THE UNIVERSITY
of EDINBURGH

Understanding Mobile Network Quality and Infrastructure with User-Side Measurements

Mah-Rukh Fida

Doctor of Philosophy

Institute of Computing Systems Architecture

School of Informatics

The University of Edinburgh

2018

Abstract

Measurement collection is a primary step towards analyzing and optimizing performance of a telecommunication service. With an Mobile Broadband (MBB) network, the measurement process has not only to track the network's Quality of Service (QoS) features but also to asses a user's perspective about its service performance. The later requirement leads to "user-side measurements" which assist in discovery of performance issues that makes a user of a service unsatisfied and finally switch to another network.

User-side measurements also serve as first-hand survey of the problem domain. In this thesis, we exhibit the potential in the measurements collected at network edge by considering two well-known approaches namely crowdsourced and distributed testbed-based measurements. Primary focus is on exploiting crowdsourced measurements while dealing with the challenges associated with it. These challenges consist of differences in sampling densities at different parts of the region, skewed and non-uniform measurement layouts, inaccuracy in sampling locations, differences in RSS readings due to device-diversity and other non-ideal measurement sampling characteristics. In presence of heterogeneous characteristics of the user-side measurements we propose how to accurately detect mobile coverage holes, to devise sample selection process so to generate a reliable radio map with reduced sample cost, and to identify cellular infrastructure at places where the information is not public. Finally, the thesis unveils potential of a distributed measurement test-bed in retrieving performance features from domains including user's context, service content and network features, and understanding impact from these features upon the MBB service at the application layer. By taking web-browsing as a case study, it further presents an objective web-browsing Quality of Experience (QoE) model.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified. Part of the material used for the contributions made by my thesis has been published in the papers listed below.

Massimiliano Molinari, Mah-Rukh Fida, Mahesh K. Marina and Antonio Pescape. "Spatial Interpolation based Cellular Coverage Prediction with Crowdsourced Measurements." *In the Proceedings of the 2015 ACM SIGCOMM Workshop on Crowdsourcing and Crowdsharing of Big (Internet) Data (C2B(1)D '15*, August 2015.

Mah-Rukh Fida, Andra Lutu, Mahesh K. Marina and Özgü Alay. "ZipWeave: Towards Efficient and Reliable Measurement based Mobile Coverage Maps." *In the Proceedings of IEEE International Conference on Computer Communications (INFOCOM) 2017,* May 2017.

Mah-Rukh Fida, Konstantinos Kousias, Andra Lutu and Mohammad Rajiullah, Özgü Alay, Anna Brunstrom and Antonios Argyriou. "FLEX-MONROE: A Unified Platform for Experiments under Controlled and Operational LTE Settings." *In the Proceeding of 11th ACM International Workshop on Wireless Network Testbeds, Experimental Evaluation & Characterization (WiNTECH2017)*, 2017.

(*Mah-Rukh Fida*)

Table of Contents

List of Figures

List of Tables

Acronyms

AAS Adaptive Algorithm Selection

BCS Bayesian Compressive Sensing

CS Compressive Sensing

CCL Connected Component Labelling

DBSCAN Density-Based Spatial Clustering of Applications with Noise

FIRE Future Internet Research Experimentation

FLEX FIRE LTE testbed for open Experimentation

HCLUST Hierarchical Clustering

QoS Quality of Services

QoE Quality of Experience

GLL Grid Likelihood

MBB Mobile Broadband

MDT Minimization of Drive Tests

MEC Minimum Circle Enclosing

MCC Mobile Country Code

MNC Mobile Network Code

MNO Mobile Network Operator

MONROE Measuring Mobile Broadband Networks in Europe

OK Ordinary Kriging

PAS Probabilistic Algorithm Selection

PLT Page Load Time

PLP Path Loss Propagation

REM Radio Environment Map

RF Random Forest

RNC Radio Network Controller

RSS Received Signal Strength

RSSI Received Signal Strength Indicator

RTT Round Trip Time

SRSS Strongest Received Signal Strength

WC Weighted Centroid

Chapter 1

Introduction

A communication network is an inter-connection of diverse set of entities and layers, requiring its operator and service provider to ensure its seamless and efficient working by constant examination and troubleshooting of the involved components. The ubiquity of smart devices and constant increase in the demand for high quality multimedia services has raised the maintenance burden in the sense that a mobile broadband (MBB) service provider has both to stay informed about network side performance issues and about users' perception of an MBB service. This is because of two primary reasons:

- User's experience with an application service is not merely the result of lower layer features such as data rate, packet loss and round trip time delay. It is also impacted by application content such as viewing of a HD streaming video or browsing of a complex web-page along with the context where the application is running e.g. time of the day, end-device capabilities and user expectations to name a few.
- An unsatisfied user usually switches to another network without prior alerts and may impacts the decision of others with word of mouth. According to an Accenture survey [1], about 90% of users do not want to complain about a low quality service rather they simply leave the provider and go to another.

To avoid customer churn it is, therefore, essential to have a means for continuous

measurement of Quality of Experience (QoE)[1]. This can be done by keeping track of customers' experience with a network via user-side measurements. These measurements can provide in-time information about when, where and why an MBB service quality is below expectation, making it easy to troubleshoot the problem in time. In this work, we quantify the potential challenges and benefits of these measurements by studying:

1. *Reliable generation of a coverage map using user-side measurements:* Coverage status of a network provides a high level insight about users' experience with a network in general. For example a report from Nokia Insight [2] shows that due to inclusion of small cells in HetNets, network coverage improves that ultimately results in enhanced video QoE. To understand coverage status of a network, its radio map can be generated from RSS measurements reported by end-users. These measurement samples may consist of characteristics that damages accurate generation of a coverage map. Our research study, therefore, initiates with the investigation of how to reliably build a coverage map when underlying user-side measurement samples have characteristics that can lead to an inaccurate radio map.

2. *Sampling strategy for coverage map generation:* Secondly, to report network measurements, a crowd user has to consume his own resources including Internet data, battery life and computing power. Furthermore he should be willing to expose his measurement locations in face of potential privacy threats. To encourage users for crowdsourcing, incentives are necessary to provide participants with enough rewards. In face of budget constraints, it is ideal to decrease the sampling cost while maximizing information gain. The next piece of our study is related to a sample selection process such that with much less sampling cost a good estimate of a cellular network's coverage status, for different parts of a region, can be obtained.

3. *Analyzing reliability of coverage map from the perspective of an end device-type:* The measurement samples, that built up coverage map, are reported from

[1] Wikipedia: *Quality of Experience (QoE) is a measure of the delight or annoyance of a customer's experiences with a service*

diverse devices which vary in their RSS readings even when at same location and connected to same tower and network. A single combined coverage map may, therefore, not represent true picture of a network's status from the perspective of a device-type. In this study we analyse if a set of measurement sources add to or degrade accuracy of expected radio map for an end-device. Such an understanding assist in better device centric coverage map generation.

4. *Quantitative assessment of user's QoE from lower-layer and application-specific features collected at end-nodes:* To analyze impact from different factors on user's satisfaction with a network service, it is important to evaluate an end user's QoE against network-oriented, content and context-based parameters. Such an analysis helps in building QoE model which points out the prospective features, proper tuning of which steers improvement in QoE. In this research work we undertook such an analysis by taking web-browsing as a case-study.

5. *Locating cell sites with collected measurements:* Cell site information is widely used in tracking and locating end-users when GPS is absent. This information is however not released publicly e.g. operators do not disclose and camouflage their towers to protect their commercial interests [3] and avoid aesthetic complaints from local residents. Third party entities like Apple, Skyhook and Google Map thus use estimated locations of cell towers along with WiFi APs, derived from their proprietary databases containing crowdsourced measurements [4], in locating end-users. The accuracy of end-user location is directly affected by the accuracy of estimated cell tower/AP location. In this study, we work on accurate estimation of cell sites with crowdsourced measurements that come with heterogeneous characteristics.

With the basic aim to examine worth of user-side measurements in explaining performance of a network from users' point of view, this chapter starts by providing an overview of different end-device based measurement collection approaches in section 1.1. Section 1.2 then outlines the specific technical problems tackled in this thesis and the corresponding research contributions. Section (i.e. 1.3) describes the organization of the rest of the thesis report.

1.1 User-side Measurements

In following subsections, we mention different methods with which user-side measurements can be collected. These methods include drive-testing, crowdsourced measurements and distributed test-beds.

1.1.1 Drive Testing

Drive-testing measures a broad range of parameters of a mobile cellular service in a defined geographic region [5]. These include a network's coverage status, capacity, and QoE levels of various MBB services e.g. voice call quality and web-browsing speed. These measurements are usually collected by special personnels like surveyors who either war-walk or drive vehicles on roads deployed with mobile radio network air interface, measurement equipment, GPS and various selected mobile hand-sets.

Drive-testing is controlled therefore can easily be focused on a problem area. By measuring what a subscriber is expected to experience, in a particular location, a mobile network operator (MNO) can then make corrective planning for network performance improvement. It is however costly due to recruiting of skilled engineers/ surveyors and using fleet of vehicles. Moreover, it is constrained in both time and space and rarely covers in-building areas which greatly limits the validity of data as it restricts real-life scenarios.

1.1.2 Crowd-sourced Measurement

Crowdsourcing is a practice of engaging a 'crowd' or group of people for a common goal [6]. In network analysis the crowd is real end-users who conduct network measurements and report it to a central server.

In this measurement method, smart-phones operate as geo-localized sensors capable of both passive monitoring and active probing of the state of an access network. The common practice by the devices is to run a measurement application such as OpenSignal [7] and MobiPerf [8] while coordinating with a designated server. The application conducts speed tests including measurement of throughput, uplink and downlink bandwidth, RTT along with signal quality, SNR, GPS and cell-related iden-

tity information.

With no device maintenance or replacement overhead from the organization, the smart-phone user provides cheap measurements diverse both in time and space. In crowdsourced measurement approach, state of the services and applications is reported from places where they are used thus providing an end-user's perspective. It is a good way to query for QoE or verify if QoS agreements are met [9].

Crowdsourcing is not devoid of challenges Being reported by a diverse set of devices, the parameters reported in measurements is limited by the 'openness' of the end-nodes. Usually the properties such as device battery status, CPU load, type and number of applications running in parallel and indoor/outdoor distinction is not known to a central server. As crowdsourced measurements are uncontrolled, the download & installation of a measurement application, the activation and deactivation of GPS, WLAN and mobile data switches are dependent upon user's discretion. Also, user-discretion is influenced by the early battery drainage, limited data capacity caps and privacy concerns. Last but not the least, crowdsourced measurements are not guaranteed to be distributed evenly in space and time, mode (e.g 3G, 4G), and technology (i.e. LTE, HSPA, UMTS etc) and is prone to redundant measurements which result in burdening the end-nodes and server with no additional information.

1.1.3 Distributed Measurement Platforms

Distributed platforms consist of measurement devices either static or mobile connected to commercial or experimental networks. These platforms vary in type of networks and characteristics they measure, the number of deployed devices and the area that they cover e.g. MONROE [10] sparsely covers four European countries and measures performance of commercial MBB networks while RIPE Atlas [11] devices are deployed across the world and they measure global Internet performance.

A distributed test platform assists in taking controlled and reproducible measurement experiments and is an efficient method to collect service performance statistics at user-end hence giving an insight about impact of a particular network configuration on its customers. Some of the concerns with these platforms are if the access to a platform and its database is open for public, if platform is scalable and how resource

maintenance is handled during collision and simultaneous requests by multiple users.

1.2 Thesis Contributions

Focusing on the measurements that are collected at user-side, we quantify how these measurements can contribute in analyzing a network's performance.

1.2.1 Reliable Measurement based Coverage Map Generation

After the deployment of a network's infrastructure, regular coverage status investigation is carried out to troubleshoot, coverage holes, if any, in time. This coverage investigation is carried out by two means i.e. analytical models and drive-testing. The analytical models need accurate values of variables and constants which are usually not feasible due to changes in terrain. Drive-testing, on the other hand, is a reliable measurement collection method but is expensive both in time and labor and as discussed earlier it does not cover off road areas. We therefore use crowdsourced measurements method which is inline with Minimization of Drive Testing (MDT) [12] specification put forth by 3GPP for UMTS and LTE networks.

Though the crowdsourced measurement samples are 'organic' representing real user experience, they come with heterogeneous characteristics e.g. samples can have different layouts i.e. uniform or non-uniform, they can have varying sampling densities in different parts of the region and may record the sampling locations with different degrees of inaccuracies. To generate a comparatively accurate coverage map, our first contribution is to identify among a list of possible interpolation processes the approach which is robust enough to estimate a network's availability status accurately within a region. The study concludes Ordinary Kriging (OK) to be the best interpolation choice while assessing radio coverage map from crowdsourced RSS measurement samples. The research outlined in this section is reported in following publication:

Massimiliano Molinari, Mah-Rukh Fida, Mahesh K. Marina and Antonio Pescape. "Spatial Interpolation based Cellular Coverage Prediction with Crowdsourced Measurements." *In the Proceedings of the 2015 ACM SIGCOMM Workshop on Crowdsourcing and Crowdsharing of Big (Internet) Data (C2B(1)D '15*, August 2015.

1.2.2 Sampling Process for Coverage Map Generation

Building up of a reliable radio map needs large sampling size and good representation from every part of the region. Such a sampling strategy however incurs more effort in sample collection and high cost in incentivizing prospective crowd [13]. To make sampling affordable, intelligent sample selection strategy is required so that less number of samples may represent well the unobserved locations. The second contribution of my research work is therefore about devising a sampling selection mechanism, called `ZipWeave`.

`ZipWeave` is based on the argument that different sub-areas of a region does not need same sampling density e.g. open places like parks and fields have smoothness in variation of their propagation characteristics while downtown like areas where there are tall buildings, the signal strength changes are abrupt even at small lag distances. Exploiting this fact, `ZipWeave` identifies these different types of areas in the region and proposes sparse sampling for areas with little variation and vice versa. This uneven distribution of sampling substantially reduces the overall sampling cost and effort while maintaining desired coverage map accuracy. Research outlined in this section is reported in following publication:

Mah-Rukh Fida, Andra Lutu, Mahesh K. Marina and Özgü Alay. "ZipWeave: Towards Efficient and Reliable Measurement based Mobile Coverage Maps." *In the Proceedings of IEEE International Conference on Computer Communications (INFOCOM) 2017,* May 2017.

1.2.3 Device-centric Accuracy of Crowdsourced Coverage Map

Measurement samples from diverse devices lead to RSS readings that differ in quality distribution even when collected at same location from same serving cell tower and network. Coverage maps presented by mobile network operators (MNO)s on their official websites or by crowdsourced databases gives a coarse view of the coverage status within a region. They mostly specify boolean signal quality information per network mode (i.e. 2G, 3G and 4G) within geographic granularity of zip code or city. From the perspective of a device model, we show in this study that, such a combined coverage map is trust worthy only if measurement samples from same device model or

device vendor have been used while generating the map.

In the analysis study, we further illustrate the level of impact on accuracy of a fine grained coverage map of a device-type when measurement samples from different sources are used. The accuracy decreases when measurement samples from different device vendors and past time-span are used. However when these measurement sources are used in conjunction with measurement samples from same device model and time span the degradation in accuracy is negligible.

A short paper on the analysis study outlined in this section has been accepted at:

Mah-Rukh Fida and Mahesh K. Marina. "Accuracy of Coverage Map with Device Diversity." *14th International Conference on Network and Service Management (CNSM)*, 2018.

1.2.4 Localizing Base Stations with Crowdsourced Measurement Samples

Knowledge of cell tower locations assists customers in finding out availability of a network in their places of interest and provides cheap end-user localization. This information also helps in location-based services such as Google map, guiding about nearby restaurants and WiFi hotspots and informing bus timings at nearby bus stops etc. Though cell tower location information has a large number of applications that can benefit both network operators and third party entities, the network operators usually do not disclose location of their infrastructure to third party entities and general public. As an alternate solution, public crowdsourced measurement databases such as OpenCelliD [14] and OpenBmap [15] estimate the cell towers with crowdsourced measurements in hand.

In this study, we demonstrate that the heterogeneous characteristics of crowdsourced measurements including correlation of RSSI and response rate of samples to their distances from cell tower locations, outliers and noise in recorded RSSI values and locations of the samples, and layouts of the measurement samples etc. obstruct a single localization algorithm to consistently achieve high accuracy in all measurement scenarios, where a 'measurement scenario' represents the set of samples with same

serving cell tower.

The contribution in this regard is a supervised learning model called `Adaptive Algorithm Selection (AAS)`. `AAS` enables selection of an localization algorithm that it considers to locate a cell tower relatively accurately among a suite of commonly used localization approaches. The choice of an algorithm is performed on the basis of crowdsourced measurement characteristics of the samples belonging to a cell tower. `AAS` substantially raises the localization accuracy of cell towers compared to when a single localization algorithm is used or when only a 'predictive' set of measurement samples are used. Predictive samples means a subset of measurement samples that are considered to help in arriving at a cell tower location accurately. Furthermore an `AAS` model trained on a mobile country code-mobile network code (MCC-MNC) pair even works relatively better when applied on a different MCC-MNC pair thus identifying infrastructure layout of a network unknown to public.

Research work outlined in this section has been accepted :

Mah-Rukh Fida and Mahesh K. Marina "Robust Cell Tower Localization with Crowdsourced Measurements." to *The Tenth International Conference on Information and Communication Technologies and Development, (ICTDX)*, 2019.

1.2.5 Web Browsing QoE Analysis via Distributed Testbed Measurements

The final piece of work is about unveiling potential of a distributed platform in enabling retrieval of performance features from different QoE impacting domains while running an MBB service. By taking web-browsing as a case study, large scale experiments were run on the distributed platform of MONROE [16]. MONROE comprises of static and mobile measurement devices attached to commercial LTE MBB networks. The study demonstrates that such a distributed platform can enable one in modelling of an objective user's QoE metric, such as page load time (PLT) in case of web-browsing, by tracking the performance features from application content, network and user's context at end-nodes of the test-bed. It further contributes in identifying intensity of impact from these different domain specific features on user's QoE, in the web-browsing paradigm, guiding on responsibility of content-designer, network operator and user's

themselves in how they can get good quality of experience by minimizing unfavourable impact of the different features.

Further with the aim to validate impact of network features identified via study conducted on commercial networks, and to discern how LTE's and future 5G networks' PLT performance can be improved with network configuration, we ran same web-browsing experimental on the LTE experimental test-bed of FLEX-MONROE [17] . By tuning some of significant network parameters such as RFSignalPower and MCS, though we were able to validate impact of signal quality and throughput on PLT, due to single MONROE test node being deployed at a static distance from FLEX LTE eNodeB we were limited in our understanding of optimal LTE network configuration. To achieve the aim, the FLEX-MONROE project then aimed to deploy a number of MONROE nodes in FLEX testbed so to have a better insight of network configuration on user's web-browsing experience.

Initial work performed in the research study, briefly illustrated in this section, is reported in following publication:

Mah-Rukh Fida, Konstantinos Kousias, Andra Lutu and Mohammad Rajiullah, Özgü Alay, Anna Brunstrom and Antonios Argyriou. "FLEX-MONROE: A Unified Platform for Experiments under Controlled and Operational LTE Settings." *In the Proceeding of 11th ACM International Workshop on Wireless Network Testbeds, Experimental Evaluation & Characterization (WiNTECH2017)*, 2017.

The work illustrating degree of influence of performance parameters on web-browsing QoE metrics has been submitted to a conference.

1.3 Thesis Organization

This thesis is organized of following chapters:

Chapter 3 provides detailed description of the work undertaken in coverage analysis using crowdsourced measurements. It demonstrates the robustness of Ordinary Kriging in comparison to other interpolation processes with measurements having various characteristics.

Chapter 3.6 mentions the working of `ZipWeave` which proposes a measurement

collection strategy so that to reduce the sampling cost required in generating a reliable coverage map.

Chapter 4 states analysis carried out in identifying the impact of different measurement sources on coverage map accuracy from the perspective of an end-user's device type. It also illustrates the scenarios where correlation between coarse grained combined coverage map and that of an end-device becomes weak and strong .

Chapter 5 presents working and performance of AAS model in raising accuracy of localized infrastructure. It illustrates how heterogeneity in crowdsourced measurements characteristics hinders a single fixed localization algorithm to behave consistently well and how these characteristics can be exploited in estimation of a suitable localization algorithm.

Chapter 6 explores the impact of context, QoS and application specific features on QoE by taking web-browsing service as a use-case. It provides an insight on how a distributed test bed based measurements, can unveil relationship between features from different domains and the corresponding experience of a user with an MBB service.

Finally, chapter 7 concludes the thesis by summarizing the work presented and discussing the limitations of the contributions as well as possible directions for future research.

Chapter 2

Reliable Measurement based Coverage Map Generation

Coverage map of a network has two basic applications. It is used by customers as a guide in adopting a network carrier. For an operator it is generated to detect coverage holes. The coverage holes thus diagnosed are dealt with either by changing antenna tilt, its height, power level or deployment of new base stations. A cheap method to generate a radio coverage map is to draw signal strength measurement samples; apply an interpolation process on the samples, and generate a coverage availability map across the region of interest.

To draw measurements, operators in the Next Generation Mobile Networks (NGMN) alliance proposed a standardized solution in 2011 called Minimization of Drive Tests (MDT) (i.e. put forth in 3GPP Rel-10 specification for LTE and UMTS networks). Key features of MDT are that a mobile device reports its location information along with performance measurements using to Trace Collection Entity (TCE) via Radio Network Controller (RNC) [18]. Since MDT is done on user end therefore measurements collected this way is a form of crowdsourcing.

In this chapter we exploit the crowdsourced measurements collected with measurement apps, that are similar to MDT in reporting coverage status information from end-devices, for generating a reliable coverage map of the surveyed region. The chapter consists of five sections. Different interpolation techniques and related work about

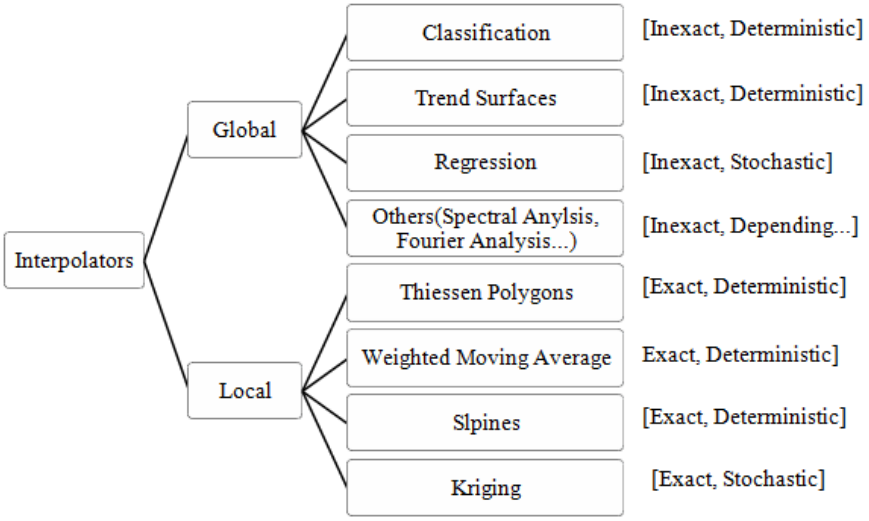


Figure 2.1: A taxonomy of spatial interpolation techniques based on [19]

.

a coverage map generation are mentioned in section 2.1 and 2.2 respectively. Section 2.3 illustrates the motivation of the study. The results from comparative study of different interpolation techniques in generating a radio map are described in section 2.4. Finally we summarize the work undertaken in this chapter in section 2.5.

2.1 Spatial Interpolation Techniques

An interpolation method is a process of using points with known values or sample points to estimate values at a set of unknown points within the boundary of the sampled region. It can be used to predict unknown values for a variable that has meaningful values at every point within a region such as elevation, rainfall, chemical concentrations and noise levels. In this study our variable of interest is signal strength of a network carrier.

Pringle [19] presents a taxonomy of spatial interpolation techniques which are summarized in Figure 2.1. The interpolation techniques can be divided into global and local techniques, exact and inexact processes, and deterministic and stochastic methods.

Global and local interpolators: Global interpolation techniques apply a single function to all the points in a study area. They calculate predictions using the en-

tire dataset. In such cases, a change in one observed value would affect the entire interpolation process. These interpolators tend to produce a surface with few abrupt changes, because, in general, they employ the principle of averaging. This reduces the influence of extreme values. Such methods are most appropriate when the surface being modelled is known to have an overall trend [20]. On the other hand, local techniques calculate predictions from the measured points within neighborhoods, which are smaller spatial areas within the larger study area. They apply the same function repeatedly to a small portion of the total set of sample points for which data have been observed. Then a surface is constructed by linking these regional observations together. It is useful when there is little or no knowledge about the overall trend in the surface.

Exact and inexact interpolators: At sampled locations, exact interpolators yield values identical to the observed values whereas inexact interpolators predict values that are different from the measured values.

Determistic and stochastic interpolators: Deterministic interpolation techniques create surfaces from measured points, based on either the extent of similarity or the degree of smoothing. It is appropriate to be used when a data set has sufficient knowledge about the phenomena to allow its behavior to be described by a mathematical function. Unfortunately, few geographical phenomena are understood in sufficient detail or obey such precise rules to permit a deterministic approach to interpolation. There is a great deal of uncertainty about what happens at unsampled locations. For such problem stochastic models are used. Stochastic models incorporate the concept of randomness suggesting that the interpolated surface is only one of an infinite number that might have been produced from the known data points [20]. Unlike deterministic, stochastic methods therefor provide uncertainty of the estimated values [21].

Below are some of the prominent interpolation techniques belonging to different afore mentioned categories.

- **Classification.** In Classification (CLSF) [19], the key idea is to infer the values of one variable attribute based upon the knowledge of the values of another attribute. The basic assumption is that the value of the variable of interest is strongly influenced by another variable that can be used to classify the study

area into zones.

- **Trend Surfaces.** It is a statistical method that finds the surface that fits the sample points using a least square regression fit. Unlike CLSF, it does not use external variable rather it uses the spatial variable of interest itself along with the spatial locations. It fits a single polynomial equation to the entire surface such that the surface variance in relation to the input values is minimized. It is an inexact interpolator so the resulting surface rarely passes through the input points, however, it is good for identifying coarse scale patterns in the data [22].

- **Regression Model (LOESS Surfaces).** LOcally wEighted Scatter plot Smooth (LOESS) performs two steps for each data point: (1) computes the regression weights for each data point in the so-called *span*, where the span controls the size of the neighborhood; (2) a weighted linear least squares regression is then performed with a second order polynomial.

- **Thiessen Polygons (THI).** These polygons are constructed by drawing lines between neighboring sample points. Next a set of lines are then constructed to bisect the first set of lines at right-angles at their mid-points. This second set of lines make the boundary of Thiessen polygons. All points lying inside a polygon are estimated to be similar to the sample point lying in its boundary. This attribute results in sharp jumps in values from one polygon to another.

- **Weighted Moving Average.** This technique, exemplified by Inverse Distance Weighting (IDW), assumes that each input point has a local influence that diminishes with distance. It weights the points closer to the processing point greater than those further away. A specified number of points or all points within a specified radius can be used to determine the output value of each location. The scheme is good for highly variable data. If samples are dense, it can estimate well the extreme changes in terrain [22].

- **Splines.** Splines consist of polynomials of degree p being local rather than global. Each polynomial describes a piece of the surface that exactly fits to a small number of data points. When all the pieces fit together, they join smoothly.

The places where the pieces join are called knots. For a degree p of 1, 2 and 3 a spline is called linear, quadratic and cubic spline respectively [23].

- **Kriging** Kriging is a class of local interpolation techniques that are quite commonly used to address spatial prediction problems in the context of mining, hydrogeology, natural resources and environmental science. The basic idea of Kriging is to estimate data at a point based on regression of observed surrounding values of that point weighted according to the spatial correlation of the field under study [24]. It assumes that the distance between sample points reflects a spatial correlation that can be used to explain variation in the surface. Here I illustrate the working of a well-known form of Kriging, i.e. *Ordinary Kriging (OK)*.

  Formula: Let z_i be the measured data value at a space location i, u represents a space location at which the data value z_u is unknown, and $V = \{1, \ldots, n\}$ be the set of points in the neighborhood of point u such that value z_i is known for each point i in V. Ordinary Kriging estimates value z_u as weighted-linear combinations of the known values at V:

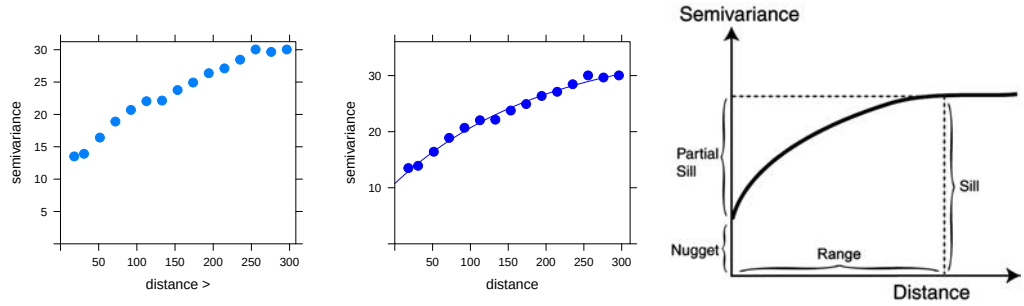
$$z_u = \sum_{i=1 \in V} w_i z_i \tag{2.1}$$

  In equation 2.1, w indicates weights that 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. The weights in Kriging formula are calculated as follows:

$$\begin{pmatrix} w_1 \\ \vdots \\ w_n \\ \lambda \end{pmatrix} = \begin{pmatrix} \beta(h_{1,1}) & \ldots & \beta(h_{1,n}) & 1 \\ \vdots & \ddots & \vdots & 1 \\ \beta(h_{n,1}) & \ldots & \beta(h_{n,n}) & 1 \\ 1 & \ldots & 1 & 0 \end{pmatrix}^{-1} \begin{pmatrix} \beta(h_{1,u}) \\ \vdots \\ \beta(h_{n,u}) \\ 1 \end{pmatrix} \tag{2.2}$$

  $\beta(h_{i,j})$ is a semivariogram's function of distance between points i and j, and λ is the Lagrange multiplier to minimize the Kriging error.

  Semi-variogram: A critical component of generating any Kriging model is creating the semivariogram, which is a plot that shows the variance in measured values between all pairs of sampled locations at different lag distances. The



(a) (b) (c)

Figure 2.2: (a) Experimental semi-variogram (b) Theoretical semi-variogram (c) Elements in semi-variogram

Fig. 2.2 (a) displays an experimental sem-variogram built from sampled points with x-axis showing distance bin between set of sampled points and y-axis showing variance between values at these sampled locations. Fig. 2.2 (b) illustrates a model fitted to the experimental semi-variogram. In the semi-variogram, points near to each other are expected to be more similar than points that are farther apart. The range is where autocorrelation exists among points based on distance. The nugget is the error or random effect. The sill is the distance at which points are no longer spatially autocorrelated. As a function sem-variogram describes the difference between field values at two locations that are h distance apart [25] as:

$$\beta(h) = \frac{1}{2(N(h))} \sum_{i,j|h_{i,j}=h} (z_i - z_j)^2 \tag{2.3}$$

N(h) indicates the number of sample point pairs that are h distance apart. As distance h between points i and j increases, the correlation between those points is expected to decrease.

In the Fig. 2.2 (b), we have used exponential model, with which the β(h) works as:

$$\begin{aligned} \beta(h) &= c0 + c1(1 - exp^{-h/a}), h > 0 \\ \beta(0) &= 0 \end{aligned} \tag{2.4}$$

Where *c0* represent nugget variance, *c0* + *c1* is sill, and a represents range of the

fitted model.

Kriging is an advanced geostatistical procedure that generates an estimated surface from spatially scattered sample points. Unlike other interpolation methods, it effectively involves an interactive investigation of the spatial behaviour of the phenomenon represented by the samples. In addition to estimated values, it also provides the un-certainty in its predictions. As a network's signal quality is a spatial variable, Kriging seems to be a suitable candidate interpolator.

Along with OK there is another well-known Kriging method called *Universal Kriging (UK)*. It is an extension of OK that incorporates local trend within the neighborhood search window as a smoothly varying function of coordinates. A search window of a location is the geographical boundary, the sampled locations within which have, correlation with the variable value at the location in question. UK estimates trend components within each search neighborhood window and then performs Simple Kriging [23] on the corresponding residuals.

2.2 Related Work

Before evaluating performance of the discussed intepolators under different crowd-sourced measurement scenarios, it is important to see how different studies have found role of different interpolation schemes in radio coverage map generation.

To built a radio map, Z. Dali et al. [26] used support vector regression (SVR) algorithm so to improve efficiency and reduce the labor cost of establishing the radio-map during offline phase. In this method, a nonlinear mapping between RSS and physical locations is constructed by SVR. To reduce finger prints collection and to have a radio-map indoors with reduced effort, J. Racko et al. [27] proposed use of linear interpolation based on weighted average of known signal strength samples and Thiessen polygons.

Some other studies have preferred different versions of Kriging for coverage prediction in wireless networks. In [28], Konak estimates path loss in wireless LANs using OK by defining the distance between two points as their Euclidean distance plus a term that represents the set of obstacles between the points. Also Phillips et al. [29]

used OK on a 2.5 GHz WiMax network setting, finding it to produce radio environment maps that are more accurate and informative than both the explicitly tuned path loss models and the basic fitting approaches.

In [30] and [31] Kolyaie et al. used drive testing to collect signal strength measurements, and compared the performance of empirical models and spatial interpolation techniques. Specifically, they use the Okumura-Hata empirical model, a common model used for cellular system planning and management. They evaluated its accuracy in comparison with IDW and two Kriging variants: OK and Universal Kriging (UK). Though Okumura-Hata empirical model was seen to yield better results than IDW, OK and UK provided best prediction results.

In [32], Sayrac et al. proposed Bayesian spatial interpolation for coverage analysis in cellular networks, specifically focusing on coverage hole prediction. Kitanidis' Bayesian Kriging (BK) interpolation method is used which automatically calculates the interpolation model parameters through a process of sub-settings and simulations. The main disadvantage of this method is its high computational complexity. In [33], Braham et al. consider a variant of Kriging called Fixed Rank Kriging (FRK), which is aimed at reducing the Kriging complexity. In fact, the computational complexity of Kriging is $O(n^3)$, where n is the number of measurements. The authors in [33] argue that FRK can reduce this computational complexity while keeping an acceptable prediction error.

2.3 Motivation

The crowdsourced measurements represent real user experience and are reported over time from parts of the region where crowd users roam. As these measurements are uncontrolled, they come with a variety of characteristics. The prominent features include layout, density and location inaccuracy of the samples.

The above enumerated features effect the accuracy of the resulting coverage map. To minimize the adverse impact of the unfavourable characteristics, goal of this study is to identify an interpolation technique that produce relatively reliable coverage map in different crowdsourced measurement scenarios.

2.4 Robust Interpolation Scheme

For the purpose of this study, we used RF Signal Tracker [34], a measurement Android application, to obtain information such as GPS/network location, cellular technology, location area code, cell ID, signal strength in Arbitray Signal Unit (ASU)[1] and exact measurement location added manually by the user.

After collecting crowdsourced measurements with various characteristics we evaluated accuracy of resulting coverage maps generated with different spatial interpolation techniques. These techniques included variants of Kriging (i.e. OK, UK, BK and FRK) and rest of the techniques outlined in section 2.1.

To quantify accuracy of different interpolation schemes, the metric used is mean absolute prediction error (MAPE) in ASU. MAPE provides the mean absolute difference between actual and estimated ASU values at the test locations. The evaluation is done via calibration-and-validation (C&V) scheme in which two-third of measurements are used to make a prediction in the remaining one-third.

2.4.1 Location Inaccuracy of Measurements

Crowdsourcing based measurements exploit network data obtained from a commodity smartphone and its built-in mechanisms to identify the device location. The duration of device location updates may not be in-line with its measurement reporting frequency. This imperfection may result in highly inaccurate locations in some cases, which in turn leads to misprediction problems. This subsection looks into this problem, focusing on the outdoor environment and relying on GPS (of smartphones) for storing measurement positions.

To assess the effect of GPS inaccuracies on the coverage status accuracy with different spatial interpolation techniques, we took a total of 75 measurements of signal strength values within a small area in the city of Edinburgh; for each measurement, we stored both the actual and reported location by the phone GPS. Actual locations were known because they were pre-emptively decided on a map, after visiting these locations their GPS detected coordinates were stored along side the ASU readings.

[1] ASU stands for arbitrary strength unit. Signal strength in ASU is on an integer scale and is linearly related to signal strength in dBm.

The location difference between both sets of measurements were, on average, of 18 meters. We evaluated the accuracy of the prediction with GPS based locations and actual locations. Fig. 2.3 displays the prediction errors in terms of MAPE with different interpolation techniques for *Actual* as well as *GPS* based locations. We see that some techniques are more adversely affected than others. Splines-Bicubic, Thiessen Polygons and LOESS regressions are negatively impacted. For LOESS, we only show the result with the span value[2] that yields the best prediction — results with other span values are similar to that with Bicubic. This is no surprise, since they strongly rely on the notion of neighboring points to make a prediction. As for Kriging-based techniques, OK and UK yield better predictions even in presence of location inaccuracies as they are more dependent upon the correlation between samples' ASU at different distance bins rather than by exact distance lags. Also these two are less sensitive to location inaccuracies than the BK and FRK kriging methods. As per other techniques, IDW and Trend surfaces are also only slightly affected, this is as expected since they mainly rely on regional trends.

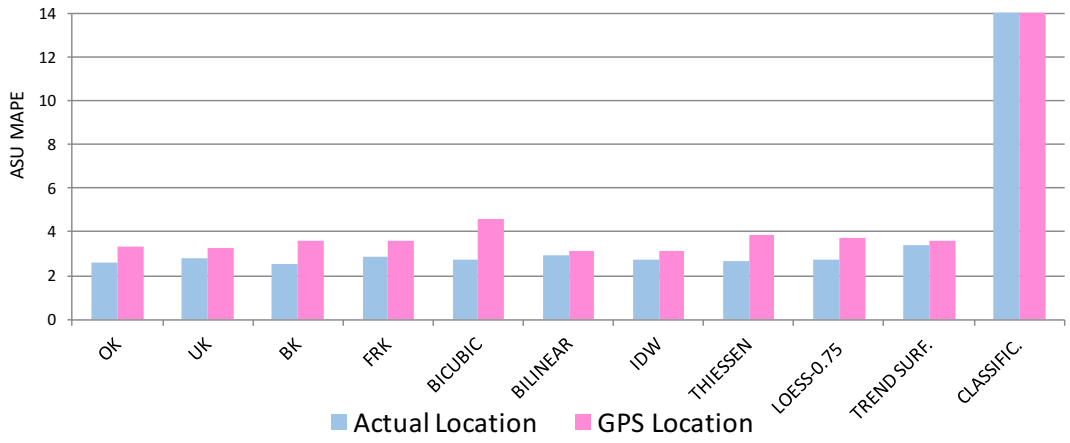


Figure 2.3: Impact of location inaccuracy on MAPE performance of different spatial interpolation techniques.

[2] A higher span smooths out the fit more, while a lower span captures more trends but introduces statistical noise if there is too little data.

2.4.2 Spatial Distribution of Measurements

With crowdsourcing based measurement, the spatial distribution of participant devices may not be uniform in the region of interest. In this subsection, we consider such scenarios with measurements distributed non-uniformly in space. To serve as a reference, we first consider a more ideal scenario with uniformly distributed measurements in space. It is to investigate how well each spatial interpolation technique can estimate the data in the gaps which are unmeasured.

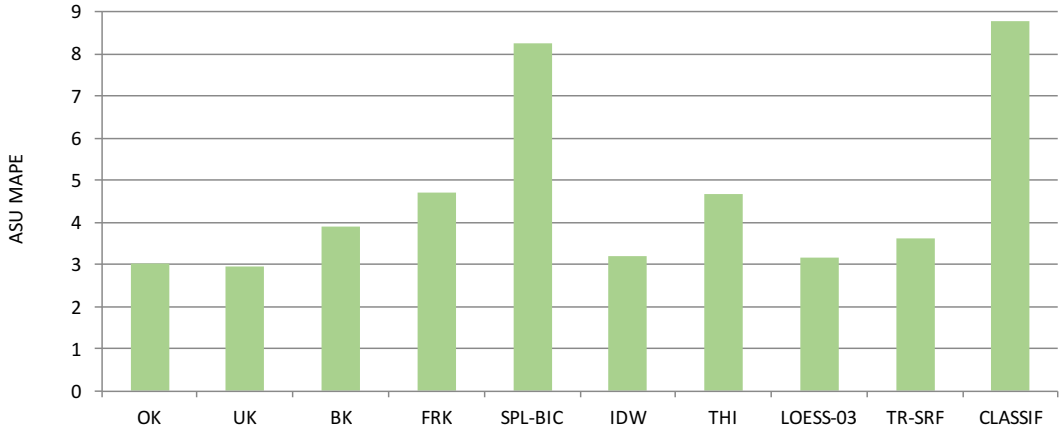


Figure 2.4: MAPE with different interpolation schemes for uniform spatial distribution of measurements.

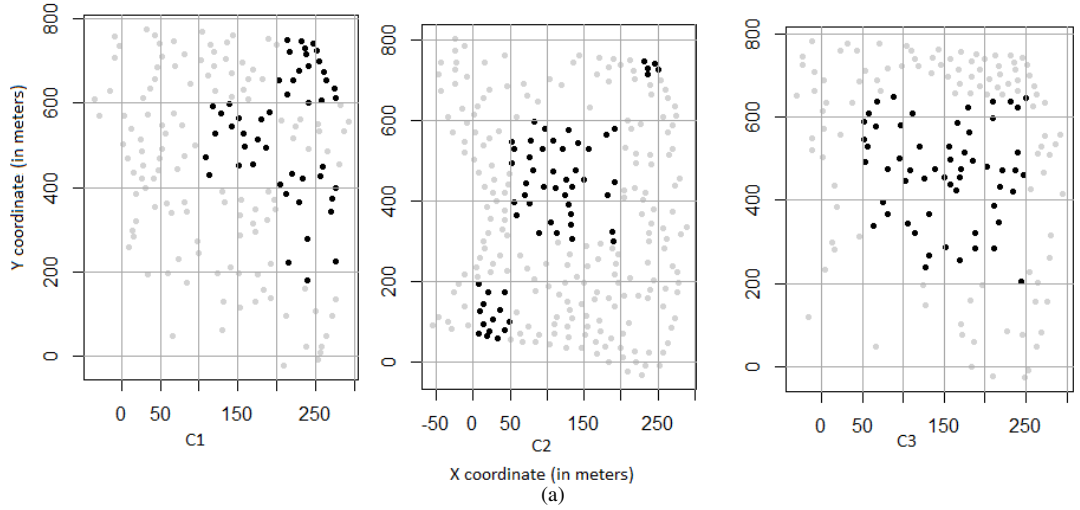
Uniformly Distributed Measurements: To assess performance of spatial interpolation techniques when measurements are uniformly distributed, we use C&V approach as before but in such a way that both calibration and validation points are chosen randomly and *uniformly* throughout the interest area. Specifically, we consider an urban outdoor open space in a park and use 479 measurements in total of which 420 measurements were used for calibration and rest for validation.This selection of calibration and validation sampling locations is done uniformly and randomly so that both the sets well represent all parts of the area. The mean absolute prediction errors , when uniformly spread calibration sampling locations are used, are shown in Fig. 2.4. The differences among several of the schemes are somewhat negligible in this scenario. Exceptions to this conclusion are Splines and Classification, which perform very poorly. Like before, a few of the schemes like BK, FRK and Thiessen Polygons

yield predictions with higher errors but not as high as Splines and Classification.

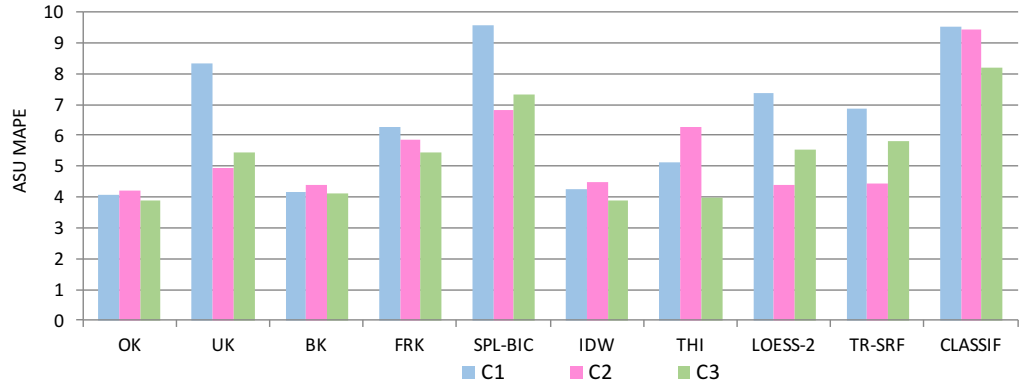
Non-Uniformly Distributed Measurements: This subsection presents the more realistic case of non-uniformly distributed measurements in space. We use measurements from the same park environment but taken in a spatially non-uniform manner, specifically to reflect *clustered measurements* and *measurements with holes*.

- *Clusters:* Clustered scenarios are shown in Fig. 2.5 (a), with black dots showing positions of calibration data and gray dots indicating validation points; the number of calibration points is around 50 in all three scenarios and there are at least 100 validation points in each of the scenarios.

  We find that prediction errors are widely different for different span values with LOESS. Though not shown in Fig. 2.5 (b), LOESS with span values less than 0.75 yield quite erroneous results. As per the other schemes, we see that OK and IDW in particular consistently give lower prediction errors across all scenarios, whereas other schemes like Splines-Bicubic and Classification give worse results as before.



(a)



(b)

Figure 2.5: (a) Three different scenarios with measurement samples clustered together at different parts of the region (b) ASU MAPE in each of the clustered measurement scenarios.

- *Holes:* Scenarios with measurement holes are shown in Fig. 2.6 (a) with black dots indicating calibration points and rest are validation points where predicted values are compared with actual values to compute the errors and MAPE. As with clustered scenarios, here we consider scenarios with different forms of holes: a smaller corner hole, a large middle hole and a bigger side hole surrounded from three sides. Prediction errors in each of these scenarios is shown in Fig. 2.6 (b). As with clustered scenarios, prediction errors vary with different types of holes. In general we see that errors in both the clustered and hole-based

sampling approaches are higher than the uniform measurements sampling case.

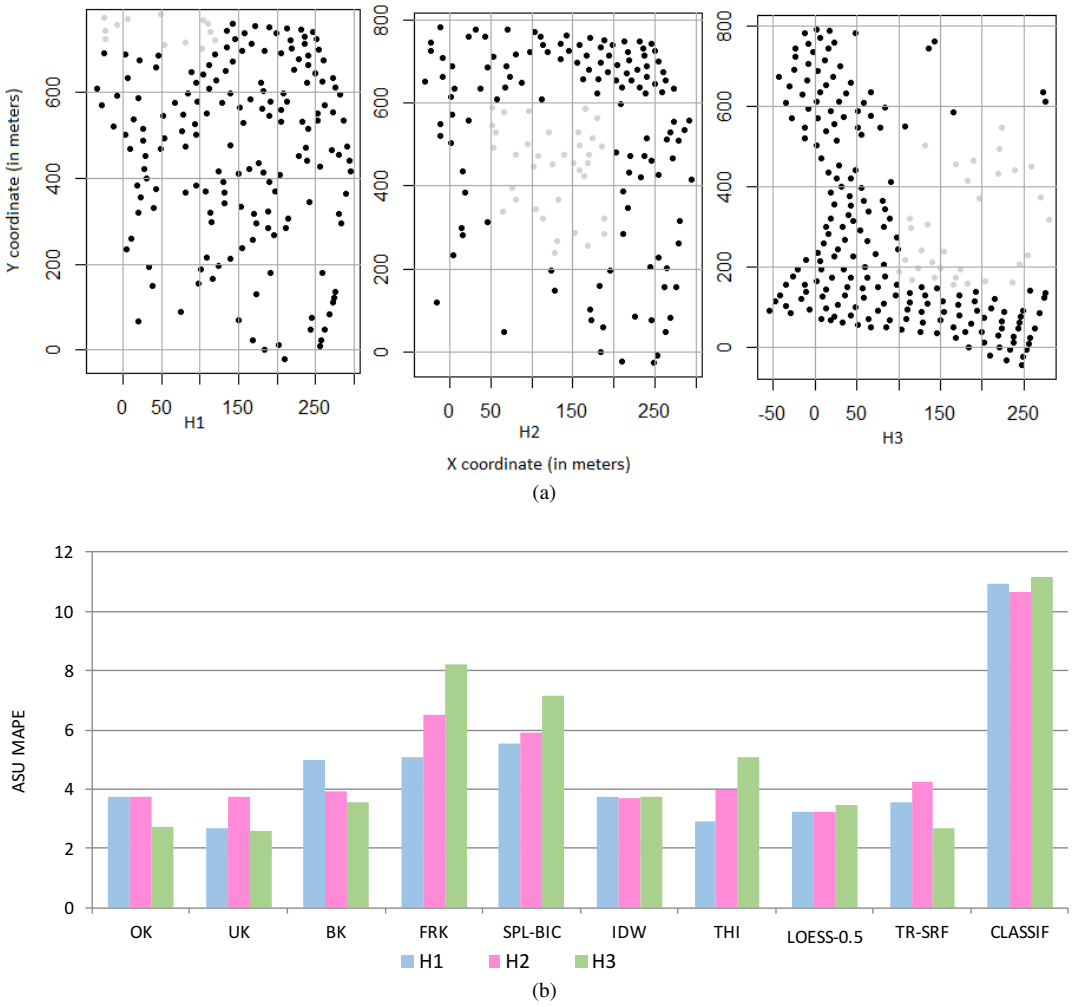


Figure 2.6: (a) Scenarios with measurement holes; (b) Prediction errors in terms of MAPE in each of the measurement hole scenarios.

2.4.3 Density of Measurements

In this subsection we assess the impact of measurement density on the prediction. For analysis we collected about 500 measurements in a park in Edinburgh and then selected 479 measurements out of these so as to have similar density throughout the area. We used 59/479 measurements as validation dataset and the remaining 420/479 as initial calibration dataset. The size of the calibration dataset is then gradually decreased

in steps to obtain different density values. We consider 9 steps. In the initial one (step 1), we have about 14 measurements per squared hectare. In the last (step 9), we have only one.

We show the prediction error results with varying density in two graphs. Fig. 2.7 focuses on Kriging techniques, IDW and Thiessen Polygons and Trend Surfaces, whereas results with LOESS, Splines and Classification are shown in Fig. 2.8. We observe that for most of the schemes, prediction error increases as measurement density decreases as one would expect. Some of the poorly performing schemes from earlier sections like Classification, Splines, FRK and BK still yield poor results largely regardless of measurement density. Prediction error gets really worse for low measurement densities with some of the schemes (e.g., LOESS, FRK). Though not shown in Fig. 2.8 to avoid clutter, higher span values with LOESS (e.g., span value of 2) are more robust at low measurement densities but come with somewhat higher prediction errors at higher measurement densities; the opposite holds for lower span values – we show the result with lowest span value of 0.3. Overall, we find OK and IDW to be most robust schemes across all measurement densities.

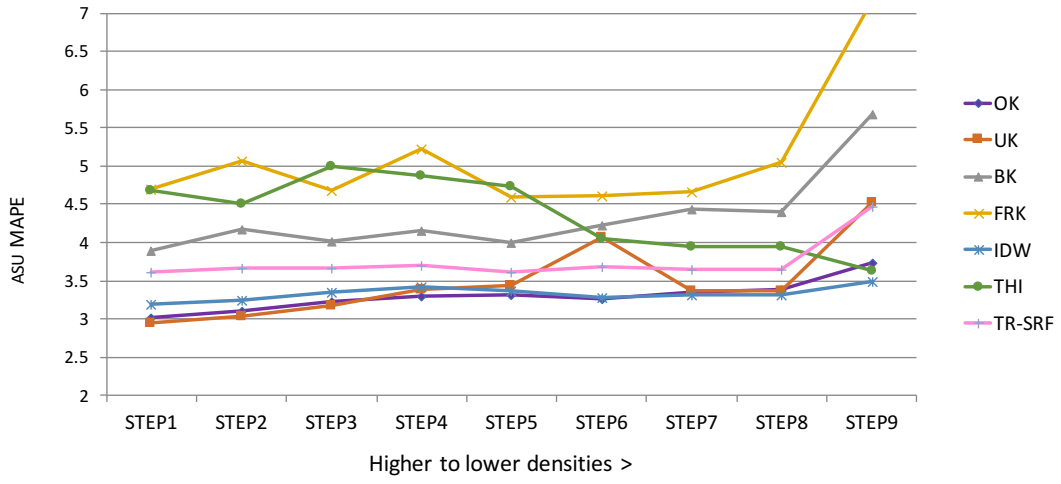


Figure 2.7: Impact of measurement density on MAPE values for Kriging techniques, IDW, THI and TR-SRF.

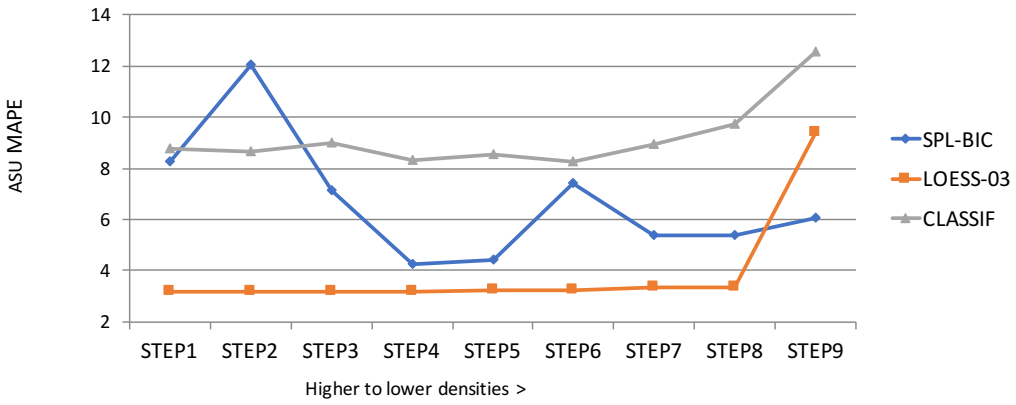


Figure 2.8: Impact of measurement density on MAPE values for LOESS Surfaces, Classification and Bicubic Splines.

2.4.4 Ordinary Kriging vs Compressive Sensing

For indoor location based services, lack of GPS is overtaken by WiFi fingerprint database. This database is generated by either a surveyor or indoor crowd. To reduce cost associated with collecting these fingerprints and to recover absent data without introducing errors Yuexing Zhang et al. [35] proposed use of Compressive Sensing (CS). CS is known to recover a signal, image, spatial and spatial-temporal data with much less measurement samples than required by the traditional methods. The basic idea is that if a signal is sparse itself or is sparse in any other domain (e.g. Fourier/Wavelet Transform etc.), it can be recovered almost accurately with a small set of samples.

Our study regarding radio map generation suggests OK as best perfoming interpolation process under various crowdsourced measurement scenarios. For the sake of completeness, it is interesting to see the behaviour of CS in recovering coverage status at unobserved locations, in comparison to OK. To inquire this, we use two well-known methods of CS namely Matrix Completion (MC) [36] and Bayesian Compressive Sensing (BCS) [37].

Matrix Completion is the task of filling the missing entries of a partially observed matrix. It often seeks to find the lowest rank matrix that matches the known entries. A low rank matrix means that there must be some similarity in rows and columns of

the matrix. Datasets related to different types of world problems can be arranged in the form of matrix and its missing information can be recovered with MC e.g. the Netflix movie ratings where ratings are available for a set of movies from different users. Matthew Roughan et al. [38], used MC for road traffic estimation with matrix columns representing road junctions and its rows representing traffic at each of these junctions at different time periods. Bo Yang et al. [39] on the other hand, used BCS for updating radio coverage map at time *t* with a small set of measurements by using the correlation between measurement points of the radio map generated at time *t-1*.

A requirement for MC is to have sporadic spatio-temporal data for hotspot locations of a region; where in radio map, need is to have spatio-temporal data for the whole region which is usually not available. To make things easier instead, we take a small rectangular area, divide it into uniform grids and take samples at some of the grid locations so that to predict signal quality at unsampled grid locations with rectangle representing matrix having spatial correlation between its rows and columns. For BCS [39] requirement is to have at least a full map of the region from the past. To meet the requirements of both the MC and BCS, we use the radio maps generated from OpenSignal HSDPA measurements of central London of operator EE for year 2012 and 2013. These maps shown in Fig. 4.2 represent coverage status in matrix of 50 x 50 dimension with signal quality represented in ASU ranging from 1-19 with 19 indicating highest signal strength.

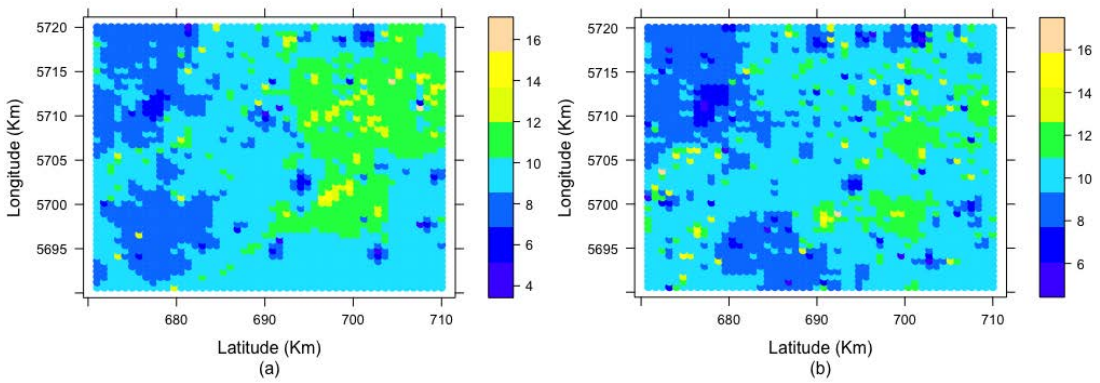


Figure 2.9: Radio maps of central London for HSDPA, EE in year (a) 2012 and (b) 2013 with signal quality represented in range 1-19 ASU.

We take training samples of 20, 40 and 60% respectively from map of year 2013, to see how MC and BCS performs in comparison with OK in recovering rest of the map. In Fig. 2.10 we show errors up to 4 MAPE for visual understanding. MC gives quite poor result for 20% of training set with more than 40 percent errors lying above 3 MAPE. It however rapidly improves with growing training size but still is the wort scheme. BCS presents 90% of errors below 2 MAPE both for the lowest and the highest training sizes, similarly OK too seems to be unaffected by raising sampling density with achieving less than 1 MAPE for 90% of the time in all three cases.

With these results we again conclude OK as the most suitable recovery process for a radio coverage map. Additionally it is neither bounded by matrix type of structure like MC nor it needs any prior signal-strength based correlation information of a region, like BCS.

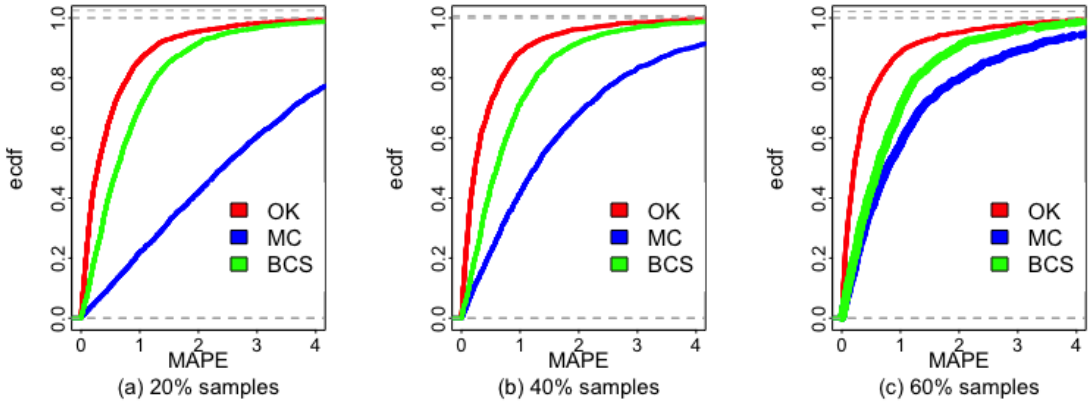


Figure 2.10: ECDFs of mean absolute prediction errors with 20, 40 and 60% of training data with each of the coverage map recovery processes of OK, MC and BCS.

2.5 Summary

In this chapter, we illustrated the mechanisms that enable generation of a reliable coverage map under different crowdsourced measurement characteristics.

Using measurements taken from commodity smartphones, we evaluated the accuracy of a number of spatial interpolation techniques along with various forms of Kriging in several scenarios capturing unique characteristics of crowdsourced measurements. The results show that basic form of Kriging called Ordinary Kriging followed by IDW, generally perform well even when measurements are non-uniformly

distributed, have inaccurate locations, and when their density is very low.

Chapter 3

Sampling Process for Coverage Map Generation

The generation of a measurement-based radio coverage map needs two basic components i.e. an interpolation process and a set of measurement samples having the basic information of signal strength of a network carrier tagged with its location information. In chapter , we performed an analysis study on identifying an interpolator that is robust enough to generate a relatively reliable coverage map under different crowdsourced measurement scenarios. This chapter is about devising a sampling strategy for crowdsourced measurement collection so that the desired accuracy of a coverage map is achieved with minimal sampling cost and effort.

We propose a sampling strategy called `ZipWeave`, that reduces the overall sampling cost by identifying the sub-areas within a region according to their sampling size requirement. To understand the whole process, this chapter illustrates six sections. Section 3.1 and section 3.2 presents the motivation behind the study and the related work respectively. After describing the datasets, used for evaluation, in section 3.3, the working of `ZipWeave` is explained in section 3.4. In section 3.5, we demonstrate that cost can be further reduced by using in combination the samples generated by diverse sources. Lastly we end up the chapter by summarizing the lessons learnt in section 3.6.

3.1 Motivation

For crowdsourced measurements, the size and locations of samples are dependent upon the crowd users. Samples are generated from locations where crowd users roam which may result in redundant information. The useless redundancy burdens end-devices, communication links and central server. On the other hand there are areas that are under-represented, most probably with users who are not willing to run measurement applications due to limited battery life, data constrains and privacy concerns. There are different recommendations for encouraging end-users to be part of the crowd [40] e.g. monetary incentives, gamification, altruism and social motivations, most effective of which is monetary incentives which means the higher the budget the greater area will be covered.

For a coverage map, the requirement is to maximize the reliability of resulting map while maintaining sampling cost under the budget. On one hand, it is to avoid unnecessary burden on crowd users brought by redundancy of samples and on the other hand to cover under-represented locations by incentivizing un-willing users run measurement applications . The aim of this study is, therefore, to devise a sample selection strategy such that with a reduced sampling cost a good representative coverage map can be generated.

3.2 Background and Related Work

3.2.1 Spatial Sampling Schemes

Generally, spatial sampling design follows a two phase process. The *first phase sampling* aims to provide a primary sense of the region of interest. In our context, this means identifying parts of the region where variation in signal strength in closer locations is comparatively higher. The *second phase sampling* then aims to complement the first phase with additional samples from the sub-areas where high variation in spatial phenomena is observed. Some of the commonly used sampling patterns for first phase, shown in Fig. 3.1 are as follows:

- **Random sampling:** It draws measurements with random selection of sampling locations, each location has equal probability of selection and is independent of

other sampled locations. The spread is uneven and may have large in-between gaps [41].

- **Systematic (or regular) sampling:** It is an unbiased scheme where once the first sample point is drawn, the other samples are drawn in a given preset order relative to the first point. It has the advantage over the random sampling as the sampling points are spread more evenly over the sampled area however it is inefficient to irregular features of the space such as stratification, inhomogeneous variances and anisotropy [42].

- **Stratified sampling:** A somewhat better scheme is stratified sampling that assumes that the population is spatially stratified and that the sub-population within each stratum is i.i.d. Once stratas are identified, random or systematic sampling is applied to each stratum.

- **Clustered sampling:** After certain distance it collects a set of samples at geographically close locations. It is desirable when need is to retrieve correlation in nearby locations with non-stationarity across the whole region.

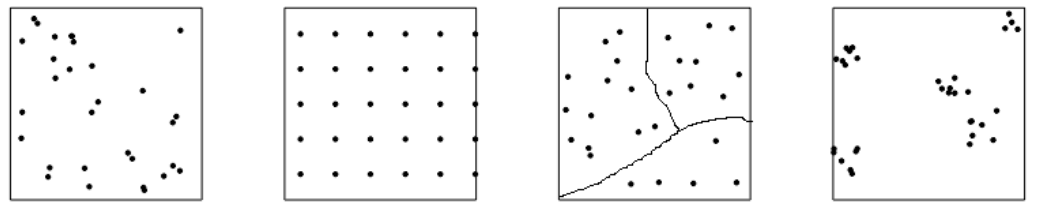


Figure 3.1: Sampling patterns for the first phase sampling

.

3.2.2 Sampling Strategy for Coverage Analysis

To generate a reliable coverage map, for *the first phase sampling*, R. Olea [43] orders different sampling patterns according to their degree of efficiency i.e. regular (or systematic) followed by stratified then random and at end clustered scheme. Delmelle

et al. [44] employs a combination of systematic and nested clustered-sampling scheme so that to guarantee a coverage of entire study region and also to have a better estimate for locations having variations at small lag distances. Philips et al. [29] also uses the same combination for generating a Carrier to Interference and Noise Ratio (CINR) map for 2.5 GHz WiMax network. Like the systematic-clustered scheme, the systematic-random approach too helps in extracting both the heterogeneity across a region by space filling and correlation between neighbors with different lag distances.

The *second phase sampling*, which aims to boost the accuracy of the coverage map within the region of interest, could take one of two different approaches for collecting additional samples: *Sequential*, where new samples are collected one at a time, for example S. Grimoud et al. [45] proposes an iterative Radio Environment Map (REM) building process based on Kriging. Each time a REM update is needed, the algorithm chooses the candidate location which minimizes the expected value of the conditional mean square error. This approach however, is susceptible to local optima. The second method is *Simultaneous*, where the whole set of additional samples are collected at the same time so that to optimize an objective function such as minimizing Mean Kriging Variance [46].

3.3 Datasets

To analyse the working of `ZipWeave`, we use two different types of user-side measurements namely crowdsourced and controlled datasets.

- **Controlled Datasets:** These are two different datasets; the first one is collected by a NorNet Edge [47] deployed in an NSB train in Oslo, Norway. A NorNet node is made up of a relatively inexpensive single-board computer that is connected to all major operators of mobile broadband in Norway and reports signal quality, cell ID, network technology and GPS coordinates with a 10 seconds granularity. This data is collected in August 2015.

  The second dataset is collected by the Lothian buses at Edinburgh. This measurement set is not the primary purpose of the bus company rather it is an outcome of WiFi provision facility to the bus passengers. When a mobile data signal

(3G/4G etc) is received by the onboard router, the router not only broadcasts a Wi-Fi signal within the bus to which passengers can connect to but also records the received signal strength along with date, time, network operator, technology and GPS position. This is a dense dataset comprising a single day's measurements from 85 buses.

Both of these datasets mimic drive-testing measurement approach due to reliability brought by using single measurement device and limitedness of their measurements to drive-ways and roads.

- **Crowdsourced Datasets:** Crowdsourced dataset is collected by OpenSignal measurement application, run by user crowd, from 1st August 2015 to 30th October 2015 for both the cities of Oslo, Norway and Edinburgh, UK. These datasets too consist of signal quality, network operator, network technology, serving cell ID and GPS coordinates of the measurement samples.

3.4 ZipWeave Sampling Strategy

We illustrate working of the `ZipWeave` measurement framework by first assuming that we have a previous realization of coverage map of the region. This realization guides `ZipWeave` in pointing out the level of sampling density required at different parts of the region. Next, we take a realistic situation where an investigator does not have prior knowledge about coverage status of the region. Here `ZipWeave` initiates with first phase sampling and proceeds to next phase by optimizing an objective function. In both of these cases `ZipWeave` adopts a non-uniform sampling strategy which outperforms in coverage map accuracy compared to the benchmarked systematic-random sampling approach.

3.4.1 Sampling Strategy with Prior Coverage Information

`ZipWeave` provides a measurement collection framework both for a geographical area and for drive ways and streets. It identifies geographical patches and road segments that are similar in signal strengths and result in large clusters/segments at places having smooth variation in signal quality and small clusters/segments where nearby locations suffer from abrupt changes in signal quality.

In both these scenarios the assumption is of having previous radio map information for the region of interest. It is a top-down approach which after demarcating clusters/segments (with similar propagation characteristics) in previous map, recommends same amount of drilling from each of the identified cluster/segment irrespective of their sizes for future coverage map generation. We explain the cluster and segment identification processes in following subsections.

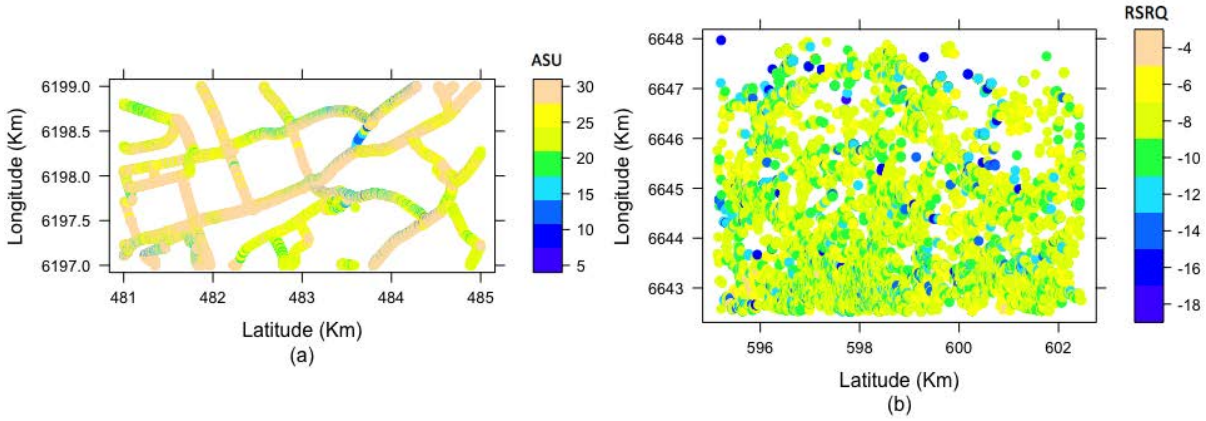


Figure 3.2: Parts of the (a) Edinburgh buses dataset (with signal quality in ASU) and (b) Oslo OpenSignal dataset (with signal quality in RSRQ) on which `ZipWeave` is applied.

3.4.1.1 Sampling in a Geographical Area

For a geographical area two datasets i.e. a dense portion of Lothian Buses dataset and Oslo OpenSignal dataset are studied (Fig. 3.2). To have a complete representation for each of the chosen area, first full enumeration dataset is generated by predicting signal strengths at regular grid locations narrowly spaced using the real set of measurement samples in that region and the OK interpolation process. The resulting full enumeration datasets of are shown in Fig. 3.3.

To identify spatially similar regions with respect to propagation features initially *Hierarchical Clustering (HCLUST)* is applied on the full enumeration dataset (i.e. on the interpolated signal strengths at gridded locations). For HCLUST, the similarity threshold is set to be within 4 signal strength units. This value is chosen empirically so that to avoid both a large number of small clusters and few number of large clusters having higher internal signal strength variation.

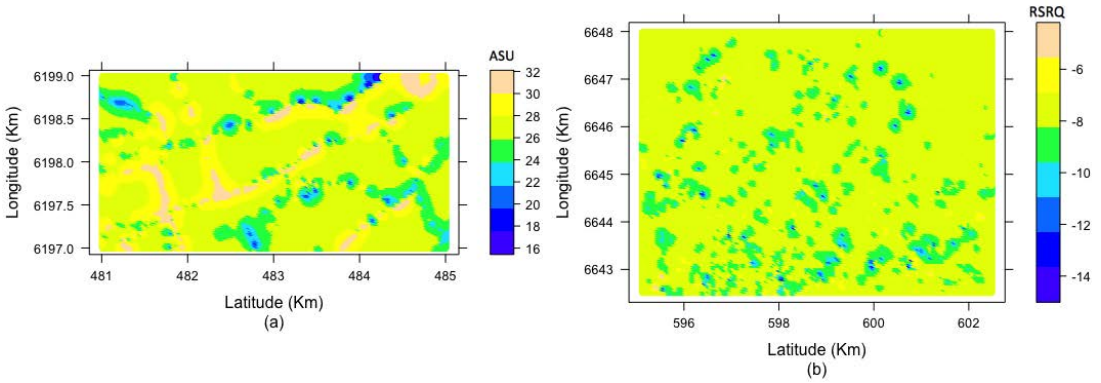


Figure 3.3: Full enumeration datasets in a rectangular region of both the (a) Edinburgh Lothian buses and (b) Oslo OpenSignal dataset scenario.

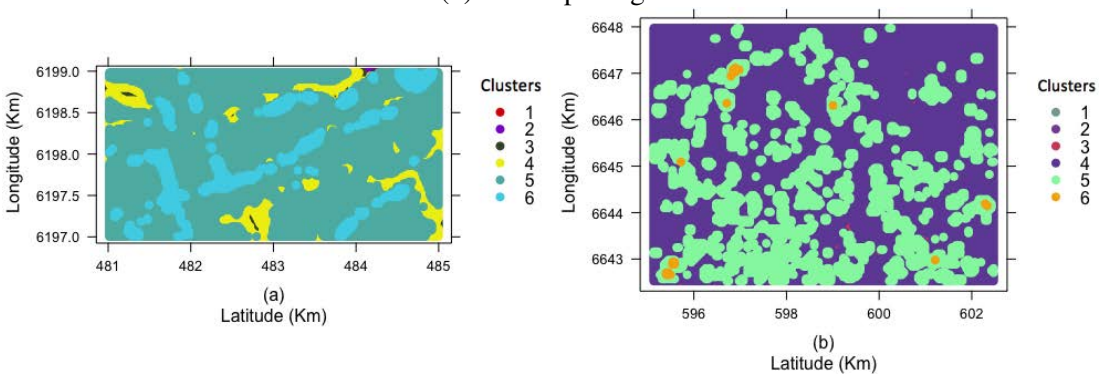


Figure 3.4: Clusters from HCLUST in(a) Edinburgh Lothian Buses (b) Oslo, OpenSignal datasets.

Fig. 3.4 shows clusters after applying HCLUST on the datasets. HCLUST clusters have similar propagation characteristics but members of most of the clusters are in the form of geographically disjoint clumps. To get clusters that are both similar in signal values and whose locations are geographically adjacent *Connected Component Labelling (CCL)* is applied on the resultant hierarchical clusters. CCL is a feature of image processing which identifies pixels in an image that have similar color code and are adjacent to each other.

CCL scans over each cluster identified by HCLUST and splits it into separate clusters such that member grid locations of each new cluster are both geographically adjacent and similar in signal quality. Other than large clusters the splitting by CCL creates many small clusters (Fig. 3.5) consisting of too few members representing great variation of its member locations from the surroundings in signal quality. The scantiness

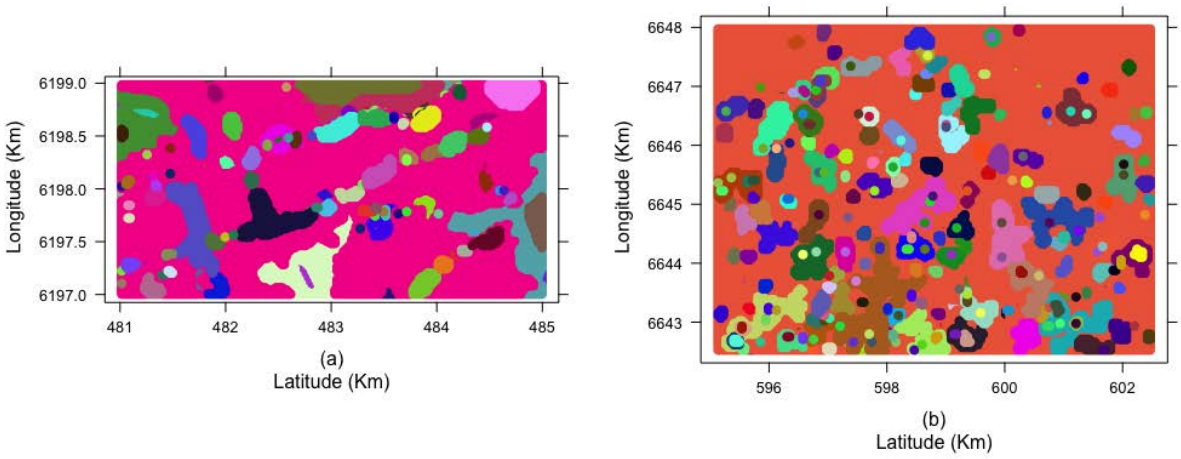


Figure 3.5: Color coded clusters from CCL in (a) Edinburgh Lothian Buses (b) Oslo OpenSignal datasets.

of these small clusters do not provide enough sampling choice and therefore need to be merged into bigger adjacent clusters so to absorb sudden changes in signal values and not to loose any probable sampling location. As a final step *K-NN* is used to solve this issue by applying 1-NN i.e. by merging a tiny cluster (in this study consisting less than 20 grid locations) into its nearest big cluster (Fig. 3.6).

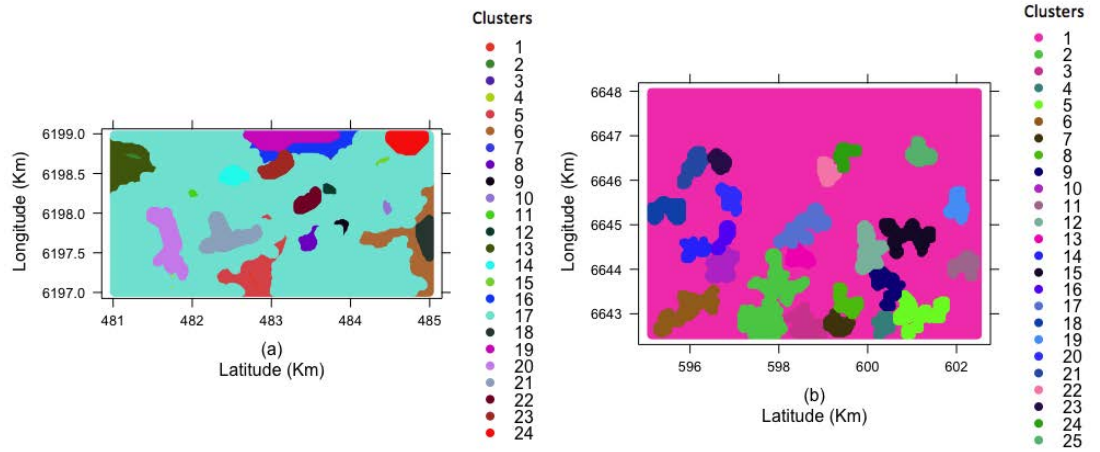


Figure 3.6: Reduced clusters after applying K-NN on tiny clusters of CLL from (a) Edinburgh Lothian Buses and (b) Oslo, OpenSignal datasets.

Evaluation: By taking systematic-random sampling technique as a benchmark, we evaluate the performance of `ZipWeave`. In systematic-random sampling we divide the whole region into equally sized squared tiles and then collect equal number of

samples randomly from each of the tiles. Tiles enable uniform sampling across the whole region and the randomness within a tile extracts correlation in both near and far samples. Table 3.1 and 3.2 shows coverage prediction results by taking 25% of the full enumeration dataset (almost 5000 samples in both cases) as validation set and 20 samples per uniform tile (in Systematic-random sampling) and per identified cluster (in `ZipWeave`) respectively (Fig. 3.7).

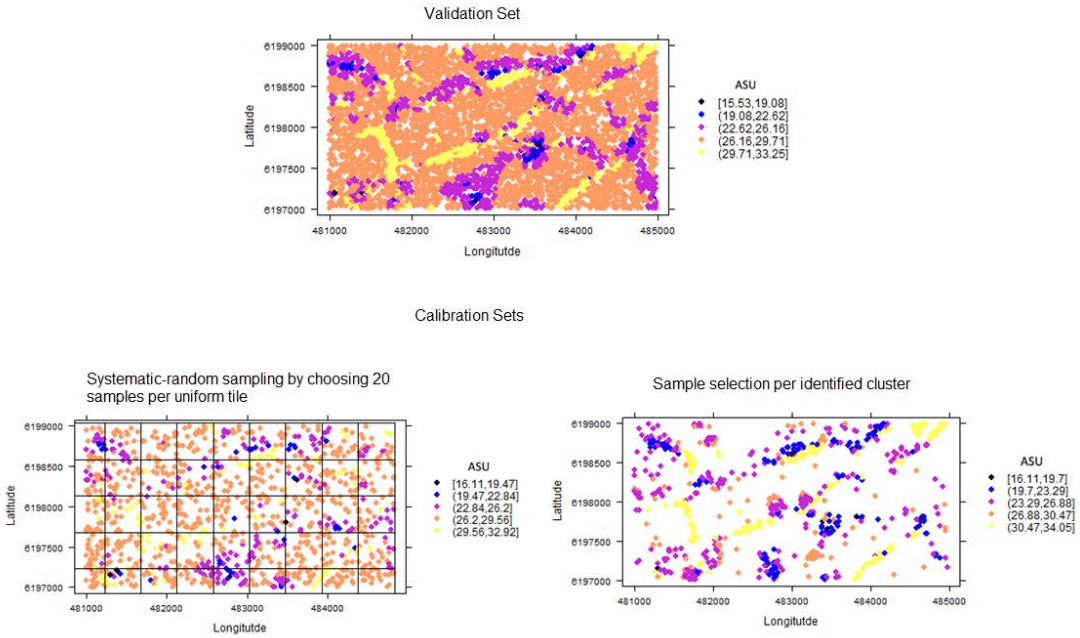


Figure 3.7: Sampling choice from the full enumeration, Edinburgh Lothian buses dataset, using Systematic-random (uniform-tile based) sampling and (`ZipWeave` based) Non-uniform sampling approaches

Each record in Table 3.1 and 3.2 shows raise in sampling density for the systematic-random sampling approach by decreasing the size of tiles (or in others increase in the number of tiles in the region). On the other hand, with ZipWeave each record shows mean prediction error by using a different set of samples per identified clusters. In both the test scenarios, `ZipWeave` outperforms with almost half the sampling size to that of systematic-random sampling approach.

Table 3.1: RSS MAPE with systematic-random sampling and ZipWeave using the controlled bus dataset from Edinburgh.

Systematic-random sampling		**ZipWeave**	
Sample Size	MAPE	Sample Size	MAPE
360 (18 tiles)	1.00	480 (24 clusters)	0.74
560 (28 tiles)	0.98	480 (24 clusters)	0.77
640 (32 tiles)	0.85	480 (24 clusters)	0.76
900 (45 tiles)	0.77	480 (24 clusters)	0.76

Table 3.2: RSS MAPE with systematic-random sampling and ZipWeave using the crowd dataset from Oslo.

Systematic-random sampling		**ZipWeave**	
Sample Size	MAPE	Sample Size	MAPE
500 (25 tiles)	0.31	500 (25 clusters)	0.26
600 (30 tiles)	0.32	500 (25 clusters)	0.27
720 (36 tiles)	0.31	500 (25 clusters)	0.27
840 (42 tiles)	0.31	500 (25 clusters)	0.26

3.4.1.2 Sampling on Drive Routes

For drive paths ZipWeave specifies segments on roads and streets where fewer number of samples per segment can predict well about the coverage status of the unsampled locations of that segment irrespective of the segment size. However due to the layout of the locations of interest (i.e. being limited to road and streets) ZipWeave is tweaked by replacing the clustering approaches that are suitable for retrieving road-based clusters or segments.

Constrained Clustering replaces HCLUST and CCL in extracting road-based clusters with members both geographically close and similar in signal strengths (a similarity threshold is specified for Constrained Clustering to work). Starting from the first sampling location on the road, CCL labels the nearest sampling location to be member of the same cluster if difference from its signal quality falls within the threshold. If the difference is higher, then the new sampling location becomes first member of a new

cluster. The process of merging next closer sampling locations into current cluster or generating a new cluster continues till the whole road (where previous sampling information is available) is traversed. Like its CCL counterpart, constrained clustering results in many tiny clusters where signal quality is highly variable from its adjacent locations (Fig. 3.8 (a)). This time merging of tiny clusters is performed by *Density-based Spatial Clustering of applications with Noise (DBSCAN)* [48] which has the characteristic to group together points that are closely packed marking as outliers points whose nearest neighbors are too far away (Fig. 3.8 (b)). It runs with two input metrics that is *MinPts* and *Eps*, where *MinPts* indicate minimum number of points that must lie within a *Eps* (i.e. radius) to make a cluster.

Since *DBSCAN-based Clustering* merges tiny nearby clusters, there is great variation within such clusters. When combined with big clusters from *Constrained Clustering* the overall prediction accuracy of `ZipWeave` decreases. Even with this decrease the prediction accuracy of `ZipWeave` still out-performs one achieved from systematic-random sampling approach (Table 3.3).

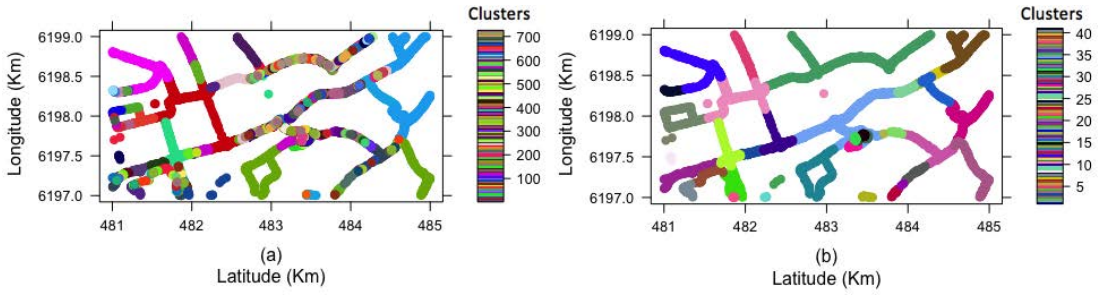


Figure 3.8: Road segments showing signal strength clusters (a) 411 in number, after applying Constrained Clustering (a) 42 in number, after applying DBSCAN to merge tiny clusters

3.4.2 Non-Uniform Sampling

This section presents a sampling strategy, which justifies the non-uniform sampling framework presented by `ZipWeave`, for a region that lacks prior coverage information.

The sampling strategy follows a multi-phase sampling process in which the first phase is performed to capture both the heterogeneity across the region and correlation in nearby locations. The sampling pattern recommended for the first phase is a systematic-random approach. Each subsequent phase of the sampling strategy then

Table 3.3: RSS MAPE with systematic-random sampling and ZipWeave using controlled bus dataset of a route.

Systematic-random sampling		**ZipWeave**	
Sample Size	MAPE	Sample Size	MAPE
660	2.64	600	2.07
720	2.44	600	1.84
980	2.26	600	1.89
1,100	2.24	600	1.85

chooses drilling areas that optimizes an objective function i.e. *Minimizing Mean Prediction Error*.

As stated earlier, unlike previous scenarios, here we do not have prior coverage information so to derive clusters/segments that are similar in propagation characteristics. Following subsections, therefore, verifies the suitability of non-uniform sampling strategy (guided by an objective function) over the systematic-random sampling approach (in later phases of sampling) for a previously un-sampled region.

3.4.2.1 Preliminaries

Before discussing the working of ZipWeave sampling process for regions with no prior coverage information, it is neccessary to understand the components used in the sampling process. These consist of a probability raster and an objective function. itemize

Probability Raster: To derive second phase samples the study exploits the idea of *probability raster* presented by Balanced Sampling scheme [49]. Probability raster is a vector with size equal to the number of sub-areas within the region of interest such that each value indicates the importance of a sub-area with respect to the sampling objective function.

Objective Function K. Krivoruchko and K. Butler [49], who worked on determining air pollution over a region, claimed that the probability raster should reflect statistical features such as the Kriging prediction stantard error or Kriging variance (KV).

Kriging Variance (KV) is specified as giving out the confidence interval of the Kriging predictions. In other words it is a means to indicate the accuracy of interpolated values.

In order to drill the region for better samples the locations which have higher KV, calculated via previous phase of samples, are selected as prospective sample locations for the next phase so that to raise prediction accuracy of a region. D. J. Brus and G.B.M Heuvelink in [50] has used spatial simulated annealing with the objective function of Minimizing Mean of Universal Kriging Variance to specify a spatial pattern for a sample selection. However J.K. Yamamoto [51], states that KV is a global variance and does not capture the local differences in values. It only depends upon the configuration of the samples and the global variogram. For example, if two set of samples have same layout but different values their KV will be same. In situation where there is lack of stationarity across a region of interest global KV is not a trust worthy indicator of prediction accuracy. J.K. Yamamoto [51] recommended use of Local Error Variance (also called as Interpolation Variance (IV)) for estimating the prediction accuracy, instead. It not only takes into account the Kriging weights but also the data values of the samples to measure the reliability of the estimated values. It is the weighted average of the experimental squared differences between data values and the Kriging estimates. Our experiments with different datasets showed that neither KV nor IV can be trusted to optimize sampling choice as both show little correlation with the absolute prediction error (shown in Fig. 3.9).

An always reliable indicator for prediction accuracy is the "Kriging prediction error" itself. It, however, needs knowledge about the actual signal strength values at the test locations so that to calculate the prediction error. Taking into account this constraint a feasible way is to use leave-one out cross validation (LOOCV) approach of Kriging. In LOOCV process the whole sample set except one sample is used as calibration set and the left out sample act as validation set. This process is repeated for each sample point in the sample set. This method only uses the samples in hand to decide about the required sampling density per sub-area.

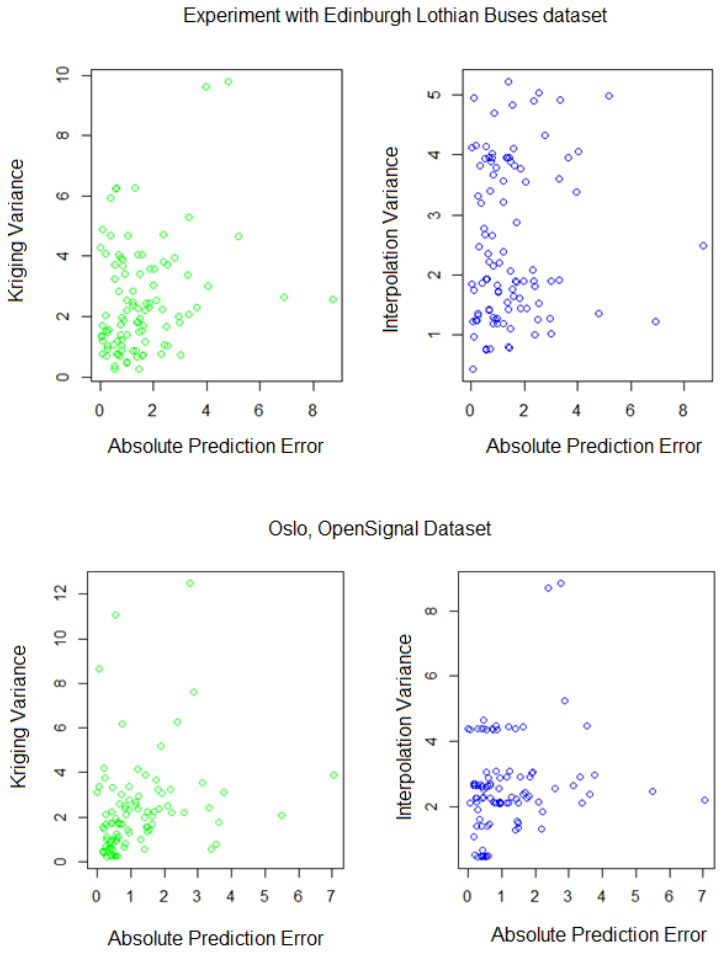


Figure 3.9: Correlation of Mean Kriging Variance and Mean Interpolation Variance with Mean Absoulte Prediction Error (at validation set of each uniform tile) .

3.4.2.2 Sample Selection Process

The proposed scheme starts with systematic-random sampling in the first phase of sampling. The whole area is divided into uniformly sized tiles and each tile is sampled equally so that to have representation from each part of the region under consideration. The sampling pattern within each tile is kept random so that correlation among close locations can be retrieved.

Next, OK with LOOCV is applied on first-phase samples of each tile to get Mean Absolute Prediction (MAPE) per tile. A probability raster is generated with size equal to number of tiles. For tiles with MAPE below the average of all the absolute prediction errors, a value of zero is assigned in the probability raster while for the rest of the tiles

normalized values relative to their MAPE errors are assigned such that the probability raster sums up to one.

During the next phase the tiles with non-zero value in the probability raster are only chosen. Number of samples taken from each of these chosen tiles is proportional to their normalized MAPE error in the probability raster (where total number of samples are equal to number of tiles with non-zero probability times N samples). Unlike this approach in Systematic-random sampling for each next phase N number of samples are retrieved from each of the tile.

Evaluation: Table 3.4 and 3.5 shows improvement in prediction accuracy by applying non-uniform sampling strategy on Oslo OpenSignal dataset (with 44 tiles and N=20 samples) and Edinburgh buses dataset (with 24 tiles and N=20 samples). With each subsequent phase the performance gap between systematic-random and non-uniform sampling raises. The results indicate `ZipWeave` to arrive at a coverage map accuracy with almost half the systematic-random sampling size.

Table 3.4: RSS MAPE with systematic-random sampling and the proposed non-uniform sampling scheme using Oslo crowd dataset.

Systematic-random sampling		**Non-Uniform sampling**	
Sample Size	MAPE	Sample Size	MAPE
880	1.19	880	1.19
1,580	1.15	1,193	1.13
2,257	1.10	1,427	1.09
2,917	1.09	1,737	1.07
3,574	1.08	1,967	1.04

3.5 Weaving Crowd and Controlled Measurements

Availability of both crowdsourced and controlled measurements at certain regions brings forth the question of their combined usage for coverage prediction. Under the impression of scalability brought by crowdsourced measurements and the reliability of controlled measurements it is assumed that utilizing both the datasets will raise accuracy of coverage prediction in a region.

Table 3.5: RSS MAPE with systematic-random sampling and the proposed non-uniform sampling scheme using Edinburgh bus dataset.

Systematic-random sampling		**Non-Uniform sampling**	
Sample Size	MAPE	Sample Size	MAPE
480	5.20	480	5.20
960	5.10	694	5.04
1,440	4.97	908	4.84
1,920	4.77	1,102	4.61

The *Weaving* part of the study recommends fusion of different types of datasources from same region. In our case study we have two different datasources for both the cities of Edinburgh (i.e. Lothian Buses dataset and OpenSignal measurements) and Oslo (i.e. NSB-train dataset and crowdsourced OpenSignal measurements). However the application of *Weaving* for Edinburgh is hindered by the difference of signal strength unit captured by each of the dataset.

Because of no compatibility issue for the Oslo datasets; weaving is done for its datasets. As we have controlled measurements from a single train route; we limit our available crowdsourced measurements to that route (shown in Fig. 3.10) and divide the route area into tiles; there are 22 such tiles having both the measurements from crowd and controlled dataset. Tiling is done to have samples from both datasets equally and from across the whole route region.

To evaluate coverage prediction from combined dataset we retrieved 1000 validation points from all these tiles such that it has equal representation from the route covered by control measurements and nearby places covered by crowdsourced measurements. Also we retrieve three different types of calibration datasets; *crowdsourced* (with a total 1000 crowdsourced samples from all the tiles), *controlled* (with 1000 controlled samples from all the tiles) and *combined* (with 1000 samples from the combination of two; such that 500 data points are chosen from each of the dataset). This process was repeated thrice (Table 3.6) and higher prediction accuracy was observed with the combined calibration dataset.

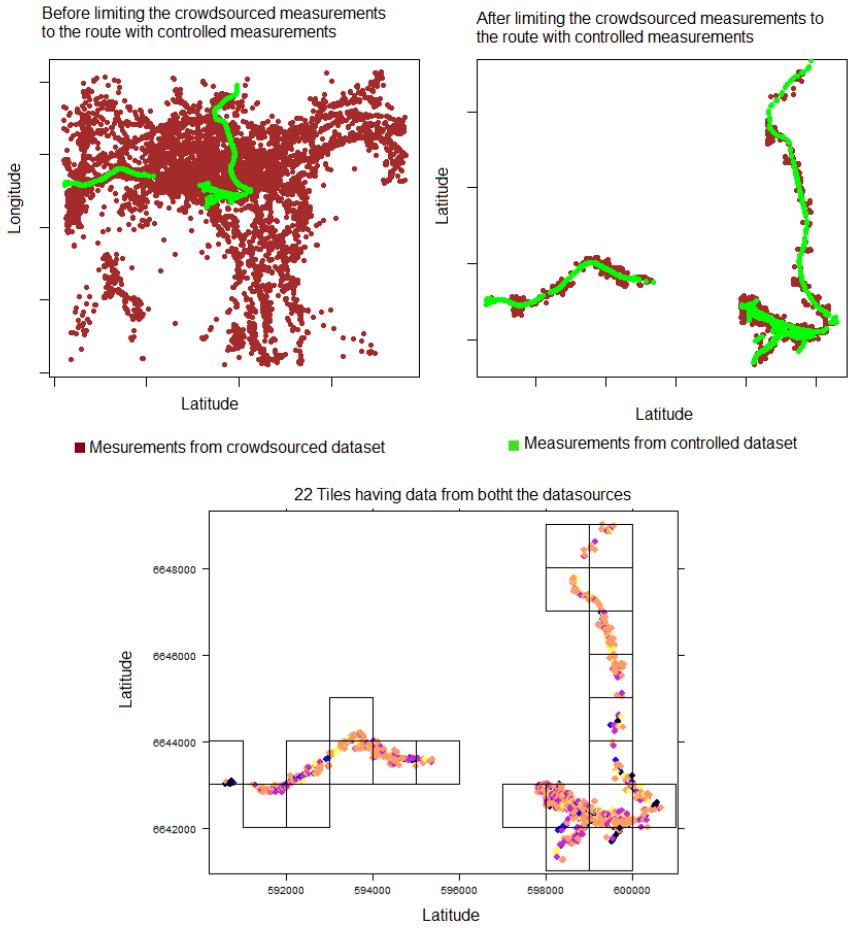


Figure 3.10: Controlled and Crowdsourced dataset of Oslo.

Though the improvement percentage is low but it may be due to the fact that in both the datasets the range of signal strength values is too narrow and similar (i.e. the first quartile of signal strength values is -6 dBm and third quartile is -9 dBm).

The results in Table 3.6 proves the benefit of combining different sources of measurements for raising accuracy of coverage prediction. We next look into the non-uniform sampling process from the perspective of combined data sources. Initiating with $N = 30$ combined samples from each of the specified 22 tiles in first phase and moving to second-phase sampling with both systematic-random sampling approach and non-uniform probability-raster based scheme we observed that (Table 3.7) the latter sampling scheme outperforms the earlier one in all four tests with smaller sample sizes.

Table 3.6: RSS MAPE, using OK on three different configurations of the calibration/ validation datasets.

Test No.	Crowdsourced	Controlled	Datasets Union
1	2.59	2.45	2.28
2	2.45	2.83	2.25
3	2.53	2.57	2.25

Table 3.7: RSS MAPE in four different tests with systematic-random sampling and non-uniform sampling strategy in second phase.

Test No.	Systematic-random sampling (Phase 1)	Systematic-random sampling (Phase 2)	Non-uniform sampling (Phase 2)
1	2.02 (547)	1.99 (968)	1.94 (650)
2	1.98 (547)	1.92 (968)	1.89 (631)
3	2.06 (547)	2.02 (968)	1.98 (669)
4	2.09 (547)	2.05 (968)	1.98 (677)

3.6 Summary

In this chapter, we presented a measurement sample selection strategy named as `ZipWeave`. `ZipWeave` reduces the sampling cost by half while retaining the desired accuracy of a network's coverage map. The sampling strategy exploits the similarity in the propagation characteristics of the sub-areas within a region. Due to terrain differences sub-areas differ in their sampling density requirement e.g. open areas need less sampling density than downtown like areas. It therefore identifies these sub-areas according to their sampling requirement. This identification is done both in regions which has at least a single realization of coverage status and in regions with no prior coverage information.

The study further recommends that weaving together of measurements from different data sources enhances the accuracy of a radio coverage map. It is demonstrated with a crowdsourced and controlled dataset which when combined adds scalability and reliability to the overall measurement set.

Chapter 4

Device-centric Accuracy of Crowdsourced Coverage Map

A common practice for operators is to use their infrastructure information and measurement based models to provide coverage maps of their networks on their respective official websites such as in [52] and [53]. Measurements to aid in this process are collected either with drive testing or increasingly via low cost user-side approaches like crowdsourcing [5]. The reported maps on such sites typically only display maps per radio-mode (2G/3G/4G) and are coarse grained in which they showcase either a boolean representation of signal quality namely 'weak' and 'strong' signal or on an unknown scale of 'low' to 'high' value at per zipcode level such as in [52]. Same can be said about the granularity of information provided by third party mobile coverage mapping sites such as OpenSignal [7]. Alternatively, signal quality over space can be mapped in a fine-grained manner. As mentioned in previous chapters, state of the art method for such mobile coverage map generation is to collect RSS measurement samples at some parts of the region of interest and interpolate signal quality at the unobserved parts.

With the proliferation of different smartphone models from different manufacturers in the market [54, 55], the question is how well the above outlined types of coverage maps represent the mobile coverage perceived by different devices when all other factors remain the same (operator, cell, location, etc.). In other words, how accurate is

the coverage map from any given device's perspective. This is a significant issue given the multitude of device types users carry and being used for collecting signal strength measurements, especially for crowdsourced coverage maps.

In view of the above discussion, this chapter examines the reliability of a coverage map generated using measurements collected from a large pool of diverse smartphone devices, locations and times from the perspective of a given device, which may or may not be represented in the set of measurement devices. In particular, using a large real-world crowdsourced mobile signal measurement dataset with over million measurements spanning an year from 80 different devices, we study:

- how well the coarse grained mobile coverage maps available from official websites of operators and crowdsourced coverage mapping sites correlate with coverage experienced for a given device type.
- how well measurements from different device models from the same vendor reflect the fine grained coverage status seen by any of those devices.
- the impact on coverage map accuracy when measurements come from devices across multiple vendors.
- how useful measurements from an earlier time span are to assess coverage map accuracy for a device at a future time.

Our key finding is that so long as a given device is among the set of devices contributing measurements underlying a coverage map, even if the measurements are from a previous time span, the map is reliable from the device's perspective. Our study also reveals guidelines for producing reliable mobile coverage maps in presence of device diversity.

This chapter consist of five sections. Section 4.1 provides background and related work on device diversity. After mentioning motivation of this analysis study in section 4.2, we describe the datasets used and methodology adopted in section 4.3. Section 4.4 starts with providing measurement based evidence to demonstrate the impact of device diversity on received signal strength distribution in both controlled and crowdsourced

settings. This section then focuses on assessing the level of difference/similarity one might confront when taking guidance from coarse coverage information, and impact on a fine grained device-centric coverage map accuracy when it is generated with samples from different crowdsourced measurement sources. Lastly we summarise the paper with key take-aways.

4.1 Background and Related Work

The large variation in devices and their characteristics result in varying degree of signal qualities at a given location even if served by the same cell of the same mobile network operator. For example, a study commissioned by the UK regulator Ofcom tested sensitivity level of ten different devices under different radios and frequency bands in free space and observed variations in their received signal strength (RSS) values [56].

Before mentioning the work done to minimize adverse impact of device diversity, we first enumerate some of the prominent reasons that result in differences in received signal strength quality of different device-types. There are a number of reasons [56,57] for RSS variation including but not limited to:

- *antenna design*: whether an internal or external antenna is used and its size can affect the gain of the device antenna;
- *device design*: materials of different devices can have different absorption effects on radio signals;
- *RF receiver design*: noise and nonlinearity at receiver circuitry of the device can affect the performance; and
- *number of frequency bands supported*: as more frequency bands are added to the device antenna the receiver design becomes more complex, which can make it more difficult to achieve good signal quality.

There are some other impacting factors for signal variation such as user's mobility, orientation of device, humidity and temperature but these can affect all types of devices.

The issue of device diversity impact and mitigation has been previously studied in the context of indoor localization based on Wi-Fi fingerprinting. Purpose is to accuratley locate an end-user when fingerprint database is generated with a device-type different from the one used by the end-user.The studies [58–60] resolve this issue by either proposing calibration or calibration-free methods. In calibration methods, before locating an end-user with fingerprint database, the signal strength values received from the surrounding APs are calibrated into readings similar to the device with which coverage map is generated using for example pair-wise linear transformation [58, 61]. On the other hand, in calibration free methods either relative RSS values of the sensed APs are used as in [59, 62] or APs are ranked according to their RSS value [63]. The relative relation or ranking is then used to locate end-user accurately avoiding the bias introduced by differences in the devices' capabilities. For calibration methods it is however supposed that enough measurements for a set of locations per device is collected in prior so that to derive an accurate mapping function, which fails when a new type/ model of device enters the building.

4.2 Motivation

To generate a device-centric view of a cellular coverage map, we can neither use the calibration-based methods [58, 61] nor the calibration-free methods [59, 62]. As cellular networks span over large geographical areas with variety of terrain properties, there is a small probability to have multiple measurements from all foreseeable devices for the same set of locations under similar environmental conditions, thus making it difficult to apply calibration based methods. Additionally mobile applications used for obtaining crowdsourced measurements in practice have access to RSS readings from only the serving cell tower, thereby making it impossible to use calibration-free methods.

Towards reliable outdoor WLAN radio map construction with diverse devices/ measurements, noise cancellation approaches have been studied [64, 65]. X. Fan et al. [64] propose a model-driven approach for taming heterogeneous noise generated by devices to produce a reliable radio map. Similarly, C. Xiang et al. [65] proposed CARM, a method to mitigate the effect of error-prone and inaccurate crowd-sensed

readings from crowd devices, to build better outdoor WLAN RSS maps. While the works of [64, 65] propose building of a single noise-free coverage map by dealing with noise or RSS variation that come with diverse devices, they do not examine the accuracy of the coverage map from the perspective of a device.

The aforementioned paragraphs indicate that neither calibration and calibration-free methods are workable nor the work done by [64,65] are about generating a reliable device-centric coverage map for a cellular network. To move forward in this direction the motivation of this study is to investigate the potential of different crowdsourced measurement sources in generating a reliable device-centric network-availability map and how these sources can be used in order of preference.

4.3 Dataset and Methodology

Datasets: For this study, we use a large crowdsourced measurement dataset for London from OpenSignal [7] spanning an year between 2012 and 2013. This dataset consists of measurements for two major UK mobile network operators, henceforth referred to as `Operator1` and `Operator2`. Moreover, the dataset consists of measurements from 80 different device-types, largely various models of Samsung and HTC. In terms of the radio mode, although the measurements span 2G, 3G and 4G measurements, we focus on measurements for 3G mode which has the most number of measurements. It is useful to note that majority of 3G measurements in our dataset correspond to HSDPA. In all, the dataset consists of over 1.129 million measurements. RSSI values in these dataset are represented in ASU [66].

It must be noted that within both the datasets from `Operator1` and `Operator2`, large proportion of samples are from the Samsung and HTC vendors. We, therefore, focused on evaluating different scenarios by using samples belonging to different models from these two vendors unless stated otherwise. In each scenario, we have tested combination of devices from these two vendors including GT-i9100, GT-i9300, GT-i9300P, GT-N7100, GT-N7105, HTC Desire, HTC ONE S, HTC ONE X, HTC EVO 3DX 515m, HTC Sensation 2710e, and Galaxy Nexus. The chapter however presents results from a single combination of these devices in each test case.

Methodology and Performance Metrics: To assess the accuracy of coarse-grained

coverage maps (with few discrete levels of signal quality like weak and strong) from a device's perspective, we use 'Overlap Coefficient', a similarity metric that measures the overlap between two sets.

For fine-grained measurement based coverage map generation, we rely on Ordinary Kriging (OK) which has been shown to be a robust spatial interpolation scheme [30] for crowdsourced measurements. And to assess the accuracy of such fine-grained maps we use mean absolute prediction error (MAPE) as the performance metric.

4.4 Device-centric Assessment of Coverage Map Accuracy

To understand the impact of device diversity on cellular RSS distribution, we take two case studies with measurements collected at same location by two different device models connected to same network and cell sector. In the first case, that we call as controlled, we collected samples by holding two set of devices, namely Samsung Galaxy S III and a device from Motrola vendor, side-by-side. Whereas in the second scenario that is crowdsourced, samples are from OpenSignal dataset collected at Edinburgh, UK by two different devices i.e. SM-G920F and SM-N9005 at same set of locations. Fig. 4.1 demonstrate that both in the controlled and crowdsourced measurement scenarios, different devices vary in their RSS distributions even if they are experimented within similar context.

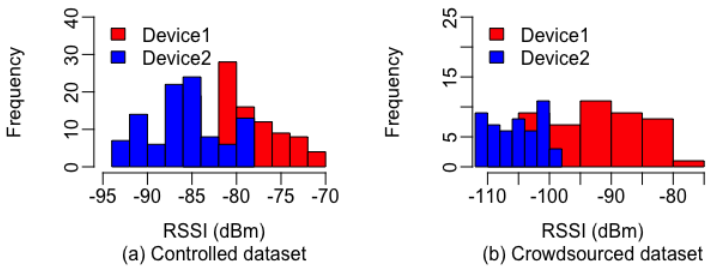


Figure 4.1: Variation in RSSI distribution due to device diversity.

The variation in RSS distribution from diverse devices impacts accuracy of resulting radio coverage map of a network. From a device-side perspective its experience

with a network has low similarity with the combined coverage map if the latter is generated ignoring samples from the device model. We demonstrate this by interpolating a combined coverage map, applying Ordinary Kriging, both on measurement sets including and excluding samples from the test device. Table 4.1 shows that ASU MAPE at test locations from that of a test device is higher when the generated combined radio map does not consider samples from the device model in question. We see substantial drop in error for different test devices when the samples from test device are included while generating a combined coverage map.

Table 4.1: Mean absolute prediction error in ASU with fine-grained coverage map

Method	**Operator1**		**Operator2**	
used	***Device 1***	***Device 2***	***Device 1***	***Device 2***
Not Included	4.49	7.28	4.92	5.69
Included	2.94	4.64	2.96	3.72
% Drop	34.5%	36%	39.8%	34.6%

In following subsections, we initially investigate the device-centric accuracy of a coarse-grained combined coverage map as is generally presented by official websites of the network operators. Secondly we evaluate the potential of different crowdsourced measurement sources in generating a fine-grained coverage map for a device-type.

4.4.1 Correlation with Coarse Grained Coverage Map

High ASU MAPE difference with the two variants of fine-grained coverage maps in Table 4.1 insist on assessment of coarse-grained coverage map that is generally displayed by operators and crowdsourced databases as a guide for customers. These maps present coverage quality in either a coarse boolean scale of weak-strong or 5-scale metric of low-to-high. Some other methods include a general description of what MBB services are available or if coverage is available indoors and outdoors. The geographic granularity of the displayed coverage information is also at coarse level of zip-code. We are interested to see how well correlated such a map can be when samples from device-type in question are not used in deriving the map.

Table 4.2: Overlap coefficient with coarse-grained coverage map

Method used	**Operator1**		**Operator2**	
	Device 1	***Device 2***	***Device 1***	***Device 2***
median RSS	81%	40%	70%	65%
max. RSS	95%	60%	77%	74%

We take a rectangular area of central London and divide it into grids of 50 sq. meter. We then specify if a grid has 'weak' or 'strong' signal on the bases of median/maximum RSSI of its enclosed samples. For 3G, OpenSingal considers a signal as weak if below -99 dBm and strong if above -85 dBm within a range of -51 to -113 dBm. Since in our dataset signal strength values are in ASU with range 1 to 20, we tag signals below 9 ASU as 'weak' and those above as 'strong'. For evaluation we take test devices from two different vendors from each of the operators i.e. `Operator1` and `Operator2`.

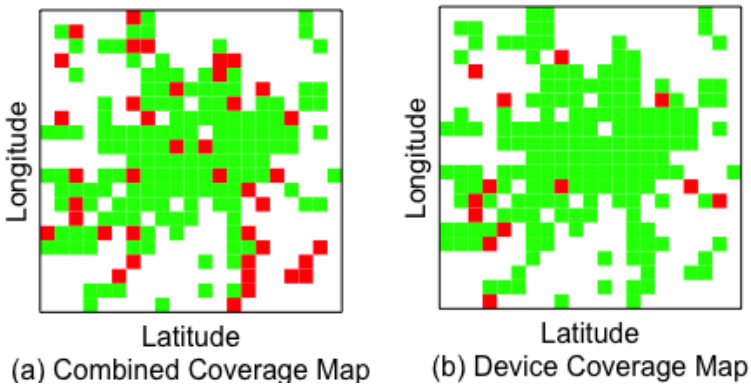


Figure 4.2: Part of a coarse radio map with weak (red) and strong (green) RSS samples generated with (a) measurement samples from devices (other than test device) and (b) samples from the test device

Fig. 4.2 displays a combined coarse map (ignoring samples from test device) and a device specific coarse-grained map with tagging done on median RSS per grid. Complete similarity results from this method are presented in Table 4.2, here we observe overlap coefficient is relatively higher for Device 1. This may be because most samples in dataset are from the vendor of Device 1, i.e. Samsung, with 47% & 56% share in

datasets from Operator1 and 2 respectively, where for the vendor of Device 2, which is HTC, this share is 9% and 14%. Secondly correlation improves substantially when instead of median, tagging uses maximum RSS per grid which however biases radio map towards good network availability. The similarity drop from Device 1 to Device 2 verifies that a single combined coverage map cannot well represent coverage status observed at a test device when it lacks measurement samples from the device and its vendor.

4.4.2 Impact of Different Measurement Sources on Fine Grained Device-Centric Coverage Map

A fine grained coverage map is desirable for correct decision making for example to spot out exact locations where network faces connectivity loss helping operator to know the cause (e.g. construction of a new building) and to act promptly for remedy. It assists users in finding places in residential and work area where network connectivity is seamless. With fine grained map, we mean a map with true signal quality at the granularity of longitude and latitude.

In this subsection we assess accuracy of a fine grained coverage map, from the perspective of a device-type, generated with measurements from diverse sources. Since we are using crowdsourced measurements, the assessed scenarios therefore include measurement samples from different crowdsourced sources. These sources specifically include samples reported from end-devices with different models, from different vendors, at different time spans and from different spatial locations. As we aim at generating a device-specific coverage map for a particular region, we consider all the samples collected within the spatial boundary of the region. For the remaining sources, we evaluate the difference in accuracy brought by each. Aim is to identify the order of preference in which samples from these different sources can be used so to maximize accuracy of the resulting coverage map.

Moreover, the tested locations in each of the following test-case depend upon the cells where both the validation and calibration devices have measurement samples. Care is taken in each test-case to have same calibration sample sizes for the device-models and vendors and time spans that are compared.

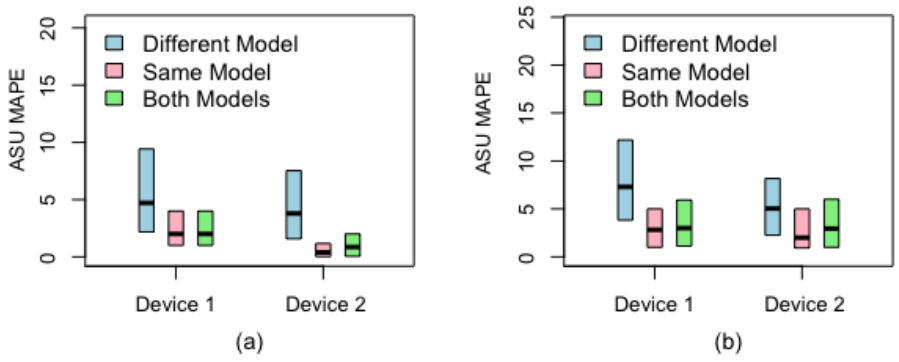


Figure 4.3: ASU MAPE with measurements from different, same and both the device models using (a) `Opearator1` and (b) `Opearator2` datasets.

Samples from different device-model Taking different case-studies we find that using measurements from a different device-model always produce a higher prediction error compared to when using measurement samples from same device-model. Fig. 4.3, illustrates results from two such devices with HSDPA radio. A useful observation, however, is that using different device-model's measurements in conjunction with samples from same device-model does not degrade accuracy considerably.

Samples from different device-vendor The accuracy worsens as samples from different device vendors are used to predict radio coverage status at a test device. Fig. 4.4 demonstrates it with results from devices belonging to two different vendors. We see that use of a different vendor generally raises prediction error by a higher percentage indicating to avoid using samples from different vendors, if possible, when generating coverage map for a device that lack samples.

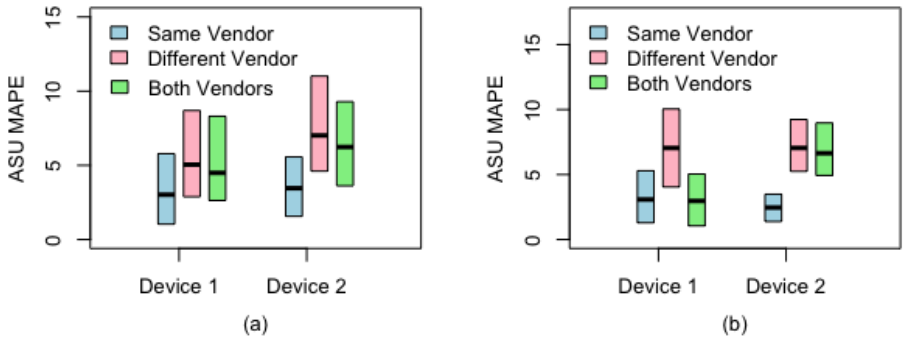


Figure 4.4: ASU MAPE with samples of devices from same and different vendor using (a) `Opearator1` & (b) `Opearator2` datasets.

Samples from past time-span To see value in samples collected in past time-span we use it in conjunction with measurements from 'target' time-span, a period for which coverage map of a device is desired. For purpose of illustration here we have split datasets from both the operator's in half on the basis of measurement reporting time.

Just like impact of using measurements from different device-model here too we find that measurements from past does not add to the accuracy of a coverage map and when used separately, produces higher error. Fig. 4.5 show results from two device models from each operator's dataset. Use of measurements from previous time span raises error above 50% with slight raise when measurements from both same and previous time span are used, recommending to use recent samples where possible.

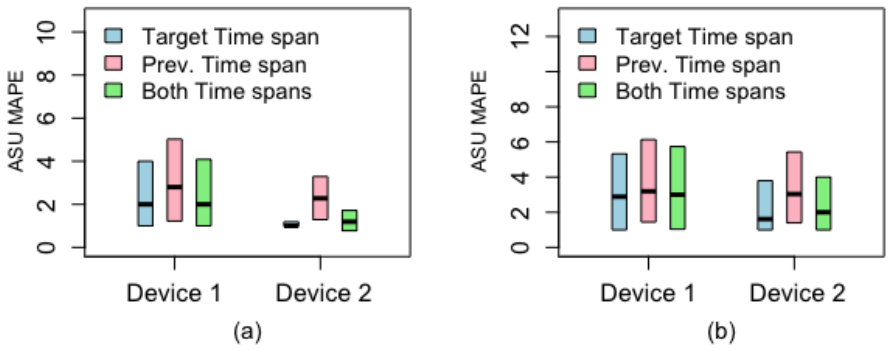


Figure 4.5: ASU MAPE with samples from target, previous and both the time spans for two device-types using (i) `Operator1` & (ii) `Operator2` datasets.

Previous coverage information vs different device-models If generating a device-centric coverage map two sets of samples are available i.e. samples from previous time-span with same device model and from same time-span but different device model, the analysis such as in Fig. 4.6 shows that good choice is to use previous information from same device-model.

4.5 Summary

Impact of device diversity has been studied widely in the context of fingerprint based indoor end-user localization. Due to multiple factors, devices vary in their RSS distribution even if held at same location, with same serving AP or base station. For

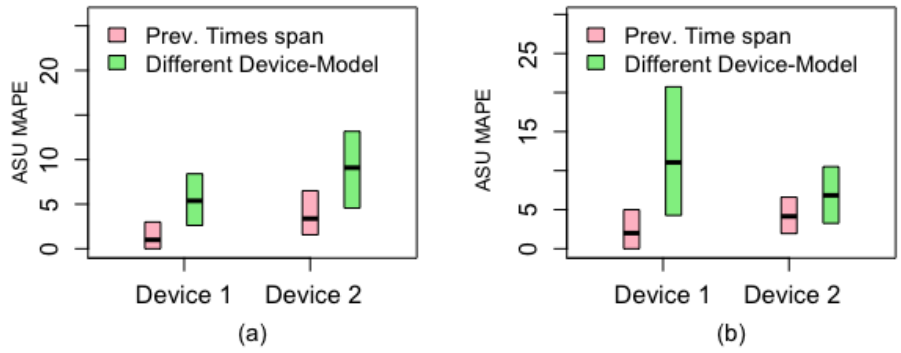


Figure 4.6: ASU MAPE with samples from previous time-span but same device-model and with different device-model but for same target time-span using (i) `Operator1` & (ii) `Operator2` datasets.

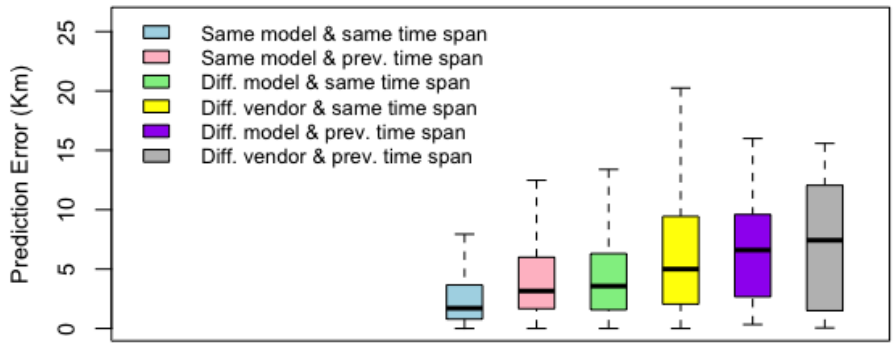


Figure 4.7: A general trend of impact on coverage map accuracy of a device-type when measurement samples from different sources are used.

indoor end-user localization use-case, pre-processing such as inter-device RSS mapping is performed on RSS readings, reported by an end device so to locate its user accurately. No such preprocessing is, however, performed when generating a cellular coverage map with crowdsourced measurement samples. Reason is its non-feasibility in the large-scale cellular network with samples coming from not only a large pool of diverse devices but also from diverse locations and at different timings. This analysis study therefore illustrates the relative impact on coverage map accuracy, from the perspective of a device-type, that is expected when using measurement samples from a different device-models, device-vendors and time-spans.

We find that the percentage increase or decrease in performance is highly variable across different case-studies, however, the general trend is similar. It is that measure-

ment samples from same device-model and same time-span always results in better coverage accuracy, followed by samples collected with same model but at different time-span e.g. previous year. For different models, we observed if they are from same vendor the error is smaller (substantially lower in some scenarios) than when samples from different device vendors are used. A general trend of impact on device-centric coverage map accuracy w.r.t. different measurement sources is listed in Fig. 4.7. Furthermore the analysis show that if measurement samples form different device-types/ models and time-spans are used in conjunction with test device-model's samples, the drop in accuracy is negligible.

In this study we also investigated the overlap correlation between coarse coverage map of a device with one similar to that displayed publicly by carrier network websites or crowdsourced databases. The correlation is high when most of the samples are from same vendor, indicating that for a combined coverage map to be useful it must take into account the measurement sample from widely used device-types.

Chapter 5

Localizing Base Stations with Crowdsourced Measurement Samples

Understanding the deployment patterns of communication infrastructure in general offers several benefits, including improving competition and quality/cost of services in the telecommunication markets to the benefit of consumers. However network operators treat their infrastructure related information,i.e. cell tower locations, as sensitive from their market position standpoint and generally do not disclose it, except to regulators and policy makers and that too with a non-disclosure agreement. Even though the knowledge of cell tower locations allows external validation of operator provided mobile coverage maps and more crucially enable identification of unserved or poorly served regions by correlating with population data [67], it is rarely available in the public domain. So measuring from the outside (say, from user devices) and making inferences about the infrastructure is therefore the only means to estimate this information.

In view of the above, our focus in this chapter is on estimating the cell tower locations from user-side measurements (i.e. crowdsourced measurements). The above outlined uses clearly indicate the value of such estimation for developing country settings, which we highlight in this chapter. However, generally speaking, knowing cell tower locations can enable several other applications. Device localization via trilateration from multiple nearby cell towers is a popular usecase, offering an alternative

to GPS when it is unavailable or for energy-efficiency reasons, as is evident from the cell tower location databases maintained by various location service providers [68–70]. Estimating cell footprint, improving accuracy of a coverage map and finding/tracking density of cellular infrastructure are some other applications. There are benefits even for mobile network operators such as locating the transmitters from rogue networks operated using SDR platforms and getting insight on where to grow their infrastructure depending on the infrastructure owned by other operators.

While relying on measurements contributed by users is a cost-effective means for measurement-based cell tower localization, accuracy and robustness become challenging due to the lack of control over the measurement process on the device side. Our analysis using a large-scale crowdsourced measurement dataset with ground-truth cell tower locations from OpenCelliD [70] reveals that none of the commonly used localization algorithms provide consistently good accuracy performance when used with crowdsourced measurements. Moreover we find that even the state-of-the-art cell tower localization approach, which we call as Filtered Weighted Centroid (FWC) [71], that filters out less predictive measurements from an accurate localization perspective is far from the best achievable localization outcome as it is limited by relying on a specific localization algorithm underneath.

Keeping in mind the above mentioned observations, we propose a novel approach called *Adaptive Algorithm Selection (AAS)* to select the localization algorithm, from a suite of algorithms, that is expected to provide the most accurate cell tower localization for a given cell and an associated set of crowdsourced measurements. AAS employs a framework based on supervised machine learning for this purpose. Through an extensive measurement based evaluation, again using the OpenCelliD dataset, we show that our AAS approach significantly outperforms FWC [71] by more than 65% in median error and reduces the mean error by more than half. At the same time, AAS achieves median error under 20% of the Oracle scheme that always picks the best performing algorithm. In addition, we show that AAS provides similar improvements even when applied for Wi-Fi AP localization. Even more crucially, we examine the applicability of AAS model trained in one setting to new settings for which there may not be ground-truth cell tower location information to build the model and obtain promising

results.

This chapter consists of five sections. Section 5.1 mentions related work regarding different categories of AP/ cell tower localization processes. Section 5.2 illustrates the datasets used, performance assessment metrics and the choice of commonly used localization algorithms that can work with crowdsourced measurement scenarios. In section 5.3 we examine the impact of crowdsourced measurement characteristics on cell tower localization and show that none of the commonly used algorithms nor the state of the art approach to filtering out less predictive measurements deliver robust localization performance. In section 5.4, we propose a novel supervised machine learning based `Adaptive Algorithm Selection (AAS)` approach for robust cell tower localization with crowdsourced measurements. Using large-scale crowdsourced cellular and WLAN measurement datasets, we show that AAS significantly outperforms existing alternatives. Before summarizing the chapter, in section 5.5 we present three case studies in three different countries in Africa showing the use of AAS based cell tower localization to reliably infer mobile infrastructure in developing countries. We also demonstrate some of the other use-cases to the benefit of different stake-holders.

5.1 Related Work

Positioning algorithms can be broadly divided into four categories on the basis of features that they use i.e. angle-of-arrival, geometry, RSS and path-loss propagation based schemes. Related work in each of the category is given below:

5.1.1 Angle of Arrival based Schemes

To determine the direction of an AP these schemes either use multiple directional antennas or a steerable beam antenna along with other measurement equipments. Such antennas help the measurement device to locate an AP by assessing the direction of highest RSS such as done by [72].

To avoid the use of expensive non-commercial off the shelf directional antenna Borealis [73] was proposed by Z. Zhan et al. in which human body was taken as a center point around which the smart phone was rotated. During a particular sector where human body acted as a blockage between an AP and the smartphone; largest

dip in signal strength was expected to be observed. The AP was supposed to be in the opposite direction of the middle of such a sector. The step size towards an AP depended upon the confidence of the predicted direction. The approach though cheap is not scalable and is time costly. In our evaluation we did not use this category of algorithms due to their non-applicability for cellular base stations where measurements are gathered from end-devices in the wild.

5.1.2 Geometry-based Schemes

These schemes are greatly influenced by the 2-D layout of measurement samples. The simplest approach in this category is *Centroid* which predicts transmitter to be lying at the mean location of the measurement samples. An improvement over Centroid is *Weighted Centroid (WC)* [74] which claims that measurement locations should not be treated equally as locations where RSS is high the chances of transmitter residing nearby is higher too. It estimates a transmitter's position by RSS weighted mean of measurement locations.

Both the Centroid and WC approaches give poor location estimate when measurements are skewed. To overcome the probable skewness of the measurement locations, *Minimum Enclosing Circle (MEC)* approach is presented by [75], it draws a circle around the measurements such that all measurements are enclosed with minimum possible radius. The center of the circle is then supposed to be the location of AP. Although it deals with the skewness problem it ignores the fact that most of the cellular antennas are directional covering a single cell sector, in which situation the MEC approach results in large errors.

5.1.3 RSS-based Schemes

RSS (Received Signal Strength) based schemes are the most studied approaches due to supposedly high correlation of RSS with the distance to its transmitter. The most simplest of which is *Strongest RSS (SRSS)* based localization scheme. It chooses transmitter location to be close to the measurement location where the value of sensed signal strength is strongest. This method may report inaccurate location when the chosen measurement location is effect of a reflection or when there are multiple spread

out locations with same highest received signal strength.

To tackle with the reflections and multiple measurement locations receiving same high signal strength, a solution pointed out by E. Neidhardt et. al [75] is to divide the whole region into grids and to calculate the mean and variance of RSS per grid block. The center of the grid block with highest probability of observing high signal strength is chosen to be the transmitter's location, we name it as *Grid Likelihood (GL)*. There are a few gradient-based approaches that derive position of a transmitter by determining its direction per grid from RSS. For example the schemes presented [76, 77] first averages locations and signal strength of measurements per grid to mitigate the effect of various sources of noise. Then a plane is fitted across the near-by grids in latitude-longitude-RSS space and gradient of this space is derived from the middle of each grid resembling an arrow pointing towards the direction of increasing RSS which denotes direction of the AP. AP is supposed to be at a location where the sum-squared angular error (weighted by RSS) to the arrows is minimum.

A somewhat different and simpler method is defined by Y. Cho et. al [78] where *Gaussian Process Regression (GPR)* is combined with the Weighted Centroid (WC) to improve the results of WC. The author claims that by interpolating signal strengths in form of a grid reduces the skewness problem that haunts WC. After interpolation the location with highest signal strength is chosen and compared with the location indicated by simple WC. If the difference is higher than the general error range of WC (in the paper [78] it is said to be upto 10 m for indoor WiFi AP) then the next higher RSS location is chosen until the error threshold is crossed. Though this approach improves the position error by 34 percent, the question is how to define the error threshold for cell tower localization and outdoor WLAN APs where the location inaccuracy of WC is highly variable depending upon the configuration of chosen samples and range of transmitters. GPR is also used in a multiple studies to localize an end-user [79, 80].

5.1.4 Path-loss Propagation (PLP) based Schemes

Path-loss propagation formula explains received signal strength at a location to be inversely related to the distance from the transmitter. The formula also incorporates other parameters such as transmitted power of the signal, path loss coefficient (that

depends upon terrain) ranging from 2 to 4, noise and unit path loss value.

WiGEM [81] proposes use of expectation–maximization (EM) model with initial parameters decided by log-distance formula for a grid-based structure, to assist in identifying most likely grid location from where a device can transmit packets with certain received signal quality at multiple known sensor locations. For AP localization [82] proposes *Monte-Carlo Path Loss (MCPL)* model by using a measurement set, log distance path loss model and possible AP locations in the form of a grid. It first estimates the unit path loss value and path loss coefficient for each possible AP location with the help of measurement set and by using least square method. Next it generates the received signal strength for each measurement location by feeding the calculated path loss coefficient and unit path loss value, in path-loss propagation formula, for each grid location. The grid location that results in minimum sum of difference between estimated and actual RSS, for all measurement locations, is termed to be the probable location of an AP. The main limitation of the scheme is the dependency on grids; as smaller grid sizes bring greater accuracy but with the expense of greater time complexity.

A more sophisticated approach is given by A. Achtzehn et. al [83], where least square method is used to identify the values of the propagation model parameters with the help of measurement samples and a set of known AP locations. The estimated locations of unknown APs are then calculated by using the propagation formula and the measurement set. The accuracy of the method raises with increase in the number of known AP locations, which is also the downside of this approach because in most of the cases the ground truth locations of APs are not known at all.

With crowdsourced measurements angle-of-arrival based approaches can not locate transmitters. From rest of the three categories section 5.2.3 refers to AP localization schemes that are applicable to locating cell towers using crowdsourced measurements.

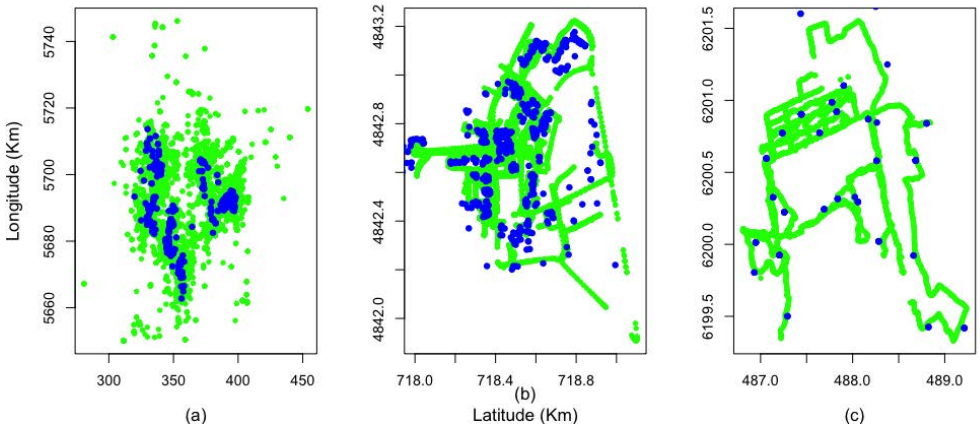


Figure 5.1: Measurement samples (in green) and ground truth transmitter locations (in blue) from (a) a LAC from OpenCelliD sub-dataset, (b) Darmouth WLAN dataset and (c) RF Signal Tracker dataset.

5.2 Preliminaries

5.2.1 Datasets

We use four different types of datasets as listed in Table 5.1 with nearly 12 million measurement samples spanning over 15000 cells. Majority of our measurement data is from OpenCelliD [70], a community project aimed at building a database of cell towers around the world based on crowdsourced signal measurements. We have used its sub-datasets from Germany, Poland, Zambia, South Africa and Morocco. In most of these cases, each cell has at least 50 measurement samples; each sample corresponds to mobile network signal strength from a user device stamped with location, time, cell id (CID), radio access technology (RAT), etc. For the OpenCelliD datasets, we use measurements for various 2G variants (GSM, GPRS and EDGE) as they are the largest in number. Ground truth cell tower locations are only available for Germany and Poland.

5.2.2 Metrics

We mainly use two metrics to characterize the accuracy & robustness of cell tower localization approaches including our proposal: (1) *Mean absolute prediction error (MAPE)*, defined in this chapter as the average Euclidean distance between estimated

and ground truth locations of cell towers (or Wi-Fi APs), across all towers (APs); (2) *Median absolute prediction error*, defined similarly but focusing instead on the *median* of the errors. We also make use of box plots, bar charts and CDFs of localization errors in some cases to draw attention to the distribution of localization errors produced by different approaches.

5.2.3 Localization Algorithms

For this study, our choice of algorithms is driven by their suitability of use with crowdsourced measurements. We, therefore, do not consider angle of arrival based approaches such as DrivebyLoc [72] and Borealis [73]; coarse-grained approaches that give zip code level estimates for transmitter location such as in [85]; as well as approaches that require meticulous orchestration of measurement collection as in CrowdWiFi [86]. We prefer commonly used algorithms mentioned by [87] and [71] in their work that are applicable to the uncontrolled crowdsourced measurements. All of these algorithms, belonging to the remaining three localization categories consists of Minimum Enclosing Circle (MEC) [75], Centroid (C) [75, 76], Weighted Centroid (WC) [74–76], Strongest RSS (SRSS) [74], Grid Likelihood (GL) [75, 88], Gaussian Progress Regression (GPR) [78] and Monte-Carlo Path Loss (MCPL) [82] approaches. We choose five of these algorithms in our study, the reason for dropping MEC and GPR is their comparatively higher localization errors with the crowdsourced cellular measurements shown in Fig. 5.2.

5.3 Motivation

5.3.1 Impact of Crowdsourced Measurement Characteristics

Crowdsourced measurements are uncontrolled as they are reported from random locations at diverse environmental situations, times and devices. Here we state some of the characteristics of crowdsourced measurements that either favor or hurt cell tower localization accuracy of a given algorithm, motivating the need for our proposed adaptive algorithm selection approach.

Inaccuracy in measurement location. We use the RF Signal Tracker dataset to

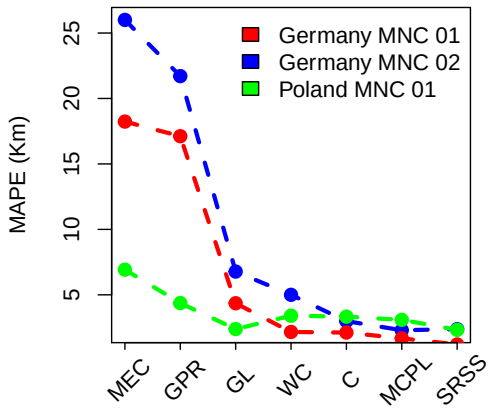


Figure 5.2: Error performance of different localization algorithms using OpenCelliD datasets with ground truth cell tower locations.

assess the impact of measurement location inaccuracy on a transmitter's predicted position. This dataset consists two sets of measurements for same route but one collected with GPS and other with network-based location information. The inaccuracy for GPS-based measurement locations ranges from 8m-16m with median value of 12m while for network based locations it ranges from 20m-40m with median of 22m. With these two sub-datasets, we observe in Table 5.2 that inaccuracy in measurement locations adversely affects all types of localization algorithms but to different degrees. RSS and Path Loss Propagation (PLP) based schemes seem to be much more impacted compared to geometric algorithms. This is because geometric schemes take into account overall spread of measurements and are not sensitive to location errors unless center changes. Localization accuracy with MCPL, on the other hand, degrades the most because of non-alignment of samples' distances from probable cell tower location and their path loss values.

Layout of measurement samples. Crowdsourced measurements come in different layouts. For example, samples generated by pedestrians and passengers are along streets and roads while those at hotspots, homes and work places may have more random locations. These layouts are highly dependent upon deployment location of a cell, surrounding landscape, crowd population density, their moving patterns in the cell, and a cell's footprint.

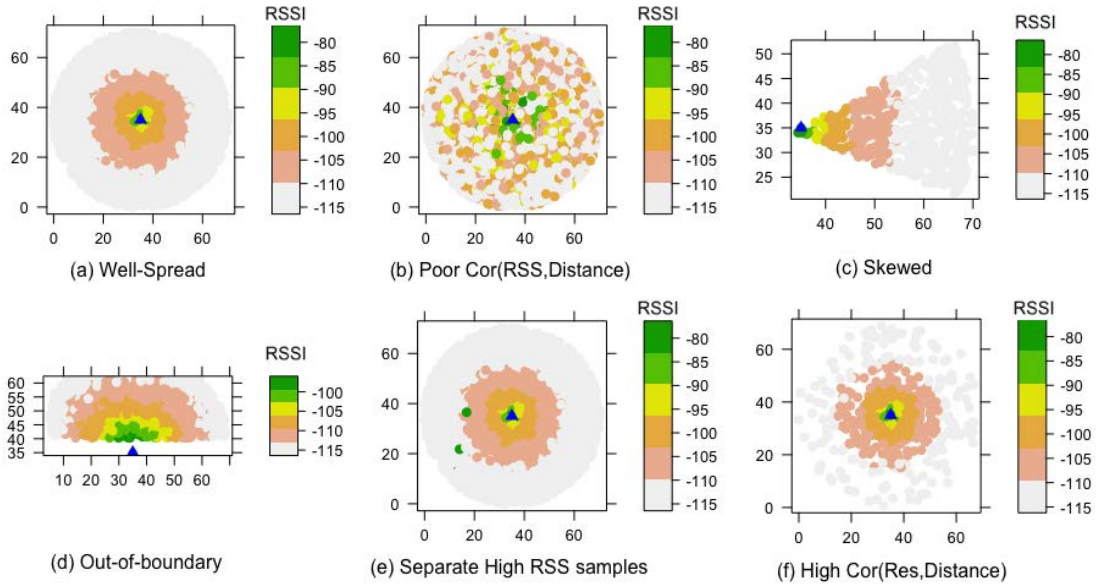


Figure 5.3: A cell tower with different spatial layout of its measurement samples.

Fig. 5.3 shows some of these layouts generated with the synthetic dataset mentioned in Table 5.1. The *well spread* case shown in Fig. 5.3 (a) is a good layout for each of the localization categories and it also exhibits high -ve correlation between samples' distances to cell site and their corresponding RSSs. When the layout is *skewed* as shown in Fig. 5.3 (c), localization errors for schemes in the geometric category are high (Table 5.2) as they incline towards the center of the sampled region. Directional tower up on a hill, restricted building or like features make measurements to be reported away from the cell tower location leading to *out of the boundary* (Fig. 5.3 (d)) layout. Such a layout is poor for all localization algorithms though to different degrees as each algorithm estimates cell location to be lying somewhere inside the measurement boundary. A favorable layout for RSS based approaches is when there is a single RSS peak while the opposite is true when there are multiple separated peaks shown in Fig. 5.3 (e). Such a *Separate High RSS samples* layout makes the tower location prediction erroneous especially for SRSS as indicated by results in Table 5.2.

Correlation of RSS to distance from cell tower. With free-space path-loss, signal strength is inversely proportional to the square of the distance from the transmitter. This ideal relation, however, does not always hold in real-world situations because

of effects (reflection, etc.) from the ground and objects in the path. Localization algorithms that depend highly on signal strengths of samples are likely to be impacted adversely as the correlation between samples' signal strength and distance to cell tower weakens. Using the synthetic dataset visualized in Fig. 5.3 (b), we observe from results in Table 5.2 that, when compared to *well spread* layout, the percentage degradation in localization accuracy is substantial for the algorithms relying on signal strength quality of the sampled locations. In contrast, accuracy with the Centroid algorithm remains unaffected. WC, on the other hand, also observes a dip in accuracy as it takes RSS values as a weight for calculating centroid of measurement locations.

Correlation of Response Rate to Distance from Cell Tower According to a general assumption an end-device hears a cell tower more when close to it. Response rate, therefore, means number of measurements fetched at a particular distance from its cell site. When correlation between response rate and distance to cell tower approaches to -1, geometric algorithms especially Centroid and WC ends up in pulling estimated location closer to the ground-truth location of the cell tower. This improvement can be seen in Table 5.2, where accuracy raise for these algorithms is substantial in comparison with *well spread* layout that has lower corresponding correlation. In *well spread* layout, response rate raises with distance from its transmitter.

Is there a clear winner among the commonly used measurement-based cell tower localization algorithms? We now examine the overall error performance of five commonly used cell tower / AP localization algorithms over three real-world crowdsourced measurement datasets: OpenCelliD datasets for Germany MNC 01 and Poland MNC 01; and Dartmouth WLAN dataset (Table 5.1). For this analysis and henceforth, *a measurement scenario* is a set of measurements (equivalently, samples) available for a cell tower (AP). For each measurement scenario in each of the three above mentioned datasets, we apply each of the five algorithms and calculate the percentage of measurement scenarios in a dataset when each algorithm provides the least localization error. If there was a clear winner among the algorithms, then that algorithm would have a percentage of 100 at the expense of other algorithms. Fig. 5.4 shows the results in the form of pie charts. While SRSS is the best performing one for majority of measurement scenarios in all three datasets, it performs really poorly when it is not the best. In

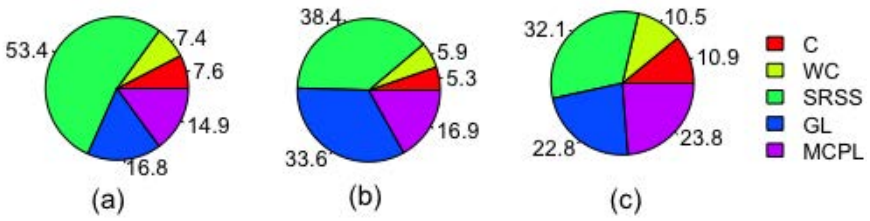


Figure 5.4: Percentage of measurement scenarios in (a) OpenCelliD Germany MNC 01; (b) OpenCelliD Poland MNC 01; and (c) Darmouth WLAN datasets for which each of the five commonly used localization algorithms performs best.

other words, SRSS has a long tail with some extremely high errors ($>$ 100Km in cellular dataset) that leads to higher mean and standard deviation of errors (above 1.5Km and 5.5Km, respectively). In summary, we conclude that there is no single consistently best performing algorithm from among the commonly used localization algorithms.

5.3.2 Limitations of Existing Localization Approaches to Deal with Crowdsourced Measurement Characteristics

As the peculiar characteristics, noise and outliers in crowdsourced measurements can negatively impact the accuracy of any given wireless infrastructure localization algorithm, some research studies employ pre-processing of measurements before application of a localization algorithm. For example to improve localization accuracy of SRSS and WC, J. Yang et. al [74] introduced three pre-processing steps, namely *RSS Thresholding*, *Boundary Filtering*, and *Tower-based Regrouping*. Concerning rationale behind the first step, the authors in [74] argue that once the strongest observed RSS drops below -60 dBm, the localization error of the SRSS algorithm increases significantly due to drop in correlation between the strongest RSS and the distance to the cell tower. Secondly if such a sample lies at the boundary of measurement layout, they suggest to exclude such samples. However we observe in our datasets that for 50%of the cases where maximum RSS is below -60 dBm, SRSS localizes cells within 500m error, which is not a very high error for cell tower localization. Moreover, rather than exclusion we believe that coarse estimation is better than no estimation as it provides a rough idea of probable deployment area of a network's infrastructure. For the third step, authors in [74] claim that merging measurements of cell sectors with same

cell tower improves WC results as it removes the ill-effect of skewed measurements. Merging cell sectors is, however, beneficial only when one knows naming pattern of CID's associated to sectors of same cell tower. In an other study, to have a reliable analysis out of crowdsourced data, F. Ricciato et al. [89] presented a few solutions to some crowdsourced measurement issues including identification of erroneous cell-IDs, unrealistic cell sizes, effect of antenna dragging and outliers.

To get the most out of a crowdsourced measurement dataset, the work of Zhijing Li et al. [71] can be regarded as the most recent work. It assesses the predictive value of a subset of measurement samples and finds that samples with high RSS standard deviation ($>$ 100k) and low RSS-weighted dispersion mean ($<$ 0.5km) correlate to high localization accuracy for WC. Based on this observation, they devised a variant of WC which we call as `Filtered WC (FWC)`, it relies on measurements that meet the above two criteria. To see if FWC offers a satisfactory alternative to the five algorithms studied in Fig. 5.4 above, Fig. 5.5 (a) shows a measurement scenario where FWC chooses a smaller subset of samples as predictive with mean dispersion of samples 635m and standard deviation of RSS double to that of the whole measurement set. While filtering measurements can be useful sometimes, there are also pitfalls underlying the FWC approach:

- One has to iteratively collect more measurement samples until RSS-weighted dispersion mean threshold and high standard deviation of RSS samples are met, which may not be practical if given a dataset that does not meet either of these criteria.
- If a measurement subset meets the two criteria used in FWC, it is not always true that the left out measurement samples perform poorer with traditional WC approach.
- Finally FWC bases its localization on a single algorithm (i.e. WC). As we saw in Fig. 5.4, none of the algorithms is clearly superior over others. This can be further verified with localization errors in Fig. 5.5 (b), where another algorithm, SRSS, exploits the available measurement samples better than the FWC algorithm.

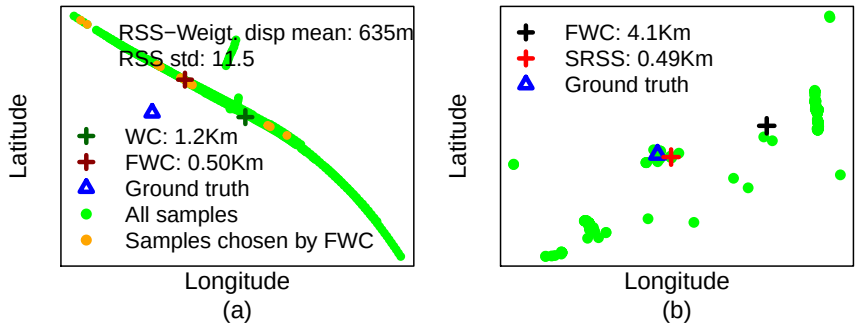


Figure 5.5: Example measurement scenarios that show (a) FWC outperforming WC via fewer carefully chosen samples; (b) SRSS yielding a significantly better localization accuracy than FWC.

5.4 Towards Robust Cell Tower Localization

Results and discussion from the previous section show that relying on a particular localization algorithm or filtering out measurement samples both limit localization accuracy that can be achieved for a given measurement scenario. So in this section we propose an alternative novel paradigm which is to use all available measurements and choose between different localization algorithms (for example, the ones studied in Fig. 5.4). This paradigm is in sharp contrast to the approach taken in FWC [71] where the algorithm to be used is fixed first (WC) and then the measurements are filtered to retain only those that are likely to help in achieving high localization accuracy. Note that we do not filter out any samples with the rationale that the algorithm if carefully chosen can exploit all the available measurement samples.

5.4.1 Probabilistic Algorithm Selection

We first examine a naive instantiation of the new paradigm to choose between different algorithms. Based on some measurement data for a set of cells (measurement scenarios) and corresponding ground-truth cell tower location information that can be viewed as *training data*, we estimate relative percentage of scenarios each algorithm yields best localization result as in Fig. 5.4. Such a pie chart is used for all subsequent cell tower location estimations for probabilistically choosing an algorithm; essentially, percentages in the pie chart serve as prior probabilities for picking different algorithms.

We refer to this approach as `Probabilistic Algorithm Selection (PAS)`.

Using 10-fold cross-validation on the three OpenCelliD datasets from Germany and Poland, we see that PAS outperforms FWC by 46 to 65% in the different datasets as shown in Fig. 5.8. It is however off from the "Oracle" results by more than 57% both in mean and median errors. By Oracle, we refer to a scheme that can always select the best performing algorithm (from among the five individual algorithms) given a measurement scenario. The results in Fig. 5.8 also show that PAS performs poorly compared to SRSS, the algorithm that yields the lowest localization error in majority of the scenarios.

5.4.2 Adaptive Algorithm Selection

The results from the previous subsection suggest that while the simple-minded PAS (reflecting the approach to choose between algorithms) is already better than FWC (that is based on a specific algorithm – WC) it leaves room for substantial improvement compared to Oracle and SRSS.

In light of the above, we propose a more sophisticated variant called `Adaptive Algorithm Selection (AAS)`. AAS views the problem of choosing a localization algorithm from among the suite of different algorithms as a classification problem in machine learning – different algorithms make up different classes for the classifier. Unlike PAS which somewhat randomly selects an algorithm with no regard to the specific characteristics of the measurement scenario for which cell tower needs to be localized, AAS classifier model considers a variety of features (outlined next) that aid in distinguishing between different measurement scenarios and algorithms.

5.4.2.1 Feature Set

In Table 5.2, we illustrated some of the features that have varying degree of impact upon the three categories of algorithms. To have an in-depth understanding of the combination of features that can serve as a guide for assessing the suitability of a localization algorithm for a given measurement scenario, we extract four types of features as listed below:

- **Measurement Spread Features:** These include size (number of measurement

samples); radius (spatial spread of the samples as determined by the radius of the minimum enclosing circle); DistTl (mean distance of all samples to the "trend" line of the samples); DispCenter statistics (i.e., mean, median, standard deviation and index of dispersion of the samples from central location of minimum enclosing circle); and Density (mean number of samples per sq. km across a measurement scenario).

- **Signal Strength Features:** These consist of received signal strength (RSS) statistics as well as highest RSS statistics. For the latter, we use number of highest RSS samples; minimum, maximum, mean and standard deviation of distances among highest RSS sample locations.

- **Weighted Measurement Spread Features:** These include DispRSS statistics (i.e., mean, median, standard deviation and index of RSS based weighted dispersion from central location) and Autocorrelation among samples.

- **Features based on Estimated Locations:** These extract correlation between signal quality of measurements and their distances to estimated locations of the five algorithms in Fig. 5.4; distance between each pair of estimated locations and distance of each estimated location to the center of the trend line.

5.4.2.2 AAS Model Generation

We take a supervised machine learning approach to the classification problem stated above. To generate the `AAS` classifier model, for a subset of measurement scenarios with ground-truth cell tower location information, we create data to train the model as follows. For each of these measurement scenarios (cells), we create tuples with features computed as in the previous subsection and the algorithm among the five in Fig. 5.4 that yields the least localization error as the class label.

Another key question to realize the classification model is the selection of a classification technique that yields the most accurate classification. We empirically address this question and compare the accuracy with seven well-known and commonly used classification techniques: K Nearest Neighbors (KNN), Naive Bayes (NB), Decision Tree (DT), Multinomial Regression (MNReg), Neural Networks (NNET), Support

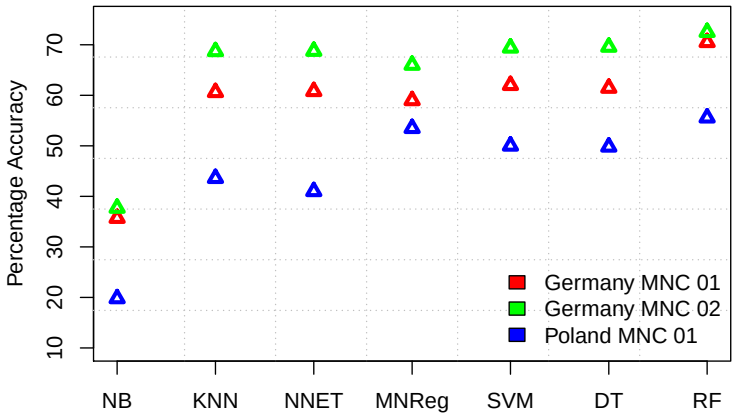


Figure 5.6: Comparison of accuracy between various classification techniques that can be used to generate AAS model.

Vector Machine (SVM) and Random Forest (RF). Results shown in Fig. 5.6 indicate RF to be the best technique with accuracy ranging between 56 and 73% for different datasets, so we use RF as the classification technique in AAS. From deeper examination, we find that there are two main reasons for the somewhat low level of classifier accuracy (in comparison with accuracies over 90%): (1) imbalance between the different classes, indicated earlier by Fig. 5.4; and (2) confusion between subsets of algorithms (classes) having similar localization inaccuracy within few tens of meters of each other. Even so, as we will see later in this section, the AAS performs significantly better than the state of the art and the simple-minded PAS.

Significant Features. As for the significant features, AAS model is highly impacted by features showing mutual distance gap between the estimated cell tower locations of different localization algorithms. Table 5.3 shows the importance of the top impacting features in the form of Mean Decrease in Accuracy (MDA). The more the accuracy of a random forest decreases due to the exclusion (or permutation) of a particular feature, more important that feature is deemed, and therefore features with a large MDA are more important for the purpose of classification.

5.4.2.3 AAS Evaluation

We now evaluate localization performance obtained with AAS in comparison with Oracle and the other alternatives discussed before. Our evaluations are based on two

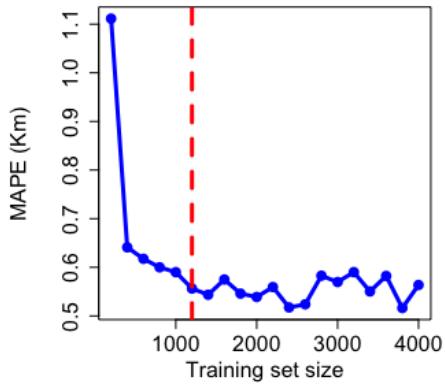


Figure 5.7: A training set size of around 1200 is sufficient for AAS to deliver low localization error.

methods: (1) 10-fold cross validation (CV); and (2) using a training set of around 1200 scenarios and test set of 200 scenarios. The selection of training and test sets is random with results an average of ten runs. We choose training set size to be 1200 scenarios as the learning curve given in Fig. 5.7 suggests this training set size is sufficient to train an AAS model. We see that the results for 10-fold CV and training size of 1200 is similar indicating around 1500 of training set size appropriate to generate an AAS model.

We present the results from 10-Fold CV in Fig. 5.8 (a). Because of its more reliable choice of a localization algorithm, AAS reduces the median localization error by 42.4%, 28% and 25.7% respectively for the three datasets, compared to PAS. For the same reason, the median localization accuracy with AAS is within 20% of the Oracle performance in all three crowdsourced datasets.

5.4.2.4 AAS Applicability to WLAN AP Localization

Given that many of the algorithms employed for cell tower localization have originally been proposed for localizing the Wi-Fi access points (APs) (e.g., [76, 78, 82]), it is natural to wonder if an approach like AAS that is seen to be effective for cell tower localization is also effective for the AP localization setting. To address this question, we use the Dartmouth WLAN dataset and compare the different schemes.

Results in Fig. 5.9 (a) indicate similar relative performance as before. However, different from the cell tower localization setting, the ratio of improvement in this setting is lower for AAS. We find that the dataset used is the key reason behind these obser-

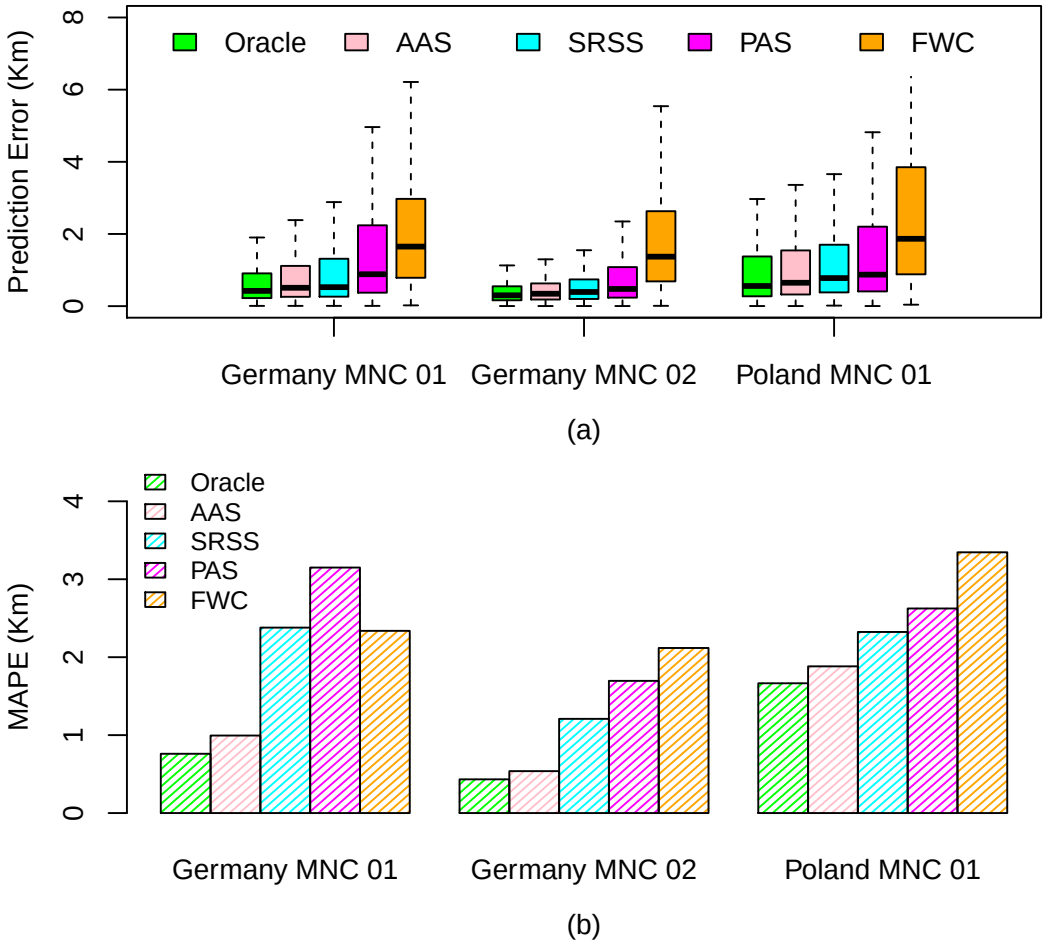


Figure 5.8: (a) Distribution of localization errors and (b) mean absolute prediction error with different schemes using OpenCelliD datasets.

vations. Dartmouth dataset is not composed of crowdsourced measurements; instead it is collected via war-driving and war-walking restricting to roads. Moreover, this dataset is relatively smaller in size compared to previously used OpenCelliD dataset, both in terms of the number of measurements and scenarios (280 APs vs. 2000-4200 cell sites). Nevertheless, these results do clearly demonstrate the robustness of the `AAS` approach across different wireless infrastructure localization settings.

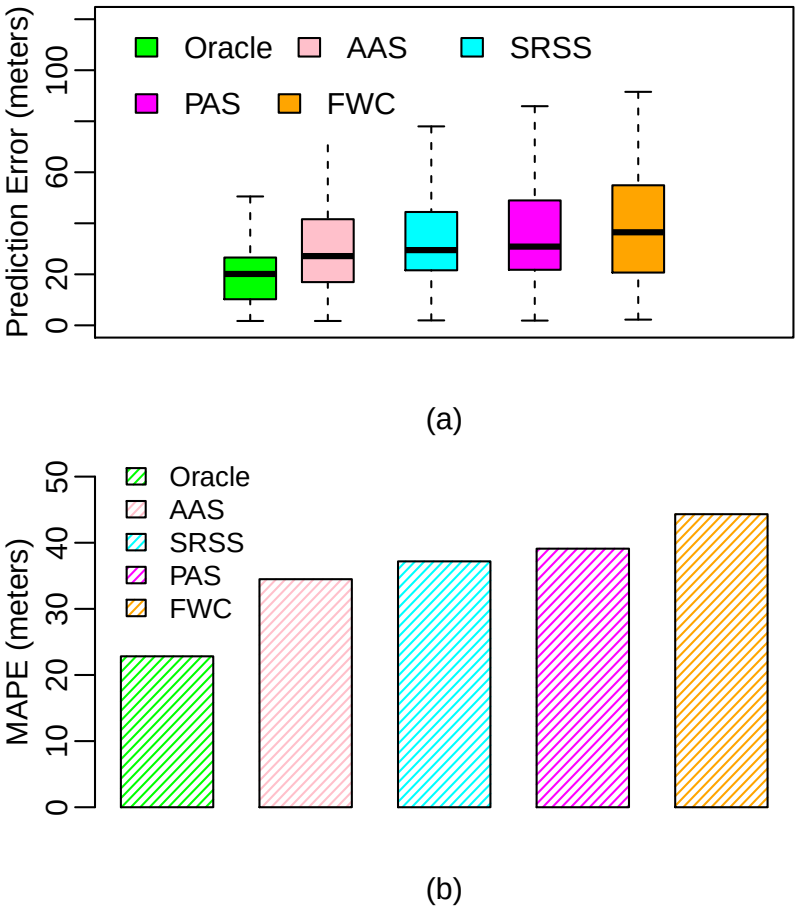


Figure 5.9: (a) Distribution of localization errors and (a) MAPE with AAS in comparison to other schemes using WLAN dataset.

5.5 Applications

As indicated at the outset, knowing the locations of cell towers can enable a variety of applications. In this section, we provide results leveraging AAS for a number of such applications.

5.5.1 Applicability of AAS in New and Diverse Settings

As stated at the outset, a key motivation behind our study into localizing cell towers with measurements is to have a means to gain insight into the reach of mobile infrastructure in developing country settings. To this end, we initially validate AAS model

in new settings before demonstrating its applicability in developing country settings under the typical and realistic assumption that ground-truth cell tower location information is unavailable.

AAS Validation in New Settings with Ground Truth Here we first examine if `AAS` model trained with data from a particular operator and country can be used in a different setting (different country and operator) assuming ground truth cell tower location information is available for the latter.

This investigation is aimed at testing the applicability of the AAS model in new and previously unseen settings. For this purpose, we use Germany's MNC 01 dataset to train our AAS model and test it over Germany's MNC 02 and Poland's MNC 01 datasets; we refer to this variant of AAS as `AAS (Diff.)`. For comparison, we also include `AAS` variant which is trained and tested on different parts of the same dataset (e.g., Germany's MNC 02) as `AAS (Same)` along with Oracle and FWC.

Table 5.4 and Fig. 5.10 show the results. The focus is on the difference in error performance between the two AAS variants (`AAS (Same)` and `AAS (Diff.)`, former indicating the best case result achievable with AAS in a new setting. Results indicate that the difference between these variants is marginal with both test datasets and close to the Oracle performance, and that `AAS (Diff.)` is significantly better than FWC or SRSS.

Moreover one should look at the distribution of features' values that provide guidance about the extent an `AAS` model trained on different dataset is trustable. For example for the case of Poland MNC 01, we observe its measurement scenarios' radii, mean dispersion of samples from center, autocorrelation and distances of algorithms' results (from each other, to trend line, and to center) are comparatively smaller while minimum and median RSS values are higher. All of these attributes are significant features effecting decision of `AAS` model where some of these distributions are shown in Fig. 5.11. In other words, smaller the difference in distribution of features' values, from training dataset, higher is the overall localization accuracy achieved by `AAS (Diff.)`.

Due to difference of the distribution we initially obtained a difference of 6% in the median error between `AAS (Diff.)` and `AAS (Same)` for the Poland's dataset. By

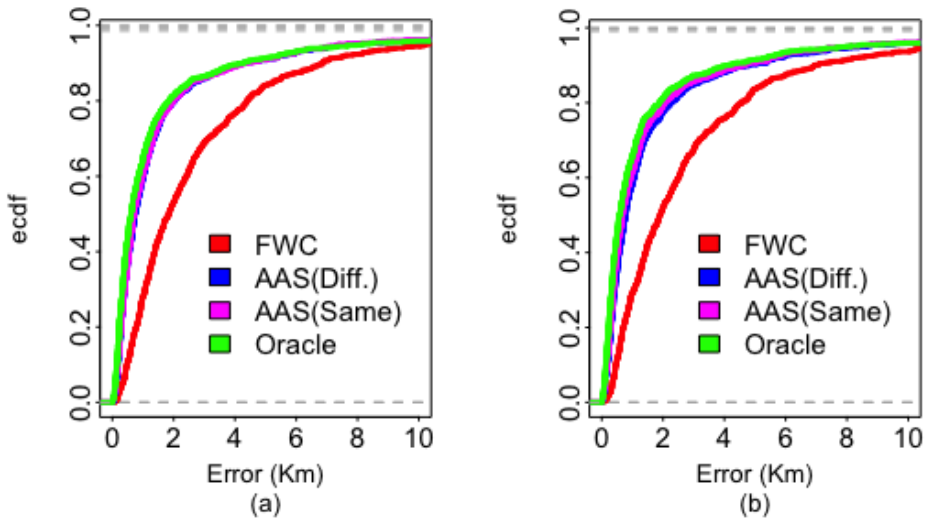


Figure 5.10: Localization errors with `AAS (Diff.)` that is trained on Germany MNC 01 dataset and tested on (a) Germany MNC 02 and (b) Poland MNC 01 datasets, relative to other schemes.

dropping some of the drifting variables [90] that cause highest covariate shift and using instances from training set similar in distribution to that of test set, the error dropped by 2%. Features with covariate-shift are the ones having very different distribution for both the test and the train sets. While dropping these features care should be taken not to remove the highly significant ones. Secondly while generating the model, accuracy can be improved by either assigning higher weight or retaining the instances from training set that are similar to those in the test dataset.

AAS Application to Developing Country Settings Now we come to our key motivating use case of estimating cell tower locations in new settings where there is no ground-truth information available, keeping developing countries in mind. Results from the preceding subsection suggest that a pre-trained `AAS` model when used in an entirely different setting still gives location estimations within around 20% of the Oracle approach, which makes it reasonably trustworthy and that too by a big margin compared to the alternative schemes from the literature.

To demonstrate the usefulness of `AAS` in inferring mobile infrastructure in developing countries via measurements, we consider three representative countries from Africa – Zambia, South Africa and Morocco – as case studies. We selected these countries

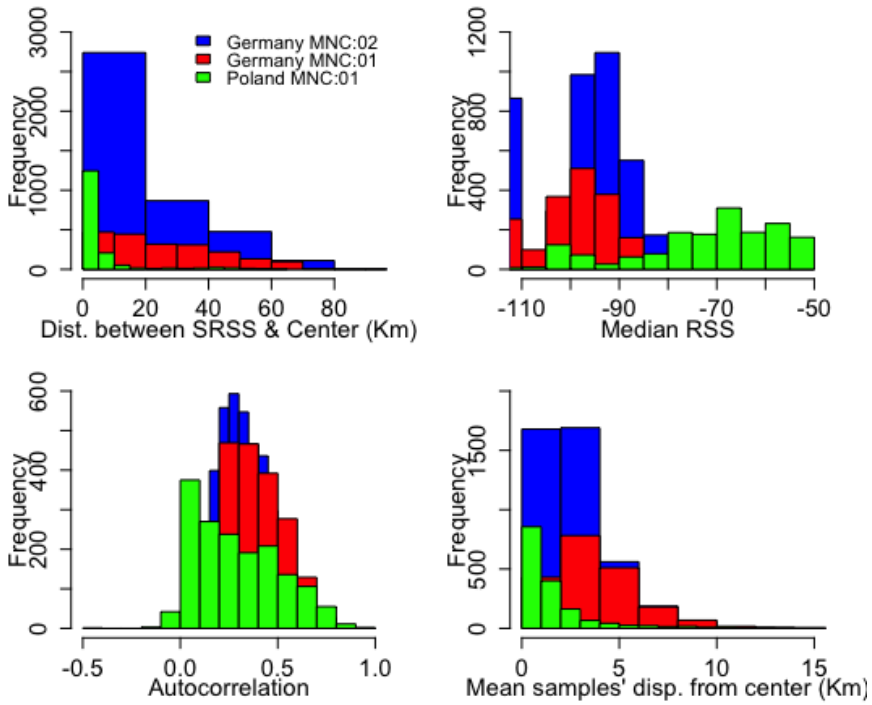


Figure 5.11: Differences in the distributions of select features used by AAS across different datasets that can impact accuracy of `AAS (Diff.)`.

keeping in mind availability of crowdsourced measurements in the OpenCelliD dataset and side information in the form of some publicly available coverage maps to visually inspect and assess the correctness of cell tower location estimations made by `AAS`.

Considering our first case study of Zambia, we focus on cell tower infrastructure for Airtel (MNC 01), which is one of the three largest operators in the country but does not even provide coverage map on its official website [91] let alone revealing its infrastructure siting information. For crowdsourced measurements for this operator, we rely on OpenCelliD's sub-dataset for Zambia. To apply `AAS` in this new setting, we train it on Poland MNC 01 dataset (in view of its similarity in distribution of features to that seen in Zambia which is apparent from Fig. 5.12). Resulting cell tower location estimations indicate the probable infrastructure layout of this operator (Fig. 5.13) (a), which shows good alignment with the coverage map information available for this operator from OpenSignal, shown in Fig. 5.13 (b).

We repeated a similar process of estimating cell tower locations for CellC (MNC

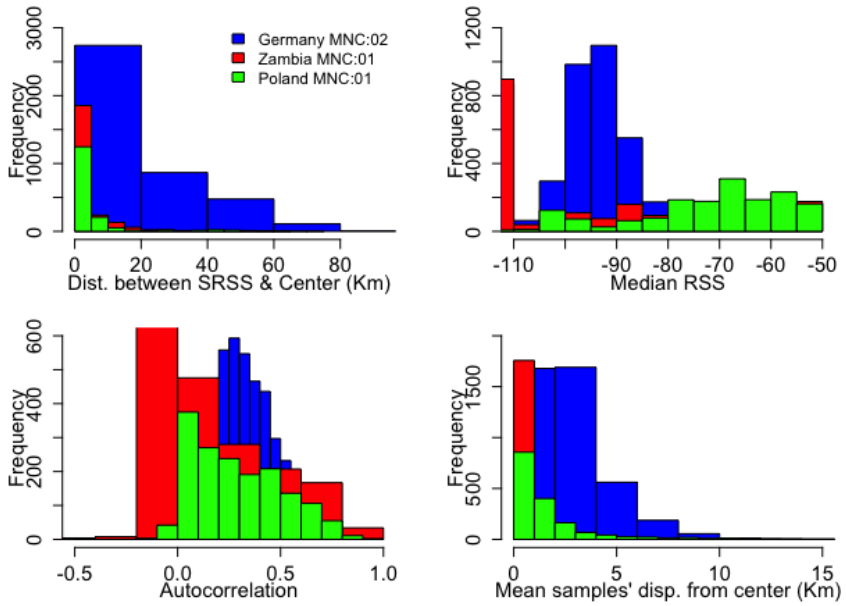
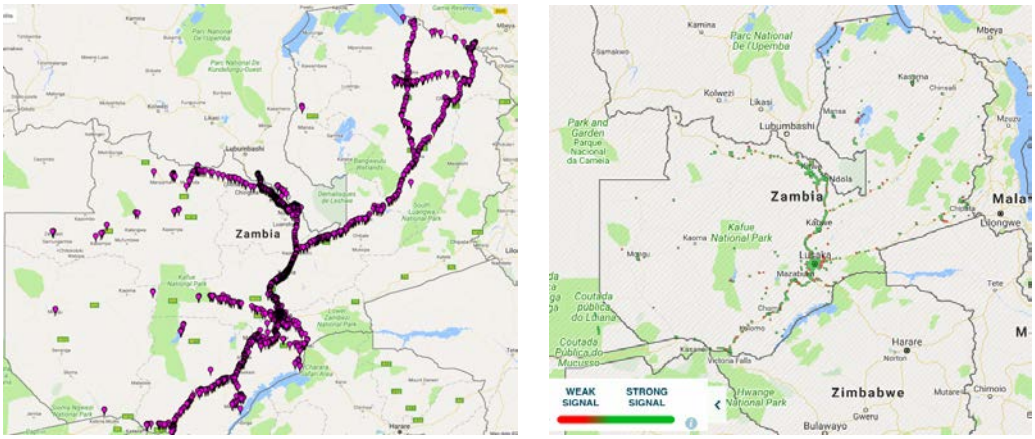


Figure 5.12: Differences in distributions of AAS model features between Zambia and potential training datasets (with ground-truth cell tower location information).

07) 2G mobile network in South Africa and IAM (MNC 01) network in Morocco. For both cases, we trained the AAS model on Germany MNC 01 dataset in view of its feature similarity to the above test networks, like above. Resulting map with inferred cell tower locations for both these networks in different countries along with corresponding but independent coverage maps from public sources adding confidence to these inferences are shown in Figs. 5.14 and 5.15, respectively. These case studies clearly demonstrate the value of AAS approach for robust measurement based cell tower localization to map/track mobile infrastructure in developing countries.

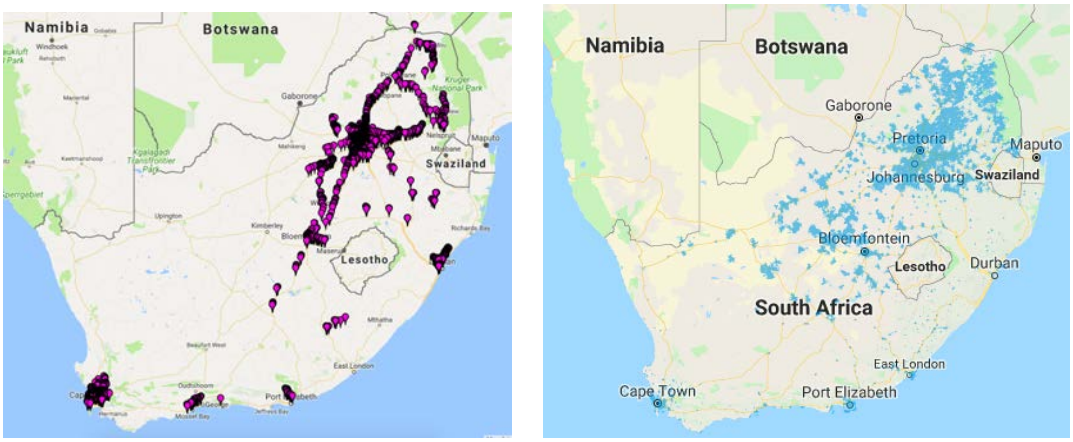
5.5.2 Cell Footprint Estimation

Localizaing infrastructure using AAS indirectly assists in estimation of cell footprint, that has several benefits. For example, it helps in determining likely cell associations along the trajectory taken by a mobile user which in turn can enable device-centric intelligent mobility management as in [92]. It can also aid in generating more reliable mobile coverage maps as demonstrated in the next subsection. Footprints of cells have been estimated by different studies in following three ways:



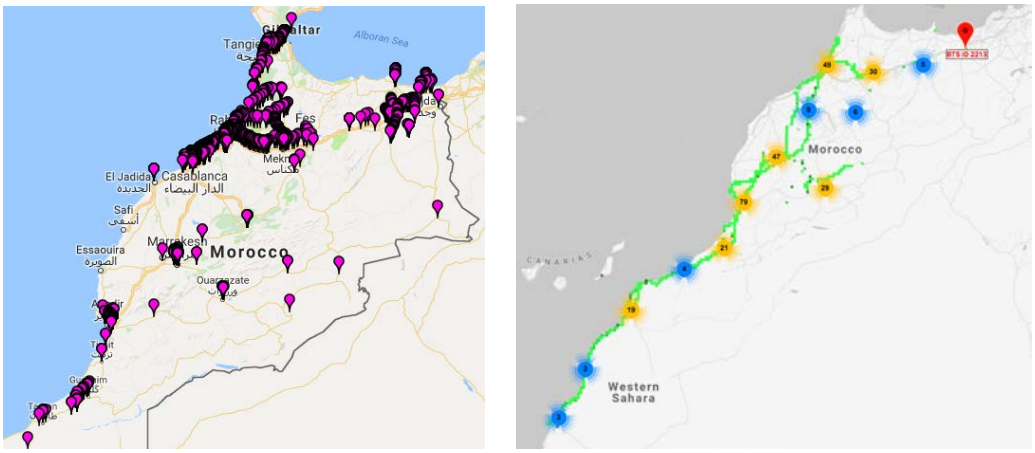
(a) (b)

Figure 5.13: (a) `AAS` model trained on OpenCelliD Poland MNC 01 dataset and tested over measurement data for Airtel MNC 01 in Zambia from OpenCelliD; (b) Publicly available coverage status for Airtel, Zambia from OpenSignal.



(a) (b)

Figure 5.14: (a) Infrastructure layout of South Africa's CellC 2G network as identified by `AAS (Diff.)` with measurement samples obtained from OpenCelliD and (b) Coverage map of South Africa's CellC 2G network from its official website.



(a) (b)

Figure 5.15: (a) `AAS (Diff.)` identifies infrastructure layout of GSM cells of Morocco's IAM network using measurement dataset from OpenCelliD and (b) Infrastructure layout of the same network as shown by cellmapper.com.

- Bounding geographic area, where signal strength from a cell tower is heard, as coverage footprint of the cell tower. This process may lead to large overlapping parts for neighboring cell towers.
- Applying Voronoi tessellation on cell tower locations [93]. The method ends up in hard triangular boundaries for cell footprints.
- Using information from operator such as sector location, orientation, beam width and the intended spatial extent of the installation (i.e., macro, micro, or femto cell) to estimate coverage area of a cell [94].

For third party entities, the cell tower related information (i.e. location, beam width, orientation and power) is hard to get; one is then left with crowd-based measurements and corresponding estimated cell tower locations. Using only cell site information for footprint analysis (via Voronoi tessellation) give hard boundaries with rather low accuracy. For example Fig. 5.16 (a) show cells from a LAC of Germany's dataset with Fig. 5.16 (b) presenting footprints applying Voronoi tessellation on `AAS` estimated cell locations. To assess accuracy of the cell footprints derived with Voronoi tessellation, we apply the process on a number of LACs and verify estimated cell as-

sociation of measurement samples in these LACs with their ground truth cell association; Fig. 5.17 indicates accuracy for Voronoi Tessellation to be 50% for almost half the cases. We instead use both the measurements and estimated cell sites in footprint boundary demarcation; and find Random Forest (RF) followed by Decision Tree (Fig. 5.17) to achieve best cell association accuracy for measurement locations. This time we observe in Fig. 5.16 (c) the detected boundaries of RF to be soft and more realistic.

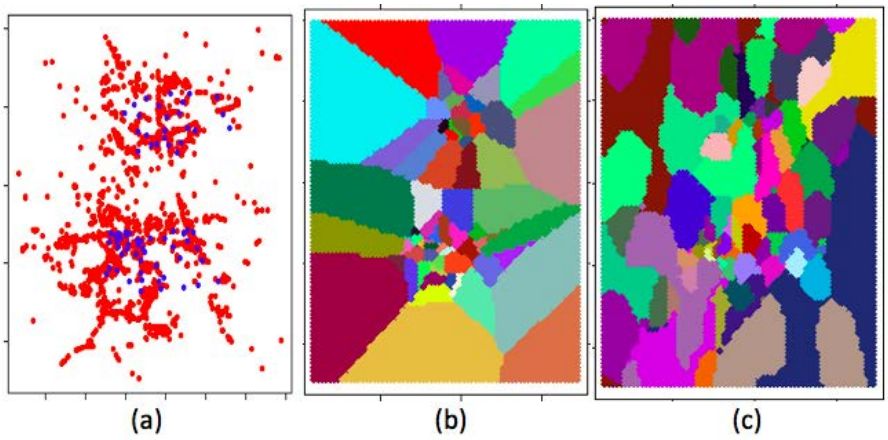


Figure 5.16: (a) Measurement samples (red) with ground truth cell towers (blue) of a LAC from Germany (b) Cell footprints with Voronoi Tessellation (c) and Random Forest.

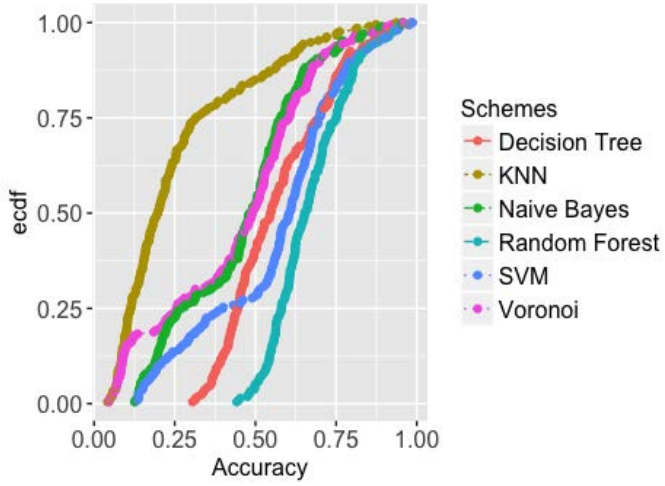


Figure 5.17: Cell footprint accuracy with different classifiers.

Once we have the boundary and shape of a cell's footprint, its estimated cell tower

location and measurement readings; we can attain the signal propagation footprint of the cell within an area using path loss propagation formula. Error between coverage footprint generated with actual cell tower location and with AAS location (shown in Fig. 5.18) is 0.12 RSSI MAPE and for FWC it is 4.15 RSSI MAPE. The maps, constructed this way, give a rough idea of expected signal quality from the cell tower at different locations of the cell's footprint.

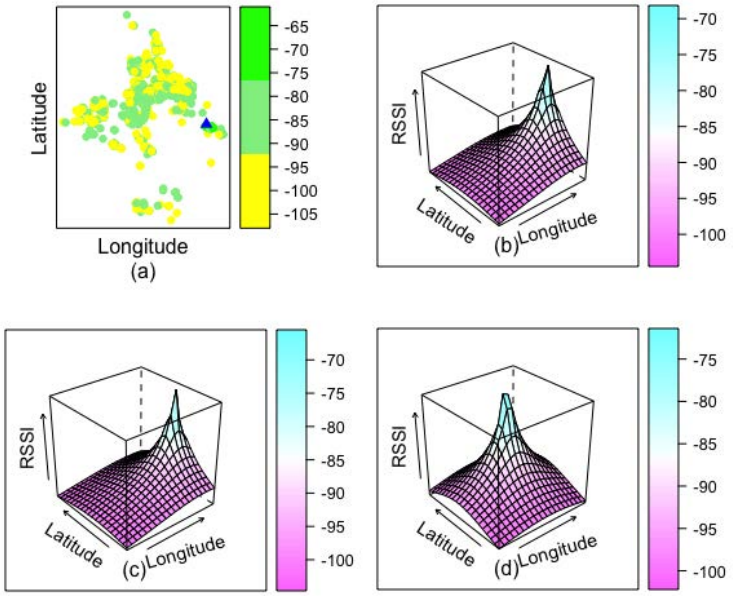


Figure 5.18: (a) Measurement samples (dots) with ground-truth cell tower location, Signal strength map across the terrain using (b) ground truth (c) AAS and (d) FWC based cell tower location.

5.5.3 Improving Coverage Map Accuracy

Mobile coverage maps are often relied on in practice by consumers in choosing a mobile access service provider, for example based on the coverage in their residential neighborhood. Equally, these are used by regulators and operators in identifying poorly served areas and coverage holes.

In an earlier chapter we observed OK as the most approriate interpolation process for coverage map generatation. OK, however does not differentiate between measurement samples and uses all of them together, overlooking the fact that at any given area

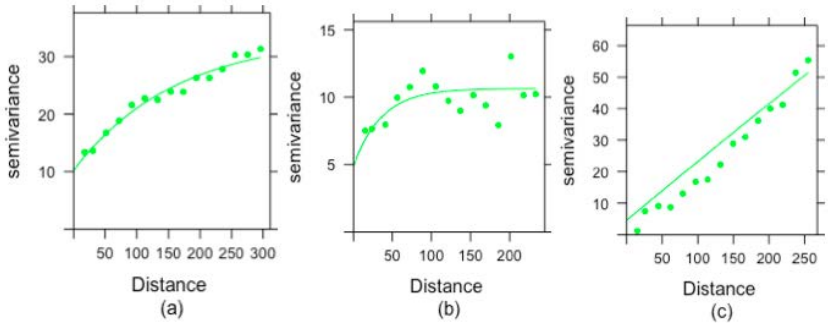


Figure 5.19: (a) Combined variogram of cells (b) Variogram of first cell (c) Variogram of second cell.

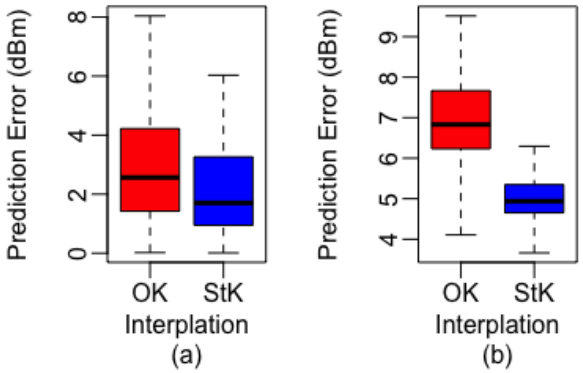


Figure 5.20: RSS prediction error with OK and StK (a) Meadows dataset (b) A LAC from Germany OpenCelliD dataset.

signal strength values from different accessible cell towers might be different from one another whereas from a same source they are usually similar. In situations where variograms [95] and signal strength ranges from measurement subsets belonging to different cells are dissimilar as is shown in Fig. 5.19 for Meadows data used in chapter 3, the interpolation with combined measurements from both serving and non-serving cell towers reduces the reliability of the predicted coverage map. The situation especially worsens when only measurements from non-serving cells are used for interpolation.

We therefore hypothesize that the adverse impact on coverage map accuracy can be reduced if interpolation process is guided by cell association of locations in question. To verify this hypothesis, we take two different datasets i.e. Meadows dataset that consist samples from twelve cells and a random selection of cells from a LAC of

Germany's OpenCelliD dataset. We compare two approaches to coverage map generation: (1) using OK based spatial interpolation as done commonly; (2) use Stratified Kriging (StK) in which available measurements are stratified (segregated from one another) based on the cell they correspond to. With StK, differently from OK, the signal strength at an unobserved location is obtained based on the likely serving cell at that location. We rely on footprint approach presented in section 5.5.2 to estimate cell association of measurement locations, for StK. Fig. 5.20 shows that for coverage map prediction, StK reduces the median prediction error of the interpolation process by 33.5% and 42.3% for the two datasets respectively; showing the value of estimating cell tower locations.

5.5.4 Cell Density Analysis

Estimation of cell tower locations naturally makes it straightforward to assess the density of cellular infrastructure and when done over time can provide insight about its growth, which can be useful in light of the importance of cell densification in scaling mobile network capacities as well as from the perspective of ubiquitous mobile service provisioning. In the spirit of [96] where cell tower information was used to estimate relative population distribution, we use AAS over crowdsourced measurements from OpenSignal [7] for a city to estimate the cell tower locations (Fig. 5.21 (a)) and also present population density across the city from the census data (Fig. 5.21 (b)) to aid in visual correlation. Along with crowdsourced-based estimated cell tower locations, Fig. 5.21 (a) provides the spatial distribution of medical stores and supermarkets in the same city with quite high spatial correlation of *0.95* and *0.96* to that of cell towers respectively. Clearly these correlations suggests the opportunities for informed network planning based on these other factors or urban planning influenced by cell infrastructure; as pointed out by [97], cell tower density can be used to quality assure the official population estimates as well as to indicate on the maximum expected population in an area.

5.5.5 Avoiding Health Hazards

Along with positive side, technology comes with a cost; and from cell towers perspective this is in the form of health hazards. People living within 50 to 300 meter

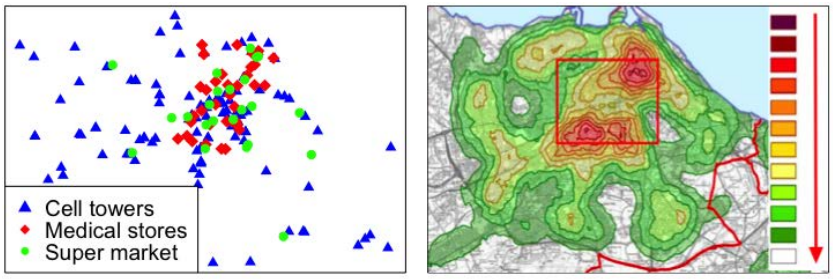


Figure 5.21: (a) Spatial correlation of cell towers with medical stores and supermarkets (b) Population density showing intensity similar to cell towers.

radius of a cell tower are in the high radiation zone and are more prone to ill-effects of electromagnetic radiation. E.g India adopts exposure limits of 9.2 W/m sq., for 1800MHz towers. A report [98] says that this power level has 'biological' effects such as cancer and genetic damage (where three cancer patients in a buidling were identified whose flats were directly exposed to three cell towers deployed on roof top of an opposite building). The study also reports that the radiation impacts environment in several ways such as decrease in dairy production, disappearance of bees, decrease in number of birds and so on. Thus along with guiding towards demographic distribution, cell infrastructure information can help people to reside at places away from dense clump of towers, to avoid its ill effects.

5.5.6 Device Localization

As noted at the outset, device localization has been a key driver behind cell tower location databases. While such location services employ mechanisms like trilateration using cell tower locations along with signal strength measurements from a mobile device, we present here a proximity based device localization [99] application that simply uses the serving cell's tower location as the device's estimated location. In Fig. 5.22, we show that there is no difference between proximity based device localization accuracy when true and estimated cell sites are used. This energy-efficient approach is especially useful when GPS is unavailable and cell density is higher, as demonstrated by the results in Fig. 5.22 (b) where small cell sizes result in better location accuracy for a device.

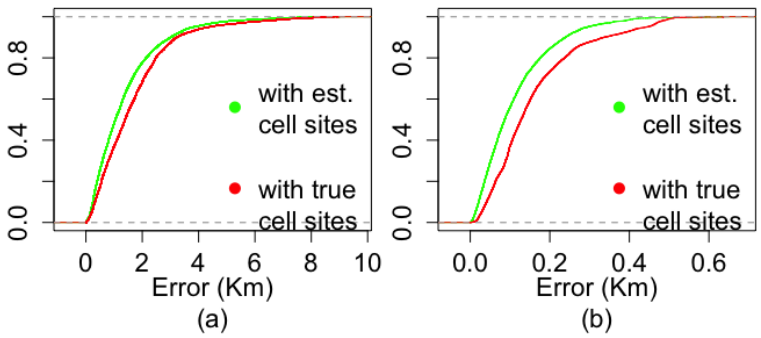


Figure 5.22: Ecdfs of proximity based device localization leveraging estimated cell tower locations in two cases: (a) OpenCelliD dataset over a wide area with large cell sizes; (b) using crowdsourced measurements in dense central part of a city.

5.6 Summary

User-side measurements can be used to anlayze a network from different angles. In earlier chapters the measurements, in the form of crowdsourced samples, proved their value in generating a network 's mobile coverage map. In this chapter, these measurements showed their potential in identifying layout of a cellular network's infrastructure.

In order to imrpove accuracy of cell tower localization, using a large-scale crowdsourced measurement dataset with ground-truth cell tower locations, we first showed that each of the commonly used localization algorithms is susceptible given the wide variations in features across different measurement scenarios. Even the recent FWC approach [71] to avoid using less predictive measurements in conjunction with a specific algorithm is found to be similarly vulnerable. Motivated by these observations, we proposed AAS, a novel localization approach that aims to adaptively select a localization algorithm. AAS is expected to yield most accurate achievable localization performance as its choice of a localization algorithm is dependent upon the characteristics of a measurement scenario. Due to its adaptive feature, AAS significantly outperforms both the commonly used strategy of fixing an algorithm and FWC that uses predictive set of measurement samples.

AAS is robust across new and different settings. This is demonstrated both by use of AAS in WLAN AP scenario and by cell tower localization in three different African

countries so to infer mobile infrastructure in developing countries.

Finally, we showed value of AAS powered cell tower localization in some other concrete use-cases. These include cell density and footprint analysis, improving coverage map reliability, end-user localization and guiding people on electromagnetic radiation sources.

Table 5.1: Summary of datasets.

Dataset	**Description**	**# Cells**	**#Samples**
OpenCelliD	MCC 262 (Germany), MNC 01	2002	1,579,120
	MCC 262 (Germany), MNC 02	4200	3,378,117
	MCC 260 (Poland), MNC 01	2000	1,164,741
	MCC 645 (Zambia), MNC 01	2000	2,453,827
	MCC 655 (South Africa) MNC 07	5379	2,766,26
	MCC 604 (Morocco) MNC 01	1676	80,213
RF Signal Tracker dataset	Collected with Samsung Galaxy S3 during 18-20 April 2017 at downtown area of a city.	68	6,000
Synthetic dataset	Generated using Okumara-Hata model[1] with range 35Km, carrier frequency 1700 MHz, antenna of height 200m and end-device at 3m depicting a cell in small and medium-size cities.	1	maximum 32,000
Dartmouth WLAN dataset [84]	Collected at Dartmouth campus using Place Lab software during 12–14 Sept 2005. We used warwalk dataset with at least 20 samples per AP.	280 APs	31,312

Table 5.2: Effect of crowdsourced measurement characteristics on prediction error (in meters).

Characteristic	**Geometric**		**RSS**		**PLP**
Schemes	C	WC	SRSS	GL	MCPL
Location inaccuracy: RF Signal Tracker dataset					
GPS locations	261	263	238	238	227
Network-based	265	266	264	277	291
% degradation	1.5	1.1	10.8	16.3	28
Measurements layout: Synthetic dataset					
Well spread	352	356	754	707	388
Skewed	22K	22K	754	1K	1K
Out-of-boundary cell	16K	16K	5K	5K	5K
High RSS samples separated	352	360	18K	707	668
Cor(RSS, distance to cell tower): Synthetic dataset					
Poor correlation	352	434	2K	2K	1K
% degradation	0	21	144	182	222
Cor(Response rate, distance to cell tower): Synthetic dataset					
High correlation	165	158	754	707	401
% improvement	53	55	0	0	–

Table 5.3: Features for AAS model in decreasing order of importance with RF as the classification technique and as per MDA.

Features	MDA
Distance between estimated locations of the algorithms	30
Maximum RSS value	24
Distance between estimated locations of algorithms and trend line	23.6
Mean RSS value	23.3
Autocorrelation of samples' locations and their RSS values	22
Mean dispersion of samples from central location	21.7
RSS standard deviation	20.9
Standard deviation of RSS weighted dispersion of samples	20.6
Mean distance of samples to trend line	20.2
Size of measurement samples	20
Median of RSS weighted dispersion of samples from central location	19
Correlation of samples' distances and RSS values to Geometric locations	18.5
Median of dispersion of samples from central location	18

Table 5.4: Key localization error statistics with AAS (Diff.) that is trained on one dataset (Germany MNC 01) and tested on different datasets, relative to Oracle and other alternative schemes.

Test Set	**Germany MNC 02**		**Poland MNC 01**	
Scheme	*Median APE*	*Mean APE*	*Median APE*	*Mean APE*
Oracle	309m	431m	547m	1.62km
AAS(Same)	11.6%	18.5%	18%	17%
AAS(Diff.)	12.6%	27%	22%	19%
SRSS	14%	148%	32%	38%
FWC	335%	386%	241%	136%

Chapter 6

Web Browsing QoE Analysis via Distributed Testbed Measurements

With a complex ecosystem of networks, smart devices and traffic-intense applications, Mobile broadband (MBB) network brings certain challenges to the operators and service providers. For example, the higher the load in a network cell, the more variance users perceive in the time to view the content of interest (e.g., webpage load time), which in turn, translates to poor quality of experience (QoE) [100].

In this complex ecosystem, finding influence on user's experience from network features, experiment context and and application content offers the promise of helping operators and application developers to make improvements that matter to customers. Using distributed measurement platforms, this study contributes in demonstrating how features from different domains, retrievable by measurements conducted at end-nodes, can assist in accurate objective assessment of web QoE by generating a PLT model. Further we present intensity of influence from non-content features. These relationships not only informs about what to expect under a network and experimental context but also where the performance gaps lies to be filled by taking corrective actions. To validate the conclusion drawn from experimenting with commercial networks, we tune some of the identified significant network-side parameters in an experimental testbed to observe its level of impact on PLT. This additional investigation gives a better understanding of how much a network configuration improves or degrades web-browsing

QoE.

We performed this study on MONROE [10, 16] and FLEX-MONROE platforms under FLEX-MONROE project [17]. The MONROE platform is designed to advance ones understanding of today's operational MBB ecosystem from the end-user's perspective. It enables performance measurements across attached commercial carrier networks. The FLEX [101] testbed, on the other hand, allows configurational changes in its experimental LTE network thus enabling experimenters to identify how variation of different LTE network parameters vary the performance of services and applications.

This chapter consists of nine sections. Section 6.1 illustrated methods used or proposed by research community in assessing QoE of an MBB service. Section 6.2 describes the motivation of the experimental study showing importance of good web QoE. Section 6.3 mentions the performance metrics that are considered to be quantitative measures of a user's web-browsing experience. The measurement platforms used and the web-browsing experiment conducted are illustrated in section 6.4 and 6.5, respectively. Section 6.6 is about modelling web performance that ultimately gives us the significant features from different domains, with section 6.7 highlighting intensity and trend of change in web QoE as different non-content features vary. Before summarizing the intuition achieved from this experimental study, section 6.8 describes role of the experimental test-bed in validating impact of a network configuration on web QoE.

6.1 Background and Related Work

6.1.1 Deriving QoE from QoS

This section mentions the work related to QoE assessment, where a user makes good or bad judgement about a network from what they experience at an application layer. QoE is dependent upon features from a number of domains including quality of services (QoS), application content, context in which the application service is running, and the user's expectation. The most important domain of these is considered to be QoS, in following subsections we therefore state the assessment methods generally used for determining Quality of Services (QoS) and QoE separately and the related work performed in identifying linkage between different QoS metrics and QoE with

an application service.

- **QoS:** QoS monitoring deals with measuring various performance factors of a network including throughput, bandwidth, round trip time (RTT), latency, bit error rate (BER) and packet loss rate to name a few. These measurements are either collected actively or passively.

- **QoE:** It deals with measuring user-centric metrics which determines level of satisfaction that a user observes while downloading a web-page, making a phone call, browsing a search engine, interacting a multimedia application or watching a streaming video. QoE measurement schemes usually operate on application layer e.g. counting stalls and examining blockiness in a video, engagement time or abort ratio of video clips, voice quality of an audio application, call success and drop rate, response time of a web-page and amount of energy and data consumption by an application etc.

To determine the effect of the underlying context on the above application level features QoE measurement has to grasp knowledge of device features (i.e. resolution, interface and power consumption), user factors (i.e. expectations, security, requirements and interests), environment (i.e. mobile/stationary, indoor/outdoor) and most importantly network level QoS as it is a strong basis for the respective QoE [102]. Only assessment in this way can identify the domain which is the source of issue.

6.1.2 QoE Measurement Tools

Literature presents two ways for judging the satisfaction level of a user with a service or a product. These are termed as *Subjective* and *Objective* methods.

- **Subjective Method:** In subjective measurement a group of users watch videos, listen to audio clips, browses web-pages or engage in some service and provide their feedback in the form of ratings i.e. Mean Opinion Score (MOS) which is a five point score defined by ITU-T Recommendation P. 800 [103]. The test is conducted in a controlled environment e.g. in a lab with certain requirements like room lighting, distance from the screen, and the length of viewing time so

that each user is tested in similar scenario. This method detects information such as loudness, echo, delay in web-pages download and distortion in an audio or video clip. The user then presents the overall perception in single score chosen from the set of {"Excellent", "Good", "Fair", "Poor", "Bad"}. Users however are usually influenced by the most recent experience and may be unable to report the opposite initial experience about a service. Also the method is time-consuming, expensive and not repeatable [104], moreover different people may have different interpretations of the terms e.g. "Poor" and may give different ratings even though they have observed similar experience in a test [105], making correct evaluation of results difficult.

- **Objective Method:** To avoid the expenses of subjective assessment and to produce repeatable and comparable results, objective assessment method emerged as an attractive approach. Instead of perception it depends upon mathematical retrieval of quality and then mapping of it into the rating format (if necessary). Examples of this scheme might include PSNR in an audio app, missing information in a video due to a compressor, buffering ratio of a video, data upload and download time etc.

  QoE Doctor [106], proposes an objective QoE measuring method that targets in finding user percieved latency in web-page download and YouTube video startup phase and amount of energy and data consumed by the different versions and tasks in Facebook. Here user QoE-related behaviour is replayed via a UI automation.

  The objective method though captures individual characteristics of the played service, it does not take into account overall perception including time of the day, user's expectations, features of the surrounding environment like noise, light intensity and disturbances.

6.1.3 Relationship between QoE metrics and Network factors

A big share in QoE is that of QoS metrics i.e. throughput, RTT, BER, packet loss and latency etc. It must be noted, however, that sometimes optimizing a feature such as data rate does not mean that achieved QoE is becoming better [107], also each

network feature has different effect on different services e.g. QoE of VoIP is influenced a lot by bit-rate, video streaming QoE is effected by initial delay [102] while web QoE is affected by SNR, inter-radio access technology (IRAT) handovers and load on the network [107]. Keeping in view the possible relationships between network features and corresponding QoEs, many research studies are conducted to estimate QoE indirectly from network factors that are tracked passively while a service is under utilization [108, 109]. The purpose of indirect evaluation can be multi-fold:

- when direct access to the user's application is not available,
- when monitoring the changes on GUI of a service is complex,
- when diagnosis down the network layers is desired,
- and to determine the network features that influence a particular service.

Literature study shows that the processes being adopted for retrieving the features are by tracking network flow using the command Tcpdump, observing the traffic flow at the passive probes installed within the network [108] and using QxDM (Qualcom eXtensible Diagnostic Model) [106] to record state transitions at Radio Control Layer (RCL) and packet data units at Radio Link Layer (RLL).

Either effect of each separate feature on QoE is observed such as the influence of throttled bandwidth on the buffering ratio or on the initial loading time of a YouTube video [106], or a combined influence of a bunch of different network metrics is computed [108]. In the later case the view is that most of the network features are inter-dependent [110], the complexity of their relationships is hard to determine by human investigator therefore instead of separate evaluation all of the retrieved features along with corresponding QoE detected at application layer is used to train some machine learning algorithm. The machine learning algorithm assigns weights to different features, and once trained it uses only the traffic flow features to estimate the corresponding QoE [108–111] e.g. V. Mekovski et al. [109] proposed creation of QoE predictor that estimates QoE on the basis of QoS metrics and improves the predictor over time as more and more user feed back is available.

6.2 Motivation

Our motivation for this study is to investigate how well a distributed measurement platform enables us in understanding the relationship of different impacting domains on web borwsing QoE. Further more to analyse if one can translate the resulting performance features from different domains into a user's experience with a web-page.

We took a case study of web browsing as web data has a large share in the Internet traffic, and more and more businesses are relying upon websites for revenue generation. Slow speed is considered a killer for revenues e.g. Amazon's calculated that a page load slowdown of just one second could cost it $1.6 billion in sales each year. Google has calculated that by slowing its search results by just four tenths of a second they can loose 8 million searches per day – meaning they'd serve up many millions fewer online adverts [112]. With MBB networks, mobile web surfing is surpassing day by day the desktop web traffic volume. According to a statistics portal [113], in last quarter of 2017, 51 percent of global web traffic originated from mobile devices.

As websites are available around the clock they are convenient for far away customers to buy products online. On the other hand, these websites provide businesses an easy means of marketing, credibility of their products and increase in sale. This importance of websites make developer and owners concerned about their speed as according to data from Akamai, the website that takes longer than three seconds to load, could loose half of the visitors.

In this context, the time a user spends to reach the web content is a critical metric the carriers aim to optimize [114, 115] mainly because of its strong correlation with user satisfaction and, subsequently, with company revenues [116–122]. The study therefore aims to identify the features from the application itself and the network QoS, in MBB network, that have significant impact upon user's web-browsing QoE.

6.3 Web-browsing QoE metric

Previous web performance measurements differ in the metric they chose to evaluate performance. Work in [123–128] used PLT, a metric primarily based on OnLoad event fired by the browser. This event is fired when all objects on the page are loaded. Google

introduced SpeedIndex [129, 130], as an alternative to PLT to better capture the user perceived experience. SpeedIndex is a measure of an average time to get all above the fold content (AFT) in the screen, in other words an average time for the visual completion of a page in the browser. More visual contents at the beginning of the page loading process lead to smaller SpeedIndex. However, measuring SpeedIndex requires to film the page loading process and is thus quite complex and can significantly inflate the measurement time.

Similar to SpeedIndex but much less computationally intensive metrics are ObjectIndex and ByteIndex, proposed in [131]. They are computed from the arrival time of all objects in the Web page waterfall. ObjectIndex tracks the time at which the content of the page is retrieved, taking into account all external images, style-sheets and scripts needed to render the page. ByteIndex operates in the same way, but weights objects by their size. A higher value indicates higher page load time. Additionally, [132] proposed the "3rd Party Trailing Ratio" metric, which measures the fraction of download time for the 3rd party assets on the webpage critical path. In our study, we however observed, that this metric shows little correlation with total amount of time all the objects in a page take to download. In our measurements, we took three different metrics, namely ByteIndex (BI), ObjectIndex (OI) and PLT. We found similar results for all three of them. Further more, Zuzana [133] shows results from 12 case studies indicating how slight increase and decrease in PLT impacts the revenue, conversion rates, search engine ranking and pages views etc. therefore in current study we focused on PLT.

6.4 Platforms Overview

To analyse impact of various factors upon web-browsing QoE, we make use of two platforms i.e. MONROE and FLEX-MONROE. In the following we give a brief description of each of these measurement platforms.

MONROE: The MONROE (Measuring Mobile Broadband Networks in Europe) system enables multihomed and large-scale experimentation on commercial cellular operators of four European countries (namely Norway, Sweden, Spain and Italy). It integrates a number of hardware devices (both mobile and stationary), a software frame-

work that enables the orchestration of experiments and the collection, analysis and visualization of results. Authenticated external users can access the platform, reserve resources and deploy their own or ready-to-use experiments [134] under a predefined quota. For the study we use MONROE, as it provides full control over its measurement nodes, allowing to systematically collect a rich and a better quality dataset over a long period of time along with the network metadata and context.

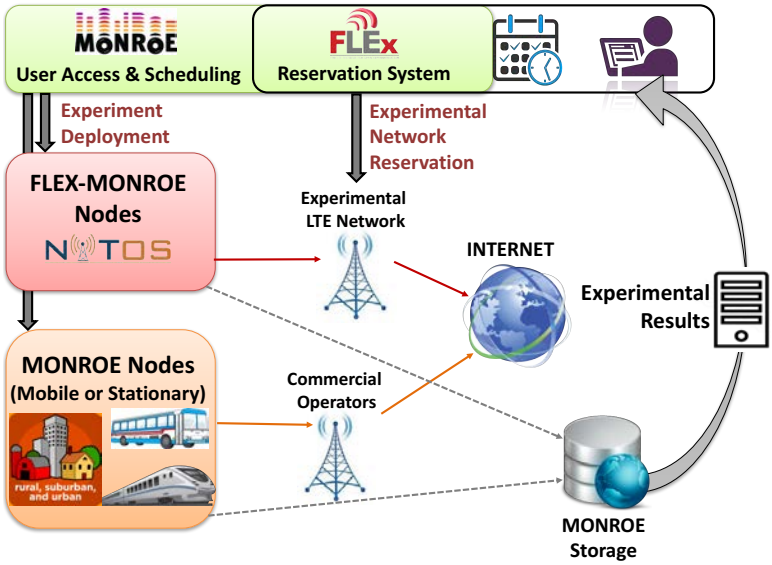


Figure 6.1: High Level Design of the FLEX-MONROE platform.

FLEX-MONROE Platform: It is an integration of FLEX (FIRE LTE Testbeds for Open Experimentation) and MONROE platforms where FLEX provides an open and remotely accessible platform for experimentation with LTE. In this study we used the NITOS [101] indoor testbed of FLEX which provides a fully controllable RF-isolated environment consisting of over 60 operational wireless nodes offering for experimentation on various technologies, including Wi-Fi and LTE. We aim to measure the experimental LTE network in FLEX in the same manner as any other commercial network in MONROE is measured. It is to verify what we observe and deduce from commercial network and additionally to help in finding better 4G network configuration so to enhance user's experience with web-browsing. We show in Figure 6.1 the high-level FLEX-MONROE system architecture.

The FLEX-MONROE node can be either a hardware native MONROE node running the MONROE software (hardware integration) or a hardware native FLEX node

Table 6.1: LTE parameters that the experimenter can modify in the NITOS experimental network.

Parameter	**Description**	**Range**
DL BW	Downlink bandwidth	5/10 MHz
UL BW	Uplink bandwidth	5/10 MHz
Power	Signal transmit power	-15 dBm to -26 dBm
Tx Mode	Enabled antennas	1/2
MCS DL	Downlink MCS profile	0-28
MCS UL	Uplink MCS profile	0-26
FQ Band	LTE band	7/13

running the MONROE node software (software integration). For this study we used hardware integration. For FLEX-MONROE integration, we therefore installed within the FLEX NITOS testbed a physical MONROE node equipped with custom FLEX SIM cards that are configured with the NITOS PLMNs. The resulting FLEX-MONROE node connects to the NITOS experimental LTE network as it would connect to any other commercial LTE network within the MONROE system. In order to interact with the platform, the FLEX-MONROE user needs to access both the standalone MONROE system, as well as the standalone FLEX NITOS system [135]. Then, a workflow with a number of steps be followed [136], to enable the synchronization of the reservations in both systems in order to deploy experiments on FLEX-MONROE.

To experiment in FLEX-MONROE, experimenter needs reservation in both FLEX and MONROE platforms. After connecting to the FLEX in NITOS testbed the experimenter can configure eNodeB parameters as are required. The list of configurable parameters are summarized in Table 6.1 with a short description and the range of possible values.

6.5 Web-browsing Experiments

This study began with an initial version of web-browsing experiment called `WebWorks` [137], for later experiments we used an enhanced version named as `ACME` [138]. Both the experiments are in the form of MONROE-compatible Docker containers that mimic mobile device browser to retrieve the mobile version of the top Alexa web-pages. The

containers mainly vary in browsers that they test for example `WebWorks` uses Selenium with Firefox and generates a detailed HTTP Archive Report (HAR) [139] for each visited page in a JSON-format, `ACME` on the other hand retrieves web-pages using both the Firefox version of 56.0.1 and Chrome version 64.0.3282.186 as browsers.

To understand impact of different factors upon web-browsing QoE we undertook two set of web-browsing measurement studies.

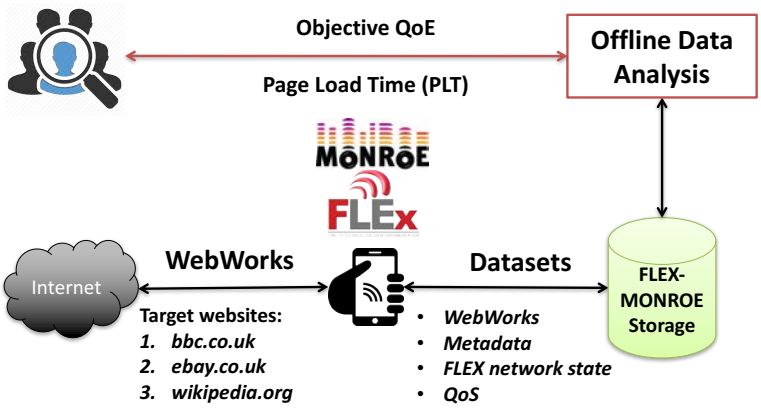


Figure 6.2: The analysis of web performance in FLEX-MONROE using MONROE `WebWorks`.

6.5.1 Measurements to model Web QoE

In the initial measurement campaign we deployed the `WebWorks` Docker container on 18 static MONROE nodes (each node measuring up to three different mobile carriers at the same time) from the four countries where MONROE coverage is available. MONROE nodes resolve the target websites using Google's public DNS resolver, not the mobile carrier's default resolver. `WebWorks` was configured to collect web performance measurements while visiting top 10 ranking Alexa [140] websites (with detail given in Table 6.2). All the target websites expect TLS with HTTP1.1 (referred as H1s from now on) connections by default, further they can work with HTTP/2.0 (referred as H2 from now on). This selection of websites also covers a wide range of user interests in terms of topics including social networking, video, career, search engine, news site, wiki or shopping. For each target website, we chose 10 different pages to visit in order to capture a wide range of resource sizes, resource counts and domains visited. For example, instead of measuring the landing page for facebook.com, we visit

Table 6.2: Characteristics of the target websites we select for our measurement campaign. These are average values over the 10 different pages we visit per website.

Site	Size (KB)	# Objects	# Domains
facebook	798	76	6
instagram	1,230	33	6
youtube	815	30	9
wikipedia	241	10	3
google	114	13	4
linkedin	232	24	5
yahoo	1,480	49	8
ebay	493	28	11
guardian	1,895	133	33
nytimes	3,131	205	55

specific target pages, such as 'facebook.com/telia/' or 'facebook.com/LeoMessi/'. We present statistics per target website in Table 6.2.

The experiment enables Firefox to cache the each page visit throughout the time it measures the same mobile operator. When moving on to a different operator connected on the same node, Firefox clears the caches before starting the measurements against the full set of targets. The measurement campaign was run in May 2017 (from 30th of April until 17th of May) and June 2017 (from 1st of June until 14th of June). In total, we monitor 11 mobile operators, which we list in Table 6.3 and collected more than 40,000 samples. The distribution of samples per operator is shown in Table 6.3.

Table 6.3: Statistics on the WebWorks dataset; the Country shows where the subscription is active.

Operator Name	Country	# measurements
Telia (SE)	Sweden	6,473
Telenor (SE)	Sweden	6,345
3 (SE)	Sweden	4,549
Telenor (NO)	Norway	6,350
Telia (NO)	Norway	2,806
ICE (NO)	Norway	2,664
TIM (IT)	Italy	4,392
Vodafone (IT)	Italy	1,961
Wind (IT)	Italy	1,883
Yoigo (ES)	Spain	3,183
Orange (ES)	Spain	727

The metrics collected in this initial campaign includes:

Web content: The HAR file that was generated during each WebWorks web-browsing experiment helped to derive a number of web-page related parameters and metrics. These include PLT, protocol, size of web pages in bytes, number of objects, size of each object (we also calculated metrics such as the standard deviation, the minimum and maximum of the object size), number of unique domains that are accessed by the web-page, object types (such as Javascript, CSS, image and HTML etc.), object load time including DNS resolution time, TCP connection time and object receive timings.

Network and user context: As a metadata, during each web-browsing experiment MONROE node also reported certain other features including frequency, frequency band, average RTT, minimum RTT, maximum RTT, RSRQ, throughput, signal strength variations and time of the day feature.

6.5.2 Measurements to observe Impact of Non-content parameters

For the second measurement campaign ACME was used as web-browsing experiment. This campaign was run on a number of nodes both under mobility (e.g., operat-

ing on-board public transport vehicles, such as trains or buses) and in stationary (e.g., in laboratories, or hosted by volunteers in homes) scenarios.

This campaign extended target websites to 18, with both H1s and H2 support. These websites include *wikipedia, coursera, facebook, youtube, twitter, reddit, flicker, imgur, instagram, kayak, live, microsoft, stackoverflow, thegaurdian, tmall, yelp, etsy* and *ebay*. The selection ensures covering of a wide range of user interests just like the first measurement campaign. The collected samples are almost 1.8 million with good representation from each of the protocols, browsers, node types and the eleven operators.

This second and comparatively recent study is undertaken during 1st April 2018 to 4th June 2018 so to better understand the individual impact of mobility, handovers, RAT technology, latency and signal quality on page rendering process. More specifically, the campaign collected the following parameters:

Experimental context: This includes the type of browser and the protocol used for one experiment run. Also, it includes the node type (stationary/mobile) and the distance the node travelled during an experiment.

Access network context: This includes parameters from the the Radio Access Technology (RAT) (more specifcally for 3G and 4G technology) such as radio status before the start of the experiment (Initial RAT, Initial RSRQ, Initial RSRP, Initial RSSI) and the radio changes throughout the experiment (median values for RSRQ, RSRP, RSSI parameters, the number of RAT handovers), and average RTT against the target webpage server (measured via ping).

6.6 Modeling Web QoE

Among the existing machine learning approaches, we select *multiple-linear regression* to find out how the PLT of a web-page varies with different parameters that characterize web-content, configuration of the LTE network, and the context in which the user surfs the web. We choose multiple-linear regression as the target parameter i.e. PLT is a continuous variable. Also the corresponding candidate predictors are either continuous or discrete variables, thus fulfilling the eligibility for linear regression. We

did not chose to run Polynomial regression as we have multiple candidate predictors to consider, for which case the regression would result in a complex and over-fitted model. Also, a previous study [141] shows that when the target web performance metric is measurable then linear regression modeling gives the best results.

Before applying the muti-linear regression, we applied normalization on the non-uniformly distributed parameters. To achieve reduced complexity of the regression model and isolate the independent variables that affect the PLT, we retain only the most significant of predictors (by re-generating the regression model and retaining features with low P value), and eliminate the multi-collinear variables (by identifying features that are correlated). For feature selection we use standard step-wise sub-scheme (as a Wrapper method) and subsequently apply the filter method to retain only most significant attributes by looking at their P values (i.e., we select predictors with P values less than 0.05). The P values indicates how confident one can be that an independent variable has some correlation with the dependent variable.

In Table 6.4, the "Predictor Variable" column lists the different parameters available as independent variables to analyze the PLT variations. Since we collect data from 11 different native operators in four countries, we follow two approaches to generate the web performance models: In the first approach, we merge all the data we collect, independent of operator and investigate how the PLT values vary with changes in the main identified predictor variables, independent of operator, target website or protocol (i.e., the general model).

In the second approach, we generate separate models for each of the operators we capture in the measurement campaign, while merging the measurements for the different websites we targeted in our measurements (i.e., per-operator models), independent of target website or protocol version.

When evaluating the resulting regression models, we investigate several performance metrics, including: the fractional bias (FB); the geometric mean bias (MG); the normalized mean square error (NMSE); the geometric variance (VG); the correlation coefficient (R); the fraction of predictions within a factor of two of observations (FAC2). A perfect model would have MG, VG, R, and FAC2 equal to 1.0, while FB and NMSE are equal to 0.0. Out of the metrics above, the most robust one is the FAC2,

Table 6.4: Predictor variables and their correlation (P value) with PLT within the context of the general regression model for the page load time. The "Frequency" column shows in how many individual per-operator models the predictor was strongly correlated with the PLT.

Type	Predictor Variable	P value	Freq.
Target website	Total # of objects	$< 2e-16$	10
	Number of JS	n/a	0
	Number of CSS	n/a	0
	Number of HTML	n/a	0
	Number of images	n/a	0
	Maximum object size	n/a	0
	Average object size	$< 2e-16$	10
	STDEV object size	$< 2e-16$	9
	Number of domains	$< 2e-16$	10
	Protocol	$6.55e-5$	0
	Page size	n/a	0
Network and User Context	Frequency	$< 2e-16$	1
	LTE frequency band	$< 2e-16$	0
	Throughput	$< 2e-16$	8
	Average RTT	n/a	0
	Minimum RTT	n/a	1
	Maximum RTT	$1.05e-9$	6
	Time of day	$< 2e-16$	8
	Signal strength variations	$< 2e-16$	8
	RSRQ	$7.03e-10$	4

since it is not excessively influenced by very low or very high outliers. Moreover, the fractional bias (a dimensionless number) is convenient for comparing the results from studies involving different concentration levels (e.g., different numbers of experiments per operator and per target website). For a model to be acceptable, there are several metrics we can monitor and interpret together:

- the fraction of predictions within a factor 2 of the real observations is about 50% or greater (FAC2 > 0.5)
- the mean bias is within $\pm 30\%$ of the mean ($|FB| < 0.3$ or $0.7 < MG < 1.3$)
- random scatter is about a factor of two to three of the mean (NMSE < 1.5 or $VG < 4$)

Regression Results In this section, we analyze the performance of the regression models we generated using the `WebWorks` dataset and interpret our results. We first focus on the general model, where we investigate the trends in the web measurements independent of operator and target website. We then discuss in more detail how we break down this analysis for each operator in order to capture how different potential network configurations might impact web performance. In Table 6.5 we show the performance metrics for all the 11 different models we generated.

We first generated a general model using all the data from the 10 operators we captured in the `WebWorks` dataset (after eliminating Orange (ES) because of limited sample size), independently of the target website or H1s/H2 protocol. In terms of model performance, we observe in Table 6.5 (first line corresponding to the general model) that the FB value is very close to 0.0 value (-0.002), showing the model is relatively free of bias. This is further supported by the other metrics we show in Table 6.5. The NMSE value close to 0.5 (0.40) indicates that the fraction of predictions within a factor of two of the actual observation value is higher than 50%. Furthermore, when evaluating the MG metric (1.06) we conclude that the mean bias is actually within $\pm 30\%$ of the mean. This tells us that the general model we built is highly reliable for further predictions. The column "P values" in Table 6.4 shows the P values of the predictors we considered in the linear regression model. In accordance with the description we have in the previous section, we removed several predictors we initially considered (marked

Table 6.5: Performance of `WebWorks` models, general and per-operator. We show the correlation coefficient (COR), the normalized mean square error (NMSE), the geometric mean bias (MG), the geometric variance (VG), the fractional bias (FB) and the fraction of predictions within a factor of two of observations (FAC2).

Model	COR	NMSE	MG	VG	FB	FAC2
General (all)	0.89	0.40	1.06	1.006	-0.002	1.23
Telenor (NO)	0.88	0.32	1.04	1.003	0.0005	1.17
Telia (NO)	0.76	0.33	1.02	1.01	0.019	1.12
ICE (NO)	0.93	0.20	1.01	1.001	0.002	1.05
TIM (IT)	0.88	0.34	1.03	1.005	-0.0005	1.15
Vodafone (IT)	0.84	0.30	1.05	1.01	0.002	1.12
Wind (IT)	0.94	0.34	1.05	1.005	0.0005	1.18
Telia (SE)	0.91	0.49	1.07	1.007	0.002	1.32
3 (SE)	0.86	0.51	1.07	1.01	0.012	1.33
Telenor (SE)	0.90	0.38	1.05	1.003	0.0012	1.22
Yoigo (ES)	0.79	0.50	1.08	1.01	0.003	1.33

with n/a in Table 6.4) after observing multi-collinearity between predictors or weak correlation with the target variable.

From this study we conclude that the most important metrics that influence the PLT values are directly dependent on the target website, namely total number of objects in each target webpage, average and standard deviation of object size and the number of domains. Moreover, we see that the network specific parameters, such as throughput, RSRQ or maximum RTT heavily impact the experience of the end-user. The context in which the end-user performs the web browsing activity also impacts the PLT, but to a lesser degree than the earlier mentioned predictors.

6.7 Analysing Impact from Non-content Parameters

By monitoring the 9 parameters, mentioned in Table 6.5 we can make good prediction of the PLT values. We see that adding up more contextual information may further improve accuracy of the model. In this section we demonstrate this by separately displaying impact of parameters both from network and experimental context on web QoE using dataset collected in the second measurement campaign.

Impact from Network Context: User's experience correlates highly with varying network conditions. We need to delve into more details to understand both the intensity and trend of impact, on web performance, when a feature varies so to make better network configuration decisions. For evaluation we use the web experiments conducted in Sweden for a single web page such as *reddit.com*.

We first look at the independent effect of signal quality, that has already proven significance in the web QoE model Table 6.4. Instead of RSRQ we take RSSI as it represents signal quality for each type of RAT. As expected the performance of the web-page clearly degrades under poor signal strength conditions as is illustrated by Figure 6.3. Interestingly, the RSSI has a non-linear effect, with sudden degradation for RSSI smaller than -80dBm. Above -60dBm, we observe no further improvement, an important observation for network operator's looking for optimal network configuration. Next, by keeping nodes type as stationary, as average RTT to the web server raises the performance gradually degrades per website, verifying RTT as an important factor in shaping PLT.

We further assess the impact of the Radio Access Technology (3G and 4G/LTE) used for accessing the Internet. Using the experiments from nodes under stationary scenarios, Figure 6.3 captures the impact of RAT on web performance. As expected, the better performance offered by 4G/LTE technology benefits the PLT.

As a last parameter, we showcase the impact of the number of handovers under mobility. Here we compare experiments with no inter-RAT handover against experiments where at least one inter-RAT (3G to 4G or vice versa) handover has been observed. Results clearly show the penalty introduced by inter-RAT handovers, with a worsening factor of 2x on the average, and worst case up to 5x slower (when several inter-RAT handovers occur).

Impact from Experiment Context: The context in which the experiment is executed cannot be ignored while estimating a user's QoE. For example with 4G high speed networks we enjoy good quality MBB services, still the quality of services degrades when user is mobile, this is clearly indicated by the higher web-page loading latency with mobile nodes in Figure 6.4. The experience with web-browsing worsens with raise in speed displayed in the figure with distance covered during a single web-experiment.

Knowledge of the effect from contextual features can help both the operators and the users in improving the experience with an MBB service by taking corrective actions. For example the study shows that one of the main factor due to which mobility gives poor performance is high inter-RAT handovers, a solution of which might be for operators to prefer mobile users to be connected to cell towers with higher footprints.

As for procotol, though the web QoE model in Table 6.5 considers it as a significant feature, the H2 gives only a slightly better performance than H1s. Also we observe that the relationship is not consistent over different web-sites, we therefore conclude that web QoE model is trying to capture marginal gains by taking protocol as one of the predictors.

Just like protocols, when understanding effect of browser type on a single website like *reddit.com*, Chrome is performing better than Firefox. However, for a different website the comparative relation changes indicating that a web QoE model generated and tested on different websites may not always work.

Takeaway: From the web QoE model and separate analysis of the non-content parameter we conclude that the most significant impact on PLT of a web-site is directly related to how complex and heavier it is. This presses a need for website owners and designers to minimize unnecessary load on their web-pages so to raise user conversion rate. Secondly along with RAT mode, mobility is the one of highest influencing factors requiring operators to upgrade their services to 4G, where it has not been done, and devise a strategy to minimize inter-RAT handover where possible. Thirdly good signal quality raises user's experience, again encouraging network operators to enhance coverage availability. Lastly find impact from protocols and browsers to be non-consistent and negligible.

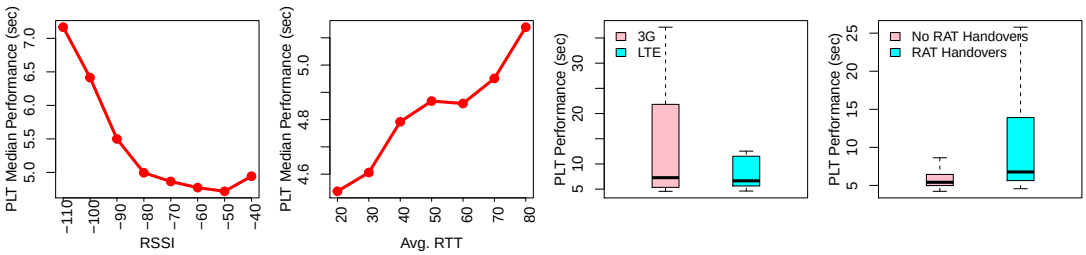


Figure 6.3: (a) Impact of network context on web QoE (PLT) of reddit.com

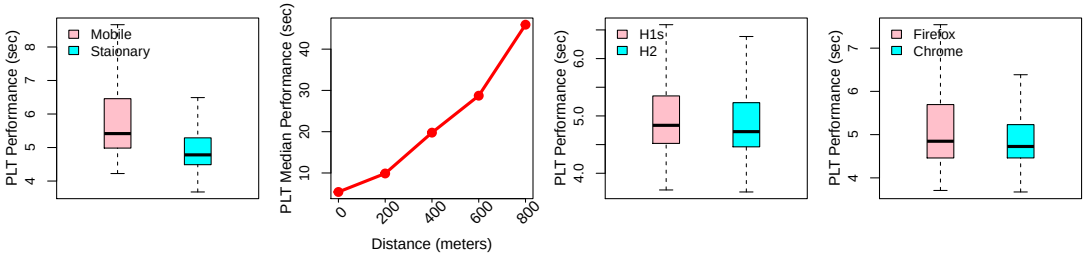


Figure 6.4: (b) Impact of experiment context on web QoE (PLT) of reddit.com

6.8 Tuning of Network Parameters

We undertook this last study in order to validate the impact of changes in the LTE network features on user's web QoE. Here we exploit the opportunities provided by the experimental FLEX-MONROE platform for tweaking the values of the network-side parameters, such as the Modulation and Coding Scheme (MCS) or the transmission power level from the eNodeB. An additional aim was to identify optimal network con-

figuration under different settings. We however were limited in taking full advantage of the experimental settings as at the time of this work, we deployed a single MONROE node at a static distance from LTE eNode. Further we were unable to control the number of simultaneous users, however we were able to validate impact of some of the network features.

Observing the impact of signal quality on the PLT in the operational networks, we configure the LTE network in FLEX-MONROE to analyze the same phenomena under the controlled settings. To achieve this, we run the `WebWorks` experiments for different transmission power levels at eNodeB as the power level changes at eNodeB affects the received signal strength values directly. In Figure 6.5, where we display results of a few websites namely *bbc.co.uk, ebay.com and wikipedia.com*, we observe that the change in PLT is small even after a gap of 2 to 3 power levels. The decrease and less variation in PLT with better power levels, in experimental network, validates the results got from operational network indicating signal quality as a significant feature in varying web-browsing PLT.

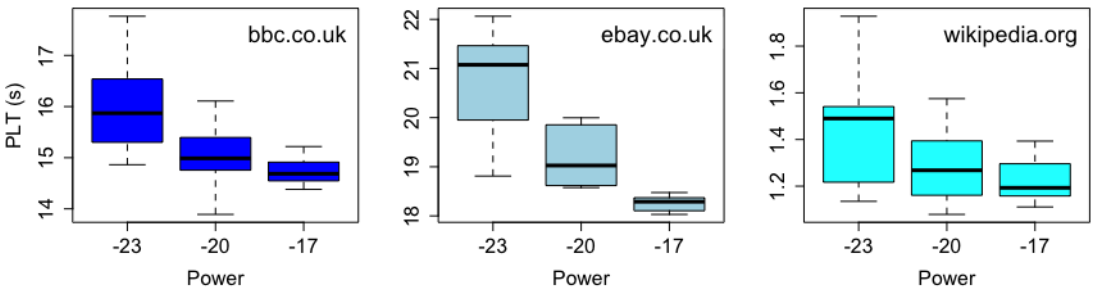


Figure 6.5: PLT at different power levels

In order to understand whether the signal strength variation is large enough to bring significant changes in corresponding PLTs of webpages, we ran further experiments. In Figure 6.6, we observe that though the quality of RSRP reduces with reduction in the RFSignalPower, the overall range of RSRP we experienced is quite small, varying only from -85 to -97 dBm. This whole range is considered *Good* for transmission where RSRP values in general can have best quality (around -44 dBm) to worst (around -140 dBm). We conjecture that by varying the distance between the FLEX-MONROE node and the eNodeB, we may observe more notable changes in RSRP and its propagating

effects on performance metrics.

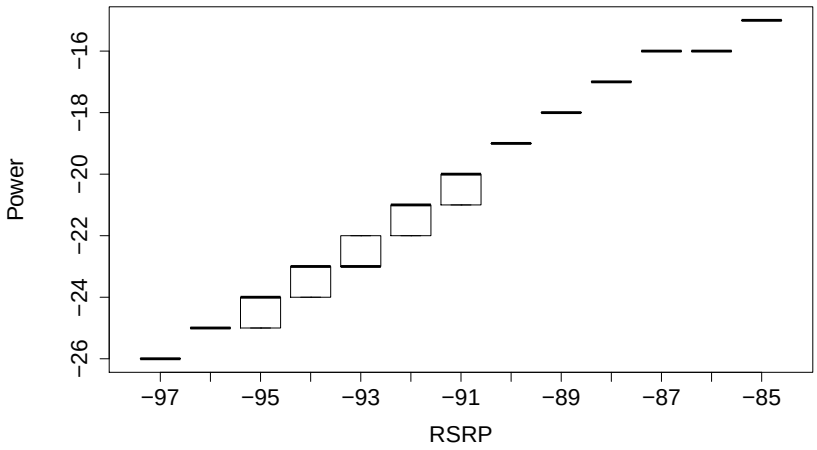


Figure 6.6: RSRP at different power levels

We further quantify the impact of RFSignalPower changes on the throughput. For throughput analysis we leverage another MONROE EaaS (the HTTP download experiment, in this case) and download 100MBytes of data from the same source in each run. Turning to impact of RFSignalPower changes on throughput, we observe that due to the propagated effect of RSRP to CQI and further to MCS, the data rate gradually changes. We show this effect in Figure 6.7, where the throughput drops from 5 MB/s to 1 MB/s with the fall in power level.

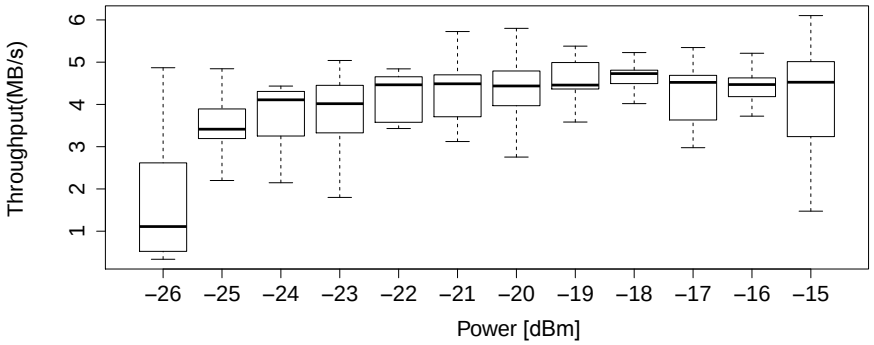


Figure 6.7: Throughput at different power levels

6.9 Summary

In this study, we investigated impact of features from application content, network context, and experiment context features on web-page load time (that translates to user experience) in LTE MBB network. We undertook two measurement studies. In the first

study, we ran web-browsing experiments on the static nodes of MONROE platform (attached to commercial LTE MBB networks) and generated an objective web QoE model out of collected parameters. The highest impact is found to be due to a web-page's own features consisting of number of objects downloaded, number of unique domains accessed by a web-page, average and standard deviation of object size. Such an outcome confirms that web-site owners and designers should minimize object-based load on their sites if they are interested in raising conversion rate.

The model also revealed importance of network QoS and user's context in influencing the web QoE i.e. PLT. To clearly understand the intensity of impact from a number of prominent non-content features, we conducted an other set of measurement study, this time both on mobile and static MONROE nodes and focused on retrieving many additional impacting factors. This analysis illustrated node type, movement speed during web-browsing, number of inter-RAT handovers, signal quality, and RTT to the web-server as main reasons of influencing a user's experience, from non-content domain. Comparative to above features we further observed impact of protocol type i.e. HTTP1.1/TLS and HTTP/2.0, and browser type i.e. Firefox and Chrome to be non-consistent across web-pages and negligible.

With the intention to validate the impact of network parameters and to identify optimal configuration under different network settings such as user load, distances from eNodeB etc., we additionally carried out a small set of web-browsing measurements on the experimental LTE test bed of FLEX-MONROE platform. Though we were able to validate impact from signal quality on the web-page PLT, due to deployment of a single MONROE node in FLEX premises we could not better exploit the test-bed.

Chapter 7

Conclusion and Future Directions

With mobile broadband networks an end-user's expectations about seamless availability of network services has raised. To maintain customer market, a network-operator has to stay informed about users' experience with the network e.g. with continual feedback from end-users. This feedback act as first hand information in identifying the problem domain. It along with back-end network-side performance aid network operators to improve their networks.

User-side measurements provides user's view of a network. These measurements range from reporting simple performance metrics including signal strength, download and upload speed, throughput and RTT to QoE with application layer service such as call quality, browsing speed, video resolution and its continuity etc. The measurements can be collected by three means i.e. drive testing, crowdsourcing and through distributed-test bed. Drive-testing is conducted by an MNO to examine its network's performance and to gain an insight into the working of other competitive carriers by deploying engineers with a set of hand-sets who walk or drive in the test region while examining the performance of the networks. Drive-testing collects a broad range of network performance metrics including network availability, call quality and web-browsing speed etc. The second method is to collect the measurements from real end-users. The method in its simplest form uses a crowdsourced application that runs in background to collect and report basic performance features such as signal quality, download and upload speed and RTT to a central server. Lastly there are distributed

platforms that have special purpose end-nodes attached to commercial cellular networks or Internet through WLAN with the goal to understand the working, performance and problems within the network by running either active tests or passively tracking of the ongoing traffic.

Aim of the research study is to inquire value in the user-side measurements collected especially with crowdsourcing and distributed test-beds. The contributions made to arrive the aim along with some of identified limitations in the work and prospective future directions are mentioned in the sections to follow:

7.1 Contributions

Network availability is the foremost requirement for the smooth working of rest of the network services, e.g. weak network signals hinders establishing a successful call and slows down browsing speed to an unbearable state. In this study, we therefore focus and start with understanding role of crowdsourced measurements in providing an accurate network availability status of a region. The research work thus consists of:

Robust interpolation process: The crowdsourced measurements are uncontrolled and reported with different location-based approaches such as GPS or positioning assisted with WLAN and cellular network databases. Moreover, depending upon user's discretion the collected measurement samples does not evenly cover the target region and are therefore non-uniformly distributed with holes and clusters within them. Similarly the density of reported samples is not under-control of the crowdsourced server either. With these characteristics, building of a reliable coverage map of the region is challenge. By analysing a set of different interpolation schemes, we find Ordinary Kriging to be a robust and efficient process in generating a radio coverage map even when the crowdsourced measurements are sparse, have noise and are non-uniformly spread. This is because Ordinary Kriging efficiently utilizes the spatial correlation among RSSI samples, also it is suitable when there is stationarity within the region.

Sampling strategy for coverage map generation: Highly accurate coverage map is achieved only when large number of samples well distributed over the region are available. In crowdsourcing, end-users are reluctant to run the crowdsourced applications due to battery consumption, limited data caps and security concerns. To have a

representative measurement set, monetary incentives are required to encourage users from under-represented areas. With the aim to maximize accuracy of the resulting coverage map by staying within budget we devise `ZipWeave`. `ZipWeave` exploits the fact that places differ in their sampling requirement e.g. open places like fields have smooth variation in the signal quality while places with tall congested buildings have high variation in signal quality even at small lag distances. For better representation latter areas need higher sampling density than the earlier ones. `ZipWeave` assists in identifying different parts of the region according to their sampling requirement, which finally translates into reducing sampling cost by half while still achieving a desired coverage map accuracy.

Generating a reliable device-centric coverage map: Demand from users and competition of vendors has lead to rapid increase in diversity of device types. This diversity has enabled wider availability of economically affordable devices and at the same time has extended the range of variations across devices' capabilities. E.g. one of these differences is exhibited by the variation in RSS distributions of the device-types or models. With differences in RSS values an obvious question is how well a single combined coverage map correlates with what a user might observe at his device-type. For coarse representation of a coverage map, a granularity with which network status is displayed on official website of a network carrier, we find that if measurement samples from same vendor have been included in coverage map estimation the correlation is high else similarity is weak.

For fine-grained device-centric coverage map best approach is to use measurement samples from same device-model, collected within same time-span. If such measurements are not available then measurements from same device model and previous-time span is the best option. Among device models, the error is lower when measurements from different device models belonging to same vendor are used than those which are from different vendors. We find the percentage raise and drop in accuracy with each of these cases is variable, however the general pattern of preference is similar. Also when measurements from unfavourable sources are used in conjunction with favourable sources the raise in error is not drastic rather it is negligible, again indicating that a combined coverage map is trust worthy, from a device -type perspective,

if it takes into account samples from same device model.

Cell-tower localization with crowdsourced measurements The differences in features across crowdsourced measurement scenarios make it impossible for a single localization algorithm to consistently estimate celllular infrastructure accurately. Also filtering out less predictive measurement samples may not always improve the accuracy of a localization process. The features of a measurement scenario that labels it un-predictive may become favourable when localization approach is changed. In this work we proposed `AAS`, a cell tower localization approach, that adaptively switches among a suite of five algorithms including Centroid, WC, SRSS, GL and MCPL based on the characteristics of a cell's measurement scenario. `AAS` reduces the localization errors substantially to that of the FWC, a state of art localization approach. Within a given dataset it outperforms the most recurring best performing localization algorithm by reducing the mean localization error by half or more. `AAS` trained on a dataset from a different operator and country applies well to a second dataset, especially when measurement feature distributions are similar for both the datasets. Even in case of difference, the achieved accuracy still outperforms than when using a single fixed localization algorithm.

Web browsing QoE analysis with features from content, context and network: This study was performed on a distributed platform of MONROE [10] with a number of static and mobile nodes, each connected to at most three commercial LTE MBB networks. The study demonstrates that the end-nodes connected to a distributed platform assist one in understanding the reasons behind a particular QoE. For example with MONROE platform we were able to run a large scale web-browsing experiments. During the web experiments the objective web QoE parameter PLT along with web-page content parameters, underlying network performance metadata and the experiment's context was collected.

The analysis show that web QoE is highly impacted by the complexity of a web-page. It is also influenced a lot by the number of inter-RAT handovers which especially happen when user is moving at high speed. The other impacting factors are time of the day feature, signal quality, RTT to the web-server, throughput of the network and radio access technology. To have good web QoE, the analysis indicate that network con-

figuration is not the only responsible entity, the web-designer should try to avoid less observable and less useful web-page objects, further a user should opt for favourable context if he cannot afford long PLT.

7.2 Limitations and Future Directions

This section mentions some of the limitations and possible future directions for the work undertaken in this research study.

Interpolation process with non-stationary region: For coverage map generation an interpolation process is chosen that achieves good overall accuracy, for example in our analysis the best approach is OK. OK is influenced by the correlation in signal strength values of samples at different lag distances with possible heterogeneity in correlation across the region at large. In the latter situation there are two issues that needs investigation:

- First, for parts of the region where no sampling is done at all and which are big enough to have the possibility of having different terrain and propagation characteristics than its surrounding sampled regions, there is high chances that interpolated values for these places are quite far away from the true values. On the basis of rate of change in terrain characteristics towards un-sampled parts, threshold distance should be defined after which samples should be collected, even if through drive testing so that interpolation process can provide true picture of every part of the region.

- Secondly, if there is heterogeneity or non-stationarity in samples signal strength correlation at different parts of the region, using a single variogram for a larger span of geographic area may be misleading. The probable solution of which might be to identify such heterogeneous areas, that is the areas with different variograms, and apply interpolation processes separately on each of them by using their corresponding variograms.

Sampling with zip-code areas: In `ZipWeave`, for the first phase of sampling we propose to have uniform-random sampling across the whole region. For uniformity

in samples the region is divided into equally sized virtual tiles. In reality however there are places within the region e.g. streams, mountains and restricted areas where sampling is not only difficult but also has little value from customers' point of view. Instead of regions, a realistic option is to use zip codes. Though zip-codes vary in sizes, the benefit of their sizes is they indirectly represent the population density and number of buildings within them. Thus they provide potential clusters to `ZipWeave` with reducing sample size even at the first phase i.e. by collecting equal number of random samples from each zip-code area irrespective of its size.

Analysis on measurement sources for device-centric coverage map generation To generate an accurate coverage map for a device-type the best set of measurements are when they are collected by same device type and within same time span. The analysis study however has some limitations, we have not analyzed the reliability of measurements from device-types (belonging to same vendor) that are close in models than those that are far in functionality. Secondly we have not investigated the reliable time-span duration beyond which the measurements are misleading rather than helpful. This is especially desirable when a device-type lacks fresh measurement samples and needs a reliable coverage assessment.

Pre-processing of crowdsourced measurement samples For cell tower localization we exploited the, as it is, features of the measurement scenarios without applying any pre-processing steps. Crowdsourced measurements, however, can come with outliers and noise that badly degrade working of the localization algorithms e.g. noise in measurement locations such that the pair wise distances between some of the measurement samples is much larger than the largest possible footprint of the cell tower, a mixture of very high and very low signal strength values present in different parts of the measurement scenarios and clusters of measurement samples separated by large distances. We expect that application of suitable pre-processing steps on measurement samples will lead to reduction in localization errors of different algorithms and will ultimately enhance accuracy of `AAS` approach.

Evaluating impact of network configuration on web browsing QoE: In the work under FLEX-MONROE project, one of the aims was to identify optimal network configuration at LTE MBB network that has to result in better user's QoE in terms of

web-browsing PLT. The deployment of a single MONROE node at FLEX premises helped us validate the impact of significant network parameters, identified by MONROE testbed experiments, including RSRP and throughput on web-page PLT.

However with a single MONROE node being deployed at a static distance from LTE eNodeB, we were limited in our understanding of threshold RSRP, throughput and RTT beyond which PLT decrease is negligible. For example varying of RF Signal Power values from -15 to -26 at eNodeB we observed change in corresponding RSRP quality. But the change only ranged from -85 to -97 dBm showing small negative impact on PLT. The reason of which is the whole range of -85 to -97dBm is considered good for communication, where in fact the complete range of RSRP is from -44 to -140 dBm. We were therefore not able to see impact of extreme RSRP values. Similar was the situation with throughput (i.e. a result of changing MCS and number of simultaneous users) and RTT (a result of node's distances from eNodeB). To get a better understanding of relation between PLT and network configuration at LTE eNodeB, more MONROE nodes are therefore needed to be deployed at different lag distances with experiments to run simultaneously on different number of these nodes. Such a deployment is expected to give a better insight on impact of changing LTE network configuration, on network-side features and ultimately, on web-browsing QoE.

Bibliography

[1] Nokia. Quality of experience (QoE) of mobile services: Can it be measured and improved? *White Paper, Nokia Corporation*, 2004.

[2] B. Krogfoss. LTE small cells greatly improve QoE for video. https://insight.nokia.com/lte-small-cells-greatly-improve-qoe-video, Nov. 2015.

[3] S. Patel. Map coordinates for all mobile phone masts in the UK. https://www.whatdotheyknow.com/request/map_coordinates_for_all_mobile_p, July 2008.

[4] J. Mena. *Machine-to-Machine Marketing (M3) via Anonymous Advertising Apps Anywhere Anytime (A5)*. CRC Press, 2012.

[5] AKAMAI. Measuring real customer experiences over mobile networks. https://www.akamai.com/jp/ja/multimedia/documents/white-paper/measuring-real-customer-experiences-over-mobile-networks-report.pdf, accessed June 2018.

[6] Crowdsourcing Week. What is crowdsourcing? https://crowdsourcingweek.com/what-is-crowdsourcing/, accessed May 2018.

[7] OpenSignal. Understand and Improve Mobile Signal. https://opensignal.com/, accessed July 2017.

[8] MobiPerf. http://www.mobiperf.com, 2014.

[9] C. Midoglu and P. Svoboda. Opportunities and challenges of using crowdsourced measurements for mobile network benchmarking: A case study on rtr open data. In *SAI Computing Conference*, 2016.

[10] MONROE. Monroe project. https://www.monroe-project.eu/, accessed June 2018.

[11] RIPE NCC. Ripe network coordination center. https://www.ripe.net/, accessed June 2018.

[12] J. Johansson, W. A. Hapsari, S. Kelley, and G. Bodog. Minimization of drive tests in 3gpp release 11. *IEEE Communications Magazine*, 50(11):36–43, 2012.

[13] A. Katmada, A. Satsiou, and I. Kompatsiaris. Incentive mechanisms for crowdsourcing platforms. In *Internet Science - Third International INSCI Conference*, pages 3–18, Sep. 2016.

[14] OpenCelliD. Frequently asked questions. http://wiki.opencellid.org/wiki/FAQ#I_know_where_cell_tower_x_exactly_is_but_OpenCellID_shows_another_position, accessed July 2017.

[15] openBmap. openBMap: free and open map of wireless communicating objects. http://www.openbmap.org/, accessed July 2017.

[16] Ö: Alay, A. Lutu, R. García, M. Peón-Quirós, V. Mancuso, T. Hirsch, Dely. T., et al. Measuring and assessing mobile broadband networks with monroe. In *IEEE 17th International Symposium WoWMoM*, pages 1–3, 2016.

[17] Shikhar. Flex-monroe: Fire lte testbeds for open experimentation. https://www.simula.no/research/projects/flex-monroe-fire-lte-testbeds-open-experimentation, accessed July 2018.

[18] Wuri A Hapsari, Anil Umesh, Mikio Iwamura, Magorzata Tomala, Bodog Gyula, and Benoist Sebire. Minimization of drive tests solution in 3gpp. *Communications Magazine, IEEE*, 50(6):28–36, 2012.

[19] D.J. Pringle. Spatial interpolation techniques. http://www.nuim.ie/staff/dpringle/gis/gis09.pdf, accessed Jan. 2015.

[20] M. Béla. Spatial analysis 4- digital elevation modeling. https://www.tankonyvtar.hu/en/tartalom/tamop425/0027_SAN4/ch01s02.html, 2010.

[21] GisResources. Classification of interpolation. http://www.gisresources.com/classification-of-interpolation_2/, 2014.

[22] GIS Resources. Types of interpolation methods. http://www.gisresources.com/types-interpolation-methods_3/, accessed Jan. 2018.

[23] J. Li and A.D. Heap. A review of spatial interpolation methods for environmental scientists. *Geoscience Australia, Record 2008/23*, 2008.

[24] M. A. Oliver and R. Webster. A tutorial guide to geostatistics: Computing and modelling variograms and kriging. *CATENA*, 113:56–69, 2014.

[25] Noel A. C. Cressie. *Statistics for Spatial Data*. Wiley Series in Probability and Statistics, 1993.

[26] Z. Dali, Z. Huihui, and F. Weimiao. Research on the construction of radio-map based on support vector regression. *Fourth International Conference on Instrumentation and Measurement, Computer, Communication and Control*, 2014.

[27] J. Rackoa, J. Machaja, and P. Bridaa. Wi-fi fingerprint radio map creation by using interpolation. *Procedia Engineering*, 192:753–758, 2017.

[28] A. Konak. Estimating path loss in wireless local area networks using ordinary kriging. In *Proceedings of the WSC*.

[29] C. T. Phillips, M. Ton, D. C. Sicker, and D. Grunwald. Practical radio environment mapping with geostatistics. In *IEEE International Symposium on DySPAN*, pages 422–433, 2012.

[30] S. Kolyaie and M. Yaghooti. Evaluation of geostatistical analysis capability in wireless signal propagation modeling. In *In Proceedings of 11th International Conference on GeoComputation*, 2011.

[31] S. Kolyaie, M. Yaghooti, and G. Majidi. Analysis and simulation of wireless signal propagation applying geostatistical interpolation techniques. *Archives of Photogrammetry, Cartography and Remote Sensing*, 22:261–270, 2011.

[32] Berna Sayraç, Ana Galindo-Serrano, Sana Ben Jemaa, Janne Riihijärvi, and Petri Mähönen. Bayesian spatial interpolation as an emerging cognitive radio application for coverage analysis in cellular networks. *Trans. Emerging Telecommunications Technologies*, 24(7-8):636–648, 2013.

[33] A. Konak. A Kriging approach to predicting Coverage in Wireless Networks. *International Journal of Mobile Network Design and Innovation*, 3(2):65–71, 2009.

[34] Kent Hunt. Rf signal tracker (donut). https://play.google.com/store/apps/details?id=com.hotrod.utility.rfsignaltrackerhl=enGB, accessed June 2016.

[35] Y. Zhang, Y. Zhu, M. Lu, and A. Chen. Using compressive sensing to reduce fingerprint collection for indoor localization. In *WCNC*, pages 4540–4545, April 2013.

[36] T. Hastie, R. Mazumder, J. D. Lee, and R. Zadeh. Matrix completion and low-rank SVD via fast alternating least squares. *Journal of Machine Learning Research*, 16:3367–3402, 2015.

[37] S. Ji, Y. Xue, and L. Carin. Bayesian compressive sensing. *IEEE Transactions on Signal Processing*, 56(6):2346–2356, 2008.

[38] M. Roughan, Y. Zhang, W. Willinger, and L. Qiu. Spatio-temporal compressive sensing and internet traffic matrices (extended version). *IEEE/ACM Trans. Netw.*, 20(3):662–676, 2012.

[39] B. Yang, S. He, and S.-H. G. Chan. Updating wireless signal map with bayesian compressive sensing. In *In Proceedings of the 19th International Conference on MSWiM*, pages 310–317, November 2016.

[40] A. Katmada, A. Satsiou, and I. Kompatsiaris. Incentive mechanisms for crowdsourcing platforms. In *Internet Science - Third International Conference, INSCI*, pages 3–18, 2016.

[41] B-B. Gao J-F. Wang, A. Stein and Y. Ge. A review of spatial statistics. *Spatial Statistics*, 2:1–14, 2012.

[42] P.M. Sahoo, R. Singh, and A. Rai. Spatial sampling procedures for agricultural surveys using geographical information system. *Journal of the Indian Society of Agricultural Statistics*, 60(2):134–143, 2006.

[43] R.A. Olea. Sampling design optimization for spatial functions. *Journal of the international Association for Mathematical Geology 16.4*, 16(4), 1984.

[44] E.M. Delmelle and P. Goovaerts. Second-phase sampling designs for non-stationary spatial variables. *Geoderma*, 153(1):205–216, 2009.

[45] S. Grimoud, D. J. Brus, B. Sayrac, S. B. Jemaa, and E. Moulines. Best sensor selection for an iterative REM construction. In *2011 IEEE Vehicular Technology Conference, Fall*, 2011.

[46] D. J. Brus and G.B.M. Heuvelink. Optimization of sample patterns for universal kriging of environmental variables. *ScienceDirect, Geoderma*, 138(1–2), 2007.

[47] NorNET Edge. https://www.nntb.no/nornet-edge/, accessed June 2018.

[48] M. Ester, J. Kriegel, H.-P.and Sander, and X. Xu. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on KDD-96*, pages 2236–231, 1996.

[49] K. Krivoruchko, K.and Butler. Unequal probability-based spatial sampling. http://www.esri.com/esri-news/arcuser/spring-2013/unequal-probability-based-spatial-sampling, Spring 2013.

[50] D. J. Brus and G. B.M. Heuvelink. Optimization of sample patterns for universal kriging of environmental variables. *Geoderma, ScienceDirect*, 138:86–95, 2007.

[51] J.K. Yamamoto. An alternative measure of the reliability of ordinary kriging estimates. *Mathematical Geology*, 32(4):489–509, 2000.

[52] Mobile Coverage Checker. https://ee.co.uk/why-ee/mobile-coverage#nav-0, Accessed May 2018.

[53] Check Coverage and Network Status. https://www.o2.co.uk/coveragechecker, Accessed May 2018.

[54] Android fragmentation (august 2015). https://opensignal.com/reports/2015/08/android-fragmentation/, 2015. [Online; accessed July 2018].

[55] E. Kroski. *On the Move with the Mobile Web: Libraries and Mobile Technologies*. ALA TechSourc, 2008.

[56] F. Razally. Mobile handset testing. https://www.ofcom.org.uk/_data/assets/pdf_file/0015/72231/mobile_handset_testing_1v01.pdf, November 2015.

[57] BoR (17) 186. BEREC Preliminary report in view of a common position on monitoring mobile coverage. 7300-draft-berec-preliminary-report-in-view-o_0.pdf, October 2017.

[58] J.-G. Park, D Curtis, S. J. Teller, and J. Ledlie. Implications of device diversity for organic localization. In *INFOCOM*, pages 3182–3190, April 2011.

[59] C. Laoudias, D. Zeinalipour-Yazti, and C. G. Panayiotou. Crowdsourced indoor localization for diverse devices through radiomap fusion. In *IPIN*, pages 1–7, Oct. 2013.

[60] H. N. Manh, C.-C. Huang, and L Hsiao-Yi. Landmark-based device calibration and region-based modeling for rss-based localization. *Wireless Communications and Mobile Computing*, pages 1726–1745, 2016.

[61] A. Haeberlen, E. Flannery, A. M. Ladd, A. Rudys, D. S. Wallach, and L. E. Kavraki. Practical robust localization over large-scale 802.11 wireless networks. In *In Proceedings of the 10th MOBICOM Conference*, pages 70–84, September 2004.

[62] Fangfang Dong, Yiqiang Chen, Junfa Liu, Qiong Ning, and Songmei Piao. A calibration-free localization solution for handling signal strength variance. In *Mobile Entity Localization and Tracking in GPS-less Environnments, Second International Workshop, MELT 2009, Orlando, FL, USA, September 30, 2009. Proceedings*, pages 79–90, 2009.

[63] J. Machaj, P. Brida, and R. Piché. Rank based fingerprinting algorithm for indoor positioning. In *IPIN 2011*, pages 1–6, September 2011.

[64] X. Fan, P. Yang, C. Xiang, and L. Shi. iMap: A Crowdsensing Based System for Outdoor Radio Signal Strength Map. In *2016 IEEE Trustcom/BigDataSE/ISPA*, August 2016.

[65] C. Xiang, P. Yang, C. Tian, L. Zhang, H. Lin, F. Xiao, M. Zhang, and Y. Liu. CARM: crowd-sensing accurate outdoor RSS maps with error-prone smartphone measurements. *IEEE Trans. Mob. Comput.*, 15(11):2669–2681, 2016.

[66] Wikipedia. Mobile phone signal. https://en.wikipedia.org/wiki/Mobile_phone_signal, accessed July 2018.

[67] Steve Stong. Mapping the unserved. https://manypossibilities.net/2017/04/mapping-the-unserved/, April 2017.

[68] Z. Ji and R. Jain. Google enables Location-aware Applications for 3rd Party Developers. http://googlemobile.blogspot.co.uk/2008/06/google-enables-location-aware.html, accessed June 2017.

[69] G. Fleishman. How the iPhone knows where you are. http://www.macworld.com/article/1159528/smartphones/how-iphone-location-works.html, accessed July 2017.

[70] ENAiKOON. OpenCelliD. http://www.opencellid.org, accessed July 2016.

[71] Z. Li, A. Nika, X. Zhang, Y. Zhu, Y. Yao, B.Y. Zhao, and H. Zheng. Identifying Value in Crowdsourced Wireless Signal Measurements. In *Proceedings of ACM WWW Conference*, 2017.

[72] A. P. Subramanian, P. Deshpande, J. Gao, and S. R. Das. Drive-By Localization of Roadside WiFi Networks. In *In Proceedings of INFOCOM*, 2008.

[73] Z. Zhang, X. Zhou, W. Zhang, Y. Zhang, G. Wang, B. Y. Zhao, and H. Zheng. I Am the Antenna: Accurate Outdoor AP Location using Smartphones. In *In Proceedings of MobiCom*, 2011.

[74] J. Yang, A. Varshavsky, H. Liu, Y. Chen, and M. Gruteser. Accuracy characterization of cell tower localization. In *In Proceedings of UbiComp*, pages 223–226, 2010.

[75] E. Neidhardt, A. Uzun, U. Bareth, and A. Küpper. Estimating Locations and Coverage Areas of Mobile Network Cells based on Crowdsourced Data. In *In Proceedings of 6th Joint IFIP WMNC, year=2013*.

[76] D. Han, D. G. Andersen, M. Kaminsky, K. Papagiannaki, and S. Seshan. Access point localization using local signal strength gradient. In *In Proceedings of PAM Conference*, 2009.

[77] Z. Fang, L. Haiyong, G. Hao, and S. Qijin. An RSSI Gradient-based AP localization algorithm. *China Communications*, 11(2):100–108, 2014.

[78] Y. Cho, M. Ji, Y. Lee, J. Kim, and S. Park. Improved Wi-Fi AP position estimation using regression based approach. In *In Proceedings of IPIN*, 2012.

[79] K. Liu, Z. Meng, and C.-Mi. Own. Gaussian process regression plus method for localization reliability improvement. *Sensors*, 16(8), 2016.

[80] M. Aravecchia and S. Messelodi. Gaussian process for rss-based localisation. In *10th International Conference WiMob*, pages 654–659, October 2014.

[81] A. Goswami, L. E. Ortiz, and S. R. Das. Wigem: a learning-based approach for indoor localization. In *In Proceedings of CoNEXT*, page 3, December 2011.

[82] Myungin Ji, J. Kim, Y. Cho, Y. Lee, and S. Park. A novel Wi-Fi AP localization method using monte carlo path-loss model fitting simulation. In *In Proceedings of IEEE PIMRC*, September 2013.

[83] A. Achtzehn, L. Simić, P. Gronerth, and P. Mähönen. A propagation-centric transmitter localization method for deriving the spatial structure of opportunistic wireless networks. In *10th International Conference of WONS*, pages 139–146, 2013.

[84] M. Kim, D. Kotz, and J. J. Fielding. Crawdad dataset dartmouth/wardriving (v. 2006-06-02). http://crawdad.org/dartmouth/wardriving/20060602, 2006.

[85] G. Ranjan, Z.-L. Zhang, S. Ranjan, R. Keralapura, and J. Robinson. Unzipping cellular infrastructure locations via user geo-intent. In *In Proceedings of INFOCOM*, 2011.

[86] Di Wu, Liu, Y. Q. Zhang, J. McCann, A. Regan, and N. Venkatasubramanian. CrowdWiFi: Efficient crowdsensing of roadside WiFi networks. In *In Proceedings of ACM Middleware Conference*, 2014.

[87] H. Nurminen, M. Dashti, and R. Piché. A survey on wireless transmitter localization using signal strength measurements. *Wireless Communications and Mobile Computing*, 2017.

[88] P. Nurmi and J. Bhattacharya, S.and Kukkonen. A grid-based algorithm for on-device GSM positioning. In *In Proceedings of UbiComp*, 2010.

[89] F. Ricciato, P. Widhalm, M. Craglia, and F. Pantisano. Estimating Population Density Distribution from Network-based Mobile Phone Data. *JRC Technical Report*, 2015.

[90] Shikhar. Train test similarity. https://www.kaggle.com/shikhar1/train-test-similarity, accessed July 2017.

[91] Airtel zambia. http://www.africa.airtel.com/wps/wcm/connect/africarevamp/Zambia/, accessed August 2018.

[92] V. Pichapati, H. Kowshik, and A. P. and Subramanian. Location assisted handoffs in dense cellular networks. In *Proc. IEEE SECON*, 2014.

[93] G. Kumar, R. Davies, T. Dwyer, and R. Soni. Methods of determining coverage areas, April 30 2013. US Patent 8,433,327.

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

[95] J.-M. Montero, G. Fernández-Avilés, and J. Mateu. *Spatial and Spatio-Temporal Geostatistical Modeling and Kriging*. Wiley, 2015.

[96] S. Sarkar. Cell Phone Towers and Population Density in the Downeast Maine. *Atlas of Maine*, 2011(1), 2011.

[97] S. William and A. Sozzi. Office for National Statistics working paper series. https://www.ons.gov.uk/, accessed December 2017.

[98] S.K. Guha, S. Neogi, and G. Kumar. Radiation hazards from cell phones/cell towers. https://www.ee.iitb.ac.in/~mwave/, published 2011.

[99] H. Liu, H. Darabi, P. P. Banerjee, and J. Liu. Survey of Wireless Indoor Positioning Techniques and Systems. *IEEE Trans. Systems, Man, and Cybernetics, Part C*, 37(6):1067–1080, 2007.

[100] Ericsson. Ericsson mobility report 2016. https://www.ericsson.com/res/docs/2016/ericsson-mobility-report-2016.pdf, 2016.

[101] N. Makris, C. Zarafetas, S. Kechagias, T. Korakis, I. Seskar, and L. Tassiulas. Enabling open access to LTE network components; the NITOS testbed paradigm. In *Proceedings of 1st IEEE Conference on NetSoft*, pages 1–6, April 2015.

[102] V. A. Siris, K. Balampekos, and M. K. Marina. Mobile quality of experience: Recent advances and challenges. In *International Conference on PerCom*, pages 425–430, March 2014.

[103] ITU-T Recommendation P.800. Methods for subjective determination of transmission quality. urlhttps://www.itu.int/rec/T-REC-P.800-199608-I/en, 1996.

[104] Zhuoqun S. and Emmanuel C. I. Voice quality prediction models and their application in voip networks. *IEEE Trans. Multimedia*, 8(4):809–820, 2006.

[105] Kuan-Ta Chen, Cheng-Chun Tu, and Wei-Cheng Xiao. Oneclick: A framework for measuring network quality of experience. In *INFOCOM 2009. 28th IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies, 19-25 April 2009, Rio de Janeiro, Brazil*, pages 702–710, 2009.

[106] Qi Alfred Chen, Haokun Luo, Sanae Rosen, Zhuoqing Morley Mao, Karthik Iyer, Jie Hui, Kranthi Sontineni, and Kevin Lau. Qoe doctor: Diagnosing mobile app qoe with automated ui control and cross-layer analysis. In *Proceedings of the 2014 Internet Measurement Conference, IMC 2014*, pages 151–164, 2014.

[107] A. Balachandran et al. Modelling web quality of experience on cellular networks. In *In Proceedings of the 20th annual international conference on Mobile computing and networking (ACM) 2014*, pages 213–224, 2014.

[108] V. Aggarwal, E. Halepovic, J. Pang, and H. Venkataraman, S.and Yan. Prometheus: toward quality-of-experience estimation for mobile apps from passive network measurements. In *15th Workshop on Mobile Computing Systems and Applications, HotMobile*, pages 18:1–18:6, Febraury 2014.

[109] V. Menkovski, G. Exarchakos, and A. Liotta. Online QoE prediction. In *Second International Workshop on QoMEX*, pages 118–123, 2010.

[110] A. Balachandran, V. Sekar, A. Akella, S. Seshan, I. Stoica, and H. Zhang. A quest for an internet video quality-of-experience metric. In *HotNets-XI*, pages 97–102, 2012.

[111] A. Balachandran, V. Sekar, A. Akella, S. Seshan, I. Stoica, and H. Zhang. Developing a predictive model of Quality of Experience for internet video. In *SIGCOMM*, pages 339–350, August 2013.

[112] K. Eaton. How one second could cost Amazon $1.6 billion in sales. https://www.fastcompany.com/1825005/how-one-second-could-cost-amazon-16-billion-sales, accessed 2018.

[113] Statistica. Percentage of mobile device website traffic worldwide from 1st quarter 2015 to 4th quarter 2017. https://www.statista.com/statistics/277125/share-of-website-traffic-coming-from-mobile-devices/, accessed Feburary 2018.

[114] M. Belshe, R. Peon, and M. Thomson. Hypertext Transfer Protocol Version 2 (HTTP/2). RFC 7540 (Proposed Standard), 2015.

[115] Google Inc. SPDY: An experimental protocol for a faster web. https://www.chromium.org/spdy/spdy-whitepaper.

[116] S. Souders. Velocity and the bottom line. https://goo.gl/8wTE2e, July 2009.

[117] Steve Souders. The performance of web applications – one-second wonders for winning or losing customers. https://goo.gl/ErNLcm, November 2008. accessed June 2017.

[118] E. (Bing) Schurman and J. (Google) Brutlag. Performance related changes and their user impact. https://goo.gl/hCr7ka, Jul 2009. accessed June 2017.

[119] A. Singla, B. Chandrasekaran, P. B. Godfrey, and B. Maggs. The internet at the speed of light. In *In Proceedings of the 13th ACM Workshop on HotNets*, pages 1:1–1:7, 2014.

[120] High Scalability. Latency is everywhere and it costs you sales - how to crush it. https://goo.gl/3PdqIP, July 2009. accessed June 2017.

[121] U. Hoelzle. The google gospel of speed. https://goo.gl/3LQyue, January 2012. accessed June 2017.

[122] P. Dixon. Shopzilla's site redo - you get what you measure. https://goo.gl/e9foNn, June 2006. accessed June 2017.

[123] X. S. Wang, A. Balasubramanian, A. Krishnamurthy, and D. Wetherall. How Speedy is SPDY? In *11th USENIX Symposium on NSDI*, pages 387–399, Seattle, WA, 2014.

[124] Y. Liu, Y. Ma, X. Liu, and G. Huang. Can HTTP/2 really help web performance on smartphones? In *2016 IEEE International Conference SCC*, pages 219–226, June 2016.

[125] H. de Saxcé, I. Oprescu, and Y. Chen. Is HTTP/2 really faster than HTTP/1.1? In *IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 293–299, April 2015.

[126] Matteo Varvello, Kyle Schomp, David Naylor, Jeremy Blackburn, Alessandro Finamore, and Kostantina Papagiannaki. To http/2, or not to http/2, that is the question. *arXiv preprint arXiv:1507.06562*, 2015.

[127] J. Erman, V. Gopalakrishnan, R. Jana, and K. K. Ramakrishnan. Towards a SPDY'er mobile web? *IEEE/ACM Transactions on Networking*, 23(6):2010–2023, 2015.

[128] T. Zimmermann, J. Rüth, B. Wolters, and O. Hohlfeld. How HTTP/2 Pushes the Web: An Empirical Study of HTTP/2 Server Push. In *2017 IFIP Networking Conference and Workshops*, 2017.

[129] D. Imms. Speed Index: Measuring page load time a different way. https://www.sitepoint.com/speed-index-measuring-page-load-time-different-way/, September 2014. accessed June 2017.

[130] Test a website's performance. https://www.webpagetest.org. accessed June 2017.

[131] E. Bocchi, L. De Cicco, and D. Rossi. Measuring the quality of experience of web users. *ACM SIGCOMM Computer Communication Review*, 46(4):8–13, 2016.

[132] U. Goel, M. Steiner, M. P. Wittie, M. Flack, and S. Ludin. Measuring what is not ours: A tale of 3rd party performance. In *ACM PAM*, pages 142–155. Springer, February 2017.

[133] Z. Padychova. How page load time affects conversion rates: 12 case studies. https://blog.hubspot.com/marketing/page-load-time-conversion-rates, accessed May 2018.

[134] Monroe-project/experiments. https://github.com/MONROE-PROJECT/Experiments , year=accessed July 2018.

[135] NITLAB.

[136] Mah-Rukh Fida, Konstantinos Kousias, Rajiullah Mohammad Lutu, Andra, Ozgu Alay, and Anna Brunstrom. Flex-monroe: A unified platform for experiments under controlled and operational lte settings. In *WinTECH'17*, October 2017.

[137] M. Rajiullah. Webworks headlessbrowsing container. https://github.com/MONROE-PROJECT/Experiments/tree/master/experiments/webworks, accessed July 2018.

[138] M. Rajiullah. Monroe-browsertime. https://github.com/MONROE-PROJECT/Experiments/tree/master/experiments/monroe-browsertime, accessed July 2018.

[139] HAR Export Trigger, 2017. http://www.softwareishard.com/blog/har-export-trigger.

[140] Alexa. The top sites on the web. http://www.alexa.com/topsites, 2017.

[141] Athula Balachandran, Vaneet Aggarwal, Emir Halepovic, Jeffrey Pang, Srinivasan Seshan, Shobha Venkataraman, and He Yan. Modeling web-quality of

experience on cellular networks. In *Proceedings of the 20th annual international conference on Mobile computing and networking, MobiCom'14*, pages 213–224. ACM, 2014.