Pedestrians Complex Behavior Understanding and Prediction with Hybrid Markov Chain

Mostafa Karimzadeh, Zhongliang Zhao, Florian Gerber, Torsten Braun
Institute of Computer Science, University of Bern, Switzerland
Email : {karimzadeh, zhao, braun}@inf.unibe.ch, florian.gerber@students.unibe.ch

***Abstract*—The prevalence of smartphones equipped with global positioning system has enabled researchers to excavate users mobility patterns in the cities. The knowledge of users' behavior, such as their locations, plays a significant role in location-based services, resource management, logistic administration and urban planning. To understand complex behavior of humans we utilize spatio-temporal analysis on collected geo-location points to exploit Individual Zone of Interests in urban areas. In addition, we designed a hybrid Markov chain model to forecast future locations of pedestrians. Compared to existing mobility prediction methodologies, our predictor can adapt it's behavior constantly based on the quality of existing traced data to switch between first-order or second-order Markov chain. Moreover, we propose a model to predict city area congestion. The model predicts the number of users in a specific area of a city by discovering the regular mobility patterns of a group of users. We conducted comprehensive empirical experiments using a real-life dataset, namely the Mobile Data Challenge dataset, which was collected in the city of Lausanne in Switzerland with around 180 participants. We found a satisfactory user future location prediction accuracy of $70-84\%$ and area congestion prediction accuracy of $65-73\%$ for the users.**

***Index Terms*—Mobile analysis, Mobility and Congestion Prediction, Mobility Behavior, Location based Services.**

I. Introduction

Extracting meaningful information from collected trace data of users to determine their movement pattern is an important part of location-based services (*LBSs*). For example, to predict future behavior of a mobile user, mobility predictors rely on clustering techniques to capture a user's Individual Zone of Interests (*I-ZOIs*) from collected trace data. Intuitively, an I-ZOI is a city area that an individual user visits frequently and the user spends considerable time in this region. Typically, LBSs are using mobility prediction as a means to improve quality of service by providing context-aware information to users before arriving them to destination.

In the last decade, with the increasing adoption of services such as *Google Now*, which proactively collects data, such as Bluetooth/WiFi connectivity, call/SMS log, information about running applications, large size of heterogeneous data is accumulated. This knowledge is essential for humans in their daily life activities. Another popular service, *Moves* enables automatic recording of any walking, cycling, and running of users and displays pertinent information, such as traveled distance, duration and calories burned for each activity. It is apparent from the above examples that location based services are prospering, giving a notable chance to collect contextual data about visited locations of users. In addition to mobility prediction, area congestion prediction in large cities is also of great importance. The past decades have witnessed a rapid development of modern cities accompanied with an increasing demand for mobility [1], accounting for the conflict between the limited resource capacities and the increase of traffic demand reflected by severe user congestion in hot spot regions. Induced by such a problem, several negative impacts arise for citizens, e.g., economic losses, reduction of travel efficiency and accessing to resources. Fortunately, accumulated data from smartphones and LBSs have been used to forecast congested areas in smart cities. The type of dataset plays an important role in accurate location prediction as the prediction algorithms learn user movement patterns from collected data [2]. To examine prediction performance of proposed models, we use a large scale, real-world dataset from the Nokia Mobile Data Challenge, collected in the city of Lausanne by almost 180 participants. In this work, we propose a dynamic Markov chain-based model, which adaptively selects the first-order or the second-order Markov chain model based on the available trace quality. We propose a clustering algorithm to discover frequently visited places by the user and integrate it with a mobility predictor to predict a user's future behavior. Moreover, in order to estimate number of users in frequently visited places we propose a congestion prediction algorithm. The system overview is depicted in Figure 1. The contributions of our paper are as follows:

- We design a mobility prediction algorithm that benefits from both the first-order Markov chain and the second-order Markov chain to forecast users' future location.
- We propose the Zone of Interest discovery scheme, which helps us to model the mobility behavior of users.
- We introduce a mechanism to predict congestion areas that are frequently visited by users.
- We evaluate our mobility and congestion predictors using a real-life dataset, obtaining consistently satisfactory results.

The remainder of the paper is organized as follows. Section II discusses research efforts. Section III presents our system model and introduces some preliminaries used in the paper. Sections IV and V discuss the methodology to evaluate the mobility model and demonstrate obtained results respectively. Finally, in Section VI we conclude our paper by sketching future research directions.

II. Related Work

Understanding human mobility by mining raw GPS logs has been a long-standing subject in academic research [3]. These approaches rely on clustering user visits to extract hot spots. The premier research in adapting the data clustering algorithms for modeling mobility behavior [4] proposes to iteratively extract hot spots of users. Montoliu et al. proposed a clustering algorithm [5], where GPS coordinates (*latitude and longitude*) are clustered in the temporal domain to detect the stay points that are used to derive frequently visited regions using a grid-based clustering approach.

The above techniques use several temporal bounds to classify a particular region as a cluster. Some parameters include maximum distance between the collected locations, maximum/minimum time bound of visited places and cluster shapes. However, if we only take into account the temporal bounds, this leads to some inaccuracies in estimating the total number of clusters belonging to a user. As opposed to using only temporal metrics to cluster the individual regions, we form the clustering algorithm by benefiting both temporal and spatial metrics (*instantaneous velocity, average velocity*) to quantify the correlation between visited regions.

Regarding the mobility prediction methods, a majority of proposed models first explore movement patterns and consequently employ this knowledge to predict next movements [6]. In recent years Mobility Markov Chain (*MMC*) algorithms have been used widely to forecast future behavior of users, due to their simplicity, low execution time and good prediction performance [7]. In [8] authors have found that Markov-based algorithms are performing better than more complex and more memory consuming algorithms such as Sampled Pattern Matching (*SPM*) or Prediction by Partial Matching (*PPM*). The author in [2] proved that Neural network based approaches suffer from high computation complexity. In [9] authors proposed Markov-based algorithm to predict next location using non-Gaussian data. Using higher order Markov-based algorithms to predict next location of users proposed in [10]. The authors in [11] integrated Markov chain-based predictors with other algorithms to improve prediction performance. Such schemes completely ignore the trace data with poor quality and just consider users with good quality of trace data. Our work consists of estimating the frequently visited locations, and then attaining the hybrid Markov chain model, which adaptively chooses from the first-order or the second order Markov chain, based on the quality of mobility trace to predict future behavior of the users. Due to continued developments in location tracking techniques, forecasting of urban congestion has become an active research topic. In [12] authors proposed to use Autoregressive Integrated Moving Average (*ARIMA*) algorithm to analyze time series data. In [13], the author uses Markov-based algorithm to calculate the probability distribution of overflow queue. The algorithm is able to estimate mean queue and its variance for stationary and non-stationary arrival processes. However, the existing techniques are not well suited to predict time of congestion. Therefore, we utilize the area congestion predictor with a tunable time threshold, which means according to our requirements the algorithm can determine time of users' congestion in scales of *seconds* or *minutes*.

III. System Model

The system overview is depicted in Figure 1. The proposed system model involves three main layers: Common Zone of Interest (*C-ZOI*) Discovery, Individual Zone of Interest (*I-ZOI*) Prediction and Common Zone of Interest (*C-ZOI*) Congestion Prediction. The relevant notations and system component definitions are explained in the following subsections.

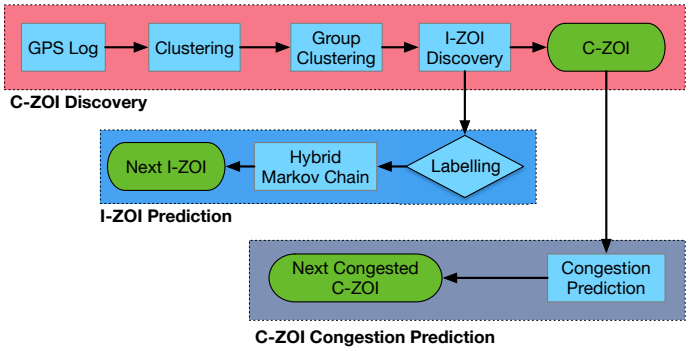

Fig. 1: Overview of the System Model.

A. Locations and Places

Mobile devices have the potential to track users' visited locations using the Global Positioning System (*GPS*). The device regularly collects a user's raw location logs as a list $L = [loc_1, loc_2, loc_3, ..., loc_n]$, where $loc_i = (\alpha, \beta, t, v)$ is a tuple representing a location point in the format $(latitude, longitude, timestamp, velocity)$. The rest of the paper uses the notations $loc.\alpha$, $loc.\beta$, $loc.t$ and $loc.v$ for the location point elements.

B. Common Zone of Interest (C-ZOI) Discovery

Intuitively, an Individual Zone of Interest (*I-ZOI*) is a cluster group that is frequently visited by a user and a Common Zone of Interest (*C-ZOI*) depicts a region covering multiple I-ZOIs. Conceptually, each C-ZOI represents a region, where several users are interested to visit frequently and spend considerable amount of time. In order to extract the C-ZOIs of users based on the location history L, we must first introduce the notations of cluster and cluster group. A cluster is a set of visited location points with similar temporal (*e.g., visiting time*) and spatial (*e.g., instantaneous velocity*) features. A cluster group is an aggregation of intersected clusters.

1) Cluster Discovery: A cluster represents a subset of successive location points in L, which are confined with similar temporal and spatial features. Δd_{max}, e_{max}, e_{min} and $v \in \mathbb{R}$ represents the distance expressed in meters, maximum velocity, minimum velocity and instantaneous velocity expressed in meters per second, respectively. In addition, $\Delta t_{min} \in \mathbb{N}$ is a time duration expressed in minutes. We introduce

two functions: $ClusterCentroid\left([loc_1, loc_2, loc_3, ..., loc_n]\right)$, to compute the centroid of visited location points, and $Distance\left([loc_i, loc_j]\right)$, which measures the Euclidean distance between the two locations loc_i and loc_j. A subset $l \subseteq L$ becomes a cluster if the following conditions in Equation 1, 2, 3 and 4 are met:
$\forall loc_i, loc_{i+1} \in l$:

$$Distance\left(centroid\left(loc_1, ..., loc_i\right), loc_{i+1}\right) \leq \Delta d_{max} \quad (1)$$

$$loc_n t - loc_1 t \geq \Delta t_{min} \quad (2)$$

$$e_{min} \leq \sum_{i=1}^{n} \frac{v_i}{n} \leq e_{max} \quad (3)$$

$$\nexists l' \neq \quad : \quad l \subset l' \quad (4)$$

A cluster is a 4-item tuple $c = (\alpha_c, \beta_c, r, l)$, where α_c and $\beta_c \in \mathbb{R}$ are the latitude and longitude coordinates of the centroid, $r \in \mathbb{R}$ is it's radius in meters, and $l \in L$ is the subset of locations belonging to c. The average of all $loc.\alpha$ and $loc.\beta$ of the locations contained in the subset l is the centroid (α_c, β_c) of the cluster, which is designated as $c.centroid$. We introduce C, the set of clusters extracted from the location log of a user as $C = \{c_1, c_2, c_3, ..., c_n\}$. Disjointness of discovered clusters can not be guaranteed by equations 1, 2, 3 and 4. Therefore, construction of a cluster group is required and explained in the next step.

2) Cluster Group Discovery: A cluster group includes a set of overlapping clusters. Thus, we define equation 5 to check whether two clusters c_i, $c_j \in C$ are intersected or not.

$$Distance\left(c_i.centroid, c_j.centroid\right) - \left(c_i r + c_j r\right) < 0 \quad (5)$$

A cluster group is a 4-item tuple $cg = (\alpha_{cg}, \beta_{cg}, r, \{c_1, c_2, c_3, ...\})$, where α_{cg}, β_{cg} and $r \in \mathbb{R}$, $\{c_1, c_2, c_3, ...\} \in C$ are latitude, longitude, radius and array of clusters constituting g respectively. $(\alpha_{cg}, \beta_{cg})$ represents the centroid of the cluster group, which is the mean of all the clusters formed g, and r must be compared to enclose all the individual clusters present in g. G contains the n cluster groups belonging to a user as $G = \{cg_1, cg_2, cg_3, ..., cg_n\}$

3) Individual Zone of Interest (I-ZOI): An I-ZOI refers to a frequently visited region by a user during daily activities. We define two constants $minCountThreshold$ and $maxTimeDifference \in \mathbb{N}$ representing the minimum threshold of visits and the maximum time difference threshold between two consecutive visits, respectively. Then, $CountVisits(cg)$ is a function that counts the number of clusters included in cluster group cg, and $timeDuration(G)$ is a function that returns the duration between two consecutive visited dates of cluster group cg in G. A cluster group $cg \in G$ is transformed into an I-ZOI z if the conditions of Equation 6 and 7 are met:

$$CountVisits(cg) \geq minCountThreshold \quad (6)$$

$$timeDifference(G) \leq maxTimeDifference \quad (7)$$

An I-ZOI z is a 6-item tuple $z = (\alpha_z, \beta_z, r, ID_{zone}, \{g_1, g_2, g_3, ..., g_n\}, T_{ID})$, where α_z, β_z and $r \in \mathbb{R}$, $ID_{zone} \in \mathbb{N}$ and $\{g_1, g_2, g_3, ..., g_n\} \in G$ are the latitude, longitude, radius, Zone-ID and group clusters becoming an I-ZOI. T_{ID} represents visiting dates of the I-ZOI by each user. The set Z is finally the set of I-ZOIs of the user such that $Z = \{z_1, z_2, z_3, ..., z_n\}$. As shown in Figure 2, the sets of discovered cluster groups are depicted by intersecting yellow circles. Finally, the cluster groups that could satisfy above conditions are considered as I-ZOIs.

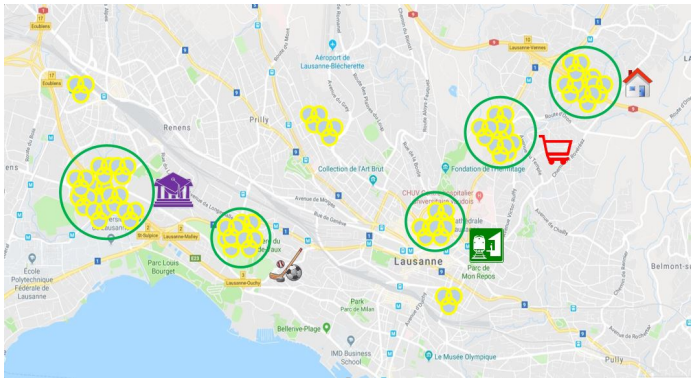

Fig. 2: I-ZOI construction from cluster groups of location points.

4) Common Zone of Interest (C-ZOI): A C-ZOI is an aggregation of adjacent I-ZOIs. We introduce a constant $distanceThreshold \in \mathbb{N}$ to represent the maximum threshold of distance between each I-ZOIs. $Distance\left(I - ZOI_i centroid, I - ZOI_j centroid\right)$ is a function to compute the Euclidean distance between the $I - ZOI_i$ and $I - ZOI_j$. I-ZOIs are grouped whenever the following condition in Equation 8 is satisfied:

$$Distance\left(I - ZOI_i centroid, I - ZOI_j centroid\right) \leq distanceThresold \quad (8)$$

A C-ZOI is a 6-item tuple $C - ZOI = (\alpha_{cz}, \beta_{cz}, r, ID_{cz}, \{ZOI_1, ZOI_2, ZOI_3, ..., ZOI_n\}, T_{ID})$, where α_{cz}, β_{cz} and $r \in \mathbb{R}$, $ID_{cz} \in \mathbb{N}$ and $\{ZOI_1, ZOI_2, ZOI_3, ..., ZOI_n\} \in Z$ are the latitude, longitude, radius, C-ZOI-ID and group of I-ZOIs, respectively. The last item of the tuple indicates visiting dates of the C-ZOI by each user. Finally, we introduce CZ, which contains the n C-ZOIs belonging to users as $CA = \{C - ZOI_1, C - ZOI_2, C - ZOI_3, ..., C - ZOI_n\}$. Figure 3 shows an example of extracted C-ZOIs for a group of users. Each C-ZOI encapsulates a set of nearby I-ZOIs.

C. Individual Zone of Interest (I-ZOI) Prediction

This module predicts the users' future locations (*I-ZOI*). The mobility model represents the movement of mobile users and how their locations change over time. Collected mobility traces of users during their movements could be used to explore some sort of regularities in their daily life. This knowledge is utilized by a mobility predictor to forecast locations that a user may visit in future. The proposed mobility prediction

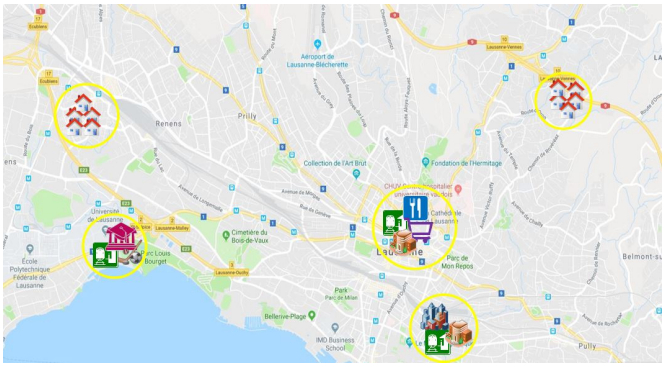

Fig. 3: C-ZOIs construction from I-ZOIs.

scheme in this paper is based on a hybrid Markov chain model, which adaptively switches between the first-order or the second-order Markov chain, depending on the availability and quality of collected user traces [14]. The proposed hybrid Markov model benefits from both the first-order Markov chain [15] and the second-order Markov chain. The rationale behind using a hybrid predictor is that the standard first-order Markov chain algorithms are memoryless models [14], which means that the mobility predictor only benefits from current temporal (*time and day of week*) and spatial (*location*) to predict next movement. However, the second-order Markov chain model benefits from the previous state in addition to the current state to predict future location. Actually, this information is really beneficial: if a user is currently at a city center, e.g., a restaurant, knowing whether he/she was at work or at home just before greatly helps in estimating his/her future behavior. However, we observe for some users trace data with discrete gaps (*ranging from a few seconds to a few minutes*). In these cases the 2-order state information conditions will not be met, which led to poor performance for the second-order Markov predictor [14].

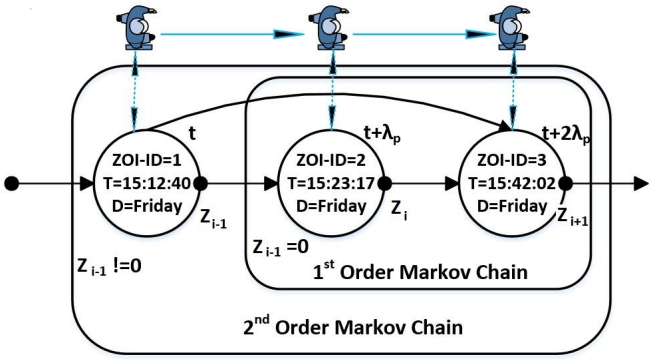

Fig. 4: Hybrid Markov Chain.

The proposed hybrid model is illustrated in Figure 4, in which a Markov chain state consists of a time step and an I-ZOI ID. Equation 9 defines the calculation of a future location probability, in which Z_i represents an I-ZOI with ID *i*, D indicates the day of the week (*e.g., Saturday*), T_i defines the time of the day D (*e.g., 13:22:43 h*), and λ_p determines the future time interval.

$$Pr(Z_{i+1}(t+\lambda_p)) = \begin{cases} Pr(Z_{i+1}|Z(t+\lambda_p)=Z_i, \\ \quad D, T(t+\lambda_p)=T_i) & Z_{i-1}=0 \\ Pr(Z_{i+1}|Z(t)=Z_{i-1}, Z(t+\lambda_p)=Z_i, \\ T(t)=T_{i-1}, T(t+\lambda_p)=T_i, D) & Z_{i-1}\neq 0 \end{cases} \tag{9}$$

Equation 9 can be considered as a location-dependent distribution and a time-dependent distribution (as expressed in Equations 10 and 13). As shown in Equation 13, both the time-dependent and location-dependent distributions in the second-order Markov chain model benefit from current and previous state information (*e.g., time, day and location*). The location dependent distribution can be modeled as a Mobility Markov Chain (*MMC*). The MMC is described by a transition matrix, which includes discovered I-ZOIs per each single user and all the calculated transitions between them. The mobility predictor obtains these transitions during training steps by considering a tunable time threshold (*e.g., each minutes*) on the given days of the week (*e.g., all Wednesdays, all Thursdays and etc.*). The second-order Markov chain calculates transition probabilities among states only if two successive states are present in traced data.

For the case of the second-order Markov chain, the counting of transition frequency happens only when the user's movement is continuously following the sequence of two states.

$$Pr(Z_{i+1}|Z(t+\lambda_p)=Z_i, T(t+\lambda_p)=T_i, D) \tag{10}$$
$$=Pr(Z_{i+1}|Z(t+\lambda_p)=Z_i) \tag{11}$$
$$+Pr(T(t+\lambda_p)=T_i, D) \tag{12}$$

$$Pr(Z_{i+1}|Z(t)=Z_{i-1}, Z(t+\lambda_p)=Z_i, T(t)=T_{i-1}, T(t+\lambda_p)=T_i, D) \tag{13}$$
$$=Pr(Z_{i+1}|Z(t)=Z_{i-1}, Z(t+\lambda_p)=Z_i) \tag{14}$$
$$+Pr(T(t)=T_{i-1}, T(t+\lambda_p)=T_i, D) \tag{15}$$

D. Common Zone of Interest (C-ZOI) Congestion Prediction

From the probability distribution of the users' future visited I-ZOIs, we can further estimate the number of users that may visit and stay in a C-ZOI together at a future moment. Therefore, next, we target at predicting the probability distribution of the number of users that may visit together a specific C-ZOI within a given time. As explained in Section III, nearby I-ZOIs are covered by C-ZOIs. In these regions users either do not move or they move very slowly and users are spending a considerable amount of time together in each C-ZOI. Each of these hot spot regions are candidates to host a significant number of users (*pedestrians*). Then, we define the *AreaCongestionThreshold (M)*, which refers to the number of predicted users in each C-ZOI. The congestion prediction model counts the number of predicted users in each C-ZOI. If the number of users exceeds the defined threshold, we assume that this region will experience congestion within the

next λ_c minutes. Predicting the number of users will help us to facilitate tasks such as resource management, logistic administration, and urban planning. For instance, if we know how many users will be in a specific C-ZOI between time t and $t + \lambda_c$ we could optimize placement of resources (*e.g., bandwidth allocation, public transportation*) in the city or dynamically adapt those resources, while taking into account the number of users. Equation 16 defines the probability of having M users visiting $C - ZOI_i$ at time $t + \lambda_c$, which is derived from the estimated number of upcoming and outgoing of users in $C - ZOI_i$ at time $t + \lambda_c$. The parameters used in this equation are listed in Table I.

$$\begin{aligned} P\left\{N_{C-ZOI_i}(t+\lambda_c)=M\right\} = \\ \sum_m P\left\{N_{C-ZOI_i}(t+\lambda_c)=M \mid N_{C-ZOI_i}(t)=m\right\} \times \\ P\left\{N_{C-ZOI_i}(t)=m\right\} = \\ \sum_m \langle \sum_{n_1,n_2,n_1-n_2=\triangle m} P\left\{N_{in,C-ZOI_i}(t+\lambda_c)=n_1\right\} \times \\ P\left\{N_{out,C-ZOI_i}(t+\lambda_c)=n_2\right\}\rangle \times P\left\{N_{C-ZOI_i}(t)=m\right\} \end{aligned} \quad (16)$$

In Equation 16, $P\left\{N_{in,C-ZOI_i}(t+\lambda_c)\right\}$ describes the probability of having n_1 users that may move into $C-ZOI_i$ at time $t+\lambda_c$ and $P\left\{N_{out,C-ZOI_i}(t+\lambda_c)\right\}$ indicates the probability of having n_2 users may moving out from $C-ZOI_i$ at time $t+\lambda_c$. These probabilities can be calculated using Equation 17.

$$\begin{aligned} P\left\{N_{in,C_i}(t+\lambda_c)=n_1\right\} = \sum_{A_1 \in F_{C_{i'}(t)}} \prod_{j_1 \in A_1} P_{j_1} \prod_{j_1' \in A_1^c} \left(1-P_{j_1'}\right) \times \\ P\left\{N_{out,C_i}(t+\lambda_c)=n_2\right\} \\ = \sum_{A_2 \in F_{C_i(t)}} \prod_{j_2 \in A_2} P_{j_2} \prod_{j_2' \in A_2^c} \left(1-P_{j_2'}\right) \end{aligned} \quad (17)$$

IV. Evaluation

In this section, we present an evaluation methodology to validate the proposed user mobility and area congestion prediction models.

1) Dataset: In order to train mobility and congestion prediction algorithms, accumulated traced data of users' movements is needed. In this research, we are relying on collected trace data in the Nokia Mobile Data Challenge (*MDC*) dataset. [16]. This dataset contains records of almost 180 smartphones carried by residents around the lake Geneva in Switzerland. The data records on which our work is based cover a duration over 17 months from October 2009 to March 2011. The basic demographic documents show that the participants are mostly young individuals and university students [16]. The dataset includes data generated from sensors and applications, such as visited locations (*GPS coordinates*), movement (*instantaneous velocity*), proximity (*Bluetooth*), communication (*Cell-IDs, WLAN-IDs*), etc. However, for mobility and congestion predictions, GPS coordinates of visited places and corresponding time stamps are required, which are nearly 10 million location points. To evaluate prediction performance of both mobility and congestion predictors, we divided collected mobility trace data of each single user to two parts: (i) dataset (L): which contains 70% of data as learning dataset. (ii) dataset (T): which contains the rest of traced data (30%) as testing dataset. The learning dataset is used to obtain states for both algorithms and to determine their transition probability matrix. The testing dataset is used to test and evaluate the accuracy of the proposed prediction algorithms.

TABLE I: Area congestion prediction algorithm parameters.

Parameter Name	Parameter Definition
Z_i	State i in the Markov chain
$Pr\left\{Z_{i+1}(t)\right\}$	Probability of at State $(i+1)$ at t
$C-ZOI_i$, U_j	C-ZOI ID i, User ID j
D, T_i	Weekday and time of being at State i
λ_c, t	Future time interval and current time
$C=\{C-ZOI_1,..,i\}, \lvert C\rvert = I$	Set of C-ZOIs, I is the total numbers
$U=\{User_1,..,j\}, \lvert U\rvert = J$	Set of Users, J is the total numbers
$N_{C-ZOI_i}(t)$, $N_{C-ZOI_i}(t+\lambda_c)$	Number of Users in C-ZOI i at time t and t+λ_c (*e.g.,* m and M)
$N_{in,C-ZOI_i}(t+\lambda_c)$	Number of Users that may move to C-ZOI i at time t+λ_c (*e.g.,* n1)
$N_{out,C-ZOI_i}(t+\lambda_c)$	Number of Users that may move from C-ZOI i at time t+λ_c (*e.g.,* n2)
$F_{C-ZOI_i(t)}$	Subset of all users in $C-ZOI_i$ at time t
$F_{C-ZOI_{i'}(t)}$	Subset of all users out of $C-ZOI_i$ at time t
P_{j_1} , $User_{j_1} \in F_{C-ZOI_{i'}(t)}$	$P\{User_{j_1}$ *is in* $F_{C-ZOI_{i'}(t)}$ *at time* $t\}$ $\times$ $P\{User_{j_1}$ *moves to* $C-ZOI_i$ *at time* t+$\lambda_c\}$
P_{j_2} , $User_{j_2} \in F_{C-ZOI_i(t)}$	$P\{User_{j_2}$ *is in* $F_{C-ZOI_i(t)}$ *at time* $t\}$ $\times$ $P\{User_{j_2}$ *moves from* $C-ZOI_i$ *at time* t+$\lambda_c\}$

2) User Trace Quality: The quality of collected traces depends on the behavior of pedestrians. Some users keep the smartphone with themself everyday. However, others sometimes forgot to carry the devices or had to charge them, such that data recordings are non-continuous. As learned from our previous experiences [17] [18], the number of valid states (*with a time stamp and I-ZOI ID*) in the derived hybrid Markov chain for each user depends on the quality of data trace in each day. Therefore, we first classify the dataset into two groups (*good or poor quality*) based on the number of recorded instances during the whole data collection period. We choose five users with good quality of trace data (*e.g., 500000-400000 records*) and five users with poor quality of trace data (*e.g., 250000-350000 records*).

3) Evaluation Metrics: Prediction accuracy measures the accuracy of the location prediction algorithm. For each single user we select states out of all the Markov chain states derived for each particular weekday (*e.g., Friday*) from the training dataset L for each user. Afterwards, the prediction algorithm is performed for each of the selected states to estimate the possible future visited I-ZOI(s) for mobility prediction in the next λ_p minutes. We check the transition probability for states during the same period of time in the testing data set T as well. Afterwards, the Mean Absolute Error (*MAE*) of the possible transitions of the corresponding testing points is calculated according to Equation 18. To evaluate performance of the area congestion predictor we define two metrics: (i) density of users, which counts the number of users that may move to each C-ZOI; (ii) area congestion prediction accuracy, which

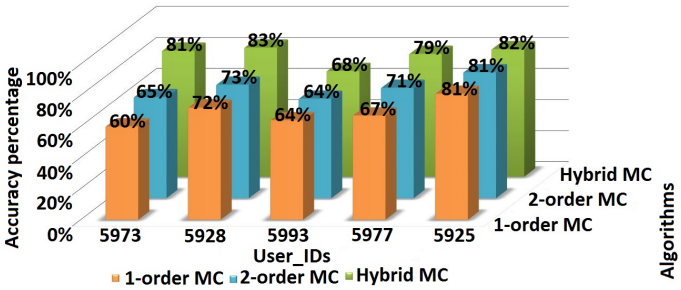

Fig. 5: Prediction accuracy for users with good quality.

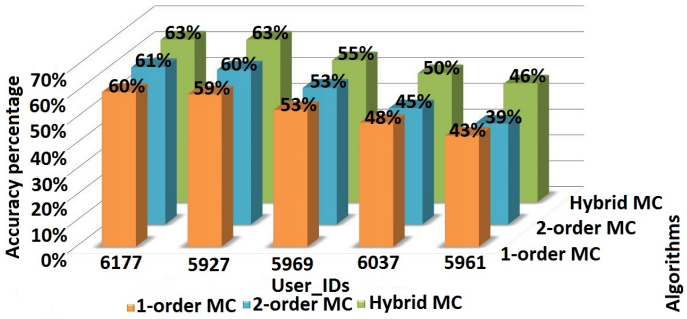

Fig. 6: Prediction accuracy for users with poor quality.

represents probability of moving users to a C-ZOI in a specific day of week and is calculated by average of future location prediction accuracies of users in each C-ZOI derived from Equation 18.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} \mid Pr_iL - Pr_iT \mid, Accuracy = (1 - MAE) \times 100 \quad (18)$$

4) Experimental Settings: We describe the experimentation parameters of the discussed clustering, mobility prediction, and congestion prediction algorithms. In order to determine the parameters we analyze traced data for users with at least 10 months duration of collected data. Then, we read the data-points sequentially according to the recorded time stamps. Table II shows the experiment parameters and the associated values in our assessment.

TABLE II: Experiments parameters.

Parameter	Definition	Value
Δd_{max}	Maximum Euclidean distance between centroid of location points and next location	60 m
Δt_{min}	Minimum time threshold of staying in each location point	15 min
e_{max}	Maximum instantaneous velocity threshold	50 *m/s*
e_{min}	Minimum instantaneous velocity threshold	0
M	Number of predicted users in each C-ZOI	6
minCountTh	Minimum number of visits of each cluster group	60
maxTimeDiff	Maximum time difference between two consecutive visits of a cluster group by a user	24 *h*
DistanceTh	Maximum Euclidean distance between two I-ZOIs	500 *m*
λ_p	Time threshold for hybrid-MC algorithm	1 *min*
λ_c	Time threshold for area congestion prediction algorithm	15 *min*

V. Evaluation Results

1) Mobility Prediction Accuracy Results: This subsection details the prediction accuracy results of the proposed hybrid predictor, the first order and the second order Markov chain. We first present the average prediction accuracy of all the users with different trace qualities. Then, we discuss more details about the prediction accuracy per day, for users with poor and good trace qualities.

Figures 5 and 6 show the prediction accuracy of different MMC predictors for users with good and poor quality of mobility traces. We define two categories of quality depending on the number of instances recorded in a user's movement traces. We randomly choose 5 User IDs (5973, 5928, 5993, 5977, 5925) from the group of good quality trace data, and

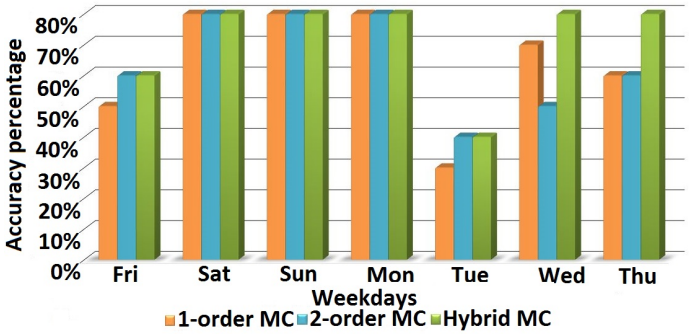

Fig. 7: Prediction accuracy per day for User-5928.

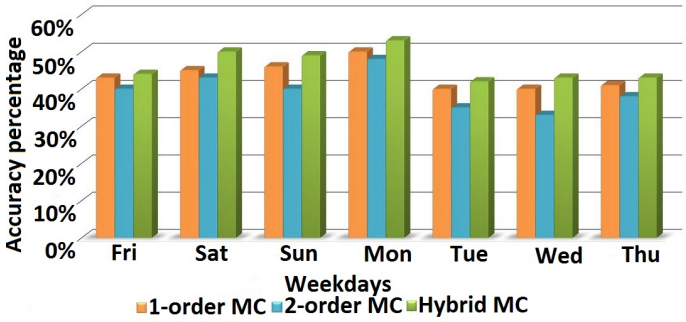

Fig. 8: Prediction accuracy per day for User-6037.

5 User IDs (6177, 5927, 5969, 6037, 5961) from the group of poor quality trace data. As we can see from Figure 5, the hybrid predictor can deliver an average prediction accuracy over all weekdays and weekends of nearly 83% for User-ID 5928. Moreover, it can be observed from Figure 6 that the estimated accuracy is improved significantly when the hybrid predictor used for users with poor quality trace data. For instance, it delivers an average accuracy of 63% for User-IDs 6177 and 5927. The results clearly demonstrate that the hybrid predictor outperform others, while using the traced data with either poor or good quality. Figures 7 and 8 show the prediction accuracy of three different predictors for each day. This helps us to explain the advantages of the hybrid predictor compared to the first-order and second-order Markov chains. From defined user categories we randomly select User-ID 5928 and User-ID 6037 as the representatives of users with good and poor qualities. The graphs show that the hybrid predictor performs better than the first-order and the

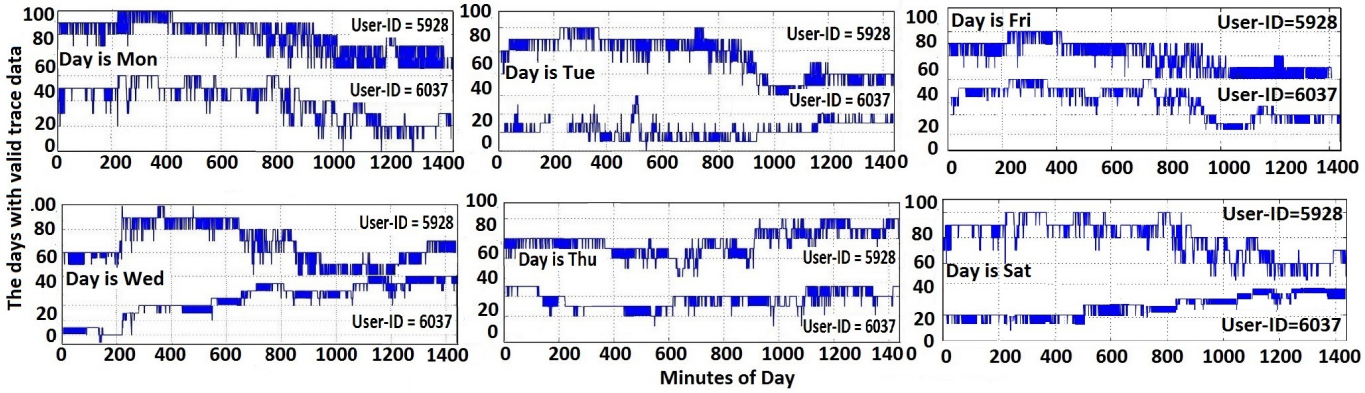

Fig. 9: Trace qualities of User_IDs 5928/6037 on weekdays.

second-order MC predictors for both categories. To explain the performance differences of mobility predictors for these two users, we next discuss the data quality of User-ID 5928 and 6037. Figure 9 depicts mobility traces of users over a year, shown as a matrix, where each column is a minute of the day and each line indicates the number of days with valid trace data (*with time stamp and GPS coordinates*). We map each interval of valid records to continuous pulses, and leave blank intervals time during which we do not have information about users' locations. To count the number of days with valid trace records, we introduce a *threshold*, which counts the days with more than 1500 records as valid days for prediction. If a user has less than 1500 records in one day, the data of that specific day will not be included in the prediction. This is because such a low number of records happen most probably because of network connectivity issues or defective sensors. Therefore, collected traces for these users are not valid and should not be included in the prediction procedure. Figure 6 illustrates that for User-ID 6037, the hybrid predictor can only deliver an accuracy of around 52% for Monday. This situation arises typically because location data are partly available. For User-ID 5928, due to having continuous intervals of collected GPS records at a high number of days with valid trace data between 80 to 100, the hybrid predictor has improved performance (81% *to* 83%) for all weekdays.

2) Congestion Prediction Accuracy Results: In addition to estimating future locations of mobile users, we are also interested in area congestion prediction. In this subsection, we present the prediction accuracy of the congestion prediction algorithm. Then we discuss more details about the number of predicted users in each C-ZOI.

We examine the predictability of congestion by employing recorded GPS coordinates of all available users in the dataset. We focus on the extracted C-ZOIs in the city of Lausanne by predicting the number of users that may move and stay together in each common hot spot. Figure 10 depicts the results of the congestion prediction algorithm for time-of-days (*08:00 h, 12:00 h, 18:00 h and 22:00 h*). The graph shows that the congestion prediction algorithm achieves accuracies exceeding 70% for C-ZOIs. In addition to congestion prediction accuracy, we also count the number of users in each C-ZOI for time-of-

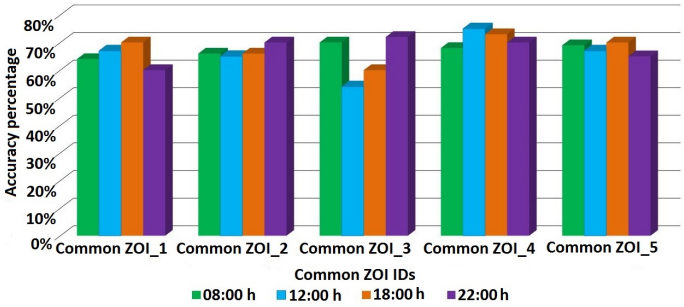

Fig. 10: Area congestion prediction accuracy.

days (*08:00 h, 12:00 h, 18:00 h and 22:00 h*). Figure 11 shows the density of pedestrians for different hours. We observe that in the evening the users have a tendency to travel towards the city suburbias (*C-ZOI 1, C-ZOI 5*), going back home for dinner. An inverse behavior is detected at 08:00 AM and 12:00 PM, when the most of flows are toward the universities or city center (*C-ZOI 2, C-ZOI 3 and C-ZOI 4*). Although we can not compare these population densities against a proper ground truth, we remark that the model represents very reasonable results that match well to the movements of inhabitants in the city of Lausanne.

VI. CONCLUSIONS

With the explosive growth of location-based service on mobile devices, predicting users' future locations is of increasing importance to support proactive information services. In this paper, we introduce a hybrid predictor to estimate future locations of a user. Further, we propose a technique to discover hot spot regions for users by relying on spatial and temporal constraints. The achieved results over real world mobility traces validate our proposed algorithms, which achieve more than 81% correct predictions for users. More important, we present a novel approach to predict congestion in hot spot regions using GPS coordinates. This achieves accuracies exceeding 70% for discovered C-ZOIs from the available dataset.

For future enhancements we will concentrate on predicting trajectories of mobile users that they use for transition among I-ZOIs. Furthermore, to foster our hybrid predictor we are planning to employ a time series-based periodicity detection

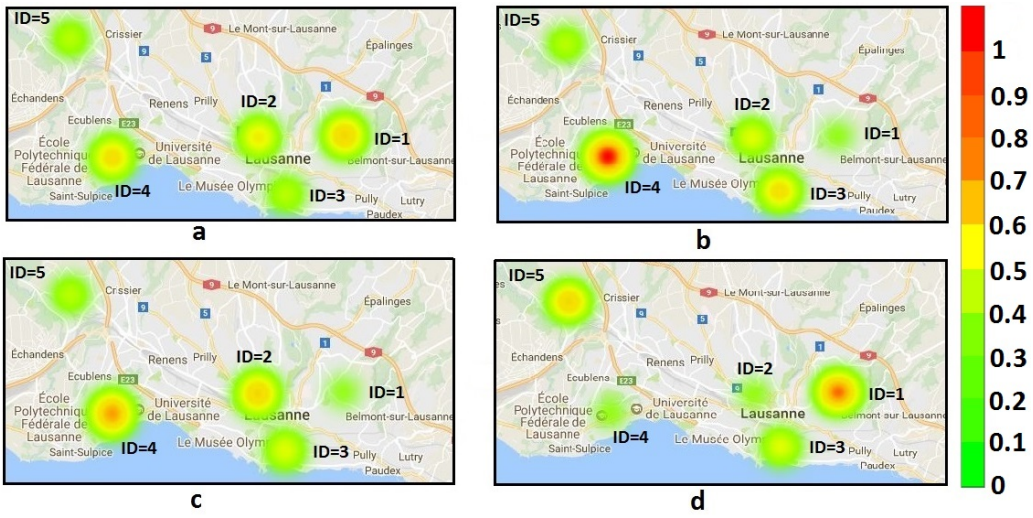

Fig. 11: Density of users in each C-ZOI. a) Wednesday at 08:00 h. b) Wednesday at 12:00 h. c) Wednesday at 18:00 h. d) Wednesday at 22:00 h.

algorithm to recognize variations in pedestrians' behaviors, and apply the appropriate mobility predictor accordingly. We will also conduct extensive practical experiments to compare our mobility predictor with other Markov chain-based algorithms on other large scale datasets.

ACKNOWLEDGMENT

This work has been supported by the Swiss National Science Foundation with project number 154458.

REFERENCES

[1] G. Khodabandelou, V. Gauthier, M. A. El-Yacoubi, and M. Fiore, "Population estimation from mobile network traffic metadata," *CoRR*, vol. abs/1610.06947, 2016. [Online]. Available: http://arxiv.org/abs/1610.06947

[2] Z. Zhao, M. Karimzadeh, F. Gerber, and T. Braun, "Mobile crowd location prediction with hybrid features using ensemble learning," *Future Generation Computer Systems*, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167739X17318058

[3] X. Chen, J. Pang, and R. Xue, "Constructing and comparing user mobility profiles," *ACM Trans. Web*, vol. 8, no. 4, pp. 21:1–21:25, Nov. 2014. [Online]. Available: http://doi.acm.org/10.1145/2637483

[4] T. N. Maeda, K. Tsubouchi, and F. Toriumi, "Next place prediction in unfamiliar places considering contextual factors," in *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ser. SIGSPATIAL'17. ACM, 2017, pp. 76:1–76:4.

[5] R. Montoliu and D. Gatica-Perez, "Discovering human places of interest from multimodal mobile phone data," in *Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia*, ser. MUM '10. ACM, 2010, pp. 12:1–12:10.

[6] J. J.-C. Ying, W.-C. Lee, T.-C. Weng, and V. S. Tseng, "Semantic trajectory mining for location prediction," in *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ser. GIS '11. New York, NY, USA: ACM, 2011, pp. 34–43. [Online]. Available: http://doi.acm.org/10.1145/2093973.2093980

[7] N. Bui, M. Cesana, S. A. Hosseini, Q. Liao, I. Malanchini, and J. Widmer, "A survey of anticipatory mobile networking: Context-based classification, prediction methodologies, and optimization techniques," *IEEE Communications Surveys Tutorials*, vol. 19, no. 3, pp. 1790–1821, thirdquarter 2017.

[8] L. Song, D. Kotz, R. Jain, and X. He, "Evaluating next-cell predictors with extensive wi-fi mobility data," *IEEE Transactions on Mobile Computing*, vol. 5, no. 12, pp. 1633–1649, Dec 2006.

[9] Y. Qiao, Z. Si, Y. Zhang, F. B. Abdesslem, X. Zhang, and J. Yang, "A hybrid markov-based model for human mobility prediction," *Neurocomputing*, vol. 278, pp. 99–109, 2018. [Online]. Available: https://doi.org/10.1016/j.neucom.2017.05.101

[10] M. H. Sun and D. M. Blough, "Mobility prediction using future knowledge," in *Proceedings of the 10th ACM Symposium on Modeling, Analysis, and Simulation of Wireless and Mobile Systems*, ser. MSWiM '07. New York, NY, USA: ACM, 2007, pp. 235–239. [Online]. Available: http://doi.acm.org/10.1145/1298126.1298167

[11] M. Chen, X. Yu, and Y. Liu, "Mining moving patterns for predicting next location," *Information Systems*, vol. 54, pp. 156 – 168, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0306437915001295

[12] L. Li, Y. Wang, G. Zhong, J. Zhang, and B. Ran, "Short-to-medium term passenger flow forecasting for metro stations using a hybrid model," *KSCE Journal of Civil Engineering*, Jul 2017. [Online]. Available: https://doi.org/10.1007/s12205-017-1016-9

[13] P. S. Olszewski, "Modeling probability distribution of delay at signalized intersections," *Journal of advanced transportation*, vol. 28, no. 3, pp. 253–274, 1994.

[14] Z. Zhao, L. Guardalben, M. Karimzadeh, J. Silva, T. Braun, and S. Sargento, "Mobility prediction-assisted over-the-top edge prefetching for hierarchical vanets," *IEEE Journal on Selected Areas in Communications*, pp. 1–1, 2018.

[15] M. Karimzadeh Motallebi Azar, Z. Zhao, L. Hendriks, R. de Oliveira Schmidt, S. la Fleur, H. van den Berg, A. Pras, T. Braun, and M. Julian Corici, *Mobility and Bandwidth Prediction as a Service in Virtualized LTE Systems*. IEEE, 10 2015, pp. 132–138, 10.1109/CloudNet.2015.7335295.

[16] I. Yaqoob, I. A. T. Hashem, A. Gani, S. Mokhtar, E. Ahmed, N. B. Anuar, and A. V. Vasilakos, "Big data," *Int. J. Inf. Manag.*, vol. 36, no. 6, pp. 1231–1247, Dec. 2016. [Online]. Available: https://doi.org/10.1016/j.ijinfomgt.2016.07.009

[17] B. Sousa, Z. Zhao, M. Karimzadeh Motallebi Azar, D. Palma, V. Fonseca, P. Simoes, T. Braun, H. van den Berg, A. Pras, and L. Cordeiro, *Enabling a Mobility Prediction-aware Follow-Me Cloud Model*. United States: IEEE Computer Society, 11 2016, pp. 486–494, eemcs-eprint-27394.

[18] Z. Zhao, M. Karimzadeh Motallebi Azar, T. Braun, A. Pras, and H. van den Berg, *Cloudified Mobility and Bandwidth Prediction in Virtualized LTE Networks*. IEEE, 7 2017.