MELT '09, Orlando, Florida, Sept 30, 2009

Predicting User-Cell Association in Cellular Networks from Tracked Data

Kateřina Dufková[1], Jean-Yves Le Boudec[2], Lukáš Kencl[1], Milan Bjelica[3]

[1] R&D Centre for Mobile Applications (RDC), Czech Technical University in Prague
Technicka 2, 166 27 Prague 6, Czech Republic
{katerina.dufkova, lukas.kencl}@rdc.cz

[2] Ecole Polytechnique Fédérale de Lausanne
EPFL, CH-1015 Lausanne, Switzerland
jean-yves.leboudec@epfl.ch

[3] Faculty of Electrical Engineering (ETF), University of Belgrade
Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia
milan@etf.rs

Abstract. We consider the problem of predicting user location in the form of user-cell association in a cellular wireless network. This is motivated by resource optimization, for example switching base transceiver stations on or off to save on network energy consumption. We use GSM traces obtained from an operator, and compare several prediction methods. First, we find that, on our trace data, user cell sector association can be correctly predicted in ca. 80% of the cases. Second, we propose a new method, called "MARPL", which uses Market Basket Analysis to separate patterns where prediction by partial match (PPM) works well from those where repetition of the last known location (LAST) is best. Third, we propose that for network resource optimization, predicting the aggregate location of a user ensemble may be of more interest than separate predictions for all users; this motivates us to develop soft prediction methods, where the prediction is a spatial probability distribution rather than the most likely location. Last, we compare soft predictions methods to a classical time and space analysis (ISTAR). In terms of relative mean square error, MARPL with soft prediction and ISTAR perform better than all other methods, with a slight advantage to MARPL (but the numerical complexity of MARPL is much less than ISTAR).

1 Introduction

Prediction of future user location is useful to a number of applications, including home automation, road traffic management, wearable computers and context aware applications [1–4]. We are interested in applying location prediction to wireless cellular networks (GSM networks). We seek to estimate the future number of users in different parts of the network, with granularity of a Base Transceiver Station (BTS).

This may have many applications, such as economizing the rental cost of virtual networks, crowd management, provision of real-time network services, or reduction of energy consumption. For example, it is shown in [5, 6] that turning off some of the BTSs when there are few users to serve, and associating these users to neighbouring cells, leads to significant energy savings while maintaining quality of service. Indeed, telephony network operators identify scaling of energy needs with traffic through sleep mechanisms as one of the research challenges of interest for them [7].

As a first step, we would like to evaluate whether it is possible to make some predictions of user association with BTSs, and which prediction methods can be of help. The time scale is 2min, motivated by typical deployment times for near real time network management. Our approach is based on mining the User-Cell association records obtained by active tracking [8]. We evaluate several prediction methods, such as Prediction by Partial Match (PPM), which was successfully used in [1] for location prediction of single users and LAST, which takes as prediction the last visited location. The results motivate us to propose a new method, called "MARPL", which uses Market Basket Analysis to separate patterns where PPM works well from those where LAST is best.

Next, we argue that, in our context, one should make a distinction between *hard* and *soft* prediction. The former predicts the most likely location, whereas the latter gives a spatial distribution. We show how one can transform the hard prediction methods of interest into soft prediction methods. We find that soft predictions are more accurate on our data when tracking an ensemble of users. As a benchmark, we also compare to a classical time and space analysis (ISTAR). The main contributions of the paper are:

- description of a hard prediction method that builds on PPM and Market Basket Analysis to improve prediction;
- transformation of a hard prediction method into a soft prediction method, better suited to the prediction of total number of users at a location;
- comparison, using operator data, of PPM, MARPL, LAST and ISTAR;
- conclusion that user cell sector association can be correctly predicted in ca. 80% of the cases. In term of relative mean square error of user ensemble location estimation, soft methods are better than hard ones, and MARPL with soft prediction and ISTAR perform better than PPM or LAST, with a slight advantage to MARPL (with the added benefit of lower numerical complexity).

The rest of the paper is organized as follows. Section 2 describes the state of the art. Section 3 describes our experimental data. In Section 4 we describe the prediction methods we use. Section 5 presents experimental results and Section 6 concludes the paper.

2 Related Work

Location is an important feature for many applications, and wireless networks can better serve their clients by anticipating client mobility.

González *et al.* in [9] study the trajectories of 100000 mobile phone users over a six-month period. They conclude that the individual travel patterns collapse into a single spatial probability distribution, indicating that it is possible to obtain the likelihood of finding a user in a given location. This further implies that it is possible to quantify the general phenomena driven by human mobility.

Some authors investigate how to obtain datasets which could reliably represent the user's mobility patterns. Sohn *et al.* in [10] showed how coarse-grained GSM data from mobile phones (e.g. readings like signal strength, cell IDs and channel numbers of nearby base station towers) could be used to recognize high-level properties of user mobility. Ashbrook and Starner showed how locations of significance could be automatically learned from GPS data at multiple scales [3]. They describe a system that clusters these data and incorporates them into a predictive Markov model of user's movements. The potential applications of such models would include both single and multi user scenarios. Zang and Bolot in [11] mine more than 300 million call records from a large cellular network operator to characterize user mobility and create mobility profiles. They use passive network monitoring namely in the form of *Per Call Measurement Data* (PCMD) analysis. PCMD records contain data about voice, SMS and data calls performed in the network together with the initial and final cell that served the call. The authors focus mainly on cells where users make call, while we focus purely on user mobility (our data set does not even contain information about calls). Contrary to all these approaches, we use a data set obtained by *active tracking* of selected users' cell associations, without any further "external" location indicators (such as GPS).

Another group of papers investigates methods for predicting user's location. Song *et al.* in [12] present extensive evaluation of location predictors, using a two-year trace of over 6000 users of a Wi-Fi campus network. Even the simplest classical predictors could obtain median prediction accuracy of about 72% over all users with sufficiently long location histories, although accuracy varied widely from user to user. The simple Markov predictors performed comparably or better than the more complicated LZ predictors, with smaller data structures.

There exists a close relation between prediction of discrete sequences and lossless compression algorithms. Begleiter *et al.* in [13] studied the performance of a number of prominent algorithms for prediction of discrete sequences over a finite alphabet, using variable order Markov models. The results show that *Prediction-by-Partial-Match* (PPM) algorithm performed the best. In this paper, we use their implementation of the so-called PPM–C method.

In [1], Burbey and Martin applied the PPM algorithm to data including both temporal and location information. Tests on data traces from IEEE 802.11 wireless network showed that a first-order PPM model had 90% success rate in predicting the user's location, while the third order model was correct 92% of the time. However the studies [12, 13, 1] were performed on data with different attributes, and an order-of-magnitude lower number of distinct locations, or general states, than in our study.

In this work, we discuss using probabilistic (soft) and aggregate predictions for tracking an ensemble of users. When forecasting the aggregate of variables measured over time and in different regions, it is plausible to assume that the individual components will be spatially correlated. Giacomini and Granger investigate forecasting of a Space-Time Autoregressive model aggregate [14]. Min *et al.* further exploit spatio-temporal correlations to road traffic prediction [4]. Their approach inspired us to formulate the *Integrated Space-Time Auto Regressive* prediction model (ISTAR) (see Section 4.3).

Amongst other papers, Hightower and Borriello used a probabilistic approximation algorithm implementing a Bayes filter, known as *particle filter*, to estimate location [15]. Like us, they also use spatial probability distributions, but they focus rather on indoor localization with an order of magnitude higher precision. Thus, their work is not directly applicable to our dataset. Bauer and Deru notice that relevance of some piece of information is connected to the places a user is likely to visit [16]. They used a variety of machine-learning techniques to derive motion profiles of WLAN users. Their primary goal was not location prediction; instead, they use these profiles to recommend the information which might become useful to the observed user in the foreseeable future.

We end this section with a brief overview of traffic prediction models for wireless networks. Shu *et al.* used seasonal autoregressive integrated moving average (ARIMA) model to capture the behavior of a GSM network traffic stream [17]. Tikunov and Nishimura use a technique known as Holt-Winter's exponential smoothing [18], while Hu and Wu use chaos theory [19].

3 Experimental Data

Mobile cellular networks contain various user data that can be used for location estimation. In the spatial domain, typically the granularity is the *user-cell association*. Finer precision may be gained using triangulation from multiple base stations, but this requires additional sophistication (such as location services platforms), either on the user terminal or on the network side.

Call Detail Records (CDRs) are stored by the telephony network operators. They contain traffic data, including cell association, but only of active users. Mobile terminals themselves may also report their GPS coordinates or currently visible cells (e.g. Google Latitude [20]) over the network, but this requires user cooperation. Cell association of passive, non-communicating users, is beyond the reach of majority of methods, as those users are reporting their location only sporadically using a procedure called location update. A location update is done when a user crosses boundaries of the so called "location areas" (those are geographically large, consisting of hundreds of cells) or after a significant time (order of hours for the network studied in this work). Thus passive users must be tracked actively — the user-cell association observations have to be polled or user-reported.

Data used in this work were obtained by *active tracking of a group of mobile phone users* (unlike in [9]), using the platform from [8]. The platform allows to

periodically poll and store cell association of a set of users in a real-time manner and without user cooperation. The users were selected from a list of users who did a location update in the studied network recently, the focus group being foreign roamers. The polling interval was set to 2 minutes, and the association was recorded for all selected users, including passive. The trace contains 72 hours of tracking in December 2008, with 2731 distinct real users moving around in an existing country-wide GSM network. The total number of cells visited by the users was 7332 (not all cells of the network were visited).

For each user, we obtain a sequence of his/her associations, each being either a *Cell identification*, or one of the special states: *Offline*, for users having switched their mobile phones off; *Rival*, for users having left the network to a rival national mobile operator; and *Abroad*, for users having left the network to a foreign operator. The state space thus contains 7335 states.

The trace of user-cell associations represents a sequence of regular location observations [12], the spatial dimension of user mobility. Although previous work has experimented with incorporating both time and space into a single sequence [1], due to the high number of distinct states and amount of data available we chose to deal just with the spatial dimension. Thus, we have removed the *Offline* state from the data, as it seems to depend rather on time of day heavily. We split the user traces around the *Offline* state.

When analyzing user mobility, we observe that the probability of staying at the same location (i.e. being associated with the same cell) is very high and only slowly decreasing over time (see Fig. 1, left), in harmony with [9]. The fashion of selecting users for the tracking implies their higher mobility at the beginning of tracking, as majority of the users are put into the tracking when they are moving, the typical case being a roamer entering country (and the studied network) traveling to a particular destination (see Fig. 1, right).

4 Predicting Location

Assume we have a finite set of users $\mathbb{I} = \{1, 2, ..., I\}$ and a finite set of cells (base stations, access points, etc.) $\mathbb{J} = \{1, 2, ..., J\}$ of a cellular network. Assume we can observe the cell association $a^i(t) \in \mathbb{J}$ for any user $i \in \mathbb{I}$ and any time $t \in \mathbb{N}$. Let $A^i = \{a^i(t)\}$, $t \in \mathbb{N}$ be a sequence of observations of cell association for a user $i \in \mathbb{I}$ over discrete equidistant time slots. Let $Y_j(t)$ be the number of users associated with cell $j \in \mathbb{J}$ in a time slot $t \in \mathbb{N}$.

4.1 Hard vs. Soft Decisions

Assume H^i is a sequence of previous associations of a user i. We define the *hard decision location prediction problem* as the task of finding a single location $j \in \mathbb{J}$ with the highest $Prob(j|H^i)$, where the user i will most likely be at the next time slot. We define the *soft decision location prediction problem* as the task of constructing a vector $U^i = [u^i_j]$, $u^i_j = Prob(j|H^i)$ of probabilities for a user $i \in \mathbb{I}$ to be at any possible location $j \in \mathbb{J}$.

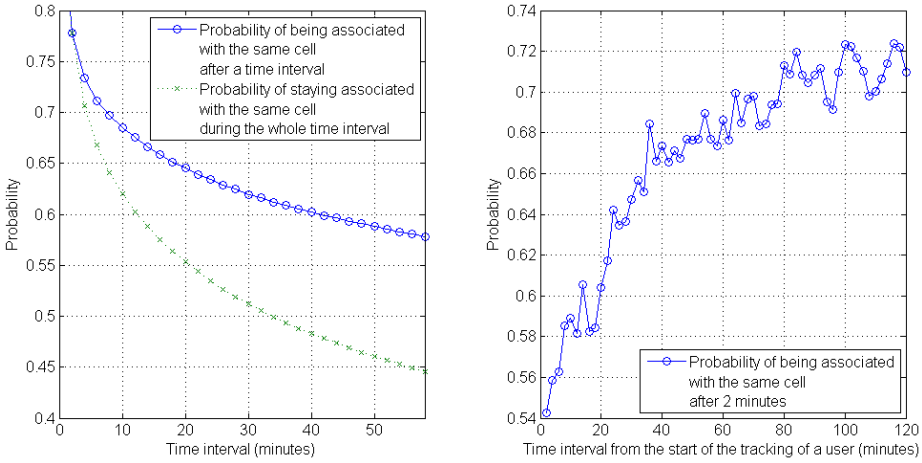

Fig. 1. Left: Probability of a user being associated with the same cell, for different time intervals, mean values over whole tracking. **Right:** 2-minutes mobility of users as function of time interval from the start of the tracking. Due to specific focus on roamers, mobility is higher at the start of the tracking.

While the predictors that provide a hard decision on the next location of the user are useful in many applications, the "winner takes all" strategy does not have to be optimal for all applications. One of them is the application we study in this paper, where *aggregation* is used to obtain network-wide statistics about numbers of users associated with individual cells (see Figure 2).

We formalize the task as follows: Knowing $a^i(s)$, $i \in \mathbb{I}$, $s \in \{1, 2, ..., t-1\}$ and $Y_j(s)$, $j \in \mathbb{J}$, $s \in \{1, 2, ..., t-1\}$ we want to predict $Y_j(t)$. For practical reasons, as we do not want to store much historical data, we want to base the prediction just on the last r values, i. e. on the values related to $s \in \{t-r, t-r+1, ..., t-1\}$.

4.2 Individual Hard Decision Methods

LAST predictor. The simplest possible predictor, which always uses the last known value as the prediction, will be used as a reference for proposed methods.

PPM predictor. The well-known *Prediction-by-Partial-Match* (PPM) algorithm that uses variable order Markov models. We use implementation of the so-called PPM–C method provided by [13].

MARPL predictor. We propose a method called MARPL (MARket basket analysis + Ppm + Last), which combines PPM and LAST predictors, after splitting the problem into subproblems according to the last few associations of the user, and choosing the best strategy for every subproblem independently.

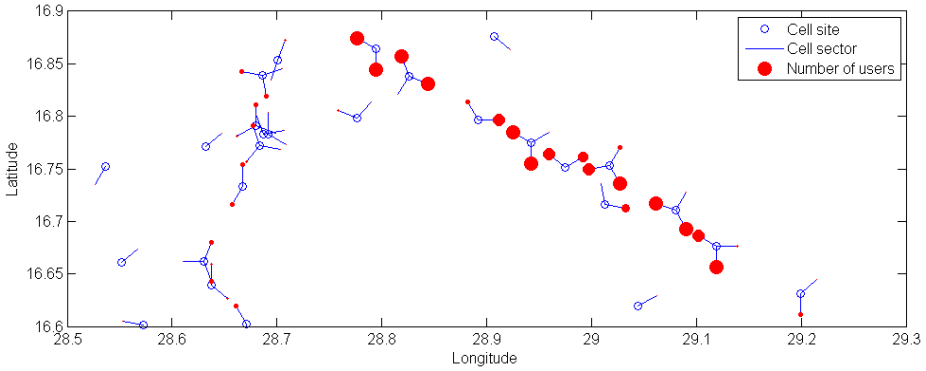

Fig. 2. Example of a cellular network with BTS sites hosting multiple cells with different transceiver directions. A road is recognizable from the higher numbers of users.

The splitting is loosely inspired by the Market Basket Analysis method [21] and its way of discovering hidden rules in the data, with the difference that the original method was intended for unordered sets of elements instead of ordered sequences. We construct set of all possible rules of order r, each rule being of the form $H_1 H_2 ... H_r \rightarrow P$, where $H_s \in \{A, B, C...\}$ represents the history of the last r associations of a user and $P \in \{A, B, C...\}$ represents the predicted association. The $A, B, C, ...$ symbols are wildcards as we are interested in generally applicable rules. For example rule $AABB \rightarrow B$ represents the situations where, after observing a cell A twice and then another cell B twice, the next cell is B.

We define applicability and reliability of a rule as follows (L(rule) denotes the left side of a rule, R(rule) the right side of a rule):

$$\text{Applicability(rule)} = \frac{\#\text{ possible usages}}{\#\text{ all predictions}} = Prob\left(L(\text{rule})\right), \tag{1}$$

$$\text{Reliability(rule)} = \frac{\#\text{ successful usages}}{\#\text{ possible usages}} = Prob\left(R(\text{rule})|L(\text{rule})\right). \tag{2}$$

We split the problem as follows:

1. Use the LAST predictor on subproblems, where the rule corresponding to the LAST predictor has strictly higher reliability than the PPM predictor success rate (54,5%, see Section 5). See Table 1.
2. Otherwise use the PPM predictor with a fallback to the LAST predictor on cases where the PPM is "not sure". The level of certainty of the PPM prediction can be obtained as the likelihood $Prob(\text{Predicted symbol}|H^i)$; we accept the PPM prediction only if its likelihood is above certain threshold.

Table 1 summarizes results of the analysis for our data and $r = 4$, which proved best in the experiments. The thresholds were set according to the reliability of the LAST predictor on the subproblem (see Table 1). The lower the percentage

of good predictions that LAST predictor would make, the lower the threshold and, consequently, the lower the number of fallbacks to the LAST predictor.

The reason we chose to use directly the LAST predictor on some subproblems (instead of using high threshold) is performance. The subproblems where we use LAST predictor together make 77% of the cases, so the MARPL achieves remarkable speedup of the prediction process, compared to the PPM predictor.

Finally, selection of the training data needs care. The staying pattern (rule $AAAA \rightarrow A$) is dominant in the dataset, but useless for the PPM predictor, as it will never be used on this kind of data. We considered three training phase strategies — using *all available data*, using *selected overlapping subsequences of length* $r+1$, and using *selected non-overlapping subsequences of variable length*. The overlapping sequences strategy omitted the sequences that contained just one symbol, the non-overlapping sequences strategy continued to grow the current subsequence until the staying pattern was recognized, and then started a new sequence, omitting the repeating symbols. The selected non-overlapping subsequences proved best in the experiments and will be used further.

Table 1. Market Basket Analysis for sequences of associations A^i for rules of order 4. Each row represents all rules with the same left side. The rules can be classified into two user behaviour patterns — *stay* and *move*. Staying (represented by the $AAAA \rightarrow A$ rule) prevails greatly, the rest of the rules relate to moving users. The star marks the subproblems where the threshold chosen according to the reliability of the LAST predictor did not perform well, and was changed to more appropriate value.

Rules	Applicability (%)	Reliability (%) A	B	C	D	E	LAST reliability	Algorithm	Threshold (%, rounded up)
$A, A, A, A \rightarrow ?$	**66.8**	*96*	4	-	-	-	*96*	LAST	-
$A, B, C, D \rightarrow ?$	**8.7**	0	1	2	*18*	79	*18*	PPM	18
$A, A, A, B \rightarrow ?$	4.0	29	*45*	26	-	-	*45*	PPM	46
$A, B, B, B \rightarrow ?$	4.0	12	*70*	18	-	-	*70*	LAST	-
$A, A, B, B \rightarrow ?$	3.0	17	*60*	23	-	-	*60*	LAST	-
$A, B, C, C \rightarrow ?$	2.7	3	4	*40*	53	-	*40*	PPM	40
$A, A, B, C \rightarrow ?$	2.6	4	6	*31*	59	-	*31*	PPM	41*
$A, B, B, C \rightarrow ?$	2.3	4	8	*30*	58	-	*30*	PPM	31
$A, A, B, A \rightarrow ?$	1.6	*67*	21	12	-	-	*67*	LAST	-
$A, B, A, A \rightarrow ?$	1.6	*72*	16	12	-	-	*72*	LAST	-
$A, B, B, A \rightarrow ?$	0.9	*55*	31	14	-	-	*55*	PPM	45*
$A, B, A, B \rightarrow ?$	0.7	41	*48*	10	-	-	*48*	PPM	49
$A, B, C, B \rightarrow ?$	0.5	8	*43*	16	32	-	*43*	PPM	43
$A, B, A, C \rightarrow ?$	0.4	16	8	*37*	39	-	*37*	PPM	37
$A, B, C, A \rightarrow ?$	0.3	*46*	12	17	25	-	*46*	PPM	46

4.3 Aggregated Soft Decision Methods

In this section we transform MARPL and PPM predictors to provide *soft decisions*. Then we propose another approach, that does not take into account individual users and predicts the number of users directly.

MARPL soft predictor. The MARPL predictor provided just the single most likely next location. Instead of it a vector $U^i = [u^i_j]$, $u^i_j = Prob(j|H^i)$, $j \in \{1, 2, ..., J\}$ of probabilities for a user i to be at all the possible locations $j \in \{1, 2, ..., J\}$ is now needed. We construct the vector as follows.

• For the subproblems where PPM is used, $u^i_j = Prob(j|H^i)$ where H^i is the association history of user i.

• For the subproblems where LAST is used, $u^i_j = 1$ if j is the prediction obtained by LAST, $u^i_j = 0$ otherwise.

• By aggregating the vectors U^i for all the users $i = \{1, 2, ..., I\}$ we obtain the prediction $\hat{Y}_j(t) = \sum_{i=\{1,2,...,I\}} u^i_j$.

PPM soft predictor. Created from the PPM predictor by the same procedure as MARPL soft predictor (the second branch is never used).

Integrated Space-Time Auto Regressive model (ISTAR). The proposed method is a time series analysis method inspired by [4] on road traffic prediction. Assume we have an adjacency matrix $A_{i,j}$ such that $A_{i,j} = 1$ if a user can move from location i to location j within one time step (at the highest possible speed). Otherwise $A_{i,j} = 0$. The matrix A is static, derived by comparing the distances between all pairs of BTS with a fixed distance threshold D. Recall that $Y_j(t)$ is the number of users at location j at time t. We apply differencing, as is common in time series analysis, and define $X_j(t) = Y_j(t) - Y_j(t-1)$. The model is:

$$X_j(t) = \sum_{i:A_{i,j}=1} \alpha_{i,j} X_i(t-1) + \beta_j X_j(t-1) + \epsilon(t) \tag{3}$$

where $\epsilon(t)$ is Gaussian white noise. The parameters to be estimated are the matrix α ($J \times J$), the vector β ($J \times 1$) and the noise variance (J is the number of locations). At time t, the prediction for $X_j(t+1)$ is $\hat{X}_j(t) = \sum_{i:A_{i,j}=1} \alpha_{i,j} X_i(t) + \beta_j X_j(t)$. The parameters α and β are estimated by minimizing

$$\hat{\sigma}^2_t := \frac{1}{tJ} \sum_j \sum_{s=2}^{t} w^{t-s} \left(X_j(s) - \hat{X}_j(s-1) \right)^2 \tag{4}$$

where w is a "forgetting" factor, close to 1 and less than 1. Finally, the one-step-ahead prediction for $Y_j(t+1)$ is $\hat{Y}_j(t) = \hat{X}_j(t) + Y_j(t)$.

4.4 Algorithm Complexity

The complexity of predicting the next state of the whole network is considered.

PPM & MARPL. Given the implementation we use, the complexity of PPM prediction for I users and histories of r associations is $O(I \cdot J \cdot r^2)$. For MARPL, the complexity of predicting is $O(r)$ for the decision between the PPM and LAST plus $O(1)$ for the 77% of cases where the LAST predictor is used, or PPM prediction complexity for the rest of the cases. For both, the time complexity of learning one sequence of length n is $O(n)$ and the space required for the worst case is $O(r \cdot n)$, where r is the order of the model [13].

ISTAR. Theoretically, the complexity of predicting the next value for all J locations is $O(J^2)$. The complexity of estimating the α and β parameters is determined by the complexity of computing Equation 4 ($O(t \cdot J^2)$ where t is number of time slots) and complexity of minimization. As minimization algorithm we use Matlab function *lsqnonlin* with default Trust-Region-Reflective algorithm (whose complexity is $O(\text{iterations} \cdot \text{parameters})$) on $O(J^2)$ parameters corresponding to the fraction of ones in adjacency matrix A. Thus the overall parameter estimation worst case complexity is $O(t \cdot J^4 \cdot \text{iterations})$ and $O(J^2)$ space is required. Practically, on large networks the matrix A will become sparse and the J^2 factor can be replaced with J^a, $a \in [1, 2)$, leading to $O(t \cdot J^{2a} \cdot \text{iterations})$ complexity.

For our data ($I = 2731$, $J = 7335$, $r = 4$, $t = 60$, $n =$ ca. 170000) the complexity (in terms of both space and time) of soft PPM and soft MARPL is one order of magnitude lower than that of ISTAR.

5 Experimental Results

5.1 Individual Hard Decision Methods

Data. To use the PPM predictor the data need to be divided to training and test groups. The original data of 2731 users were pseudo-randomly split to 20 groups and experiments were repeated 20 times, each time with one group as test data and the rest of groups used as training data. Each test group contained 96681 subsequences of length 4 with correct next association for evaluation purposes.

Comparing MARPL, PPM and LAST predictors. Figure 3 compares the hard predictors by means of both percentage of correct predictions and distribution of distances between the real and predicted cell. Note that 0m distance between the real and predicted cell occurs in two cases — when correct sector on correct base station is predicted (denoted as *OK BTS+sector*), and when another sector on correct base station is predicted (denoted as *OK BTS*). The difference stems from the cellular network architecture, where a base station often holds more transceivers, serving different sectors and cells, most commonly three.

The MARPL predictor performs best, achieving 79.3% success rate when the exact prediction of BTS and sector is required, and 84.4% success rate when

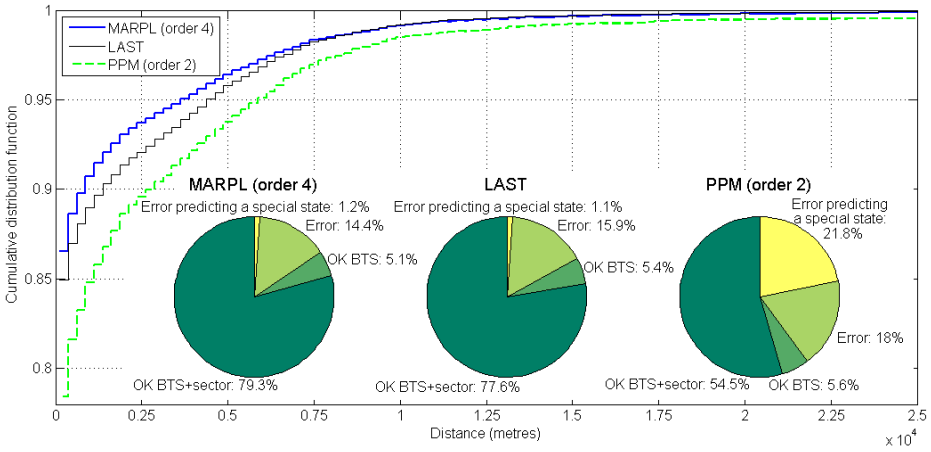

Fig. 3. Comparison of hard predictors, cumulative distribution function of distances between the real and predicted cell. Inset are pie charts showing overall success rate wrt. exact next cell-ID prediction. For both PPM and MARPL only the results of the best performing model are shown for brevity (order 2 for PPM, order 4 for MARPL). PPM is markedly the worst of the predictors, LAST and MARPL provide similar results, with slight advantage of MAPRL. However both PPM and MARPL can be improved by introducing soft decisions, while LAST has no soft decision variant.

the prediction of BTS suffices. From the perspective of predicting user location to switch off under-utilized hardware, the above results are encouraging, as the lower distance errors prevail markedly. We can conclude that the MARPL is able to predict correctly 94% associations with error up to 2500 metres, which is acceptable given the typical cell overlays in cellular networks.

Surprisingly the PPM predictor performs worse than the LAST predictor. The reason is that the LAST predictor builds on the low mobility of users (see Figure 1), while PPM has to deal with problems related to the character of our data — the *high number of distinct cells* to associate with, the consequent *training data shortage* and finally the *PPM predictor behavior when "not sure"*. Here PPM predicts the most frequent symbol of the training data (universal *Rival* state for our data), while having in mind the Figure 1, the best strategy is to predict the last known value. The MARPL predictor overcomes these problems by using PPM on the subset of data coming from moving users, and LAST on the data from staying users.

5.2 Aggregated Soft Decision Methods

Data. The splitting to training and test data was the same as in previous section. From the test data, just the users with associations history long enough to predict 60 consecutive time slots were selected, which makes 1296 users and total of 77760 predictions in all 60 time slots.

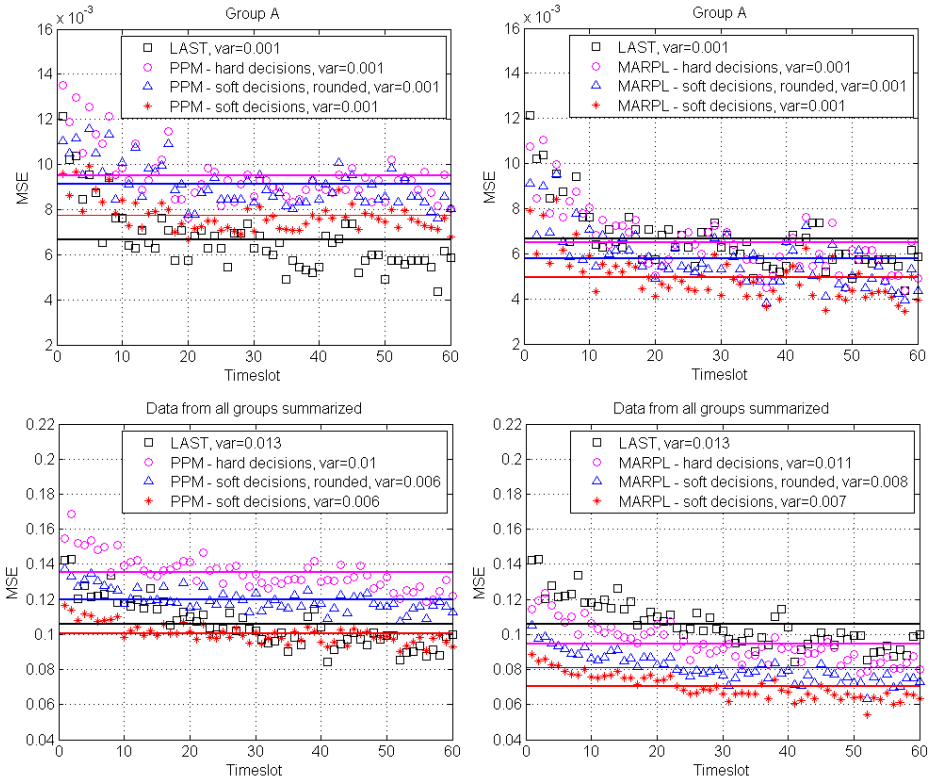

Fig. 4. Comparison of aggregated soft and hard predictors over 60 consecutive time slots, MSE. Top graphs show values for a single group of test data (96 users), bottom graphs for all test groups (1296 users). The left graphs show variants of the MARPL predictor, the right graphs of the PPM predictor. The reason why the MSEs are generally low, especially for single group of test data, is that we have only 96 (or 1296) users moving around 7332 cells, which implies large number of empty cells (where all predictors succeed), pushing the MSE down. Perhaps also surprising is that all predictors improve over time, even though the LAST predictor obviously does not learn from past data. This is due to the diminishing mobility of users over time (see Fig. 1).

Comparing soft and hard predictors. Fig. 4 compares the aggregated predictions from soft and hard versions of MARPL and PPM predictors by means of mean squared error (MSE) between the vectors $[\hat{Y}_j(T)]$, $j \in \mathbb{J}$ obtained using the predictors, and the real vector $[Y_j(T)]$, $j \in \mathbb{J}$. MARPL consistently achieves lower MSE than the PPM predictor, and soft predictors consistently achieve lower MSE than the hard predictors, both for single group of test users and for all groups. The mean MSE for MARPL soft predictor is 0.070, which is just 66.4 % of the mean MSE of LAST (0.106) and 69.8 % of the mean MSE of PPM soft (0.101).

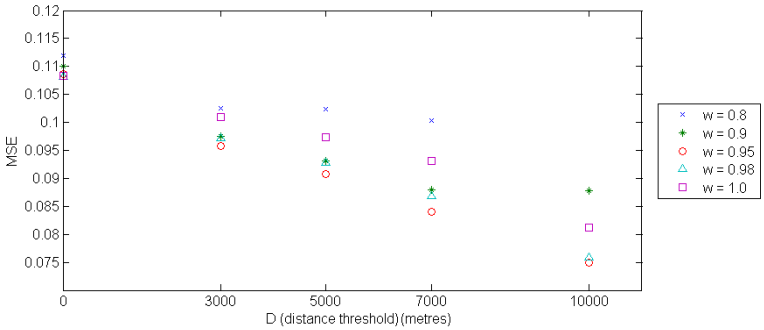

Fig. 5. The ISTAR model performance given by means of MSE for different combinations of parameters. The model improves with higher D and works best for $w = 0.95$.

On our dataset, the growing size of population does not affect the results. While the absolute MSE grows with the number of users in the population, the MSE relative to the number of users remains approximately the same, making the order of the methods stable for all population sizes we considered.

Optimal parameters of the ISTAR model. The parameters of the model are the "forgetting" factor w and distance threshold D, which determines the number of ones in adjacency matrix and thus the computation complexity. Figure 5 concludes that ISTAR improves with higher D and works best for $w = 0.95$.

Comparing aggregated location predictors and ISTAR model. Finally we compare the aggregated results of the location predictors and of ISTAR with optimal parameters. Due to the computational requirements of ISTAR (see Section 4.4), the comparison was feasible on only a subset of 59 cells in one geographical district. The results of location predictors were obtained by restricting the results from the experiment over the entire dataset to the selected cells. This raises the question if it is fair to compare models trained on larger data to ISTAR, but why ignore MARPL's and PPM's capability to train on larger datasets. Regarding test data, the neighborhood errors at the region's borders may influence ISTAR, but not enough users associated to the selected cells for 60 consecutive time slots were available to fairly scale down the location predictors tests. The results (see Table 2) conclude that the MARPL soft predictor performs best out of the studied methods.

6 Conclusions

We show that predicting user location within a cellular network in the next time interval, with the granularity of the associated BTS, is a feasible task with acceptable performance. On our experimental data, best results are achieved using a novel prediction method, MARPL, which combines Prediction by partial match (PPM) and LAST location predictor, using Market Basket Analysis. This

Table 2. The overall MSE achieved by the studied methods (ordered from best to worst). We specify the type of results for each method, for real number predictors (soft predictors and ISTAR) rounding is considered.

Method	MARPL	ISTAR	MARPL	ISTAR	MARPL	PPM	LAST	PPM	PPM
Decisions	Soft	-	Soft	-	Hard	Soft	Hard	Soft	Hard
Result	$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R} \to \mathbb{N}$	$\mathbb{R} \to \mathbb{N}$	$\mathbb{N}$	$\mathbb{R}$	$\mathbb{N}$	$\mathbb{R} \to \mathbb{N}$	$\mathbb{N}$
MSE	0.0715	0.0750	0.0864	0.0890	0.0949	0.1228	0.1263	0.1537	0.2144

is an initial result on a limited (size) and specific (roaming clients) data set — general applicability to arbitrary cellular network mobility data will need to be verified in the future.

Further, we argue that the soft, probabilistic prediction methods are more useful in predicting the aggregate location of a user ensemble, as shown using mean square error comparison. Predicting location as a probabilistic vector, or aggregate location of an ensemble of users, makes sense due to a number of potential applications focusing on network resource optimization. We show that the soft methods in general outperform the hard ones, with MARPL requiring fewer resources. In our future work, we intend to focus on practical applications of the predictions for tasks such as economizing cellular network energy consumption.

Acknowledgment. We wish to thank Vodafone Czech Republic a.s. for their generous support of the project.

References

1. Burbey, I., Martin, T.L.: Predicting future locations using prediction-by-partial-match. In: MELT '08: Proceedings of the first ACM international workshop on mobile entity localization and tracking in GPS-less environments, San Francisco, California, USA, ACM (September 2008) 1–6
2. Das, S.K., Cook, D.J., Battacharya, A., Heierman, Lin, T.Y.: The role of prediction algorithms in the MavHome smart home architecture. Wireless Communications, IEEE **9** (2002) 77–84
3. Ashbrook, D., Starner, T.: Using GPS to learn significant locations and predict movement across multiple users. Personal and Ubiquitous Computing **7** (2003) 275–286
4. Min, W., Wynter, L., Amemiya, Y.: Road traffic prediction with spatio-temporal correlations. In: Proceedings of the Sixth Triennial Symposium on Transportation Analysis, Phuket Island, Thailand (June 2007)
5. Chiaraviglio, L., Ciullo, D., Meo, M., Marsan, M., Torino, I.: Energy-aware UMTS access networks. In: Proceedings of the 11th International Symposium on Wireless Personal Multimedia Communications (WPMC '08), Lapland, Finland (September 2008)
6. Marsan, M.A., Chiaraviglio, L., Ciullo, D., Meo, M.: Optimal Energy Savings in Cellular Access Networks. In: Proceedings of GreenComm '09 — First International Workshop on Green Communications, Dresden, Germany (June 2009)

7. Lister, D.: An operator's view on green radio. Keynote presentation at First International Workshop on Green Communications (GreenComm '09) (June 2009)
8. Dufková, K., Ficek, M., Kencl, L., Novák, J., Kouba, J., Gregor, I., Danihelka, J.: Active GSM cell-id tracking: "Where did you disappear?". In: MELT '08: Proceedings of the first ACM international workshop on mobile entity localization and tracking in GPS-less environments, San Francisco, California, USA, ACM (September 2008) 7–12
9. González, M.C., Hidalgo, C.A., Barabasi, A.L.: Understanding individual human mobility patterns. Nature **453**(7196) (June 2008) 779–782
10. Sohn, T., Varshavsky, A., Lamarca, A., Chen, M., Choudhury, T., Smith, I., Consolvo, S., Hightower, J., Griswold, W., de Lara, E.: Mobility detection using everyday GSM traces. In: Proceedings of the Eight International Conference on Ubiquitous Computing (UbiComp '06), California, USA (September 2006) 212–224
11. Zang, H., Bolot, J.C.: Mining call and mobility data to improve paging efficiency in cellular networks. In: MobiCom '07: Proceedings of the 13th annual ACM international conference on Mobile computing and networking, Montréal, Québec, Canada, ACM (September 2007) 123–134
12. Song, L., Kotz, D., Jain, R., He, X.: Evaluating location predictors with extensive Wi-Fi mobility data. In: Proceedings of the 23rd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '04). Volume 2., Hong Kong, China (March 2004) 1414–1424
13. Ronbeg, R.B., Yona, G.: On prediction using variable order Markov models. Journal of Artificial Intelligence Research **22** (2004) 385–421
14. Giacomini, R., Granger, C.W.: Aggregation of space-time processes. Boston College Working Papers in Economics 582 (July 2002)
15. Hightower, J., Borriello, G.: Particle filters for location estimation in ubiquitous computing: A case study. In: Proceedings of the Sixth International Conference on Ubiquitous Computing (UbiComp '04), Nottingham, England, UK (September 2004)
16. Bauer, M., Deru, M.: Motion-based adaptation of information services for mobile users. In: Proceedings of the User Modeling 2005 (UM '05), Edinburgh, Scotland, UK (July 2005) 271–276
17. Shu, Y., Yu, M., Liu, J., Yang, O.: Wireless traffic modeling and prediction using seasonal ARIMA models. In: Proceedings of the IEEE International Conference on Communications, 2003 (ICC '03). Volume 3., Anchorage, Alaska, USA (May 2003) 1675–1679
18. Tikunov, D., Nishimura, T.: Traffic prediction for mobile network using Holt-Winter's exponential smoothing. In: Proceedings of the 15th International Conference on Software, Telecommunications and Computer Networks (SoftCOM '07), Portsmouth, UK (September 2007) 1–5
19. Hu, X., Wu, J.: Traffic forecasting based on chaos analysis in GSM communication network. In: Proceedings of the International Conference on Computational Intelligence and Security Workshops (CISW '07), Harbin, Heilongjiang, China (December 2007) 829–833
20. Google: Latitude project, www.google.com/latitude/
21. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: SIGMOD '93: Proceedings of the 1993 ACM SIGMOD international conference on management of data, Washington, D.C., United States, ACM (May 1993) 207–216