On the Inappropriateness of Static Measurements for Benchmarking in Wireless Networks

Vaclav Raida, Philipp Svoboda, Markus Rupp
Institute of Telecommunications
Technische Universitt Wien
Vienna, Austria
{firstname.lastname}@nt.tuwien.ac.at

***Abstract*—A state-of-the-art method of characterizing a mobile network operator's performance or of benchmarking multiple operators is to measure the achievable throughput along a particular route. It does not matter if the measuring devices are mounted on a train, car, bus, or truck. During longer measurement campaigns, there will undoubtedly be some intervals, during which the vehicle remains stationary. In these static phases, the impact of small-scale fading on the throughput is uneven.**

Although small-scale fading is a well-known phenomenon, the body of literature that examines its impacts on the measured throughput is scarce. Our lab experiments in LTE have shown that the throughput differences between nearby points reach tens of Mbit/s while the small-scale throughput pattern remains stable over several days.

Furthermore, we examine the impact of static measurements on the throughput distributions collected from several LTE drive tests. We find that the small-scale patterns remain stable even in an outdoor environment; thus, the static samples introduce a significant bias in the overall distribution. If we exclude the zero-speed intervals, then we obtain identical distributions from two receivers, which are 20 cm apart.

***Index Terms*—LTE, measurement, benchmarking, throughput, fading, small-scale, shadowing, path loss, decorrelation, drive test, live network, cellular, mobile**

I. Introduction

A typical way of characterizing the performance of a mobile network operator (MNO) or of benchmarking multiple MNOs is by measuring the throughput repeatedly along a certain route—sometimes measuring on foot, but more often while onboard a vehicle. The more recent methods include crowdsourced measurements [1]–[3] and dedicated distributed measurement platforms [4], with some of the measurement devices aboard public transport vehicles. All of these solutions have one thing in common: besides dynamic measurements, they also contain static measurements. Car drivers need to take breaks to refuel and to refresh, trains and buses spend considerable amount of time at stops and stations, and users of the crowdsourcing tools conduct their measurements under various conditions.

In this paper, we demonstrate that due to small-scale fading,static measurements are not representative to characterize network coverage; thus, static measurements cannot be used for any kind of operator benchmarking or measurement tool benchmarking. We validate our conclusions using vehicular LTE measurements, and measure MNO's throughput with two identical receivers simultaneously.

Numerical analysis reveals that with static samples, the throughput distributions differ significantly. After removing the static intervals, the distributions are evaluated as equal.

At LTE wavelengths with short sampling period (0.5 s), a slow movement already leads to an improvement as the small-scale pattern evens out. Although we validate our assumptions only for two different LTE frequencies, the impact of static measurements will be similar in any wireless network. The only difference is that the minimum speed (or the sampling period) required to suppress the small-scale effects increases at lower frequencies (longer wavelengths).

A. Related Work

Researchers often evaluate the performance of wireless networks in terms of empirical throughput distributions. T. Berisha et al. [5] measured, among others, the throughput distributions of two MNOs in UMTS and LTE networks while onboard a high-speed train. J. Landre et al. [6] performed LTE drive tests in two urban scenarios and compared the measured throughput distributions with the outcome of simulated channels. J. Beyer et al. [7] collected drive test data in a trial heterogeneous LTE network consisting of one macrocell and eight picocells; they compared the throughput distributions of different cells.

The presence of static measurements is also relevant for distributed measurement platforms like MONROE [4], which contains tens of LTE nodes—many of them are mounted on buses, trains, and trucks—in several European countries.

In dense traffic, where traffic lights and pedestrian crossings are present, a nonnegligible portion of the drive test time is likely to be spent in a steady state. For example, trains stop for several minutes at each station to wait for passengers. In some cases, the impact of static measurements on the overall distribution may be small enough to be ignored. In other cases, the static measurements can result in a bias that leads to, for example, an unfair MNO comparison or to a worse match with a channel model. In the above-mentioned publications, it is difficult to asses the severity of the problem because the authors do not give enough details about the measurement duration and speed distribution.

Some campaigns are based on static measurements per-design. A. Khatouni et al. [8] benchmarked the performance of LTE MNOs in four countries by comparing the throughput

distributions obtained from static nodes. Specifically, they used about 7–28 measurement nodes for every MNO. It is not clear, however, whether such a count of static locations yields a meaningful benchmark or whether the differences among the MNOs are only random because of the temporarily stable small-scale fading. No detailed placement strategy was specified.

To our knowledge, no publication dealing with performance benchmarking in wireless networks has considered the impact of the zero-speed intervals in the analysis.

In [9], we performed repeated measurements on an xy-table to examine small-scale variations of measured wireless signal characteristics. In [10], we introduced the throughput "ground truth" reconstruction, and approximated the throughput that a user equipment (UE) would achieve if it would get all resources in the given cell; this then has allowed us to measure the throughput with multiple UEs in parallel.

II. Small-Scale Spatial Throughput Variations

As we have experienced in our previous static LTE 800 and LTE 2600 downlink (DL)[1] measurements [9], the throughput varies significantly due to small-scale fading (caused by interference of multipath components). Fig. 1 presents the throughput measured in LTE 800 on an xy-table with the dimensions of 75×75 cm and a step size of 1 cm ($76^2 = 5776$ measurement points in each repetition).

At first, we can see that the throughput can change dramatically by shifting the measuring UE by a few centimeters. In Fig. 1, we obtain up to 30 Mbit/s differences for points that are not more than 20 cm apart. We achieve 50% decorrelation [9] by moving about 0.3 wavelengths λ. [2]

Secondly, we see that the small-scale throughput pattern remains stable over time; after one week, we still observed a very similar pattern in our indoor environment. What varies throughout the day is the interference and a spatially constant (location-independent) part caused by the cell load [12]—both can be modeled as a cyclo-stationary process superimposed on the time-invariant pattern.

A. Averaging out the Small-Scale Effects

In dynamic measurements, the small-scale fading can be averaged out if the throughput sampling period, i.e., the interval over which the throughput is averaged, is larger than the channel coherence time T_c [13, (4.40.c)]

$$T_c = \sqrt{\frac{9}{16\pi f_m^2}}, \tag{1}$$

where $f_m = \frac{v}{c} f$ is the maximum Doppler spread.

At a pedestrian speed $v = 5\,\mathrm{km/h}$, the coherence times are 0.11 s and 0.03 s in LTE 800 and LTE 2600, respectively. At 20 km/h, we obtain $T_{c,\,\mathrm{LTE\,800}} \approx 29\,\mathrm{ms}$ and $T_{c,\,\mathrm{LTE\,2600}} \approx 9\,\mathrm{ms}$. These values are relevant in Section III-B, where we compare them with the sampling period T_s.

[1] For the exact frequencies, see [11].

[2] In LTE 800, $\lambda \approx 37$ cm. In LTE 2600, $\lambda \approx 11$ cm.

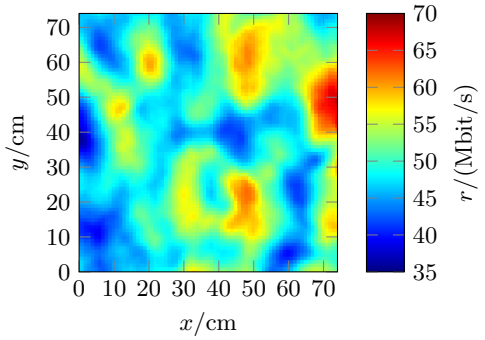

Fig. 1. IP-layer DL throughput r measured by FLARP [14] on an xy-table, as described in [9]. LTE 800; 2×2 MIMO; indoor environment (lab at the Institute of Telecommunications, TU Wien); 22 November, 2017.

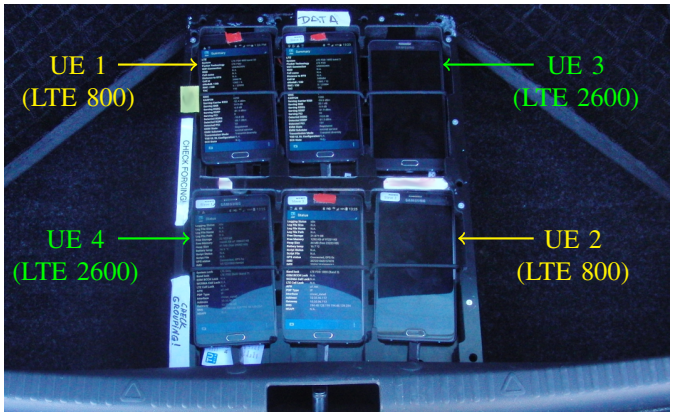

Fig. 2. A NEMO backpack in the car trunk fixed with a rubber band. Two phones were locked to LTE 800 (band 20) and another two to LTE 2600 (band 7). The remaining two phones were not used during this measurement campaign.

B. Limits of Static Measurements for Benchmarking

At a single point during the static measurements, we consistently measure a higher or lower throughput depending on a relative position. Such static measurement does not reflect the MNO's coverage. To obtain a reasonable estimate that represents an operator's performance, we need to sample multiple points in a given area.

Moreover, benchmarking different MNOs is not possible with samples from a singular static point. Since each MNO uses its own base station antennas, we necessarily observe a different small-scale throughput pattern for each MNO.

Let us assume that we measure 110 Mbit/s for MNO A and 70 Mbit/s for MNO B at a given point. The 40 Mbit/s difference can be completely random (a peak in the small-scale map of MNO A coinciding with a minimum in the small-scale map of MNO B).

By moving a little bit, we may obtain an opposite result. We then need to systematically sample an area that is big enough to allow us to compare without bias the performance of MNO A with that of MNO B based on static measurements (similar to that done in Fig. 1 and [9]).

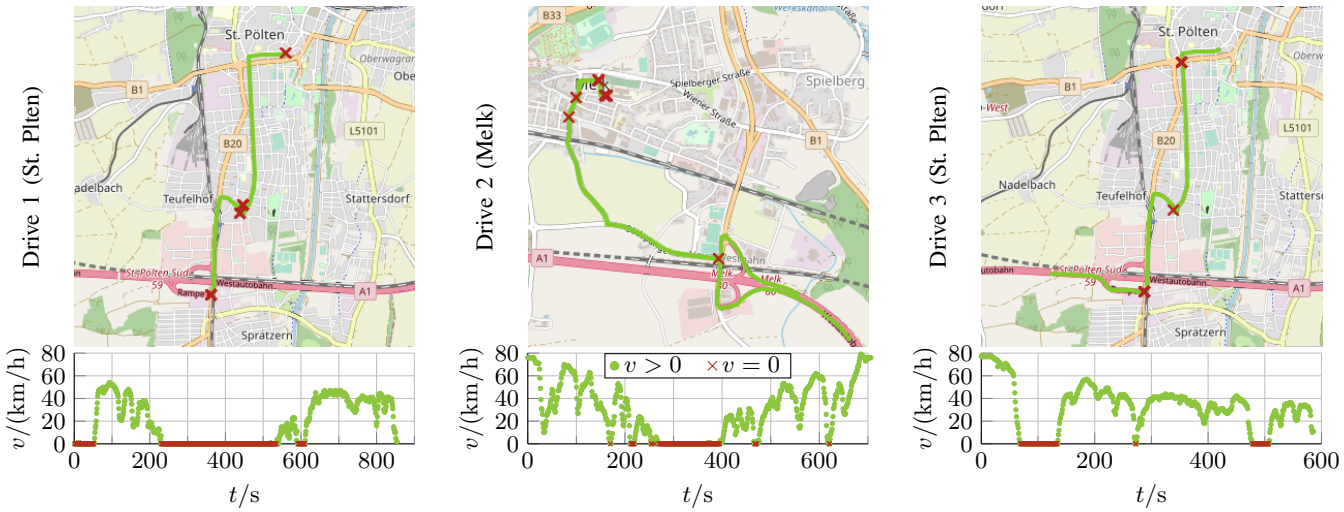

Fig. 3. Three measurement drives. In each drive, the static intervals are marked with red "×" symbols in both map tiles and speed profiles. See also Table I.

TABLE I
SUMMARY OF ALL THREE MEASUREMENT DRIVES TOGETHER WITH THE RESULTS OF THE KOLMOGOROV-SMIRNOV TEST

Drive	Measurement start	Duration / min	All samples	Static samples	UE 1 & 2 (LTE 800)						UE 3 & 4 (LTE 2600)					
					All samples			Nonstatic only			All samples			Nonstatic only		
					$\mathcal{H}$	d_{KS}	p	$\mathcal{H}$	d_{KS}	p	$\mathcal{H}$	d_{KS}	p	$\mathcal{H}$	d_{KS}	p
1	14 Jan 2019, 15:32	14.2	1686	45%	$\mathcal{H}_1$	0.08	0.0%	$\mathcal{H}_0$	0.04	40.5%	$\mathcal{H}_1$	0.15	0.0%	$\mathcal{H}_0$	0.04	57.4%
2	17 Jan 2019, 13:49	11.8	1399	22%	$\mathcal{H}_1$	0.14	0.0%	$\mathcal{H}_0$	0.03	51.8%						
3	17 Jan 2019, 14:17	9.8	1160	17%	$\mathcal{H}_1$	0.06	1.7%	$\mathcal{H}_0$	0.03	89.2%	$\mathcal{H}_1$	0.07	0.3%	$\mathcal{H}_0$	0.03	89.4%

$\mathcal{H}_0$ = null hypothesis: data come from the same continuous distribution. If $p \geq \alpha$, $\mathcal{H}_0$ cannot be rejected.
$\mathcal{H}_1$ = alternative hypothesis: data come from two different distributions. If $p < \alpha$, $\mathcal{H}_0$ is rejected in favor of $\mathcal{H}_1$.
Note: In Melk (drive 2), there was no LTE 2600 coverage for the given MNO.

III. Drive Tests: Measurement Setup

In the previous section, we discussed the throughput variations caused by small-scale fading. What we want to characterize here when we examine mobile network coverage are the effects caused by large-scale fading (path loss and shadowing).

We conducted all the measurements in the LTE network of an Austrian operator. We picked the lowest and the highest available frequency: LTE 800, band 20, and LTE 2600, band 7 (consistent with [9]). The measuring UEs used in Fig. 2 were Keysights NEMO phones [15] with two receive antennas (i.e., 2×2 MIMO). The channel bandwidth was set at 20 MHz.

We took three measurement drives in Lower Austria (one of the nine states in Austria)—two in the city of Sankt Plten and another one in Melk.[3] Fig. 3 depicts the GPS coordinates and the speed profiles of all three drives. Table I presents the time ranges and other details together with the numerical evaluation, which is explained in Section IV-B.

A. Logging, Scheduling, and "Ground Truth" Throughput

From the quantities logged by the NEMO cell phones, we are interested in the following: resource block (RB) utilization[4] $\eta[k]$; MAC layer throughput[5] $r[k]$; and GPS latitude, longitude, and speed v. The RB utilization and the throughput are logged approximately every $T_s = 0.5\,s$ ($\pm 20\,\mathrm{ms}$), whereas the GPS data are logged every ca. 1 s ($\pm 50\,\mathrm{ms}$).

We then generate the DL throughput through continuous and uninterrupted HTTP download (a 40 GB file) using a high-speed server (gigabit connection; located at Institute of Telecommunications, TU Wien). Thus, both the measuring UEs, which use the same band (UE 1 and 2 in LTE 800, UE 3 and 4 in LTE 2600), share their resources all the time. There is no guarantee, however, that they get the same amount of resources—their throughputs can differ significantly.

In this context, the "ground truth" throughput reconstruction (as introduced in [10]) is an essential concept that allows us to estimate the throughput that would be achieved by the measuring UE if it would get all the resources, i.e., if it would be the only active user in the cell.

According to [10], we estimate the "ground truth" throughput $r_{\mathrm{GT}}[k]$ at the kth interval

$$r_{\mathrm{GT}}[k] \approx \frac{r[k]}{\eta[k]}. \tag{2}$$

[3] In fact, these measurements were part of a larger measurement campaign. We selected all the measurement intervals that contain static measurements during instances such as safety breaks, refueling, or waiting at traffic lights.

[4] This denotes the quotient of the number of RBs scheduled by an eNodeB to the measuring UE and the number of all RBs transmitted by this eNodeB (in the measured LTE channel during the kth measurement interval). See [10].

[5] Medium access control layer throughput, see [16] and [10].

As opposed to the measured $r[k]$, the reconstructed throughputs $r_{\mathrm{GT}}[k]$ should be approximately equal in both UEs.

If the RB utilization is not known (conventional UEs do not report this quantity), we can then attempt to compare the throughputs by slotting the measurements into disjoint time intervals (as we did in [17]). In certain slots, only UE 1 measures; in others, only UE 2 measures—that is, however, out of the scope of this paper. In general, the duration of the slots, together with the UEs' speed and slot assignment strategy, will play a role. Moreover, the activity of the other users in the cell will still impact the measurements.

B. Measuring Network Coverage

The shadowing decorrelation distance reaches tens to hundreds of wavelengths (J. Weitzen and T. J. Lowe [18] report values from 25 m to 100 m at 1900 MHz). The measuring UEs are $d \approx 20\,\mathrm{cm}$ apart (Fig. 2). Since d is much smaller than the shadowing decorrelation distance, the shadowing experienced by both UEs is highly correlated. Because $d \approx \lambda$, the small-scale fading should be effectively decorrelated.

In Fig. 3, $v > 20\,\mathrm{km/h}$ most of the time (in the nonstatic intervals), and thus $T_s \gg T_c$ in both frequencies. At pedestrian speeds $v \leq 5\,\mathrm{km/h}$, we obtain $T_s \approx T_c$ or even $T_s < T_c$. In our study, the amount of such slow-speed measurements is so low that they have no impact.[6]

With this constellation, we are able to demonstrate the problem outlined in Section II. Both UEs experience notably different throughputs during static phases. At the same time, both UEs are very close to each other with respect to the shadowing decorrelation distance. Therefore, when the whole setup is moving, the small-scale fading averages out; both UEs measure comparable throughput, which reflects the large-scale fading (network coverage).

IV. Drive Tests: Measurement Results

A. Time-Series and Cumulative Distributions

Fig. 4 visualizes the "ground truth" throughput r_{GT} measured by UE 1 and UE 2 (LTE 800) in drive 2. The traces of both UEs overlap quite nicely when the car is moving. However, we see a consistent offset in the long interval that corresponds to the static measurements. Such an offset that has accumulated over a longer period of time causes a significant difference between the empirical cumulative distribution functions (CDF). The CDFs $F_{\mathrm{UE}\,1}(r_{\mathrm{GT}})$ and $F_{\mathrm{UE}\,2}(r_{\mathrm{GT}})$ (LTE 800, drive 2) are plotted in Fig. 5 a) with 95% lower and upper confidence bound[7] (c.b.). When we compare 90 percentiles of the throughput, we obtain 110 Mbit/s for UE 2 but only 99 Mbit/s for UE 1. By excluding the samples collected at zero speed, we remove this difference—see Fig. 5 b).

[6] In scenarios where slow-speed measurements occupy a larger share of the entire measurement, we should either exclude them or we might want to increase the averaging interval by re-binning the measured data into longer resampling intervals.

[7] The lower and upper confidence bounds are calculated using Greenwoods formula [19].

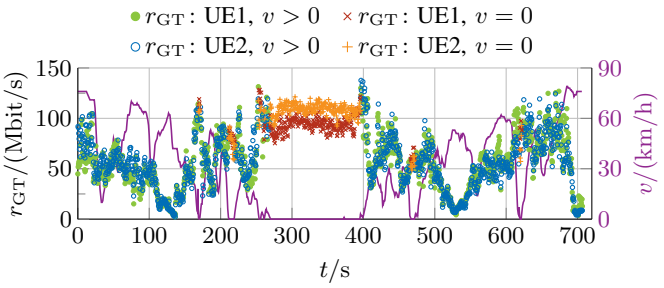

Fig. 4. "Ground truth" throughput r_{GT} measured by UE 1 and UE 2 (LTE 800) during the second drive (left axis) and the vehicle speed profile (right axis). The time on the x-axis is relative to the beginning of the measurement.

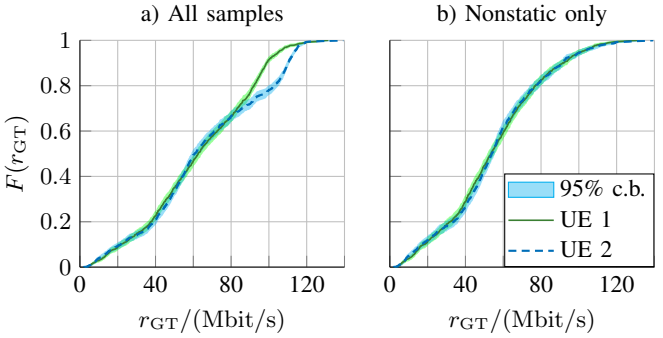

Fig. 5. Empirical CDFs of the r_{GT} samples collected by UE 1 and UE 2 (LTE 800, drive 2). a) Considering all samples, the CDFs differ significantly. b) After excluding the static measurements, the CDFs become nearly identical. By excluding the static samples, the mean throughput has changed from 63.1 Mbit/s to 54.8 Mbit/s (13%) for UE 1 and from 65.6 Mbit/s to 55.0 Mbit/s (16%) for UE 2.

B. Numerical Evaluation

For each drive, we compare $F_{\mathrm{UE}\,1}(r_{\mathrm{GT}})$ with $F_{\mathrm{UE}\,2}(r_{\mathrm{GT}})$ (LTE 800), and $F_{\mathrm{UE}\,3}(r_{\mathrm{GT}})$ with $F_{\mathrm{UE}\,4}(r_{\mathrm{GT}})$ (LTE 2600). First, we compare the CDFs considering all measurement samples, and then we examine the impact of excluding the static samples.

To quantify the similarity between two CDFs, we use the Kolmogorov-Smirnov test [20], which returns the maximum absolute difference d_{KS} between the two CDFs and the corresponding approximated p-value (see [17, Section III-A]).

Furthermore, based on the p-value, we decide between the null hypothesis $\mathcal{H}_0$ and the alternative hypothesis $\mathcal{H}_1$

$\mathcal{H}_0$: the samples come from the same distribution,

$\mathcal{H}_1$: the samples come from two different distributions,

as follows:

$p \geq \alpha$: $\mathcal{H}_0$ is not rejected,

$p < \alpha$: $\mathcal{H}_0$ is rejected in favor of $\mathcal{H}_1$.

We set the significance level to $\alpha = 5\%$.

Table I summarizes the results. In all cases, the throughput distributions would be evaluated as different ($\mathcal{H}_1$) if we consider all samples and as identical ($\mathcal{H}_0$) if we consider the nonstatic samples only.

Interestingly, there is no clear connection between the share of static samples and the distance d_{KS}. In LTE 800, we see similar distances in drives 1 and 3: $d_{\mathrm{KS}} = 0.08$ in drive 1 (45% samples static) and $d_{\mathrm{KS}} = 0.06$ in drive 3 (17% samples static). Drive 2, which has 22% of static samples (i.e., more than what drive 3 has and less than what drive 1 has), results in ca. double distance $d_{\mathrm{KS}} = 0.14$. This is because, as explained, the throughputs measured by UE 1 and 2 in static scenarios are uncorrelated; their difference is merely random. Thus, in some static intervals, the throughputs differ significantly, whereas they match in others. The intervals, in which the UEs experience similar throughputs, do not increase the distance between CDFs.

V. Conclusion

With multiple measurements in live LTE network, we have verified that the throughput in static scenarios is sensitive to the relative position of the measuring device due to small-scale fading. Changing the position by a few centimeters can change the result by 50% (Fig. 1). The small-scale fading patterns remain stable over long periods of time (100 s in Fig. 4) even in an outdoor environment. This consequently leads to the significantly different throughput distributions recorded by individual receivers when conducting repeated measurements.

The results can be summarized as follows:

1) Static measurements at a single point cannot be used to benchmark MNOs. Even in the case of a single receiver with multiple SIM slots, the measured throughputs are not comparable—each MNO has a different small-scale fading pattern that is caused by differently placed sources of radiation.

2) Either a dynamic measurement or a series of static measurements that systematically sample certain route or area are necessary. In the case of drive tests, the static measurements should be excluded (see the impact in Fig. 5).

3) In a dynamic scenario, it is possible to achieve the same throughput distributions even if we measure with non-collocated receivers.[8] During a drive test, we can thus measure the throughput of two MNOs simultaneously using two UEs and still obtain a good benchmark.

Acknowledgment

This work has been funded by the ITC, TU Wien. The research has been cofinanced by the Austrian FFG, Bridge Project No. 871261. We thank A1 Telekom Austria AG for their support and Kei Cuevas for the proofreading.

References

[1] RTR – NetTest. [Online]. Available: https://www.netztest.at/en

[2] OpenSignal Apps – OpenSignal. [Online]. Available: https://opensignal.com/apps

[3] Speedtest by Ookla – The Global Broadband Speed Test. [Online]. Available: http://www.speedtest.net

[4] Ö. Alay *et al.*, "Measuring and assessing mobile broadband networks with MONROE," in *Proc. WoWMoM 2016*, Jun. 2016, pp. 1–3.

[5] T. Berisha, P. Svoboda, S. Ojak, and C. F. Mecklenbrauker, "Cellular network quality improvements for high speed train passengers by on-board amplify-and-forward relays," in *Proc. ISWCS*, Sep. 2016, pp. 325–329.

[6] J. Landre, Z. E. Rawas, and R. Visoz, "LTE performance assessment prediction versus field measurements," in *Proc. IEEE 24th PIMRC*, Sep. 2013, pp. 2866–2870.

[7] J. Beyer *et al.*, "Performance measurement results obtained in a heterogeneous LTE field trial network," in *Proc. 77th VTC Spring*, June 2013, pp. 1–5.

[8] A. S. Khatouni *et al.*, "Speedtest-like measurements in 3G/4G networks: The MONROE experience," in *Proc. 29th ITC*, vol. 1, Sep. 2017, pp. 169–177.

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

[10] V. Raida, P. Svoboda, and M. Rupp, "Repeatability for spatiotemporal throughput measurements in LTE," in *89th VTC Spring*, Apr. 2019.

[11] LTE frequency bands – Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/LTE_frequency_bands

[12] V. Raida, M. Lerch, P. Svoboda, and M. Rupp, "Deriving cell load from RSRQ measurements," in *Proc. 2018 TMA Conf.*, Jun. 2018, pp. 1–6.

[13] T. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2001.

[14] M. Rindler, P. Svoboda, and M. Rupp, "FLARP, fast lightweight available rate probing: Benchmarking mobile broadband networks," in *Proc. 2017 IEEE ICC*, May 2017.

[15] Keysight Technologies. (2019) Nemo handy handheld measurement solution. [Online]. Available: https://www.keysight.com/en/pd-2767485-pn-NTH00000A/nemo-handy

[16] V. Buenestado, J. M. Ruiz-Avils, M. Toril, S. Luna-Ramrez, and A. Mendo, "Analysis of throughput performance statistics for benchmarking LTE networks," *IEEE Commun. Lett.*, vol. 18, no. 9, pp. 1607–1610, Sep. 2014.

[17] V. Raida, P. Svoboda, M. Kruschke, and M. Rupp, "Constant rate ultra short probing (CRUSP): Measurements in live LTE networks," in *2019 IEEE ICC*, May 2019, to be published.

[18] J. Weitzen and T. J. Lowe, "Measurement of angular and distance correlation properties of log-normal shadowing at 1900 MHz and its application to design of PCS systems," *IEEE Trans. Veh. Technol.*, vol. 51, no. 2, pp. 265–273, Mar. 2002.

[19] The MathWorks, Inc. (2019) Empirical cumulative distribution function – MATLAB ecdf. [Online]. Available: https://www.mathworks.com/help/stats/ecdf.html

[20] ——. (2019) Two-sample Kolmogorov-Smirnov test – MATLAB kstest2. [Online]. Available: https://www.mathworks.com/help/stats/kstest2.html

[8] Still, the receivers should have the same hardware and should not be too far apart. The distance between them should be much smaller than the shadowing decorrelation distance.