IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 May 2023, accepted 24 May 2023, date of publication 29 May 2023, date of current version 6 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3280828

RESEARCH ARTICLE

Unbiased Benchmarking in Mobile Networks: The Role of Sampling and Stratification

SONJA TRIPKOVIC[1,2], **(Graduate Student Member, IEEE),**
LUKAS ELLER[1,2], **(Graduate Student Member, IEEE),**
PHILIPP SVOBODA[1,2], **(Senior Member, IEEE), AND MARKUS RUPP**[1], **(Fellow, IEEE)**
[1]Institute of Telecommunications, Technische Universität Wien, 1040 Vienna, Austria
[2]Christian Doppler Laboratory for Digital Twin Assisted AI for Sustainable Radio Access Networks, Institute of Telecommunications, Technische Universität Wien, 1040 Vienna, Austria

Corresponding author: Sonja Tripkovic (sonja.tripkovic@tuwien.ac.at)

This work has been funded by the Christian Doppler Laboratory for Digital Twin assisted AI for sustainable Radio Access Networks, Institute of Telecommunications, TU Wien. The financial support by the Austrian Federal Ministry for Labour and Economy and the National Foundation for Research, Technology and Development and the Christian Doppler Research Association is gratefully acknowledged. The authors acknowledge TU Wien Bibliothek for financial support through its Open Access Funding Programme.

ABSTRACT Cellular operators tightly monitor their networks to keep up with the market demand and frequently benchmark their performance against competitors. Typical benchmarking tests compare key performance indicators, quality of service, and quality of experience parameters on the city- and regional levels using user-collected crowdsourced data or drive test measurements. However, time-variant parameters and different user mobility patterns can bias the performance comparison. Designing a measurement sampling strategy that deals with such issues is critical for achieving a valid benchmark. Whether we would like to determine how many tiles of a map have to be measured in drive tests or how many samples we need from crowdsourced data to reach an estimate with the required accuracy, sampling theory can provide us with an answer. Since propagation conditions depend on user mobility and measurement environment, splitting the data set into groups or strata allows us to attain an unbiased estimate with fewer samples, thus allowing for a fair comparison to other mobile network operators with minimum effort measurements. In this work, we characterize the performance of different sampling methods on the simulated data set while investigating specific use cases to reveal scenarios where the stratification method pays off. We further analyze the sampling methods on two real-world crowdsourced data sets from a major Austrian operator. By stratifying the data into meaningful strata, we obtain the required number of areas and measurements in each stratum while remaining under the pre-set estimation error level. To our knowledge, this is the first study on sampling methodologies applied to real-world crowdsourced cellular measurements.

INDEX TERMS 5G, cellular network benchmarking, crowdsensing, crowdsourcing, LTE, MDT, measurements, RSRP, sampling, signal strength, stratification.

I. INTRODUCTION

In the realm of mobile network operators (MNOs), monitoring the evolution of coverage and mobile service quality is a primary task. In order to estimate MNO network performance, continuous data collection is typically performed either through extensive, time-consuming drive tests or via crowdsourced data. However, obtaining an up-to-date and unbiased performance comparison among different MNOs is constrained by time- and space-variant key performance indicators (KPIs) and quality of service (QoS) parameters, as well as differences in MNO users' mobility patterns. Therefore, designing a measurement sampling strategy that addresses these issues is essential for achieving a valid benchmark.

Regulatory bodies in each country typically establish the minimum requirements for MNOs. For example, in Austria,

The associate editor coordinating the review of this manuscript and approving it for publication was Adao Silva.

the Regulatory Authority for Broadcasting and Telecommunications (RTR) sets minimum requirements that include coverage, quality of service, and network security [1]. To ensure compliance with these requirements, MNOs seek to estimate network performance in a representative and cost-efficient manner. Sampling theory provides a framework to achieve this goal by determining the minimum number of measurements required to satisfy a particular estimation error bound. This allows MNOs to optimize their performance estimates while minimizing measurement areas, costs, time, and resources, thereby reducing their environmental impact.

In the following parts of this section, we will introduce the state of the art in MNO benchmarking and sampling methodologies. The remainder of this paper is organized as follows. Section II summarizes the related work on stratification in cellular communications and provides further motivation for this work. Section III provides a theoretical background on simple random sampling (SRS) and stratified sampling (SS) methods. Section IV exemplifies the performance difference between SRS and SS and clarifies the advantages of using stratification under specific parameter conditions. In Section V we apply the stratification methodologies on real-world minimization of drive tests (MDT) data sets. To our knowledge, this is the first paper that applies stratification in mobile communications to extensive crowdsourced data. Section VI concludes the paper and summarizes the findings.

A. STATE OF THE ART BENCHMARKING

Benchmarking cellular network coverage is a fundamental task in ensuring that mobile networks deliver reliable and consistent service to their users [2], [3]. Sampling methodologies play a pivotal role in this task by enabling network operators to obtain accurate and representative data on network performance across different geographic regions [4]. In recent years, there have been significant advances in sampling methodologies for benchmarking cellular network coverage, driven by the increasing availability of high-quality geospatial data [5] and the development of advanced adaptive sampling mechanisms [6].

1) DRIVE TESTS

One of the most widely used sampling methods for benchmarking mobile networks is drive testing, where a vehicle equipped with specialized measurement equipment travels through a geographic area and collects data on the network performance [7]. While drive tests yield highly detailed and accurate data on network performance, they are expensive and time-consuming to implement on a large scale and face challenges regarding speed dynamics and measurement repeatability [8].

Most drive test campaigns focus only on a subset of major streets in the area of interest, often disregarding other outdoor spaces, such as minor streets, parks, or squares, where satisfactory mobile network quality is equally crucial and should impact the final mean estimate. Since the measurements are limited to the on-street scenario they tend to be highly correlated, making this sampling method far from random. Therefore, computing representative mean KPI values based solely on such a subset of measurements is a challenging task.

2) CROWDSENSING

On the other hand, mobile crowdsensing involves collecting data from mobile users through specialized applications that measure network performance and location information. Today, many mobile applications allow users to trigger a measurement at the touch of a button [9], [10], [11]. The term mobile crowdsensing was introduced in [12], while also providing several examples of its applications. In [13], the concept was extended to include mobile computing, where mobile devices are used not only for the collection of data but also for its processing. However, a significant challenge associated with crowdsensing is the lack of sufficient measurements at desired locations and times. To motivate users to perform tests in locations of interest, researchers [14], [15], [16], [17], [18], [19], [20] introduce different cost functions while optimizing reward and recruitment strategies. Alternatively, Tutela [21] incentivizes app developers to integrate its software development kit into their applications, which in turn collects anonymized network data from end-users. While crowdsourced/crowdsensed[1] data can provide a wealth of information on network performance across a wide geographic area, it may be subject to biases and may not be as accurate as data collected through drive testing [22]. This type of passive monitoring involves collecting data on network performance from devices that are already in use, such as smartphones or IoT devices. While it can provide highly granular data on network performance, it can be limited by the availability of compatible devices and the need to protect user privacy [23].

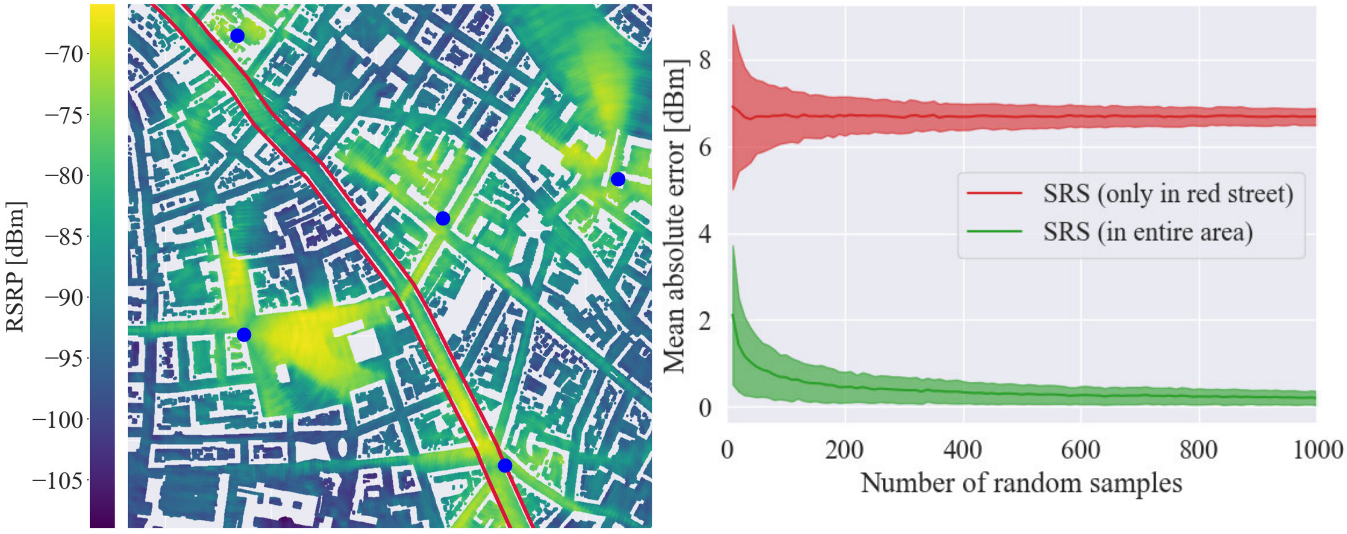
3) SIMULATED DATA: SAMPLING EXAMPLE

To demonstrate the importance of random sampling and compare its effectiveness in crowdsensed and drive test scenarios, we analyze the simulated map shown in Fig. 1(a), representing the serving reference signal received power (RSRP) for a realistic network layout in an urban area. The generated map is based on the Deep Learning Network Planner (DLNP) presented in [24] – simulation details are outlined in Appendix A. Our goal is to estimate the mean RSRP in the aforementioned area. To this end, we evaluate two sampling methods – a single street drive test campaign and the crowdsensing of the entire region, based on the following two scenarios:

- 🟥 *Drive test*: SRS of n samples only from the marked red street in Fig. 1(a) (red curve in Fig. 1(b)),
- 🟩 *Crowdsensing*: SRS of n samples from the entire RSRP data set (green curve in Fig. 1(b))

In both cases, the population mean of all simulated RSRP values (1m × 1m tile = one sample) acts as the ground truth. The number of random samples n (or 1m × 1m tiles of

[1] We use these two terms interchangeably.



(a)

(b)

FIGURE 1. (a) Simulated outdoor RSRP map in the third Vienna district. Red outlined street is the Landstrasser Hauptstrasse. Blue scatter points are the BS locations. (b) Absolute error of the mean estimation based on SRS.

the simulated map) is depicted on the x-axis in Fig. 1(b). For each n the sampling was performed in 500 independent iterations, to attain the error bounds. Shaded areas of the curves represent $[\mu_e - \sigma_e, \mu_e + \sigma_e]$, where μ_e is the mean absolute error, and σ_e is the standard deviation (SD) of the absolute error for each n.

Out of the 567 209 available samples[2] in the population, a random collection of 200 samples from the entire area is sufficient for a mean estimation well below 1 dB absolute error. Conversely, if the focus is solely on a single large street that may be covered in a measurement campaign, an error bias of nearly 7 dB is observed. The bias persists even by increasing the sample size to encompass all available samples in that street. Moreover, this analysis reveals that crowdsourced data is better suited for mean estimates in comparison to biased drive test data sets, which primarily capture the propagation characteristics of the street canyons. In real-world measurement scenarios, the advantage of crowdsourcing is even more significant as it extends beyond outdoor measurements to also include indoor measurements (which were omitted in this simulation).

The current state-of-the-art in sampling methodologies for benchmarking cellular network coverage is characterized by a wide range of approaches that can be customized to meet the specific requirements and available resources of network operators. The selection of a sampling methodology depends on several factors, such as the level of granularity needed, the geographic coverage of the study, and the available resources. Thus, adapting the measurement collection or sampling strategy to the individual propagation environments is essential to achieve the best possible results. As new technologies and analytical techniques emerge, it is probable that sampling methodologies for cellular network benchmarking will continue to progress and improve.

B. BENCHMARKING WITH SAMPLING THEORY

Sampling theory is concerned with selecting a subset of n samples out of a population with N units, such that the estimates for the population as a whole can be computed [25]. Since we deal with distinct scenarios in cellular networks, such as rural vs. urban, mobile vs. stationary, or close vs. far from the base station (BS), differing in propagation conditions, sampling theory suggests that these heterogeneous regions should be partitioned into homogeneous strata. By dividing the measurement collection into strata, it becomes possible to use fewer measurements in total while achieving the same level of accuracy for mean estimation.

Moreover, for benchmarking different MNOs, in a particular large-scale area of interest, e.g., country, we have to acquire randomly distributed measurements across that area. Provided periodic radio measurements with global positioning system (GPS) positions, collected by MDT [26] of different operators, the comparison may suffer from a user distribution bias since some operators might have more measurements in rural areas, while for others urban areas might be over-represented. Moreover, in the case of limited sample sizes, variations in network deployments by different operators may introduce bias due to potential differences in the distance to the BS. Similarly, comparing one operator's indoor measurements to another's outdoor measurements is not a valid benchmark.

To mitigate various biases, it is possible to restrict the analysis to regions where all MNOs have a sufficient number of users, as was the case for urban areas, highways,

[2] The number of samples is not rounded due to the missing tiles in the building indoor areas (outdoor simulation only).

or railways in the observed data sets. These regions can be assigned to distinct strata, and the cellular performance can be compared across the strata. Since the strata are assumed to be homogeneous and have low variance, fewer samples are required for estimating the mean within confidence intervals than when not distinguishing among the strata. In essence, the variance is reduced by dividing the data set into strata.

Stratification is a useful approach when planning a drive test campaign since it allows us to divide the area of interest into tiles or geographical units (GUs) and assign each GU to a specific stratum based on expected propagation conditions. By spatially assigning measurements to the corresponding GUs and computing the average over them, a ground truth value is assigned to each GU. This division allows us to determine the number of tiles of each stratum that must be covered to obtain an accurate mean estimate for the entire area of interest, thus reducing the time and costs associated with the measurement campaign.

II. RELATED WORK

Performance monitoring in mobile cellular networks is of utmost importance for ensuring the quality of service and network optimization. In current 5G deployments, the instantaneous performance figures are required to conduct slicing, and network orchestration [27]. In [28] and [29], the 3GPP defines the measurement methodology and KPIs for physical, medium access control (MAC), and service level. The network-wide optimization, monitoring, and root cause analysis of KPI performance are essential for operators and regulatory bodies. Generally, a specified KPI evaluation requires large-scale measurements in the form of drive tests. A common compromise is using sampled KPI values from drive tests, e.g., a set of sampled call drop rate measurements [3]. In cellular 4G and 5G networks, the RSRP as an indicator of the received signal strength, is used by regulators to evaluate coverage obligations, such as defined in [30].

In the current literature, many experimental and exploratory solutions are presented [31], [32]. However, there is a limited in-depth analysis of sampling strategies for cellular data for either minimizing the collection of samples, removing confounding, or validating existing data samples, e.g., crowdsource data.

The International Telecommunication Union (ITU) suggests that the overall measurement campaign should be split into two steps: (1) using SS to select the geographic areas to be measured, and (2) using SRS to determine the required number of samples per measured area [2], [4]. Although this approach is commonly used for drive test planning, stratification can also be used in crowdsourced scenarios to eliminate measurement bias introduced by unevenly distributed measurement points [33]. By converting the measurement data set into a two-dimensional map of GUs through grouping and averaging, the impact of the measurement distribution can be removed, providing a truthful map of the performance indicator being measured. A similar approach was presented in [34], where stratification with weights proportional to stratum areas was applied to remove bias in mean area temperature estimation.

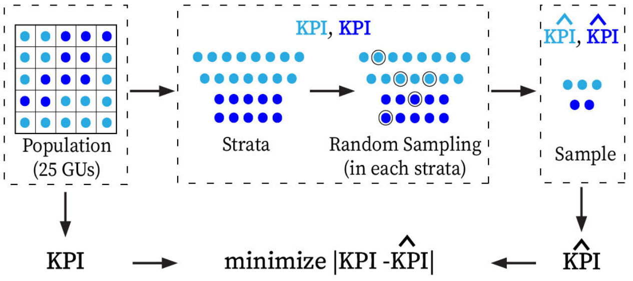


FIGURE 2. Stratification sampling of GUs in area of interest.

A. SAMPLING IN MOBILE NETWORKS

Stratification is a statistical technique utilized to reduce the estimation error bounds in comparison to a simple random sample of equivalent size [35]. The process involves partitioning large areas into smaller subregions, commonly referred to as GUs, and subsequently assigning them into predetermined strata or groups. Through the application of stratified sampling, we can determine the number of GUs required to be measured in each stratum to achieve reliable estimates, as illustrated in Fig. 2.

Initially, the area of interest is split into N tiles or GUs of equal size. We can apply the same segmentation strategy whether we aim to acquire the sample mean for the railways, car routes, or entire countries of one particular network KPI. Next, each GU is allocated to one of the predetermined strata based on predefined criteria, such as population or BS density, traffic demand, interference levels, or inter-BS distances. The assignment of a GU to a stratum needs to be unique, i.e., the strata need to be mutually exclusive. Also, note that the strata assignment needs to fulfill the criteria that the variable of interest should be homogeneous within a stratum. Correspondingly, the assignment needs to be specifically adapted to the measured KPI.

After the stratification of the measurement area, we can apply the proportional or optimal allocation method, discussed in Section III, to determine the necessary number of GUs to be measured in each stratum. Finally, the GUs to be measured in each stratum are randomly selected to ensure the validity of the results.

In simple random sampling (SRS), we draw n samples without replacement from N units of the whole population. The samples are thereby selected at random following a uniform distribution [35]. If the given population is heterogeneous, the randomly drawn samples will not faithfully represent the true parameter estimates. If our population can be divided into mutually exclusive subgroups that will take on different mean values for the variable being studied, then using stratification will allow for more precise statistical estimates that have lower variance, compared to SRS.

In stratified sampling (SS), we consider a heterogeneous population, which we split into a total of L groups or strata, such that the units in each individual stratum are as similar as possible. We aim for strata that are homogeneous within, but heterogeneous among each other [25]. A typical example of such strata can be observed in different geographic areas that exhibit different propagation conditions for wireless signals and consequently varying signal quality, for instance, rural vs. urban areas. Such an approach allows for a drastic reduction of the required samples n for a given estimation error limit.

III. SAMPLING METHODOLOGIES

Sampling methodologies in geospatial applications, such as cellular coverage, are crucial for obtaining accurate and representative data. These methodologies involve selecting a subset of locations or data points from a larger population in order to make inferences about the population as a whole. In the context of cellular coverage, sampling can be used to determine the strength and quality of the signal at different locations, which can be useful for optimizing network infrastructure and improving coverage. We introduce the following methods: simple random sampling and stratified sampling. Note that depending on the use case in Section V, the population samples discussed in Section III-A, and III-C correspond either to the KPI measurements in our data set, or the GU averages of the KPI measurement values.

A. SIMPLE RANDOM SAMPLING (SRS)

Consider a population of size N, where each sample from the population is associated with a value of the variable of interest y. The sample mean of that population μ and the sample variance σ^2 are calculated as:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} y_i, \tag{1}$$

and

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\mu)^2. \tag{2}$$

By drawing n samples at random following a uniform distribution from the given population, we can define the sample mean as:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \tag{3}$$

which is an unbiased estimator of the population mean μ [35]. Equivalently, we can also estimate the population variance σ^2, by computing the sample variance s^2:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2 . \tag{4}$$

It can again be shown, that s^2 is an unbiased estimator of σ^2 [35].

Further, we derive expressions for the expected variance of the estimators themselves. The variance of the population mean estimator $\bar{y}$ is given by:

$$\mathrm{var}(\bar{y}) = \left(\frac{N-n}{N}\right)\frac{\sigma^2}{n} \tag{5}$$

and can be estimated using the following expression:

$$\widehat{\mathrm{var}}(\bar{y}) = \left(\frac{N-n}{N}\right)\frac{s^2}{n}. \tag{6}$$

To obtain a confidence interval I for a given estimate, we select a small number α, which denotes the probability of our population mean being outside of the confidence interval I. Therefore, for the estimation of the sample mean $\bar{y}$, the population mean μ should lie in the interval I with probability $1-\alpha$, i.e., we require that

$$P(\mu \in I) = 1-\alpha, \tag{7}$$

where we consider all possible samples of size n.
Under the assumption that the population mean estimates are normally distributed under random sampling, we obtain the interval as:

$$I = \left[\bar{y} - t\sqrt{\left(\frac{N-n}{N}\right)\frac{s^2}{n}},\quad \bar{y} + t\sqrt{\left(\frac{N-n}{N}\right)\frac{s^2}{n}}\right], \tag{8}$$

where t denotes the upper $\alpha/2$ point of the Student-t distribution with $n-1$ degrees of freedom. In nearly all practical scenarios, Eq. (8) holds due to the central limit theorem. As a rule of thumb, we can replace the Student-t distribution with the standard normal distribution whenever $n > 50$. For a more detailed discussion, we refer the interested reader to [25], [35], and [36].

B. REQUIREMENTS ON SAMPLE SIZE IN SRS

Suppose that one wishes to estimate a population parameter θ, for example the population mean, with an estimator $\hat{\theta}$. With a basic understanding of SRS, we can address the question of determining the required sample size n for an estimate with a predefined accuracy.

We specify the maximum allowable absolute or relative error d between the estimate and the true value, while allowing for a small probability α that the error may exceed threshold d. We define these probabilities as being less than α for absolute and relative error respectively as:

$$\alpha > \begin{cases} P\left(|\hat{\theta}-\theta| > d\right) & \text{for absolute error,} \\ P\left(|\hat{\theta}-\theta| > d|\theta|\right) & \text{for relative error.} \end{cases} \tag{9}$$

Under the assumption of an unbiased normally distributed estimator, the distribution of error normalized by the square root of estimator variance

$$\frac{\hat{\theta}-\theta}{\sqrt{\operatorname{var}(\hat{\theta})}} \tag{10}$$

approaches a standard normal distribution for large n. This allows us to reformulate the inequalities from Eq. (9) into Eq. (11), by introducing z as the upper $\alpha/2$ point of the standard normal distribution:

$$P\left(\frac{|\hat{\theta}-\theta|}{\sqrt{\operatorname{var}(\hat{\theta})}} > z\right) = P\left(|\hat{\theta}-\theta| > z\sqrt{\operatorname{var}(\hat{\theta})}\right) = \alpha. \tag{11}$$

For the case of the population mean estimator under SRS, we have $\theta = \mu$ and $\hat{\theta} = \bar{y}$, we thus need to solve:

$$z\sqrt{\operatorname{var}(\bar{y})} = \begin{cases} d & \text{for absolute error,} \\ d|\mu| & \text{for relative error,} \end{cases} \tag{12}$$

with the variance of our population mean estimator specified in Eq. (5). Solving Eq. (12) for n gives us:

$$n = \frac{1}{1/n_0 + 1/N} := n_{\text{srs}}, \tag{13}$$

with

$$n_0 = \begin{cases} \frac{z^2\sigma^2}{d^2} & \text{for absolute error,} \\ \frac{z^2\sigma^2}{d^2\mu^2} & \text{for relative error,} \end{cases} \tag{14}$$

for absolute and relative error respectively. To distinguish n in SRS from n in Section III-D, we denote it as n_{srs}.
Note, that the main challenge in this setup is the selection of the population variance σ^2, which has to be done beforehand.

C. STRATIFIED SAMPLING (SS)

Now we consider a heterogeneous population, which we split into a total of L groups or strata, such that the samples in each individual stratum are as similar as possible. We denote the number of population samples per stratum h as N_h, such that the total number of samples in the population is given by:

$$N = \sum_{h=1}^{L} N_h. \tag{15}$$

Equivalently, we define n_h, as the number of randomly drawn samples from stratum h — then the total number of drawn samples is given by:

$$n = \sum_{h=1}^{L} n_h. \tag{16}$$

Further, we introduce the mean in a stratum h as μ_h:

$$\mu_h = \frac{1}{N_h}\sum_{i=1}^{N_h} y_i. \tag{17}$$

Depending on our use case, y_i represents either the KPI of interest in the i^{th} GU or the i^{th} KPI measurement sample.

By summarizing the estimates for the individual strata we can derive an unbiased estimator of the population mean μ. We denote this estimate as the stratified sample mean $\bar{y}_{\text{st}}$, that is given by:

$$\bar{y}_{\text{st}} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{y}_h. \tag{18}$$

Note, that Eq. (18) assumes SRS estimates $\bar{y}_h$ for each stratum.

The variance of the estimator in Eq. (18) is given by:

$$\operatorname{var}(\bar{y}_{\text{st}}) = \sum_{h=1}^{L}\left(\frac{N_h}{N}\right)^2\left(\frac{N_h-n_h}{N_h}\right)\frac{\sigma_h^2}{n_h}. \tag{19}$$

The corresponding estimator of this variance can be derived by replacing the population variance σ^2 with the sample variance s^2 for each stratum:

$$\widehat{\operatorname{var}}(\bar{y}_{\text{st}}) = \sum_{h=1}^{L}\left(\frac{N_h}{N}\right)^2\left(\frac{N_h-n_h}{N_h}\right)\frac{s_h^2}{n_h}. \tag{20}$$

Given the total number of samples to be drawn n, we need to specify how to allocate these samples across different strata. In the following, we distinguish between two different types of allocation:

1) *Proportional Allocation*
If the strata differ in size, the proportional allocation could be used to maintain a steady sampling fraction throughout the population. Here, we simply select the number of samples in accordance with the overall units N_h per stratum such that:

$$n_h = \frac{nN_h}{N}. \tag{21}$$

2) *Optimal Allocation*
Optimal allocation results in the population mean estimate with the lowest variance for a fixed total number of samples n. Here, we have for the number of samples per stratum h:

$$n_h = \frac{nN_h\sigma_h}{\sum_{k=1}^{L} N_k\sigma_k}. \tag{22}$$

For this, we again have to estimate the stratum variances σ_h^2 in advance. Typically, this is done with past data.

D. REQUIREMENTS ON SAMPLE SIZE IN SS

Similar to SRS, we first specify the maximum allowable error d between the estimate and the true value, while allowing for a small probability α that the error may exceed the threshold d. For the estimation of the sample mean $\overline{\mu}_{\text{st}}$, this means that the population mean μ should lie in interval I with probability $1-\alpha$, i.e., we require that

$$\mathrm{P}(\mu \in I) = 1-\alpha, \tag{23}$$

where we consider all possible samples of size n.

Assuming $\bar{y}_{\mathrm{st}}$ to be normally distributed, using the central limit theorem the confidence interval can be constructed as follows:

$$I = \left[\bar{y}_{\mathrm{st}} - z\sqrt{\operatorname{var}\left(\bar{y}_{\mathrm{st}}\right)}, \bar{y}_{\mathrm{st}} + z\sqrt{\operatorname{var}\left(\bar{y}_{\mathrm{st}}\right)}\right], \tag{24}$$

where z is the upper $\alpha/2$ point of the standard normal distribution. For a more detailed discussion, we refer the interested reader to [25] and [36].

In analogy with Eqs. (9), and (10) we can derive the required sample size for fixed choices of d and α. W.l.o.g. we define $d_{\mathrm{abs}} := d$ and $d_{\mathrm{rel}} := d|\mu|$, and solve for $d_{\mathrm{X}} = z \cdot \sqrt{\operatorname{var}\left(\bar{y}_{\mathrm{st}}\right)}$, where d_{X} denotes either d_{abs} or d_{rel}. Thereby, we insert the expression for variance from Eq. 19 and corresponding allocation according to Eqs. (21), and (22), to get for:

1) *Proportional Allocation*

$$\begin{aligned} \frac{d_{\mathrm{X}}^2}{z^2}N^2 + \sum_{h=1}^{L} N_h\sigma_h^2 &= \sum_{h=1}^{L} \frac{N_h^2\sigma_h^2}{n_h} \\ &= \frac{1}{n} \cdot \sum_{h=1}^{L} N_h\sigma_h^2\, N. \end{aligned} \tag{25}$$

Hence, the required number n to achieve the error bound d with probability $1-\alpha$ using proportional allocation is given by:

$$n = \frac{\sum_{h=1}^{L} N_h\sigma_h^2 N}{\frac{d_{\mathrm{X}}^2}{z^2}N^2 + \sum_{h=1}^{L} N_h\sigma_h^2} := n_{\mathrm{prop}}. \tag{26}$$

2) *Optimal Allocation*

$$\begin{aligned} \frac{d_{\mathrm{X}}^2}{z^2}N^2 &= \sum_{h=1}^{L} N_h^2\left(\frac{N_h - n_h}{N_h}\right)\frac{\sigma_h^2}{n_h} \\ &= \sum_{h=1}^{L} \frac{N_h^2\sigma_h^2}{n_h} - \sum_{h=1}^{L} N_h\sigma_h^2, \\ \frac{d_{\mathrm{X}}^2}{z^2}N^2 + \sum_{h=1}^{L} N_h\sigma_h^2 &= \sum_{h=1}^{L} \frac{N_h^2\sigma_h^2}{n_h} \\ &= \frac{1}{n} \cdot \sum_{h=1}^{L} N_h\sigma_h \cdot \left(\sum_{k=1}^{L} N_k\sigma_k\right) \\ &= \frac{1}{n} \cdot \left(\sum_{h=1}^{L} N_h\sigma_h\right)^2. \end{aligned} \tag{27}$$

Hence, the required number n to achieve the error bound d with probability $1-\alpha$ using optimal allocation is given by:

$$n = \frac{\left(\sum_{h=1}^{L} N_h\sigma_h\right)^2}{\frac{d_{\mathrm{X}}^2}{z^2}N^2 + \sum_{h=1}^{L} N_h\sigma_h^2} := n_{\mathrm{opt}}. \tag{28}$$

Finally, by substituting d_{X} with d_{abs} or d_{rel} in Eqs. (26) and (28), depending on the error measure in question, the number of samples (n_{prop}, n_{opt}) required to remain below the given error bound is calculated. As was the case in SRS, the main challenge in stratification setup is the selection of the strata variances σ_h^2, which has to be done beforehand.

The main reason for using SS is that it requires a smaller n to achieve the same error bound, i.e., the required n is typically significantly smaller than what we obtain for SRS. Thereby allowing the MNOs to achieve accurate estimates of the overall network quality while covering fewer GUs or having fewer measurement samples. In the following section, we investigate the influence of the strata means, variances, and sizes on the performance among SRS, SS with proportional allocation, and SS with optimal allocation.

IV. SAMPLING OF SIMULATED DATA SET

Three parameters influence the sampling performance with respect to the given error bound d:

- strata means μ_h,
- strata variances σ_h^2 and
- strata sizes N_h,

where $h = 1, 2, \ldots, L$. After extensive simulation analysis, we have found that the key parameter influencing the required sample size for each method is the stratum variance. The higher the variance per stratum (or population in the case of SRS), the larger the sample size required to achieve the desired level of precision. Furthermore, stratified sampling outperforms simple random sampling when there is a significant difference in the strata means. In terms of stratum size, optimal performance is achieved when the strata are of the same order of magnitude. However, as stratum sizes become increasingly unbalanced, with a majority of samples concentrated in a single stratum, the performance of all three methods becomes almost indistinguishable. In Section IV-A, IV-B we examine the impact of these parameters on sampling performance by analyzing a general use case and two boundary special cases to gain a comprehensive understanding of their influence.

Looking at Eqs. (13), (26), and (28), a causality dilemma arises, as the variance of either the total population or each stratum must be known or properly estimated before calculating the required number of samples for a given acceptable error. Thus, if we wish to determine the number of samples in yet unseen regions, we must estimate variances based on similar previously measured areas. Despite this challenge, by simulating an artificial data set and examining sampling performance, we can gain an understanding of how these parameters influence the sampling process. Depending on the use case, such an artificial sample corresponds either to a single measurement or to the measurement average in a single GU. In Section V we will resolve the two cases, however, the following general simulation results apply to both.

A. SAMPLING PERFORMANCE: GENERAL USE CASE

Using a set of parameters $S = \{\mu_h, \sigma_h, N_h\}$ we generate an artificial stratified data set, by drawing samples from a

Normal distribution with mean μ_h and standard deviation σ_h of size N_h for each stratum, finally combining these strata into a full artificial data set. Here we wish to point out, that different distributions can be used for the data set generation, e.g., Student-T, Gumball, and even multi-modal Gaussian, as long as the sample mean estimate is approximately normally distributed.

W.l.o.g., we show the results simulated using Normal distribution. The simulation parameters we used for generating an artificial RSRP data set consisting of two strata are provided in Table. 1 (general use case). Next, we choose $\alpha = 0.05$ and generate a range for acceptable sampling error d, which is represented on the x-axis of Fig. 3. For each of the d-values, we use Eqs. (13), (26) and (28) to calculate the required number of samples for each sampling method (solid curves in Fig. 3). Thereby we assume perfect knowledge of N_h and σ_h for both strata – something we will need to estimate for real-world data. For proportional and optimal SS we further use Eqs. (21) and (22) respectively, to calculate the number of samples required in each stratum (dashed curves in Fig. 3). Horizontal dashed lines illustrate N_h – the population total for each stratum. Considering Fig. 3(a), assume we want to predict the mean value of the simulated RSRP data set. If we are willing to accept that the absolute error between the mean estimate (calculated from the sample) and the population mean is in 95% of the cases $(1-\alpha)$ below $d = 10^{-1}$, then we require $n_{\mathrm{SRS}} \approx 11\,800$ samples (≈90%) using the SRS method. On the other hand, using the proportional stratification method, this number reduces to $n_{\mathrm{prop}} \approx 10\,600$ (≈80%), while the optimal stratification method requires only $n_{\mathrm{opt}} \approx 9\,000$ (less than 70%) samples in total (solid lines in Fig. 3(a)). While in proportional stratification the number of samples in each stratum proportionally rises until reaching its boundary at respective N_h (red dashed curves), the optimal scheme exploits the knowledge of strata variances.

For lower strata variance, fewer samples are required for accurate stratum mean estimation (*S2 SS opt* curve in Fig. 3(a)). However, if the variance in a stratum is too high, the algorithm would require more samples than we have available (*S1 SS opt* curve overshoots the N_1 level). This might represent a problem if we are, for instance, trying to determine how many tiles need to be measured in the area of interest, as we cannot simply introduce more tiles in that area. However, if we are determining the number of required measurements, we merely would have to measure more in those areas where the variance is high. The dotted part of the n_{opt} (*total SS opt* curve in Fig. 3(a)) curve symbolizes the overshot of one of the stratum curves.

For verifying that under SRS of n_h samples in each stratum, we truly remain under set estimate error level d in $(1-\alpha)100\%$ of the cases, we apply the verification algorithm provided in Appendix C.

B. SPECIAL CASES

To help us understand when the increased complexity of stratification proves beneficial and when it is preferable to stay in the simpler SRS domain, we consider the following specific scenarios.

Scenario I: equal variances

Let us assume all stata variances are equal, i.e., the SD are equal $\sigma_1 = \sigma_2 = \cdots = \sigma_L := \sigma_s$, then we can simplify Eq. (26) and (28) to:

$$n_{\mathrm{prop}} = \frac{\sum_{h=1}^{L} N_h \sigma_h^2 N}{\frac{d_{\mathrm{X}}^2}{z^2} N^2 + \sum_{h=1}^{L} N_h \sigma_h^2} = \frac{N \sigma_s^2 \sum_{h=1}^{L} N_h}{\frac{d_{\mathrm{X}}^2}{z^2} N^2 + \sigma_s^2 \sum_{h=1}^{L} N_h} = \frac{\sigma_s^2 N^2}{\frac{d_{\mathrm{X}}^2}{z^2} N^2 + \sigma_s^2 N}, \tag{29}$$

$$n_{\mathrm{opt}} = \frac{\left(\sum_{h=1}^{L} N_h \sigma_h\right)^2}{\frac{d_{\mathrm{X}}^2}{z^2} N^2 + \sum_{h=1}^{L} N_h \sigma_h^2} = \frac{\sigma_s^2 \left(\sum_{h=1}^{L} N_h\right)^2}{\frac{d_{\mathrm{X}}^2}{z^2} N^2 + \sigma_s^2 \sum_{h=1}^{L} N_h} = \frac{\sigma_s^2 N^2}{\frac{d_{\mathrm{X}}^2}{z^2} N^2 + \sigma_s^2 N}. \tag{30}$$

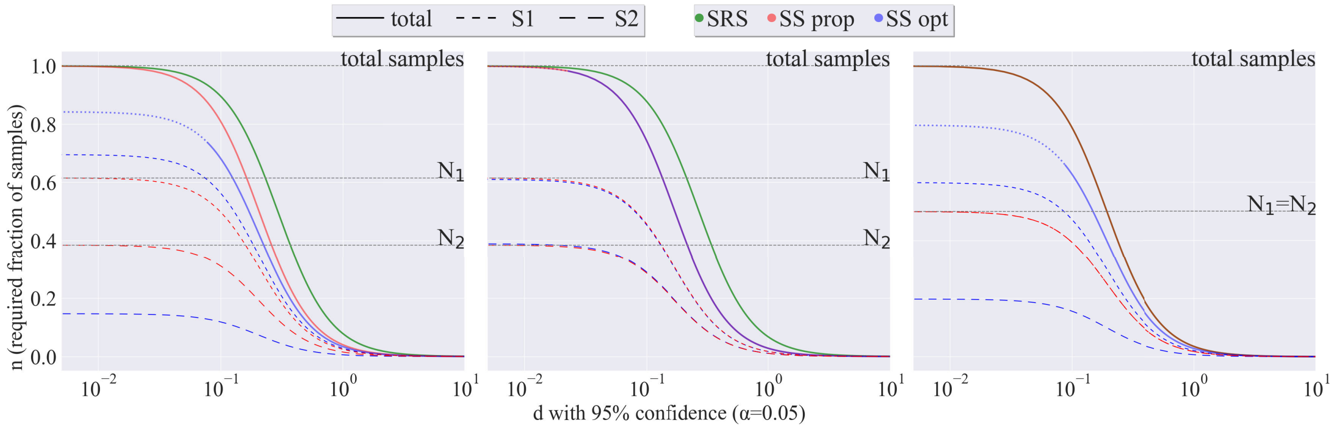
From Eqs. (29) and (30), we notice that n_{prop} and n_{opt} are equal as long as there is no difference in variance among strata. This is also supported by the results of our simulations, as illustrated in Fig. 3(b) and based on the simulation parameters outlined in Table. 1 (equal strata variances). In the simulation, two strata were employed, each with differing mean values and sizes but with equal variances. The maximum available samples in each stratum from which sampling can be conducted are depicted by horizontal dashed lines. The curves portray the number of samples required to remain below the absolute error bound d. In situations where the variances of the strata are identical, the n_{prop} and n_{opt} curves coincide, and both outperform the SRS approach. Thus, the use of the optimal allocation scheme, which requires greater computational effort than the simpler proportional allocation scheme, is unnecessary when strata variances are known to be equal. However, employing stratification in cases where the strata means and sizes differ surpasses the simple SRS approach.

Scenario II: equal means and sizes

Let us now assume that strata variances are not equal, but strata means and strata sizes are, i.e., $\mu_1 = \mu_2 = \cdots = \mu_L$ and $N_1 = N_2 = \cdots = N_L = \frac{N}{L}$. By substituting the second condition into Eq. (26) we get:

$$n_{\mathrm{prop}} = \frac{\sum_{h=1}^{L} N_h \sigma_h^2 N}{\frac{d_{\mathrm{X}}^2}{z^2} N^2 + \sum_{h=1}^{L} N_h \sigma_h^2} = \frac{\sum_{h=1}^{L} \frac{N}{L} N \sigma_h^2}{\frac{d_{\mathrm{X}}^2}{z^2} N^2 + \frac{N}{L} \sum_{h=1}^{L} \sigma_h^2} = \frac{1}{\frac{d_{\mathrm{X}}^2}{z^2} \frac{L}{\sum_{h=1}^{L} \sigma_h^2} + \frac{1}{N}}. \tag{31}$$

Since the means μ_h and strata sizes N_h are equal, we can prove that under the assumption of large N_h, the averaged variances of different strata will amount to the variance of the entire



(a) (b) (c)

FIGURE 3. Simulation results for two strata (S1 and S2) for a general use case (a) and two special use cases, namely strata with equal variances (b) and equal strata means and sizes (c). N_1 and N_2 correspond to the total number of samples in S1 and S2 respectively, while their sum $N_1 + N_2 = N$ corresponds to *total samples* that we can sample from. Note that all curves and vertical lines are normalized by N to represent n as a fraction of total samples N.

TABLE 1. Simulation parameters for three use cases represented in Fig. 3 in analogous order. The abbreviation SD stands for standard deviation and is equal to the squared root of the variance.

GENERAL USE CASE		EQUAL STRATA VARIANCES		EQUAL STRATA MEANS AND SIZES	
stratum 1 (S1)	stratum 2 (S2)	stratum 1 (S1)	stratum 2 (S2)	stratum 1 (S1)	stratum 2 (S2)
mean μ_1 = -110 dBm SD σ_1 = 15 dBm size N_1 = 8000	mean μ_2 = -85 dBm SD σ_2 = 5 dBm size N_2 = 5000	mean μ_1 = -110 dBm SD σ_1 = 10 dBm size N_1 = 8000	mean μ_2 = -85 dBm SD σ_2 = 10 dBm size N_2 = 5000	mean μ_1 = -95 dBm SD σ_1 = 15 dBm size N_1 = 6500	mean μ_2 = -95 dBm SD σ_2 = 5 dBm size N_2 = 6500

combined sample (see Appendix B), i.e.:

$$\sigma^2 = \frac{\sum_{h=1}^{L} \sigma_h^2}{L}. \tag{32}$$

By inserting Eq. (32) into Eq. (31) the SRS result from Eq. (13) is again obtained:

$$n_{\text{prop}} = \frac{1}{\frac{d_X^2}{z^2\sigma^2} + \frac{1}{N}} = n_{\text{srs}}. \tag{33}$$

It follows that the proportional stratification provides no benefit to SRS in cases where the means and sizes of strata are identical. Hence, in such a scenario, proportional stratification is unnecessary as it does not improve estimate accuracy while only bringing higher computational costs. The optimal allocation remains the only viable option that can bring further improvement, as it still depends on different strata variances.

This behavior is also noticed in our simulation results in Fig. 3(c) with two strata S1 and S2. Simulation parameters are given in Table. 1 (equal strata means and sizes). Here, the n_{srs} and n_{prop} curves overlap. However, n_{opt} outperforms them both as it requires fewer samples in total to achieve the same estimate. The dotted part of n_{opt}, where $d < 10^{-1}$, results from high variance in strata S2. The optimal scheme here requires more sampled segments than we have available in that strata, making such sampling unfeasible. However, for $d > 10^{-1}$ the performance of the optimal scheme is better than the proportional one. Thus, depending on the mean estimate error d we are willing to accept, we can use simulations to choose the sampling scheme for a specific use case at hand.

TABLE 2. MDT data sets.

	Vienna	Austria
Frequency [MHz]	800/1800/2600	800/1800/2600
Duration	8 days	3 days
Area [km^2]	414	83 871
Total samples	10 000 000	10 000 000
Filtered samples	5 771 426	4 592 394
BS type	Macro only	Macro only
Number of cells	693	3 468

Assuming further equal variances among the strata, the methods of SRS, SS with optimal allocation, and SS with proportional allocation can be reduced to a common formula. The reduction results in the same performance for all three methods in terms of error bounds. However, the methods still differ in their complexity. In such cases, it is recommended to use the simplest method, i.e., the SRS.

V. SAMPLING THE CROWDSOURCED MDT DATA SET

In this section, we apply the sampling methodologies to a real-world MDT data set. First, we describe the MDT measurements in Section V-A. Then, we apply and compare the

TABLE 3. Austria, RSRP statistics per frequency band.

frequency	800 MHz	1800 MHz	2600 MHz
RSRP mean	-98.66 dBm	-94.96 dBm	-95.48 dBm
RSRP SD	12.33 dBm	10.95 dBm	11.29 dBm
measurement count	1 476 179	2 560 008	556 207

TABLE 4. Vienna, RSRP statistics per frequency band.

frequency	800 MHz	1800 MHz	2600 MHz
RSRP mean	-96.77 dBm	-93.88 dBm	-96.21 dBm
RSRP SD	11.85 dBm	10.87 dBm	10.73 dBm
measurement count	1 100 826	3 546 656	1 123 944

sampling methods to MDT data on a measurement level in Section V-B and on a GU-level in Section V-C.

A. MDT MEASUREMENTS DATA SET

The MDT data sets evaluated in this study were provided to us by a major MNO in Austria. The measurements were collected in a live long-term evolution (LTE) network on a large number of end-user consumer-grade devices through a dedicated Android application using the Netscout Geo-Analytics tool for data collection [37]. The first data set is gathered in the city of Vienna, Austria, consisting of a total of 10 000 000 samples. The second data set contains 10 000 000 samples in the entire country of Austria in the same network. Features of both data sets are summarized in Tables. 2 to 4.

Both data sets contain only measurements from the user equipments (UEs) connected to LTE macro BSs in the area of interest. Before applying sampling algorithms, we first clean up the data, by removing all measurements with erroneous or inaccurate attributes. To this end, we apply the following data processing:

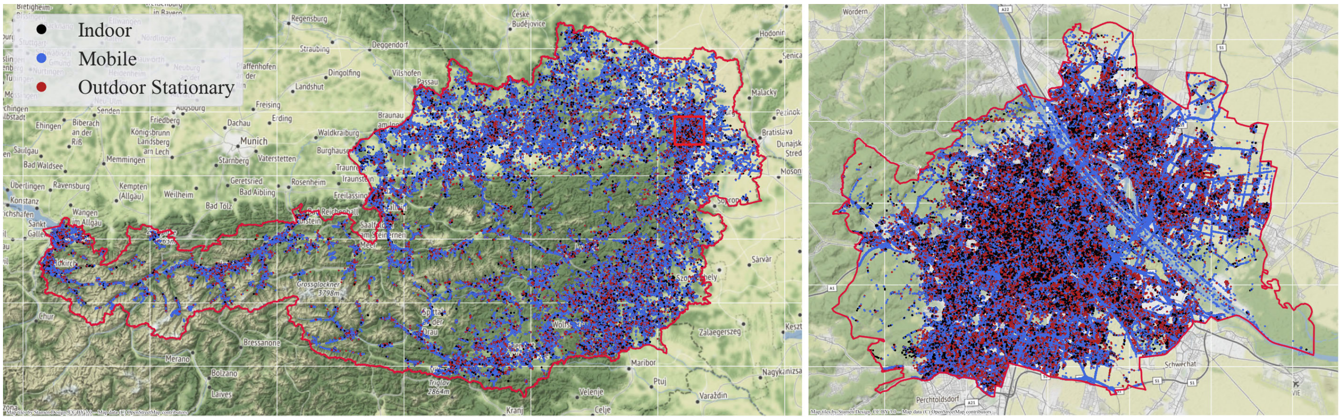
- Using the reported GPS location, we filter out the points that are outside of the area of interest, even though they are connected to the BS in the area. We are interested in predicting the mean KPI in the area of interest only. To achieve this, we acquire country and city borders from OpenStreetMap (OSM) using Python Overpass API [38] and use Python Geopandas library [39] for geospatial manipulations and filtering.
- Next, we filter out all points where the in meters reported timing advance (TA) is smaller than the calculated line-of-sight (LOS) Euclidean distance to the BS. TA corresponds to the time a signal takes to reach the BS from a mobile phone. The BS can use precise arrival time to determine the distance to the mobile phone [40]. Since it is physically impossible for the TA values to be smaller than the minimum LOS distance, we exclude such measurement points from our data sets. For calculating the distance to the serving BS, we use precise BS locations in Austria, provided by the MNO.
- Since indoor/outdoor mapping depends on the GPS accuracy, we remove the measurements with imprecise GPS values, by filtering based on the reported GPS uncertainty (in meters). In particular, we limit our analysis only to measurements with a reported GPS uncertainty below 45m.
- Finally, we split the data sets based on frequency bands and look at each band individually.

To test and compare different sampling methodologies on each of the MDT data sets, we have to split the data into sensible strata. Our goal is to obtain an accurate mean RSRP estimate with a minimum number of measurements and/or a minimum number of GUs used for the mean estimate calculation in each method. We investigate both alternatives in the following sections.

B. REQUIRED NUMBER OF MEASUREMENTS

In a real-world context, effectively leveraging the benefits of stratification necessitates identifying the specific conditions that strongly influence the reported KPI values. To achieve a suitable separation of strata, it is important to base the split on the statistical characteristics of the KPI in question. For the case of RSRP, one possible approach is splitting the data into groups based on factors such as radio conditions, position, and motion of the UE, since the reported value correlates with these parameters. The resulting groups exhibit distinctive RSRP statistics with different means and smaller standard variations when compared to the full data set. On the other hand, for other KPIs such as reported throughput value, the relevant parameters for the stratification split decision are expected to be user tariffs, available bandwidth, and cell load. If there is no prior knowledge regarding the statistics and dependencies of the KPI, clustering methods can be employed to identify underlying correlations in the data set and determine the groups or strata. In a practical realization, the split can be based on the various regions of interest (e.g. tunnels, stadiums, stations, industry sites, rural and urban areas) that share common propagation conditions.

Fig. 4 depicts the filtered measurement data sets. Classification to *Indoor*, *Mobile*, and *Outdoor Stationary* measurements is provided by the operator and is based on radio conditions at the time of the measurement, as well as the GPS details. We use this classification to separate our measurement data sets into three corresponding groups or strata. Figs. 5 and 6 provide more detailed characteristics of the measurement data sets in the form of boxenplots [41]. Since we are only considering UEs connected to LTE macro BS, the lowest mean RSRP value is observed in the *Indoor* scenario, due to high building penetration loss (BPL) [42], [43]. UEs in *Mobile* scenario experience many physical cell identity (PCI) changes and handovers (HOs) with lower signal qualities on the cell edges, while *Outdoor Stationary* case ensures the best signal quality, as it does not have to deal with HOs and BPL. This trend is visible in Fig. 5 for both Austria and Vienna data sets, where indicated values represent the mean in the particular stratum. The speed profiles in Fig. 6 further



(a) Austria

(b) Vienna, Austria

FIGURE 4. Filtered MDT measurement data sets of a large MNO in Austria. In the map on the left, the red square marks the city of Vienna. Different colors represent TrueCall Netscouts' radio condition-dependent *environment* classification, with the following categories: *Indoor*, *Mobile*, and *Outdoor Stationary*. We use these classes as three separate strata in the following.

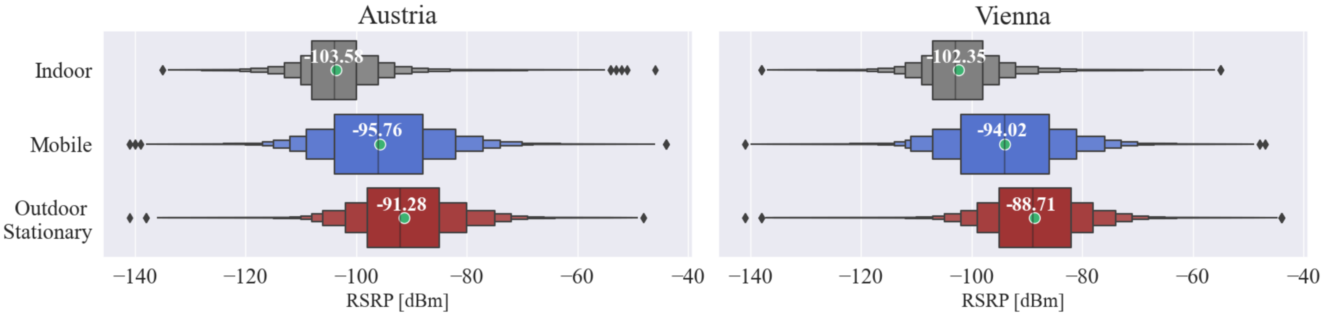


FIGURE 5. RSRP distribution per environment and data set in 1800 MHz frequency band. Marked green points represent the mean RSRP values. Strata tendencies observed in the 800 and 2600 MHz bands are similar.

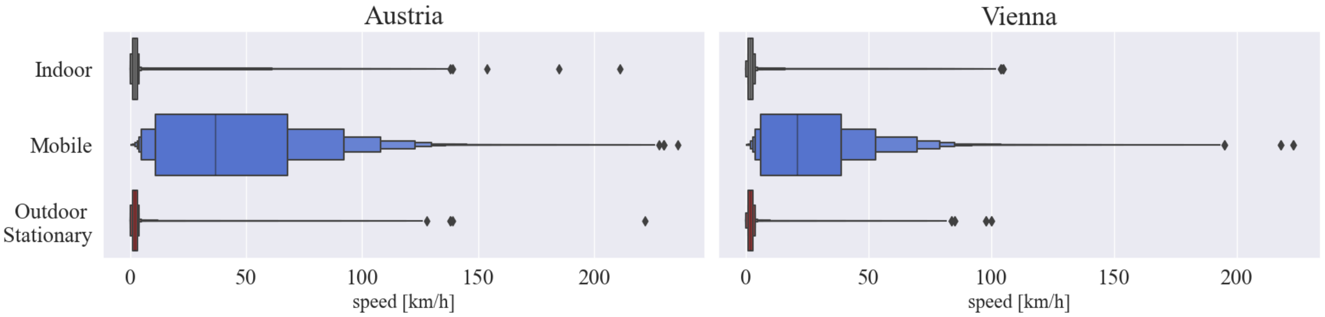


FIGURE 6. Speed profile per environment strata and data set for all three frequency bands combined.

illustrate that only *Mobile* stratum has average speeds higher than 5 km/h, with few outliers in both static strata. We can also notice the speed difference between the country and city data set, with the city (right) having lower average speeds compared to the country (left) as expected.

We use these three strata, to calculate how many measurements from each group and in total are required to achieve a mean RSRP estimate with an error below level d in over 95% of the cases ($\alpha = 0.05$). The statistics of each stratum are provided in Table. 5.

Fig. 7 shows the comparison between different sampling methods. The x-axis represents the acceptable error d, while the y-axis indicates how many measurements are required per strata (dashed) and in total (solid curves) to achieve the mean RSRP estimate under the error bound d with 95% accuracy. Notice that in Table. 5 for both data sets, the *Mobile* stratum has the highest SD. Therefore, to achieve the estimation error of less than 10^{-2} dB we require more samples than we have available in that stratum in our crowdsourced data sets, which is why we have an overshoot over N_2 level in both

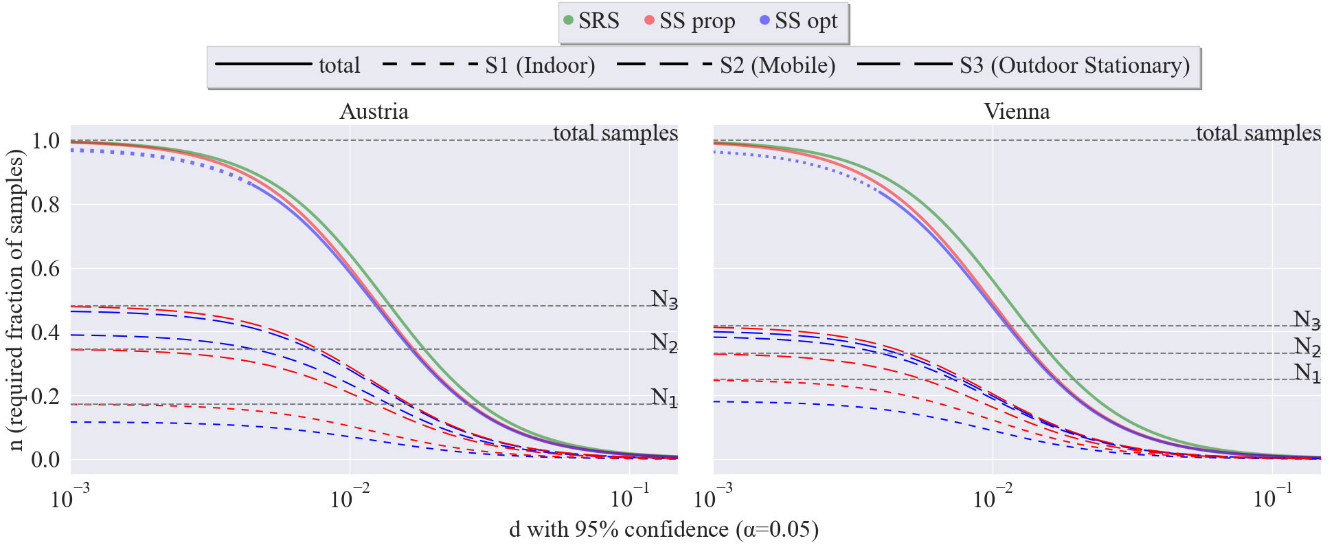


FIGURE 7. Required fraction of samples *n* plotted over acceptable estimation error *d* for different sampling schemes in Austria (left) and Vienna (right) data set (1800 MHz), based on radio-condition and speed-dependent strata split. All curves are normalized by *N*.

TABLE 5. Strata RSRP statistics for both data sets in the 1800 MHz frequency band.

Austria (1800 MHz)			Vienna (1800 MHz)		
Indoor	Mobile	Outdoor Stationary	Indoor	Mobile	Outdoor Stationary
μ_1 = -103.5 dBm	μ_2 = -95.76 dBm	μ_3 = -91.28 dBm	μ_1 = -102.35 dBm	μ_2 = -94.02 dBm	μ_3 = -88.71 dBm
σ_1 = 6.87 dBm	σ_2 = 11.49 dBm	σ_3 = 9.82 dBm	σ_1 = 6.99 dBm	σ_2 = 11.12 dBm	σ_3 = 9.25 dBm
N_1 = 442 864	N_2 = 884 781	N_3 = 1 232 363	N_1 = 885 728	N_2 = 1 180 224	N_3 = 1 480 704

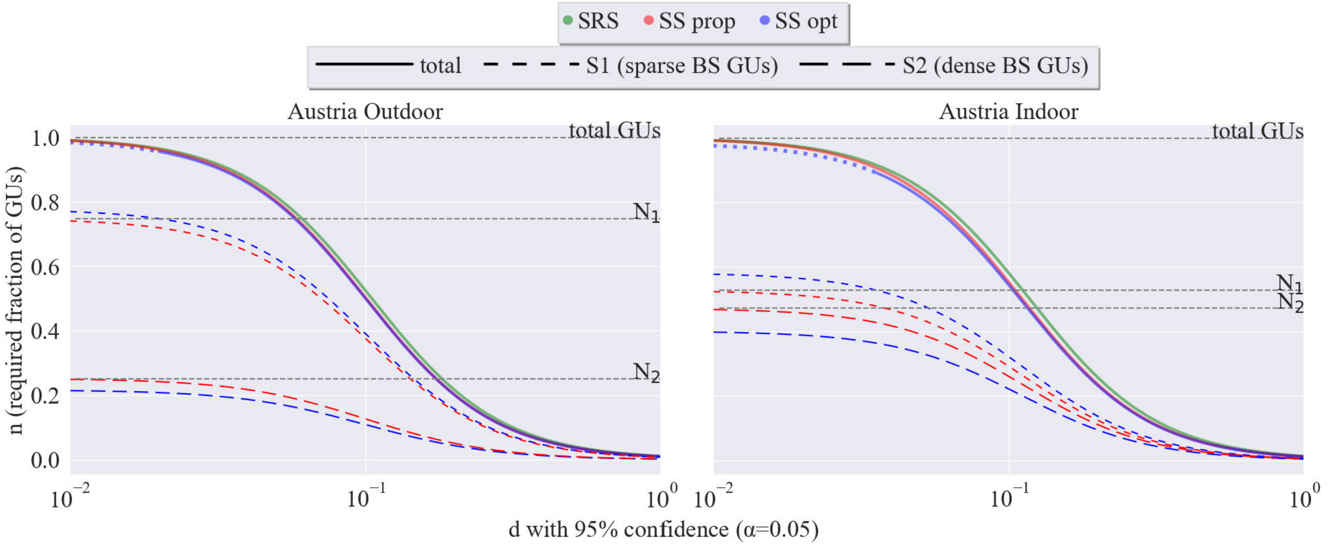


FIGURE 8. Required fraction of GUs *n* plotted over acceptable estimation error *d* for different sampling schemes in Austria outdoor (left) and indoor (right) data set (1800 MHz), based on BS-density strata split. All curves are normalized by *N*.

cases for *optimal* allocation scheme. On the other hand, the *Indoor* stratum has the lowest SD in both data sets. Therefore, the *optimal* scheme does not require all available *Indoor* measurements for any requested estimation error level and remains well below the N_1 level. On the total (solid lines), *proportional* (red) and *optimal* (blue) allocation schemes performed almost identically since the difference among the SD levels in all three strata was much smaller compared to our simulated data set in Section IV. However, both stratification schemes outperform the SRS (green). Considering that the

TABLE 6. Strata RSRP statistics for outdoor (left) and indoor (right) Austria data set in the 1800 MHz frequency band.

Austria Outdoor (1800 MHz)		Austria Indoor (1800 MHz)	
S1 (*BS-sparse* GUs)	S2 (*BS-dense* GUs)	S1 (*BS-sparse* GUs)	S2 (*BS-dense* GUs)
μ_1 = -102.75 dBm	μ_2 = -96.49 dBm	μ_1 = -109.29 dBm	μ_2 = -105.11 dBm
σ_1 = 9.41 dBm	σ_2 = 7.79 dBm	σ_1 = 6.62 dBm	σ_2 = 5.10 dBm
N_1 = 23 102	N_2 = 7 778	N_1 = 5 791	N_2 = 5 177

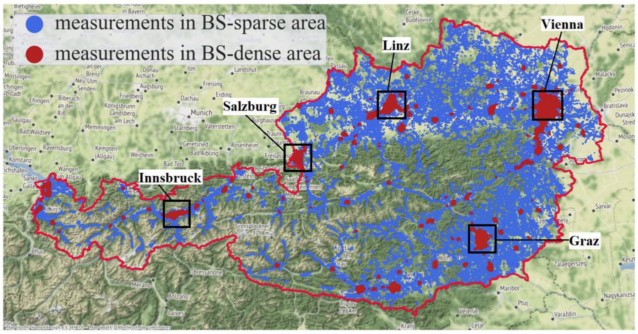

● measurements in BS-sparse area
● measurements in BS-dense area
Linz
Vienna
Salzburg
Innsbruck
Graz

FIGURE 9. Filtered MDT measurement data set of a large MNO in Austria. Different colors represent measurement classification based on the BS-density in a 2 km radius of the measurement GU. *BS-dense* deployments in red overlap with larger cities in Austria (denoted by black squares), while the blue corresponds to more suburban and rural areas.

performances of proportional SS and SRS would overlap for equal strata means and sizes, we notice that this difference originates in the discrepancy among different strata means. However, since their discrepancy is not as high as in the simulated scenario, we see only a minor advantage to SRS, which is valid for both city and country levels. Note, however, by finding an even better-suited strata split, than the one the operator is providing, the strata mean differences may increase, such that stratification schemes show a higher advantage to SRS than the one we are currently seeing. The findings demonstrate that, in the present configuration, obtaining a mean RSRP estimate with a mean absolute estimation error lower than 10^{-2} dB requires the utilization of 50-65% of the available measurements in both data sets. The specified threshold for mean absolute estimation error of 10^{-2} dB was chosen as an indication of a substantially high degree of accuracy in the RSRP estimation.

C. REQUIRED NUMBER OF GUs

Given a measurement data set, whether it is crowdsourced, drive- or train-test data, it often happens that measurements get accumulated in certain areas, and are infrequent in others. For illustration, a driving train in an ongoing measurement campaign makes longer stops at train stations, while on some parts of the tracks, it reaches speeds of 250 km/h. Under such conditions measurement data set gets very confounded in time [8]. To solve this problem, we can bin the measurement data into GUs, take the average, and get representative KPI values per train length, independent of the train speed [44]. To remove measurement bias created by having more measurements in very crowded areas, compared to very few measurements in suburban/rural regions, we can apply the same binning strategy with crowdsourced data and then work with GU averages instead of the measurement samples.

Stratification sampling can also be used to determine which environment types and in what amount should be covered in a measurement campaign. For instance, rural and urban regions have different BS deployments, in terms of BS density, propagation loss, and LOS connectivity. This fact is particularly relevant in developing countries, where the population is more concentrated in rural areas despite poorer BS deployment [45]. To have a better understanding of the overall network quality in such conditions, and to be able to plan a measurement campaign more efficiently, we can split the area of interest, e.g. country, into fixed-sized GUs, and assign rural or urban property (strata) to each. With the addition of having an expected SD of the KPI in each stratum, we can determine how many GUs of each stratum we should cover to gain an accurate KPI mean on the whole.

To this end, we test sampling techniques on a GU-level, and we only take a look at a single data set – Austria. To reduce the influence of the BPL we split the data into *Indoor* and *Outdoor* data sets and look at these scenarios separately. By splitting the entire country area into, e.g. 500 m × 500 m large, GUs, we can determine how many of them need to be taken into account for an accurate mean RSRP estimation.

Since we do not have a clear rural/urban area split in Austria, we use the macro BS locations in Austria (provided by the operator), to determine how many BSs are in a 2 km radius of each GU. If more than 50 BSs are found in the radius, then the GU is classified as urban or *BS-dense*, otherwise, we classify it as rural or *BS-sparse*. Measurements are then mapped to their belonging GU and thus to their corresponding stratum – Fig. 9 depicts the mapping of the outdoor measurement data set. Notice that the red *BS-dense* areas overlap with larger cities in Austria, e.g. Vienna, Graz, Linz, Salzburg, Innsbruck.

All crowdsourced data are then binned by their GUs and the mean for each GU is calculated. These GU averages represent our population ground truth in the following sampling schemes. Table. 6 presents the RSRP GU statistics in the 1800 MHz band. The statistics show as expected, that indoor and outdoor measurements have ≈ 7 dB discrepancy, while the strata alone in each data set, show a difference of 4 – 6 dB. In the Outdoor scenario, we have predominantly *BS-sparse* GUs, while in the Indoor scenario number of *BS-sparse* and *BS-dense* GUs is in the same order of magnitude. This indicates, that many GUs in Austria are missing Indoor measurements in the crowdsourced data. If we look at the performance comparison between applied sampling strategies in Fig. 8, we can hardly notice a difference in the performance of the three sampling methods. This is the case due to the very small SD difference among the strata, as well as insufficient mean discrepancies. This testifies to the fact that the coverage of the operator in question is as good in rural areas as it is in urban, as separating the data set into strata brings almost no advantage. Due to its higher variance, the *BS-sparse* stratum (S1) experiences an overshoot of the N_1 level when approaching the 10^{-2} dB error bound in both data sets. In the Outdoor scenario, we found that we would need to cover approximately 50% (15,440 GUs) of the currently measurement-covered GUs in Austria to maintain a mean estimation error of less than 10^{-1} dB in 95% of cases, regardless of the sampling method employed. Similarly, in the Indoor scenario, we would need to cover around 55% (6,032 GUs) of the currently measurement-covered GUs to achieve the same level of accuracy.

VI. CONCLUSION

Accurate estimation of KPIs in mobile networks is critical for improving network performance and customer satisfaction. Sampling methods can be used to estimate KPIs with an acceptable error level while minimizing the number of measurements required. In this work, we investigated the behavior of three sampling methods for accurate KPI mean estimation in mobile networks: SRS, SS with proportional allocation, and SS with optimal allocation.

We characterized the performance of all three methods on the simulated KPI data set while investigating specific cases to reveal scenarios where stratification pays off. We then analyzed the same sampling methods on two MDT crowdsourced data sets from a major Austrian operator. To test and compare different sampling methods on both MDT data sets, we stratified the data into meaningful strata to obtain an accurate mean RSRP estimate with a minimum number of measurements and/or a minimum number of GUs used for mean estimate calculations in each method. Our analysis showed that the first strata split, based on the GPS position, speed, and radio channel conditions, offered a subtle advantage of the SS methods over the SRS method. All three approaches resulted in between 50 and 65% of the total measurements being required to remain below a mean absolute estimation error of 0.01 dB in both data sets.

We further binned the data into equally sized GUs, removing the confounding by determining a single representative KPI value for each GU. Using BS-density-based stratification, we determined how many rural and urban GUs are required for accurate mean prediction. Again, we compared three sampling techniques while using calculated mean GU KPI values as our population ground truth. The analysis revealed that we would have to cover around 50% of the GUs to remain below a mean absolute estimation error of 0.1 dB. Stratification provided a minimal advantage to SRS due to the comparable coverage of this operator in rural and urban regions in Austria, with minor differences in mean and SD among these two strata.

Although we observed only a slight advantage of stratification in real-world data sets for RSRP mean estimation, we can utilize these methods to determine how many samples or areas are required for determining the mean of any KPI in the network. For instance, considering throughput and cell load, possibly more distinct strata can be found to utilize stratification to its full benefit. The findings of this study can help network operators determine the required number of measurements and measurement areas for accurate KPI estimation while minimizing costs and time. The values for the strata variances we derived from the real-world data can allow other researchers to initialize their methods.

APPENDIX

A. RSRP SIMULATION

To illustrate the importance of random sampling and compare it in crowdsourced and drive test scenarios, we simulate an outdoor RSRP map for a part of Vienna's third district using the deep learning network planner (DLNP) from [24], with the following simulation parameters: 15 sectors (three sectors at each of the five BS locations) with $P_{\text{TX}} = 15$ W, $f = 1\,800$ MHz, sector down-tilt of 10° and BS height of 30 m. The simulated area has a dimension 1 000 m × 1 000 m, with an RSRP map resolution of 1 m. The DLNP utilizes the geospatial building model, obtained from the Geodatenviewer der Stadtvermessung Wien [46] and a realistic network layout. The serving RSRP map is obtained by computing the maximum RSRP value across all 15 sectors at each location of the map grid and is depicted in the final map shown in Fig. 1(a). Blue scatter points represent five BS locations, while the red line outlines the Landstrasser Hauptstrasse street in Vienna, obtained from OSM using Python Overpass API [38].

B. VARIANCE APPROXIMATION

Assuming two separate sample sets or stata $S_1 = \{x_1^{(1)}, x_2^{(1)}, \ldots, x_{N_h}^{(1)}\}$, $S_2 = \{x_1^{(2)}, x_2^{(2)}, \ldots, x_{N_h}^{(2)}\}$ that have same mean μ and sample size N_h, we define their variances as:

$$\sigma_1^2 = \frac{\sum_{i=1}^{N_h}(x_i^{(1)} - \mu)^2}{N_h - 1},$$
$$\sigma_2^2 = \frac{\sum_{i=1}^{N_h}(x_i^{(2)} - \mu)^2}{N_h - 1}. \tag{34}$$

Then the average variance of the two groups is given as

$$\frac{\sigma_1^2 + \sigma_2^2}{2} = \frac{\sum_{i=1}^{N_h}\left[(x_i^{(1)} - \mu)^2 + (x_i^{(2)} - \mu)^2\right]}{2(N_h - 1)}$$
$$= \frac{\sum_{i=1}^{2N_h}(x_i^{(1,2)} - \mu)^2}{2N_h - 2}. \tag{35}$$

In comparison, if we combine these two strata into one set $S_{1,2} = S_1 \cup S_2 = \{x_1^{(1)}, x_2^{(1)}, \ldots, x_{N_h}^{(1)}, x_1^{(2)}, x_2^{(2)}, \ldots, x_{N_h}^{(2)}\}$,

Algorithm 1 Algorithm for Stratification Verification

```
Input: population samples $d, R, y_i, n_h$ for $h = 1, 2, \ldots, L$
Output: $\alpha$
 1: Compute strata mean $\mu_h = 1/N \sum_{i=1}^{N_h} y_i$ for $h = 1, 2, \ldots, L$
 2: Compute population mean as $\mu = 1/N \sum_{h=1}^{L} \mu_h N_h$
 3: for $h = 1$ to $L$ do
 4:     if $(n_h = 0)$ then
            $n_h = 1$
 5:     end if
 6:     if $(n_h > N_h)$ then
            $n_h = N_h$
 7:     end if
 8: end for
 9: Initialize error list: error = []
10: for $i = 1$ to $R$ do
11:     SRS of $n_h$ from $N_h$ in each strata $h = 1, 2, \ldots, L$
12:     Compute sample means $\hat{\mu}_h = 1/n_h \sum_{i=1}^{n_h} y_{h_i}$
13:     Compute SS mean $\bar{y}_{st} = 1/N \sum_{h=1}^{L} \hat{\mu}_h N_h$
14:     error[i] = $|\bar{y}_{st} - \mu|$
15: end for
16: return $\alpha = \frac{\sum_{i=1}^{R} \mathbf{1}(\text{error}[i] > d)}{R}$
```

then the variance of the combined set is given by

$$\sigma = \frac{\sum_{i=1}^{2N_h} (x_i^{(1,2)} - \mu)^2}{2N_h - 1}. \tag{36}$$

In the limit $N_h \longrightarrow \infty$, the denominator terms in Eq. (35) and Eq. (36) can be approximated with $2N_h$. This approximation results in $\sigma = \frac{\sigma_1^2 + \sigma_2^2}{2}$, the equation that can be easily generalized to account for arbitrary L strata:

$$\sigma = \frac{\sum_{h=1}^{L} \sigma_h^2}{L}. \tag{37}$$

Hence, we can use the approximation from Eq. (37) for sufficiently large strata sizes.

C. VERIFICATION ALGORITHM

After calculating how many samples n_h are required in each stratum $h = 1, 2, \ldots, L$ for remaining under a certain error bound d with 95% accuracy, we can verify this result by using calculated n_h in each of R sampling iterations. In each iteration, we compute the strata sample means, the stratified mean estimate and the absolute error between the true population mean and the stratified mean estimate. Finally, we compute the α, representing the percentage for which the condition estimation error level is violated. If α indeed lies below 5% (100-95), then randomly sampling previously calculated n_h samples from the corresponding h stratum, results in the stratified mean estimate under the estimation error bound d in 95% of the cases. Note that in the optimal allocation scheme, it may happen that the calculated n_h is higher than N_h. We address this in step 6. Similarly, step 4. corrects for the n_h values rounded to zero, which can happen in rare cases. Clearly, a sample size of zero or drawing more samples than are available in the population is unfeasible. The verification algorithm is summarized in Algorithm 1.

REFERENCES

[1] (2022). *RTR Net Neutrality Report*. [Online]. Available: https://www.rtr.at/TKP/aktuelles/publikationen/publikationen/netzneutralitaetsbericht/RTR_Net_Neutrality_Report_2022.pdf

[2] *Crowdsourcing Approach for the Assessment of End-to-End Quality of Service in Fixed and Mobile Broadband Networks*, Standard E812, ITU-T, May 2020.

[3] GSMA. *4G/5G Network Experience Evaluation Guideline*. Accessed: Mar. 2023. [Online]. Available: https://www.gsma.com/futurenetworks/wp-content/uploads/2020/02/4G5G-Network-Experience-Evaluation-Guideline-_GSMA.pdf

[4] *Measurement Campaigns, Monitoring Systems and Sampling Methodologies to Monitor the Quality of Service in Mobile Networks*, Standard E806, ITU-T, Jun. 2019.

[5] M. P. Armstrong, S. Wang, and Z. Zhang, "The Internet of Things and fast data streams: Prospects for geospatial data science in emerging information ecosystems," *Cartography Geographic Inf. Sci.*, vol. 46, no. 1, pp. 39–56, Jan. 2019.

[6] P. Katsikouli, D. Madariaga, A. C. Viana, A. Tarable, and M. Fiore, "DuctiLoc: Energy-efficient location sampling with configurable accuracy," *IEEE Access*, vol. 11, pp. 15375–15389, 2023.

[7] S. Farthofer, M. Herlich, C. Maier, S. Pochaba, J. Lackner, and P. Dorfinger, "An open mobile communications drive test data set and its use for machine learning," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 1688–1701, 2022.

[8] V. Raida, P. Svoboda, and M. Rupp, "On the inappropriateness of static measurements for benchmarking in wireless networks," in *Proc. IEEE 91st Veh. Technol. Conf. (VTC-Spring)*, May 2020, pp. 1–5.

[9] (2019). *Apps | Opensignal*. [Online]. Available: https://opensignal.com/apps

[10] (2019). *Ookla LLC—Speedtest by OoklabThe Global Broadband Speed Test*. [Online]. Available: http://www.speedtest.net/

[11] (2018). *Alladin IT—The Alladin Nettest*. [Online]. Available: https://nettest.alladin.at/home

[12] R. K. Ganti, F. Ye, and H. Lei, "Mobile crowdsensing: Current state and future challenges," *IEEE Commun. Mag.*, vol. 49, no. 11, pp. 32–39, Nov. 2011.

[13] B. Guo, Z. Wang, Z. Yu, Y. Wang, N. Y. Yen, R. Huang, and X. Zhou, "Mobile crowd sensing and computing: The review of an emerging human-powered sensing paradigm," *ACM Comput. Surv.*, vol. 48, no. 1, pp. 1–31, Sep. 2015.

[14] F. Ma, X. Liu, A. Liu, M. Zhao, C. Huang, and T. Wang, "A time and location correlation incentive scheme for deep data gathering in crowdsourcing networks," *Wireless Commun. Mobile Comput.*, vol. 2018, pp. 1–22, Jan. 2018.

[15] C.-K. Tham and T. Luo, "Quality of contributed service and market equilibrium for participatory sensing," *IEEE Trans. Mobile Comput.*, vol. 14, no. 4, pp. 829–842, Apr. 2015.

[16] S. Reddy, D. Estrin, and M. Srivastava, "Recruitment framework for participatory sensing data collections," in *Proc. 8th Int. Conf.* Helsinki, Finland: Springer, 2010, pp. 138–155.

[17] F. Campioni, S. Choudhury, K. Salomaa, and S. G. Akl, "Improved recruitment algorithms for vehicular crowdsensing networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1198–1207, Feb. 2019.

[18] Y. Wang, Z. Cai, G. Yin, Y. Gao, X. Tong, and G. Wu, "An incentive mechanism with privacy protection in mobile crowdsourcing systems," *Comput. Netw.*, vol. 102, pp. 157–171, Jun. 2016.

[19] Z. Song, C. H. Liu, J. Wu, J. Ma, and W. Wang, "QoI-aware multitask-oriented dynamic participant selection with budget constraints," *IEEE Trans. Veh. Technol.*, vol. 63, no. 9, pp. 4618–4632, Nov. 2014.

[20] F. Zhang, B. Jin, H. Liu, Y. Leung, and X. Chu, "Minimum-cost recruitment of mobile crowdsensing in cellular networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1–7.

[21] (2023). *Tutela*. [Online]. Available: https://www.tutela.com

[22] F. Wamser, A. Seufert, A. Hall, S. Wunderer, and T. Hoßfeld, "Valid statements by the crowd: Statistical measures for precision in crowdsourced mobile measurements," *Network*, vol. 1, no. 2, pp. 215–232, Sep. 2021.

[23] C. Midoglu and P. Svoboda, "Opportunities and challenges of using crowdsourced measurements for mobile network benchmarking a case study on RTR open data," in *Proc. SAI Comput. Conf. (SAI)*, Jul. 2016, pp. 996–1005.

[24] L. Eller, P. Svoboda, and M. Rupp, "A deep learning network planner: Propagation modeling using real-world measurements and a 3D city model," *IEEE Access*, vol. 10, pp. 122182–122196, 2022.

[25] D. R. Shalabh, *Sampling Theory*. Kanpur, India: Indian Institute of Technology Kanpur, 2011. [Online]. Available: http://home.iitk.ac.in/~shalab/course1.htm and http://home.iitk.ac.in/~shalab/sampling/chapter4-sampling-stratified-sampling.pdf

[26] *Technical Specification Group Radio Access Network; Universal Terrestrial Radio Access (UTRA) and Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Measurement Collection for Minimization of Drive Tests (MDT); Overall Description; Stage 2*, document TS 37.320, 3GPP, Jun. 2022.

[27] *5G; Management and Orchestration; 5G Performance Measurements*, document TS 28.552, 3GPP, Oct. 2022.

[28] *Digital Cellular Telecommunications System (Phase 2+) (GSM); Universal Mobile Telecommunications System (UMTS); LTE; Telecommunication management; Performance Management (PM); Performance Measurements; Core Network (CN) Circuit Switched (CS) Domain; UMTS and Combined UMTS/GSM*, document TS 32.407, 3GPP, Apr. 2022.

[29] *Universal Mobile Telecommunications System (UMTS); LTE; Telecommunication Management; Key Performance Indicators (KPI) for Evolved Universal Terrestrial Radio Access Network (E-UTRAN): Definitions*, document TS 32.450, 3GPP, Apr. 2022.

[30] *Ausschreibungsunterlage Im Verfahren Betreffend Frequenzzuteilungen in Den Frequenzbereichen 800 MHz, 900 MHz und 1800 MHz*, Telekom-Control-Kommission, Vienna, Austria, Mar. 2013.

[31] R. V. Akhpashev and V. G. Drozdova, "Spatial interpolation of LTE measurements for minimization of drive tests," in *Proc. 19th Int. Conf. Young Specialists Micro/Nanotechnologies Electron Devices (EDM)*, Jun. 2018, pp. 6403–6405.

[32] R. Enami, S. Gupta, D. Rajan, and J. Camp, "LAIK: Location-specific analysis to infer key performance indicators," *IEEE Trans. Veh. Technol.*, vol. 70, no. 5, pp. 4406–4418, May 2021.

[33] S. Tripkovic, P. Svoboda, V. Raida, and M. Rupp, "Cluster density in crowdsourced mobile network measurements," in *Proc. IEEE 93rd Veh. Technol. Conf. (VTC-Spring)*, Apr. 2021, pp. 1–7.

[34] I. Koukoutsidis, "Estimating spatial averages of environmental parameters based on mobile crowdsensing," *ACM Trans. Sensor Netw.*, vol. 14, no. 1, pp. 1–26, Feb. 2018.

[35] W. G. Cochran, *Sampling Techniques*. Hoboken, NJ, USA: Wiley, 1977.

[36] S. K. Thompson, *Sampling*. Hoboken, NJ, USA: Wiley, 2012.

[37] (2023). *Netscout Geoanalytics*. [Online]. Available: https://www.netscout.com/success-stories/optimize-cell-performance-geo-analytics

[38] (2023). *Python Overpass API*. [Online]. Available: https://python-overpy.readthedocs.io

[39] (2023). *Python Geopandas Library*. [Online]. Available: https://geopandas.org

[40] L. Eller, V. Raida, P. Svoboda, and M. Rupp, "Localizing basestations from end-user timing advance measurements," *IEEE Access*, vol. 10, pp. 5533–5544, 2022.

[41] H. Hofmann, H. Wickham, and K. Kafadar, "Letter-value plots: Boxplots for large data," *J. Comput. Graph. Statist.*, vol. 26, no. 3, pp. 469–477, Jul. 2017.

[42] R. Hoppe, G. Wolfle, and F. M. Landstorfer, "Measurement of building penetration loss and propagation models for radio transmission into buildings," in *Proc. Gateway 21st Century Commun. Village. VTC-Fall., IEEE VTS 50th Veh. Technol. Conf.*, 1999, pp. 2298–2302.

[43] M. Rindler, S. Caban, M. Lerch, P. Svoboda, and M. Rupp, "Swift indoor benchmarking methodology for mobile broadband networks," in *Proc. IEEE 86th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2017, pp. 1–5.

[44] S. Tripkovic, P. Svoboda, and M. Rupp, "Benchmarking of mobile communications in high-speed scenarios: Active vs. passive modifications in high-speed trains," in *Proc. IEEE 95th Veh. Technol. Conference: (VTC-Spring)*, Jun. 2022, pp. 1–6.

[45] L. T. Gwaka, J. May, and W. Tucker, "Towards low-cost community networks in rural communities: The impact of context using the case study of Beitbridge, Zimbabwe," *Electron. J. Inf. Syst. Developing Countries*, vol. 84, no. 3, May 2018, Art. no. e12029.

[46] (2023). *Geodatenviewer Der Stadtvermessung Wien*. [Online]. Available: https://www.wien.gv.at/ma41datenviewer/public

SONJA TRIPKOVIC (Graduate Student Member, IEEE) received the Dipl.Ing. degree in telecommunications from Technische Universität Wien (TU Wien), in 2020. She is currently a University Assistant with the Institute of Telecommunications, TU Wien, with a focus on end-user performance estimates for 4G and 5G networks. Her research interests include the use of crowdsensing for performance measurements in 4G and 5G mobile networks, data-driven estimation of user mobility patterns, and network performance estimation for high-speed mobility scenarios, augmented by geospatial environment integration.

LUKAS ELLER (Graduate Student Member, IEEE) received the Dipl.Ing. degree in telecommunications from Technische Universität Wien (TU Wien), in 2020. He is currently a Project Assistant with the Institute of Telecommunications, TU Wien, with a focus on end-user performance estimates for 4G and 5G networks. His research interest includes assessing the deployment of data-driven deep learning methods for the generation of performance models based on crowdsourced measurements.

PHILIPP SVOBODA (Senior Member, IEEE) received the Dr.Ing. degree in electrical engineering from Technische Universität Wien (TU Wien). He is currently a Senior Scientist with TU Wien, with a focus on the performance aspects of mobile cellular technologies. He is also examining the feasibility of using crowdsensing to conduct performance measurements on 4G and 5G mobile networks. His research interests include a common framework for evaluating the performance of mobile networks, guaranteeing reliable, and fair connectivity for end-users.

MARKUS RUPP (Fellow, IEEE) received the Dipl.Ing. degree from the University of Saarbrucken, Germany, in 1988, and the Dr.Ing. degree from Technische Universität Darmstadt, Germany, in 1993. Until 1995, he was a Postdoctoral Researcher with the University of California at Santa Barbara, Santa Barbara, CA, USA. From 1995 to 2001, he was with the Wireless Technology Research Department, Nokia Bell Laboratories, Holmdel, NJ, USA. Since 2001, he has been a Full Professor of digital signal processing in mobile communications with Technische Universität Wien.