2015 4th IEEE International Conference on Advanced Logistics and Transport (ICALT)

Characterization of mobile phone localization errors with OpenCellID data

Michael Ulm, Peter Widhalm, Norbert Brändle
Austrian Institute of Technology
A-1210 Vienna, Austria
michael.ulm@ait.ac.at, peter.widhalm@ait.ac.at, norbert.braendle@ait.ac.at

***Abstract*—We examine the publicly available OpenCellID database for the purpose of estimating the position of mobile devices along with the positioning accuracy in any given geographic region. We present a method coping with typical distortions and artifacts in the data to compute robust estimates of cell positions and localization error distributions. Applying the method to analyze different regions and mobile networks reveals that the localization errors follow a fat-tailed distribution with similar shapes and properties. We experimentally validate the approach using an independent set of GPS and cell ID data collected by 250 individuals over a period of several weeks. The localization error estimated from OpenCellID data agrees well with the experimental error distribution. While there are significant differences between urban and rural areas, we show that there is only a weak correlation between antenna density and empirical localization errors.**

I. Introduction

The communication infrastructure of a cell phone network provides extensive information about the travel movements of a large proportion of a population. Every mobile device connecting to the cellular network (GSM, GPRS, UMTS or LTE) generates digital traces which may serve as input for analyzing traffic indicators and mobility behavior. Analyzing cell phone data for extracting spatial-temporal information and patterns of traffic and mobility behavior without additional sensor infrastructure is therefore an active research field: Cell phone data were already used to compute origin-destination matrices (e.g. [1], [2], [3]), traffic flow and volumes [4], trip modeling [5], land use [6], and disaster relief [7]. A survey can also be found in [8].

Cell phone data can provide device locations only with low spatial resolution, with typically sparse and irregular temporal updates. Existing work usually represents spatial resolution of cellular localization as a Voronoi tesselation of cell centroids, i.e. the area is partitioned into a set of disjunct convex polygons. Such a Voronoi tesselation assumes that the density of cell towers is the only factor determining spatial accuracy in an area. Spatial errors, however, depend on many more factors than cell tower density, e.g. type and make of the antenna, signal strength, system load, reflections, make and model of the mobile device – all of which influence the connection between the device and antenna. As a consequence, a cell phone device not necessarily connects to the nearest antenna. Fig. 1 depicts a set of locations measured during an experiment. They are all connected to one particular antenna, which illustrating the invalidity of the assumption of a Voronoi tesselation.

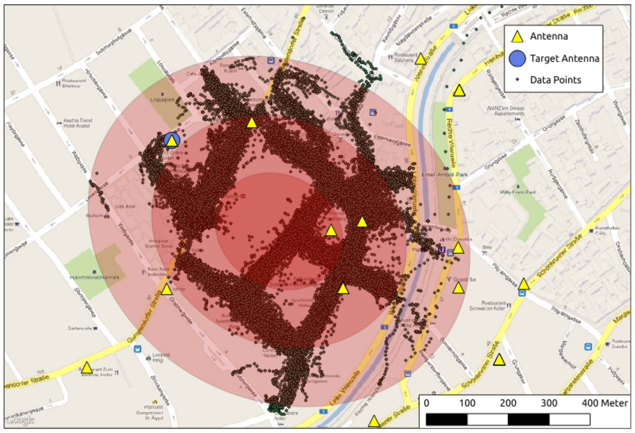


Fig. 1. Measured GPS points of devices connected to a given target antenna (large blue dot) and the σ, 2σ, and 3σ ellipses of this point cloud.

Theoretical studies provide some insight into the reach of an antenna with featured characteristics, e.g. [9] or [10]. Given that the reach of an antenna largely depends on the type – with low-powered pico and micro cells in urban areas having a small reach (see Fig. 2a) and macro cells having higher power output and significantly larger reach (see Fig. 2b), knowledge about the antenna types is a valuable cue for estimating spatial resolution of cellular localizations in a particular area. Telecommunications providers, however, usually do not release information about antenna characteristics, not even information about antenna location. Consequently, theoretical studies based on antenna characteristics are of limited practical value in this context. Some empirical aspects of the spatial distribution of the localization error were examined in [11]. Due to limited sample sizes and restricted contexts of this study, a more thorough analysis is still necessary. The work in [12] characterizes location accuracy of cell tower localization by analyzing the signal strength captured during wardriving in the greater Los Angeles area. However, for many applications, signal strength is not available as a data source.

OpenCellID is a large collaborative project collecting GPS location data for cell identifiers (Cell-ID), with the main application of providing power-efficient and fast location information to mobile devices. As of August 2014, over a billion measurements [13] were collected, which are publicly available under a free Creative Commons license. The data are collected fully automatically by registered users via various

This work was partly funded by the Austrian Federal Ministry for Transport, Innovation and Technology (BMVIT) within the strategic program FIT-IT under grant 835946 (SEMAPHORE).

978-1-4799-8400-8/15/$31.00 ©2015 IEEE

smart phone apps. To the best of our knowledge, there currently exists no work using OpenCellID data for modeling the spatial resolution for a given region. In an effort to close this gap, this paper proposes robust location error models estimated from the OpenCellID data. Since the sample of OpenCellID volunteers might be biased towards the behavior of heavy contributors and the location measurements might be distorted by peculiarities in the sampling process and measurement technique, we validate the model with an independent experimental dataset of GPS positions and cell IDs collected from 250 individuals over a period of several weeks.

This paper is organized as follows: Section II examines the OpenCellID measurements for each cell and infers information on the reaches of antennas in order to model positioning accuracy. In Section III we validate the positioning accuracy model using an independent data set from an experiment where cell phone ids and GPS positions were recorded simultaneously and investigate the correlation of location error with antenna density and Voronoi cell size.

II. Localization Error Modeling

Estimating robust localization error models from the OpenCellID data involves an initial preprocessing step for data cleansing and a modeling step identifying location and size of point clouds assigned to Cell-IDs.

A. Data Preprocessing

The automatically collected measurement data typically include the following phenomena and artifacts which potentially affect cell position estimates and the localization error model:

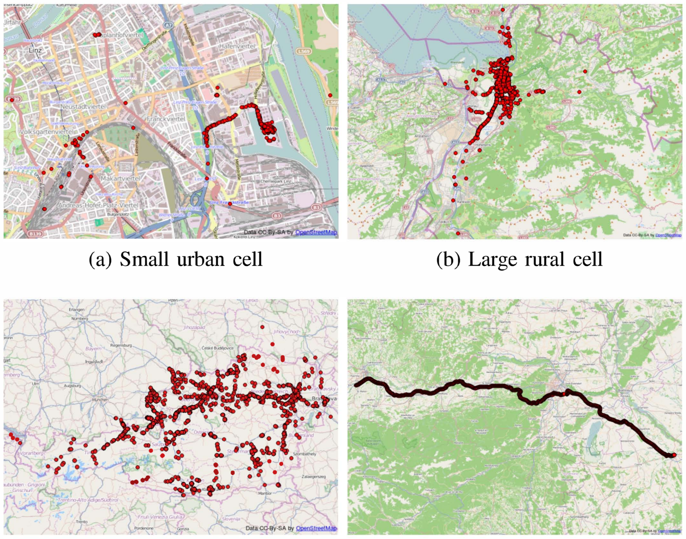
- *Erroneous Cell IDs* : The density of cell towers is unrealistically high in some regions. Examining such areas and the corresponding antennas reveals that this is typically due to erroneous cell IDs, caused by a mix up of cell ID, local area code (LAC) and mobile network code (MNC). Typically, only few measurements are attached to such erroneous Cell IDs.

- *Antenna dragging* is caused whenever during a trip a device does not update the cell ID, but reports the original cell ID, illustrated in Fig. 2d.

- *Outliers:* Some cells with plausible measurement points have additional measurements attached which are clearly wrong – far away from the correct position, often not even in the same country.

- *Unrealistic cell sizes:* The GPS measurements of some cells are distributed across a whole country which is clearly unrealistic (see Fig. 2c). The reason for this type of errors is not known to the authors. One possible explanation for such phenomena are mobile antennas used by mobile phone providers to cover temporarily increased demand during specific events. Such mobile antennas provide identical cell IDs for their different positions.

In order to remove erroneous cell IDs and reduce cell dragging to just a few additional points, we employ the following pre-processing scheme on the raw data:



(a) Small urban cell

(b) Large rural cell

(c) Unrealistic cell size

(d) Antenna dragging

Fig. 2. Typical point clouds

1) Allow only one measurement per hour for each antenna. If several measurements are available within an hour, aggregate them to one position using the center of all those positions.
2) If the resulting point cloud of measurements has less than a given threshold of t entries, discard the antenna.

Determining the threshold t for the number of measurements for an admissible antenna involves the usual information retrieval trade off between precision and recall: If the threshold t is low, many spurious antennas will remain in the data. If t is high, valid antennas will be lost. This work aims at restricting the number of erroneous entries and sets a relatively high threshold of $t = 100$ measurements.

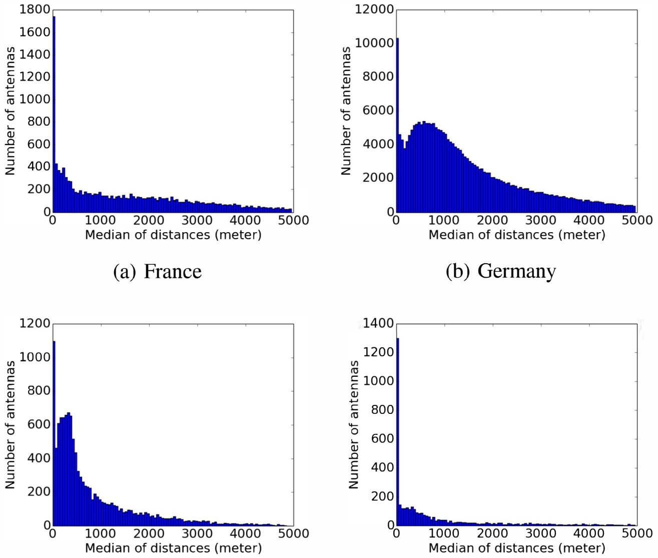
The resulting data are, however, still rather noisy. Consequently, robust statistics for estimating cell position and size are required to cope with remaining outliers and noise.

B. Robust estimation of cell position and size

Each antenna serves a particular area in the landscape. The location and size (scale) of this area varies by antenna strength, type, and external influences such as obstruction and reflection zones.

Estimating the cell position requires to define a robust centroid for each point cloud of GPS measurements collected with OpenCellID. While the median is a common robust measure for one-dimensional location, it does not generalize easily to higher dimensions. Several such generalizations are known [14], and we use the *centerpoint*: it is defined as a point for which each hyperplane through the centerpoint divides the point cloud into two subsets such that the smaller of these subsets has at least a $\frac{1}{d+1}$ fraction of the points. The algorithm provided in [15] provides a fast probabilistic approach for computing centerpoints.

We characterize the size of a cell by the distances of each point in the point cloud to the centerpoint, and use the median of these distances as robust statistic. Computing the median distances for each antenna in a region (e.g. a country) one obtains a distribution over cell sizes. Examples for such distributions are given in Fig. 3.



(a) France

(b) Germany

(c) Austria

(d) Russian Federation

Fig. 3. Examples of distributions of median distances

The distance distributions for various countries show rather similar characteristics i.e. a large share of antennas with point cloud radius less than 50 meters mixed with a fat tailed unimodal or nearly unimodal distribution with a mode a approximately 300 meters. We characterize these distributions by three numbers: the share of small antennas, the median of the distribution and the scaling parameter (or exponent) of the fat tail. The fat tailed distribution p is given by the relation

$$p(x) \propto L(x)x^{-\alpha}, \tag{1}$$

where L is a slowly varying function, and the scaling parameter α is computed with a modified Hill estimator [16] as

$$\alpha = 1 + \frac{n}{\sum_{i=1}^{n} \log \frac{x_i}{x_{\min}}}. \tag{2}$$

Here, x_i are the observed distances larger than $x_{\min}$, and $x_{\min}$ is chosen to minimize the distance between the observed and the modeled distribution for $x > x_{\min}$, measured by the Kolmogorov-Smirnov statistic.

The results of the distribution parameter computations for a selection of countries is given in Table I. While the median and the share can vary widely between countries, the scaling parameter is stable at a value around 2.0 for all countries.

III. Location Accuracy Results

We validated the positioning accuracy model with an independent experimental data set of GPS positions and cell IDs collected from 250 individuals who volunteered to log their cell positions and GPS points. This experiment took place between May and August 2014 in Austria and provided over 7 million data points.

TABLE I. Distribution characteristics by country

Country	Nr. Antennas	Share larger 50m	Median	Scaling parameter
Germany	239496	0.043	1189.285	1.831
Netherlands	16076	0.073	419.899	2.126
France	14337	0.121	1358.620	1.685
Austria	12208	0.090	466.647	2.019
Denmark	11593	0.035	1330.300	1.763
Belgium	9796	0.110	449.224	2.091
Russian Federation	8412	0.492	53.533	1.836
Switzerland	7628	0.081	587.327	1.937
United Kingdom	7435	0.127	386.601	2.184
Czech Rep.	6921	0.094	413.253	2.102
Poland	6356	0.107	663.926	1.797
Brazil	5873	0.411	134.301	2.233
South Africa	5687	0.071	950.271	1.897
Norway	5050	0.068	829.736	1.816
United States	4398	0.299	400.448	1.672
Sweden	3923	0.111	905.208	1.687
Italy	3141	0.236	485.632	1.780
Spain	3007	0.228	302.662	1.753
Slovakia	2725	0.101	383.568	2.096
New Zealand	2468	0.091	390.453	1.819
Mexico	2407	0.451	73.255	1.902
Zambia	1663	0.239	525.866	1.585
Luxembourg	1611	0.094	439.496	1.934
Turkey	1448	0.442	81.644	1.693
Ukraine	1335	0.519	42.242	1.878
Australia	1178	0.199	391.454	1.795
Canada	1053	0.318	218.082	1.659

Of the 21075 distinct cell IDs recorded in the experiment, 8633, or 40.96 percent, were found in the OpenCellID database. Consequently, of the 7065824 data points recorded in the experiment, 3119391, or 44.15 percent could be matched with a cell position estimate based on OpenCellID data.

For the points that could be matched and located with OpenCellID data, we determined their experimental position accuracy by comparing the recorded GPS position with the cell position estimate given in the OpenCellID database. The histogram of the resulting error sizes is shown in Fig. 4. There, we also distinguished between antennas in the urban region of Vienna, where antenna density is high, and those outside of Vienna, where average antenna density is low. This distinction already hints at a correlation between antenna density and localization error.

The resulting error distribution is a fat tailed distribution with a mode of about 200 meters, median at 507.6 meters, 75th percentile at 1574.6 meters and 90th percentile at 5685.6 meters. This is comparable to the positioning accuracy obtained when using cell position estimates provided by mobile operators [11].

We next address the question if Voronoi cells can be used to characterize positioning accuracy and to what extent higher antenna density translates to lower localization error. To this end, we computed two proxies for antenna density. Firstly, we counted for each antenna the number of neighbouring antennas in a 500 meter radius. This was then translated into an antenna density measure estimating the average area (in square meters) for each antenna in this radius. Secondly, we computed the size of the Voronoi cell for each antenna.

To examine the influence of antenna density on location error, we computed the Spearman rank coefficient between the antenna density measures and the median localization error of the antenna. The scatter plot in Fig. 5 shows a weak correlation between antenna density and localization error. This is also

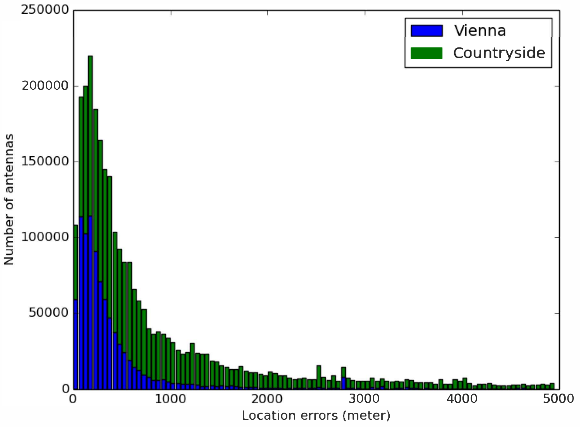
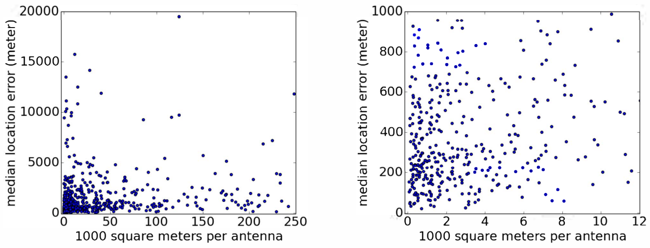


Fig. 4. Distribution of location errors

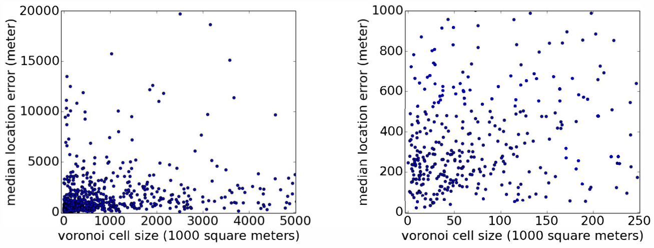
supported by Spearman's rank correlation coefficient of 0.458. Similarly, for Voronoi cell sizes the scatter plot is given in Fig. 6, and the Spearman correlation coefficient computes as 0.552. Higher antenna density will therefore in general have a positive influence on the distribution of the location error, yet large errors remain common even for very high antenna densities.



(a) All datapoints

(b) Detail

Fig. 5. Antenna density vs. median location error

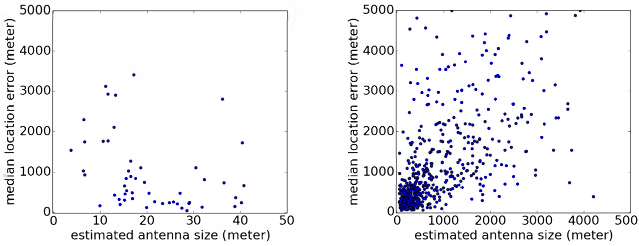


(a) All datapoints

(b) Detail

Fig. 6. Voronoi cell size vs. median location error

For comparison, we examine the correlation between the estimated cell size using the approach described in Section II and experimental localization errors. Examining the data, it turns out that cells with estimated sizes of less than 50 meters (the spikes in the distributions of Fig. 3) do show a different behaviour than those with larger estimated size: Fig. 7a shows that in the first group the experimental localization error varies greatly and shows no correlation with the estimated cell sizes. On the other hand, the cells with larger estimated sizes correlate quite well with the experimental location error (Fig. 7b), with a Spearman correlation coefficient of 0.670. These results suggest that for cells with estimated size less than 50 meters the information contained in the OpenCellID database is either insufficient or too biased to allow for accurately estimating cell size and positioning accuracy. As a consequence, these cells have to be treated as outliers and discarded. For the remaining cells, however, the estimated cell size estimates provide a significantly more accurate characterization of the localization error than Voronoi cell sizes.



(a) estimated size $<$ 50 meters

(b) estimated size $>$ 50 meters

Fig. 7. Estimated cell size vs. experimental localization error

IV. Conclusion

OpenCellID data provide a useful tool for localizing when only cell information is given. The localization error can still be quite big, which, however, does to a large extent not stem from a problem in the OpenCellID data, but from the general problem of positioning using cell data. One drawback of the OpenCellID data set is that not all cells can be found in the data. The number of measurements and coverage of cells varies greatly by country.

We showed that the localization errors follow fat tailed distributions with a scaling parameter of about 2.0. The distribution of the estimated cell sizes agrees well with the experimental error distribution, suggesting that the source of the error does not stem from bad cell localization in the OpenCellID database but from the way the network is organized. The estimated cell sizes showed a significantly higher correlation with the empirical localization error than antenna density and therefore provide a more accurate characterization of positioning accuracy than Voronoi tesselation of cell centroids. However, analysis of experimental results also suggests that very small estimated cell sizes (less than 50 meters) do not actually correspond to low localization errors but are instead probably due to a bias in the OpenCellID data.

In this study we were not able to easily distinguish between antenna types using the point clouds. Further investigation is necessary to find the cause and maybe a remedy for this problem, which is left for future work.

References

[1] M.-H. Wang, S. D. Schrock, N. Vander Broek, and T. Mulinazzi, "Estimating dynamic origin-destination data and travel demand using cell phone network data," *International Journal of Intelligent Transportation Systems Research*, vol. 11, no. 2, pp. 76–86, 2013.

[2] N. Caceres, L. M. Romero, F. G. Benitez, and J. M. Del Castillo, “Traffic flow estimation models using cellular phone data,” *Intelligent Transportation Systems, IEEE Transactions on*, vol. 13, no. 3, pp. 1430–1441, 2012.

[3] F. Calabrese, G. Di Lorenzo, L. Liu, and C. Ratti, “Estimating origin-destination flows using mobile phone location data,” *Pervasive Computing, IEEE*, 2011.

[4] P. Wang, T. Hunter, A. M. Bayen, K. Schechtner, and M. C. Gonzalez, “Understanding road usage patterns in urban areas,” *Sci. Rep.*, vol. 2, Dec 2012. [Online]. Available: http://dx.doi.org/10.1038/srep01001

[5] L. F. Huntsinger and K. Ward, “Using mobile phone location data to develop external trip models,” in *Proceedings of the 94th Transportation Research Board Annual Meeting*, 2015.

[6] J. L. Toole, M. Ulm, M. C. Gonzalez, and D. Bauer, “Inferring land use from mobile phone activity,” in *UrbComp ’12*, 2012, pp. 1–8.

[7] J. P. Bagrow, D. Wang, and A.-L. Barabsi, “Collective response of human populations to large-scale emergencies,” *PLoS ONE*, no. 6(3), 2011.

[8] F. Calabrese, L. Ferrari, and V. D. Blondel, “Urban sensing using mobile phone network data: a survey of research,” *ACM Computing Surveys (CSUR)*, vol. 47, no. 2, p. 25, 2014.

[9] T. Z. Qiu, P. Cheng, J. Jin, and B. Ran, “State of the art and practice: cellular probe technology applied in advanced traveler information system,” in *Proceedings of the 86th Transportation Research Board Annual Meeting*, 2007.

[10] A. LaMarca, Y. Chawathe, S. Consolvo, J. Hightower, I. Smith, J. Scott, T. Sohn, J. Howard, J. Hughes, F. Potter, J. Tabert, P. Powledge, G. Borriello, and B. Schilit, “Place lab: device positioning using radio beacons in the wild,” in *Pervasive Computing*, ser. Lecture Notes in Computer Science, H.-W. Gellersen, R. Want, and A. Schmidt, Eds. Springer Berlin Heidelberg, 2005, vol. 3468, pp. 116–133. [Online]. Available: http://dx.doi.org/10.1007/11428572_8

[11] M. Ulm and P. Widhalm, “Properties of the positioning error of cell phone trajectories,” in *NetMob 2013*, 2013.

[12] J. Yang, A. Varshavsky, H. Liu, Y. Chen, and M. Gruteser, “Accuracy characterization of cell tower localization,” in *Proceedings of the 12th ACM international conference on Ubiquitous computing*. ACM, 2010, pp. 223–226.

[13] “OpenCellID,” http://opencellid.org, accessed: 2015-01-10.

[14] H. Edelsbrunner, *Algorithms in combinatorial geometry*. Springer Verlag, 1987.

[15] K. L. Clarkson, D. Eppstein, G. L. Miller, C. Sturtivant, and S.-H. Teng, “Approximating center points with iterative radon points,” *Int. J. Comput. Geom. Appl.*, vol. 357, no. 06, 1996.

[16] A. Clauset, C. R. Shalizi, and M. Newman, “Power-law distributions in empirical data,” *SIAM Review*, no. 51, pp. 661–703, 2009.