



Article

Valid Statements by the Crowd: Statistical Measures for Precision in Crowdsourced Mobile Measurements

Florian Wamser ^{1,*}, Anika Seufert ¹, Andrew Hall ², Stefan Wunderer ³ and Tobias Hoßfeld ¹

- ¹ Chair of Communication Networks, University of Würzburg, 97074 Würzburg, Germany; anika.seufert@informatik.uni-wuerzburg.de (A.S.); tobias.hossfeld@uni-wuerzburg.de (T.H.)
² Tutela Technologies, Ltd., Victoria, BC V8W 1H8, Canada; ahall@tutelatechnologies.com
³ Nokia Networks, 89081 Ulm, Germany; stefan.wunderer@nokia.com
* Correspondence: florian.wamser@informatik.uni-wuerzburg.de

Abstract: Crowdsourced network measurements (CNMs) are becoming increasingly popular as they assess the performance of a mobile network from the end user's perspective on a large scale. Here, network measurements are performed directly on the end-users' devices, thus taking advantage of the real-world conditions end-users encounter. However, this type of uncontrolled measurement raises questions about its validity and reliability. The problem lies in the nature of this type of data collection. In CNMs, mobile network subscribers are involved to a large extent in the measurement process, and collect data themselves for the operator. The collection of data on user devices in arbitrary locations and at uncontrolled times requires means to ensure validity and reliability. To address this issue, our paper defines concepts and guidelines for analyzing the precision of CNMs; specifically, the number of measurements required to make valid statements. In addition to the formal definition of the aspect, we illustrate the problem and use an extensive sample data set to show possible assessment approaches. This data set consists of more than 20.4 million crowdsourced mobile measurements from across France, measured by a commercial data provider.

Keywords: mobile networks; crowdsourced measurements; statistical validity



Citation: Wamser, F.; Seufert, A.; Hall, A.; Wunderer, S.; Hoßfeld, T. Valid Statements by the Crowd: Statistical Measures for Precision in Crowdsourced Mobile Measurements. *Network* **2021**, *1*, 215–232. <https://doi.org/10.3390/network1020013>

Academic Editor: Amitava Datta

Received: 28 July 2021

Accepted: 31 August 2021

Published: 13 September 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Mobile internet is increasingly used in every-day life, and end users expect to have the same quality as when they are at home. For this reason, service and network operators are interested in monitoring the current state of quality perceived by end users with their service or network. While operators so far collected measurement data on the physical, data transmission and network layers to which they have direct access, more and more companies and operators are striving to measure network quality from a user perspective. Measurements from the end user perspective are essential to detect or to understand upcoming problems in networks, and are therefore essential for improving Quality of Service (QoS) and enhancing Quality of Experience (QoE). It is, however, not possible to ask the customers about their satisfaction every time they use an app or service. Consequently, the measurement method of crowdsourced network measurements (CNMs) emerged. According to [1], CNMs are defined as “[...] actions by an initiator who outsources tasks to a crowd of participants to achieve the goal of gathering network measurement-related crowd data.” Using the end user devices for gaining crowdsourced measurements on the user side, operators can gain a much better holistic understanding of the impact of network challenges or issues on the quality experienced by end-users. CNMs, in combination with traditional quality measurement methods in the network layer and on a QoS basis, were proven to be a promising approach for a comprehensive quality view of mobile networks.

In general, the term *crowdsourcing* includes the active participation of volunteers in an outsourced campaign [2]. In the context of network measurements, this is the active participation of users in measurements with deliberate user actions; for example, the use of

a typical speed test application in the mobile network [1]. Similar mobile measurements are carried out within an app in the background without the active involvement of end-users, but also involve the mass of end users in a crowdsourced sense. Crowdsourced measurements take into account a special type of crowdsourcing: the collection and processing of data that are measured by the crowd [3].

Network operators and data provider companies (i.e., companies that collect CNM data and then resell it, e.g., to network operators) use this approach to take large-scale measurements of mobile networks. The measurement apps, tools, or mobile websites, usually provided by the measurement provider such as a network operator, regulatory agencies, or a big data company, typically measure one or more aspects of the network: signal strength, mobile app performance such as video streaming quality, QoE ratings on a numeric scale (e.g., 1–5), or QoS. 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 differs due to the uncontrolled measurement environment, which is especially true in case of crowdsourced mobile measurements. It is typical for crowdsourced measurements that a lot of data originate from large cities, that measurements are taken during the day, and that the measurements per device are at different intervals, which can be attributed to the concentration of mobile phone users in cities and the times when devices are commonly used. There is always the question of the validity [4] of such measurements, i.e., is it permissible to make this statement based on the available data? In the simplest case, an outcome 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.

The Super Bowl 2019 can serve as an example of such a challenge. After the match, different CNM companies, for example Ookla and Tutela, published statistics about the mobile network speed during the Super Bowl. Ookla quantified the throughput during the Super Bowl at about 100 Mbit/s, with T-Mobile as the fastest network [5]. In contrast, according to Tutela, Verizon delivered the fastest average download speeds in the stadium at around 40 Mbit/s. The evaluations of the two providers thus differ significantly. 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 the validity of CNM data. However, little research was performed in this area to date. The only works that go towards validity of crowdsourced data are [1,6]. The authors of [6] state that, for example, end device related issues, resource consumption, and privacy versus reliability are challenges that CNMs bring. In [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, there is a lack of detailed discussion on the validity of crowdsourced data and, in particular, a lack of guidelines or metrics on how to test data for validity. Hence, the basics are needed to understand its importance, avoid errors, and carry out crowdsourced measurements in a meaningful way.

To address this issue, in this paper a large-scale commercial CNM data set from July 2019 to December 2019 for France with 20.4 M crowdsourced mobile measurements throughout the country and its overseas territories is analyzed. We tackle the following aspect for analyzing the validity of crowdsourced mobile network measurements: 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. This paper is a full extension of [7] and, to the best of the authors' knowledge, shows for the first time the application and definition of a comparable score to quantify the precision of a statement from a CNM data set.

The remainder of the paper is structured as follows. In Section 2, different ways of measuring mobile network quality are presented, and the methodology of crowdsourced

network measurements is introduced. Section 3 summarizes related work in the field of CNMs. Definitions for validity are summarized in Section 4, while an explanation of the used data set is given in Section 5. The aspect of precision, including the definition of the metric called CNM Precision Validity Score, is dealt with in Section 6. Section 7 illustrates the importance and the applicability of the CNM Precision Validity Score by showing some exemplary results based on the given data set. Section 8 concludes the paper and summarizes the findings.

2. Measuring Mobile Network Quality

There are different ways how operators can monitor the quality of their network at the user side. This starts with the collection of subjective ratings directly from the end user, continues with monitoring mobile applications and services, and ends with network measurements. In the following, we will present the different methods and also classify and describe the emerging technique of CNMs.

Subjective User Studies are always required for modeling the user's QoE. Here, people are asked about their satisfaction with a given service under specific network conditions. Using their results, models can be created to identify the key performance indicators (KPIs) of a service or application. These KPIs can later be measured automatically, and can then be mapped to an estimated QoE value. Here, the advantage is that real user experience is included in the evaluation. Nevertheless, this method is very cost-intensive as the participants have to be paid. Furthermore, it is not possible to conduct subjective user studies on a larger scale. Best practices and recommendations for crowdsourced QoE assessment are summarized in [8,9].

In-Service Monitoring is another another way of measuring the networks' quality by passively measure the speed of incoming and outgoing data of an application, for example, a mobile messaging application or smartphone game. In addition, user behavior can be monitored to get deeper insights in the QoE. Negative aspects of in-service monitoring are that the access to use the network information has to be requested and allowed by the smartphone holders. Furthermore, depending on the service in which the measurement tool is included, it is not easy to reach a large number of people, and thus, monitor the mobile network in a large scale for different purposes.

Measurement Applications are used to monitor the current status of the network by using a standalone measurement application, which can be freely downloaded by smartphone holders who are interested in network statistics. This kind of application offers the possibility to the user to run network speed tests at any time they want to evaluate the current network conditions. In addition, it is also possible to start small network tests at regular intervals, for example daily, to to receive continuous information. Disadvantages of this way of collecting network data is that it is hard to get results on a large scale, as the incentive for the users is limited. It follows that only interested users download this app, and thus, only network statistics from them are collected. The users therefore rather reflect a nonrepresentative group of the population. Furthermore, using measurement apps, only QoS parameters can be monitored; the user satisfaction (QoE) can only be estimated using QoE models.

Hybrid Applications combine advantages of in-service monitoring and measurement applications. Here, different applications like smartphone games or messaging services can trigger active measurements in addition to passive monitoring. This is especially interesting if the same service provider can address different target groups, e.g., people who play online games and people who use messaging services, to collect QoS values from a heterogeneous group of people. Using different apps, especially if widespread applications cooperate with network measurement companies, it is relatively easy to monitor the mobile network in a large scale.

In-network Measurements are probably the simplest measurement method for Internet service providers. Here, providers do network measurements within their own network. The biggest advantage of this way of collecting network data is that the measurements can

be completed fully automated, and the status of the whole network can be evaluated at regular intervals. Nevertheless, this is also the biggest disadvantage, as only statistics from one provider can be collected, and thus, no comparison of different providers is possible. Furthermore, for internet service providers, it is not possible to measure the network's quality down to the end user, but only until the last hop under their control (e.g., base station). Thus, the QoS, and especially also the QoE, of the end users can only be estimated.

Crowdsourced Network Measurements use crowdsourcing to gather information about the quality of the network. Crowdsourcing is the methodology of processing a task by a large group of people instead of a designated agent [2]. For network measurements, crowdsourcing has three major advantages: it makes it much easier to cover a wide range of situations and users, it allows entities other than the network operator to assess the performance and other characteristics of a network independently, with a coverage that is not feasible using other methods such as drive testing, and it offers the possibility to collect statistics from end-user perspective. Thus, CNMs make it possible to get insights into the real network behavior as it is experienced by the end-user, as they use realistic hardware and software settings with heterogeneous devices, access networks, and load situations. A comparison of crowdsourcing with traditional measurement techniques and best practices how to design crowdsourced network measurements issues is made in [3]. There are two ways of doing crowdsourcing studies: active or passive measurements. Either workers are paid to actively process a task or applications on the end users' smartphones are used to collect KPIs in an active way using measurement applications. In the second case, CNMs can be seen as a special case of crowdsensing, where user devices act as environmental sensors, and thus, passively monitor the network using in-service monitoring. Crowdsourced measurement data (crowd data) offers new possibilities and can be used for various applications, such as the benchmarking of network operators, providers, technologies, or countries, as well as, e.g., for monitoring, planning, and optimization of the network. In this way, crowd data provides insights beyond the network layer, that is, at the application and user level. This makes crowd data very valuable and extends the current practice of operators to evaluate networks. The ultimate goal is to use crowd data—combined with other network and user data—to improve QoE, but also for regulatory purposes, e.g., to identify issues with coverage or network settings. Challenges, drawbacks, and benefits of CNMs are listed in [1].

3. Related Work on the Usage of CNMs

In recent years, CNMs became increasingly relevant in research and practice, as they enable the fast and relatively cheap collection of information on network and application level. Fundamentals on CNMs are specified in [1]. In their white paper, the authors provided definitions of the terms crowdsourced network and QoE measurements, defined use cases for CNMs, and discussed challenges. The three main use cases the authors mention are network planning, network monitoring, and benchmarking. Thus, the following related work is grouped into these three categories. In addition, research which focuses on challenges of CNMs are discussed.

One area of application of CNMs is network planning, which includes, for example, the creation of coverage maps for mobile networks. In [10], the authors analyzed different estimation approaches for base station positions using crowdsourced data. They found that a grid-based approach provides the best estimates when compared with that of their real locations. Another approach to estimate base station localization using CNMs was done by [11]. Here, the authors evaluated the applicability of crowdsourced cellular signal measurements in this context and showed that feature clustering leads to good results.

For internet service providers (ISPs), network monitoring is essential. With the help of CNMs, they are able to monitor network quality from a user perspective. This can, for example, be done by collecting information during the use of specific smartphone applications. Here, different KPIs can be measured on several layers, from context parameters such as cultural background through network parameters such as signal strength up to

application parameters including number of stalling and user-focused parameters, such as browser session time. Especially video streaming applications are well-used options to collect crowd data on the smartphone of the end-users. For example, in [12], the authors designed a smartphone application to analyzing the QoE of YouTube HTTP Adaptive Streaming in mobile networks. Another approach was performed by the authors of [13]. Here, an active measurement framework to collect video streaming KPIs was designed to monitor the quality of mobile networks in Europe [14]. A statistical report of the mobile internet experience for Germany based on CNMs data report can be found in [15]. In addition to some scientific work in this field, the number of commercial CNM service providers increased in the last few years. Examples of such providers are Tutela, Ookla, Umlaut, QoSi, Opensignal, and Rohde & Schwarz, which all use the smartphones of the end users as measurement devices. These companies regularly publish reports on the mobile network experience, for example [16–19]. In these reports, a comprehensive evaluation of the current state of the mobile networks is given. Furthermore, they compare network operators, coverage, and speed of their networks.

Another use case of CNMs is benchmarking, and thus, to measure and compare different ISPs. As previously mentioned, commercial CNMs service providers regularly publish reports that can also include the comparison of the network quality of different ISPs. In research, other benchmarking approaches are presented. For example, in [20], the authors used crowd data collected from peer-to-peer BitTorrent users to compare the performance of ISPs from end-user perspective. Using transfer rates as well as network and geographic location information, they showed that this approach is a feasible way to characterize the service that subscribers can expect from a particular ISP. Another model for evaluating the performance of different ISPs using CNMs was presented by [21]. In their work, they introduce a model which characterizes throughput as a function of signal power.

In addition to a wide range of applications, however, CNMs also involve a number of challenges. In [1], the key challenges of CNMs are named as validity, reliability, and representativeness, which 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. Other challenges inherent to CNMs via smartphones were presented by [6]. Here, for example, end device related issues, resource consumption, and privacy versus reliability were discussed and shown by the example of a CNMs data set. While these two articles describe in detail various challenges in designing, collecting, and analyzing CNM data, they do not provide specific solutions or guidelines. To the best of our knowledge, a detailed discussion of the validity of crowdsourced data and a guideline on how to check data for validity is still missing.

4. Defining Statistical Validity for CNMs

The problem of validity of measurements was generally extensively studied in various research in different domains [4,22–24]. *Validity* is, in addition to reliability and objectivity, a quality criterion for models, measurement, or test procedures [23,25].

Validity: a measurement is valid if it actually measures what it is intended to measure, and thus, delivers credible results.

Reliability: reliability relates to whether your research produces reliable results when done repeatedly.

Objectivity: research is objective if there are no unwanted influences from people involved.

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 [4,25]. 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 [23,25]. Fundamental general work on sampling and sample theory is given, for example, in [26].

The accuracy of a measurement is further given by the *precision* and *trueness* of a measurement [4,27]. The International Standard ISO 5725 [4] defines them as follows.

“The general term accuracy is used in ISO 5725 to refer to both trueness and precision. (...) Trueness refers to the closeness of agreement between the arithmetic mean of a large number of test results and the true or accepted reference value. Precision refers to the closeness of agreement between test results.”

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 the literature, the concept of validity is commonly further divided into several empirical and theoretical validity aspects for measurements [23]. These include construct validity, convergent validity, discriminant validity, or content validity [28]. In the following, we assume that the measured values were selected in the sense of the characteristic to be recorded (construct validity). Furthermore, in this work we only deal with questions about the degree of precision.

In psychology [23] and medicine [24], studies on medication or treatment programs are regularly carried out. Generalized statements are drawn there from a finite number of observations, and 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 [29] by a representative selection of the surveyed citizens to satisfy the validity [22].

Given a CNM S with scope m , i.e., a measurement can be seen as a sample with m observations. Let $S \subseteq U$ be the CNM with U as the finite underlying population $U = \{1, \dots, n\}$ with $n \in \mathbb{N}$. 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 $\mu = \bar{y}_U = \sum_{i \in U} \frac{y_i}{N}$ 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: $p : S \rightarrow [0, 1]$.

CNMs result in uncontrolled observations without statistical certainty. The values observed in the measurement $(y_{i_1}, \dots, y_{i_n})$ 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 [4] 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 [4]. Both are attributed to unavoidable random errors inherent in every CNM measurement procedure. 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 .

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”.

Hence, we can evaluate precision and trueness. In this paper, we will focus solely on precision in the following when analyzing CNM data. Table 1 summarizes the notations.

Table 1. Key variables and notations used in the paper.

Notation	Description
U, n	Underlying population, $U = \{1, \dots, n\}$ with $n \in \mathbb{N}$ being the size of the population U
S, m	Sample population, i.e., CNM data set, $S \subseteq U$ with $S = \{1, \dots, m\}$ and $m \leq n$
y_U	Values of population U (e.g., all download throughput values in U)
y_S	Values of all measurements in S (e.g., all measured download throughput values in S)
$\Theta(y_U)$	Characteristic to be evaluated on U (e.g., mean value)
$T(y_S)$	Estimator function of the given characteristic Θ using y_S
\bar{y}_U, \bar{y}_S	Population mean of U , resp. S
$p(S)$	Measurement plan on S , assigns a measurement probability to each sample, $p : S \rightarrow [0, 1]$
$\sigma(\Theta)$	Standard deviation of a specific evaluation or characteristic Θ
$\sigma(\bar{y}_U)$	Standard error of the mean (SEM), see Equation (1)
s	Sample standard deviation, see Equation (2)
CI_α	Confidence interval with a significance level of α , see Equation (3)
z_β	Quantile function for probability β for a given distribution (e.g., Normal or Student's t distribution)
t^*	Target precision (e.g., $\delta^* = 100$ kbit/s or $\gamma^* = 0.01$)
δ^*	Target precision as maximum absolute difference
γ^*	Target precision as maximum relative difference
$n_{abs.}^{min}(\delta^*)$	Minimum number of measurements to achieve an absolute precision of δ^* , see Equations (4) and (5)
$n_{rel.}^{min}(\gamma^*)$	Minimum number of measurements to achieve a relative precision of γ^* , see Equations (6) and (7)
q	Target precision type ($q = abs.$ for absolute precision, resp. $q = rel.$ for relative precision)
$Val. Score_{prec.}(t^*)$	CNM Precision Validity Score for target precision t^* , see Equation (10)

5. Data Set

For the investigation of validity of CNMs, a commercial data set from Tutela Ltd. (Victoria, Canada) is used. Tutela collects data and conducts network tests through software embedded in a variety of over 3000 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 users are throttled because they exceeded the data volume of their contract. 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 data used were collected over six months from July 2019 to December 2019 in France and in its overseas departments of the French territorial collectivity. Within the used data set, 20,486,257 CNMs are included.

Figure 1 shows the location of the measurements in France. The color within the plot represents the number of measurements per square kilometer. The more crowd-sourced measurements were made at a location, the brighter the point. The differences are particularly noticeable for the Paris region in the inner city, which becomes clear in the subfigure at the bottom-left. The measurements for the region around Lyon are shown at the bottom-right. Overall, these figures show where most of the measurements are carried out, namely in cities or in busy places such as main roads and highways. The mean number of measurements per square kilometer is 48.29. In addition to meta information like date and geo-coordinates, the data set includes information on current network performance, including, amongst other variables, download throughput. The question now arises as to whether the number of measurements in a region of interest is sufficient to be able to make a valid statement. This is examined in the following sections.

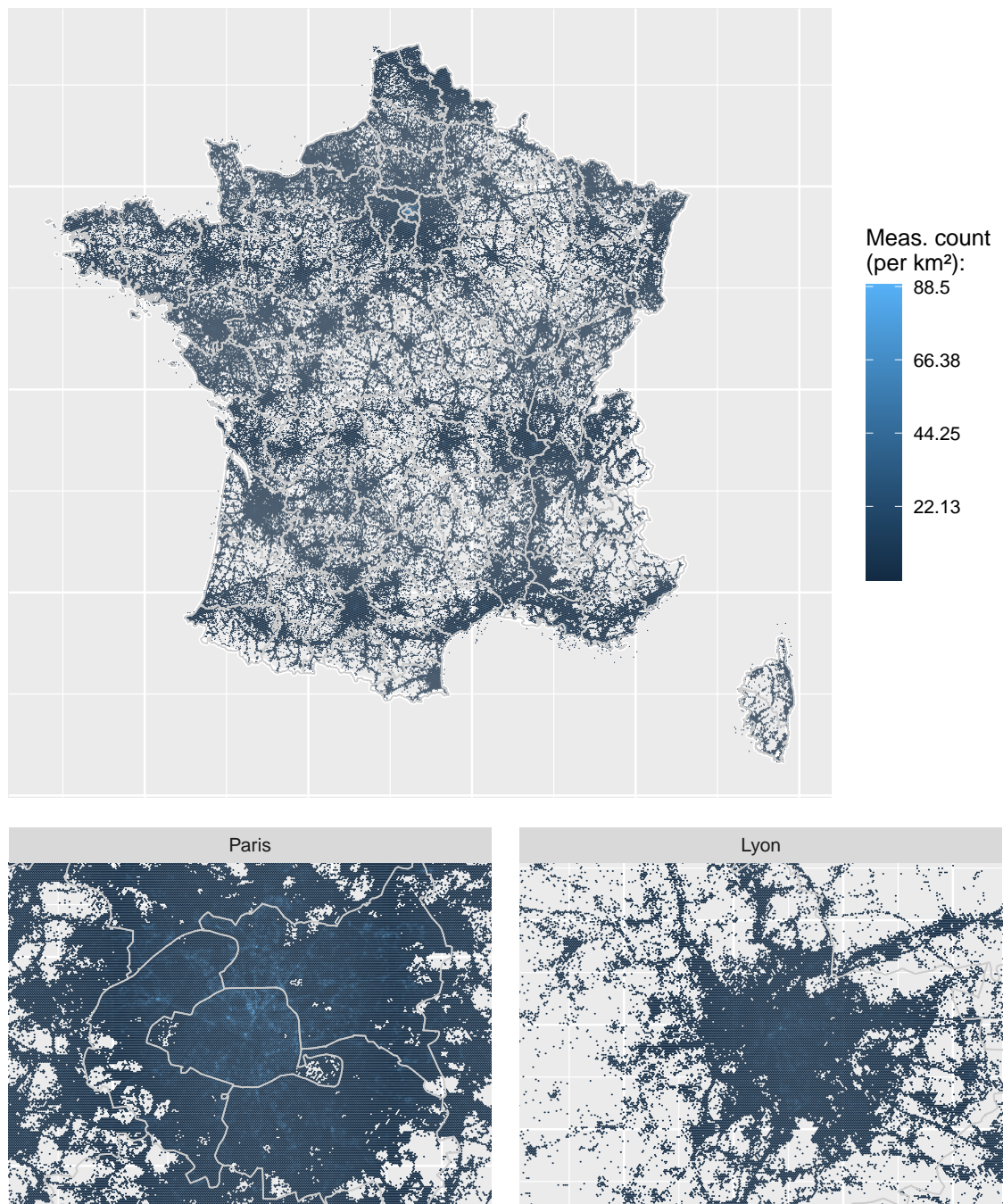


Figure 1. Number of crowdsourced network measurements (CNMs) per km² from July to December 2019 in data set of France with additional subfigures for regions Paris and Lyon at bottom-left and -right, respectively. Absence of a measurement is indicated in gray, whereas different shades of blue show density of the measurements.

6. Precision

This part of the investigation is devoted to precision, which is the description of the spread in values 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.

6.1. Standard Error and Confidence Intervals in the Context of CNMs

The standard error (SE) is the standard deviation for a measured characteristic Θ on the sampling distribution, i.e., it is a measure of how much an observed parameter in a sample deviates on average from the true parameter of the population. Speaking for CNMs, this corresponds to the variability of the measurement results of the users evaluating the same characteristic Θ with estimator $T(y_S)$. The variability of the characteristic is firstly given by the spread of the values in the population U itself, i.e., the variance of y_U with $Var(y_U) = \mathbb{E}[(y_U - \bar{y}_U)^2]$ and, secondly due to the nonexhaustive 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 $\sigma(\Theta) = \sqrt{Var(\Theta)}$. Please note that we use the symbol σ in our work for the standard deviation of a specific evaluation or characteristic Θ of CNM data. Other standard deviations are indicated by lowercase letters, for example s , to distinguish between the two standard deviations with different data. If the characteristic to be measured is the mean value ($\Theta = \bar{y}_U$), σ is called standard error of the mean (SEM).

The standard deviation of the population being sampled is seldom known. Thus, SEM on the sampling distribution S is estimated by

$$\sigma(\bar{y}_U) \approx \frac{s}{\sqrt{m}}, \quad (1)$$

where s is the standard deviation calculated by an estimator on sample S , and $m = |S|$ is the size of the sample. m is inversely included in the SEM, which means that the SEM decreases with increasing sample size. The estimator, i.e., the sample standard deviation s of the observations y_i , is defined as

$$s = \sqrt{\frac{1}{m-1} \sum_{i \in S} (y_i - \bar{y}_S)^2}, \quad (2)$$

where y_i are the measured values, \bar{y}_S is the sample mean, and m is the size of the sample. $\frac{1}{m-1}$ ensures that s is an unbiased estimator. Using s , SEM $\sigma(\bar{y}_U)$ can be estimated as $\frac{s}{\sqrt{m}}$, resulting in an absolute value for the degree of dispersion for a characteristic Θ when sampling.

Using SEM, confidence intervals (CIs) propose a range of plausible values for an unknown parameter of the real population (e.g., the mean \bar{y}_U). The interval has an associated significance level that the exact parameter \bar{y}_U is in the proposed range CI_α . The confidence interval for the mean is defined as

$$CI_\alpha = \left[\bar{y}_S - z_{\frac{\alpha}{2}} \frac{s}{\sqrt{m}}, \bar{y}_S + z_{\frac{\alpha}{2}} \frac{s}{\sqrt{m}} \right], \quad (3)$$

with \bar{y}_S as sample mean, $z_{\frac{\alpha}{2}}$ as quantile at $\frac{\alpha}{2}$ for a given distribution, and α is the chosen significance level.

For crowdsourced measurements, this gives the possibility to quantify how precise a characteristic Θ can generally be determined in terms of the number of measurements and a given significance level [30]. 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.

To maintain a precision given by the maximum absolute difference $\delta^* = |\bar{y}_S - \bar{y}_U|$ [30] between the estimated mean value \bar{y}_S of the CNM S and the ex-

act one \bar{y}_U of the underlying population, the minimum number of measurements $n_{abs.}^{min}$ can be estimated as

$$n_{abs.}^{min}(\delta^*) = \arg \min_{\hat{m} \in \mathbb{N}} \left\{ z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{\hat{m}}} \leq \delta^* \right\}. \tag{4}$$

For certain distributions such as the Standard Normal Distribution, this formula can be solved to

$$n_{abs.}^{min}(\delta^*) = \left(\frac{z_{\frac{\alpha}{2}} \cdot s}{\delta^*} \right)^2. \tag{5}$$

Please note that this is not directly possible for the Student's t distribution because quantile $z_{\frac{\alpha}{2}}$ depends on the number of measurements, which makes the formula an estimate that overestimates the minimum number of measurements required. For this reason, we will later give two algorithms for the calculation: (1) the calculation with the direct transformation (Equation (5)) and (2) the iterative, exact calculation for Student's t distribution (Equation (4)), if one does not want to use the approximation.

Another possibility, which is often required in practice, would be the relative difference according to the mean value instead of the absolute difference, i.e., the error to the exact mean value relative to the population mean. Given the maximum relative error $\gamma^* = \frac{|\bar{y}_S - \bar{y}_U|}{|\bar{y}_U|}$ for the estimated mean value \bar{y}_S and the exact one \bar{y}_U with $\bar{y}_S, \bar{y}_U \neq 0$, the number of required measurements can be estimated as follows

$$n_{rel.}^{min}(\gamma^*) = \arg \min_{\hat{m} \in \mathbb{N}} \left\{ \frac{z_{\frac{\alpha}{2}} \cdot s / \sqrt{\hat{m}}}{|\bar{y}_S|} \leq \frac{\gamma^*}{1 + \gamma^*} \right\}. \tag{6}$$

Similar to Equation (5), it also applies here that a direct solution with

$$n_{rel.}^{min}(\gamma^*) = \left(\frac{z_{\frac{\alpha}{2}} \cdot s \cdot (1 + \gamma^*)}{|\bar{y}_S| \cdot \gamma^*} \right)^2 \tag{7}$$

is possible, except when using the Student's t distribution.

The particular inequation in Equation (6) can be derived as follows. Given the condition for the absolute error $\delta^* = |\bar{y}_S - \bar{y}_U|$ with $z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{m}} \leq \delta^*$ in $n_{abs.}^{min}(\delta^*)$, it applies

$$z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{m}} \leq |\bar{y}_S - \bar{y}_U| \iff \frac{z_{\frac{\alpha}{2}} \cdot s / \sqrt{m}}{|\bar{y}_S|} \leq \frac{|\bar{y}_S - \bar{y}_U|}{|\bar{y}_S|}. \tag{8}$$

With $\gamma^* = \frac{|\bar{y}_S - \bar{y}_U|}{|\bar{y}_U|}$ and $\frac{\gamma^*}{1 + \gamma^*} = \frac{|\bar{y}_S - \bar{y}_U|}{|\bar{y}_S|}$ when using γ^* , the condition can be written as

$$\frac{z_{\frac{\alpha}{2}} \cdot s / \sqrt{m}}{|\bar{y}_S|} \leq \frac{\gamma^*}{1 + \gamma^*}, \tag{9}$$

which corresponds to Equation (6).

6.2. CNM Precision Validity Score

With the help of the previous definitions, a comparable score is now defined to indicate whether sufficient measurements are available for a given criterion to meet a certain precision. On the one hand, this helps to compare data sets of different sizes, whether the accuracy is statistically different or not. On the other hand, the score can be used to quantitatively indicate for a CNM what percentage of the required measurements were already made to achieve a specified precision.

The measure is defined as follows. Given a target precision t^* , e.g., $t^* = \delta^* = 100$ kbit/s or $t^* = \gamma^* = 0.01$ (i.e., for the latter, the deviation of \bar{y}_S from \bar{y}_U corresponds to maximum 1% of the mean value \bar{y}_U), the *CNM Precision Validity Score* is defined as

$$Val. Score_{prec.}(t^*) = \min \left\{ \frac{m}{n_q^{min}(t^*)}, 1 \right\} \quad (10)$$

with $m = |S|$, $S \subseteq U$, as the number of measurements done within the CNM S with measurement plan p , target precision t^* (δ^* , resp. γ^*), required number of measurements n_q^{min} to meet precision of type q as $q = 'abs.'$ for absolute precision, resp. $q = 'rel.'$ for relative precision, as defined in Section 6.1. It corresponds to the percentage number of measurements in CNM S compared to the number required to achieve the desired precision. The minimum condition within the formula with 1 ($\geq 100\%$) ensures that the score results in $0 < Val. Score_{prec.}(t^*) \leq 1$.

In case enough measurements are contained in the CNM ($\geq 100\%$), the score thus reflects the same value with 1 (=100%). If there are too few measurements for the target precision, it shows what percentage of the measurements are already included until the desired precision is achieved. The validity score is intended to be mentioned additionally in connection with a CNM result to prove the given precision and error margin, for example, in the case of throughput calculations for a region, which are customary in practice.

With a target precision given by the maximum absolute difference, the calculated CNM Precision Validity Score depends largely on the estimated standard deviation of the underlying population by sample S . This means that if the sample size is small, a poor estimate of the actual standard deviation has a significant influence on the score. In case of a calculation of the validity score with relative target precision, the estimate of the mean value of the underlying population also plays a role. In the case of small samples, the estimator for the standard deviation and the mean value can therefore be poor and both can falsify the result. Based on our experience in applying the score with our CNM data, a sample size of at least 100 is recommended in practice to avoid falsified results. A practical example of the influence of a small sample size and its estimators on the validity score can be found in Section 7.

We differentiate between five different types of the CNM Precision Validity Score as listed in Table 2. They differ in the method to estimate the distribution of the statistics for the confidence interval, i.e., when calculating the quantile $z_{\frac{\alpha}{2}}$. This means that they stand out in terms of their requirements, such as their computational effort, the minimum number of measurements required, and whether the data set has to be approximately normally distributed or not.

Depending on the assumed underlying distribution for the measured parameter, $z_{\frac{\alpha}{2}}$ can be derived according to (1) the Standard Normal Interval, (2) the Studentized t Interval, (3) the Basic Bootstrap Confidence Interval, (4) the Percentile Confidence Interval, or (5) the Bias Corrected and Accelerated (BCa) Confidence Interval. Compared to that of method (1) with the Standard Normal Interval, (2) takes into account the correction of the standardized estimator of the sample mean of normally distributed data with a small sample size. (3), (4), and (5) are based on the bootstrapping method. Bootstrapping is generally useful for estimating the distribution of a statistic (e.g., mean, variance) without using normal theory.

Bootstrapping [31,32] accounts for the exact distribution of the underlying measurement parameter and falls under the broader class of resampling methods. This is particularly necessary if arbitrary distributed measured values are obtained from the CNM. Bootstrapping estimates the properties of a distribution by measuring those properties from a sample. Bootstrapping and jackknife methods were proven to be powerful tools for approximating the sample distribution and variance.

The bootstrap values can be determined as follows: (1) B random bootstrap samples are generated, (2) a parameter estimate is calculated from each bootstrap sample, (3) all B bootstrap parameter estimates are ordered from lowest to highest when calculating the Percentile Confidence Interval, and (4) the CI is constructed accordingly.

BCa confidence intervals adjust for skewness in the bootstrap distribution, but since CNMs in particular often have to evaluate huge amounts of data, this can be computationally intensive for many users. According to our preliminary investigations and the given

the size of our data set, the results on our server, namely a Super Micro server with 96 CPU cores and 1008 GB RAM, were practically incalculable. For individual regions with more than 100k measurements, the calculation took more than a day with results without any significant difference compared to the other bootstrapping methods for the downlink throughput in France.

Table 2. Overview of the different versions of the CNM Precision Validity Score. The most relevant method for practice is highlighted in gray.

Name	Confidence Interval Method	⚙️	☑️	⚠️	Remarks
Val. Score (Norm.)	Standard Normal Interval	↓	—	✓	Version with the lowest computational complexity, Normal theory must be applicable
Val. Score (Stud-t)	Studentized <i>t</i> Interval	○	✓	✓	Iterative calculation with correction of the standardized estimator of the sample mean in order to be applicable to few data values
Val. Score (Boot Basic)	Basic Bootstrap Confidence Interval	○	—	—	Basic bootstrapping procedure for taking into account any underlying distribution of the measured values
Val. Score (Boot Perc.)	Percentile Confidence Interval	○	—	—	Bootstrapping where the bootstrap estimates are lined up from lowest to highest
Val. Score (BCa)	Bias Corrected and Accelerated Confidence Interval	↑	—	—	Second-order accurate interval, adjusts for skewness in the bootstrap distribution

⚙️ = Computational effort (↓/○/↑ = low/medium/high), ☑️ = Applicable with few numbers of measurements (✓/— = yes/no), ⚠️ = Normal theory must be applicable to estimate the distribution of the statistics (✓/— = yes/no).

The pseudocode for calculating the CNM Precision Validity Score can be found in Algorithm 1. One input parameter is the type of calculation with $type = \{Normal \mid Stud-t \mid Bootstrap\}$. Basically, this algorithm can be used for all types. If *Stud-t* is used, however, the result is an estimate, as described in Section 6.1. Therefore, a further algorithm is given in Algorithm 2, which allows the exact calculation on the basis of an iterative method. Here, a *repeat..until* loop is used to calculate the exact value of the quantile of the Student’s *t* distribution after every increase in \hat{m} , since it depends on it (cf. degrees of freedom). Note that a further extension for the practical implementation of the algorithms would be a binary search instead of the *repeat..until* loop to be able to determine the parameter \hat{m} more quickly. The given pseudocode is intended to represent the computation in pertinent notation to understand the basic idea for calculation.

Algorithm 1 *Val. Score_{prec.}* with Standard Normal Interval, Studentized *t* Interval (approx. only), or Bootstrapping.

Input: CNM $S \subseteq U$, $m = |S|$, α (signif. level), $type = \{Normal \mid Stud-t \mid Bootstrap\}$, δ^* or γ^* (absolute or relative diff.)

Output: *Val. Score_{prec.}*

- 1: **if** $type == Normal$ **then** ▷ Standard Normal Interval
- 2: Calculate quantile $z_{\frac{\alpha}{2}}$ according to Standard Normal distribution
- 3: **else if** $type == Stud-t$ **then** ▷ Approximation only, Student’s *t* distribution
- 4: Calculate quantile $z_{\frac{\alpha}{2}}$ according to Student’s *t* distribution with m degrees of freedom
- 5: **else if** $type == Bootstrap$ **then** ▷ Bootstrapping methods
- 6: Generate B random bootstrap samples from $S \subseteq U$
- 7: A parameter estimate is calculated from each bootstrap sample
- 8: Parameter estimates are ordered from low to high ▷ If Perc. Conf. Interval
- 9: **end if** ▷ Quantile $z_{\frac{\alpha}{2}}$ was calculated
- 10: $\bar{y}_S \leftarrow \frac{\sum_{i \in S} y_i}{m}$ ▷ Assumption: mean value \bar{y}_S and sample standard deviation s is stable
- 11: $s \leftarrow \sqrt{\frac{1}{m-1} \sum_{i \in S} (y_i - \bar{y}_S)^2}$
- 12: Direct calculation of $n_{abs.}^{min}(\delta^*)$ or $n_{rel.}^{min}(\gamma^*)$ according to Equation (5), resp. Equation (7)
- 13: Calculate *Val. Score_{prec.}* with $\frac{m}{n_q^{min}(t^*)}$ with $q \in \{abs., rel.\}$ according to Equation (10) ▷ Output *Val. Score_{prec.}*

Algorithm 2 *Val. Score_{prec.}* iterative approach with Studentized *t* Interval for CNM *S*.**Input:** CNM $S \subseteq U$ with $m = |S|$, α (significance level), δ^* or γ^* (max. absolute or relative difference)**Output:** *Val. Score_{prec.}*

- 1: $\bar{y}_S \leftarrow \sum_{i \in S} \frac{y_i}{m}$ ▷ Assumption: mean value \bar{y}_S and sample standard deviation s is stable
- 2: $s \leftarrow \sqrt{\frac{1}{m-1} \sum_{i \in S} (y_i - \bar{y}_S)^2}$ (otherwise re-calculate it inside *repeat...until* with \hat{m} loop after incrementing \hat{m})
- 3: Temporary \hat{m} with $\hat{m} \leftarrow 0$ (or reasonable low start value m_0)
- 4: **repeat** ▷ Determine min. measurements required: $n_{rel.}^{min}$
- 5: **Increment** \hat{m} (since condition has not yet been reached)
- 6: Calculate $z_{\frac{\alpha}{2}}$ according to Student's *t* distribution with \hat{m} degrees of freedom
- 7: **until** $z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{\hat{m}}} \leq \delta^*$, resp. $\frac{z_{\frac{\alpha}{2}} \cdot s / \sqrt{\hat{m}}}{|\bar{y}_S|} \leq \frac{\gamma^*}{1+\gamma^*}$ ▷ Depending on or absolute (δ^*) or relative precision (γ^*)
- 8: $n_q^{min} \leftarrow \hat{m}$
- 9: Calculate *Val. Score_{prec.}* with $\frac{m}{n_q^{min}(t^*)}$ according to Equation (10) ▷ Output *Val. Score_{prec.}*

7. Practical Application of the CNM Precision Validity Score

To show the practical applicability of the CNM Precision Validity Score, we first discuss the score for different sample sizes. For our data set, the mean downlink throughput is 23.83 Mbit/s, having a standard deviation of 19.56 Mbit/s and a maximum of 167.95 Mbit/s. To precisely quantify the effect for this data set, in addition to the mean value of the sample, the standard error of mean, the confidence interval, the required minimum number of samples $n_{abs.}^{min}(\delta^*)$, resp. $n_{rel.}^{min}(\gamma^*)$, and the corresponding *Val. Score_{prec.}* are evaluated for a confidence level of 95% ($\alpha = 0.05$) of exemplary sample sizes from 10–10,000,000 measurements of the data set with Studentized *t* Intervals in Table 3.

If a precision of $\delta^* = 1$ Mbit/s is desired, the table shows how many measurements are needed to fulfill this precision: For $n_{abs.}^{min}(\delta^*) \geq 1491 \Rightarrow z_{0.025} \cdot \frac{s}{\sqrt{m}} \leq 0.99$ and thus, the precision is higher than 1 Mbit/s. The validity score shows the number of measurements made as a percentage of the measurements required for the desired precision. Once the precision was reached, the validity score is ' $\geq 100\%$ '.

If, instead, you prefer to tolerate at most a relative error of $\gamma^* = 1\%$, the following condition must hold: $\frac{z_{0.025} \cdot s / \sqrt{m}}{\bar{y}} \leq \frac{0.01}{1+0.01} = 0.0099$. In our example, this condition is fulfilled for $n_{rel.}^{min}(\gamma^*) \geq 26,576$. Thus, in this case, a sample with 26,576 measurements would lead to a high accuracy of at most 1% inaccuracy relative to the exact mean value when evaluating the mean.

Table 3. Precision of different sample sizes. *Validity Score_{prec.}* for $\delta^* = 1$ Mbit/s or $\gamma^* = 1\%$ with Studentized *t* Intervals.

Sample Size m	\bar{y}_S	$\frac{s}{\sqrt{m}}$	$CI_{0.05}$	$z_{0.025} \frac{s}{\sqrt{m}}$	$n_{abs.}^{min}(\delta^*)$	<i>Val. Score_{prec.}</i> (abs.: δ^*)	$n_{rel.}^{min}(\gamma^*)$	<i>Val. Score_{prec.}</i> (relative: γ^*)
10	33.54	9.26	[12.59; 54.49]	20.96	4390.61	0.002 (0.2%)	39,823.66	<0.001 (<0.1%)
100	21.20	1.84	[17.54; 24.86]	3.66	1338.36	0.074 (7.4%)	30,368.21	0.003 (0.3%)
1000	23.89	0.62	[22.67; 25.11]	1.22	1473.70	0.679 (67.9%)	26,348.74	0.038 (3.8%)
1491	23.51	0.51	[22.52; 24.50]	($\delta^* = 1$ Mbit/s) 0.99	1489.89	$\geq 100\%$	27,491.78	0.054 (5.4%)
10,000	23.51	0.19	[23.13; 23.89]	0.38	1462.00	$\geq 100\%$	26,976.23	0.371 (37.1%)
26,576	23.57	0.12	[23.34; 23.80]	($\gamma^* = 1\%$) 0.23	1447.80	$\geq 100\%$	26,575.75	$\geq 100\%$
100,000	23.82	0.06	[23.70; 23.94]	0.12	1473.98	$\geq 100\%$	26,495.38	$\geq 100\%$
1,000,000	23.85	0.02	[23.81; 23.89]	0.04	1473.52	$\geq 100\%$	26,431.12	$\geq 100\%$
10,000,000	23.83	0.01	[23.82; 23.84]	0.01	1470.10	$\geq 100\%$	26,405.05	$\geq 100\%$

$\bar{y}_S, \frac{s}{\sqrt{m}}, CI_{0.05}, z_{0.025} \frac{s}{\sqrt{m}}$ are in [Mbit/s].

The values of $n_{abs.}^{min}(\delta^*)$ and $n_{rel.}^{min}(\gamma^*)$ differ. This is especially the case with a small sample size compared to that of a larger sample size. As described in Section 6.2, this is due to the estimation of the standard deviation from the different samples for the absolute case and to the estimation of the standard deviation and the mean value in the relative case. With more than 100 measured CNM values, however, the result significantly improves here. We therefore point out the uncertainty of the validity score for small sample sizes and recommend its calculation for larger sample sizes. Furthermore, based on the estimation of the parameters of the population, a certain small difference might always occur. In practice, CNMs with a very small number of measurements are rarely carried out, so the score should be suitable for practical use.

To illustrate the defined measure, Figure 2 shows the number of measurements, precision in terms of the confidence interval, and validity scores for $\delta^* = 100$ kbit/s for selected departments and islands of France. The subfigures are arranged according to the number of available measurement results in the specified area. Each point in the upper map of each subfigure represents a CNM measurement result. Below are the key figures that correspond to the figures defined in the paper in terms of precision. The validity score is given according to the Standard Normal Interval, Studentized t Interval, and various bootstrapping methods to show the differences.

In the top row, departments with sufficient measurements are shown. All regions have a validity score of $\geq 100\%$ and precision according to the confidence interval of at most $\bar{y}_S \pm 50$ kbit/s for the average throughput. In the bottom row, islands and departments are listed for which too few measurements are available to determine the average downlink value with a precision of at least 100 kbit/s. For Île-de-France, for example, 8.5 million measurement results are available, which is sufficient to ensure that the real average throughput is within a confidence interval of $\bar{y}_S \pm 13.3$ kbit/s. The mean value of the measurements is 23.78 Mbit/s. More precisely, according to the statistical analysis based on the 8.5 million samples and their distribution, the real mean of the population is actually between 23.76 Mbit/s and 23.79 Mbit/s (confidence level $\alpha = 0.05$). This range is less than 100 kbit/s, which results in a validity score of ' $\geq 100\%$ '. For Brittany, 421k measurements are available, which corresponds to a confidence interval of $\bar{y}_S \pm 52.1$ kbit/s. Here, the mean value \bar{y}_S is 21.61 Mbit/s but it can only be limited to 21.556 Mbit/s to 21.660 Mbit/s according to the Studentized t Interval. This corresponds to a range of 104.20 kbit/s, which is above the desired precision. For this reason, the subfigure has a yellow background. The validity score shows what percentage of the measurement results were obtained to maintain the precision. In this case, about 7–10% of the measurements are missing for Brittany, depending on the calculation method and estimate of the standard deviation from the sample. In the case of the bottom row, we recommend taking more measurements to maintain the desired precision and to ensure comparability between the different regions in terms of the average throughput.

The validity score indicates the relative number of measurements required to achieve the desired precision. The description with the key figures below each map also shows the differences between the individual calculation methods of the validity score. The score varies from 89.96% for Bootstrapping Basic to 92.14% for Studentized t Intervals for Brittany, for example. In general, it makes the most sense to trust the bootstrapping intervals, as they provide a good estimate for the underlying variance of the distribution of the population. However, bootstrapping methods are slower to compute. Based on our experience and the small differences in practice, we recommend using the method with Standard Normal Interval to calculate the validity score.

In the following, we present another practical example. We show that the calculation of the average downlink throughput for France is (almost) possible for regions, but not at the departmental level, based on our given data set. At the departmental level, the calculation is only valid for large cities like Paris or Lyon if you want to maintain the precision of $\delta = 100$ kbit/s.

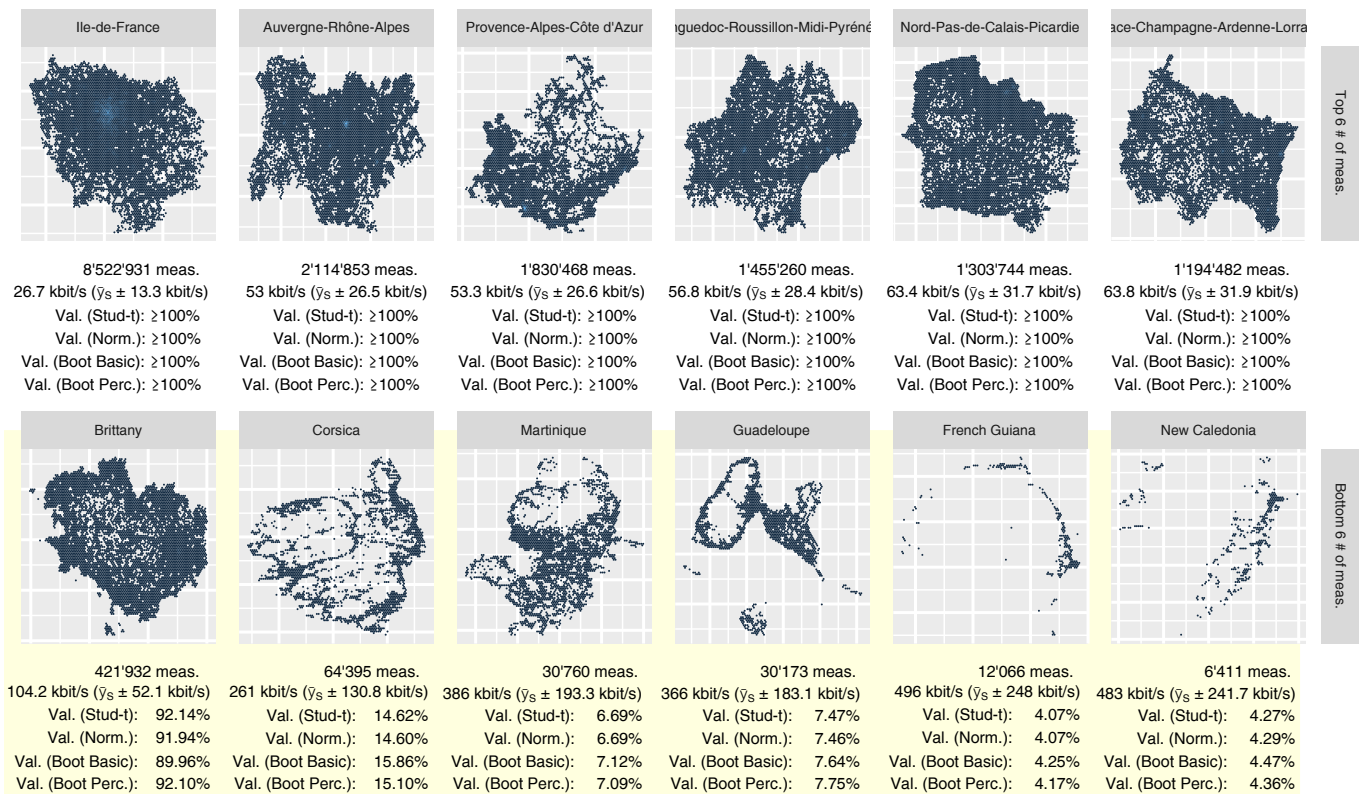


Figure 2. Amount of measurements, precision, and validity scores for $\delta^* = 100$ kbit/s for selected regions and islands of France. In the top row, regions with sufficient measurements are shown. All regions have a validity score of $\geq 100\%$ and precision according to an interval with at most $\bar{y}_S \pm 50$ kbit/s for average throughput in CNM data S . In the bottom row, islands and regions are listed for which too few measurements are available to determine average downlink value with an precision of at least 100 kbit/s.

In Figure 3, the scores are calculated once for the larger regions (left) and once for individual departments (right) in France. Everything is shown on two maps next to each other to highlight where enough measurements were made for which case. The figure depicts the average downlink throughput with different colors. All areas with sufficient measurements are colored in blue, as the precision here corresponds to the target value of at least $\delta^* = 100$ kbit/s, i.e., the actual average throughput lies within an interval of 100 kbit/s.

In this case, the $Val. Score_{prec.}$ is calculated with Bootstrapping and Percentile Confidence Intervals to take into account the real distribution of all downlink measurements in our data set for a region or department. The annotations include the number of measurements and the associated validity score in percent, which shows whether enough CNM samples were taken or whether more measurements need to be made to avoid throughput calculations with large ranges. Especially for the investigation of the average downlink throughput according to departments, the number of measurements is not sufficient to guarantee the precision due to the smaller division compared to regions. As a result, there are enough measurements available for the analysis by region, but not for the analysis by department.

The Auvergne-Rhône-Alpes region is the third largest region, and it contains, for example, 13 individual departments. Our data set contains 21 million throughput measurements for the entire region. The validity score is ' $\geq 100\%$ ' with a mean throughput value according to the measurement samples of 25.49 Mbit/s and a calculated mean value between 25.46 Mbit/s and 25.51 Mbit/s. For department Ain, 111k measurements are available, with a sample average throughput of 22.37 Mbit/s. If you calculate $n_{abs.}^{min}(\delta^*)$ here, about 509k measurements are required to maintain the precision. The validity score

is 21.80%, which indicates that too few measurements were made for the consideration at the department level.

The right subfigure shows three more departments in the southwest of France that have too few measurements; see yellow labels. Only the departments around Paris, Lille, Lyon, and Marseilles contain enough data to maintain the target precision at the departmental level.

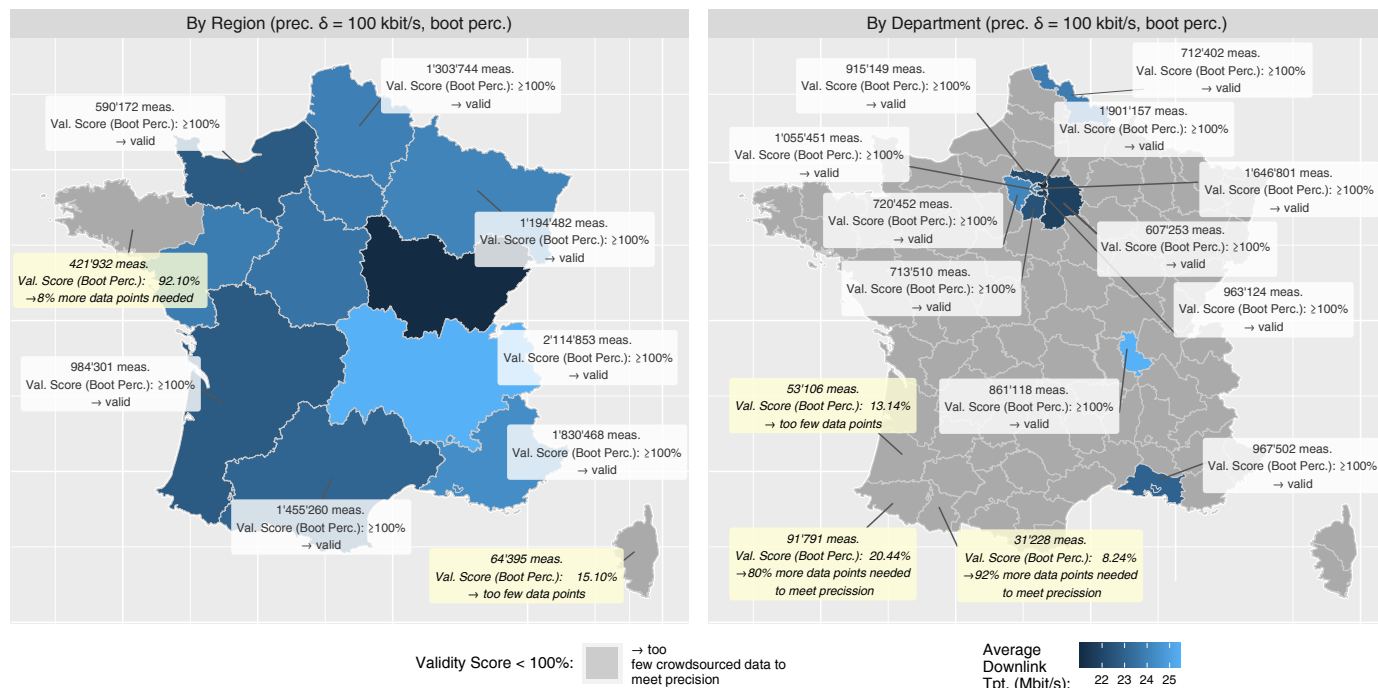


Figure 3. Depiction of the average downlink throughput for regions (left) and departments (right). All areas with sufficient measurements are colored, as precision there corresponds to target value of at least $\delta^* = 100$ kbit/s, i.e., real average throughput lies actually within an interval of 100 kbit/s. Annotations include for selected areas the number of measurements and associated validity score in percent, which shows whether enough CNM samples were taken or whether more measurements need to be made to avoid throughput calculations with large error margins.

8. Conclusions

When using crowdsourced network measurements (CNMs), network operators, regulators, and big data companies are faced with the challenge of making valid statements out of measurements in uncontrolled environments of the crowd. There is always the question of validity of such measurements, as the temporal and spatial coverage as well as the total number of measurements can fluctuate strongly. Thus, this article defines concepts and guidelines for analyzing the validity of crowdsourced mobile network measurements with statistical measures.

We consider CNMs to be a mathematical sampling process and, as a result, derive from this the need for high-precision and validity. Therefore, we define a measure called CNM Precision Validity Score to indicate whether a sufficient number of measurements is available for a sample statistic like the mean throughput to meet a certain precision. This score can be used to quantify what percentage of the required CNMs were already conducted to achieve a specified precision. To satisfy different types of measurements, e.g., small number of samples or skewed data distributions, we present different versions of the CNM Precision Validity Score, including different confidence interval methods, namely Standard Normal Interval, the Studentized *t* Interval, the Basic Bootstrap Confidence Interval, and the Percentile Confidence Interval. In addition to the theoretical background of the score, we illustrate its applicability by applying it to a large CNM dataset. Using the example of a data set from France, we showed for which regions the data are sufficient to

achieve an accuracy of at least 100 kbit/s. We show that the measurements are sufficient for regions, but not at the departmental level. For the consideration of individual departments, more measurements need to be made to achieve the same precision.

In future work, we would like to further explore the methodology of evaluating CNM data and define metrics for the representativeness, e.g., the spatial and temporal distribution of the data. This could later be used to write comprehensive guidelines on how to deal best with CNM data.

Author Contributions: Conceptualization, F.W. and A.S.; methodology, F.W., A.S. and T.H.; software, F.W.; validation, F.W., A.S., A.H., S.W. and T.H.; formal analysis, F.W., A.S. and T.H.; investigation, F.W. and A.S.; resources, A.H.; data curation, F.W., A.S. and A.H.; writing—original draft preparation, F.W. and A.S.; writing—review and editing, A.H., S.W. and T.H.; visualization, F.W. and A.S.; supervision, A.H., S.W. and T.H.; project administration, A.H. and S.W.; funding acquisition, A.H., S.W. and T.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Bavarian Ministry of Economic Affairs, Regional Development and Energy under grant number DIK-2010 5GQMON within the framework “5G” of the BAYERN DIGITAL strategy of the Information and Communication Technology R&D program of the State of Bavaria.

Data Availability Statement: 3rd Party Data. Restrictions apply to the availability of these data. Data was obtained from Tutela Technologies, Ltd., Victoria, BC, Canada.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hoßfeld, T.; Wunderer, S.; Beyer, A.; Hall, A.; Schwind, A.; Gassner, C.; Guillemin, F.; Wamser, F.; Wascinski, K.; Hirth, M.; et al. *White Paper on Crowdsourced Network and QoE Measurements—Definitions, Use Cases and Challenges*; Technical Report 10.25972/OPUS-20232; University of Würzburg: Würzburg, Germany, 2020.
2. Howe, J. The Rise of Crowdsourcing. *Wired Magazine*, 14 June 2006; pp. 1–4.
3. Hirth, M.; Hoßfeld, T.; Mellia, M.; Schwartz, C.; Lehrieder, F. Crowdsourced Network Measurements: Benefits and Best Practices. *Comput. Netw.* **2015**, *90*, 85–98. [CrossRef]
4. 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: Geneva, Switzerland, 1994.
5. Dano, M. Who Won the Super Bowl Speed Game? It Depends on Who You Ask. 2019. Available online: <https://www.lightreading.com/testing/who-won-the-super-bowl-speed-game-it-depends-on-who-you-ask/d/d-id/749241> (accessed on 1 January 2021).
6. Midoglu, C.; Svoboda, P. Opportunities and challenges of using crowdsourced measurements for mobile network benchmarking a case study on RTR open data. In Proceedings of the 2016 SAI Computing Conference (SAI), London, UK, 13–15 July 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 996–1005.
7. Seufert, A.; Wamser, F.; Wunderer, S.; Hall, A.; Hoßfeld, T. Trust but Verify: Crowdsourced Mobile Network Measurements and Statistical Validity Measures. In Proceedings of the 2021 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit), Porto, Portugal, 8–11 June 2021.
8. Hossfeld, T.; Keimel, C.; Hirth, M.; Gardlo, B.; Habigt, J.; Diepold, K.; Tran-Gia, P. Best Practices for QoE Crowdttesting: QoE Assessment with Crowdsourcing. *IEEE Trans. Multimed.* **2013**, *16*, 541–558. [CrossRef]
9. Hoßfeld, T.; Hirth, M.; Redi, J.; Mazza, F.; Korshunov, P.; Naderi, B.; Seufert, M.; Gardlo, B.; Egger, S.; Keimel, C. *Best Practices and Recommendations for Crowdsourced QoE—Lessons Learned from the Qualinet Task Force “Crowdsourcing”*; Technical Report hal-01078761; Hindustan Aeronautics Limited: Nantes, France, 2014.
10. Neidhardt, E.; Uzun, A.; Bareth, U.; Küpper, A. Estimating Locations and Coverage Areas of Mobile Network Cells based on Crowdsourced Data. In Proceedings of the 6th Joint IFIP Wireless and Mobile Networking Conference (WMNC), Dubai, United Arab Emirates, 23–25 April 2013; pp. 1–8.
11. Li, Z.; Nika, A.; Zhang, X.; Zhu, Y.; Yao, Y.; Zhao, B.Y.; Zheng, H. Identifying Value in Crowdsourced Wireless Signal Measurements. In Proceedings of the 26th International Conference on World Wide Web, Perth, Australia, 3–7 April 2017; pp. 607–616.
12. Wamser, F.; Seufert, M.; Casas, P.; Irmer, R.; Tran-Gia, P.; Schatz, R. YoMoApp: A tool for Analyzing QoE of YouTube HTTP Adaptive Streaming in Mobile Networks. In Proceedings of the European Conference on Networks and Communications (EuCNC), Paris, France, 29 June–2 July 2015; pp. 239–243.
13. Schwind, A.; Midoglu, C.; Alay, Ö.; Griwodz, C.; Wamser, F. Dissecting the Performance of YouTube Video Streaming in Mobile Networks. *Int. J. Netw. Manag.* **2020**, *30*, e2058. [CrossRef]

14. Schwind, A.; Wamser, F.; Wunderer, S.; Gassner, C.; Hoßfeld, T. Mobile Internet Experience: Urban vs. Rural—Saturation vs. Starving? *arXiv* **2019**, arXiv:1909.07617.
15. Schwind, A.; Wamser, F.; Hossfeld, T.; Wunderer, S.; Tarnvik, E.; Hall, A. Crowdsourced Network Measurements in Germany: Mobile Internet Experience from End User Perspective. In Proceedings of the Broadband Coverage in Germany (14. ITG Symposium), Berlin, Germany, 23–24 March 2020; pp. 1–7.
16. 4Gmark. Data Mobile Barometer. White Paper. Available online: <https://www.5gmark.com/news/2017/4Gmark-Barometer-2017-DE.pdf> (accessed on 28 February 2018).
17. Tutela Technologies. *Germany Mobile Experience Report—Country-Level Mobile Experience and Usage Results from Tutela’s Crowdsourced Mobile Network Testing (May–July 2019)*; Technical Report; Tutela Technologies: Victoria, BC, Canada, 2019.
18. P3 Communications. The Great 2019 Mobile Network Test. Available online: http://p3-networkanalytics.com/wp-content/uploads/2018/12/Network-Test-2019-connect-2019-01-English_sp.pdf (accessed on 15 January 2019).
19. Opensignal. Germany—Erfahrungsbericht mit mobilem Netzwerk (November 2019). Available online: <https://www.opensignal.com/de/reports/2019/11/germany/mobile-network-experience> (accessed on 15 November 2019).
20. Bischof, Z.S.; Otto, J.S.; Sánchez, M.A.; Rula, J.P.; Choffnes, D.R.; Bustamante, F.E. Crowdsourcing ISP Characterization to the Network Edge. In *First ACM SIGCOMM Workshop on Measurements up the Stack*; ACM: New York, NY, USA, 2011; pp. 61–66.
21. Raida, V.; Svoboda, P.; Lerch, M.; Rupp, M. Crowdsensed performance benchmarking of mobile networks. *IEEE Access* **2019**, *7*, 154899–154911. [CrossRef]
22. Rich, R.; Brians, C.; Willnat, L. *Empirical Political Analysis: Quantitative and Qualitative Research Methods*; Routledge: Abingdon-on-Thames, UK, 2018.
23. Eid, M.; Schmidt, K. *Testtheorie und Testkonstruktion*; Hogrefe: Göttingen, Germany, 2014.
24. Chow, S.; Liu, J. *Design and Analysis of Clinical Trials: Concepts and Methodologies*; Wiley Series in Probability and Statistics; Wiley: Hoboken, NJ, USA, 2004.
25. Messick, S. Validity. *ETS Res. Rep. Ser.* **1987**, *1987*, i-208. [CrossRef]
26. Stuart, A. *Basic Ideas of Scientific Sampling*; Griffin’s Statistical Monographs; Hafner Press: Royal Oak, MI, USA, 1976.
27. Deutsches Institut für Normung. *Qualitätsmanagement und Statistik: Normen. Begriffe*; Number Bd. 1 in DIN-Taschenbuch; Beuth: Berlin, Germany, 2001.
28. Wirtz, M.; Strohmmer, J. *Dorsch-Lexikon der Psychologie*; Dorsch—Lexikon der Psychologie, Hogrefe: Göttingen, Germany 2016.
29. Adcock, R.; Collier, D. Measurement Validity: A Shared Standard for Qualitative and Quantitative Research. *Am. Political Sci.* **2001**, *95*, 529–546. [CrossRef]
30. Law, A.M.; Kelton, W.D. *Simulation Modeling and Analysis*; McGraw-Hill: New York, NY, USA, 2000; Volume 3.
31. Efron, B. *The Jackknife, the Bootstrap and Other Resampling Plans*; SIAM: Philadelphia, PA, USA, 1982.
32. Efron, B. Bootstrap methods: Another look at the jackknife. In *Breakthroughs in Statistics*; Springer: Berlin/Heidelberg, Germany, 1992; pp. 569–593.