On optimal spatial probability density estimation of passive mobile positioning events

Toivo Vajakas
Reach-U Ltd;
Institute of Computer Science, University of Tartu
Tartu, Estonia
e-mail: tvajakas@gmail.com

Joosep Rõõmusaare
Reach-U Ltd;
Institute of Computer Science, University of Tartu
Tartu, Estonia

***Abstract* — In passive mobile positioning the cell-level measurements must be translated into geographical location, which can be expressed as spatial probability density function (SPDF).**

In this paper we present the results of a study where we compared different methods to estimate SPDF of passive mobile positioning. The mobile operators provide spatial data about network cells. It is called cellplan. Cellplan can be provided in various formats: location of cell towers and azimuth of antennas, or radio signal strength levels from radio propagation models, etc. Each such source requires specific processing to infer SPDF for passive mobile positioning. We investigated the probabilistic properties of different processing algorithms on different variants of input data. Some investigated methods apply Bayes rules to take into account the effects from overlapping neighbor cells. The results indicate that even on relatively small dataset one can clearly see different accuracy from different processing parameters. The accuracy of SPDF varies significantly depending on processing parameters. and sophisticated radio propagation models did not have significant advantages over simple procedures using Voronoi diagram. (*Abstract*)

Keywords — passive mobile positioning; Bayesian location estimate; spatial probability density function; PDF, heat map, location estimation accuracy

I. Introduction

A. Motivation

Passive mobile positioning data, gathered as a by-product by mobile operators, has gained much popularity in human geography studies due to availability of large samples [1]. Mobile positioning data describes the location of a mobile station (MS). A MS can be a phone or a modem for a device such as a security or environmental sensor. Mobile positioning determines the position of a MS with significant spatial uncertainty [2]. Spatial uncertainty can be generally described as spatial probability density function (SPDF) of location

The mobile operators provide spatial data about network cells. It is called cellplan. Cellplan can be provided in various formats: location of cell towers and azimuth of antennas, radio signal strength levels from radio propagation models, etc. Each such source requires specific processing to infer SPDF for passive mobile positioning. In majority of mobile positioning papers the effects of cell area overlap are ignored. We considered applying Bayesian rule to take into account, importance of this effect needed verification.

The aim of this paper is to describe a methodology for such investigation and compare various cellplan SPDF preparation methods.

B. General characteristics of mobile positioning data

Each passive mobile positioning data record has an attribute identifying the mobile network cell the phone was connected to, known as the Cell Global Identity (CGI). A cell is the geographical area where it is possible to connect to one transceiver of a base station. Each cell has limited capacity and therefore operators design smaller cells in regions of high population density. Neighboring cells have considerable overlap. When a mobile phone disconnects from one cell and connects to another the event is called a ‘handover’. For mobile positioning it is important to know the geographical shape of each cell. The actual cell shape depends on many factors, such as antenna radiation pattern and height, network load, signal attenuation on landscape and indoors, signal reflections, radio interference and noise, network configuration parameters such as handover threshold and neighbor cell lists [3]. Fig 1 illustrates the uncertainty present when determining location from the fact that phone is connected to particular cell.

Other location-related attributes in addition to CGI can be collected for improved location accuracy, such as distance to the antenna or signal strength from neighboring cells. CGI data is however still the most scalable passive positioning approach and puts the least load on a network, including for the recently introduced LTE (Long-Term Evolution) networks [4]. In this paper we consider positioning using only CGI data.

Mobile positioning data can be exported from different nodes in the mobile operator’s network, resulting in different levels of detail of the data. The most notable options of passive positioning data are call detail records (CDR) and network subsystem (NSS) event stream [2]. CDR data has been the most widely used option in mobile positioning research. Each

CDR describes a billing-related subscriber activity like starting a call or sending a text message. In some configurations mobile operator provides also periodic update events that report (typically every hour or two) the current location of a mobile. The network also generates location update events when a phone moves from one location area (a group of closely situated base stations) to another [2].

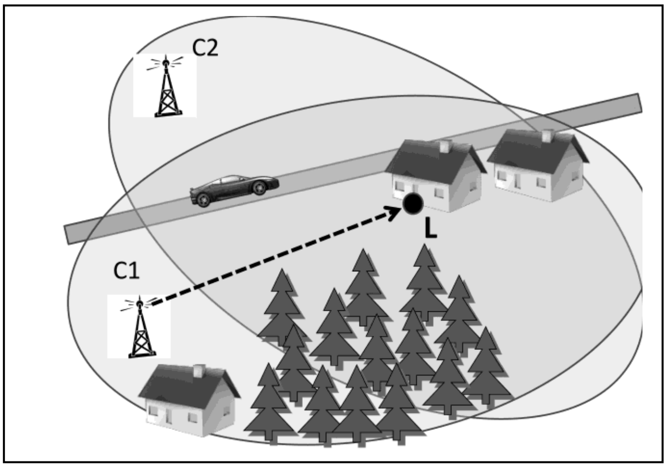

Figure 1. Mobile station location uncertainty problem illustration. Suppose we know that a MS was connected to cell C1 and want to estimate the probability that MS was in location L. The spatial probability distribution of each cell (as defined by cellplan) is shown as filled oval areas. The probability estimate is affected by cell C2 that also covers location L and by prior knowledge that people stay mostly in houses and move on roads and nobody lives in forest.

C. Related work

Based on available information the shape of each cell has to be defined to give location estimates for mobile positioning. Cell data provided by mobile operators can be translated to cell shapes as Voronoi polygons by using the assumption that a phone connects to the nearest tower [5]; as best server data polygons by using the assumption that a mobile phone connects to the cell with the strongest signal [6]; or as a raster model based on the assumption that the probability to connect to a cell is a function of distance from the antenna tower [7].

A methodology applying Bayesian methods for defining the SPDF for mobile positioning was given by Zang et al. (2010) [8]. That paper provided a solution for the situation where neighbor antennas are present. The estimate was based on signal-to-interference-and-noise ratio (SINR) calculations. The paper provided a Bayesian probability estimate for situations where CGI is the only information known and additionally it is known that no other cells had good enough SINR. The method was tested against the subset of emergency call data limited to situations where only one cell was within the reach of the MS. They found that the difference between the location measured by GPS and the MLE provided by that method was improved by 20%, compared to baseline method. No direct PDF tests were performed by Zang et al. (2010) [8]. The Bayesian probability formulas in the paper by Zang et al. do not consider the effect that the probability to connect to a concrete cell will be reduced in locations where other cells are also reachable.

D. Research problem

We are investigating how the SPDF estimation quality is affected by following factors

- Different input data (tower coordinates vs radio propagation data)
- Post-processing of data (e.g. enlarging and blurring cell boundaries by given factor)
- How much is result affected by cell overlap effects
- How much the result depends on MS and location

II. Methods

A. Mathematical model of SPDF

Given a mobile operator's event logs for some time period, we want to map each event probabilistically to geographical space where the mobile event occurred. We assign spatial PDF to each cell such that PDF defines the probability that the event occurred in any given location. We consider here only discrete probability densities obtained by dividing the area of interest into pixels of appropriate size.

The radio area network (RAN) consists of a finite set of cells, $\mathcal{C}$. For each cell C in the cellplan $\mathcal{C}$ and each pixel x, we want to compute $\mathrm{P}(x|C) = \mathrm{P}\left(\underline{x} = x \mid \underline{C} = C\right)$, the probability that the MS is at location x if it generates an event in cell C. In the following we will describe a method how to do it. (The raster $\mathrm{P}\left(\underline{x} = \cdot \mid \underline{C} = C\right)$ is then what we call the spatial PDF of the cell C.)

The probability $\mathrm{P}(x|C)$ is determined by the Bayes' formula:

$$\mathrm{P}(x|C) = \frac{\mathrm{P}(C|x)}{\mathrm{P}(x)} \qquad (1)$$

where

- ○ $\mathrm{P}(C|x) = \mathrm{P}(\underline{C} = C \mid \underline{x} = x)$ is the probability probability that if a MS at location x then it is connected to cell C (rather than any other cell or no cell);
- ○ $\mathrm{P}(x) = \mathrm{P}(\underline{x} = x)$ is the Bayesian prior density, i. e. the probability for a person (or more precisely, a mobile station) to be in pixel x;
- ○ $\mathrm{P}(C) = \mathrm{P}(\underline{C} = C)$ is the probability for mobile station (at a random location) to be connected to cell C, i. e. $\mathrm{P}(C) = \sum_x \mathrm{P}(C|x)\mathrm{P}(x)$ where x ranges theoretically over the world (in practice, over an area of interest, e. g. one country).

The Bayesian prior $P(x)$, representing our prior belief, can BE constructed e. g. from population density data, road and building layers (people are more likely to be on road or in a building).

The probability $P(C|x)$ is computed as follows:

$$\mathrm{P}(C|x) = \mathrm{P}(\text{connected} \mid x) \cdot \mathrm{P}(C \mid x, \text{connected}) \qquad (2)$$

where

- $\mathrm{P}(\text{connected} \mid x)$ is the probability that the MS is connected to the RAN at all (i.e., is able to generate an event) if it is at location x,
- $\mathrm{P}(C \mid x, \text{connected})$ is the probability that if a MS is at location x and is connected then it is connected to cell C.

Let $\underline{S}$ denote the active set of the MS, i. e. the set of cells actually detectable by the MS at a given time moment. A rough estimate of the conditional probability $\mathrm{P}\big(C \in \underline{S} \,\big|\, x\big)$ (that a certain cell C is detectable by the MS if the MS is at point x) is provided by the mobile operator in the form of cell polygon: $P\big(C \in \underline{S} \,\big|x\big)$ is 1 if location x lies inside the cell's polygon and 0 if outside, so the active set depends deterministically on x. In a more refined model, $P\big(C \in \underline{S} \,\big|x\big)$ could have non-zero values to reflect our ignorance of the precise coverage area and the stochastic nature of cell coverage caused by effects like Rayleigh fading and weather changes.

We can only receive CDR events when the mobile phone is connected to network. Also, Estonia is very well covered and connection probability is almost 100%. We don't know actual connection rate. Therefore, we simplify and calculate for each raster pixel $\mathrm{P}'(C \,\big|\, x) \approx \mathrm{P}(C, \mathit{connected} \,\big|\, x)$.

For the probability $\mathrm{P}(C \mid x, \text{connected})$ we propose an ad-hoc formula (in contrast to the other formulas in this paper, which have some theoretical justification): namely, we take the probability to be proportional to the square of the probability $\mathrm{P}\big(C \in \underline{S} \,\big|x\big)$, i. e.

$$\mathrm{P}(C \mid x, \text{connected}) = \frac{\mathrm{P}\big(C \in \underline{S} \,|x\big)^2}{\prod_{C \in \mathcal{C}} \mathrm{P}\big(C \in \underline{S} \,|x\big)^2} \qquad (3)$$

B. *Description of test data*

Data consists of GPS measurements of individual persons who have installed GPS track recording software into their mobile phones, and CDR data for same persons from mobile operator.

For each CDR record we found from GPS track the location of person for given time moment. The records not covered by GPS track were ignored. We used 4 personal tracks from same 8 month time period. Example of data is given on Fig 2.

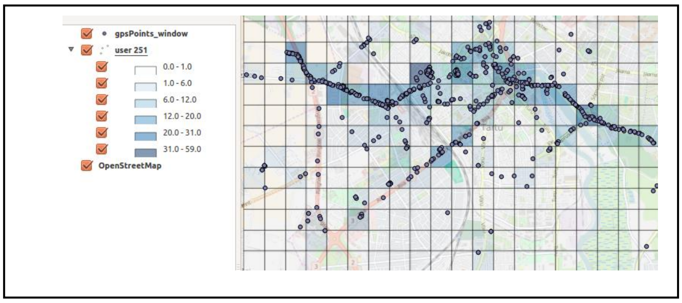

Figure 2. Measured data used for model accuracy estimation. On the left is colorscale for count of measurements in grid cell. Each grid cell is 630m square

C. *SPDF model quality assessment criteria*

We can only receive CDR events when the mobile phone is connected to network. Also, Estonia is very well covered and connection probability is almost 100%. We don't know actual connection rate. Therefore, we simplify and calculate for each raster pixel $\mathrm{P}'(C \,\big|\, x) \approx \mathrm{P}(C, \mathit{connected} \,\big|\, x)$.

The quality of SPDF estimate is evaluated with logarithm of likelihood

$$\textstyle\sum_i \log \mathrm{P}'(C_i \mid x_i) \qquad (4)$$

where *i* is index of measurement (one measurement is one CDR with established GPS location).

III. Results

A. *SPDF variants from cellplans*

We divided test area into quadratic pixels of size 630 meters. For all cellplan variants the values of raster $\mathrm{P}'(C \,\big|\, x)$ were calculated. As illustration on Fig 3 is cross-section of area, showing relative probability of each cell in given point.

B. *Model likelihood calculations*

Using the calculated SPDF rasters we calculated with formula (4) the likelihood of data given each SPDF variant tested. There were two cellplan datasets for same network, one based on RSSI (Received Signal Strength Indication) data and another on tower+azimuth (used to construct Voronoi polygons). Derived variations include:

- rssi_default –RSSI data, unchanged
- rssi_A1AP2 –twice stretched beam width
- rssi_A2AP1 –twice stretched beam along azimuth
- rssi_A3AP1 –three times stretched along azimuth
- rssi_convexhull –convex hull over original geometry
- rssi_ellipse –original approximated with ellipse
- voronoi_default – tower coordinate data, Voronoi constructed
- voronoi_A1AP2 – stretched twice width of beam
- voronoi_A2AP1-- stretched twice along the beam
- voronoi_A3AP1—stretched three times along the beam
- voronoi_convexhull -- original
- voronoi_ellipse

The results are shown on Fig 4.

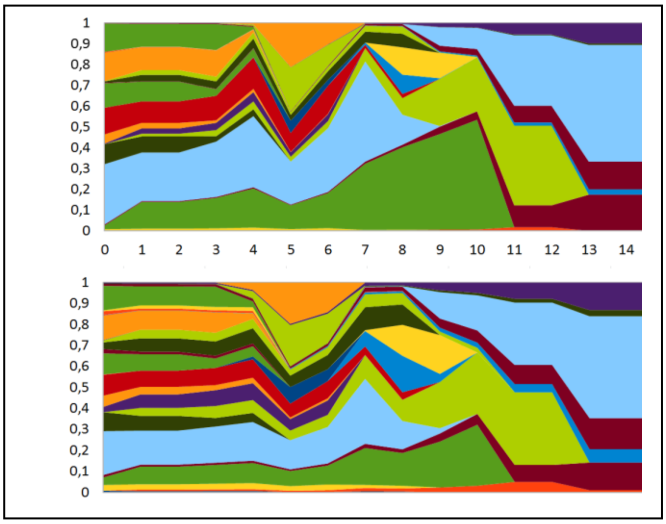

Figure 3. Illustration of SPDF of all cells along cross-section of test area along a line, showing relative probability of each cell in given location. Horizontal axis is pixel number (each pixel is 630m) and vertical axis is stacked probabilities of cells. Upper chart is generated with applying Bayesian overlapping cell model, lower drawing without considering overlapping effects.

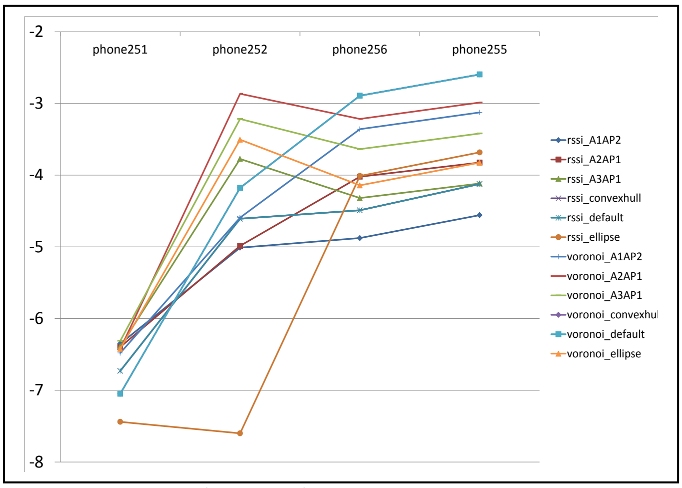

Figure 4. Relative performance of various cellplan variants. Horizontal axis – different test phone tracks (subsets of positioning data, with different spatial distribution). Vertical axis – average log P(C|x) for CDR records of given track.

IV. Discussion

The main findings are

- Some processing variants performed significantly better than the others for certain situations, but there was no single best method for all situations. When backed with large test dataset one could develop a mixed processing procedure using different method for different areas but it is limited by overfitting concerns.
- SPDF calculated from RSSI data was not superior to simplistic Voronoi-based SPDF. We had expected that RSSI input enables much better estimation than tower location and azimuth data.
- The modest dataset consisting of four personal tracks characterizes some location sufficiently but is not sufficient to give overall picture.
- Accounting for cell overlap effect with Bayes rule had in majority of cases positive effect. The likelihood value depended much more on location than Bayes, but this need not mean that applying overlap correction with Bayesian rule is insignificant. Due to the methodology of comparison likelihood is related to probability density, and in areas with larger cells the SPDF values are expected to be significantly lower, thus it need not be the flaw of SPDF estimation.

In future work we plan to analyze specific situations where performance of one or other SPDF estimation variant degrades and optimize the methods accordingly. Also we plan to investigate effects of previous state, e.g. MS approaching cell, or stationary in neighborhood of cell.

Acknowledgment

This research was partially supported by traffic applications research project in Reach-U Ltd and ELIKO and by Location-Based Big Data Algorithms project in STACC. This research has been supported by the European Union through the European Regional Development Fund.

References

[1] J. Steenbruggen, E. Tranos and P. Nijkamp, "Data from mobile phone operators: a tool for smarter cities?," Telecommunications Policy, 2014.

[2] E. Saluveer and R. Ahas, "Using Call Detail Records of Mobile Network Operators for Transportation Studies," in Mobile Technologies for Activity-Travel Data Collection and Analysis, Hershey, PA, IGI Global, 2014, p. 224.

[3] M. Amirijoo, "Neighbor Cell Relation List and Physical Cell Identity Self-Organization in LTE," in 2008 IEEE International Conference on Communications Workshop, 2008.

[4] S. Cherian and A. Rudrapatna, "LTE Location Technologies and Delivery Solutions," Bell Labs Technical Journal, vol. 18, no. 2, p. 175–194., 2013.

[5] R. Ahas, A. Aasa, A. Roose, S. Silm and Ü. Mark, "Evaluating passive mobile positioning data for tourism surveys: an Estonian case study," Tourism Management, vol. 29, no. 3, pp. 469-486, 2008.

[6] F. Calabrese, "Urban sensing using mobile phone network data.," in Ubicomp 2011 Tutorial, 2011.

[7] F. Calabrese, Using cell-phone data to understand urban dynamics in the city of Amsterdam, MIT SENSEable City Laboratory., 2008.

[8] H. Zang, F. Baccelli and J. Bolot, "Bayesian inference for localization in cellular networks," in 2010 Proceedings IEEE INFOCOM, 15–19 March 2010, San Diego, CA, USA, 2010.