{C}$. We note that this scheme is likely to work well in small cells environments since matrix $\boldsymbol{C}$ 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 $\boldsymbol{C}$:

$$\begin{aligned} &\text{minimize} \quad \text{rank}\{\boldsymbol{P}\} \\ &\text{subject to} \quad P_{ij} = C_{ij} \quad (i,j) \in \Psi \end{aligned} \tag{4.1}$$

However, the problem in (4.1) 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 [300]. However, the analysis presented in [300] proves that the coverage values in vacant bins can be obtained with high accuracy by solving the following alternate convex optimization problem:

$$\begin{aligned} &\text{minimize} \quad ||\boldsymbol{P}||_* \\ &\text{subject to} \quad P_{ij} = C_{ij} \quad (i,j) \in \Psi \end{aligned} \tag{4.2}$$

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

$$||\boldsymbol{P}||_* = \sum_{k=1}^{n} \sigma_k(\boldsymbol{P}) \tag{4.3}$$

In (4.3), $\sigma_k(\boldsymbol{P})$ denotes the kth largest singular value of $\boldsymbol{P}$. (4.2) therefore aims to determine the matrix with minimum nuclear norm that fits the data.

The problem in (4.2) can be solved with the singular value-based threshold (SVT) algorithm presented in [301]. The SVT algorithm solves the following problem:

$$\begin{aligned} &\text{minimize} \quad \eta||\boldsymbol{P}||_* + \frac{1}{2}||\boldsymbol{P}||_F^2 \\ &\text{subject to} \quad \mathcal{O}_\Psi(\boldsymbol{P}) = \mathcal{O}_\Psi(\boldsymbol{C}) \end{aligned} \tag{4.4}$$

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 $\mathcal{O}_\Psi(\boldsymbol{P})$ is equal to P_{ij} if $(i,j) \in \Psi$ and zero otherwise. It is shown in [301] that the solution of the problem of (4.4) converges to that of (4.2) as $\eta \to \infty$. The SVT algorithm is iterative and produces a sequence of matrices $\{\boldsymbol{P}, \boldsymbol{Q}\}$. At each step, a soft-thresholding operation is performed on the singular values of the matrix $\boldsymbol{Q}^t$. Thus, by selecting a large value of the parameter, η in (4.4), the sequence of iterates, $\{\boldsymbol{P}^t\}$ converges to a matrix which nearly minimizes (4.2). Starting with $\boldsymbol{Q}^0 = \boldsymbol{0} \in \mathbb{R}^{(n \times n)}$, the algorithm inductively defines

$$\boldsymbol{P}^t = \text{shrink}(\boldsymbol{Q}^{t-1}, \eta) \tag{4.5}$$

$$\boldsymbol{Q}^t = \boldsymbol{Q}^{t-1} + \Delta_i \mathcal{O}_\Psi(\boldsymbol{C} - \boldsymbol{P}^t) \tag{4.6}$$

where $\{\Delta_i\}, i \geq 1$ is a sequence of scalar step sizes, until a stopping criteria is reached. The shrink function in (4.5) applies a soft-thresholding rule at level η to the singular values of the input matrix. It is defined as

$$\text{shrink}(\boldsymbol{Q}^{t-1}, \eta) = \mathcal{S}_\eta(\boldsymbol{Q}^{t-1}) := \boldsymbol{U}\mathcal{S}_\eta(\boldsymbol{\Sigma})\boldsymbol{V}^* \tag{4.7}$$

$$\mathcal{S}_\eta(\boldsymbol{\Sigma}) = \text{diag}(\{(\sigma_k - \eta)_+\}) \tag{4.8}$$

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

$$\boldsymbol{P} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*, \quad \boldsymbol{\Sigma} = \text{diag}(\{\sigma_k\}), 1 \leq k \leq r \tag{4.9}$$

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

$$||\mathcal{O}_\Psi(\boldsymbol{P}^t - \boldsymbol{C})||_F^2 \leq (1 + \zeta) m \phi^2 \tag{4.10}$$

where ζ is a fixed tolerance. The SVT algorithm is stopped as soon as $\boldsymbol{P}^r$ is consistent with the data and obeys (4.10). Therefore, our reconstruction matrix, $\hat{\boldsymbol{C}}$ is the first $\boldsymbol{P}^t$ obeying (4.10). This algorithm is outlined in Algorithm 1.

Another similar rank minimization based algorithm used to recover the matrix $\boldsymbol{C}$ is the fixed point continuation (FPC) algorithm [302]. 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 [302]. To solve this problem, FPC-based algorithm is proposed in [302]. FPC-based algorithm has some similarity with the SVT algorithm in that it makes use of matrix shrinkage as in (4.5)-(4.8). However, it solves (4.4) by leveraging operator splitting technique [303].

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

Input : sampled set Ψ and sampled entries $\mathcal{O}_\Psi(\boldsymbol{C})$, 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: $\boldsymbol{P}^{opt}$

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

4.2.2 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 $\boldsymbol{C}$, weighted average of N known signal strength values, c_k from N adjacent bins are used, where $k = 1 \ldots N$. 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:

$$\hat{c} = \begin{cases} \frac{\sum_{k=1}^{N} \frac{1}{d_k^p} c_k}{\sum_{k=1}^{N} \frac{1}{d_k^p}} & \text{if } d_k \neq 0 \\ c_k & \text{if } d_k = 0 \end{cases} \tag{4.11}$$

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 [304]. When $p = 0$ in (4.11), 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 [294],[304].

However, the advantages of simple IDW method include the 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 [305], [304] and [306] 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 [304] 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 $i, j = 1 \ldots K$. 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 [304] 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], [304].

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 [307] suggested an improvement based on the weighted median of data in the neighborhood of missing data point. The weighting function in [307] 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 [308] 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 [306]. 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.

Table 4.1: Improvements to IDW interpolation

Improvement	References
Intelligent selection of data in neighborhood	[305], [304], [306]
Addition of directionality	[304]
Reduction of computational complexity	[305],[304],[306], [1], [304]
Reduction of computational errors	[1], [304], [308]
Addition of desirable features	[1], [304]
Extension to clustered/non-uniformly distributed data	[307]
Addition of temporal dimension	[306], [309]
Reduction of "bulls-eyes" effect	[306] [294]

Authors in [306] and [309] 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 [306], 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 [306]. Authors in [309] 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 [309].

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:

$$R = r_o/r_e, \quad r_e = \frac{1}{2(M/A)^{0.5}} \tag{4.12}$$

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. 4.2. 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 outperfom 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].

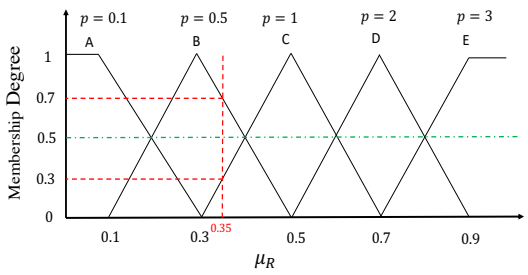


Fig. 4.2: Triangular membership function for different adaptive distance-decay parameters (modified from [1]).

4.2.3 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, 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 [295]:

$$\hat{c} = \frac{\sum_{k=1}^{N} \left(c_k + C_x(x - x_k) + C_y(y - y_k) + C_z(z - z_k)\right) / d_k^2}{\sum_{o=1}^{N} 1/d_o^2} \tag{4.13}$$

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 [294].

4.2.4 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 [295], [273]:

$$\hat{c} = \frac{\sum_{k=1}^{N} W_k Q_k}{\sum_{k=1}^{N} W_k} \tag{4.14}$$

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

$$W_k = \begin{cases} [R_w - d_k]/R_w d_k]^p & \text{if } d_k < R_w \\ 0 & \text{if } d_k \geq R_w \end{cases} \tag{4.15}$$

Then, another radius, R_v around each known data point is considered and the weights are again calculated using (4.15), 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_q and p. It is also shown to perform poorly if measurements lie in a low-dimensional subspace [294]. However, this method can reduce the "bull's eye' effect as compared to classical IDW methods.

4.2.5 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 $\min\{d(D, D_k)\} = d(D, D_l)$, $k = 1 \ldots N$. 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 [292]:

$$\lambda_k = \begin{cases} 1 \text{ if } k = l \\ 0 \text{ if } k \neq l \end{cases} \tag{4.16}$$

which leads to the missing point prediction as:

$$\hat{c} = \sum_{k=1}^{N} \lambda_k c_k = c_l \tag{4.17}$$

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 [294], [310].

4.2.6 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 [285]:

$$\lambda_{\mathbf{D}_k}(\mathbf{D}) = \frac{l_{\mathbf{D}_k}(\mathbf{D})}{\sum_k l_{\mathbf{D}_k}(\mathbf{D})} \tag{4.18}$$

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 [294].

4.2.7 Splines

The spline method is also referred to as the radius basis function and "rubber sheeting' [294]. 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 [285]:

$$\hat{c} = \sum_{k=1}^{N} w_k ||D - D_k||^2 \ln(||D - D_k||) \tag{4.19}$$

where $||.||$ is the Euclidean norm. w_k can be obtained by solving $\mathbf{Ow} = \mathbf{i}$, 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 [299]. Moreover, solving the system of linear equations in order to determine weights adds to the computational complexity of this method.

4.2.8 Kriging

Kriging, unlike the other methods discussed above, also takes into account the statistical relationships in additional 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 [311]. The weight coefficients are calculated by minimizing the variance of the estimation error, σ_e^2:

$$\sigma_e^2 = \mathbb{V}\left[\hat{C_m} - C_m\right] \tag{4.20}$$

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 (4.20), 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 [290], [287]. Kriging weights then come from the semivari-

ogram that was developed by analyzing the spatial nature of the data. These weights are a result of minimizing the variance in (4.20), which yield the following solution [292]:

$$\begin{bmatrix} \boldsymbol{\lambda} \\ \delta \end{bmatrix} = \mathbf{X^{-1}y} \tag{4.21}$$

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

$$\mathbf{X} = \begin{bmatrix} X_{1,1} & \cdots & X_{1,N} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ X_{N,1} & \cdots & X_{N,N} & \vdots \\ 1 & \cdots & 1 & 0 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \\ 1 \end{bmatrix} \tag{4.22}$$

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 (4.21), 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 is 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 [311] [299] due to its higher precision. As Kriging is geostatistical method, it also can estimate the variances of predicted values in the unsampled location.

4.3 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.

4.3.1 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. 4.3. This is particularly suitable in ulta-dense deployment scenarios [312], 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.

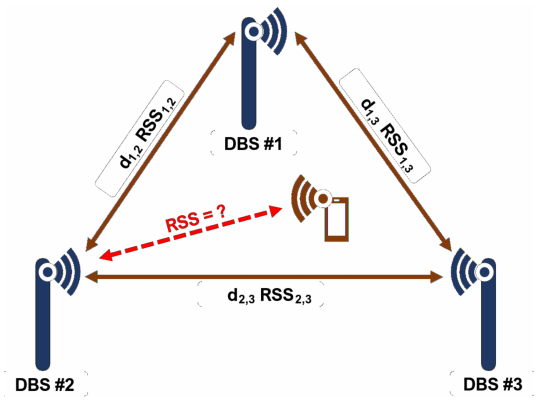


Fig. 4.3: Leveraging dense base station deployment to enrich sparse data.

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. 4.4. 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 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. 4.4). Conversely, if we traverse a path radially outwards, we can notice that the azimith angle remains the same but there is variation in Tx-Rx distance (d_1 to d_3 in Fig. 4.4 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 [313]. However, authors in [313] has not used this approach to address the data sparsity challenge. In [313], 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.

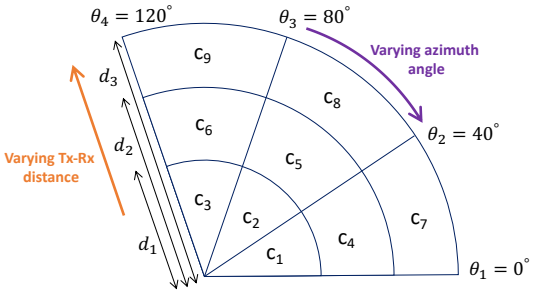


Fig. 4.4: Leveraging cluster geometry to enrich sparse data.

4.3.2 Through propagation modeling and transmitter parameter estimation

Received signal strength (RSS) based method

The RSS based method to recover sparse data is based on a combination of analytical models with statistical evaluation through measurements [314]. The RSS at a particular receiver, i located at a distance, d can be represented as:

$$P_i(d) = P_t - L - 10p\log_{10}(d) + \phi \tag{4.23}$$

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:

$$P_i^{av}(d) \approx P_t - L - 10p\log_{10}(d), \text{ where } P_i^{av}(d) = \sum_{k=1}^{N} P_i^k(d)/N \tag{4.24}$$

After performing some algebraic manipulations, taking the anti-log of (4.24) and representing d is cartesian coordinates, (4.24) can be transformed into a regression problem which can be expressed as a system of linear equation as follows [315]:

$$\begin{bmatrix} 10^{\frac{-L-P_1^{av}(d)}{5p}} & 2x_1 & 2y_1 & -1 \\ 10^{\frac{-L-P_2^{av}(d)}{5p}} & 2x_2 & 2y_2 & -1 \\ \vdots & \vdots & \vdots & \vdots \\ 10^{\frac{-L-P_N^{av}(d)}{5p}} & 2x_N & 2y_N & -1 \end{bmatrix} \begin{bmatrix} 10^{\frac{P_t}{5p}} \\ x_t \\ y_t \\ x_t^2 + y_t^2 \end{bmatrix} = \begin{bmatrix} x_1^2 + y_1^2 \\ x_2^2 + y_2^2 \\ \vdots \\ x_N^2 + y_N^2 \end{bmatrix} \tag{4.25}$$

where x_t, y_t is the transmitter location and(x_i, y_i) is the i-th receiver location. Therefore, by solving (4.25) 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 using 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 [315] 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.

Received signal strength difference (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 [314]:

$$P_{ab} = 5p \log_{10} \left(\frac{(x_t - x_a)^2 + (y_t - y_a)^2}{(x_t - x_b)^2 + (y_t - y_b)^2} \right) \tag{4.26}$$

The transmitter location in (4.26) can then be estimated by solving a linear system of equations of the following form:

$$\begin{bmatrix} 1-\beta_{12} & -2(x_2 - \beta_{12}x_1) & -2(y_2 - \beta_{12}y_1) \\ 1-\beta_{13} & -2(x_3 - \beta_{13}x_1) & -2(y_3 - \beta_{13}y_1) \\ \vdots & \vdots & \vdots \\ 1-\beta_{1N} & -2(x_N - \beta_{1N}x_1) & -2(y_N\beta_{1N}y_1) \end{bmatrix} \begin{bmatrix} x_t^2 + y_t^2 \\ x_t \\ y_t \end{bmatrix} = \begin{bmatrix} \beta_{12}(x_1^2 + y_1^2) - (x_2^2 + y_2^2) \\ \beta_{13}(x_1^2 + y_1^2) - (x_3^2 + y_3^2) \\ \vdots \\ \beta_{1N}(x_1^2 + y_1^2) - (x_N^2 + y_N^2) \end{bmatrix} \tag{4.27}$$

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 (4.27) 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.

Angle of arrival (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 [316]:

$$\mathbf{R}_i = \sqrt{\alpha}(d_i)\mathbf{h}(\theta_i)s + \mathbf{n}_i \tag{4.28}$$

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:

$$\mathbf{h}(\theta_i) = \begin{bmatrix} 1 \\ \exp(j\frac{\pi}{2}\sin(\theta_i)) \end{bmatrix}, \quad \alpha(d_i) = \phi\left(\frac{c}{4\pi f}\right) d_i^{-p} \tag{4.29}$$

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 [316] 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:

$$\begin{bmatrix} \hat{x}_t^i \\ \hat{y}_t^i \end{bmatrix} = \begin{bmatrix} x_i \\ y_i \end{bmatrix} + \begin{bmatrix} \hat{d}_i \cos(\hat{\theta}_i) \\ \hat{d}_i \sin(\hat{\theta}_i) \end{bmatrix} \tag{4.30}$$

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 [317] obtains the following over-conditioned system from the estimates:

$$\begin{bmatrix} -x_1\sin(\hat{\theta}_1) + y_1\cos(\hat{\theta}_1) \\ \vdots \\ -x_N\sin(\hat{\theta}_N) + y_N\cos(\hat{\theta}_N) \end{bmatrix} \approx \begin{bmatrix} -\sin(\hat{\theta}_1) & \cos(\hat{\theta}_1) \\ \vdots & \vdots \\ -\sin(\hat{\theta}_N) & \cos(\hat{\theta}_N) \end{bmatrix} \begin{bmatrix} \hat{x}_t \\ \hat{y}_t \end{bmatrix} \tag{4.31}$$

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.

Signal to noise ratio (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 [316]:

$$\begin{bmatrix} \hat{x}_t \\ \hat{y}_t \end{bmatrix} = \sum_{i=0}^{N} \frac{\Gamma_i}{\sum_{k=1}^{N} \Gamma_k} \begin{bmatrix} \hat{x}_t^i \\ \hat{y}_t^i \end{bmatrix} \tag{4.32}$$

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

$$\Gamma_i(d) = E\left[\frac{\alpha(d_i)P_t}{N_o B}\right] \tag{4.33}$$

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 [318]. 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:

$$\begin{aligned} P_r = P_t - A_o - A_1 \log_{10}(d) - A_2 \log_{10}(H_e) - \\ A_3 \log_{10}(d) \log_{10}(H) + 3.2(\log_{10}(11.75 H_m))^2 - \\ 44.49 \log_{10}(f) + 4.78(\log_{10}(f))^2 - L_d - L_c + G \end{aligned} \tag{4.34}$$

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 [318]:

$$G = G_{\max} - F_\theta + F_\theta \mid \cos^{p_1}\left(\frac{\theta_a zi - \theta_u}{2}\right) \mid - F_\phi + F_\phi \mid \cos^{p_2}\left(\frac{\phi_{tilt} - \phi_u}{2}\right) \mid \tag{4.35}$$

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 [279] can be as follows:

$$\begin{aligned} G = \lambda_\phi \left(G_{\max} - \min\left(12 \left(\frac{\phi_u - \phi_{tilt}}{B_\phi} \right)^2, A_{\max} \right) \right) + \\ \lambda_\theta \left(G_{\max} - \min\left(12 \left(\frac{\theta_u - \theta_{azi}}{B_\theta} \right)^2, A_{\max} \right) \right) \end{aligned} \tag{4.36}$$

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, transmitter and propagation environment parameters can then be obtained by minimizing the mean squared error between the measured and estimated signal strengths. Authors in [318] 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 [318]. Also, antenna parameters knowledge through antenna datasheets or antenna diagrams is required in this method.

4.4 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 ways to generate synthetic data through simulators.

4.4.1 Simulators

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 4.2. As observed from Table 4.2, none of the simulators is based on comprehensive 5G standard incorporating all aspects outlined in the standard. To tackle this problem, we have developed a simulator called SyntheticNET built on Python platform. The SyntheticNET 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, SyntheticNET 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 [319].

Python based platform and the flexibility of different input and output data formats in SyntheticNET simulator can assist in solving the data scarcity challenge by generating

[1]

Table 4.2: Comparison of different simulators for solving data sparsity problem.

Feature	Simulator													
	OpenAirInterface GTEC [320]	X. Wang et al. 5G-K [321]	V. V. Diaz et al. [322]	[323]	OMNeT++ ns-3 [324]	[325]	MATLAB/SIMULINK NYUSIM [326]	[327]	C-RAN [328]	OPNET [329]	Vienna 5G [330]	Atoll [274]	SyntheticNET [319]	
Scheduling support	✔	✔	✔	✔		✔	✔	✔		✔	✔	✔	✔	✔
mm-Wave support					✔	✔	✔	✔		✔		✔	✔	✔
Adaptive numerology									✔			✔	✔	✔
QCI support									✔				✔	✔
Parallelized offline traces and time-independent KPIs pre-generation for reduced online computational cost													✔	✔
Realistic antenna patterns modeling											✔		✔	✔
Signaling overhead modelling														✔
Realistic mobility modeling														✔
AI based pathloss modeling														✔
500+ COPs modeling														✔
Realistic HO management														✔
Realistic mobility pattern														✔
Python based to enable data processing and easy incorporation of ML libraries														✔
Free license*	✔	✔	✔	✔	✔	✔	✔	✔		✔	✔	✔		✔

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.

4.5 Enriching real data using machine learning

Several machine learning techniques can be leveraged to address the data sparsity challenge. These include generative adversarial networks, autoencoders, transfer learning and few shot learning techniques.

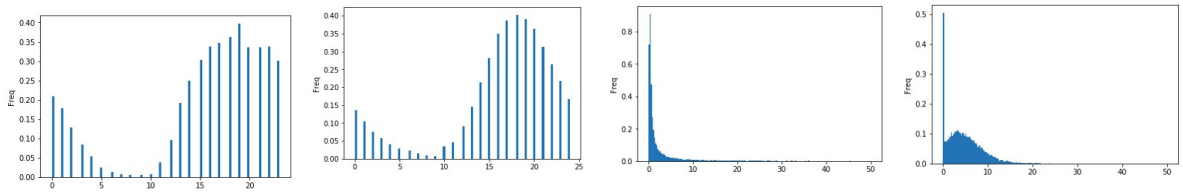
[1]Free license means free for academia use and in some cases under a signed contract by the lab head.

4.5.1 Generative adversarial networks

Generative adversarial networks (GANs) success in image processing has been well established [332]-[336]. Although this concept has widely been used in image processing, it can also be used in wireless communications. In wireless communications context, the works that utilize GANs are limited to [2, 337, 338, 339, 340, 341, 342, 343]. While GANs have been widely used for image data, its application to tabular data remains relatively limited. The works that use GANs on tabular data in a non-cellular network data context include [344, 345, 346, 347, 348, 349]. However, similiar concepts can be applied to wireless data domain too.

The basic idea 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). In GAN, a two-player minimax game is set between the discriminator DNN and generator DNN. In each training epoch, the generator iterates its weights to produce synthetic data trying to fool the discriminator DNN. The discriminator DNN on the other hand, tries to discriminate between real data and generated data. 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. 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 assess the efficacy of GAN-based approach outlined above, as a preliminary study recently published in [2], GAN was leveraged 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



(a) Real Calls Start Hours **(b)** Synthetic Calls Start Hours **(c)** Real Calls Durations (min) **(d)** Synthetic Calls Durations (min)

Fig. 4.5: Leveraging GAN for enriching the sparse training data [2].

CDR data can be used by a large number of SON solutions such as in [350], [351]. 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. 4.5a and 4.5c and represent the distribution of the real data used to train the discriminator. Figs. 4.5b and 4.5d 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.

Other GAN-based approaches in cellular networks context include the use of GANs to address the imbalance data issue in cell outage detection [337] Authors in [337] use an LTE simulator to get RSRP and RSRQ data and combine GAN wirh Adaboost to improve classification performance of imbalanced data for cell outage detection in self-organizing cellular networks.

A radio environment maps estimation algorithm leveraging a GAN-based pixel regression framework (PRF) for underlay cognitive radio networks using incomplete training data is proposed in [338, 339]. In these works, the authors first transform the radio environment maps estimation task into a pixel regression through color mapping. Then they

extract helpful information from the incomplete training data, design a feature enhancing module for the PRF algorithm, which intelligently learns and emphasizes the important features from the training images. Finally, they train the PRF to reconstruct the radio environment maps in the target area. Three indicators are used to test the proposed algorithm: the visual display of the radio environment maps, the estimated power spectrum of primary users, and the average REMs estimating error against different numbers of secondary users. Results are bench-marked with IDW and Kriging with the exponential semi-variogram estimation.

In the context of mitigating wireless jamming attacks, the work in [340] explores a GAN for the jammer to reduce the time to collect the training dataset by augmenting it with synthetic samples. In the context of wireless signal spoofing, GAN is used to generate and transmit synthetic signals that cannot be reliably distinguished from intended signals in [341]. Over here, a GAN-based spoofing attack that generates synthetic data that is transmitted by an adversary transmitter and distinguishes real and synthetic data at an adversary receiver. The minimax game between the adversary transmitter and receiver tunes both the generator and the discriminator. Then the signals generated by the GAN generator are transmitted for spoofing attack.

Using a RSS dataset in an indoor environment, the authors in [342] use GAN-generated synthetic data to enhance the accuracy of fingerprint-based localization.

Moreover, authors in [343], while drawing inspiration from image processing design a deep-learning architecture tailored to mobile networking, which combines Zipper Network (ZipNet) and GAN models. Using Telecom Italia's dataset, they infer fine-grained mobile traffic patterns to monitor city-wide mobile traffic via the GAN.

However, GANs suffer from many challenges, such as vanishing gradients, oscillations, modal collapse and the design of suitable evaluation metrics to evaluate their performance.

4.5.2 Autoencoders

Another machine learning based generative method that can be utilized for addressing data sparsity challenge is autoencoders. Unlike GANs, which come in the class of implicit density methods (where the prior distribution of latent features is not known), autoencoders fall under explicit density method, meaning that the distribution of latent features is explicitly defined.

Some works in this direction include [352, 353, 354]. An adversarial autoencoder using power spectral density data is demonstrated in [352] for detecting anomalous behavior in wireless spectrum. Along with anomaly detection, the proposed model in [352] shows a semi-supervised signal classification accuracy close to 100% on datasets using 20% of the labeled samples. Autoencoders for radio map estimation task using sparse measurements are investigated in [353]. Data-driven spectrum cartography via autoencoders is explored in [354]. After learning the propagation phenomena from data, the autoencoders based algorithm in [354] can construct a spectrum map from a significantly smaller number of measurements.

4.5.3 Transfer learning

The interpolation methods in section 4.2 are expected to work well when the sparse data has at least some latent features. Same is true for methods such as GANs 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, the transfer-learning paradigm [355, 356] can be leveraged. Given a source domain and source learning task, a target domain and a target learning task, transfer learning aims to help improve the learning of the target predictive function using the source knowledge. In cellular network context, similarities among cells can be leveraged for determining when to use transfer learning.

Several works have been carried out using transfer learning to address various aspects of cellular networks. Authors in [357] propose a transfer learning-based caching procedure at small cell base stations. They do so by exploiting the rich contextual information extracted from device-to-device interactions, referred to as source domain. This prior information is incorporated in the target domain where the goal is to optimally cache strategic contents at the small cells. A transfer learning data correlation-aware resource management approach in wireless virtual reality is presented in [358]. After formulating data correlation-aware resource management problem as an optimization problem, the authors in [358] developed a neural network reinforcement learning algorithm that uses echo state networks along with transfer learning to find the most suitable resource block allocations. Their simulation show an improved performance of the proposed approach compared to a classical Q-learning algorithm. Another transfer learning approach for parameter configuration of cellular networks is proposed in [359]. In this work, contextual bandit algorithm is leveraged along with transfer learning to optimize parameter configurations in cellular networks. Experimental results based on simulations and real network tests demonstrate the effectiveness of the proposed scheme. A transfer actor-critic learning framework for energy saving in cellular radio access networks is proposed in [360]. This work utilizes the transferred learning expertise in historical periods or neighboring regions for predicting traffic load variations. The problem of predicting the signal strength in the downlink of a real LTE network, where the antennas can be tuned to operate with different antenna tilt configurations is addressed using transfer learning in [361]. The authors show that augmenting the data from the source domain by adding data available from other tilts configurations of the same antenna improves the performance of the proposed transfer learning approaches. Transfer learning for channel quality prediction is proposed in [362], using dataset from a commercial LTE network. The results show how transfer learning can be carried out across pairs of cells working at different frequencies, or at the same frequency in different locations. In [363], authors presented a method for reducing the cost of data augmentation during the transfer learning of neural networks

on embedded devices. This is especially important when large models are implemented on embedded devices with limited computational and energy resources.

Transfer learning can also be adapted for algorithms that are not neural network based. For example, in a non-cellular network context, the XGBoost algorithm is expanded as a transfer learning framework [364].

4.5.4 Few shot learning

Few-shot learning (FSL) is another branch of machine learning that addresses the performance degradation problem of deep learning algorithms when the training dataset size is small. Using prior knowledge, FSL can master new tasks from a limited number of examples [365]. This type of learning is primarily motivated from the ability of humans to learn from only a few examples. Therefore, FSL can eliminate expensive data collection efforts and help in building suitable models for rare cases of limited supervised data [365].

FSL can be used for classification, regression and even reinforcement learning tasks using only few labeled, input-output and state-action examples respectively. However, the most common application scenario for FSL is *"N-way-K-shot classification"*, where a classifier is built for distinguishing between N classes, each having only K examples per class. When only one example with supervision is available, it is referred to as One-Shot Learning and when no example is available, it is called Zero-Shot Learning.

FSL is a very active area of research these days and the methods being proposed in the literature for solving the few-shot problem can be broadly classified in two different branches: 1) Meta learning, and 2) Metric learning. The key idea in Meta learning-based methods is to learn a learning strategy to adjust well to a new few-shot learning task by developing a task-agnostic learner (learn to learn). Examples of methods include Model Agnostic Meta Learning [366], Task-Agnostic Meta Learning [367] and Meta-transfer Learning [368]. These methods are good at out-of-distribution tasks and can handle varying and large shots well, but their model and architecture are intertwined and their optimization

process is challenging [369]. On the other hand, Metric learning-based methods learn to compare query set (test set) with support set (few-shot training set) by learning transferable representations in semantic embedding space using a distance loss function (learn to compare). Examples include Siamese Neural Networks [370], Matching Networks [371], Prototypical Networks [372], Relation Networks [373] and Graph Neural Networks [374]. As compared to meta learning-based methods, these are relatively simple, entirely feed-forward, computationally fast and easy to optimize, but harder to generalize to varying shots and to scale to very large shots [369].

A few works have been carried out using few shot learning to address training data sparsity issue in cellular networks. Authors in [375] use prototypical networks, a few shot learning-based algorithm for performance metrics analysis in LTE networks. They used eNodeB trace data from live network and classified individual eNodeBs into different performance classes based on their KPIs. Their results show an improved performance as compared to baseline DNN, 1-D CNN and 2-D CNN. Authors in [376] use self-imitation via transfer learning to achieve few-shot learning for the network power minimization problem in Cloud Radio Access Networks (C-RAN), for the scenario where the target domain (sparse dataset) only has unlabeled training samples. Their simulation results show the performance gap when compared to other learning-based methods. Authors in [377] use Siamese networks for achieving few shot learning for gesture recognition in indoor WiFi networks using channel state information (CSI) from commercial routers. Furthermore, the authors claimed that their proposed model achieved satisfactory results for spatio-temporal representation learning even under one-shot conditions.

4.6 Real data generation

The preceding 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.

4.6.1 Phone applications and parametric subscriber/third-party data

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 [351], used a novel methodology of utilizing smartphone application, 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 [351]. 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.

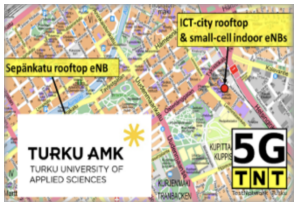
The quality of data gathered through smartphone applications, however, depends on a number of factors, including measurement capabilities of different smartphones and GPS error inaccuracy for measuring heights and positions. Smartphones equipped with barometers are likely to give a better estimate of 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. When the network operator is involved, it is possible for the subscriber to obtain parametric data from them. However, that type of data may be limited to a certain number of possible configurations. For this reason and for potential new scenarios, the solution may lie in resorting to testbeds.

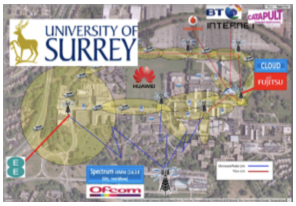
4.6.2 Testbeds

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. 4.6. 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 4.3 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.

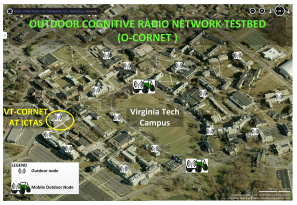
Apart from individual testbeds, several federations or consortiums of testbeds have been formed around the world. Some key federated testbeds comprising of the testbeds in Table 4.3 are presented in Fig. 4.7.



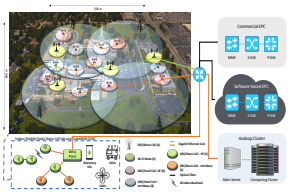
(a) Turku Testbed



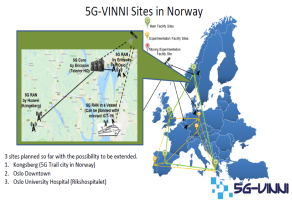
(b) Surrey Testbed



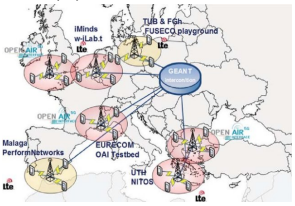
(c) CORNET Testbed



(d) OU-Tulsa Testbed

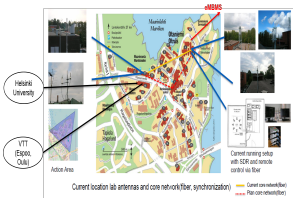


(e) 5G-VINNI Norway

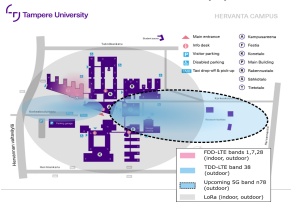


(f) FLEX Testbed

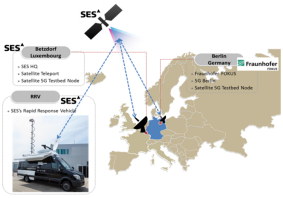
(g) 5G Playground Testbed



(h) 5G Test Network, Espoo



(i) 5G Test Network, Tampere



(j) 5G-VINNI Berlin

(k) 5G-VINNI Greece

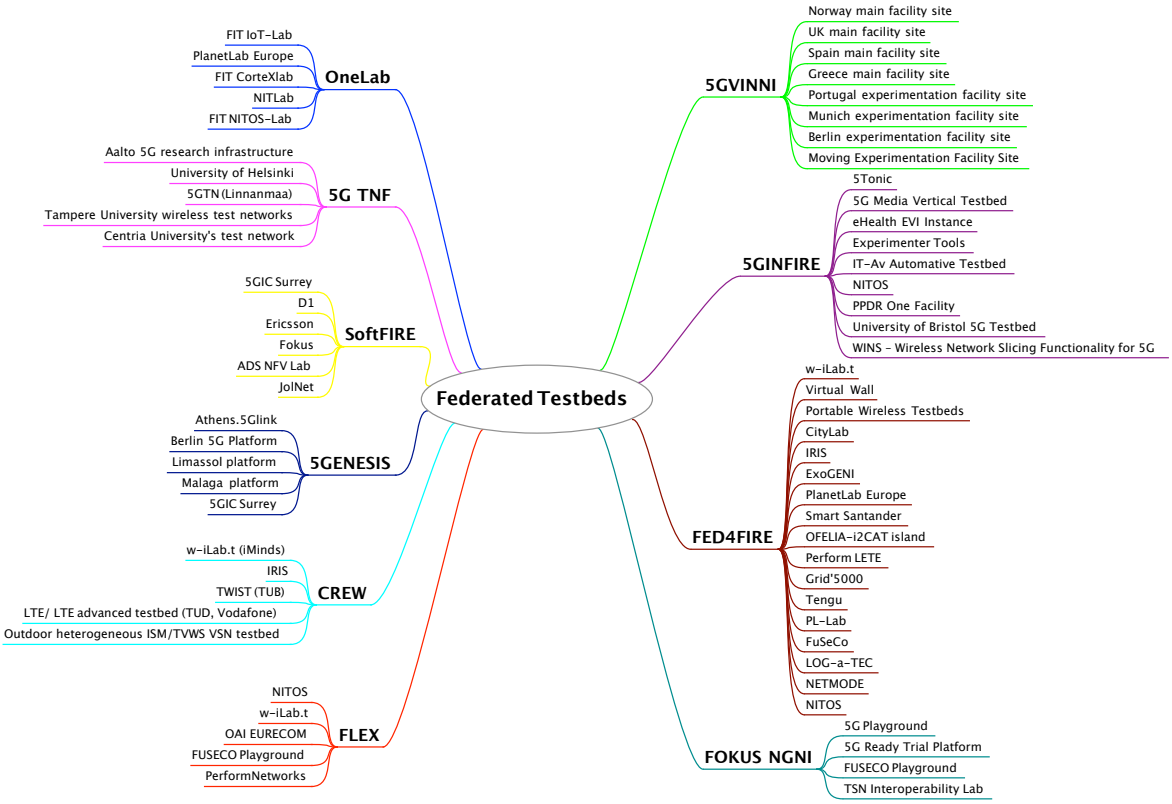


Fig. 4.7: Federated testbeds.

Table 4.3: Worldwide existing and emerging testbeds for solving data sparsity problem.

Testbed	Location	Key Features
NITOS [378] [379]	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 wensor 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 [380]
5GIC [381] - [384]	ICS, University of Surrey, Guildford, UK	- 4G LTE and 5G NR - 4km^2 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 [385] [386] [387]	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 [388] [389]	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 [390] [391] [392]	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 [393]
Firecycle [394] [395]	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 [396] [397] [398]	Center of Berlin, operated from Fraunhofer HHI, Deutsche Telekom Laboratories and 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 5GNR [399]	IITMadras 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 10 MHz bandwidth and can be extended to 20MHz - 5G NR for sub 6GHz and mm wave under development
TitanMIMO-6 [400] [401]	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 [402]	Otaniemi, Espoo, Finland	- Network slicing - Support for NB-IOT to be used for IoT hackathon - Mobile and edge computing, VR/AR, Gaming, Industrial Internet - Part of 5G TNF
University of Helsinki Test Network [403]	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 [404] [405] [406]	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 [407]
CORNET [408] [409]	Virginia Tech University, USA	- University-wide testbed - Software-defined radios, cognitive radio and 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 the grant from DURIP for upgrading to LTE and LTE-A - Outdoor network of 15 radio nodes and 2 mobile nodes

5G Playground [410]	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, like portable 5G edge nodes in progress - 3 Toolkits: Open5GCore, OpenSDNCore and Open5GMTC
Tampere University Wireless Test Networks [411] [412]	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 [413]	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 [414]	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 and management
Ericsson 5G [415] [416]	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 [417] [418] [419]	SK Telecom R&D Center, Bundang, Korea	- Developing a centimeter-wave (cmWave) 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 and AR/VR
5GTN (Linnanmaa) [420] [421] [422]	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 [423]	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 m^2 area - 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 the 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 [424]-[427]	EURECOM, France	- Open-source platform - 8-node testbed, equipped OAI compatible RF front-ends, UEs and VMs - 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 [428] [429]	TU Munich, Munchen, Germany	- 5G RAN with two sectors, each having carrier frequency: 3.4 GHz, bandwidth: 40 MHz, transmission power: 5 W antennas: up to 8 - 5G Mobile Terminals with vehicular speeds up to 50 km/h, enablingV2X - 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 as needed - 5G Core network supporting functional split – SDN – NFV Orchestration - Distributed data centres for mobile edge computing use cases
Perform Networks [430] [431] [432]	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 [433]	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 [434] [435] [436]	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 [437]	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 (OSM) - Dedicated NFVI for 5GINFIRE: 3 server computers, each with six cores, 32GB of memory, 2TB NLSAS, network card with 4 GbE ports, DPDK support - Second NFVI: 2 high-profile servers, each equipped with eight cores in a NUMA architecture, 128GB RDIMM RAM, 4TB SAS and eight 10Gbps Ethernet optical transceivers with SR-IOV capabilities
University of Bristol 5G [438]	University of Bristol, England	- Multi-site network connected through a 10 km fibre - Core network is located at HPN Lab at the University of Bristol - Extra edge computing node is available at Watershed - Access technologies are located at Millennium Square for outdoor coverage and "We The Curious" science museum for 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
D-15 Labs [439]	Ericsson, Santa Clara, CA, USA	- Validation and development platform for 5G use-cases, leverages cloud edge support, core network, and AI-based management and orchestration
ENCQOR 5G [440]	Ontario Region, Canada	- iPaaS Services: 5G connectivity of 5 Gbps Mobile Throughput and sub 5ms latency, cloud services of IoT Accelerator, emulation cloud, edge computing - iPaaS Infrastructure: 5G mobile user equipment (android-based Qualcom terminals operating at 3.5 GHz), 5G radio access technology (NR/LTE/CAT-M1/NB-IoT), 5G transport/backhaul, distributed core network and programmable data plane - Future features expected by 2021 include: 5 Gbps 5G NR, sub 5ms latency, predictive analytics, federated network slicing, real time machine learning / AI - Technology partners: Ericsson, Thales, CGI, IBM, Ciena

4.7 Conclusion

In this chapter, a systematic framework to address the data sparsity challenge was proposed and emerging new techniques were introduced that can be applied to wireless communication domain.

Table 4.4 summarizes the data augmentation techniques for handling sparse datasets in

mobile networks. The typical use cases targeted in existing literature include mobile traffic maps generation using sparse CDR data, spectrum sensing, MDT-based outage detection, CSI/RSS for localization, BS trace data for performance analysis, WiFi CSI for indoor localization, network power minimization, optimizing BS Tx power using UE SINR data, network parameter configuration optimization for power control and user scheduling, resource allocation, traffic load based energy saving, CQI and RSS prediction, radio environment map reconstruction, channel estimation in Massive MIMO systems and discovering user patterns using user trajectory data. The tools in existing literature to address these use cases include GANs and its variants, transfer learning, autoencoders, interpolation techniques, simulators and testbeds. While these techniques have proved to be beneficial for particular use cases, the generalization ability of a particular technique to different scenarios remains a challenge. Another notable challenge is the applicability of these techniques to highly dynamic or mobile environments. Efforts are also being made to reduce the training time of machine learning based models and modifying them for more robustness.

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 report. 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 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 access Kriging accuracy, it would always yield zero, since this type of interpolant satisfies the unbiased-ness 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. 4.8 is aimed to assist the researchers and network operators in choosing the best possible techniques, based on available information.

Table 4.4: Review of data augmentation techniques for handling sparse datasets in mobile networks.

Reference	Year	Data Augmentation Technique	Dataset Type	Use-Case	Task (Use-Case)	Labels
[2]	2019	GAN	Tabular	CDR Data (Call Start Hour and Call Duration)	Prediction	-
[343]	2020	ZipNet-GAN	Spatio-temporal	Fine Mobile Traffic Maps generation using coarse CDR Data	Generation	-
[340]	2018	C-GAN	Tabular	Spectrum Sensing Data	Classification	Required
[337]	2020	GAN	Tabular	MDT Data for Outage Detection	Classification	Required
[342]	2020	GAN	Tabular	CSI/RSS Data for Localization	Classification	Required
[375]	2020	Few-Shot Learning	Tabular	eNodeB Trace data for Performance Metrics Analysis	Classification	Required
[377]	2019	Few-Shot Learning	Tabular	WiFi CSI Data for Gesture Recognition / Indoor Localization	Classification	Required
[376]	2020	Few-Shot Learning + Transfer Learning	Tabular	Network Power Minimization in C-RAN	Prediction	-
[356]	2019	Transfer Learning	Tabular	UE SINR Data for optimizing BS Tx Power	Prediction	-
[357]	2015	Transfer Learning	Tabular	D2D Interactions Data for creating Content Popularity Matrix for Proactive Edge Caching	Prediction	-
[359]	2019	Transfer Learning	Tabular	Network Parameter Configuration Optimization for Power control and User Scheduling	Prediction	-
[358]	2019	Transfer Learning	Tabular	Tracking data of VR users for Correlation-aware Resource Allocation	Prediction	-
[360]	2014	Transfer Learning	Tabular	Traffic Load Arrival Rate Data for Energy Saving using BS ON/OFF Switching	Prediction	-
[362]	2019	Transfer Learning	Tabular	Hourly CQI Data for Inter/Intra-Cell CQI Prediction	Prediction	-
[441, 361]	2018-2020	Transfer Learning	Tabular	UE RSS (RSRP) data for prediction under different Antenna Tilt	Prediction	-
[338]	2020	GAN	Spatial	Radio environment Map Generation	Imputation	-
[442, 443]	2020	Deep Completion Autoencoders	Spatial	Radio environment Map Generation	Imputation	-
[339]	2020	Transfer Learning + GAN	Spatial	Radio environment Map Generation	Imputation	-
[444, 445, 446, 447, 448]	2015-2020	Context-Aware Interpolation	Spatial	REM construction using BS location estimated through reverse triangulation	Imputation	-
[449, 450, 451, 452, 453, 454]	2018-2020	Kriging Interpolation + Variants	Spatial	Radio environment Map Generation	Imputation	-
[455]	2019	Correlation-based Interpolation	Spatio-temporal	Spatio-temporal REM Generation	Imputation	-
[456]	2019	Adaptive Spatial Interpolation	Spatial	Uplink Channel Estimation in 3-D Massive MIMO Systems	Imputation	-
[457]	2019	Adaptive Triangulation-Induced Interpolation	Spatial	Multiple REM Generation	Imputation	-
[458]	2019	NN-enhanced Kriging Interpolation	Spatial	Radio environment Map Generation	Imputation	-
[459]	2018	Congregate group pattern	Spatio-temporal	Signaling data (User trajectory data) for discovering congregate group patterns	Prediction	-

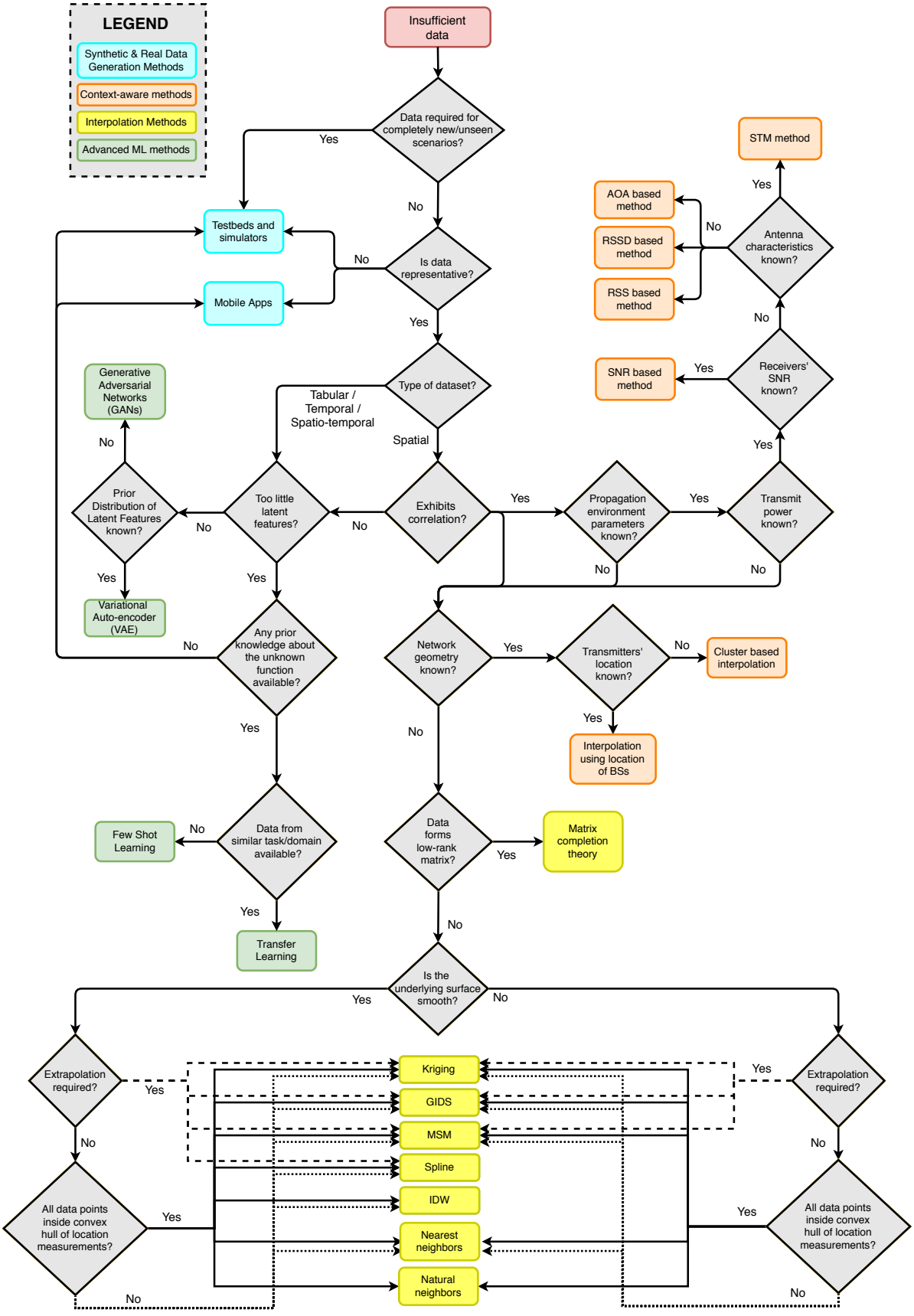


Fig. 4.8: Decision flowchart for the selection of data augmentation technique for handling sparse datasets in mobile networks.

CHAPTER 5

Towards interpretable models using machine learning to model complex systems

5.1 Introduction

As discussed in Section 1.1.2, traditional ML approaches suffer from lack of interpretability. In this chapter, using a hybrid approach consisting of analytical modeling, domain knowledge and machine learning, we move towards realistic and inherently interpretable mathematical models. Instead of using machine learning as a black box, the approach in this chapter allows insights and discovery of relationships between different KPIs and COPs, which can then to be used to optimize networks autonomously. The key idea behind the proposed approach is that instead of blindly selecting a hypothesis i.e., mathematical form of the model and then train it, as is the standard practice in ML— first leverage analytical modeling and domain knowledge to build a crude mathematical model, and then apply ML to learn some or all parameters in that model. This is a leap over conventional blindfold search in the hypothesis space when applying ML where one generally starts the training with a simple hypothesis e.g. linear equations, and move to gradually more complex ones.

After the model has been developed, COP-KPI analysis can be performed on the model to gather knowledge, such as determining what parameter ranges are more crucial to the model. These ranges can then be exploited for selective collection and enrichment of training data. This can provide the useful range of KPIs for intelligent optimization and provide a potential solution to solve the data scarcity problem. For example, as compared to a uniform or random collection for enrichment of training data, based on insights gained from the interpretable model, instead of uniform or random measurement

campaigns, more resources can be dedicated to data collection for KPI data corresponding to those KPI ranges that contribute the most in the model.

5.1.1 Related work

There are some existing studies that aim to address some of the challenges of traditional machine learning, namely, making machine learning models interpretable [460]-[468], reducing the model's training time and increasing its accuracy as new training data become available [469], [470].

Authors in [460] highlight the use of dimensionality reduction techniques that can lead to methods of information visualization in order to make the machine learning models more interpretable. By seeing interpretability as a problem of knowledge extraction from data regularity patterns, authors in [460] present visualization techniques as one of the forms of achieving knowledge extraction. Another work [461] studies the task of making some black-box models (support vector machines) interpretable, through constant B-spline kernel functions and sparsity constraints. This work [461] uses graphical tools and is medical domain focused. In [462], a cartogram-based method to reintroduce the local distortion into the low-dimensional data visualization provided by the batch-SOM (Self-organizing Maps) algorithm is provided. Reintroducing this distortion explicitly enables factorization of non-linearity of mapping in the visualization, which eases the SOM interpretation.

Other techniques that attempt to make ML-based models interpretable are discussed in [463]. These consist of adding interpretability constraints to the model, for example, a model is encouraged to use relatively fewer features for prediction and features are kept such so that they have monotonic relations with the prediction. Adding such constraints makes a model simpler and could increase the model's comprehensibility by users. However, it compromises on the features space and accuracy. Another approach to achieve this goal is mimic learning or surrogate model [465] , which means to approximate a

complex model using an easily interpretable model such as a decision tree, rule-based model, or linear model. The idea behind this approach is that if the approximation is sufficiently close, the statistical properties of the complex model will be reflected in the interpretable model. In order to assess the importance of a specific feature to the overall performance of a model to enhance interpretability, the idea of model-agnostic explanation is used, that is based on permutation feature importance, i.e., by measuring how the prediction accuracy changes when that feature is altered. Another approach to this end is the model-specific explanation, in which the model is explained by examining internal model structures and parameters. However, interpretability of an explanation will decrease when the feature dimensions become too large. A major limitation of existing work on interpretable machine learning highlighted in [463], [460] is that the explanations are designed based on the intuition of researchers [463]. Since current explanations are usually given in the form of feature importance vectors, they are a complete causal attribution and hence a low-level explanation [463]. In addition, by introducing approaches such sparsity and dimensionality reduction in models, they compromise on the explicit features or KPI to model a COP, in search for simpler/interpretable models. Moreover, these approaches do not output an explicit mathematical relationship between COPs and KPIs.

Using the Monte Carlo dropout technique and modeling the mis-specification distribution for Uber data, authors in [467] make the ML model more insightful by providing a way to model uncertainty estimation for a neural network forecast. Insights into prediction uncertainty, may help determine how much we can trust the forecast and also assist in reliable anomaly detection [467]. Another approach for interpreting ML-based model using tree ensembles is by finding prototypes in tree space utilizing the naturally-learned similarity measure [468].

Authors in [465] classify the existing interpretability models according three criteria: 1) the complexity of interpretability, which covers a class of methods that offer a post-hoc

explanation by using reverse engineering to provide some explanations of the model without even knowing the inner works of the model; 2) the scoop of interpretability, which further covers the classes of global and local interpretability; 3) the level of dependency from the used ML models, which covers model-agnostic and model-specific interpretability. Model-agnostic methods, being model-free, are preferred, and cover methods based on visualization, surrogate models, knowledge extraction, sensitivity analysis, feature importance and example-based explanations [465]. Subsections of the methods in [465] are also discussed in [466]. However, most of these approaches are still non-parametric and aimed at explaining the existing black-box models, rather than creating models that are inherently interpretable and thus able provide their own explanations [471].

Several hybrid machine learning/analytical models are discussed in [469], [470] that leverage both analytical modeling and machine learning with the goal of reducing the model's training time and increasing its accuracy over time as new data from the system becomes available. These techniques can be classified into the following seven different categories: 1) *Estimate and model*: This approach first uses ML to perform system model characterization and then uses this information to instantiate an analytical model. However, the accuracy of these solutions is ultimately dependent on the accuracy of the adopted analytical model technique. 2) *Bootstrapping*: In contrast to estimate and model approach, this technique first uses analytical modeling followed by machine learning. Analytical model is used to produce a synthetic training set over which black box learners are trained to correct analytical model's inaccuracies.

3) *Divide and conquer*: This approach consists in building specialized models, for different components of the target system, that rely either on analytical modeling or on ML. In the end, these sub-models are combined according to some formula in order to achieve the prediction output of the system as a whole. However, this approach suffers from long training phases and inaccuracies due to approximations.

4) *Ensembling:* In this technique, first an analytical model is built to predict the target

KPI, then, a chain of several black box learners is incrementally trained on both the KPI and the residual prediction errors of the previous model. 5) *K Nearest Neighbors*: This approach takes in a training set, analytical model and several black box machine learning models. It then selects which black-box model is best for a given configuration of training data by calculating the average error between the analytical model and ML-based learners across K-nearest neighbor configurations in the validation test. 6) *Hybrid Boosting*: Based on the intuition that the residual error of an analytical model may be learned more easily than the original target function that describes the relation between input and output variables, this approach aims to learn the residual error of an analytical model instead of the actual COP-KPI relationship. Unlike K Nearest Neighbors method, the ML learners are not used to build alternative models of the performance of the target system. Conversely, the learners are stacked in a chain and used to learn the error introduced by the previous learner in the chain. 7) *Probing*: The key idea of this approach is to use ML to perform predictions exclusively on the regions of those features in which analytical model does not achieve sufficient accuracy (rather than across the whole space). After a classifier identifies these regions of the feature space, a second black box regressor is trained on those regions to learn the desired performance function. However, these hybrid approaches [469], [470] despite improving the training time and accuracy, still remain predominately black-box approaches, without the ability to output an explicit mathematical COP-KPI relationship.

Our approach differs from other hybrid approaches [469], [470] in that it offers interpretability, while reducing training time. The models in [469], [470] still remain predominantly black-box models, without the ability to model explicit relations between COP-KPI relationships. Moreover, as compared to the interpretable models in [460]-[464], our proposed approach aims at an interpretable model without compromising on the feature space or searching for alternate ML-based simpler models or tweaking the parameters of existing ML-based models. Instead, it designs ML-based models based on starting analytical models to guide us to the right hypothesis class and learns the system-

specific parameters. Therefore, rather than trying to create a method for explaining the black box models, our approach aims at creating a model that is inherently interpretable in the first place.

5.2 Proposed approach

In our proposed approach, we address the challenges of COP-KPI modeling highlighted in the preceding section by leveraging a combination of domain knowledge of COP-KPI interactions, analytical modeling and ML. Firstly, using domain knowledge and available data sets, we build a crude analytical model of a complex system, such as one of the cellular network KPIs. This is done by using statistical analysis on available datasets to uncover some of the COPs-KPI dependencies while using domain knowledge and some assumptions in order to develop a parametric analytical model. Parametric models are preferable as they allow use of more deterministic (gradient based) optimization methods thanks to their tractable mathematical forms. Therefore, instead of blindly pre-supposing the hypothesis class, we use analytical modeling and subsequent mathematical insights to guide us to right hypothesis class (crude analytical model) and then apply ML by designing a custom neural network that is based on the crude analytical model. We then train this custom-designed neural network using training data. We will use the example of cellular network system to illustrate the approach.

A key challenge in this approach is the complex form of COP-KPI modeling if we consider the top-level KPIs. Hence, we start from the bottom and illustrate the proposed approach using the simplest indicator of coverage, RSRP. From RSRP, we can then calculate other top-level KPIs, such as SINR or capacity.

From domain knowledge, RSRP at the user can be written as:

$$P_r\,[\text{dBm}] = P_t - F + G - L + X \tag{5.1}$$

where P_t is the transmit power of the base station (BS), F is the free space path loss,

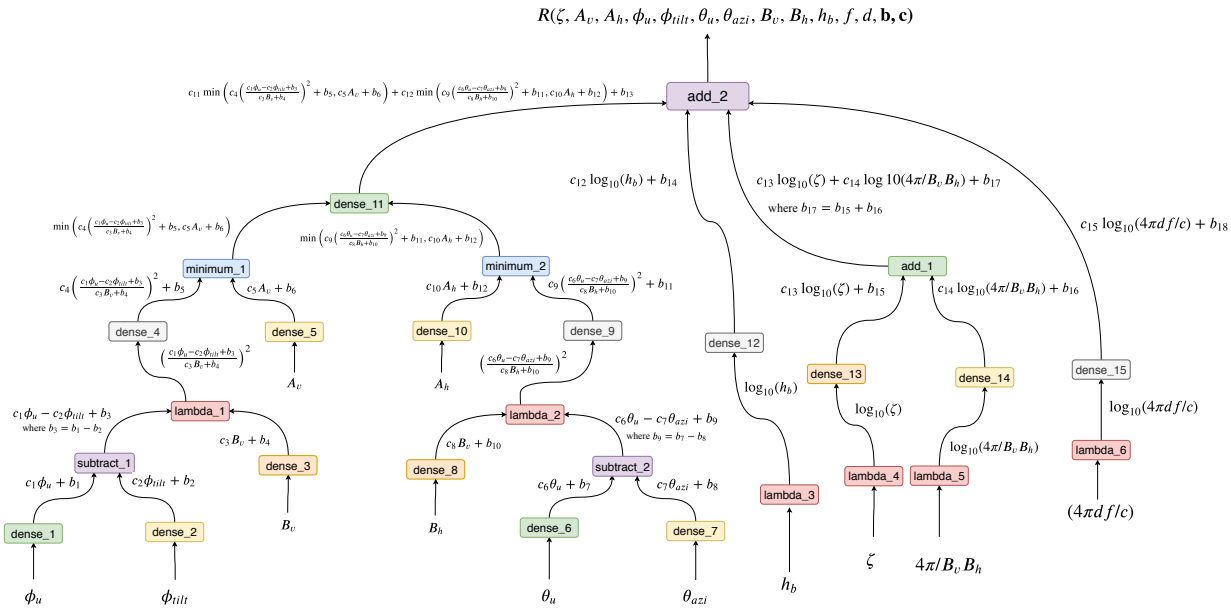


Fig. 5.1: Custom neural network variant to tune empirical parameters of RSRP.

G represents antenna gain, L denotes additional losses, such as cable, feeder or other equipment losses and X is a zero mean normal random variable with a certain standard deviation to model shadowing.

The free space path loss, F can be represented as a function of frequency and distance as follows:

$$F(f,d)[\text{dB}] = 10\eta \log_{10}\left(\frac{4\pi f d}{c}\right) \tag{5.2}$$

where η is an environment dependent constant, referred to as the path loss exponent.

We model the antenna gain, G as a combination of the maximum antenna gain, G_{max} and the antenna attenuation, A_{att}. G_{max} is a function of antenna efficiency, ζ and antenna directivity, D. The antenna directivity can be further modeled as a function of horizontal and vertical beamwidths of the directional antenna, B_h and B_v, respectively, as follows:

$$G_{max}(B_h, B_v, \zeta) = \zeta D = \zeta \frac{4\pi}{B_h B_v} \tag{3}$$

The three dimensional antenna attenuation can be approximated according to 3GPP recommendation [472]:

Table 5.1: Description of neurons and layers with activation functions in the custom designed neural network in Fig. 5.1.

Neuron/layer name	Brief description with activation function
dense_1 to dense_10, dense_12 to dense_15	core layer of type dense with custom activation function: K.abs
dense_11	core layer of type dense with custom activation function: K.sum
add_1, add_2	merge layer that adds a list of inputs
subtract_1	merge layer that subtracts a list of inputs
minimum_1, minimum_2	merge layer that computes the minimum of a list of inputs.
lambda_1, lambda_2	custom core layer of type lambda that wraps the expression K.square (input1/input2) as a layer object
lambda_3 to lambda_6	custom core layer of type lambda that wraps the expression K.log (input) as a layer object.

$$A_{att}(B_h, B_v, \phi_u, \phi_{tilt}, \theta_u, \theta_{azi}, \lambda_h, \lambda_v, A_h, A_v)[dB] =$$
$$\lambda_v \min\left(12\left(\frac{\phi_u - \phi_{tilt}}{B_v}\right)^2, A_v\right) + \lambda_h \min\left(12\left(\frac{\theta_u - \theta_{azi}}{B_h}\right)^2, A_h\right) \quad (4)$$

where ϕ_{tilt} is the tilt angle in degrees of the BS antenna, ϕ_u is the vertical angle in degrees 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 and θ_u is the angular distance of user from the horizontal reference axis. B_h and B_v represent the horizontal half power beamwidth and the vertical half power beamwidth of the BS antenna, in degrees respectively, while λ_h and λ_v represent the weighting factors for the beam pattern in both directions respectively. A_h and A_v denote the maximum attenuation at the sides and back of bore sight in horizontal and vertical dimensions respectively.

For more realistic RSRP modeling, we also add an additional term, BS height, h_b, since the RSRP at user from BS would be affected if the height of BS changes. Thus, RSRP in (5.1) can now be written as a function of several COPs as in Eq. 5.3.

$$P_r(\zeta, A_v, A_h, \phi_u, \phi_{tilt}, \theta_u, \theta_{azi}, B_v, B_h, h_b, f, d, \lambda_h, \lambda_v, A_h, A_v) = P_t - 10\eta \log_{10}\left(\frac{4\pi f d}{c}\right) + 10\log_{10}\left(\zeta \frac{4\pi}{B_h B_v}\right) - \lambda_v \min\left(12\left(\frac{\phi_u - \phi_{tilt}}{B_v}\right)^2, A_v\right) - \lambda_h \min\left(12\left(\frac{\theta_u - \theta_{azi}}{B_h}\right)^2, A_h\right) + 10\log_{10}(h_b) - L + X \tag{5.3}$$

$$R = k_1 \log_{10}\left(\frac{4\pi f d}{c}\right) + k_2 \log_{10}(\zeta) + k_3 \log_{10}\left(\frac{4\pi}{B_h B_v}\right) + k_4 \min\left(k_5\left(\frac{k_6\phi_u - k_7\phi_{tilt} + k_8}{k_9 B_v + k_{10}}\right)^2 + k_{11}, k_{12}\right) + k_{13} \min\left(k_{14}\left(\frac{k_{15}\theta_u - k_{16}\theta_{azi} + k_{17}}{k_{18}B_h + k_{19}}\right)^2 + k_{20}, k_{21}\right) + k_{22}\log_{10}(h_b) + k_{23} \tag{5.4}$$

Eq. 5.3 is a crude approximation of real RSRP that would be observed at a point. Now we will use RSRP data to fine-tine (5.3) using ML. This fine tuning can be in form of adjusting the values of empirical parameters such as pathloss co-efficient and exponents, antenna efficiency, system design parameters such as antenna parameters e.g., weights of vertical/horizontal beamwidths, or predictable parameters such as user locations. For this, we take part of the equation from (5.3) in which we can introduce tuneable/trainable coefficients that can represent empirical and system design parameters as in (5.4).

Now this expression in (5.4) for RSRP can be used as a starting point by the ML. We then proceed to design a custom neural network variant based on (5.4) as shown in Fig. 5.1. The activation functions for various layers or neurons are listed in Table 5.1. We use the abstract Keras backend API (the backend of which is imported as K) to design custom layers based on (5.4).

This ML model is then trained to tune values of $\mathbf{k} = (k_1 \ldots k_{23})$ in (5.4) using input training data. Therefore, instead of assuming values of $\mathbf{k}$ based on mathematical analysis that is often either too presumptuous to be accurate or asymptotic [9], ML will learn actual system and scenario specific values of $\mathbf{k}$ to give a more realistic COP-KPI model. Once RSRP is known, it can be translated into the top level KPIs, such as SINR or capacity analytically or using ML or combination of both.

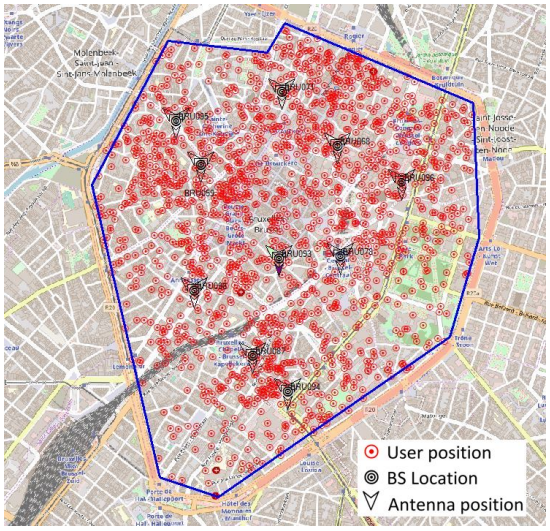


Fig. 5.2: Network topology.

Table 5.2: Key simulation parameters.

Parameter	Value/Type
Path loss model	Aster propagation
Propagation matrix resolution [m]	10
BS height [m]	20-40
Sectors per BS	1-3
Antenna tilt [^{o}]	0-8
Antenna azimuth [^{o}]	0-360
User distribution	Poisson
BS Transmit power [dBm]	30-43

5.3 Numerical results and analysis

5.3.1 Testing the proposed approach using simulated data

To evaluate the efficacy of the proposed idea, we generated RSRP values for various combination of COPs in a commercial cell planning tool, Atoll [274] based on 3-D ray tracing. The network topology used is shown in Fig. 5.2. Key simulation parameters and antenna types used are summarized in Table 5.2 and 5.3 respectively. Using the simulated data as ground truth, we implemented hybrid approach as described in the previous section and compared two approaches: (i) Pure analytical model-based approach (P_r)

Table 5.3: Key antenna types used in simulations.

B_h [^{o}]	B_v [^{o}]	G_{max} [dBi]	ϕ_{tilt} [^{o}]	θ_{azi} [^{o}]	f_{min} [MHz]	f_{max} [MHz]
30	14	18	0-8	0-360	1710	1900
33	9	21	0-8	0-360	1920	2170
60	10	16	0-8	0-360	2620	2690
65	8	18	0-8	0-360	1920	2170
90	6	16.5	0-8	0-360	1920	2170

Table 5.4: Coefficients in proposed approach obtained through neural network.

Coefficients	**Value**	**Coefficients**	**Value**
k_1	-33.47	k_{13}	-1.81
k_2	10.00	k_{14}	8.37
k_3	5.33	k_{15}	0.83
k_4	-4.63	k_{16}	1.00
k_5	5.69	k_{17}	-13.90
k_6	1.23	k_{18}	1.17
k_7	0.74	k_{19}	11.64
k_8	2.08	k_{20}	3.01
k_9	1.47	k_{21}	11.66
k_{10}	4.71	k_{22}	15.41
k_{11}	-4.76	k_{23}	7.99
k_{12}	-1.16	-	-

and (ii) Analytical model with coefficients learned through neural network ($P_t + R + X$).

The parameters learned through the custom designed neural network are shown in Table 5.4. Another clear advantage of this approach is its ability to learn system design and empirical parameters, even if they are unknown. Therefore, exact information about parameters like path loss exponents, antenna attenuation, antenna efficiency or even frequency or beamwidths is not required for this approach, since the neural network will

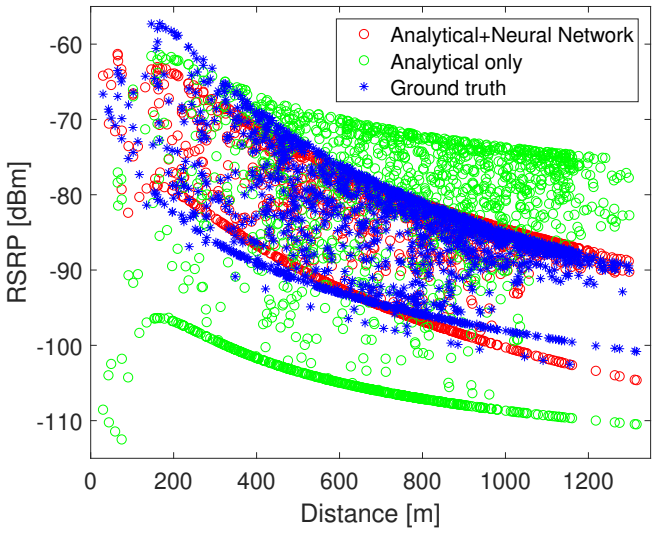


Fig. 5.3: RSRP versus distance of proposed approach compared with analytical modeling.

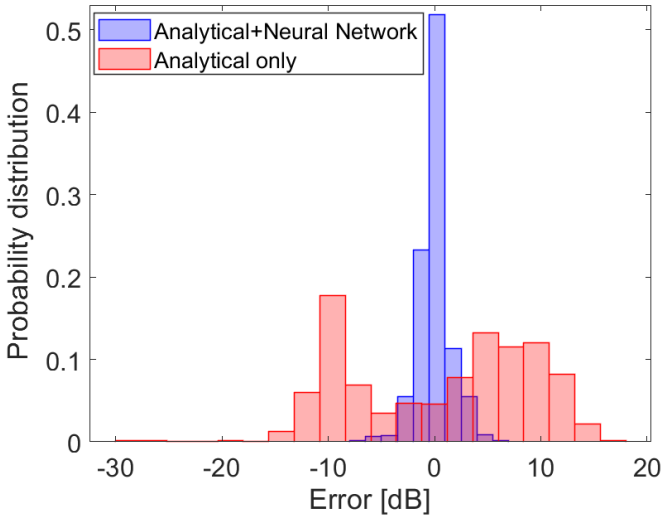


Fig. 5.4: Histograms of RSRP error using the two approaches.

learn or adjust these parameters when it is custom designed using appropriate starting equation from domain knowledge. For example, from Fig. Table 5.4, it can be seen that the path loss exponent, $-k_1/10 = 3.347$ and the horizontal antenna side and back lobe gain equals $k_{21} = 11.66$ dB

In Fig. 5.3, we show a comparison of RSRP estimated using both approaches with the ground truth (Atoll data) for a particular scenario, where $B_h = 65^o, B_v = 9^o, f = 2100$ MHz, $\phi_{tilt} = 8^o, \theta_{azi} = 180^o, P_t = 43$ dBm, $h_b = 30$ m.

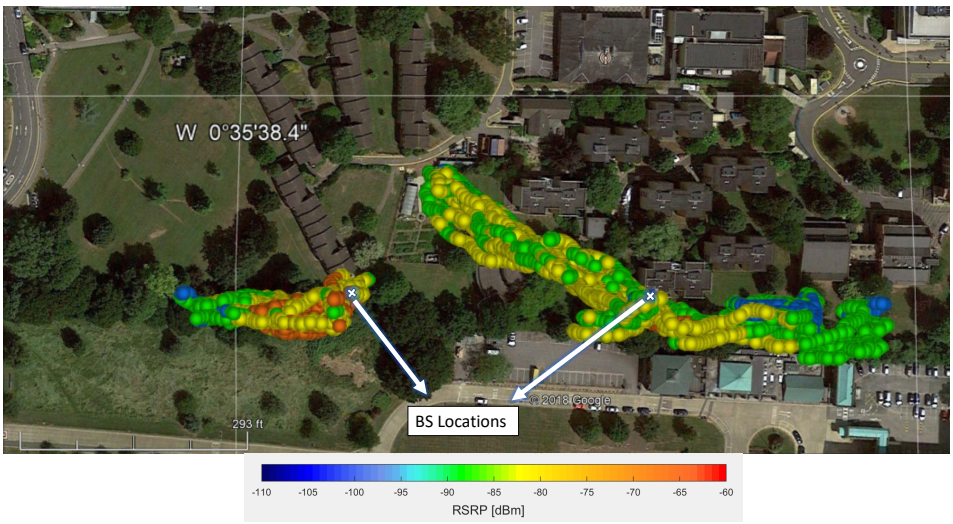


Fig. 5.5: Real data RSRP traces from two base stations.

The error in predicted RSRP using the two approaches with ground truth is presented in Fig. 5.4 wherein it is observed that our proposed approach of leveraging crude analytical approximation for guiding neural network to learn various parameters outperforms conventional analytical only approach with mean square error (MSE) decreasing from 71.36 to 2.16.

5.3.2 Testing the proposed approach using real data

In order to test the proposed approach on real data, we gathered RSRP data from two LTE pico cells from the University of Surrey 5GIC testbed. Fig. 5.5 shows a snapshot of data collected. One base station is mounted on top of the building and the other one is on ground as shown in Fig 5.6. We use the RSRP values gathered from BS on top of the building as training data for the custom designed neural network in order to learn environment-specific parameters as well as other system parameters. The final model obtained after training is then tested using data from the other base station. From Fig. 5.6, we note that these base stations have omni-directional antennas. Therefore, $\phi_{tilt} = 0$ in this case. Further, note that with $\phi_{tilt} = 0$, the azimuth plane becomes perpendicular to the boresight and the horizontal antenna attenuation term in (4) is no longer applicable. Therefore, the neural network for training can be either left unaltered as in Fig. 2 or the

(a) Base station on ground **(b)** Base station on building top

Fig. 5.6: Base stations used to collect real data

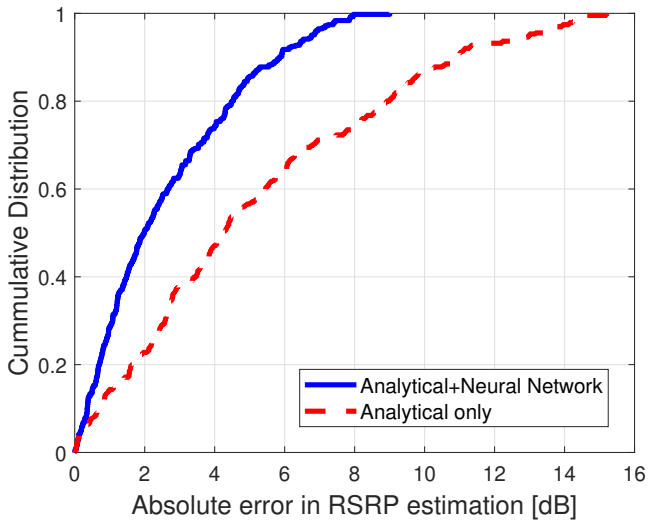


Fig. 5.7: Empirical CDFs of RSRP error using real test data.

modeling of horizontal antenna attenuation can be eliminated from it. In the former case, the neural network will automatically assign near-zero weight to the horizontal antenna attenuation term, while following the approach in the later case will lead to decreased computational complexity.

The CDF plots of absolute error in RSRP estimation utilizing both approaches are plotted in Fig. 5.7. It can be seen from Fig. 5.7 that 90% of RSRP values have error of less

than 5.9 dB using proposed approach, while this error value is 11 dB using analytical only approach. The MSE using proposed approach is reduced to 11.02 as compared to MSE of 41.79 using analytical only approach. Note, however, that a greater MSE of proposed approach using real data as compared to simulated data is observed. This can be attributed to the presence of GPS uncertainties in real data, that range from 1m-11m and consequently affect the distance between BS and receiver.

5.4 Conclusion

By combining domain knowledge and analytical modeling, we propose to design ML-based models such that instead of blindly selecting a hypothesis i.e., mathematical form of the model and then train it, as is the standard practice in ML— first leverage analytical modeling and domain knowledge to build a crude mathematical model, and then apply ML to learn some or all parameters in that model. Using RSRP data from both simulators and real world scenario, it is observed that the proposed model outperforms existing mathematical models, while also remaining interpretable. This RSRP model can be also be used to derive models of top-level KPIs, such as SNR or capacity. Thus, this approach can used to derive better mathematical models of complex systems. Motivated by the success of proposed approach to derive a mathematical model of an example KPI, RSRP, where some part of the RSRP equation as a function of COPs is already known through domain knowledge, future work will consider applying this approach for scenarios in which the analytical equation is completely unknown, since the results of this work show that a neural network model is capable of deriving a mathematical model of the KPI as a function of COPs.

CHAPTER 6

Conclusions and future work

6.1 Conclusions

This dissertation first identifies key communication requirements for advanced use cases in emerging networks. Several challenges and gaps in existing approaches to meet the requirements for emerging networks use cases are also identified. To meet the diverse and stringent communication requirements for emerging networks usecases, zero-touch artificial intelligence (AI) based deep automation in cellular networksis envisioned. However, traditional ML approaches to enable AI-based automation suffer from various challenges, including data sparsity challenge and lack of interpretability challenge. This dissertation addresses these two major challenges.

To address the training data sparsity challenge, a framework comprising of classical interpolation techniques, like inverse distance weighted and kriging to more advanced ML-based methods, like transfer learning and generative adversarial networks, several new techniques, such as matrix completion theory and leveraging different types of network geometries, and simulators and testbeds, among others is developed. Case studies are also demonstrated that show the success of these new techniques in enriching sparse data in cellular networks. The proposed framework can assist network operators and researchers 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. MDT is a key enabling feature for data and artificial intelligence driven autonomous operation and optimization in current and emerging cellular networks. However, although standardized since 3GPP Release 10, the full potential

of MDT still remains thwarted. This dissertation addresses the issues in MDT-based coverage estimation. For the first time, this dissertation reveals the existence of an optimal bin width for coverage estimation in the presence of inaccurate user positioning, scarcity of user reports and quantization error, that minimize the effect of all these errors concurrently. The presented framework can enable network operators to configure the bin size for given positioning accuracy and user density that results in the most accurate MDT based coverage estimation.

In order to address the lack of interpretability challenge, machine learning, big data and domain knowledge are leveraged derive mathematical models of complex systems. The results of this work show that a neural network model is capable of deriving a mathematical model of a KPI as a function of different COPs and hence can be used to derive interpretable equations of complex systems. The findings from this dissertation can solve the challenges in AI-based cellular networks and thus aid in their design, operation and optimization to meet the requirements of advanced use cases and applications.

6.2 Future works

In chapter 4, the idea of GANs has been leveraged to solve the data sparsity challenge. However, both the generator and discriminator in traditional GANs rely on back propagation, that in turn, rely on gradient descent method to reach to a good optima. Convergence of gradient descent becomes a major issue particularly when the distribution/process to be modeled has large dimensions (as is often the case in cellular network data), due to greater chances of different variables undoing each other's progress. This problem is aggravated if the training data is sparse, which is going to be the case for most data streams in mobile data. As a future work, I plan to solve this problem by incorporating domain knowledge to kick start the GAN training instead of relying on random hyper parameter initialization.

In chapter 5, a proof of concept was presented: a neural network is capable of deriving

an equation for a KPI as a function of different COPs. The neural network in chapter 5 is based on a starting equation. Motivated by this study, future work will explore the possibility of using a neural network to derive equations of complex systems, i.e, extracting an analytical equation from a black box neural network, when the starting equation is not completely known. One possibility is to expand different mathematical functions by taylor series expansion and try to trace those mathematical functions from the neural network by their taylor series expansion. I plan to leverage domain knowledge to custom select different parameters of the neural network, such as activation functions and layers.

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)*, Feb. 2019, pp. 282–286, iSSN: 2325-2626.

[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. Lecoeuvre, "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] 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.

[19] 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.

[20] GSMA, "Two-Thirds of Mobile Connections Running on 4G/5G Networks by 2025, Finds New GSMA Study," Accessed on: September 6, 2020 [Online]. Available: https://www.gsma.com/newsroom/press-release/two-thirds-mobile-connections-running-4g-5g-networks-2025-finds-new-gsma-study/.

[21] A. Ahad, M. Tahir, and K.-L. A. Yau, "5G-based smart healthcare network: architecture, taxonomy, challenges and future research directions," *IEEE Access*, vol. 7, pp. 100 747–100 762, 2019.

[22] L. Mucchi, S. Jayousi, S. Caputo, E. Paoletti, P. Zoppi, S. Geli, and P. Dioniso, "How 6G Technology Can Change the Future Wireless Healthcare," in *2020 2nd 6G Wireless Summit (6G SUMMIT)*, 2020, pp. 1–6.

[23] D. M. West, "How 5G technology enables the health internet of things," *Brookings Center for Technology Innovation*, vol. 3, pp. 1–20, 2016.

[24] ITU, " ITU-R Recommendation M.2083-0. IMT Vision D Framework and overall objectives of the future development of IMT for 2020 and beyond, September 2015 ."

[25] M. Xie, Q. Zhang, A. J. Gonzalez, P. Grønsund, P. Palacharla, and T. Ikeuchi, "Service Assurance in 5G Networks: A Study of Joint Monitoring and Analytics," in *2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2019, pp. 1–7.

[26] E. Marilly, O. Martinot, H. Papini, and D. Goderis, "Service level agreements: a main challenge for next generation networks," in *2nd European Conference on Universal Multiservice Networks. ECUMN'2001 (Cat. No. 02EX563)*, 2002, pp. 297–304.

[27] L. Wu and R. Buyya, "Service level agreement (sla) in utility computing systems," in *Performance and dependability in service computing: Concepts, techniques and research directions*. IGI Global, 2012, pp. 1–25.

[28] Y. Gao, H. Guan, Z. Qi, T. Song, F. Huan, and L. Liu, "Service level agreement based energy-efficient resource management in cloud data centers," *Computers & Electrical Engineering*, vol. 40, no. 5, pp. 1621–1633, 2014.

[29] F. L. Pires and B. Barán, "Multi-objective virtual machine placement with service level agreement: A memetic algorithm approach," in *2013 IEEE/ACM 6th Inter-*

national Conference on Utility and Cloud Computing. IEEE, 2013, pp. 203–210.

[30] Y. Gao, H. Guan, Z. Qi, T. Song, F. Huan, and L. Liu, "Service level agreement based energy-efficient resource management in cloud data centers," *Computers & Electrical Engineering*, vol. 40, no. 5, pp. 1621–1633, 2014.

[31] E. Wustenhoff and S. BluePrints, "Service level management in the data center," *Sun BluePrints Online*, 2002.

[32] L.-j. Jin, V. Machiraju, and A. Sahai, "Analysis on service level agreement of web services," *HP June*, vol. 19, 2002.

[33] H. Ludwig, A. Keller, A. Dan, R. P. King, and R. Franck, "Web service level agreement (wsla) language specification," *Ibm corporation*, pp. 815–824, 2003.

[34] M. B. Chhetri, J. Lin, S. Goh, J. Yan, J. Y. Zhang, and R. Kowalczyk, "A coordinated architecture for the agent-based service level agreement negotiation of web service composition," in *Australian Software Engineering Conference (ASWEC'06)*, 2006, pp. 10–pp.

[35] A. Dan, D. Dias, and J. Hellerstein, "Electronic service level agreement for web site and computer services hosting," Sep. 14 2006, uS Patent App. 11/434,096.

[36] S. Nepal, J. Zic, and S. Chen, "Wsla+: Web service level agreement language for collaborations," in *2008 IEEE International Conference on Services Computing*, vol. 2. IEEE, 2008, pp. 485–488.

[37] I. S. Monga, B. A. Schofield, and R. Tyagi, "System, device, and method for managing service level agreements in an optical communication system," Oct. 14 2008, uS Patent 7,437,449.

[38] B. A. Schofield, W. R. Hawe, P. D. Callahan, I. S. Monga, S. Suryaputra, and A. N. Fredette, "System, device, and method for managing communication services in an optical communication system," Dec. 7 2010, uS Patent 7,849,225.

[39] P. Patel, A. H. Ranabahu, and A. P. Sheth, "Service level agreement in cloud computing," 2009.

[40] P. Wieder, J. M. Butler, W. Theilmann, and R. Yahyapour, *Service level agreements*

for cloud computing. Springer Science & Business Media, 2011.

[41] S. Mubeen, S. A. Asadollah, A. V. Papadopoulos, M. Ashjaei, H. Pei-Breivold, and M. Behnam, "Management of service level agreements for cloud services in IoT: A systematic mapping study," *IEEE Access*, vol. 6, pp. 30 184–30 207, 2017.

[42] S. Girs, S. Sentilles, S. A. Asadollah, M. Ashjaei, and S. Mubeen, "A Systematic Literature Study on Definition and Modeling of Service-Level Agreements for Cloud Services in IoT," *IEEE Access*, vol. 8, pp. 134 498–134 513, 2020.

[43] I. Mesogiti, E. Theodoropoulou, K. Filis, G. Lyberopoulos, R. C. Palancar, N. S. Linares, D. Camps-Mur, J. Gutirrez, and A. Tzanakaki, "Network Services SLAs over 5G Infrastructure Converging Disaggregated Network and Compute Resources," in *2018 IEEE 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, 2018, pp. 1–5.

[44] M. A. Habibi, B. Han, M. Nasimi, and H. D. Schotten, "The structure of service level agreement of slice-based 5G network," *arXiv preprint arXiv:1806.10426*, 2018.

[45] E. Kapassa, M. Touloupou, A. Mavrogiorgou, and D. Kyriazis, "5G & SLAs: Automated proposition and management of agreements towards QoS enforcement," in *IEEE 21st Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN)*, 2018, pp. 1–5.

[46] E. Kapassa, M. Touloupou, and D. Kyriazis, "SLAs in 5G: A complete framework facilitating VNF-and NS-tailored SLA management," in *2018 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA)*. IEEE, 2018, pp. 469–474.

[47] M. Touloupou, E. Kapassa, C. Symvoulidis, P. Stavrianos, and D. Kyriazis, "An integrated SLA management framework in a 5G environment," in *2019 22nd Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN)*. IEEE, 2019, pp. 233–235.

[48] M. Touloupou, E. Kapassa, "5GTANGO SLA Manager Demo ," Accessed on: August 15, 2020 [Online]. Available: https://www.youtube.com/watch?v=_dtSAi6KjN4.

[49] B. Khodapanah, A. Awada, I. Viering, D. Oehmann, M. Simsek, and G. P. Fettweis, "Fulfillment of service level agreements via slice-aware radio resource management in

5G networks," in *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*. IEEE, 2018, pp. 1–6.

[50] G. Biczók, M. Dramitinos, H. Lønsethagen, L. M. Contreras, G. D. Stamoulis, and L. Toka, "Towards multi-operator IPTV services over 5G networks," *IPTV Delivery Networks: Next Generation Architectures for Live and Video-on-Demand Services*, pp. 283–314, 2018.

[51] J. Datta, I. Pan, and S. Bhattacharyya, "TSLA: Turing based service level agreement assessment model over diverse cloud deployments," in *2017 Third International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*. IEEE, 2017, pp. 157–162.

[52] A. G. D. of Communications and the Arts, "Impacts of 5G on productivity and economic growth," in *Bureau of Communications and Arts Research*. Australian Government Department of Communications and the Arts, 2018.

[53] H.-Y. Chang, J.-J. Jeng, S. Kumaran, H. Li, and L.-J. Zhang, "Systems and methods for monitoring and controlling business level service level agreements," Dec. 25 2007, uS Patent 7,313,533.

[54] V. C. Emeakaroha, I. Brandic, M. Maurer, and S. Dustdar, "Low level metrics to high level SLAs-LoM2HiS framework: Bridging the gap between monitored metrics and SLA parameters in cloud environments," in *2010 International Conference on High Performance Computing & Simulation*, 2010, pp. 48–54.

[55] P. Unterharnscheidt and A. Kieninger, "Service level management-challenges and their relevance from the customers' point of view." in *AMCIS*, 2010, p. 540.

[56] M. O. Al Kalaa and S. J. Seidman, "5-GHz Band LTE-LAA Signal Selection for Use as the Unintended Signal in ANSI C63. 27 Wireless Coexistence Testing," *IEEE Transactions on Electromagnetic Compatibility*, 2020.

[57] N. Bitar, M. O. Al Kalaa, S. J. Seidman, and H. H. Refai, "On the coexistence of LTE-LAA in the unlicensed band: Modeling and performance analysis," *IEEE Access*, vol. 6, pp. 52 668–52 681, 2018.

[58] FDA, "Medical Device Interoperability ," Accessed on: September 25, 2020 [Online]. Available: https://www.fda.gov/medical-devices/digital-health-center-excellence/medical-device-interoperability.

[59] "The WSLA framework: Specifying and monitoring service level agreements for web services, author=Keller, Alexander and Ludwig, Heiko," *Journal of Network and Systems Management*, vol. 11, no. 1, pp. 57–81, 2003.

[60] Klipfolio, "What are Business Metrics? ," Accessed on: September 25, 2020 [Online]. Available: https://www.klipfolio.com/resources/articles/what-are-business-metrics.

[61] P. Resnick, K. Kuwabara, R. Zeckhauser, and E. Friedman, "Reputation systems," *Communications of the ACM*, vol. 43, no. 12, pp. 45–48, 2000.

[62] D. Bates, "Samsung ordered to pay $340,000 after it paid people to write negative online reviews about HTC phones ," Accessed on: September 25, 2020 [Online]. Available: https://www.dailymail.co.uk/sciencetech/article-2476630/Samsung-ordered-pay-340-000-paid-people-write-negativeonline-reviews-HTC-phones.html.

[63] Y. Wu, E. W. Ngai, P. Wu, and C. Wu, "Fake online reviews: Literature review, synthesis, and directions for future research," *Decision Support Systems*, p. 113280, 2020.

[64] H. N. Qureshi, M. Manalastas, S. M. A. Zaidi, A. Imran, and M. O. Al Kalaa, "Service Level Agreements for 5G and beyond: Overview, Challenges and Enablers of 5G-Healthcare Systems," *IEEE Access*, 2020.

[65] G. Cisotto, E. Casarin, and S. Tomasin, "Requirements and Enablers of Advanced Healthcare Services over Future Cellular Systems," *IEEE Communications Magazine*, vol. 58, no. 3, pp. 76–81, 2020.

[66] F. Schaich, M.-H. Hamon, M. Hunukumbure, J. Lorca, K. Pedersen, M. Schubert, E. Kosmatos, G. Wunder, and K. Reaz, "The one5g approach towards the challenges of multi-service operation in 5g systems," in *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, 2018, pp. 1–6.

[67] 5GPPP, "5G Development and Validation Platform for Global Industry-Specific Network Services and Apps ," Accessed on: September 25, 2020 [Online]. Available: https://www.5gtango.eu/.

[68] 5G-Monarch, "Documentation of Requirements and KPIs and Definition of Suitable Evaluation Criteria ," Accessed on: June 30, 2021 [Online]. Avail-

able: https://5g-monarch.eu/wp-content/uploads/2017/10/5G-MoNArch_761445_D6.1_Documentation_of_Requirements_and_KPIs_and_Definition_of_Suitable_Evaluation_Criteria_v1.0.pdf.

[69] F. Krasniqi, L. Gavrilovska, and A. Maraj, "The Analysis of Key Performance Indicators (KPI) in 4G/LTE Networks," in *Future Access Enablers for Ubiquitous and Intelligent Infrastructures*, V. Poulkov, Ed. Cham: Springer International Publishing, 2019, pp. 285–296.

[70] 3rd Generation Partnership Project, "TR36.814 v9.0.0: Further advancements for E-UTRA physical layers aspects (release 9)," 3GPP, Sophia Antipolis, France, Tech. Rep., March 2010.

[71] S. Dwivedi, R. Shreevastav, F. Munier, J. Nygren, I. Siomina, Y. Lyazidi, D. Shrestha, G. Lindmark, P. Ernström, E. Stare *et al.*, "Positioning in 5G networks," *arXiv preprint arXiv:2102.03361*, 2021.

[72] D. M. Gutierrez-Estevez, M. Gramaglia, A. De Domenico, N. Di Pietro, S. Khatibi, K. Shah, D. Tsolkas, P. Arnold, and P. Serrano, "The path towards resource elasticity for 5G network architecture," in *2018 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*. IEEE, 2018, pp. 214–219.

[73] FDA, "Radio Frequency Wireless Technology in Medical Devices, Guidance for Industry and Food and Drug Administration Staff ," Accessed on: October 29, 2020 [Online]. Available: https://www.fda.gov/media/71975/download.

[74] "Global Connected Wearable Devices," Accessed on: June 27, 2021 [Online]. Available: https://www.statista.com/statistics/487291/global-connected-wearable-devices/.

[75] Q. Zhang, J. Liu, and G. Zhao, "Towards 5G enabled tactile robotic telesurgery," *arXiv preprint arXiv:1803.03586*, 2018.

[76] M. A. Usman, N. Y. Philip, and C. Politis, "5G enabled mobile healthcare for ambulances," in *2019 IEEE Globecom Workshops (GC Wkshps)*, 2019, pp. 1–6.

[77] C. Thuemmler, A. Gavrasm, A. Jumelle, A. Paulin, A. Sadique, A. Schneider, C. Fedell, D. Abraham, and D. Trossen, "5G and e-Health," *5G-PPP white paper*, pp. 1–24, 2015.

[78] M. D. Fabrlzio, B. R. Lee, D. Y. Chan, D. Stoianovici, T. W. Jarrett, C. Yang, and L. R. Kavoussi, "Effect of time delay on surgical performance during telesurgical manipulation," *Journal of endourology*, vol. 14, no. 2, pp. 133–138, 2000.

[79] R. Rayman, S. Primak, R. Patel, M. Moallem, R. Morady, M. Tavakoli, V. Subotic, N. Galbraith, A. Van Wynsberghe, and K. Croome, "Effects of latency on telesurgery: an experimental study," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2005, pp. 57–64.

[80] J. Marescaux, J. Leroy, F. Rubino, M. Smith, M. Vix, M. Simone, and D. Mutter, "Transcontinental robot-assisted remote telesurgery: feasibility and potential applications," *Annals of surgery*, vol. 235, no. 4, p. 487, 2002.

[81] S. Xu, M. Perez, K. Yang, C. Perrenot, J. Felblinger, and J. Hubert, "Determination of the latency effects on surgical performance and the acceptable latency levels in telesurgery using the dV-Trainer® simulator," *Surgical endoscopy*, vol. 28, no. 9, pp. 2569–2576, 2014.

[82] M. Perez, S. Xu, S. Chauhan, A. Tanaka, K. Simpson, H. Abdul-Muhsin, and R. Smith, "Impact of delay on telesurgical performance: study on the robotic simulator dV-Trainer," *International journal of computer assisted radiology and surgery*, vol. 11, no. 4, pp. 581–587, 2016.

[83] M. Eid, J. Cha, and A. El Saddik, "Admux: An adaptive multiplexer for haptic-audio-visual data communication," *IEEE Transactions on Instrumentation and Measurement*, vol. 60, no. 1, pp. 21–31, 2010.

[84] A. Marshall, K. M. Yap, and W. Yu, "Providing QoS for networked peers in distributed haptic virtual environments," *Advances in Multimedia*, vol. 2008, 2008.

[85] NSF, "NSF Follow-on Workshop on Ultra-Low Latency Wireless Networks," *NSF Workshop on Ultra Low-Latency Wireless Networks, Arlington, VA*, 2016.

[86] B. Cizmeci, X. Xu, R. Chaudhari, C. Bachhuber, N. Alt, and E. Steinbach, "A multiplexing scheme for multimodal teleoperation," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 13, no. 2, pp. 1–28, 2017.

[87] J. Marescaux, J. Leroy, M. Gagner, F. Rubino, D. Mutter, M. Vix, S. E. Butner, and M. K. Smith, "Transatlantic robot-assisted telesurgery," *Nature*, vol. 413, no.

6854, pp. 379–380, 2001.

[88] B. Zhen, M. Patel, S. Lee, E. Won, and A. Astrin, "TG6 technical requirements document (TRD)," *IEEE P802*, pp. 15–08, 2008.

[89] M. Patel and J. Wang, "Applications, challenges, and prospective in emerging body area networking technologies," *IEEE Wireless communications*, vol. 17, no. 1, pp. 80–88, 2010.

[90] Z. Shi, H. Zou, M. Rank, L. Chen, S. Hirche, and H. J. Muller, "Effects of packet loss and latency on the temporal discrimination of visual-haptic events," *IEEE Transactions on Haptics*, vol. 3, no. 1, pp. 28–36, 2009.

[91] Y. Makino, Y. Furuyama, S. Inoue, and H. Shinoda, "HaptoClone (Haptic-Optical Clone) for Mutual Tele-Environment by Real-time 3D Image Transfer with Midair Force Feedback." in *CHI*, 2016, pp. 1980–1990.

[92] T. Hachisu and H. Kajimoto, "Vibration feedback latency affects material perception during rod tapping interactions," *IEEE transactions on haptics*, vol. 10, no. 2, pp. 288–295, 2016.

[93] D. P. Bertsekas, "Traffic behavior and queuing in a QoS environment," *OPNETWORK 2005, Session 1813*, 2005.

[94] J. Kim, H. Kim, B. K. Tay, M. Muniyandi, M. A. Srinivasan, J. Jordan, J. Mortensen, M. Oliveira, and M. Slater, "Transatlantic touch: A study of haptic collaboration over long distance," *Presence: Teleoperators & Virtual Environments*, vol. 13, no. 3, pp. 328–337, 2004.

[95] R. T. Souayed, D. Gaiti, W. Yu, G. Dodds, and A. Marshall, "Experimental study of haptic interaction in distributed virtual environments," in *Proceedings of EuroHaptics*, 2004, pp. 260–266.

[96] K. M. Yap, A. Marshall, W. Yu, G. Dodds, Q. Gu, and R. T. Souayed, "Characterising distributed haptic virtual environment network traffic flows," in *International Conference on Network Control and Engineering for QoS, Security and Mobility*. Springer, 2005, pp. 297–310.

[97] K. S. Park and R. V. Kenyon, "Effects of network characteristics on human perfor-

mance in a collaborative virtual environment," in *Proceedings IEEE Virtual Reality (Cat. No. 99CB36316)*, 1999, pp. 104–111.

[98] P. Dev, D. Harris, D. Gutierrez, A. Shah, and S. Senger, "End-to-end performance measurement of Internet based medical applications." in *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2002, p. 205.

[99] D. Soldani, F. Fadini, H. Rasanen, J. Duran, T. Niemela, D. Chandramouli, T. Hoglund, K. Doppler, T. Himanen, J. Laiho *et al.*, "5G mobile systems for healthcare," in *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, 2017, pp. 1–5.

[100] S.-B. Xia and Q.-S. Lu, "Development status of telesurgery robotic system," *Chinese Journal of Traumatology*, vol. 24, no. 3, pp. 144–147, 2021.

[101] L. B. Valdez, R. R. Datta, B. Babic, D. T. Müller, C. J. Bruns, and H. F. Fuchs, "5G mobile communication applications for surgery: An overview of the latest literature," *Artificial Intelligence in Gastrointestinal Endoscopy*, vol. 2, no. 1, pp. 1–11, 2021.

[102] M. Dohler, "The Internet of Skills: How 5G-Synchronized Reality Is Transforming Robotic Surgery," *Robotic Surgery*, pp. 207–215, 2021.

[103] A. Acemoglu, G. Peretti, M. Trimarchi, J. Hysenbelli, J. Krieglstein, A. Geraldes, N. Deshpande, P. M. V. Ceysens, D. G. Caldwell, M. Delsanto *et al.*, "Operating from a distance: Robotic vocal cord 5G telesurgery on a cadaver," *Annals of Internal Medicine*, vol. 173, no. 11, pp. 940–941, 2020.

[104] L. B. Valdez, R. R. Datta, B. Babic, D. T. Müller, C. J. Bruns, and H. F. Fuchs, "5G mobile communication applications for surgery: An overview of the latest literature," *Artificial Intelligence in Gastrointestinal Endoscopy*, vol. 2, no. 1, pp. 1–11, 2021.

[105] A. Lacy, R. Bravo, A. Otero-Piñeiro, R. Pena, F. De Lacy, R. Menchaca, and J. Balibrea, "5G-assisted telementored surgery," *British Journal of Surgery*, vol. 106, no. 12, pp. 1576–1579, 2019.

[106] W. Tian, M. Fan, C. Zeng, Y. Liu, D. He, and Q. Zhang, "Telerobotic spinal surgery based on 5G network: the first 12 cases," *Neurospine*, vol. 17, no. 1, p. 114, 2020.

[107] J. Zheng, Y. Wang, J. Zhang, W. Guo, X. Yang, L. Luo, W. Jiao, X. Hu, Z. Yu, C. Wang *et al.*, "5G ultra-remote robot-assisted laparoscopic surgery in China," *Surgical Endoscopy*, vol. 34, no. 11, pp. 5172–5180, 2020.

[108] S. Sedaghat and A. H. Jahangir, "RT-TelSurg: Real Time Telesurgery Using SDN, Fog, and Cloud as Infrastructures," *IEEE Access*, vol. 9, pp. 52 238–52 251, 2021.

[109] M. A. Usman, N. Y. Philip, and C. Politis, "5G enabled mobile healthcare for ambulances," in *2019 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2019, pp. 1–6.

[110] J. M. Lippman, S. N. C. Smith, T. L. McMurry, Z. G. Sutton, B. S. Gunnell, J. Cote, D. G. Perina, D. C. Cattell-Gordon, K. S. Rheuban, N. J. Solenski *et al.*, "Mobile telestroke during ambulance transport is feasible in a rural EMS setting: the iTREAT Study," *Telemedicine and e-Health*, vol. 22, no. 6, pp. 507–513, 2016.

[111] F. Ehrler and J. N. Siebert, "PedAMINES: a disruptive mHealth app to tackle paediatric medication errors," *Swiss Medical Weekly*, vol. 150, no. 3536, 2020.

[112] B. Almadani, M. Bin-Yahya, and E. M. Shakshuki, "E-AMBULANCE: real-time integration platform for heterogeneous medical telemetry system," *Procedia Computer Science*, vol. 63, pp. 400–407, 2015.

[113] M. A. R. Bin-Yahya, "E-AMBULANCE: A Real-time Integration Platform for Heterogeneous Medical Telemetry System of Smart Ambulances," Ph.D. dissertation, King Fahd University of Petroleum and Minerals (Saudi Arabia), 2015.

[114] Z. C. Antoniou, A. S. Panayides, M. Pantzaris, A. G. Constantinides, C. S. Pattichis, and M. S. Pattichis, "Real-time adaptation to time-varying constraints for medical video communications," *IEEE journal of biomedical and health informatics*, vol. 22, no. 4, pp. 1177–1188, 2017.

[115] S. Yu, F. Yi, X. Qiulin, and S. Liya, "A Framework of 5G Mobile-health Services for Ambulances," in *2020 IEEE 20th International Conference on Communication Technology (ICCT)*. IEEE, 2020, pp. 528–532.

[116] S. S. Rajan, S. Baraniuk, S. Parker, T.-C. Wu, R. Bowry, and J. C. Grotta, "Implementing a mobile stroke unit program in the United States: why, how, and how much?" *JAMA neurology*, vol. 72, no. 2, pp. 229–234, 2015.

[117] L. Yperzeele, R.-J. Van Hooff, A. De Smedt, A. V. Espinoza, R. Van Dyck, R. Van de Casseye, A. Convents, I. Hubloue, D. Lauwaert, J. De Keyser *et al.*, "Feasibility of AmbulanCe-Based Telemedicine (FACT) study: safety, feasibility and reliability of third generation in-ambulance telemedicine," *PloS one*, vol. 9, no. 10, p. e110043, 2014.

[118] A. V. Espinoza, R.-J. Van Hooff, A. De Smedt, M. Moens, L. Yperzeele, K. Nieboer, I. Hubloue, J. de Keyser, A. Convents, H. F. Tellez *et al.*, "Development and pilot testing of 24/7 in-ambulance telemedicine for acute stroke: prehospital stroke study at the Universitair Ziekenhuis Brussel-Project," *Cerebrovascular diseases*, vol. 42, no. 1-2, pp. 15–22, 2016.

[119] A. Itrat, A. Taqui, R. Cerejo, F. Briggs, S.-M. Cho, N. Organek, A. P. Reimer, S. Winners, P. Rasmussen, M. S. Hussain *et al.*, "Telemedicine in prehospital stroke evaluation and thrombolysis: taking stroke treatment to the doorstep," *JAMA neurology*, vol. 73, no. 2, pp. 162–168, 2016.

[120] M. P. LaMonte, J. Cullen, D. M. Gagliano, R. Gunawardane, P. Hu, C. Mackenzie, and Y. Xiao, "TeleBAT: mobile telemedicine for the Brain Attack Team," *Journal of Stroke and Cerebrovascular Diseases*, vol. 9, no. 3, pp. 128–135, 2000.

[121] F. Geisler, A. Kunz, B. Winter, M. Rozanski, C. Waldschmidt, J. E. Weber, M. Wendt, K. Zieschang, M. Ebinger, H. J. Audebert *et al.*, "Telemedicine in prehospital acute stroke care," *Journal of the American Heart Association*, vol. 8, no. 6, p. e011729, 2019.

[122] T.-C. Wu, S. A. Parker, A. Jagolino, J.-M. Yamal, R. Bowry, A. Thomas, A. Yu, and J. C. Grotta, "Telemedicine can replace the neurologist on a mobile stroke unit," *Stroke*, vol. 48, no. 2, pp. 493–496, 2017.

[123] T.-C. Wu, C. Nguyen, C. Ankrom, J. Yang, D. Persse, F. Vahidy, J. C. Grotta, and S. I. Savitz, "Prehospital utility of rapid stroke evaluation using in-ambulance telemedicine: a pilot feasibility study," *Stroke*, vol. 45, no. 8, pp. 2342–2347, 2014.

[124] T. G. Liman, B. Winter, C. Waldschmidt, N. Zerbe, P. Hufnagl, H. J. Audebert, and M. Endres, "Telestroke ambulances in prehospital stroke management: concept and pilot feasibility study," *Stroke*, vol. 43, no. 8, pp. 2086–2090, 2012.

[125] J.-C. Hsieh, B.-X. Lin, F.-R. Wu, P.-C. Chang, Y.-W. Tsuei, and C.-C. Yang, "Ambulance 12-lead electrocardiography transmission via cell phone technology to cardiologists," *TELEMEDICINE and e-HEALTH*, vol. 16, no. 8, pp. 910–915, 2010.

[126] C. Terkelsen, B. Nørgaard, J. Lassen, J. Gerdes, J. Ankersen, F. Rømer, T. Nielsen, and H. Andersen, "Telemedicine used for remote prehospital diagnosing in patients suspected of acute myocardial infarction," *Journal of internal medicine*, vol. 252, no. 5, pp. 412–420, 2002.

[127] I. U. Rehman, M. M. Nasralla, A. Ali, and N. Philip, "Small cell-based ambulance scenario for medical video streaming: a 5G-health use case," in *2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT (HONET-ICT)*. IEEE, 2018, pp. 29–32.

[128] M. D. Kamal, A. Tahir, M. B. Kamal, and M. A. Naeem, "Future Location Prediction for Emergency Vehicles Using Big Data: A Case Study of Healthcare Engineering," *Journal of Healthcare Engineering*, vol. 2020, 2020.

[129] M. Roddy, T. Truong, P. Walsh, M. Al Bado, Y. Wu, M. Healy, and S. Ahearne, "5G Network Slicing for Mission-critical use cases," in *2019 IEEE 2nd 5G World Forum (5GWF)*. IEEE, 2019, pp. 409–414.

[130] Q. Wang, J. Alcaraz-Calero, R. Ricart-Sanchez, M. B. Weiss, A. Gavras, N. Nikaein, X. Vasilakos, B. Giacomo, G. Pietro, M. Roddy *et al.*, "Enable advanced QoS-aware network slicing in 5G networks for slice-based media use cases," *IEEE transactions on broadcasting*, vol. 65, no. 2, pp. 444–453, 2019.

[131] Y. Zhai, X. Xu, B. Chen, H. Lu, Y. Wang, S. Li, X. Shi, W. Wang, L. Shang, and J. Zhao, "5G-Network-Enabled Smart Ambulance: Architecture, Application, and Evaluation," *IEEE Network*, vol. 35, no. 1, pp. 190–196, 2021.

[132] H. Audebert, K. Fassbender, M. S. Hussain, M. Ebinger, G. Turc, K. Uchino, S. Davis, A. Alexandrov, and J. Grotta, "The PRE-hospital stroke treatment organization," *International Journal of Stroke*, vol. 12, no. 9, pp. 932–940, 2017.

[133] J. Kandimalla, A. R. Vellipuram, G. Rodriguez, A. Maud, S. Cruz-Flores, and R. Khatri, "Role of Telemedicine in Prehospital Stroke Care," *Current Cardiology Reports*, vol. 23, no. 6, pp. 1–5, 2021.

[134] EU 5G PPP Trials Working Group (including J. Alcaraz Calero and Q. Wang), "The 5G PPP Infrastructure -Trials and Pilots Brochure ," Accessed on: June 24, 2021 [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2019/09/5GInfraPPP_10TPs_Brochure_FINAL_low_singlepages.pdf.

[135] Ignacio Martinez-Alpiste, Jose M. Alcaraz Calero, Qi Wang, Gelayol Golcarenarenji, Enrique Chirivella-Perez and Pablo Salva-Garcia), "5G Can Shape Mission-Critical Healthcare Services ," Accessed on: June 24, 2021 [Online]. Available: https://https://www.comsoc.org/publications/ctn/5g-can-shape-mission-critical-healthcare-services.

[136] MIoT, "Internet of Medical Things Revolutionizing Healthcare," Accessed on: July 26, 2021 [Online]. Available:https://aabme.asme.org/posts/internet-of-medical-things-revolutionizing-healthcare.

[137] P. Lukowicz, U. Anliker, J. Ward, G. Troster, E. Hirt, and C. Neufelt, "AMON: a wearable medical computer for high risk patients," in *Proceedings. Sixth International Symposium on Wearable Computers,*. IEEE, 2002, pp. 133–134.

[138] K. M. Diaz, D. J. Krupka, M. J. Chang, J. Peacock, Y. Ma, J. Goldsmith, J. E. Schwartz, and K. W. Davidson, "Fitbit®: An accurate and reliable device for wireless physical activity tracking," *International journal of cardiology*, vol. 185, p. 138, 2015.

[139] B. Reeder and A. David, "Health at hand: a systematic review of smart watch uses for health and wellness," *Journal of biomedical informatics*, vol. 63, pp. 269–276, 2016.

[140] T. Q. Trung, S. Ramasundaram, B.-U. Hwang, and N.-E. Lee, "An all-elastomeric transparent and stretchable temperature sensor for body-attachable wearable electronics," *Advanced materials*, vol. 28, no. 3, pp. 502–509, 2016.

[141] Y. Yamamoto, D. Yamamoto, M. Takada, H. Naito, T. Arie, S. Akita, and K. Takei, "Efficient skin temperature sensor and stable gel-less sticky ECG sensor for a wearable flexible healthcare patch," *Advanced healthcare materials*, vol. 6, no. 17, p. 1700495, 2017.

[142] R. Adiputra, S. Hadiyoso, and Y. S. Hariyani, "Internet of things: Low cost and wearable SpO2 device for health monitoring," *International Journal of Electrical and Computer Engineering*, vol. 8, no. 2, p. 939, 2018.

[143] A. Azhari, S. Yoshimoto, T. Nezu, H. Iida, H. Ota, Y. Noda, T. Araki, T. Uemura, T. Sekitani, and K. Morii, "A patch-type wireless forehead pulse oximeter for SpO 2 measurement," in *2017 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2017, pp. 1–4.

[144] P. J. Chacon, L. Pu, T. H. da Costa, Y.-H. Shin, T. Ghomian, H. Shamkhalichenar, H.-C. Wu, B. A. Irving, and J.-W. Choi, "A wearable pulse oximeter with wireless communication and motion artifact tailoring for continuous use," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 6, pp. 1505–1513, 2018.

[145] G. Surrel, F. Rincón, S. Murali, and D. Atienza, "Low-power wearable system for real-time screening of obstructive sleep apnea," in *2016 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*. Ieee, 2016, pp. 230–235.

[146] M. Shilaih, B. M. Goodale, L. Falco, F. Kübler, V. De Clerck, and B. Leeners, "Modern fertility awareness methods: wrist wearables capture the changes in temperature associated with the menstrual cycle," *Bioscience reports*, vol. 38, no. 6, 2018.

[147] R. Xie, Q. Du, B. Zou, Y. Chen, K. Zhang, Y. Liu, J. Liang, B. Zheng, S. Li, W. Zhang *et al.*, "Wearable leather-based electronics for respiration monitoring," *ACS Applied Bio Materials*, vol. 2, no. 4, pp. 1427–1431, 2019.

[148] A. Mizuno, S. Changolkar, and M. S. Patel, "Wearable Devices to Monitor and Reduce the Risk of Cardiovascular Disease: Evidence and Opportunities," *Annual review of medicine*, vol. 72, 2020.

[149] C. Holz and E. J. Wang, "Glabella: Continuously sensing blood pressure behavior using an unobtrusive wearable device," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, pp. 1–23, 2017.

[150] M. Kuwabara, K. Harada, Y. Hishiki, and K. Kario, "Validation of two watch-type wearable blood pressure monitors according to the ANSI/AAMI/ISO81060-2: 2013 guidelines: Omron HEM-6410T-ZM and HEM-6410T-ZL," *The Journal of Clinical Hypertension*, vol. 21, no. 6, pp. 853–858, 2019.

[151] T. Arakawa, "Recent research and developing trends of wearable sensors for detecting blood pressure," *Sensors*, vol. 18, no. 9, p. 2772, 2018.

[152] P. Escobedo, C. E. Ramos-Lorente, A. Martínez-Olmos, M. A. Carvajal, M. Ortega-Munoz, I. de Orbe-Paya, F. Hernández-Mateo, F. Santoyo-González, L. F. Capitán-Vallvey, A. J. Palma *et al.*, "Wireless wearable wristband for continuous sweat pH monitoring," *Sensors and Actuators B: Chemical*, vol. 327, p. 128948, 2021.

[153] S. Nakata, M. Shiomi, Y. Fujita, T. Arie, S. Akita, and K. Takei, "A wearable

pH sensor with high sensitivity based on a flexible charge-coupled device," *Nature Electronics*, vol. 1, no. 11, pp. 596–603, 2018.

[154] J. Wijsman, B. Grundlehner, H. Liu, H. Hermens, and J. Penders, "Towards mental stress detection using wearable physiological sensors," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2011, pp. 1798–1801.

[155] G. Valenza, M. Nardelli, A. Lanata, C. Gentili, G. Bertschy, R. Paradiso, and E. P. Scilingo, "Wearable monitoring for mood recognition in bipolar disorder based on history-dependent long-term heart rate variability analysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 5, pp. 1625–1635, 2013.

[156] A. Gruwez, A.-V. Bruyneel, and M. Bruyneel, "The validity of two commercially-available sleep trackers and actigraphy for assessment of sleep parameters in obstructive sleep apnea patients," *PLoS One*, vol. 14, no. 1, p. e0210569, 2019.

[157] C.-T. Lin, L.-W. Ko, M.-H. Chang, J.-R. Duann, J.-Y. Chen, T.-P. Su, and T.-P. Jung, "Review of wireless and wearable electroencephalogram systems and brain-computer interfaces–a mini-review," *Gerontology*, vol. 56, no. 1, pp. 112–119, 2010.

[158] A. J. Casson, D. C. Yates, S. J. Smith, J. S. Duncan, and E. Rodriguez-Villegas, "Wearable electroencephalography," *IEEE engineering in medicine and biology magazine*, vol. 29, no. 3, pp. 44–56, 2010.

[159] J. E. Ip, "Wearable devices for cardiac rhythm diagnosis and management," *Jama*, vol. 321, no. 4, pp. 337–338, 2019.

[160] B. Jeon, J. Lee, and J. Choi, "Design and implementation of a wearable ECG system," *International Journal of Smart Home*, vol. 7, no. 2, pp. 61–69, 2013.

[161] S. Beniczky, I. Conradsen, O. Henning, M. Fabricius, and P. Wolf, "Automated real-time detection of tonic-clonic seizures using a wearable EMG device," *Neurology*, vol. 90, no. 5, pp. e428–e434, 2018.

[162] Y. Tsubouchi and K. Suzuki, "BioTones: A wearable device for EMG auditory biofeedback," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, 2010, pp. 6543–6546.

[163] V. Nathan and R. Jafari, "Particle filtering and sensor fusion for robust heart rate monitoring using wearable sensors," *IEEE journal of biomedical and health informatics*, vol. 22, no. 6, pp. 1834–1846, 2017.

[164] J.-H. Park, D.-G. Jang, J. W. Park, and S.-K. Youm, "Wearable sensing of in-ear pressure for heart rate monitoring with a piezoelectric sensor," *Sensors*, vol. 15, no. 9, pp. 23 402–23 417, 2015.

[165] F. El-Amrawy and M. I. Nounou, "Are currently available wearable devices for activity tracking and heart rate monitoring accurate, precise, and medically beneficial?" *Healthcare informatics research*, vol. 21, no. 4, p. 315, 2015.

[166] C.-W. Tsai, C.-H. Li, R. W.-K. Lam, C.-K. Li, and S. Ho, "Diabetes care in motion: Blood glucose estimation using wearable devices," *IEEE Consumer Electronics Magazine*, vol. 9, no. 1, pp. 30–34, 2019.

[167] G. Cappon, G. Acciaroli, M. Vettoretti, A. Facchinetti, and G. Sparacino, "Wearable continuous glucose monitoring sensors: a revolution in diabetes treatment," *Electronics*, vol. 6, no. 3, p. 65, 2017.

[168] J. C. Pickup, "Insulin-pump therapy for type 1 diabetes mellitus," *New England Journal of Medicine*, vol. 366, no. 17, pp. 1616–1624, 2012.

[169] J. Weissberg-Benchell, J. Antisdel-Lomaglio, and R. Seshadri, "Insulin pump therapy: a meta-analysis," *Diabetes care*, vol. 26, no. 4, pp. 1079–1087, 2003.

[170] M. Gadaleta, A. Facchinetti, E. Grisan, and M. Rossi, "Prediction of adverse glycemic events from continuous glucose monitoring signal," *IEEE journal of biomedical and health informatics*, vol. 23, no. 2, pp. 650–659, 2018.

[171] A. Angelucci, D. Kuller, and A. Aliverti, "A home telemedicine system for continuous respiratory monitoring," *IEEE Journal of Biomedical and Health Informatics*, 2020.

[172] M. Scherer, K. Menachery, and M. Magno, "SmartAid: A Low-Power Smart Hearing Aid For Stutterers," in *2019 IEEE Sensors Applications Symposium (SAS)*. IEEE, 2019, pp. 1–6.

[173] B. Sudharsan and M. Chockalingam, "A microphone array and voice algorithm

based smart hearing aid," *arXiv preprint arXiv:1908.07324*, 2019.

[174] S. DJordjevic, S. Stancin, A. Meglc, V. Milutinovic, and S. Tomazic, "Mc sensor—A novel method for measurement of muscle tension," *Sensors*, vol. 11, no. 10, pp. 9411–9425, 2011.

[175] B. Mansuri, F. Torabinejhad, A. A. Jamshidi, P. Dabirmoghaddam, B. Vasaghi-Gharamaleki, and L. Ghelichi, "Transcutaneous electrical nerve stimulation combined with voice therapy in women with muscle tension dysphonia," *Journal of Voice*, vol. 34, no. 3, pp. 490–e11, 2020.

[176] R. Velázquez, "Wearable assistive devices for the blind," in *Wearable and autonomous biomedical devices and systems for smart environment*. Springer, 2010, pp. 331–349.

[177] J. A. Garcia-Macias, A. G. Ramos, R. Hasimoto-Beltran, and S. E. P. Hernandez, "Uasisi: a modular and adaptable wearable system to assist the visually impaired," *Procedia Computer Science*, vol. 151, pp. 425–430, 2019.

[178] H. P. Savindu, K. Iroshan, C. D. Panangala, W. Perera, and A. C. De Silva, "BrailleBand: Blind support haptic wearable band for communication using braille language," in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2017, pp. 1381–1386.

[179] M. Sun, L. E. Burke, Z.-H. Mao, Y. Chen, H.-C. Chen, Y. Bai, Y. Li, C. Li, and W. Jia, "eButton: a wearable computer for health monitoring and personal assistance," in *Proceedings of the 51st annual design automation conference*, 2014, pp. 1–6.

[180] A. Kapur, S. Kapur, and P. Maes, "Alterego: A personalized wearable silent speech interface," in *23rd International conference on intelligent user interfaces*, 2018, pp. 43–53.

[181] N. Marjanovic, G. Piccinini, K. Kerr, and H. Esmailbeigi, "TongueToSpeech (TTS): Wearable wireless assistive device for augmented speech," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2017, pp. 3561–3563.

[182] W. Huo, S. Mohammed, J. C. Moreno, and Y. Amirat, "Lower limb wearable robots for assistance and rehabilitation: A state of the art," *IEEE systems Journal*, vol. 10,

no. 3, pp. 1068–1081, 2014.

[183] A. Hadi, K. Alipour, S. Kazeminasab, and M. Elahinia, "ASR glove: A wearable glove for hand assistance and rehabilitation using shape memory alloys," *Journal of Intelligent Material Systems and Structures*, vol. 29, no. 8, pp. 1575–1585, 2018.

[184] M. Gandolla, A. Antonietti, V. Longatelli, and A. Pedrocchi, "The effectiveness of wearable upper limb assistive devices in degenerative neuromuscular diseases: A systematic review and meta-analysis," *Frontiers in bioengineering and biotechnology*, vol. 7, p. 450, 2020.

[185] B. Chen, C.-H. Zhong, X. Zhao, H. Ma, X. Guan, X. Li, F.-Y. Liang, J. C. Y. Cheng, L. Qin, S.-W. Law *et al.*, "A wearable exoskeleton suit for motion assistance to paralysed patients," *Journal of orthopaedic translation*, vol. 11, pp. 7–18, 2017.

[186] Y. S. Delahoz and M. A. Labrador, "Survey on fall detection and fall prevention using wearable and external sensors," *Sensors*, vol. 14, no. 10, pp. 19 806–19 842, 2014.

[187] D. Chen, W. Feng, Y. Zhang, X. Li, and T. Wang, "A wearable wireless fall detection system with accelerators," in *2011 IEEE international conference on robotics and biomimetics*. IEEE, 2011, pp. 2259–2263.

[188] W.-J. Yi and J. Saniie, "Design flow of a wearable system for body posture assessment and fall detection with android smartphone," in *2014 IEEE international technology management conference*. IEEE, 2014, pp. 1–4.

[189] E. Bruno, A. Biondi, S. Thorpe, M. Richardson, R.-C. Consortium *et al.*, "Patients self-mastery of wearable devices for seizure detection: a direct user-experience," *Seizure*, vol. 81, pp. 236–240, 2020.

[190] J. Jeppesen, A. Fuglsang-Frederiksen, P. Johansen, J. Christensen, S. Wüstenhagen, H. Tankisi, E. Qerama, A. Hess, and S. Beniczky, "Seizure detection based on heart rate variability using a wearable electrocardiography device," *Epilepsia*, vol. 60, no. 10, pp. 2105–2113, 2019.

[191] P. Pierleoni, A. Belli, L. Palma, M. Pellegrini, L. Pernini, and S. Valenti, "A high reliability wearable device for elderly fall detection," *IEEE Sensors Journal*, vol. 15, no. 8, pp. 4544–4553, 2015.

[192] L. Atallah, B. Lo, R. King, and G.-Z. Yang, "Sensor positioning for activity recognition using wearable accelerometers," *IEEE transactions on biomedical circuits and systems*, vol. 5, no. 4, pp. 320–329, 2011.

[193] H. Ouyang, Z. Liu, N. Li, B. Shi, Y. Zou, F. Xie, Y. Ma, Z. Li, H. Li, Q. Zheng *et al.*, "Symbiotic cardiac pacemaker," *Nature communications*, vol. 10, no. 1, pp. 1–10, 2019.

[194] A. Eicken, C. Kolb, S. Lange, S. Brodherr-Heberlein, B. Zrenner, C. Schreiber, and J. Hess, "Implantable cardioverter defibrillator (ICD) in children," *International journal of cardiology*, vol. 107, no. 1, pp. 30–35, 2006.

[195] S. van der Kroft, "Design and validation of an implantable actuator for use in a novel dynamic arteriovenous shunt system," 2021.

[196] K. Shiba, T. Tsuji, and K. Koshiji, "Direct drive of an implantable actuator using a transcutaneous energy transmission system," *Journal of Life Support Engineering*, vol. 18, no. 1, pp. 17–24, 2006.

[197] "BAN Applications Matrix, Document 15-07-0735-08-0," 2008.

[198] G. Rong, Y. Zheng, and M. Sawan, "Energy Solutions for Wearable Sensors: A Review," *Sensors*, vol. 21, no. 11, p. 3806, 2021.

[199] A. Kos, V. Milutinović, and A. Umek, "Challenges in wireless communication for connected sensors and wearable devices used in sport biofeedback applications," *Future generation computer systems*, vol. 92, pp. 582–592, 2019.

[200] S. Ullah, P. Khan, N. Ullah, S. Saleem, H. Higgins, and K. S. Kwak, "A review of wireless body area networks for medical applications," *arXiv preprint arXiv:1001.0831*, 2010.

[201] S. Movassaghi, M. Abolhasan, J. Lipman, D. Smith, and A. Jamalipour, "Wireless body area networks: A survey," *IEEE Communications surveys & tutorials*, vol. 16, no. 3, pp. 1658–1686, 2014.

[202] "TG6 Applications Matrix, Document 15-08-0406-00-0006, IEEE P802," 2008.

[203] R. W. Jones and K. Katzis, "5G and wireless body area networks," in *2018*

IEEE Wireless Communications and Networking Conference Workshops (WCNCW), 2018, pp. 373–378.

[204] P. J. Soh, G. A. Vandenbosch, M. Mercuri, and D. M.-P. Schreurs, "Wearable wireless health monitoring: Current developments, challenges, and future trends," *IEEE Microwave Magazine*, vol. 16, no. 4, pp. 55–70, 2015.

[205] G. E. Santagati and T. Melodia, "A software-defined ultrasonic networking framework for wearable devices," *IEEE/ACM Transactions on Networking*, vol. 25, no. 2, pp. 960–973, 2016.

[206] C. Garcia-Perez, A. Diaz-Zayas, A. Rios, P. Merino, K. Katsalis, C.-Y. Chang, S. Shariat, N. Nikaein, P. Rodriguez, and D. Morris, "Improving the efficiency and reliability of wearable based mobile eHealth applications," *Pervasive and Mobile Computing*, vol. 40, pp. 674–691, 2017.

[207] Y. Sahni, J. Cao, S. Zhang, and L. Yang, "Edge mesh: A new paradigm to enable distributed intelligence in internet of things," *IEEE access*, vol. 5, pp. 16 441–16 458, 2017.

[208] A. A. Mutlag, M. K. Abd Ghani, N. a. Arunkumar, M. A. Mohammed, and O. Mohd, "Enabling technologies for fog computing in healthcare IoT systems," *Future Generation Computer Systems*, vol. 90, pp. 62–78, 2019.

[209] T. N. Gia, M. Jiang, A.-M. Rahmani, T. Westerlund, P. Liljeberg, and H. Tenhunen, "Fog computing in healthcare internet of things: A case study on ecg feature extraction," in *2015 IEEE international conference on computer and information technology; ubiquitous computing and communications; dependable, autonomic and secure computing; pervasive intelligence and computing*. IEEE, 2015, pp. 356–363.

[210] CEN/CENELEC, "Medical devices - Application of risk management to medical devices (ISO 14971:2019."

[211] FDA, "Content of Premarket Submissions for Management of Cybersecurity in Medical Devices: Draft Guidance for Industry and Food and Drug Administration Staff ," Accessed on: September 25, 2020 [Online]. Available: https://www.fda.gov/media/119933/download.

[212] M. Motani, K.-K. Yap, A. Natarajan, B. de Silva, S. Hu, and K. C. Chua, "Network characteristics of urban environments for wireless BAN," in *2007 IEEE Biomedical*

Circuits and Systems Conference, 2007, pp. 179–182.

[213] M. O. Al Kalaa, W. Balid, H. H. Refai, N. J. LaSorte, S. J. Seidman, H. I. Bassen, J. L. Silberberg, and D. Witters, "Characterizing the 2.4 GHz spectrum in a hospital environment: Modeling and applicability to coexistence testing of medical devices," *IEEE Transactions on Electromagnetic Compatibility*, vol. 59, no. 1, pp. 58–66, 2016.

[214] Qualcomm Technologies, Inc.), "VR and AR pushing connectivity limits ," Accessed on: June 25, 2021 [Online]. Available: https://www.qualcomm.com/media/documents/files/vr-and-ar-pushing-connectivity-limits.pdf.

[215] A. P. Pozo, M. Toksvig, T. F. Schrager, J. Hsu, U. Mathur, A. Sorkine-Hornung, R. Szeliski, and B. Cabral, "An integrated 6DoF video camera and system design," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 6, pp. 1–16, 2019.

[216] N. Pavón-Pulido, J. A. López-Riquelme, J. Ferruz-Melero, M. Á. Vega-Rodríguez, and A. J. Barrios-León, "A service robot for monitoring elderly people in the context of ambient assisted living," *Journal of Ambient Intelligence and Smart Environments*, vol. 6, no. 6, pp. 595–621, 2014.

[217] M. Bonaccorsi, L. Fiorini, F. Cavallo, R. Esposito, and P. Dario, "Design of cloud robotic services for senior citizens to improve independent living and personal health management," in *Ambient Assisted Living*. Springer, 2015, pp. 465–475.

[218] Y. Ma, Y. Zhang, J. Wan, D. Zhang, and N. Pan, "Robot and cloud-assisted multimodal healthcare system," *Cluster Computing*, vol. 18, no. 3, pp. 1295–1306, 2015.

[219] H.-M. Gross, S. Mueller, C. Schroeter, M. Volkhardt, A. Scheidig, K. Debes, K. Richter, and N. Doering, "Robot companion for domestic health assistance: Implementation, test and case study under everyday conditions in private apartments," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 5992–5999.

[220] A. Manzi, L. Fiorini, R. Limosani, P. Sincak, P. Dario, and F. Cavallo, "Use case evaluation of a cloud robotics teleoperation system (short paper)," in *2016 5th IEEE International Conference on Cloud Networking (Cloudnet)*. IEEE, 2016, pp. 208–211.

[221] M. Bonaccorsi, L. Fiorini, F. Cavallo, A. Saffiotti, and P. Dario, "A cloud robotics

solution to improve social assistive robots for active and healthy aging," *International Journal of Social Robotics*, vol. 8, no. 3, pp. 393–408, 2016.

[222] L. Fiorini, R. Esposito, M. Bonaccorsi, C. Petrazzuolo, F. Saponara, R. Giannantonio, G. De Petris, P. Dario, and F. Cavallo, "Enabling personalised medical support for chronic disease management through a hybrid robot-cloud approach," *Autonomous Robots*, vol. 41, no. 5, pp. 1263–1276, 2017.

[223] T. Cádrik, P. Takáč, J. Ondo, P. Sinčák, M. Mach, F. Jakab, F. Cavallo, and M. Bonaccorsi, "Cloud-based robots and intelligent space teleoperation tools," in *Robot Intelligence Technology and Applications 4*. Springer, 2017, pp. 599–610.

[224] F. Cavallo, R. Limosani, L. Fiorini, R. Esposito, R. Furferi, L. Governi, and M. Carfagni, "Design impact of acceptability and dependability in assisted living robotic applications," *International Journal on Interactive Design and Manufacturing (IJIDeM)*, vol. 12, no. 4, pp. 1167–1178, 2018.

[225] A. Brunete, E. Gambao, M. Hernando, and R. Cedazo, "Smart Assistive Architecture for the Integration of IoT Devices, Robotic Systems, and Multimodal Interfaces in Healthcare Environments," *Sensors*, vol. 21, no. 6, p. 2212, 2021.

[226] M. Tröbinger, C. Jähne, Z. Qu, J. Elsner, A. Reindl, S. Getz, T. Goll, B. Loinger, T. Loibl, C. Kugler *et al.*, "Introducing GARMI-A Service Robotics Platform to Support the Elderly at Home: Design Philosophy, System Overview and First Results," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 5857–5864, 2021.

[227] D. Feil-Seifer and M. J. Mataric, "Defining socially assistive robotics," in *9th International Conference on Rehabilitation Robotics, 2005. ICORR 2005.* IEEE, 2005, pp. 465–468.

[228] K. Witrisal, P. Meissner, E. Leitinger, Y. Shen, C. Gustafson, F. Tufvesson, K. Haneda, D. Dardari, A. F. Molisch, A. Conti *et al.*, "High-accuracy localization for assisted living: 5G systems will turn multipath channels from foe to friend," *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 59–70, 2016.

[229] RADIO Project, "Unobtrusive, Efficient, Reliable and Modular Solutions for Independent Ageing ," Accessed on: June 24, 2021 [Online]. Available: http://www.radio-project.eu/.

[230] N. Ramoly, A. Bouzeghoub, and B. Finance, "A framework for service robots in

smart home: an efficient solution for domestic healthcare," *IRBM*, vol. 39, no. 6, pp. 413–420, 2018.

[231] S. Kaneriya, J. Vora, S. Tanwar, and S. Tyagi, "Standardising the use of duplex channels in 5G-WiFi networking for ambient assisted living," in *2019 IEEE international conference on communications workshops (ICC workshops)*. IEEE, 2019, pp. 1–6.

[232] S. Henry, A. Alsohaily, and E. S. Sousa, "5G is real: Evaluating the compliance of the 3GPP 5G new radio system with the ITU IMT-2020 requirements," *IEEE Access*, vol. 8, pp. 42 828–42 840, 2020.

[233] European 5G Observatory), "5G trials that have been publicly announced in EU27, UK, Norway, Russia, Switzerland and Turkey ," Accessed on: June 25, 2021 [Online]. Available: https://5gobservatory.eu/5g-trial/major-european-5g-trials-and-pilots/.

[234] sdx central), "5G Trials in the United States — Steps Toward Standardization ," Accessed on: June 25, 2021 [Online]. Available: https://www.sdxcentral.com/5g/definitions/5g-trials/.

[235] Verizon), "Verizon will rapidly integrate C-band spectrum with mmWave for customers ," Accessed on: June 25, 2021 [Online]. Available: https://www.verizon.com/about/news/verizon-c-band-spectrum-mmwave.

[236] Verizon, "Explore 4G LTE and 5G network coverage in your area ," Accessed on: July 26, 2021 [Online]. Available: https://www.verizon.com/coverage-map/.

[237] AT&T, "Wireless coverage ," Accessed on: July 26, 2021 [Online]. Available: https://www.att.com/maps/wireless-coverage.html.

[238] T-Mobile, "Coverage maps ," Accessed on: July 26, 2021 [Online]. Available: https://www.t-mobile.com/coverage/coverage-map.

[239] Opensignal), "How AT&T, Sprint, T-Mobile and Verizon differ in their early 5G approach ," Accessed on: June 25, 2021 [Online]. Available: https://www.opensignal.com/2020/02/20/how-att-sprint-t-mobile-and-verizon-differ-in-their-early-5g-approach.

[240] Digital Trends), "5G vs. 4G: How will the newest network improve on the last? ," Accessed on: June 25, 2021 [Online]. Available: https://www.digitaltrends.com/mobile/5g-vs-4g/.

[241] Forbes), "5G Latency Improvements Are Still Lagging ," Accessed on: June 25, 2021 [Online]. Available: https://www.forbes.com/sites/bobodonnell/2020/02/18/5g-latency-improvements-are-still-lagging/?sh=6d74337548f1.

[242] A. Asghar, H. Farooq, and A. Imran, "Concurrent CCO and LB optimization in emerging HetNets: a novel solution and comparative analysis," in *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE, 2018, pp. 1–6.

[243] 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*, vol. 55, no. 6, pp. 2805–2821, 2019.

[244] S.-H. Park, N.-G. Kang, C. Cho, E.-T. Won, R. K. Patro, G. Goyal, A. Bhatia, K. Bynam, and A. Naniyat, "System Simulation Metrics for BAN - Samsung," *Project: IEEE P802.15 Working Group for Wireless Personal Area Networks (WPANs)*, 2008.

[245] 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.

[246] 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.

[247] 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.

[248] 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.

[249] 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.

[250] 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.

[251] 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.

[252] 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.

[253] 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.

[254] 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.

[255] 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.

[256] 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.

[257] 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.

[258] F. Sohrabi and E. Kuehn, "Construction of the RSRP map using sparse MDT mea-

surements by regression clustering," in *IEEE International Conference on Communications (ICC)*, 2017, pp. 1–6.

[259] 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.

[260] 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.

[261] 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.

[262] 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.

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

[264] 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.

[265] 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.

[266] H. Braham, S. B. Jemaa, B. Sayrac, G. Fort, and E. Moulines, "Low complexity spatial interpolation for cellular coverage analysis," in *2014 12th International Sym-*

posium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), 2014, pp. 188–195.

[267] ——, "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.

[268] 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.

[269] 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.

[270] 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.

[271] 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.

[272] 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.

[273] 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.

[274] "Atoll, [online] available:https://www.forsk.com/."

[275] 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.

[276] 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.

[277] 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.

[278] 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.

[279] 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.

[280] 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.

[281] 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.

[282] E. Balevi and J. G. Andrews, "Online antenna tuning in heterogeneous cellular networks with deep reinforcement learning," *arXiv preprint arXiv:1903.06787*, 2019.

[283] 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.

[284] 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.

[285] 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.

[286] 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.

[287] 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.

[288] L. Mitas and H. Mitasova, "Spatial interpolation," *Geographical information systems: principles, techniques, management and applications*, vol. 1, no. 2, 1999.

[289] F. Susanto, P. de Souza, and J. He, "Spatiotemporal interpolation for environmental modelling," *Sensors*, vol. 16, no. 8, p. 1245, 2016.

[290] J. Li and A. D. Heap, "A review of spatial interpolation methods for environmental scientists," 2008.

[291] 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.

[292] 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.

[293] 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.

[294] 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.

[295] M. Angjelicinoski, V. Atanasovski, and L. Gavrilovska, "Comparative analysis of spatial interpolation methods for creating radio environment maps," in *2011 IEEE 19thTelecommunications Forum (TELFOR) Proceedings of Papers*, 2011, pp. 334–337.

[296] 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.

[297] M. Höyhtyä, 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.

[298] 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.

[299] 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.

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

[301] 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.

[302] 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.

[303] 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.

[304] D. Shepard, "A two-dimensional interpolation function for irregularly-spaced data," in *Proceedings of the 1968 23rd ACM national conference*, 1968, pp. 517–524.

[305] R. Franke and G. M. Nielson, "Scattered data interpolation and applications: A tutorial and survey," in *Geometric Modeling*. Springer, 1991, pp. 131–160.

[306] F. Susanto, P. de Souza, and J. He, "Spatiotemporal interpolation for environmental

modelling," *Sensors*, vol. 16, no. 8, p. 1245, 2016.

[307] S. Henley, "Nonparamteric geostatistics," p. 145, 1981.

[308] D. Weber and E. Englund, "Evaluation and comparison of spatial interpolators," *Mathematical Geology*, vol. 24, no. 4, pp. 381–391, 1992.

[309] 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.

[310] 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.

[311] 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.

[312] 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.

[313] E. Balevi and J. G. Andrews, "Online antenna tuning in heterogeneous cellular networks with deep reinforcement learning," *arXiv preprint arXiv:1903.06787*, 2019.

[314] 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.

[315] 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.

[316] 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.

[317] 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.

[318] 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.

[319] H. F. Syed Muhammad Asad Zaidi, Marvin Manalastas and A. Imran, "AI4Networks Simulator - A True 3GPP Compliant 5G Network Simulator with Support of AI," *IEEE Access (Submitted)*, 2019.

[320] 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.

[321] OpenAirInterface, "OpenAirInterface: 5G Software Alliance for Democratising Wireless Innovation," Tech. Rep. [Online]. Available: http://opnetprojects.com/opnet-simulator/

[322] 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.

[323] 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.

[324] 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.

[325] ns 3, "mmWave Cellular Network Simulator," Tech. Rep. [Online]. Available: https://omnetpp.org/

[326] OMNeT++, "OMNeT++: Discrete Event Simulator," Tech. Rep. [Online]. Available: https://apps.nsnam.org/app/mmwave/

[327] 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.

[328] Matlab, "Why Use MATLAB and Simulink for 5G?" Tech. Rep. [Online]. Available: https://www.mathworks.com/solutions/wireless-communications/5g.html

[329] 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.

[330] OPNET, "OPNET: Optimum Network Performance," Tech. Rep. [Online]. Available: https://www.openairinterface.org/

[331] 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.

[332] A. Antoniou, A. Storkey, and H. Edwards, "Data augmentation generative adversarial networks," *arXiv preprint arXiv:1711.04340*, 2017.

[333] 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.

[334] 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.

[335] H. Huang, P. S. Yu, and C. Wang, "An introduction to image synthesis with generative adversarial nets," *arXiv preprint arXiv:1803.04469*, 2018.

[336] 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.

[337] T. Zhang, K. Zhu, and D. Niyato, "A Generative Adversarial Learning-Based Approach for Cell Outage Detection in Self-Organizing Cellular Networks," *IEEE Wireless Communications Letters*, vol. 9, no. 2, pp. 171–174, Feb. 2020, conference Name: IEEE Wireless Communications Letters.

[338] X. Han, L. Xue, Y. Xu, and Z. Liu, "A Radio Environment Maps Estimation Algorithm based on the Pixel Regression Framework for Underlay Cognitive Radio Networks Using Incomplete Training Data," *Sensors*, vol. 20, no. 8, p. 2245, Jan. 2020, number: 8 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/20/8/2245

[339] ——, "A Two-Phase Transfer Learning-Based Power Spectrum Maps Reconstruction Algorithm for Underlay Cognitive Radio Networks," *IEEE Access*, vol. 8, pp. 81 232–81 245, 2020, conference Name: IEEE Access.

[340] T. Erpek, Y. E. Sagduyu, and Y. Shi, "Deep Learning for Launching and Mitigating Wireless Jamming Attacks," *arXiv:1807.02567 [cs, stat]*, Dec. 2018, arXiv: 1807.02567. [Online]. Available: http://arxiv.org/abs/1807.02567

[341] Y. Shi, K. Davaslioglu, and Y. E. Sagduyu, "Generative Adversarial Network for Wireless Signal Spoofing," *arXiv:1905.01008 [cs, eess, stat]*, May 2019, arXiv: 1905.01008. [Online]. Available: http://arxiv.org/abs/1905.01008

[342] M. Nabati, H. Navidan, R. Shahbazian, S. A. Ghorashi, and D. Windridge, "Using Synthetic Data to Enhance the Accuracy of Fingerprint-Based Localization: A Deep Learning Approach," *IEEE Sensors Letters*, vol. 4, no. 4, pp. 1–4, Apr. 2020, conference Name: IEEE Sensors Letters.

[343] C. Zhang, X. Ouyang, and P. Patras, "ZipNet-GAN: Inferring Fine-grained Mobile Traffic Patterns via a Generative Adversarial Neural Network," *arXiv:1711.02413 [cs]*, Nov. 2017, arXiv: 1711.02413. [Online]. Available: http://arxiv.org/abs/1711.02413

[344] J. Engelmann and S. Lessmann, "Conditional Wasserstein GAN-based Oversampling of Tabular Data for Imbalanced Learning," *arXiv:2008.09202 [cs]*, Aug. 2020, arXiv: 2008.09202. [Online]. Available: http://arxiv.org/abs/2008.09202

[345] F. H. K. d. S. Tanaka and C. Aranha, "Data Augmentation Using GANs," *arXiv:1904.09135 [cs, stat]*, Apr. 2019, arXiv: 1904.09135. [Online]. Available: http://arxiv.org/abs/1904.09135

[346] N. Gao, H. Xue, W. Shao, S. Zhao, K. K. Qin, A. Prabowo, M. S. Rahaman, and F. D. Salim, "Generative Adversarial Networks for Spatio-temporal Data: A Survey," *arXiv:2008.08903 [cs, eess]*, Aug. 2020, arXiv: 2008.08903. [Online]. Available: http://arxiv.org/abs/2008.08903

[347] R. D. Camino, C. A. Hammerschmidt, and R. State, "Improving Missing Data Imputation with Deep Generative Models," *arXiv:1902.10666 [cs, stat]*, Feb. 2019, arXiv: 1902.10666. [Online]. Available: http://arxiv.org/abs/1902.10666

[348] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling Tabular data using Conditional GAN," *arXiv:1907.00503 [cs, stat]*, Oct. 2019, arXiv: 1907.00503. [Online]. Available: http://arxiv.org/abs/1907.00503

[349] L. Xu and K. Veeramachaneni, "Synthesizing Tabular Data using Generative Adversarial Networks," *arXiv:1811.11264 [cs, stat]*, Nov. 2018, arXiv: 1811.11264. [Online]. Available: http://arxiv.org/abs/1811.11264

[350] 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.

[351] 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.

[352] S. Rajendran, W. Meert, V. Lenders, and S. Pollin, "SAIFE: Unsupervised Wireless Spectrum Anomaly Detection with Interpretable Features," in *2018 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, Oct. 2018, pp. 1–9, iSSN: 2334-3125.

[353] Y. Teganya and D. Romero, "Deep completion autoencoders for radio map estimation," *arXiv preprint arXiv:2005.05964*, 2020.

[354] ——, "Data-driven spectrum cartography via deep completion autoencoders," in *ICC 2020-2020 IEEE International Conference on Communications (ICC)*. IEEE, 2020, pp. 1–7.

[355] 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, Oct. 2010, con-

ference Name: IEEE Transactions on Knowledge and Data Engineering.

[356] A. Zappone, M. Di Renzo, and M. Debbah, "Wireless Networks Design in the Era of Deep Learning: Model-Based, AI-Based, or Both?" *IEEE Transactions on Communications*, vol. 67, no. 10, pp. 7331–7376, Oct. 2019, conference Name: IEEE Transactions on Communications.

[357] E. Baştuğ, M. Bennis, and M. Debbah, "A transfer learning approach for cache-enabled wireless networks," in *2015 13th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, May 2015, pp. 161–166.

[358] M. Chen, W. Saad, C. Yin, and M. Debbah, "Data Correlation-Aware Resource Management in Wireless Virtual Reality (VR): An Echo State Transfer Learning Approach," *IEEE Transactions on Communications*, vol. 67, no. 6, pp. 4267–4280, Jun. 2019, conference Name: IEEE Transactions on Communications.

[359] J. Chuai, Z. Chen, G. Liu, X. Guo, X. Wang, X. Liu, C. Zhu, and F. Shen, "A Collaborative Learning Based Approach for Parameter Configuration of Cellular Networks," in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, Apr. 2019, pp. 1396–1404, iSSN: 2641-9874.

[360] R. Li, Z. Zhao, X. Chen, J. Palicot, and H. Zhang, "TACT: A Transfer Actor-Critic Learning Framework for Energy Saving in Cellular Radio Access Networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 4, pp. 2000–2011, Apr. 2014, conference Name: IEEE Transactions on Wireless Communications.

[361] C. Parera, Q. Liao, I. Malanchini, C. Tatino, A. E. C. Redondi, and M. Cesana, "Transfer Learning for Tilt-Dependent Radio Map Prediction," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 2, pp. 829–843, Jun. 2020, conference Name: IEEE Transactions on Cognitive Communications and Networking.

[362] C. Parera, A. E. Redondi, M. Cesana, Q. Liao, and I. Malanchini, "Transfer Learning for Channel Quality Prediction," in *2019 IEEE International Symposium on Measurements Networking (M N)*, Jul. 2019, pp. 1–6, iSSN: 2639-5061.

[363] M. S. Abrishami, A. E. Eshratifar, D. Eigen, Y. Wang, S. Nazarian, and M. Pedram, "Efficient Training of Deep Convolutional Neural Networks by Augmentation in Embedding Space," in *2020 21st International Symposium on Quality Electronic Design (ISQED)*, Mar. 2020, pp. 347–351, iSSN: 1948-3287.

[364] W. Fang, C. Chen, B. Song, L. Wang, J. Zhou, and K. Q. Zhu, "Adapted tree boosting for Transfer Learning," *arXiv:2002.11982 [cs, stat]*, Apr. 2020, arXiv: 2002.11982. [Online]. Available: http://arxiv.org/abs/2002.11982

[365] Y. Wang, Q. Yao, J. Kwok, and L. M. Ni, "Generalizing from a Few Examples: A Survey on Few-Shot Learning," *arXiv:1904.05046 [cs]*, Mar. 2020, arXiv: 1904.05046. [Online]. Available: http://arxiv.org/abs/1904.05046

[366] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," *arXiv preprint arXiv:1703.03400*, 2017.

[367] M. A. Jamal and G.-J. Qi, "Task agnostic meta-learning for few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 719–11 727.

[368] Q. Sun, Y. Liu, T.-S. Chua, and B. Schiele, "Meta-transfer learning for few-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 403–412.

[369] C. Finn and S. Levine, "Meta-learning: from few-shot learning to rapid reinforcement learning," in *ICML*, 2019.

[370] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, vol. 2. Lille, 2015.

[371] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," in *Advances in neural information processing systems*, 2016, pp. 3630–3638.

[372] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in neural information processing systems*, 2017, pp. 4077–4087.

[373] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1199–1208.

[374] V. Garcia and J. Bruna, "Few-shot learning with graph neural networks," *arXiv preprint arXiv:1711.04043*, 2017.

[375] S. Aoki, K. Shiomoto, C. L. Eng, and S. Backstad, "Few-shot Learning for eNodeB Performance Metric Analysis for Service Level Assurance in LTE Networks," in *NOMS 2020 - 2020 IEEE/IFIP Network Operations and Management Symposium*, Apr. 2020, pp. 1–4, iSSN: 2374-9709.

[376] Y. Shen, Y. Shi, J. Zhang, and K. B. Letaief, "LORM: Learning to Optimize for Resource Management in Wireless Networks With Few Training Samples," *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 665–679, Jan. 2020, conference Name: IEEE Transactions on Wireless Communications.

[377] J. Yang, H. Zou, Y. Zhou, and L. Xie, "Learning Gestures From WiFi: A Siamese Recurrent Convolutional Architecture," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10 763–10 772, Dec. 2019, conference Name: IEEE Internet of Things Journal.

[378] D. Giatsios, "FLEX - FIRE LTE testbeds for open experimentation: Flex overview," in *3RD INTERNATIONAL NORNET USERS WORKSHOP, OSLO*, 2015.

[379] "Nitos - network implementation testbed using open source platforms, [online] available: http://nitlab.inf.uth.gr."

[380] "5g virtual infrastructure provisioning over nitos testbed, [online] available: https://5ginfire.eu/nitos/."

[381] K. Kondepu, F. Giannone, S. Vural, B. Riemer, P. Castoldi, and L. Valcarenghi, "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.

[382] "5G innovation centre, university of surrey, [online] available: https://www.surrey.ac.uk/5gic."

[383] 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.

[384] *Surrey Platform.* [Online]. Available: https://5genesis.eu/surrey-platform/

[385] 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.

[386] 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.

[387] "Open-access research testbed for next-generation wireless networks (orbit) , [online] available: http://www.orbit-lab.org/."

[388] "Open-access research testbed for next-generation wireless networks (orbit) , [online] available: https://www.phantomnet.org/."

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

[390] E. Luther, "5G massive MIMO testbed: From theory to reality," *white paper*, 2014.

[391] 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.

[392] 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.

[393] S. Mattisson, "Overview of 5G requirements and future wireless networks," in *ESSCIRC 2017-43rd IEEE European Solid State Circuits Conference*, 2017, pp. 1–6.

[394] 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.

[395] J. L. Jermyn, "Discovering network control vulnerabilities and policies in evolving networks," Ph.D. dissertation, Columbia University, 2017.

[396] 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.

[397] "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."

[398] 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.

[399] "5g testbed, [online] available: https://cewit.org.in/testbed/."

[400] *5G Ready Trial Platform*. [Online]. Available: https://www.nutaq.com/products/titanmimo/titanmimo-6/technology

[401] 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

[402] *ESPOO Aalto 5G research infrastructure*. [Online]. Available: http://5gtnf.fi/sites/espoo/

[403] *HELSINKI*. [Online]. Available: http://5gtnf.fi/sites/helsinki/

[404] *VODAFONE CHAIR MOBILE COMMUNICATION SYSTEMS*. [Online]. Available: https://www.vodafone-chair.org/

[405] 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.

[406] 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.

[407] *VODAFONE CHAIR & RESEARCH*. [Online]. Available: https://www.vodafone-chair.org/chair+research#projects

[408] 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.

[409] 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.

[410] *5G Playground*. [Online]. Available: https://www.fokus.fraunhofer.de/go/en/fokus_testbeds/5g_playground

[411] *TAMPERE Tampere University wireless test networks (Hervanta)*. [Online]. Available: http://5gtnf.fi/sites/tampere/

[412] 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.

[413] *FUSECO Playground*. [Online]. Available: https://www.fokus.fraunhofer.de/go/en/fokus_testbeds/fuseco_playground

[414] *5G Ready Trial Platform*. [Online]. Available: https://www.fokus.fraunhofer.de/go/en/5GRTP

[415] *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

[416] A. E. Bjorn Halvarsson, Arne Simonsson, R. Chana, P. Machado, and H. Asplund, "5g nr testbed 3.5 ghz coverage results," in *IEEE 87th Vehicular Technology Conference: VTC2018-Spring*, 2018.

[417] *SK Telecom*. [Online]. Available: https://www.sktelecom.com/index.html

[418] *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

[419] "SK telecom's 5G architecture design and implementation guidelines (Version 1.35)," *5G Tech Lab Corporate R&D Center, SK telecom*, 2015.

[420] *OULU 5GTN (Linnanmaa)*. [Online]. Available: http://5gtnf.fi/sites/oulu/

[421] 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.

[422] 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.

[423] *TurboRAN*. [Online]. Available: http://bsonlab.com/TurboRAN/

[424] "Openairinterface testbed, [online] available: https://oailab.eurecom.fr/oai-testbed."

[425] 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.

[426] "Openairinterface massive mimo testbed : A 5G innovation platform, [online] available: https://www.openairinterface.org/."

[427] 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.

[428] *Munich Experimentation Facility Site*. [Online]. Available: https://www.5g-vinni.eu/munich-experimentation-facility-site/

[429] 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.

[430] "Performnetworks testbed, [online] available: http://morse.uma.es/performnetworks."

[431] 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.

[432] A. Diaz, C. A. Garcia-Perez, A. Martin, P. Merino, and A. Rios, "Performnetworks: a testbed for exhaustive interoperability and performance analysis for mobile networks," *Building the Future Internet Through FIRE, River Publishers*, pp. 1–250, 2017.

[433] *YLIVIESKA Centria University of Applied Sciences test network*. [Online]. Available: http://5gtnf.fi/sites/ylivieska/

[434] 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.

[435] "w-ilab.t (iminds), [online] available: http://www.crew-project.eu/wilabt.html."

[436] B. J. Bouckaert Stefan, Wim Vandenberghe, I. Moaerman, and P. Demesster, "The wilab. t testbed," in *International Conference on Testbeds and Research Infrastructures*, Springer Berlin Heidelberg, 2010, pp. 145–154.

[437] *5TONIC: an open research and innovation laboratory focusing on 5G technologies*. [Online]. Available: https://www.5tonic.org

[438] *UNIVERSITY OF BRISTOL 5G TESTBED*. [Online]. Available: https://5ginfire.eu/university-of-bristol-5g-testbed/

[439] "Ericsson D-15 Labs, [online] Available: https://www.ericsson.com/en/about-us/experience-centers/d-15/ericsson-d-15-labs ."

[440] "Accessing the 5G innovation platform as a service (IPAAS) testbed, [online] available: https://ontario.encqor.ca/accessing-5g-innovation-platform-as-a-service-ipaas-testbed/."

[441] C. Parera, A. E. C. Redondi, M. Cesana, Q. Liao, L. Ewe, and C. Tatino, "Trans-

ferring knowledge for tilt-dependent radio map prediction," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, Apr. 2018, pp. 1–6, iSSN: 1558-2612.

[442] Y. Teganya and D. Romero, "Data-Driven Spectrum Cartography via Deep Completion Autoencoders," in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, Jun. 2020, pp. 1–7, iSSN: 1938-1883.

[443] ——, "Deep Completion Autoencoders for Radio Map Estimation," *arXiv:2005.05964 [eess]*, May 2020, arXiv: 2005.05964. [Online]. Available: http://arxiv.org/abs/2005.05964

[444] K. Tsukamoto, M. Kitsunezuka, and K. Kunihiro, "Highly accurate radio environment mapping method based on transmitter localization and spatial interpolation in urban LoS/NLoS scenario," in *2018 IEEE Topical Conference on Wireless Sensors and Sensor Networks (WiSNet)*, Jan. 2018, pp. 5–7, iSSN: 2473-4624.

[445] 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, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcm.2367. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/wcm.2367

[446] G. Sun and J. van de Beek, "Simple distributed interference source localization for radio environment mapping," in *2010 IFIP Wireless Days*, Oct. 2010, pp. 1–5, iSSN: 2156-972X.

[447] 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, Mar. 2015. [Online]. Available: https://doi.org/10.1186/s13638-015-0297-2

[448] N. Mezhoud, M. Oussalah, A. Zaatri, and Z. Hammoudi, "Hybrid Kriging and multilayer perceptron neural network technique for coverage prediction in cellular networks," *International Journal of Parallel, Emergent and Distributed Systems*, vol. 0, no. 0, pp. 1–25, Aug. 2020, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/17445760.2020.1805609. [Online]. Available: https://doi.org/10.1080/17445760.2020.1805609

[449] A. M. Alam, S. Benjemaa, and T. Romary, "Clustering for High Accuracy Coverage

Mapping," in *2018 IEEE International Conference on Communications (ICC)*, May 2018, pp. 1–6, iSSN: 1938-1883.

[450] 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)*, Oct. 2018, pp. 1–10, iSSN: 2334-3125.

[451] A. M. Alam, S. Benjemaa, and T. Romary, "Performance Evaluation of Covariance Tapering for Coverage Mapping," in *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, Jun. 2018, pp. 1–5, iSSN: 2577-2465.

[452] H. Xia, S. Zha, J. Huang, and J. Liu, "Radio environment map construction by adaptive ordinary Kriging algorithm based on affinity propagation clustering," *International Journal of Distributed Sensor Networks*, vol. 16, no. 5, p. 1550147720922484, May 2020, publisher: SAGE Publications. [Online]. Available: https://doi.org/10.1177/1550147720922484

[453] Z. Han, J. Liao, Q. Qi, H. Sun, and J. Wang, "Radio Environment Map Construction by Kriging Algorithm Based on Mobile Crowd Sensing," Feb. 2019, iSSN: 1530-8669 Pages: e4064201 Publisher: Hindawi Volume: 2019. [Online]. Available: https://www.hindawi.com/journals/wcmc/2019/4064201/

[454] M. Suchanski, P. Kaniewski, J. Romanik, E. Golan, and K. Zubel, "Radio Environment Maps for Military Cognitive Networks: Deployment of Sensors vs. Map Quality," in *2019 International Conference on Military Communications and Information Systems (ICMCIS)*, May 2019, pp. 1–6.

[455] M. S. Rahman, H. Gupta, A. Chakraborty, and S. Das, "Creating Spatio-temporal Spectrum Maps from Sparse Crowdsensed Data," in *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, Apr. 2019, pp. 1–7, iSSN: 1558-2612.

[456] Y. Wang, A. Liu, X. Xia, and K. Xu, "Learning the Structured Sparsity: 3-D Massive MIMO Channel Estimation and Adaptive Spatial Interpolation," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 11, pp. 10 663–10 678, Nov. 2019, conference Name: IEEE Transactions on Vehicular Technology.

[457] Y. Liu, W. Huangfu, H. Zhang, and K. Long, "Multi-Criteria Coverage Map Construction Based on Adaptive Triangulation-Induced Interpolation for Cellular Networks," *IEEE Access*, vol. 7, pp. 80 767–80 777, 2019, conference Name: IEEE Access.

[458] K. Sato, K. Inage, and T. Fujii, "On the Performance of Neural Network Residual Kriging in Radio Environment Mapping," *IEEE Access*, vol. 7, pp. 94 557–94 568, 2019, conference Name: IEEE Access.

[459] T. Chen, Y. Zhang, Y. Tuo, and W. Wang, "Online Discovery of Congregate Groups on Sparse Spatio-temporal Data," in *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Sep. 2018, pp. 1–7, iSSN: 2166-9589.

[460] A. Vellido, J. D. Martín-Guerrero, and P. J. Lisboa, "Making machine learning models interpretable." in *ESANN*, vol. 12. Citeseer, 2012, pp. 163–172.

[461] V. Van Belle12, S. Van Huffel12, J. Suykens12, and S. Boyd, "Interval coded scoring systems for survival analysis," 2012.

[462] A. Tosi and A. Vellido, "Cartogram representation of the batch-som magnification factor." in *ESANN*, 2012.

[463] M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," *arXiv preprint arXiv:1808.00033*, 2018.

[464] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," *arXiv preprint arXiv:1703.00810*, 2017.

[465] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[466] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Interpretable machine learning: definitions, methods, and applications," *arXiv preprint arXiv:1901.04592*, 2019.

[467] L. Zhu and N. Laptev, "Deep and confident prediction for time series at uber," in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, 2017, pp. 103–110.

[468] H. F. Tan, G. Hooker, and M. T. Wells, "Tree space prototypes: Another look at making tree ensembles interpretable," *arXiv preprint arXiv:1611.07115*, 2016.

[469] D. Didona, P. Romano, I.-i. I. Técnico, and U. Lisboa, "Hybrid machine learn-

ing/analytical models for performance prediction: A tutorial." in *ICPE*, 2015, pp. 341–344.

[470] D. Didona, F. Quaglia, P. Romano, and E. Torre, "Enhancing performance prediction robustness by combining analytical modeling and machine learning," in *Proceedings of the 6th ACM/SPEC international conference on performance engineering*. ACM, 2015, pp. 145–156.

[471] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[472] 3rd Generation Partnership, "TR36.814 v9.0.0: Further advancements for E-UTRA physical layers aspects (release 9)."