Improving the map matching on passive mobile positioning data by minimizing the effects of cell geometry data

Toivo Vajakas
Institute of Computer Science
University of Tartu
Tartu, Estonia
tvajakas@gmail.com

***Abstract*—This paper addresses cell data quality problems encountered in map matching based on passive mobile positioning data. The passive positioning data is typically on cell global identifier level, i.e. data record tells that the mobile station was connected to given cell at given time. Each cell has attributes that can be used to estimate the effective area of the cell. The assumed cell area can very significantly differ from the reality which would result in very unrealistic map matching results. The paper describes a technique to find conflicting cell data and proposes an algorithm for utilizing this knowledge so that conflicting cell data is suppressed in map matching process. The technique significantly improves the map matching results.**

Keywords—mobile positioning; cellplan; data quality; map matching

I. INTRODUCTION AND THEORY

A. *Event data*

Large datasets of mobile positioning data are becoming available for commercial use. The passive positioning trajectory is a list of records <*ms_id*, *cell_id*, *t*> where one record expresses the fact that mobile station (MS) *ms_id* was connected to cell *cell_id* at time *t*.

The *cell_id* must be translated into location in geographical space, as our goal is investigation of movements in geographical space. In reality the exact location cannot be determined but we can estimate the probability density function (PDF) of location based on *cell_id*.

B. *Cellplan data*

Each cell has attributes that can be used to estimate the effective area of the cell. This data is called cellplan. Typically the effective working area of a mobile cell is not well known. The assumed cell area can very significantly differ from the reality which would result in very unrealistic map matching results.

In low resolution analysis one can use simplification that the event is connected directly to location of cell tower, optionally distributed omnidirectionally around the tower [1] or to Voronoi area around the tower [2]. In reality different cells have very different size (diameter of active area) and large cells have overlap with many smaller cells. Cellplan does not provide good information about size of cell area. For increased spatial resolution it is important to have some means to improve cell area estimates.

C. *Map matching*

The process of converting the incoming data into trajectories in real landscape is called trajectory reconstruction or map matching [3]. Map matching uses road network as prior knowledge. Road network defines most probable paths that the observed MS did take, as we assume that MS movement over significant distances is in most cases related to road vehicle movement. This assumption is valid in most cases, but there are exceptions – train traffic, marine traffic, off-road activities like agriculture on large fields, hiking, military exercises etc.

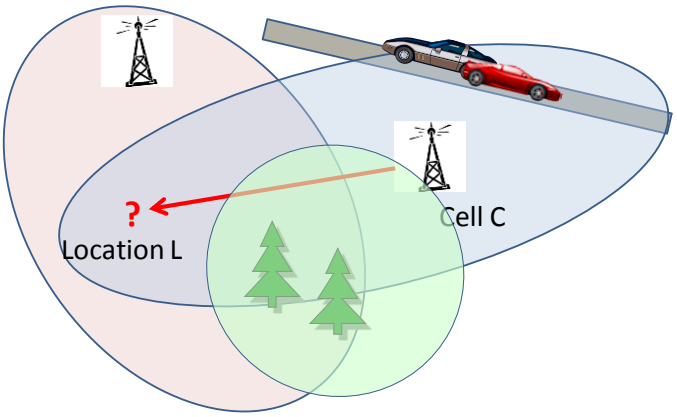

Figure 1. Spatial probability distribution problem for point measurement: If connected to cell C, what is the probability of being at location L, if we know that nobody lives in forest and road has highest concentration of people?

Map matching is sensitive to spatial distortions in radio network model – if the location or the size of effective area for given mobile network cell is wrong then all trajectories referencing the cell can snap to wrong roads and have serious distortions. This type of spatial errors is related to location

error – we assumed that MS is within estimated cell area but in reality it was elsewhere.

D. Clustering-based stop episode detection

Another type of errors is related to estimation of movement and stop episodes. Event sequences generated by stationary MS can be interpreted falsely as movement. For reliable functioning of radio access network (RAN) each location must be covered by several cells. Stationary MS can connect to different cell without any physical movement. During map matching it is not easy to distinguish between smaller real movements of MS from handover of MS from one cell to another in RAN. In map-matching algorithm we introduced clustering that considers sequential trajectory events as one cluster if cell areas have separating gap distance less than some parameter *g*. Of all events in cluster only one is selected to be used in map-matching.

If cell area is in wrong location or smaller than real cellplan then some cell handovers are considered movements. If cell area is estimated too big then one can do opposite error and interpret real movements as cell handovers.

E. Improvement to clustering based on statistical data

In previous chapter we described how clustering can be used to detect stop episodes and reduce phantom movements between cells. This clustering does not use cell areas directly but does depend on distances between cell areas. Our tests showed that most significant distortion was introduced by cases where area of very big cell was severely underestimated. In this case the stop episode could be interpreted as movements, generating movements with span up to tens of kilometers when in reality the MS did not move at all.

Distance between cell areas of any two given cells can be evaluated from trajectory more directly than real cell areas. Distance estimate does not require knowledge which of the two cells had wrong area in cellplan data. .

We developed a statistical procedure that can reveal the situations where the distance between two cells *C1, C2* is considerably overestimated. We do know that vehicles typically do not exceed allowed speed more than 10-20% [4]. It is a well-known data cleansing technique to omit parts of trajectory data with unrealistic speed [5]. In our technique we do not omit the data but enhance the distance estimate between the cells. Distance estimate improvement is based on observation that for any two cells *C1, C2* the maximum velocity of object *v_max* and the distance *d(C1, C2)* define the minimum time delay *delta_t_min=d(C1,C2) / v_max* between two positioning events *e1, e2* where *e1* is connected to *C1* and *e2* to *C2*. So if one observes in data that *delta_t<delta_t_min* then it is contradicting our assumptions and the real distance is less than stated in cellplan. One can use this knowledge to enhance the distance value estimate by replacing the original cellplan-based value with new value based on observed minimum time delay and assumed maximum velocity. The improved estimates are used in clustering of stop episodes.

F. Handling statistical uncertainties

When areas are defined in cellplan, one might assume that all observed events did appear inside the cell area, if the area is correct. Areas defined at 100% confidence level would be very large and therefore not optimal. One defines areas smaller with confidence *a*, for example *a*=0.95. So one gets more compact area where majority of events happen but has to accept that with probability (1-*a*) the event can happen outside the area.

On Fig. 2 is an illustration to the situation. Suppose in trajectory we observe two events *e1, e2* where *e1* is in left orange cell and the next one in right blue cell. In most cases it is true that the object was at *e1* in defined cell area (solid blue bar) in left cell an then at time of *e2* in right cell area, for example as green arrow indicates. The minimum length movement from one area to another is illustrated with black arrow. Due to finite confidence *a* there exists finite probability that at least one of events was outside of events. So one cannot fully exclude that between *e1, e2* the spatial movement was much less, e.g. as indicated by red arrow.

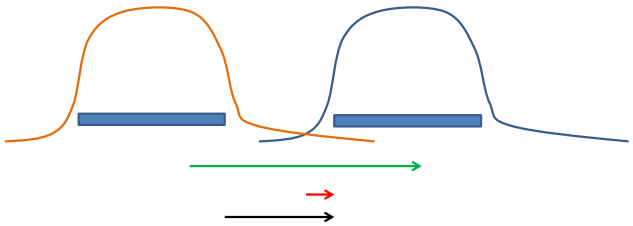

Figure 2. Illustration to probabilistic nature of spatial movement distance between cells. Cells afre presented as 1-dimensional cross-section of spatial probability distribuition field that in reality is 2-dimensional. Orange and blue curved lines are spatial probability distribution functions of two neighbor cells. They have relatively long heavy tails. Dark blue bars indicate the extents of defined cell areas at given confidence level.

When working with real data we observed that the data are not perfect. Explanation could be that a small fraction of events have wrong timestamp or very small fraction of vehicles do move significantly faster than expected (e.g. small planes, extremely fast cars). Therefore it is possible that infrequently the events are wrong, not the cellplan data.

Due to probabilistic nature of cell-to-cell movement data we based our estimates not on minimal observed transition time but used 5% quantile.

II. SOFTWARE DESIGN

Experimental software tool was developed to calculate improved distance estimates. The software for statistical calculations and for trajectory reconstruction was written in java. All intermediate results were stored in RAM. Where possible we used Trove High Performance Collections library [6] instead of standard java collections, to reduce memory footprint and garbage collection overhead.

III. EXPERIMENTS

Some descriptive statistics were generated about distance correction based on real data.

We ran the software on cell trajectory data where there were total 406232 combinations of two cells that had transitions of user from one cell to another. In 83936 cases it was just a single transition, in 40557 cases two transitions. Distribution of transition count shows that small number of cell pair combinations have very big transition count and there are very cell pair combinations with a few transitions (Fig 3).

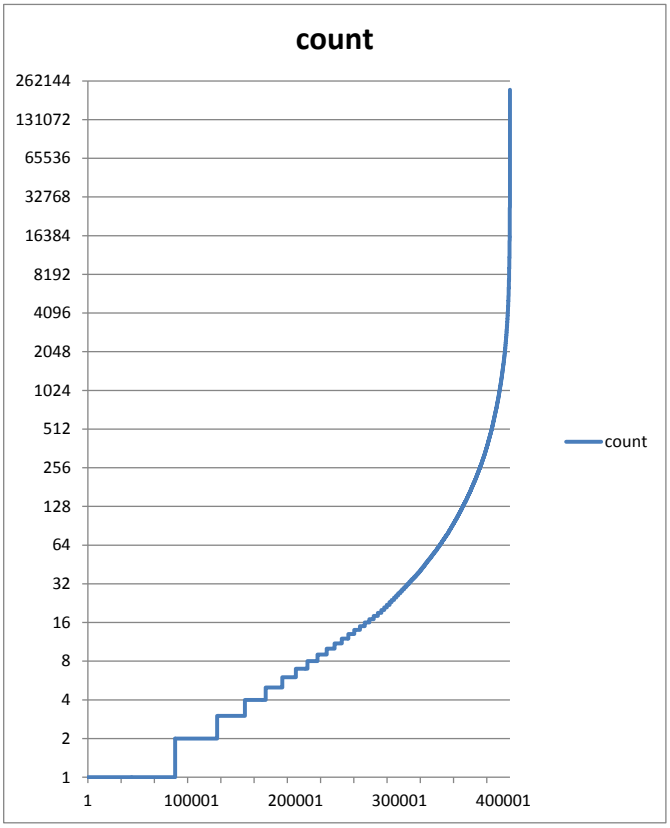

Figure 3. Cumulative graph of transaction count for each observed cell pair combination, sorted by count of transactions. Vertical axis is count of transactions for given cell combination, horizontal axis is number of cell combinations.

On Fig.4 we show how 5% quantile of transition time depends on distance between cell areas, considering only cell combinations with large transaction (count>99). Among these high-traffic cell pairs there are no such cell pairs that the distance is small but 5% quantile of transition time is large. Slope of cut-off limit is ca' 60 km/h. For any strongly connect cell pairss at least 5% of MS moved from one to another at speed at least 60 km/h. **One explanation to this result could be that there are no big-traffic cell pairs that have underestimated distance, but there are numerous cells that have severely overestimated distance.**

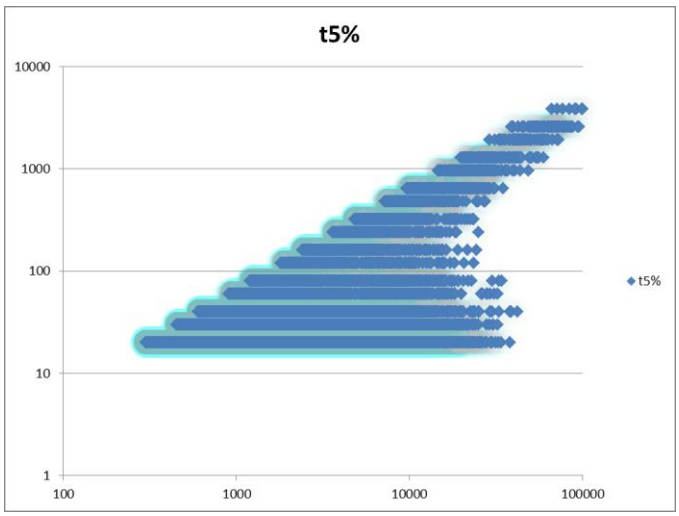

Figure 4. 5% quantile of transition time vs distance between cell areas for cell pairs with transition count 100 or above. Horizontal axis: distance between cell areas, meters. Vertical axis:5% quantile of transition time, seconds. Time was rounded down to values 0, 20, 30, 40, 60, 80, 120, 160, 240, 320, 480, 640, 960, 1280, 1920, 2560, 3840, 5120, 7680, 10240 seconds. This caused forming horizontal stripes in scatterplot. Points with time value rounded to 0 are missing from drawing (because of logarithmic time axis)

IV. DISCUSSION

Some descriptive statistics was created and one interesting observation was found suggesting systematic frequent overestimation of cell distances without any underestimations of distances.. Real test is still pending. We plan to test soon the effects of enhanced estimates on map-matching quality. Before these tests it is too early to draw conclusions.

V. FURTHER THOUGHTS AND FUTURE WORK

Main focus of this presentation is on improved distance estimates. In many cases one does need not only good distance estimates but also would benefit from enhanced cell area geometry estimates. The technique described above is sufficient to improve distance estimates but it does not indicate which of the two cells involved has incorrect area definition. Therefore it is not directly applicable to area enhancement. So-called umbrella cells have very large area. It creates significant trajectory reconstruction distortion if they are not marked in cellplan as having large area.

We are also developing a statistical procedure where one searches for such trajectory segments where cluster of several nearby cells are involved (indicating high probability that MS was practically stationary) and only one distinctive outlier cell is also involved. We know that most of cells in radio network are relatively small and radio network prefers to connect MS to local small cells so we can assume that with very big probability if there are several compactly located cells in cluster then it will consists of local cells and the outlier is reaching to cluster and not the other way round.

Below Fig.5 shows an early result from this algorithm that still needs considerable work to become practical tool in data mining.–

Fig 5. Example of cell area correction prototype. The blue lines are relations "C1 and C2 in same cluster" and white are relations "C1 was observed outlier when given cell C2 was part of local compact cluster". Cell relation lines are drawn from area centroid to area centroid (using uncorrected area definitions as given in cellplan)

VI. REFERENCES

[1] Reades, Jonathan, Francesco Calabrese, Andres Sevtsuk, and Carlo Ratti. "Cellular census: Explorations in urban data collection." *Pervasive Computing, IEEE* 6, no. 3 (2007): 30-38.

[2] Kang, Chaogui, Yu Liu, Xiujun Ma, and Lun Wu. "Towards estimating urban population distributions from mobile call data." *Journal of Urban Technology* 19, no. 4 (2012): 3-21.

[3] 1. Mohammed A. Quddus, Washington Y. Ochieng, Robert B. Noland, Current map-matching algorithms for transport applications: State-of-the art and future research directions, *Transportation Research Part C: Emerging Technologies, Volume 15, Issue 5, October 2007, Pages 312-328*

[4] Destia Eesti AS. Liikluskäitumise monitooring 2008, MAANTEEAMET Tallinn 2008. http://www.mnt.ee/failid/Limo_2008_loppparuanne.pdf

[5] Chung, Edward, Sarvi, Majid, Murakami, Yasunori, Horiguchi, Ryota, Kuwahara, Masao. AustStab; 2003. Cleansing of probe car data to determine trip OD.

[6] Trove High Performance Collections. http://trove.starlight-systems.com/