Trust but Verify: Crowdsourced Mobile Network Measurements and Statistical Validity Measures

Anika Seufert*, Florian Wamser*, Stefan Wunderer†, Andrew Hall‡, Tobias Hoßfeld*
**University of Würzburg, Chair of Communication Networks*, Würzburg, Germany
†*Nokia Networks*, Ulm, Germany
‡*Tutela Technologies, Ltd.*, Victoria, Canada
{anika.seufert | florian.wamser | tobias.hossfeld}@uni-wuerzburg.de, stefan.wunderer@nokia.com, ahall@tutela.com

***Abstract*—Network operators, regulators, and big data companies use crowdsourced measurements to study the performance of mobile networks on a large scale. Such a type of measurement is defined as the collection and processing of data measured by the crowd, here the crowd of mobile subscribers. Crowdsourced network measurements make it relatively easy and inexpensive to obtain large amounts of network data that also reflect the quality actually received by the end user. However, this measurement method also involves some uncertainties, since, for example, it is not possible to precisely control when, where and with which devices measurements are taken. Thus, there is a trade-off between the reliability of the individual measurement and the scope of the measurements. Therefore, how data of this type is analyzed is particularly important in order to obtain valid results. To address this issue, our paper defines concepts and guidelines for analyzing the validity of crowdsourced mobile network measurements. In particular, we address precision, for example the number of measurements needed to make valid statements, and also representativeness, for example the spatial and temporal distribution of the data. In addition to the formal definition of these two aspects, we illustrate the issue and possible evaluation approaches with the help of an extensive example data set. This data set consists of more than 11.7M crowdsourced mobile measurements from all over France, measured by a commercial mobile data provider. In the end, we provide an evaluation guideline and two possible use cases.**

I. Introduction and Related Work

For today's network operators, measurements are an integral part of quality monitoring. While operators so far have collected measurement data at the network layer, where they have direct access to, more and more companies and operators are striving to measure network quality from a user perspective. Consequently, the measurement method of crowdsourced network measurements (CNMs) has emerged. In general, the term *crowdsourcing* includes the active participation of volunteers in an outsourced campaign. In the context of network measurements, this is the active participation of users with deliberate user actions, for example the use of a typical speed test application in the mobile network [1]. Crowdsourced network measurements in combination with traditional quality measurement methods in the network layer and on a QoS basis have proven to be a promising approach for a comprehensive quality view of mobile networks. Fundamentals on CNMs as well as a classification of use usages and challenges are provided in [2] and [1].

There are different ways how to use CNM data to evaluate the networks quality. A survey about end-to-end mobile network measurement testbeds, tools, and services can be found in [3]. To measure and compare Internet service providers (ISPs), the authors of [4] used crowd data collected from peer-to-peer BitTorrent users to compare the performance of ISPs from an end user perspective. Other well used options are, for example, the use of video streaming applications to collect crowd data on the smartphone of end users [5, 6]. Since this way of getting large amount of network data is relatively simple and cheap, the number of CNM service providers increased in the last few years, for example Tutela, Ookla, Umlaut, QoSi, Opensignal, and Rohde & Schwarz use the smartphone of end users as measurement device. Their focus is to offer data sets and device data collected from millions of users every day to inform the mobile industry and improve the world's mobile Internet. These companies regularly publish reports on the mobile network experience, e.g., in Germany [7–11], in which they compare network operators, coverage, and speed of their networks. Their advantage is that these measurements reflect the actual course of mobile quality on the user side, directly experienced by the end user. A statistical report for Germany from a research perspective can be found in [12].

The measurement apps, tools, or websites typically measure special aspects of the network, for example, signal strength, Quality of Service (QoS) parameter, or Quality of Experience (QoE) ratings. As a result, however, a large amount of data is obtained from uncontrolled measurements through the crowd. The density, number, and accuracy of the measurements vary due to the uncontrolled measurement environment.

There is always the question of validity of such measurements. In the simplest case, a final output metric is based on one measuring point that is one year old. In the best case, there is high temporal and spatial coverage. This raises the question of the number of measurements required for a meaningful and generally applicable statement. This is particularly clear when looking at the results from various data providers that determine network performance. In 2019, for example, Ookla quantifies the throughput during the Super Bowl 2019 at 102 Mbps with T-Mobile as the fastest network. According to Tutela, on the other hand, Verizon delivered the fastest average download speeds in the stadium at around 38 Mbps. AT&T finished second, Sprint third and T-Mobile came last [13]. Possible reasons for this include the different spatial distribution of the respective measurement data as well as the number of measurements collected and the used evaluation methods. This shows that it is important to define guidelines on how to evaluate CNM data to get valid results. The only work that goes towards validity of crowdsourced data is the work [1]. The authors state that validity, reliability, and representativeness play an important role in all stages of a crowdsourcing campaign: in the design and methodology, the data capturing and storage, and the data analysis. Nevertheless, a detailed discussion of the validity of crowdsourced data and a guideline on how to check data for validity is still missing.

Thus, in this paper, a large-scale commercial CNM data set from July 2019 to December 2019 in France with 11.7M crowdsourced mobile measurements throughout the country is analyzed. We tackle two different aspects for analyzing the validity of crowdsourced mobile network measurements: First, we consider the precision of an evaluation, in particular the precision of a certain metric such as the downlink throughput. We analyze the mean mobile downlink throughput for certain regions and derive the number of measurements required to achieve high precision. Second, we elaborate on the problem of systematic measurement bias that occurs with CNMs due to the influence of user-generated measurements in certain times, such as busy hours, or locations. CNMs should provide results that are (1) true to expectation with (2) the highest possible precision. In this context, true to expectation means that the results are representative, i.e., that no bias (= systematic error) exists compared to reality, although based on sample measurements from crowdsourcing.

The reminder of the paper is structured as follows. Section II presents terms and background information. Definitions for validity are made in Section III. An explanation of the dataset is given in Section IV. The aspect of precision is dealt with in Section V, while measurement bias is discussed in Section VI. An evaluation guideline and exemplary use cases are given in Section VII and Section VIII concludes the paper.

II. Terms and Background

The problem of validity of measurements has generally been extensively studied in various research in different domains [14–16]. *Validity* is, in addition to reliability and objectivity, a quality criterion for models, measurement, or test procedures [15, 17]. Validity is fulfilled if the measurement method measures the characteristic with sufficient *accuracy* that it is supposed to measure or that it pretends to measure [17]. In empirical terms, validity denotes the agreement of the content of an empirical measurement with the logical measurement concept in reality. In general, this is the degree of accuracy with which the feature that is to be measured is actually measured. Definitions can be found in [15, 17]. Fundamental general work on sampling and sample theory is given, for example, in [18].

The *accuracy* of a measurement is further given by the *precision* and *trueness* of a measurement [19, 20]. The precision describes the spread of the results. The trueness ensures that the results also correspond to the correct or true value and are not distorted by the measurement concept, i.e., the representativeness must be ensured in such a way that *no bias* or *systematic errors* occur due to the measurement concept, even if the results are already precise.

In psychology [15] and medicine [16], studies on medication or treatment programs are regularly carried out. Generalized statements are drawn there from a finite number of observations, in this case a sample. The studies are commonly performed (i) as representative as possible and (ii) until the desired precision prevails. In addition, the systematic error in election polls is kept low in electoral research [21] by a representative selection of the surveyed citizens in order to satisfy the validity [14].

III. Statistical Validity for CNMs

Given a CNM S with scope n, i.e., a measurement can be seen as a sample with n observations. Let $S \subseteq U$ be the CNM, with U as the finite underlying population $U = \{1, ..., N\}$ with $N \in \mathbb{N}$ different elements for the measurement. For each element $i \in U$ the value of a variable y can be measured. The vector of these values y_i is denoted by y_U. The aim of the measurement is now to estimate a characteristic $\Theta(y_U)$ of U with the help of a sample S. The characteristic to be estimated is often the population mean $\bar{y}_U = \sum_{i \in U} \frac{y_i}{N}$ or the absolute sum with $y_{U+} = \sum_{i \in U} y_i$. The measurement plan $p(S)$ on S of the possible samples $S \subseteq U$ assigns a measurement probability to each sample: $p : S \to [0, 1]$.

CNMs result in uncontrolled observations without statistical certainty. The values observed in the measurement $(y_{i_1}, ..., y_{i_n})$ are denoted by y_S. This means that $\Theta(y_S)$, given from the sample observations, only reproduces exactly the characteristic relating to the sample subset. Generalized statements, i.e., conclusions in relation to the population U can only be estimated. Thus, valid CNMs are required to have an estimation function (estimator) $T = T(y_S)$ for a characteristic considering the fact that the evaluation is based on samples. A pair $(measurement\ plan,\ estimator)$, i.e. (p, T), is called a measurement strategy or concept. A good estimator is precise and unbiased.

The *quality of a CNM* is defined by measurement trueness and precision according to [19], see Section II, of the concept (p, T). Precision is expressed in terms of the degree of dispersion of y_S. Trueness is expressed in terms of measurement bias [19]. Both are attributed to unavoidable random errors inherent in every CNM measurement procedure.

(1) For precision, the degree of dispersion indicates the spread of data when using sample observations for evaluations. In sample theory, standard error is the measure of dispersion for an estimator T.

(2) A measurement with no bias means that the results are representative or "true" (trueness), i.e., that there is no systematic error. Although sometimes the true value cannot be known exactly, it may be possible to have an accepted reference value for the property being measured with CNMs. The expected value of the estimator with the measurement plan

p is $E[T(y_S)] = \sum_S p(S)T(y_S)$. The bias of an estimator is therefore the mean deviation from the characteristic to be estimated: $E[T(y_S)] - \Theta(y_U)$. An estimator with bias 0 is called unbiased or "true".

IV. Dataset

For the investigation of validity of crowdsourced network measurements, a commercial data set from Tutela Ltd. is used. Tutela collects data and conducts network tests through software embedded in a variety of over 3 000 consumer applications. Although started at random times, measurements are performed in the background in regular intervals if the user is inactive, and information about the status of the device and the activity of the network and the operating system are collected. The data is correlated, grouped, and evaluated according to device and network status (power saving mode, 2G/3G/4G connectivity). Tests are conducted against the same content delivery network. Tutela measures the network quality based on the real performance of the actual network user, including situations when a network is congested, or the user is throttled by tariff. The results in this paper are based on throughput testing in which 2 MB files are downloaded via Hypertext Transfer Protocol Secure (HTTPS). The chosen size reflects the median of the web page size on the Internet.

The used data was collected during half a year from July 2019 to December 2019 in France. Within the used data set, 20 486 257 crowdsourced network measurements are included. After filtering incomplete entries, 17 620 984 measurements remain from which we selected 11 799 577 measurements for further analysis. These measurements were selected based on their geo-coordinates, as we want to study the validity of crowdsourced measurements based on a single large area with sufficient measurements. Therefore, we selected measurements which were conducted in a specific area in France. Thus, the coordinates range between 0 and 5.5 for longitude and between 44.2 and 49.2 for latitude.

V. Precision

The first part of the investigation is devoted to *precision*, which is the description of random errors in the crowdsourcing measurement process due to the use of samples. More precisely, it is the measurement deviation from the exact value due to the scatter of the individual measured values. It is a measure of the statistical variability, expressed in terms of the degree of dispersion.

A. Standard Error and Confidence Intervals

The standard error (SE) for a measured characteristic is the standard deviation of its sample distribution. Speaking for crowdsourcing, this corresponds to the variability of the measurement results of the users evaluating the same characteristic Θ with estimator $T(y_S)$. The variability is first given by the spread of the values in the population U, i.e., the variance $Var(y_U) = \mathbb{E}\left(|y_U - \overline{y}_U|\right)$ and, second, due to the non-exhaustive measurement methodology with sample observations $S \subseteq U$ on the population U. Thus, the standard error decreases as the population variance decreases. Furthermore, it decreases the more individual values are measured.

SE is defined as standard deviation σ for the measured characteristic Θ with $\sigma_\Theta = +\sqrt{Var(\Theta)}$. If the characteristic to be measured is the mean value ($\Theta = \sum_{i \in U} y_i \cdot P(Y = y_i) = \mu$), it is called standard error of the mean (SEM). SEM is defined as $\sigma_\mu = \frac{\sigma_U}{\sqrt{n}}$, where σ_U is the standard deviation of the population U and n is the size of the sample. n is inversely included in the SEM, which means that the SEM decreases with increasing sample size.

In the case of crowdsourced network measurements, the standard deviation of the entire underlying population is usually not known. Therefore, SEM can be estimated by using the sample standard deviation s of the observations y_i, which is defined as

$$s = \sqrt{\frac{1}{n-1} \sum_{y_i \in S} (y_i - \overline{y}_S)^2} \ ,$$

where y_i are the measured values, $\overline{y}_S$ is the sample mean, and n is the size of the sample. Using s, SEM can be estimated as

$$s_{\overline{y}} = \frac{s}{\sqrt{n}} \ ,$$

resulting in an absolute value for the degree of dispersion in the given unit of the measurement.

Using SE (or in the case of CNMs: by using its sample version $s_{\overline{y}}$), confidence intervals (CIs) propose a range of plausible values for an unknown parameter of the real population (e.g., the mean μ). The interval has an associated confidence level (statistical probability) that the exact parameter μ is in the proposed range CI_α. The confidence interval for the mean is defined as

$$CI_\alpha = [\overline{y}_S - MOE_{\frac{\alpha}{2}}, \overline{y}_S + MOE_{\frac{\alpha}{2}}],$$

with $\overline{y}_S$ as sample mean and

$$MOE_{\frac{\alpha}{2}} = z_{\frac{\alpha}{2}} \cdot s_{\overline{y}}$$

is called the margin of error with $z_{\frac{\alpha}{2}}$ is the z-score at position $\frac{\alpha}{2}$ and α is the chosen confidence level.

For crowdsourced measurements, this gives the possibility to quantify how precise a characteristic Θ can generally be determined based on the number of measurements and a given probability [22]. We use this in the following to define the minimal number of crowdsourced measurements (i.e., CNM observations) needed to achieve a certain precision of the data with respect to the pure number of measurements at a given confidence level.

In order to maintain a precision given by the maximum absolute difference $\delta^* = |\overline{y}_S - \mu|$ [22] between the estimated mean value $\overline{y}_S$ of the CNM S and the exact one μ of the underlying population, the minimum number of measurements n_a^{min} can be calculated as

$$n_a^{min}(\delta^*) = \min_{m \geq n} \left\{ MOE_{\frac{\alpha}{2}} \leq \delta^* \right\} \quad (1)$$

$$= \min_{m \geq n} \left\{ z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{m}} \leq \delta^* \right\} . \quad (2)$$

Table I
PRECISION OF DIFFERENT SAMPLE SIZES

Sample size	$\overline{y}_S$	SEM	$CI_{0.05}$	$MOE_{0.025}$
10	33.54	9.26	[12.59; 54.49]	20.96
100	21.20	1.84	[17.54; 24.86]	3.66
1 000	23.89	0.62	[22.67; 25.11]	1.22
1 491	23.51	0.51	[22.52; 24.50]	0.99
10 000	23.51	0.19	[23.13; 23.89]	0.38
26 576	23.57	0.12	[23.34; 23.80]	0.23
100 000	23.82	0.06	[23.70; 23.94]	0.12
1 000 000	23.85	0.02	[23.81; 23.89]	0.04
10 000 000	23.83	0.01	[23.82; 23.84]	0.01

Given the maximum relative error $\gamma^* = \frac{|\overline{y}_S - \mu|}{|\mu|}$ for the estimated mean value $\overline{y}_S$ and the exact one μ, the number of required measurements can be determined as follows

$$n_r^{min}(\gamma^*) = \min_{m \geq n} \left\{ \frac{MOE_{\frac{\alpha}{2}}}{|\overline{y}_S|} \leq \frac{\gamma^*}{1+\gamma^*} \right\}. \qquad (3)$$

B. Example: Determining Precision for Crowdsourced Data

To illustrate the influence of precision on the evaluation results, Figure 1a shows the cumulative distribution function (CDF) of the download throughput of the whole data set. Here, the mean download throughput is 23.83 Mbps, having a standard deviation of 19.56 Mbps and a maximum of 167.95 Mbps. To emphasize the need to determine the minimum number of measurements, Figure 1b shows the mean values for random samples of different sizes, starting from samples with only 10 measurements up to the full data set. For each sample size, 300 random observations were drawn, which are intended to serve as random, crowdsourced measurements. In addition to the estimated values, the black line indicates μ, the exact mean of 23.83 Mbps. While the estimated mean values for sample sizes below 1 000 show high deviations, the mean values for larger sample sizes are getting closer to the real mean.

To precisely quantify the effect for this data set, the standard error of mean, the confidence interval, and the margin of error is evaluated for a confidence level of 95% of exemplary sample sizes from 10 to 10 000 000 measurements of the data set in Table I. If, for example, a precision of $\delta^* = 1\,Mbps$ is desired, the table shows how many measurements are needed to fulfill this precision: For $n_a^{min}(\delta^*) \geq 1\,491 \Rightarrow MOE_{0.025} \leq 0.99$ and thus, the precision is lower than $1\,Mbps$.

If, instead, you prefer to tolerate at most a relative error of $\gamma^* = 1\%$, the following condition must hold: $\frac{MOE_{0.025}}{\overline{y}} \leq \frac{0.01}{1+0.01} = 0.0099$. This condition is fulfilled for $n_r^{min}(\gamma^*) \geq 26\,576$. Thus, in this case, a sample with 26 576 measurements would lead to a high accuracy of at most 1% inaccuracy in relation to the exact mean value when evaluating the mean.

The development of the absolute as well as the relative error for different sample sizes can also be seen in Figure 1c. Here, the brown solid line shows the corresponding absolute error $MOE_{0.025}$, while the blue solid line shows the corresponding relative error $\frac{MOE_{0.025}}{|\overline{y}_S|}$.

VI. MEASUREMENT BIAS

Although having a high degree of precision, it is still possible that measured values do not represent what is to be investigated. Any factor that systematically affects the measurement of a variable across a given sample can cause systematic errors.

A sample is called representative or *true* if it corresponds to the population in the distribution of all statistical characteristics of interest [19]. It is therefore necessary to design the sampling plan in such a way that representativeness can be expected. For this purpose, one needs knowledge about the targeted population. If the statement is to be made from a user perspective, the distribution of the samples must also reflect the distribution of the active users. If the evaluation examines a certain geographical area, for example the mean throughput in France, the data must be collected evenly distributed over France. To prevent bias, data may then not only be collected in densely populated areas, but also in rural areas.

Two frequently occurring factors which can cause measurement biases are the *spatial* and the *temporal* distribution of the measurement data. Thus, in the following, both factors are investigated in detail including examples of real CNM data.

A. Example: Spatial Distribution

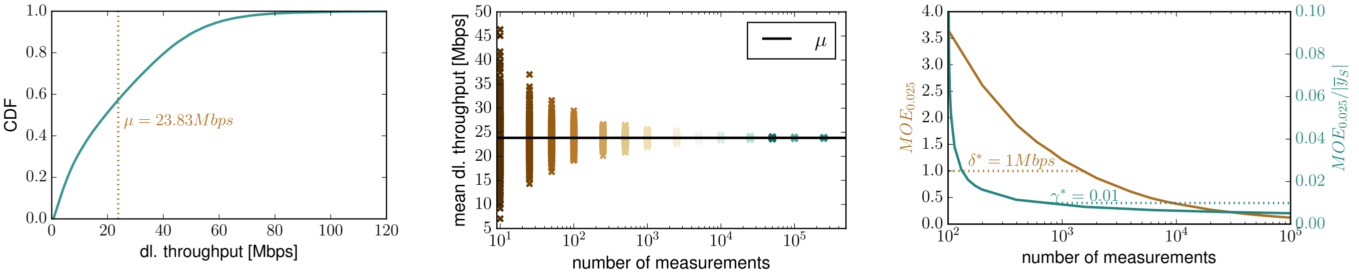
Our data set can be divided in n non-overlapping, equal sized subareas. When looking at the mean throughput for each region, clear differences are visible. By comparing the distributions of values of each equal sized subarea set using a k-sample Anderson-Darling test, the null hypotheses that these samples are drawn from the same populations can be rejected at the 5% level, even for a small number of subareas. This effect is clearly visible in Figure 2a. The mean values per number of subareas as well as the overall mean μ as black line is shown. The mean values differ significantly, having a maximum difference of 5,26 Mbps. The trend becomes even more evident for a larger number of subareas. With CNM, the measuring points are never geographically evenly represented.

To investigate the spatial distribution, Figure 2b shows the distribution of measurements' latitude in brown and longitude in blue. Focusing the latitude, an accumulation of measurement points can be seen at a latitude of about 45.7° and about 48.8°, as well as at a longitude of about 2.3° and 4.8°. That is not surprising since at these locations two large cities are located, Paris (48.86°, 2.35°) and Lyon (45.75°, 4.84°), in which many people live and thus, a high amount of crowdsourced data points can be collected.

B. Example: Temporal Distribution

Another possible factor for biases in the evaluation of crowd data is the temporal aspect. As the measurement times at CNM are often chosen by the user, not all times of the day, weekdays or months are represented to the same extent. However, the time has a decisive influence on the bandwidth.

Figure 2c shows the mean download throughput and the corresponding 95% confidence intervals grouped by the hour of the day. In addition, gray circles indicate the number

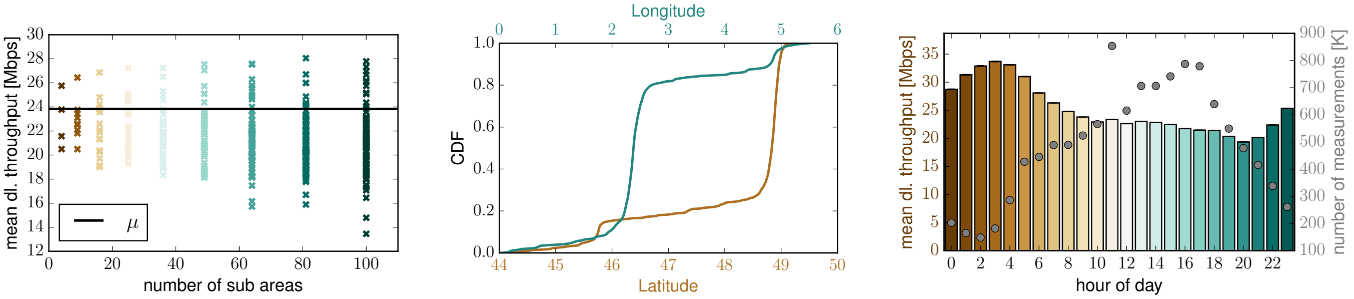


(a) CDF of the download throughput.

(b) Estimated mean throughput.

(c) Absolute and relative errors.

Figure 1. Determining precision for crowdsourced data by evaluating the estimated mean throughput and absolute/relative errors for different sample sizes.



(a) Spatial distribution of the mean throughput.

(b) Geographical distribution of the data.

(c) Temporal distribution of the mean throughput.

Figure 2. Determining spatial and temporal measurement biases for CNM data.

of measurements in thousands. The lowest mean download throughput is measured at 8pm (19.36 Mbps) while the highest means is measured at 3am (33.68 Mbps). Here it becomes clear that the download throughput is significantly influenced by the time of day. The total mean of the values per daytime is 25.13 Mbps, which is 1.3 Mbps higher than the overall mean discussed earlier. For example, at 2am, which is the time with the second highest mean download throughput, only 148 575 measurement values were collected, while a 11am, 853 743 measurements, which is more than five times as many, were collected. To avoid this bias, it is possible to adjust the weighting of the measured values per time.

C. Representativeness

The systematic error caused by the incorrect selection of samples can be eliminated by ensuring that they are representative regarding to the desired statement. (1) Given the desired metric $\Theta(y_U)$ (for example, the mean throughput for a region) the measurement plan $p(S)$ must be set such that $E[T(y_S)] - \Theta(y_U) = \Delta$ with $\Delta \approx 0$. (2) This is achieved by selecting reference statistics for spatial and temporal distribution, such as the population density of a region or country or the number of active users in the mobile network. The distribution of the reference statistics now determines $p(S)$ and guarantees compliance with the sample distribution.

VII. Evaluation Guideline and Examples

From Section V and VI, the following approach for precise and unbiased CNMs can be derived. Given a research question RQ with an evaluation characteristic Θ and a target population V, the following steps can be performed:

1) *Precision*
   a) Define thresholds for precision with absolute error δ^* or relative error γ^*
   b) Calculate n_{min} according to δ^* or γ^*, see Eq. (1), (3)
2) *Measurement Bias (representativeness)*
   a) Determine reference statistics according to V
   b) Calculate sample distribution in temporal and spatial way given by the distribution of reference statistics

A. Example: User-expected Throughput

Given RQ with "What is the expected throughput for an average active user for provider *xy*?". For throughput measurements, an appropriate threshold is an absolute error of at most 1 Mbps, i.e., $\delta^* = 1\,Mbps$. Here, V is given by the active users of provider *xy*. More precisely, either the mobile contracts or the population density might be appropriate as reference statistics V. The calculation of n_a^{min} can be done with Eq. (1) with the help of MOE. Sample observations are chosen or measured in such a way that the temporal and spatial distribution fits to the temporal and spatial distribution of V. This means that, for example, in peak hours and urban areas more measurements have to be taken into account.

B. Example: Network Optimization (Spatial Coverage)

Considering RQ with "How is the average signal strength in Texas, US?". Signal strength is commonly given as RSSI with a precision of 1 dBm, $\delta^* = 1\,dBm$. The calculation of

n_a^{min} can be done with Eq. (1). The CNM should measure in a uniformly distributed way per km^2 all over Texas, V. Here, observations should be uniformly distributed over all Texas, i.e. the weighting of the measurement should be adjusted so that no region is overrepresented. This means that, for example, accumulations of measurements due to peak times or densely populated areas must be balanced.

VIII. Conclusion

When using crowdsourced network measurements (CNMs), network operators, regulators, and big data companies are faced with the challenge of making valid statements out of uncontrolled measurements. Thus, this article defines concepts for analyzing the validity of CNMs.

In the first part, we consider CNMs to be a mathematical sampling process and, as a result, derive from this the need for high precision and trueness (representativeness). For precision, we derive the minimum number of measurements required for a given confidence level and maximum tolerable absolute or relative error. This gives operators and big data companies the opportunity to determine the number of measurements required in advance to ensure a certain precision. Next, we elaborate on the trueness. Even precise measurements can show a systematic error if their measurement process and its samples are not drawn representatively. We point out that this is often the case with CNMs. Measurements are taken more often during the day and in metropolitan areas as people are directly involved. We elaborate on the problem and point out the need for reference data to ensure representativeness. We illustrate the importance of both factors, precision and representativeness, using a large CNM data set and showed possible evaluation approaches. In the end, we provide a guideline on how to evaluate CNM data to get valid results.

References

[1] T. Hoßfeld, S. Wunderer, and et al., "White Paper on Crowdsourced Network and QoE Measurements – Definitions, Use Cases and Challenges," Würzburg, Tech. Rep. 10.25972/OPUS-20232, Mar. 2020.

[2] M. Hirth, T. Hoßfeld, M. Mellia, C. Schwartz, and F. Lehrieder, "Crowdsourced Network Measurements: Benefits and Best Practices," *Computer Networks*, vol. 90, pp. 85–98, 2015.

[3] U. Goel, M. P. Wittie, K. C. Claffy, and A. Le, "Survey of end-to-end mobile network measurement testbeds, tools, and services," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 105–123, 2015.

[4] Z. S. Bischof, J. S. Otto, M. A. Sánchez, J. P. Rula, D. R. Choffnes, and F. E. Bustamante, "Crowdsourcing ISP Characterization to the Network Edge," in *First ACM SIGCOMM Workshop on Measurements Up the Stack*, Toronto, Canada: ACM, 2011, pp. 61–66.

[5] F. Wamser, M. Seufert, P. Casas, R. Irmer, P. Tran-Gia, and R. Schatz, "YoMoApp: A tool for Analyzing QoE of YouTube HTTP Adaptive Streaming in Mobile Networks," in *European Conference on Networks and Communications (EuCNC)*, IEEE, 2015, pp. 239–243.

[6] A. Schwind, C. Midoglu, Ö. Alay, C. Griwodz, and F. Wamser, "Dissecting the Performance of YouTube Video Streaming in Mobile Networks," *International Journal of Network Management*, 2019.

[7] 4Gmark, *Data Mobile Barometer*, Feb. 2018.

[8] Speedtest, *Speedtest Market Report Germany 2016*, Dec. 2016.

[9] Tutela Technologies, "Germany Mobile Experience Report – Country-level mobile experience and usage results from Tutela's crowdsourced mobile network testing (May-July 2019)," Tech. Rep., Jul. 2019.

[10] P3 Communications, *The Great 2019 Mobile Network Test*, Jan. 2019.

[11] Opensignal, *Germany – Erfahrungsbericht mit mobilem Netzwerk (November 2019)*, Nov. 2019.

[12] A. Schwind, F. Wamser, T. Hoßfeld, S. Wunderer, E. Tarnvik, and A. Hall, "Crowdsourced Network Measurements in Germany: Mobile Internet Experience from End User Perspective," *arXiv preprint: 2003.11903*, 2020.

[13] M. Dano. (2019). Who Won the Super Bowl Speed Game? It Depends on Who You Ask, [Online]. Available: https://www.lightreading.com/testing/who-won-the-super-bowl-speed-game-it-depends-on-who-you-ask/d/d-id/749241 (visited on 01/28/2021).

[14] R. Rich, C. Brians, and L. Willnat, *Empirical Political Analysis: Quantitative and Qualitative Research Methods*. Routledge, 2018.

[15] M. Eid and K. Schmidt, *Testtheorie und Testkonstruktion*. Hogrefe Verlag, 2014.

[16] S. Chow and J. Liu, *Design and Analysis of Clinical Trials: Concepts and Methodologies*, ser. Wiley Series in Probability and Statistics. Wiley, 2004.

[17] S. Messick, "Validity," *ETS Research Report Series*, vol. 1987, no. 2, 1987.

[18] A. Stuart, *Basic Ideas of Scientific Sampling*, ser. Griffin's Statistical Monographs. Hafner Press, 1976.

[19] International Organization for Standardization, *Accuracy (trueness and precision) of measurement methods and results - Part 1: General principles and definitions (ISO 5725-1:1994)*. International Organization for Standardization, 1994.

[20] Deutsches Institut für Normung, *Qualitätsmanagement und Statistik: Normen. Begriffe*, ser. DIN-Taschenbuch Bd. 1. Beuth, 2001.

[21] R. Adcock and D. Collier, "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research," *American political science*, pp. 529–546, 2001.

[22] A. M. Law and W. D. Kelton, *Simulation modeling and analysis*. McGraw-Hill New York, 2000, vol. 3.